Math/Stat 416, Fall 2020, Lecture Notes[1]

Rodrigo Bañuelos
Purdue University

August 31, 2020

[1] **Acknowledgment:** This is standard material that can be found in most introductory text books in probability. The notes are an expanded version of some notes I got from Phanuel Mariano during his time at Purdue. I believe his notes are based on Lecture Notes of Richard Bass.

Contents

Chapter 1

Week 1: Basic Principles of Counting

1.1 Combinatorics

Probability theory is a fundamental mathematical tool used in science, engineering, social sciences, business, medicine, and in many aspects of day-to-day life. It is part of our common experience and intuition. Basic questions we always ask are: what is the probability that such and such event will occur?" This course provides the tools to put these these questions in solid mathematical foundation.

Probability gives the precise notion of *proportionality.* In order to understand the notion of *"picking something at random"* we need a sophisticated way to count different possible outcomes of experiments. These techniques come from the field of *Combinatorics* widely use in many areas of discrete mathematics and its applications, especially in computer science.

1.2 Counting

Definition 1.2.1: Basic Counting Principle

The basic counting principle will be fundamental to all our work in this course. It states that if one experiment can result in any of m possible outcomes and if another experiment can result in any of n possible outcomes, then there are

$$\boxed{m \times n = m \cdot n} \tag{1.2.1}$$

possible outcomes for both experiments.

Think of these as boxes (box method). Fort each of the m outcome of box 1 there are n outcomes of box 2. Resulting in a total of mn outcomes. Can also think of it as an $m \times n$ matrix, where we say that the outcome is (i, j) if experiment 1 results in its ith possible outcome and experiment 2 results in its jth possible outcome. Hence, the set of possible outcomes consists of m rows, each containing n elements which is an $m \times n$ matrix with m rows and n columns. The toral number of entries is $m \cdot n$.

1.2.1 Examples and general counting

Example 1.2.1. *A community consists of 10 couples each of whom has 4 children. If one couple and one of their children are to be chosen as couple and child of the year, how many different choices are possible?*

Solution. *By regarding the choice of the couple as the outcome of the first experiment and the subsequent choice of one of her children as the outcome of the second experiment, we see from the basic principle that there are* $10 \times 4 = 40$ *possible choices.*

Example 1.2.2. *There are 20 teachers and 100 students in a school. How many ways can we pick a teacher and student of the year?*

Solution. *Same as previous problem. Again, box method (or matrix method) gives* $20 \cdot 100 = 2000$.

Example 1.2.3. *A fair die is rolled and the number on the top face is noted. Then another fair die is rolled, and the number on its top face is noted. What is the total number of outcomes for the pair of dice?*

Solution. *Thinking of each possible outcome as a pair of numbers we have a* 6×6 *matrix with 36 entries.*

$$\begin{array}{rl}
Row1: & (1,1),\ (1,2),\ (1.3),\ (1,4),\ (1,5),\ (1,6) \\
Row2: & (2,1),\ (2,2),\ (2,3),\ (2,4),\ (2,5),\ (2,6) \\
Row3: & (3,1),\ (3,2),\ (3,3),\ (3,4),\ (3,5),\ (3,6) \\
Row4: & (4,1),\ (4,2),\ (4,3),\ (4,4),\ (4,5),\ (4,6) \\
Row5: & (5,1),\ (5,2),\ (5,3),\ (5,4),\ (5,5),\ (5,6) \\
Row6: & (6,1),\ (6,2),\ (6,3),\ (6,4),\ (6,5),\ (6,6)
\end{array}$$

As we shall see, there are many question we can ask whose answers follow trivially from this table.

- What is the proportion of outcomes where the sum of the two numbers showing is 5?
- What is the proportion of outcomes where one of the dice shows 2, and the other shows 4?
- What is the proportion of outcomes where the second number rolled is greater than the first number?

Example 1.2.4. *How many functions defined on n points are possible if each functional value is either 0 or 1?*

Solution. *Let the points be* $1, 2, \ldots, n$*. Since* $f(i)$ *must be either* 0 *or* 1 *for each* $i = 1, 2, \ldots, n$ *there are* 2^n *possible functions.*

Example 1.2.5. *How many different* $6-$ *places license plates are possible with 2 letters followed by 4 numbers?*

Solution. *Thinking of the set of 2 letters and the set of 4 numbers as the two experiments we see that the box method gives* $(26)^2 \cdot (10)^4$.

Last example leads us to the generalized counting principle.

Definition 1.2.2: Generalized Counting Principle

If k experiments are to be performed with the number of possible outcomes being n_i for the ith experiment, then there are

$$\boxed{n_1 \times n_2 \times \cdots \times n_k = n_1 \cdot n_2 \cdots n_k} \tag{1.2.2}$$

possible outcomes for all k experiments.

1.2.2 Examples

Example 1.2.6. *A college planning committee consists of 3 freshmen, 4 sophomores, 5 juniors, and 2 seniors. A subcommittee of 4, consisting of 1 person from each class, is to be chosen. How many different subcommittees are possible?*

Solution. *There are four separate experiments of choosing a single representative from each of the classes. From the generalized version of the basic principle that there are* $3 \times 4 \times 5 \times 2 = 120$ *possible subcommittees.*

Example 1.2.7. *How many different* 7*-place license plates are possible if the first 3 places are to be occupied by letters and the final 4 by numbers?*

Solution. *By the generalized version of the basic principle, the answer is* $26 \cdot 26 \cdot 26 \cdot 10 \cdot 10 \cdot 10 \cdot 10 = 175,760,000$

Example 1.2.8. *In the previous example, how many license plates would be possible if repetition among letters or numbers is not permitted?*

Solution. *In this case, there would be* $26 \cdot 25 \cdot 24 \cdot 10 \cdot 8 \cdot 7 = 78,624,000$ *possible license plates.*

1.3 Permutations

Definition 1.3.1: Permutations

A permutation of n objects is an ordered sequence of those n objects. Two permutations only differ according to the order of the objects

How many different ordered arrangements of the letters a, b, and c are possible? We can list them:

$$\boxed{abc, acb, bac, bca, cab, cba}$$

Each arrangement is known as a permutation. Thus, there are 6 possible permutations of a set of 3 objects. Could also have been obtained from the basic principle, since the first object in the permutation can be any of the 3, the second object in the permutation can then be chosen from any of the remaining 2, and the third object in the permutation is then the remaining 1. Thus, there are

$$\boxed{3 \cdot 2 \cdot 1 = 6}$$

possible.

1.4 General Counting

With the same argument we just used for 3 letters we obtain

Definition 1.4.1: Counting Permutations

Suppose we have n objects. The number of permutations of these objects is

$$\boxed{n(n-1)(n-2)\cdots 3 \cdot 2 \cdot 1 = n!} \tag{1.4.1}$$

Recall that the quantity $n!$ is called “n-factorial” and that 0! is define to be 1.

1.4.1 Examples, counting general permutations

Example 1.4.1. *What is the number of possible batting orders with 9 players?*

Solution. *9! = 362880*

Example 1.4.2. *A soccer team is composed of 11 players. If each of the 11 players is assigned to one (and only one) position on the field of play, how many different configurations of team can the coach make up?*

Solution. *There are* 11! *permutations of the 11 players.*

Example 1.4.3. *How many ways can we arrange 4 math books, 3 chemistry books, 2 physics books, and 1 biology book on a bookshelf so that all the math books are together, all the chemistry books are together, and all the physics books are together.*

Solution. *We can arrange the math books in* $4!$ *ways, the chemistry books in* $3!$ *ways, the physics books in* $2!$ *ways, and the biology book in* $1! = 1$ *way which gives* $4! \cdot 3! \cdot 2! \cdot 1!$. *But we also have to decide which set of books go on the left, which next, and so on. That is the same as the number of ways of arranging the letters* M, C, P, B, *and there are* $4!$ *ways of doing that. This gives* $4! \cdot (4! \cdot 3! \cdot 2! \cdot 1!) = 6912$.

Example 1.4.4. *How many different letter arrange can be form the letters* a, a, b, c*?*

Solution. *Let us label them as* A, a, b, c. *There are* $4!$*, or 24, ways to arrange these letters. But there are repeats. We could have* Aa *or* aA. *So we have a repeat for each possibility. Thus the answer should be* $4!/2! = 12$.

Example 1.4.5. *In the same way, if there were 3 a's, 4 b's, and 2 c's, we would have*

$$\frac{9!}{3!4!2!} = 1260.$$

Example 1.4.6. *How many different letter arrangements can be formed from the letters PEPPER?*

Solution. *The letter* P *is repeated 3 times and the letter* E *is repeated 2 times. The total number of different arrangements is* $\frac{6!}{3!2!} = 60$

Definition 1.4.2: Counting General Permutations

There are

$$\boxed{\frac{n!}{n_1! \cdot n_2! \cdots n_k!}} \tag{1.4.2}$$

permutations of n objects, of which n_1 are alike, n_2 are alike, . . . , n_k are alike.

Two permutations only differ according to the order of the objects. Note: Short notation for product of k-numbers $a_1, a_2, \ldots a_k$ is

$$\prod_{j=1}^{k} a_j = a_1 \cdot a_2 \cdots a_k$$

1.5 Combinations, binomial and multinomial coefficients

Definition 1.5.1: Definition–Combinations

A combination of k objects among n objects is a none-ordered subset of k objects. Property: Two combinations only differ according to nature of their objects

Definition 1.5.2: Counting Combinations–Binomial Coefficients

For $0 \leq k \leq n$, the number of different groups of k objects chosen from a total of n objects is given by

$$\boxed{\binom{n}{k} = \frac{n!}{k!(n-k)!} = \frac{n(n-1)\cdots(n-k+1)}{k!}} \tag{1.5.1}$$

Remark 1.5.1. *The terminology for* $\binom{n}{k}$ *is "n choose k". These numbers are also called the "binomial coefficients" for reasons we shall see soon! Note that*

$$\binom{n}{n} = \binom{n}{0} = \frac{n!}{0!n!}$$

as it should be. Indeed, in a set of n objects there is exactly one subset of size n (the set itself) and exactly one subset of size 0 *(the empty set).*

The solution to the following example gives the proof of formula (1.5.1).

1.5.1 Examples

Example 1.5.1. *How many different groups of 3 can we select from the 5 items A, B, C, D, and E?*

Solution. *There are 5 ways to select the initial item, 4 ways to select the next item, and 3 ways to select the final item for total of xxxx ways of selecting the group of 3 when the order in which the items are selected is*

In general: The formula

$$n(n-1)\cdots(n-k+1)$$

gives the number of different ways that a group of k items can be selected from n items when the order of selection is relevant. Since each group of k items will be counted $k!$ times in this count, if we divide this by $k!$. This gives (1.5.1).

Example 1.5.2. *A committee of 5 is to be formed from a group of 20 people. How many different committees are possible?*

Solution. *By the formula with $n = 20$ and $k = 1$ we have*

$$\binom{20}{5} = \frac{20!}{5! \cdot 15!} = \frac{20 \cdot 19 \cdot 18 \cdot 17 \cdot 16}{5!}$$

Example 1.5.3. *From a group of 7 women and 5 men, how many different committees consisting of (i) 3 women and 2 men can be formed? (ii) What if 2 of the men can't stand each other and their presence of both in the committee will be extremely disruptive–hence cannot serve together*

Solution. *(i) Groups of 3 women=*$\binom{7}{3} = \frac{7 \cdot 6 \cdot 5}{3 \cdot 2 \cdot 1}$*, and groups of 2 men=*$\binom{5}{2} = \frac{5 \cdot 4}{2 \cdot 1}$ *and from the counting principle we have*

$$\frac{7 \cdot 6 \cdot 5}{3 \cdot 2 \cdot 1} \cdot \frac{5 \cdot 4}{2 \cdot 1} = 350$$

committees.
(ii) Out of the $\binom{5}{2} = 10$ *groups of 2 men there exactly 1 or* $\binom{2}{2}$ *containing both unpleasant characters and therefore there are* $10 - 1 = 9$ *groups that do not contain both men. There are still* 35 *groups of 3 women, and so the total number of committees is* $35 \cdot 9 = 315$‘.

Example 1.5.4. *A person has* 8 *friends, of whom* 5 *will be invited to a party. (We've all been through this)*

(i) How many choices are there if 2 *of the friends are feuding and will not attend together?*

(ii) How many choices if 2 *of the friends will only attend together?*

Solution. *Observe that*

(i) We can remove the two friends and invite five of the remaining 6 friends or invite one of the two fiends and from the other six invite 4. (Remember that as with the problem of the "4 digit password" that or means we add.) We have

$$\binom{6}{5} + \binom{2}{1} \cdot \binom{6}{4}$$

(ii) For this case, we can remove both friends from the group again and invite all 5 from the remaining 6 or invite both friends and then invite 3 from the remaining 6. This gives

$$\binom{6}{5} + \binom{2}{2} \cdot \binom{6}{3}$$

1.6 The Binomial Theorem

Theorem 1.6.1: The Binomial Theorem

$x, y \in \mathbb{R}$, $n \geq 1$,

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

Hence the reason why $\binom{n}{k}$ are called the binomial coefficients.

Proof. (Counting argument) To see this, the left hand side is $(x+y)(x+y)\cdots(x+y)$. This will be the sum of 2^n terms, and each term will have n factors. How many terms have k x's and $(n-k)$ y's? This is the same as asking: in a sequence of n positions, how many ways can one choose k of them in which to put x's? (Box it) The answer is $\left(\begin{array}{c} n \\ k \end{array} \right)$, so the coefficient of $x^k y^{(n-k)}$ should be $\left(\begin{array}{c} n \\ k \end{array} \right)$. ■

1.6.1 Examples

Example 1.6.1. *Suppose*

- *we have a set A with n elements. That is, if we use the notation $|A|$ for the number of elements in A, then $|A| = n$.*
- *Let $\mathcal{P}_n$=collection of all subsets of A_n. Then*

$$|\mathcal{P}_n| = 2^n$$

Solution. *Write*

$$\begin{aligned} |\mathcal{P}_n| &= \sum_{k=0}^{n} |\textit{subsets of } A \textit{ with } k \textit{ elements}| && (1.6.1) \\ &= \sum_{k=0}^{n} \binom{n}{k} = (1+1)^n = 2^n, && (1.6.2) \end{aligned}$$

where we have applied the binomial Theorem with $x = y = 1$.

Example 1.6.2. *Expand* $(x+y)^3$.

Solution. $(x+y)^3 = y^3 + 3xy^2 + 3x^2y + x^3$.

Example 1.6.3. *Verify the identity using counting arguments–no algebra*

$$\binom{10}{4} = \binom{9}{3} + \binom{9}{4}$$

Solution. *The LHS represents the number of committees having 4 people out of the 10. Let's pick one individual from this group (does not matter who) and call her/him x. The number of committees is the sum of those committees where x is a member and those where x is not a member. The first is $1 \cdot \binom{9}{3}$ while the second is $\binom{9}{4}$.*

Example 1.6.4. *The more general formula*

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

is verified in the same way.

This is called "Pascal's identity." It can be use to give a proof of the binomial theorem using mathematical induction. This is done in your book.

1.7 Multinomial Coefficients

Properties: Combinations–Multinomial Coefficients

The number of ways to divide a n objects into one group of n_1 objects, one group of n_2 objects ... and kth group of n_k objects, where $n = n_1 + n_2 + \cdots + n_k$ is equal to

$$\boxed{\binom{n}{n_1, n_2, \ldots, n_k} = \frac{n!}{n_1! n_2! \cdots n_k!}} \qquad (1.7.1)$$

The numbers

$$\binom{n}{n_1, n_2, \ldots, n_k}$$

are called the multinomial coecienciets for reasons we shall see below.

Proof. Use the formulas from above.

- Number of choices for the ith group:

$$\binom{n - \sum_{j=1}^{i-1} n_j}{n_j}$$

- Number of divisions: # Divisions of n objects into k groups of size $n_1, \ldots, n_k$

$$\prod_{j=1}^{k} \binom{n - \sum_{j=1}^{i-1} n_j}{n_j} = \binom{n}{n_1, n_2, \ldots, n_k}$$

The last equality follows by "telescoping the product." Note that $(n - \sum_{j=1}^{i-1} n_j = n$ for $i = 1$.)

■

1.7.1 Examples

Example 1.7.1. *A police department in a small city consists of 10 officers. If the department policy is to have 5 of the officers patrolling the streets, 2 of the officers working full time at the station, and 3 of the officers on reserve at the station, how many different divisions of the 10 officers into the 3 groups are possible?*

Solution. *There are* $\frac{10!}{5!2!3!} = 2520$ *possible divisions.*

Example 1.7.2. *Ten children are to be divided into an A team and a B team of 5 each. The A team will play in one league and the B team in another. How many different divisions are possible?*

Solution. *there are* $\frac{10!}{5!5!} = 252$ *possible divisions.*

Example 1.7.3. *In Example 1.4.2, suppose there are 4 main categories of positions: goalkeeper, defensive, midfield, and attacking positions. Assume that the coach wants 1 goalkeeper, 4 defensive players, 3 midfield players and 3 attacking players. In how many ways can the coach assign each of the 11 players to one of the 4 categories?*

Solution. *This is a perfect application of the multinomial coefficient: we want to split the 11 players into 4 different groups. This can be done in* $\binom{11}{1,4,3,3} = \frac{11!}{1!\,4!\,3!\,3!} = 46,200$

Example 1.7.4. *The game of bridge is played by 4 players, each of whom is dealt 13 cards. How many bridge deals are possible?*

Solution. *Recall that a deck of cards has 52 cards. By* (1.7.1) *the number of bridge deals possible is*

$$\binom{52}{13, 13, 13, 13}$$

Example 1.7.5. *If 8 new teachers are to be divided among 4 schools, (i) how many divisions are possible? (ii) What if each school must receive 2 teachers?*

Solution. *(i) Assuming all teachers are distinct, the general counting principle gives* $(8)^4$*=4096. (ii) The multinomial equation* (1.7.1) *gives*

$$\binom{8}{2, 2, 2, 2} = \frac{8!}{(2!)^4} = 2520.$$

1.8 The Multinomial Theorem

Theorem 1.8.1: The Multinomial Theorem

$$(x_1 + x_2 + \cdots x_k)^n = \sum_{\{(n_1, n_2, \ldots, n_k) : n_1 + n_2 + \cdots + n_k = n\}} \binom{n}{n_1, n_2, \ldots, n_k} x_1^{n_1} x_2^{n_2} \cdots , x_k^{n_k}, \quad (1.8.1)$$

for all $x_j \in \mathbb{R}$, $n \geq 1$.
The sum is over all vectors $(n_1, n_2, \ldots, n_k)$ with nonnegative integer components such that their sum is n.

Hence the reason why $\binom{n}{n_1, n_2, \ldots, n_k}$ are called the multinomial coefficients. "bi" for two "multi" for many. This is not proved here.

Two useful applications of these notions are the following. These are Exercise* 1.3 and Exercise* 1.4 in your book. Will not prove these. You may use freely when appropriate.

Proposition: Proposition

(i) There are

$$\binom{n-1}{k-1}$$

distinct positive integer-valued vectors satisfying

$$x_1 + x_2 + \cdots + x_k = n, \quad x_j > 0, \; j = 1, 2, \ldots, k. \quad (1.8.2)$$

(ii) There are

$$\binom{n+k-1}{k-1}$$

distinct positive integer-valued vectors satisfying

$$x_1 + x_2 + \cdots + x_k = n, \quad x_j \geq 0, \; j = 1, 2, \ldots, k. \quad (1.8.3)$$

1.8.1 Examples

Example 1.8.1. *An elevator starts at the basement with 8 people (not including the elevator operator) and discharges them all by the time it reaches the top floor, number 6. In how many ways could the operator have perceived the people leaving the elevator if all people look alike to him?*

What if the 8 people consisted of 3 men and 5 women and the operator could tell a man from a woman?

Solution. *Think of the problem in terms of nonnegative solutions. Let x_i, $i = 1, 2, 3, 4, 5, 6$ be the number of people that get off on floor i.*

(a) We need the number of solutions to $x_1 + \cdots + x_6 = 8$ with $x_i \geq 0$. We use the formula with $n = 8$, $k = 6$ to get

$$\binom{13}{5}.$$

(b) In this case we have

$$(\text{\# solutions of } x_1 + \cdots + x_6 = 5) \cdot (\text{\# solutions of } x_1 + \cdots + x_6 = 3) = \binom{10}{5}\binom{8}{5}.$$

Example 1.8.2. *We have 20 thousand dollars that must be invested among 4 possible opportunities. Each investment must be integral in units of 1 thousand dollars, and there are minimal investments that need to be made if one is to invest in these opportunities. The minimal investments are 2, 2, 3, and 4 thousand dollars. How many different investment strategies are available if (a) an investment must be made in each opportunity? (b) investments must be made in at least 3 of the 4 opportunities?*

Solution. *As in the previous problem we have:*

(a)

$$x_1 + x_2 + x_3 + x_4 = 20,$$

but this time with $x_1 \geq 2$, $x_2 \geq 2$, $x_3 \geq 3$, $x_4 \geq 4$. Make the change of variables (which does not affect the number of investments)

$$y_1 = x_1 - 1,\ y_2 = x_2 - 1,\ y_3 = x_3 - 2,\ y_4 = x_4 - 3.$$

We now need to find the number of solutions for

$$y_1 + y_2 + y_3 + y_4 = 13,\ y_i > 0.$$

With $n = 13$ and $k = 4$ we have $\binom{12}{3} = 220$ possible investment strategies.

Note that we could have used

$$y_1 = x_1 - 2,\ y_2 = x_2 - 2,\ y_3 = x_3 - 3,\ y_4 = x_4 - 4.$$

and compute the solutions for

$$y_1 + y_2 + y_3 + y_4 = 9,\ y_i \geq 0$$

which gives the same answer.

(b) Same reasoning: $\binom{15}{2}$ *investments in 1, 2, 3,* $\binom{14}{2}$ *investments in 1, 2, 4,* $\binom{13}{2}$ *investments in 1, 3, 4,* $\binom{13}{2}$ *investments in 2, 3, 4. If the statement asked for any three investments, we would just add these quantities. However, it asks for at* <u>*at least 3*</u> *of the 4 opportunities, and hence we also need to include the answer to (a). We get*

$$\binom{15}{2} + \binom{14}{2} + 2\binom{13}{2} + \binom{12}{3}$$

Chapter 2

Week 2: Basic notions in probability

2.1 Set operations

Definition 2.1.1

By a set, call it S, we mean a collection of distinct objects. The individual objects in the set S are called elements. These are denoted by x, y, z, etc. A set A is a subset of S, we write $A \subset S$, if every element of A is an element of S. That is, if $x \in A$, then $x \in S$. Given two subsets $A \subset S$ and $B \subset S$ of S, we define:

- Union: $A \cup B = \{x \in S : x \in A \,\text{or}\, x \in B\}$–elements in S that are in A or B or both.

- Intersection: $A \cap B = \{x \in S : x \in A \,\text{and}\, x \in B\}$–elements in both A and B.

- Complement: $A^c = \{x \in S : x \notin A\}$–elements not in A.

- Empty set: The subset of S with no elements is called the empty set and denoted by $\varnothing$.

- The empty set $\varnothing$ in the operation of unions is similar to the 0 in the operation of sums of numbers. That is $A \cup \varnothing = A$ just as $a + 0 = a$. Similarly, S in the operation of intersection is similar to the 1 in the operation of multiplications of numbers. That is, $A \cap S = A$ as $a \times 1 = a$.

- Note that $\varnothing^c = S$ and $S^c = \varnothing$.

- Other common notation (labeling) of universal sets S used in probability are Ω, $\mathcal{U}$, etc..

Remark 2.1.1. *Definitions extend to any number of sets: $A_1, \ldots, A_n$.*

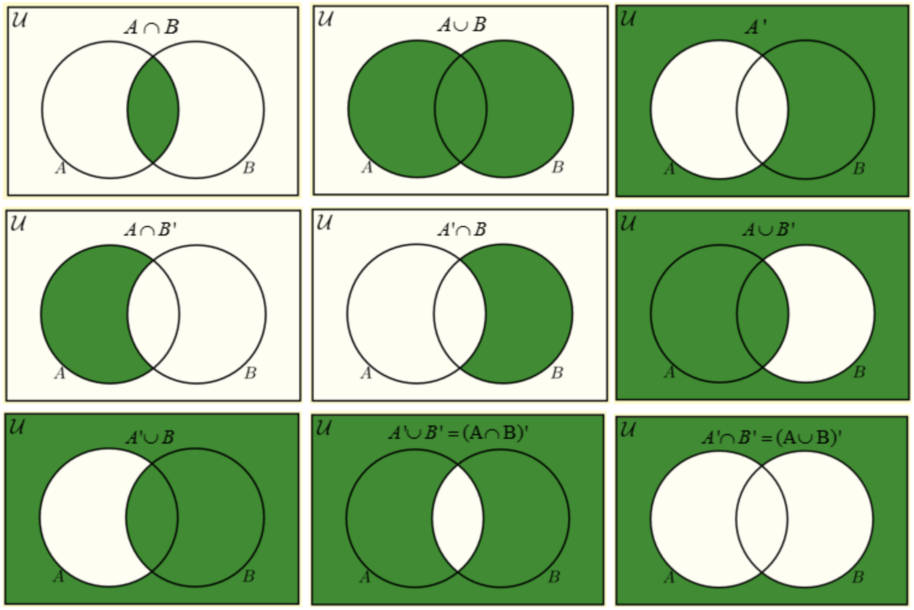
$$\bigcup_{i=1}^{n} A_i = \{x \in S : x \in A_j, \mathit{for\ any} j = 1, \ldots .n\}$$

and

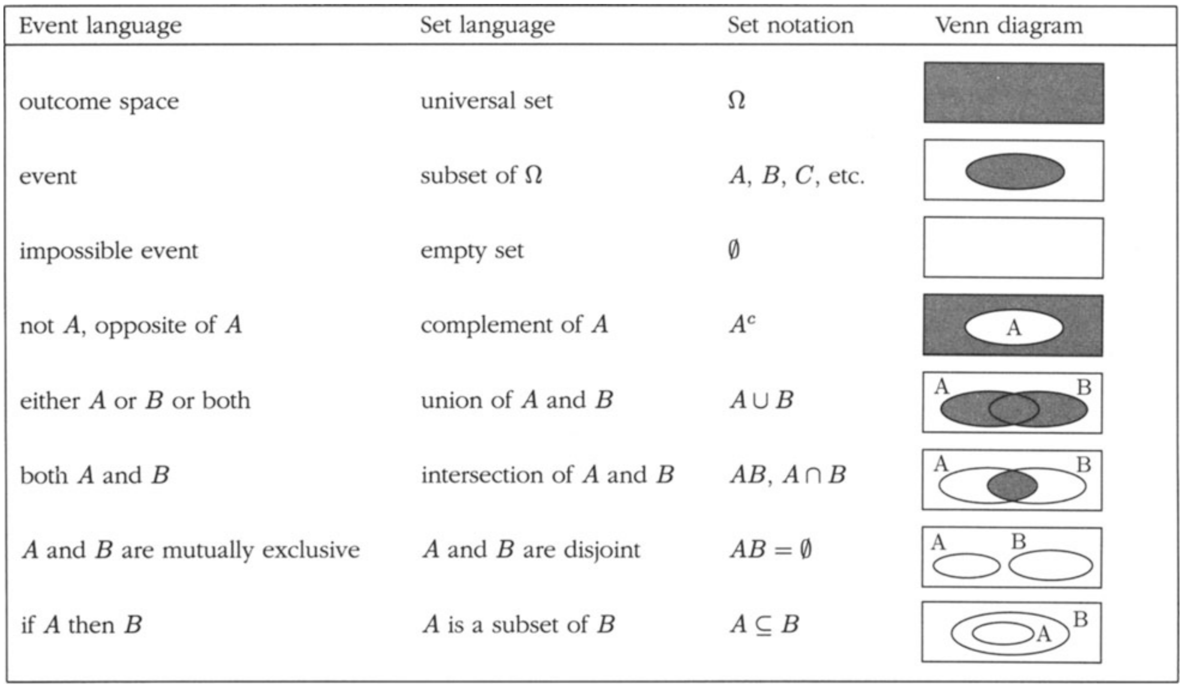
$$\bigcap_{i=1}^{n} A_i = \{x \in S : x \in A_j, \mathit{for\ all} j = 1, \ldots .n$$

the elements in all the sets.

Here are a couple of Venn diagrams to help you visualize the various operations.



In this diagram, $S = \mathcal{U}$ and $A' = A^c$.



Event language	Set language	Set notation	Venn diagram
outcome space	universal set	Ω	
event	subset of Ω	A, B, C, etc.	
impossible event	empty set	$\emptyset$	
not A, opposite of A	complement of A	A^c	A
either A or B or both	union of A and B	$A \cup B$	A B
both A and B	intersection of A and B	AB, $A \cap B$	A B
A and B are mutually exclusive	A and B are disjoint	$AB = \emptyset$	A B
if A then B	A is a subset of B	$A \subseteq B$	A B

Here, $S = \Omega$ and $AB = A \cap B$, as indicated.

Here is another summary of notation:

Summary notation and vocabulary

$a \in A$: a is an element of a set A.

$A \subseteq B$: A is a subset of set B; every element of A is an element of B.

$A \subsetneq B$: $A \subseteq B$ and $A \neq B$; A is a **proper** subset of B.

$A = B$: $A \subseteq B$ and $B \subseteq A$.

$A \cap B$: the set of all elements that are in A and in B.

A and B are **disjoint**: $A \cap B = \emptyset$.

$A \cup B$: the set of all elements that are either in A or in B.

$A \setminus B$: the set of all elements of A that are not in B.

A^c: the set of all elements in the universal set that are not in A.

Here are some examples of various sets of real numbers often used:

Summary of example sets, and their notation

$\emptyset = \{\}$: the set with no elements.

$\{a, b, c\}$, $\{a, b, \ldots, z\}$.

$\{x : x$ can be written as a sum of three consecutive integers$\}$.

$\mathbb{N}$: the set of all **N**atural numbers. Depending on the book, this could be the set of all positive integers or it could be the set of all non-negative integers. The symbols below are unambiguous:

$\mathbb{N}_0$: the set of all non-negative integers;

$\mathbb{N}^+$: the set of all positive integers.

$\mathbb{Z}$: the set of all integers ("**Z**ahlen" in German).

$\mathbb{Q}$: the set of all rational numbers (**Q**uotients).

$\mathbb{R}$: the set of all **R**eal numbers.

$\mathbb{C}$: the set of all **C**omplex numbers (more about them starts in Section 3.1).

Properties 2.1.1: Commutative, Associate and Distributive Laws of sets

- $A \cup B = B \cup A$ and $A \cap B = B \cap A$ (commutative)
- $(A \cup B) \cup C = A \cup (B \cup C)$ and $(A \cap B) \cap C = A \cap (B \cap C)$ (associative)
- $A \cup (B \cap C) = (A \cup B) \cap (B \cup C)$ and $A \cap (B \cup C) = (A \cap C) \cup (A \cap C)$ (distributive).

Properties 2.1.2: DeMorgan's Laws

Given two subsets A and B of S we have

- $(A \cup B)^c = A^c \cap B^c$ (**DeMorgan's Law of unions**–(Draw a picture).
- $(A \cap B)^c = A^c \cup B^c$ (**DeMorgan's Law of intersections**–(Draw a picture).
- More generally: $A_1, \ldots, A_n$: $(\cup_{i=1}^n A_i)^c = \cap_{i=1}^n A_i^c$ and $(\cap_{i=1}^n A_i)^c = \cup_{i=1}^n A_i^c$.

Proof. To prove that two sets, say E and F, are equal, we use their definitions to show that if we take any $x \in E$, then $x \in F$. This shows $E \subset F$. We also need to show that if we take any $x \in F$, then $x \in E$. This shows that $F \subset E$. The two containments show that $E = F$.

Using this we illustrate the proof DeMorgan's Law of unions. Let $x \in (A \cup B)^c$. This means that x is not in their union so is not in A and not in B. That is, $x \in A^c$ and $x \in B^c$. The latter is the same as $x \in A^c \cap B^c$. Thus, $(A \cup B)^c \subset A^c \cap B^c$. In the same way we show that $A^c \cap B^c \subset (A \cup B)^c$. ■

2.2 Sample spaces and events

The above definitions and properties of sets can apply in generality. In probability we work with special sets and special subsets call Sample Spaces and Events.

Definition 2.2.1

- A **Sample Space**, denoted it by S (sometimes Ω, or $\mathcal{U}$), consists of all possible outcomes of an experiment.
- An **event** E is a subset of S. That is, $E \subset S$.
- The **collection of all events** is usually denoted by $\mathcal{F}$.
- We say that events A and B are **mutually exclusive** if $A \cap B = \varnothing$. That is, if the sets are disjoint.

Remark 2.2.1. *The collection of events $\mathcal{F}$ has the property that both $\varnothing$ and S are in $\mathcal{F}$ and whenever $A_1, A_2 \dots$ are each in $\mathcal{F}$, their union and intersection are in $\mathcal{F}$. This is called a σ-filed in your book.*

2.2.1 Examples of sample spaces

Example 2.2.1. **Role a Die:** *The Sample Space is $S = \{1, 2, 3, 4, 5, 6\}$. The sample space has 6 elements. An event could be, for example, $A = 3$ or $A = \{2, 5, 6\}$.*

Example 2.2.2. **Toss a coin twice.** *The sample space is $S = \{HH, HT, TH, TT\}$ containing 4 elements. $A = \{HH, HT\}$ is an event.*

Example 2.2.3. **Roll two dice.** *The Sample Space: $S =$ would be all possible pairs made up of the numbers one through six. $S = \{(i, j) : i, j = 1, ...6\}$.. More explicitly,*

$$
\begin{aligned}
S = \quad & \big\{(1,1),\ (1,2),\ (1.3),\ (1,4),\ (1,5),\ (1,6), \\
& (2,1),\ (2,2),\ (2,3),\ (2,4),\ (2,5),\ (2,6), \\
& (3,1),\ (3,2),\ (3,3),\ (3,4),\ (3,5),\ (3,6), \\
& (4,1),\ (4,2),\ (4,3),\ (4,4),\ (4,5),\ (4,6), \\
& (5,1),\ (5,2),\ (5,3),\ (5,4),\ (5,5),\ (5,6), \\
& (6,1),\ (6,2),\ (6,3),\ (6,4),\ (6,5),\ (6,6)\big\}
\end{aligned}
$$

Example 2.2.4. *If the outcome of an experiment is the order of finish in a race among the 5 horses, say horse 1, horse 2, . . . , horse 5. Then*

$$S = \{\textit{all } 5! \textit{ permutations of } (1, 2, 3, 4, 5)\}.$$

The outcome $(3, 4, 2, 1, 5)$ *means horse 3 came first, horse 4 second, horse 2 third, and so on.*

Example 2.2.5. Car Accidents. *Measuring the number of accidents of a random person in their life time. Sample Space* $S = \{0, 1, 2, \ldots\}$.

Example 2.2.6. Others

- *Let* S *be the possible price of some stock at closing time* $S = [0, \infty)$.
- *The age at which someone dies,* $S = [0, \infty)$.

2.2.2 Examples of events

Example 2.2.7. *Let* E *be event that two dice come up even and equal. That is,* $E = \{(2, 2), (4, 4), (6, 6)\}$ *and* F *the event that the sum of the two dice is* 8*. That is,* $F = \{(2, 6), (3, 5), (4, 4), (5, 3), (6, 2)\}$*. Then*

$$E \cup F = \{(2, 2), (2, 6), (3, 5), (4, 4), (5, 3), (6, 2), (6, 6)\}$$

and $E \cap F = \{(4, 4)\}$. F^c *is all the* 31 *other ways that does not include*

$$\{(2, 6), (3, 5), (4, 4), (5, 3), (6, 2)\}.$$

Example 2.2.8. *Let* $S = [0, \infty)$ *the age someone dies.* $A =$ *the event a person dies before they reached* 30*. Then* $A = [0, 30)$*. The set* $A^c = [30, \infty)$ *represents the event that the person dies on or after their thirty birthday. If we let* $B = (15, 45)$*, then* $A \cup B = [0, 45)$ *and* $A \cap B = (15, 30)$.

Example 2.2.9. *Let* S *be the sample space in Example 2.2.4. Let*

$$E = \{\textit{all outcomes in } S \textit{ starting with a } 2\}.$$

Then

$$E = \{\textit{event that horse number 2 wins the race}\}.$$

2.3 Axioms of probability

Definition 2.3.1: Axioms of Probability

A probability $\mathbb{P}$ is a rule (a "function") which assigns a number (chances to occur) to any event E in a Sample Space S that satisfies the following **axioms:**

(i) For any $E \subset S$,
$$0 \leq \mathbb{P}(E) \leq 1,$$

(ii)
$$\mathbb{P}(S) = 1, \text{ The probability if the whole space is 1,}$$

(iii) For any sequence of events E_j, $j \geq 1$ which are pairwise disjoint (mutually exclusive), that is $E_i \cap E_j = \varnothing$,
$$\mathbb{P}\left(\bigcup_{j=1}^{\infty} E_j\right) = \sum_{j=1}^{\infty} \mathbb{P}(E_j)$$

Think of $\mathbb{P} : \mathcal{F} \to \mathbb{R}$ as a "function" with "domain" $\mathcal{F}$ and "range" in the closed interval $[0, 1]$ satisfying (ii) and (iii). The triple $(S, \mathcal{F}, \mathbb{P})$ is called a **probability space.**

Remark 2.3.1. *Note that a probability $\mathbb{P}$ acts on subsets of S, not on points of S. However, when x is element of S, we take the subset $\{x\}$ consisting of the single element x. Then it is common to write $\mathbb{P}(x)$ for $\mathbb{P}(\{x\})$. For example, if the experiment is that tossing a coin, then $S = \{H, T\}$. The probability of heads should be written as $\mathbb{P}(\{H\})$, but it is common to see $\mathbb{P}(H)$.*

Example 2.3.1. *Toss a coin. $S = \{H, T\}$*

- *If the coin is fair $\mathbb{P}(\{H\}) = \mathbb{P}(\{T\}) = \frac{1}{2}$. We may write $\mathbb{P}(H) = \mathbb{P}(T) = \frac{1}{2}$.*
- *If the coin is biased then one could have a different assignment of probability such as $\mathbb{P}(H) = \frac{2}{3}, \mathbb{P}(T) = \frac{1}{3}$.*
- *In general, for any $0 < p < 1$ we can imagine a biased coin for which $\mathbb{P}(H) = p$ and $\mathbb{P}(T) = 1 - p$.*

Example 2.3.2. *When rolling a die, the probability space consists of $S = \{1, 2, 3, 4, 5, 6\}$, with each element having equal probability $\frac{1}{6}$. Form this we can compute the probability of other events. For example, what is the probability of rolling an even number? This can be done as follows: By property (iii) in the axioms we have*

$$\begin{aligned}\mathbb{P}(\{even\}) &= \mathbb{P}(\{2, 4, 6\}) \\ &= \mathbb{P}(2) + \mathbb{P}(4) + \mathbb{P}(6) = \frac{1}{2}.\end{aligned}$$

2.4 Properties of probabilities

We begin with a simple but extremely useful properties (keep this in mind as you work on problems).

Proposition 2.4.1

(a) $\mathbb{P}(\varnothing) = 0$
(b) If $A_1, \dots, A_n$ are pairwise disjoint (mutually exclusive), $\mathbb{P}\left(\cup_{j=1}^n A_j\right) = \sum_{j=1}^n \mathbb{P}(A_i)$.
(c) $\mathbb{P}(E^c) = 1 - \mathbb{P}(E)$.
(d) If $E \subset F$, then $\mathbb{P}(E) \leq \mathbb{P}(F)$.
(e) $\mathbb{P}(E \cup F) = \mathbb{P}(E) + \mathbb{P}(F) - \mathbb{P}(E \cap F)$, for all E and F.
(f) $\mathbb{P}(E) = \mathbb{P}(E \cap F) + \mathbb{P}(E \cap F^c)$, for all E and F.

Proof. The proof follows by using the axioms. (c)–(f) are easy to show just by looking at van diagrams.
(a) Let $A_j = \varnothing$ for each $j \geq 1$. Of course, this sets are disjoint. By (i), $0 \leq \mathbb{P}(\varnothing) \leq 1$. By (iii),

$$\mathbb{P}(\varnothing) = \mathbb{P}\left(\bigcup_{j=1}^{\infty} A_j\right) = \sum_{j=1}^{\infty} \mathbb{P}(A_j) = \sum_{j=1}^{\infty} \mathbb{P}(\varnothing),$$

Since the left side is nonnegative and less than 1, so is the right hand side. But this can only be possible if each term in the sum is 0. That is, if $\mathbb{P}(\varnothing) = 0$

(b) Let $A_{n+1} = A_{n+2} = \cdots = \varnothing$ so that $\cup_{j=1}^{\infty} A_j = \cup_{j=1}^{n} A_j$ hence

$$\begin{aligned}\mathbb{P}\left(\bigcup_{j=1}^{n} A_i\right) &= \mathbb{P}\left(\bigcup_{j=1}^{\infty} A_j\right) \\ &= \sum_{j=1}^{n} \mathbb{P}(A_j) + \sum_{j=n+1}^{\infty} \mathbb{P}(\varnothing) \\ &= \sum_{j=1}^{n} \mathbb{P}(A_j) + \sum_{j=n+1}^{\infty} 0 \\ &= \sum_{i=1}^{n} \mathbb{P}(A_j)\end{aligned}$$

(c) Use $S = E \cup E^c$. By Axioms (ii) & (iii),

$$1 = \mathbb{P}(S) = \mathbb{P}(E) + \mathbb{P}(E^c),$$

hence $\mathbb{P}(E^c) = 1 - \mathbb{P}(E)$.

(d) If $E \subset F$, write $F = E \cup (F \cap E^c)$, which are disjoint.

$$\mathbb{P}(F) = \mathbb{P}(E \cup (F \cap E^c)) = \mathbb{P}(E) + \mathbb{P}(F \cap E^c) \geq \mathbb{P}(E) + 0 = \mathbb{P}(E).$$

(e) Write $E \cup F = E \cup (E^c \cap F)$, (picture of venn diagram of both). By disjointness

$$\mathbb{P}(E \cup F) = \mathbb{P}(E) + \mathbb{P}(E^c \cap F).$$

Now write F (picture) as $F = (E \cap F) \cup (E^c \cap F)$ and using disjointness

$$\mathbb{P}(F) = \mathbb{P}(E \cap F) + \mathbb{P}(E^c \cap F) \implies \mathbb{P}(E^c \cap F) = \mathbb{P}(F) - \mathbb{P}(E \cap F),$$

substitute into first equation to get

$$\begin{aligned} \mathbb{P}(E \cup F) &= \mathbb{P}(E) + \mathbb{P}(E^c \cap F) \\ &= \mathbb{P}(E) + \mathbb{P}(F) - \mathbb{P}(E \cap F), \end{aligned}$$

as needed.

(f) Note that $E = (E \cap F) \cup (E \cap F^c)$ ("add and subtract F") and the union is disjoint. So, (f) follows. ■

Part (e) of the above Proposition has a generalization to an arbitrary number of sets and has a special name. We state its version for three sets.

Proposition 2.4.2: The Inclusion-Exclusion Identity

For any three events $A, B, C \subseteq S$,

$$\begin{aligned} \mathbb{P}(A \cup B \cup C) = {} & \mathbb{P}(A) + \mathbb{P}(B) + \mathbb{P}(C) \\ & - \mathbb{P}(A \cap B) - \mathbb{P}(A \cap C) - \mathbb{P}(B \cap C) + \mathbb{P}(A \cap B \cap C). \end{aligned} \tag{2.4.1}$$

Proof. This follows by an application (e) by grouping $A \cup B \cup C$ as $A \cup (B \cup C)$ and applying (2) with A and a "new B" as $B \cup C$. ■

2.4.1 Examples of inclusion-exclusion

Here are a couple of examples of applications of the inclusion-exclusion identity.

Example 2.4.1. *Purdue will play Wisconsin next year. With probability* 0.5 *Purdue will win the Home game. With probability* 0.4 *Purdue will win the Away game and with probability* 0.3 *will win both games. What's the probability that Purdue loses both games?*

Solution. *Let A_1 be the event that Purdue wins games 1 and A_2 the event that it wind gave 2. We are given: $\mathbb{P}(A_1) = 0.5$, $\mathbb{P}(A_2) = 0.4$ and $\mathbb{P}(A_1 \cap A_2) = 0.3$. We want to find $\mathbb{P}(A_1^c \cap A_2^c)$. Simplify:*

$$\begin{aligned} \mathbb{P}(A_1^c \cap A_2^c) &= \mathbb{P}((A_1 \cup A_2)^c) \ \text{(DeMorgan's Law)} \\ &= 1 - \mathbb{P}(A_1 \cup A_2). \ \text{(Proposition–part (c))} \end{aligned}$$

Using Proposition 1(e), we have

$$\mathbb{P}(A_1 \cup A_2) = 0.5 + 0.4 - 0.3 = 0.6,$$

Hence $\mathbb{P}(A_1^c \cap A_2^c) = 1 - 0.6 = 0.4$, as needed.

Example 2.4.2. *Mom is taking two books along on her vacation. With probability 0.5, she will like the first book, with probability 0.4, she will like the second book and with probability 0.3, she will like both books. What is the probability that she likes neither book?*

Solution. *Solution. Let B_i denote the event that Mom likes book i, $i = 1, 2$. The event that she likes neither book is $B_1^c \cap B_2^c$. DeMorgan's Law gives*

$$\mathbb{P}(B_1^c \cap B_2^c) = \mathbb{P}((B_1 \cup B_2)^c) = 1 - \mathbb{P}(B_1 \cup B_2).$$

From (e) we have:

$$\mathbb{P}(B_1 \cup B_2) = \mathbb{P}(B_1) + \mathbb{P}(B_2) - \mathbb{P}(B_1 \cap B_2) = 0.5 + 0 + 0.4 - 0.3 = 0.6$$

The probability that she likes neither book $= 0.4$

Example 2.4.3. *In a certain population, 10% of the people are rich, 5% are famous, and 3% are rich and famous. Picked a person at random from this population. (a) What is the chance that the person is not rich? (b) What is the chance that the person is rich but not famous? (c) What is the chance that the person is either rich or famous?*

Solution. *(a)*

$$\mathbb{P}(\textit{not rich}) = 1 - \mathbb{P}(\textit{rich}) = 1 - 0.1 = 0.9, \ \ \textit{or 90\% chance}$$

(b)

$$\mathbb{P}(\textit{rich but not famous}) = \mathbb{P}(\textit{rich}) - \mathbb{P}(\textit{rich and famous}) = 0.1 - 0.03 = 0.07, \ \ \textit{or 7\% chance}$$

(c)

$$\begin{aligned} \mathbb{P}(\textit{rich or famous}) &= \mathbb{P}(\textit{rich}) + \mathbb{P}(\textit{famous}) - \mathbb{P}(\textit{rich and famous}) = 0.1 + 0.05 - 0.03 = 0.12 \\ & \quad \textit{or 12\% chance.} \end{aligned}$$

2.5 The Uniform discrete distribution

Many of the examples of in this course deal with discrete (combinatorial/counting) probability.

Definition 2.5.1: Uniform discrete distribution

In an experiment with finite many outcomes all equally like if S is the sample space and $E \subset S$ is an event, we define the uniform distribution by

$$\mathbb{P}(E) = \frac{\#\,\text{of outcomes in}\,E}{\#\,\text{of outcomes in}\,S}$$

Of course, the number of outcomes in S is the total number of outcomes of the experiment, let say N. Writing

$$S = \{s_1, s_2, \dots s_N\}$$

rthe assumption that each outcome is equally likely means

$$\mathbb{P}\{s_i\} = \frac{1}{N}.$$

If $E \subset S$ is an event,

$$\mathbb{P}(E) = \frac{\#\,\text{of elements in}\,E}{\#\,\text{of elements in}\,S} = \frac{|E|}{|S|} = \frac{|E|}{N}$$

Notation: For any event E (including the whole Sample Space), $|E| = \#\,\text{of elements in}\,E$

2.5.1 Examples

Example 2.5.1. *Roll a die:* $S = \{1,2,3,4,5,6\}$, $|S| = 6 = N$. *Let* $E = \{\textit{even outcomes}\}$. *Then* $\mathbb{P}(E) = \frac{|E|}{6} = \frac{1}{2}$

Example 2.5.2. *Three balls are randomly withdrawn without replacement from a bowl containing* 6 *white and* 5 *blue balls. What is the probability that one ball is white and the other two are blue?*

Solution. *(1) Regard the order in which the balls are selected as relevant.*

$$S = \{\textit{Ordered triples of balls, numbered from 1 to 11}\}$$

$|S| = 11 \cdot 10 \cdot 9 = 990$ *outcomes. The event with one white ball and two blue balls is*

$$E = WBB \cup BWB \cup BBW,$$

where $WBB = \{\textit{outcomes where first ball is white and second two blue}\}$, *and similarly for* BWB *and* BBW. *Since* $|WBB| = 6 \cdot 5 \cdot 4 = 120$, *# of outcomes of first ball white, other two are blue;* $|BWB| = 5 \cdot 6 \cdot 4 = 120$ *# of outcomes of first ball blue, second white, third blue; and*

$BBW = 5 \cdot 4 \cdot 6 = 120$ # of outcomes of first ball first two are blue and the third is white. E is the union of three disjoint events and

$$\mathbb{P}(E) = \frac{|E|}{|S|} = \frac{360}{990} = \frac{4}{11}.$$

(2) Can also regard the outcome of the experiment as the unordered set of drawn balls. Then $S = \{$three objects choses from 11$\}$. Then $|S| = \binom{11}{3} = 165$ and $|E| = \binom{6}{1}\binom{5}{2}$. This gives

$$\mathbb{P}(E) = \frac{|E|}{|S|} = \frac{\binom{6}{1}\binom{5}{2}}{\binom{11}{3}} = \frac{4}{11}$$

Remark 2.5.1. *When an experiment consists of selection (drawing) of k objects from n objects, two choices:*

(1) Considered the ordered set of selections (draws)

(2) Consider the selections (draws) as unordered

Example 2.5.3. *Suppose 5 people are to be randomly selected from a group of 20 individuals consisting of 10 married couples. Find the probability of E, the event that the 5 chosen are all unrelated. That is, no two are married to each other.*

Solution. Method 1–unordered selection: *Regard S, the sample space, as the set of 5 people chosen. Then $|S| = \binom{20}{5}$*

An outcome that does not contain a married couple is the result of a six-stage experiment. Stage 1: 5 of the 10 couples to have a member in the group are chosen. This is gives $\binom{10}{5}$ choices. In the next five stages one of the members of the each of the selected couples is seclude. This gives a total of $2^5\binom{10}{5}$. This gives

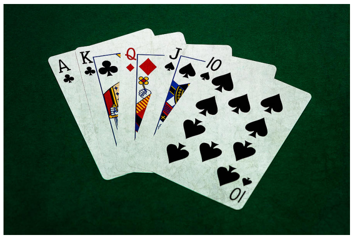
$$\mathbb{P}(E) = \frac{|E|}{|S|} = \frac{2^5\binom{10}{5}}{\binom{20}{5}} \tag{2.5.1}$$

Method 2–ordered selection: *If we let the outcome of the experiment be the ordered selection of the 5 people, there are* $20 \cdot 19 \cdot 18 \cdot 17 \cdot 16$ *equally likely outcomes with* $20 \cdot 18 \cdot 16 \cdot 14 \cdot 12$ *outcomes with unrelated (not married) people.*

$$\mathbb{P}(E) = \frac{|E|}{|S|} = \frac{20 \cdot 18 \cdot 16 \cdot 14 \cdot 12}{20 \cdot 19 \cdot 18 \cdot 17 \cdot 16}. \tag{2.5.2}$$

You can check that (2.5.1) *and* (2.5.2) *are equal.*

Example 2.5.4. *A poker hand consists of 5 cards. A straight is a hand that contains five cards of sequential rank, not all of the same suit, such as in pictures. What is the probability that you are dealt a straight?*



.

Solution. *Sample Space* $S = \{$*non ordered hand*$\}$. *Event* $E = \{$*straight hand*$\}$. $|S| = \binom{52}{5}$. *To compute* $|E|$ *we need to think more carefully. For example, let's start with an ace:*

- *# of possible hands of* $\text{ace}, 2, 3, 4, 5$*:* 4^5
- *# of possible hands of* $\text{ace}, 2, 3, 4, 5$ *not of same suit:* $4^5 - 4$
- *# of possible straights: 10.*

$$\mathbb{P}(E) = \frac{|E|}{|S|} = \frac{10(4^5 - 4)}{\binom{52}{5}} \approx 0.0014$$

Not terribly good odds, would you agree!

Problem 2.5.2. *With this example in mind*

1. *Compute the probability of a Straight flush. Recall that this means five cards of the same suit in sequence, such as* $3\heartsuit 4\heartsuit 5\heartsuit 6\heartsuit 7\heartsuit$.

2. *Compute the probability of a Royal Flush. (The best hand possible, a royal flush consists of:* **10, jack, queen, king and ace**, *all of the same suit.)*

2.5.2 The birthday problem

This is a "famous" little problem that you might have heard of before.

Problem 2.5.3. *(Birthday Problem) In a lecture room with n students, what is the probability that at least two people have the same birthday?*

Solution. *The year has 365 days (no Leap Year) and all days are equally likely. There are* $(365)^n$ *equally likely outcomes in the sample space. With*

$$E = \{\textit{Everyone different birthday}\},$$

we have $|E| = 365 \cdot 364 \cdots (365 - (n+1))$. *Thus,*

$$\mathbb{P}(E) = \frac{|E|}{|S|} = \frac{365 \cdot 364 \cdots (365 - (n+1))}{(365)^n}$$

and

$$\begin{aligned} \mathbb{P}\,(\textit{at least 2 people same birthday}) &= 1 - \mathbb{P}\{\textit{Everyone different birthday}\} \\ &= 1 - \frac{365 \cdot 364 \cdots (365 - (n+1))}{(365)^n} \end{aligned}$$

You can test that with $n \geq 23$*, this probability is greater than a* $\frac{1}{2}$ *and with* $n = 32$*,*

$$\mathbb{P}\{\textit{at least 2 people same birthday}\} \approx 0.752374.$$

Really High!!! With $n = 50$ *the probability of having two students in the lecture hall with the same birthday increases to approximately 0.970, that is, nearly a* 100% *chance. Most people find this surprising the first time they hear it.*

2.5.3 Some problems

Problem 2.5.4. *A committee of 5 is to be selected from a group of 6 men and 9 women. If the selection is made randomly, what is the probability that the committee consists of 3 men and 2 women?*

Problem 2.5.5. *An urn contains n balls, one of which is special. If k of these balls are withdrawn one at a time, with each selection being equally likely to be any of the balls that remain at the time, what is the probability that the special ball is chosen?*

Problem 2.5.6. *A 5-card poker hand is said to be a full house if it consists of 3 cards of the same denomination and 2 other cards of the same denomination (of course, different from the first denomination). Thus, one kind of full house is three of a kind plus a pair. What is the probability that one is dealt a full house?*

Problem 2.5.7. *Two cards are randomly selected from an ordinary playing deck. What is the probability that they form a blackjack? That is, what is the probability that one of the cards is an ace and the other one is either a ten, a jack, a queen, or a king?*

2.6 Probability as a continuous set function

Consider an increasing sequence of sets

$$E_1 \subset E_2 \subset E_3 \dots$$

Observe that $E_n = \cup_{k=1}^{n} E_k$ and so we set $E = \lim_{n\to\infty} E_n$.

Similarly, for a decreasing sequence

$$E_1 \supset E_2 \supset E_3 \dots$$

we have $E_n = \bigcap_{k=1}^{n} E_k$ and define $\lim_{n\to\infty} E_k = \bigcap_{k=1}^{\infty} E_k$.

Proposition 2.6.1

With the notation as above we have:

$$\lim_{n\to\infty} \mathbb{P}(E_n) = \mathbb{P}(\lim_{n\to\infty} E_n)$$

Chapter 3

Week 3: Independent Events

Of fundamental importance I probability is the notion of independence. This intuitively clear property of the independence of one out come from another outcome has a precise and beautiful mathematical formulation.

3.1 Independence

Definition 3.1.1

We say two events E and F (i.e. $E, F \in \mathcal{F}$) are independent events if

$$\mathbb{P}(E \cap F) = \mathbb{P}(E)\,\mathbb{P}(F).$$

Example 3.1.1. *Suppose you flip two coins. The event that you get heads on the second coin is independent of the event that you get tails on the first. This is why: Let $A = \{\{T, H\}, \{T, T\}\}$ be the event of getting tails for the first coin and $B = \{\{T, H\}, \{H, H\}\}$ the event of getting heads for the second coin. Then*

$$\begin{aligned} \mathbb{P}(A \cap B) &= \frac{1}{4} \\ \mathbb{P}(A)\,\mathbb{P}(B) &= \frac{1}{2}\frac{1}{2} = \frac{1}{4}. \end{aligned}$$

Example 3.1.2. *Draw a card from an ordinary deck of cards of 52 cards. Let $A = \{draw\ ace\}$, $B = \{draw\ a\ spade\}$. These are independent events since you're taking one at a time, so one doesn't effect the other. To see this using the definition we have compute $\mathbb{P}(A)\mathbb{P}(B) = \frac{1}{13}\frac{1}{4}$. White $\mathbb{P}(A \cap B) = \frac{1}{52}$, since there is only 1 Ace of spades.*

Proposition 3.1.1

If E and F be independent, then E and F^c are independent.

Proof. Draw a Venn Diagram to help with the computation, but note that writing $E = (E \cap F) \cup (E \cap F^c)$ (disjoint union) gives

$$\begin{aligned} \mathbb{P}(E \cap F^c) &= \mathbb{P}(E) - \mathbb{P}(E \cap F) \\ &= \mathbb{P}(E) - \mathbb{P}(E)\mathbb{P}(F) \\ &= \mathbb{P}(E)(1 - \mathbb{P}(F)) \\ &= \mathbb{P}(E)\mathbb{P}(F^c). \end{aligned}$$

■

Definition 3.1.2: Jointly independent events

Let $A_1, \dots, A_n$ be any collection of events. We say events $A_1, \dots, A_n$ are *jointly independent* if for all subcollections $i_1, \dots, i_r \in \{1, \dots, n\}$ we have that

$$\mathbb{P}\left(\bigcap_{j=1}^{r} A_{i_j}\right) = \prod_{j=1}^{r} \mathbb{P}\left(A_{i_j}\right).$$

In particular, if we have three events, E, F, G, they are *jointly independent* if E, F are independent, E, G are independent, F, G are independent, and $\mathbb{P}(E \cap F \cap G) = \mathbb{P}(E)\mathbb{P}(F)\mathbb{P}(G)$.

Definition 3.1.3: Pairwise Independent events

A collection $A_1, \dots, A_n$ of events is said to be pairwise independent if any two of the events are independent. That is, for any i, j, $i \neq j$,

$$\mathbb{P}(A_i \cap A_j) = \mathbb{P}(A_i \cap A_j)$$

3.1.1 Examples

Example 3.1.3. *Roll two dice. Define the following events:* $A_7 = \{\textit{sum is 7}\}$, $B_4 = \{\textit{first die is a 4}\}$, $C_3 = \{\textit{second die is a 3}\}$ *Are these events independent?*

Solution. *As already noted, the sample space* $S = \{(i,j) : i,j = 1,\ldots,6\}$ *has 36 elements*

$$\begin{aligned} S = \quad & \{(1,1),\ (1,2),\ (1.3),\ (1,4),\ (1,5),\ (1,6), \\ & (2,1),\ (2,2),\ (2,3),\ (2,4),\ (2,5),\ (2,6), \\ & (3,1),\ (3,2),\ (3,3),\ (3,4),\ (3,5),\ (3,6), \\ & (4,1),\ (4,2),\ (4,3),\ (4,4),\ (4,5),\ (4,6), \\ & (5,1),\ (5,2),\ (5,3),\ (5,4),\ (5,5),\ (5,6), \\ & (6,1),\ (6,2),\ (6,3),\ (6,4),\ (6,5),\ (6,6)\} \end{aligned}$$

From this we see (look long the diagonal from top right corner to bottom left corner) that

$$\mathbb{P}\left(A_7 \cap B_4 \cap C_3\right) = \mathbb{P}\left(\{(4,3)\}\right) = \frac{1}{36}.$$

But (look at 4th row and 3rd column)

$$\mathbb{P}\left(A_7\right)\mathbb{P}\left(B_4\right)\mathbb{P}\left(C_3\right) = \frac{6}{36}\frac{1}{6}\frac{1}{6} = \frac{1}{36 \cdot 6}.$$

Hence sets are not independent.

On the other hand, we also have

$$\mathbb{P}(A_7 \cap B_4) = \mathbb{P}(B_4 \cap C_3) = \mathbb{P}(A_7 \cap C_3) = \frac{1}{36}$$

and

$$\mathbb{P}(A_7) = \mathbb{P}(B_4) = \mathbb{P}(C_3) = \frac{1}{6}.$$

Remark 3.1.1. *This is an example of a collection of sets which is* **pairwise independent but not jointly independent**.

Example 3.1.4. *We have two urns. One contains 10 balls: 4 red and 6 blue. The other contains 16 red balls and an unknown number of blue balls. A single ball is drawn from each urn. The probability that both balls are the same color is 0.44. Calculate the number of blue balls in the second urn.*

Solution. *Let* R_i = *event that a red ball is drawn from urn i and let* B_i = *event that a blue ball is drawn from urn i. Let x be the number of blue balls in urn* 2, *Note that drawing from urn 1 and independent from drawing from urn 2. They are completely different urns! Then*

$$\begin{aligned} 0.44 &= \mathbb{P}\left((R_1 \cap R_2) \cup (B_1 \cap B_2)\right) = \mathbb{P}\left(R_1 \cap R_2\right) + \mathbb{P}\left(B_1 \cap B_2\right) \\ &= \mathbb{P}\left(R_1\right)\mathbb{P}\left(R_2\right) + \mathbb{P}\left(B_1\right)\mathbb{P}\left(B_2\right), \text{ by independence} \\ &= \frac{4}{10}\frac{16}{x+16} + \frac{6}{10}\frac{x}{x+16}. \end{aligned}$$

Solve for x! You will get $x = 4$.

Example 3.1.5. *Suppose you roll 50 dice. What is the probability that you get a 2 exactly 10 times?*

Solution. *Think of this as an experiment where you roll one single die 50 times. The rolls are independent. One possible outcome is that you get a 2 in the first 3 rolls and in the last 7 rolls and no 2 in the other 40 rolls. This probability will be* $(\frac{1}{6})^{10}(\frac{5}{6})^{40}$*. Since there are* $\begin{pmatrix} 50 \\ 10 \end{pmatrix}$ *such outcomes, we have*

$$P(\text{get 2 in 10 rolls}) = \begin{pmatrix} 50 \\ 10 \end{pmatrix} \left(\frac{1}{6}\right)^{10} \left(\frac{5}{6}\right)^{40}$$

This is a typical example of a what is called a *Bernoulli Trial.*

3.2 The binomial distribution

Definition 3.2.1: The Bernoulli Distribution (Binomial distribution)

If we have an experiment where the probability of some outcome occurring (we say this is a success) is $0 < p < 1$ and the probability of not occurring (we say failure) is $1 - p$. The experiment is repeated independently n times (n independent trials), the the probability of exactly k success in the n-trials is given by

$$\begin{pmatrix} n \\ k \end{pmatrix} p^k(1-p)^{n-k}$$

Remark 3.2.1. *The probability there are k successes is the number of ways of putting k objects in n slots which is* $\begin{pmatrix} n \\ k \end{pmatrix}$ *times the probability that there will be k successes and $n - k$ failures in exactly one of a given order. So the probability is* $\begin{pmatrix} n \\ k \end{pmatrix} p^k(1-p)^{n-k}$.

Remark 3.2.2. *Note that if we set $E_k = \{k \text{ success in } n \text{ trials}\}$ we can write the sample space $S = \bigcup_{k=0}^n E_k$ and by the Binomial Theorem with $p = x$, $y = 1 - p$,*

$$\mathbb{P}(S) = P(\bigcup_{k=0}^{n} E_k) = \sum_{k=0}^{n} \mathbb{P}(E_k) = \sum_{k=0}^{n} \begin{pmatrix} n \\ k \end{pmatrix} p^k(1-p)^{n-k} = (p - (1-p))^n = 1$$

hence the name Binomial Distribution used sometimes.

Remark 3.2.3. *This distribution is often denoted as Binom(n, p) and simply refer to it as "the binomial distributions of parameters n and p."*

3.2.1 Examples

Example 3.2.1. *Suppose you take a multiple choice test with 25 questions with 4 possible answers for each question. Suppose you randomly guess on each question with which are all independent. What is the probability that you will guess 8 correct?*

Solution. *We can model this experiment with the Bernoulli distribution* $n = 25$, $k = 8$, $p = \frac{1}{4}$.

$$\mathbb{P}\{\textit{exactly 8 correct}\} = \binom{25}{8}\left(\frac{1}{4}\right)^8\left(\frac{3}{4}\right)^{17}.$$

If you calculate this with a calculator you will see that the probability is quite low and it will get smaller as you increase your desired number of correct answers. (Don't Guess on Multiple Choice Exams!)

Example 3.2.2. *You go to Walmart and buy a box with 2 dozen lightbulbs. It is known that* 1% *of all the bulbs from this company are defective. What is the probability that you will find at least 4 defective lightbulbs in your box.*

Solution. *This another example of a Bernoulli distribution. But the words "at least" are important. We have* $n = 24$ *and* $p = 0.01$. *Want* $\mathbb{P}(\textit{at least 4 defective})$. *If we let* $\mathbb{P}(i)$ *with* $= 0, 1, 2, 3,$ *be the probability that exactly i lightbulbs are defective, we want*

$$\begin{aligned}\mathbb{P}(\textit{at least 4 defective}) &= 1 - (\mathbb{P}(0) + P(1) + \mathbb{P}(2) + \mathbb{P}(3)) \\ &= 1 - \binom{24}{0}(0.01)^0(0.99)^{24} - \binom{24}{1}(0.01)^1(0.99)^{23} \\ &- \binom{24}{2}(0.01)^2(0.99)^{22} - \binom{24}{3}(0.01)^3(0.99)^3\end{aligned}$$

Example 3.2.3. *What is the probability that exactly 4 two's will show up if you roll 10 dice?*

Solution. *These are independent. The probability that the 1st, 2nd, 3rd, and 10th dice will show a three and the other 6 will not is* $\left(\frac{1}{6}\right)^3\left(\frac{5}{6}\right)^7$.

Independence is used here: the probability is $\frac{1}{6}\frac{1}{6}\frac{1}{6}\frac{5}{6}\frac{5}{6}\frac{5}{6}\frac{5}{6}\frac{5}{6}\frac{5}{6}\frac{1}{6}$. *Note that the probability that the 10th, 9th, 8th, and 7th dice will show a* two *and the other 6 will not has the same probability.*

So to answer our original question, we take $\left(\frac{1}{6}\right)^4\left(\frac{5}{6}\right)^6$ *and multiply it by the number of ways of choosing 4 dice out of 10 to be the ones showing the* two's. *There are* $\binom{10}{3}$ *ways to do this* $\binom{10}{4}\left(\frac{1}{6}\right)^4\left(\frac{5}{6}\right)^6$.

Chapter 4

Week 4: Conditional Probability

4.1 Conditional probability and independence

The following question naturally arises: How should we modify the probability of an event when some information concerning the outcome of an experiment is available? The concept of conditional probability, which as we shall see is closely related to independence, will help answer this question.

Example 4.1.1. *Suppose there are 200 men of which 100 are smokers and 100 women of which 20 are smokers. What is the probability that a person chosen at random from these 300 people will be a smoker? The answer is* $\frac{120}{300}$. *But now, let us ask a different question: what is the probability that a person chosen at random is a smoker* **given** *that the person is a women?* $\frac{20}{100}$ *right?*

Note that this is

$$\frac{\#\,(\textit{women smokers})}{\#\,(\textit{women})} = \frac{P\,(\textit{smoker and a women})}{P\,(\textit{woman})}.$$

Definition 4.1.1: Conditional probability

Let F be a fixed event in our probability space S with $\mathbb{P}(F) > 0$. We define the conditional probability that an event E given (or knowing) F has occurred by

$$\mathbb{P}(E|F) = \frac{\mathbb{P}(E \cap F)}{\mathbb{P}(F)}$$

We read $\mathbb{P}\,(E \mid F)$ simply as “the probability of E given F.”

4.1.1: Remark

(i) It stands to reason that if the outcome of F (i.e., whether F has occurred or not) has no influence on the outcome of E, then the probability of E given F should be just the probability of E. This of course is the case. Recall Definition (1.1) from "Lecture notes week 1": two events E and F are independent if

$$\mathbb{P}\left(E \cap F\right) = \mathbb{P}\left(E\right)\mathbb{P}\left(F\right).$$

(ii) Suppose you know $\mathbb{P}(E|F)$ but you want $\mathbb{P}(F|E)$. Since $\mathbb{P}(E|F)\mathbb{P}(F) = \mathbb{P}(E \cap F) = \mathbb{P}(F \cap E)$

we have that
$$\mathbb{P}(F|E) = \frac{\mathbb{P}(F \cap E)}{\mathbb{P}(E)} = \frac{\mathbb{P}(E|F)\mathbb{P}(F)}{\mathbb{P}(E)}$$

We highlight these properties

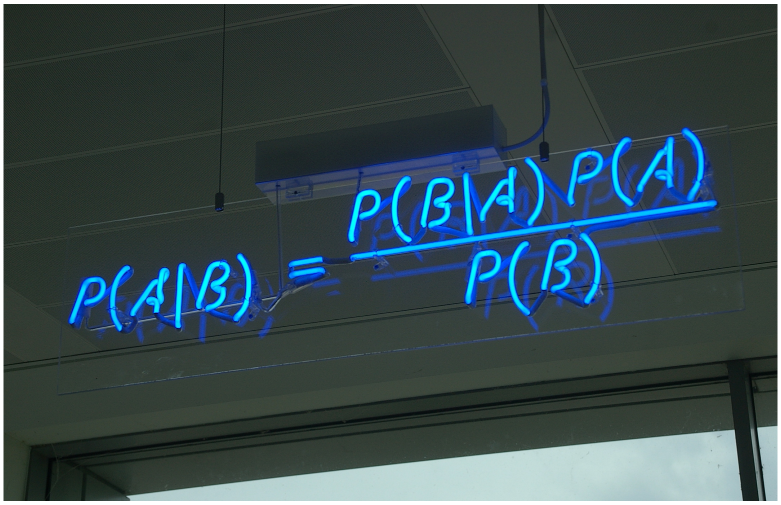
Properties 4.1.1

(Independence) If E and F are independent, then

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

$(\mathbb{P}(F|E)\, from\, \mathbb{P}(E|F))$
$$\mathbb{P}(F|E) = \frac{\mathbb{P}(F \cap E)}{\mathbb{P}(E)} = \frac{\mathbb{P}(E|F)\mathbb{P}(F)}{\mathbb{P}(E)}$$

(This formula is one of many called "Bayes' Formula." More to come in the sections below.) We see that $\mathbb{P}(F \mid E) = \mathbb{P}(E \mid F)$ if and only if $\mathbb{P}(E) = \mathbb{P}(F)$.



Offices of H.P. Autonoma sign–http://timtyler.org/bayesianism

4.1.2: Remark

Here is a concrete example concerning the statement on equality. In the dining hall a student picks salads that contain A) Artichokes, B) Broccoli, or C) Cauliflower with the following information. $\mathbb{P}(B) = 0.4$, $\mathbb{P}(C) = 0.44$, $\mathbb{P}(B \cap C) = 0.13$

$$\mathbb{P}(B|C) = \frac{\mathbb{P}(B \cap C)}{\mathbb{P}(C)} = \frac{13}{44}, \quad \mathbb{P}(C|B) = \frac{\mathbb{P}(B \cap C)}{\mathbb{P}(B)} = \frac{13}{40}$$

4.1.1 Examples

Example 4.1.2. *Peter is 80% sure he forgot his iPhone at the Purdue Union or at his Dorm. 40% sure that it is at the union, and 40% sure that it is at the dorm. Given that Peter already went to dorm and noticed his phone not there, what is the probability that it's at the union?*

Solution. *Let U = event phone in union, D = event phone in dorm*

$$\begin{aligned}\mathbb{P}\left(U \mid D^c\right) &= \frac{\mathbb{P}\left(U \cap D^c\right)}{\mathbb{P}\left(D^c\right)} \\ &= \frac{\mathbb{P}\left(U\right)}{1 - \mathbb{P}\left(D\right)}, \text{ since } U \subset D^c \\ &= \frac{4/10}{6/10} = \frac{2}{3}.\end{aligned}$$

Example 4.1.3. *Phan wants to take a Biology course or a Chemistry course. Given that a student take Biology, the probability that they get an A is is $\frac{4}{5}$. While the probability of getting an A given* *that a student take Chemistry is $\frac{1}{7}$. If Phan makes a decision on the course to take randomly, what's probability of getting an A in Chem?*

Solution. *Let $B = \{$Takes Biology$\}$ and $C = \{$Takes Chemistry$\}$ and $A = \{$"gets an A"$\}$, then*

$$\begin{aligned}\mathbb{P}(A \cap C) &= \mathbb{P}(C)\,\mathbb{P}(A \mid C) \\ &= \frac{1}{2} \cdot \frac{1}{7} = \frac{1}{14}.\end{aligned}$$

Example 4.1.4. *Three cards are randomly selected, without replacement from an ordinary deck of 52 playing cards. What is the conditional probability that the first card selected is a spade given that the second and third cards are spades.*

Solution. *Let S_1, S_2 and S_3 be the events that the first, the second and the third card selected is a spade, respectively. Then*

$$\mathbb{P}(S_1|S_2 \cap S_3) = \frac{\mathbb{P}(S_1 \cap S_2 \cap S_3)}{\mathbb{P}(S_2 \cap S_3)}$$

and

$$\mathbb{P}(S_1 \cap S_2 \cap S_3) = \frac{13 \cdot 12 \cdot 11}{52 \cdot 51 \cdot 50} = a$$

Next apply (f) of Proposition 3.1 in "Lecture notes for week 2" with $E = S_2 \cap S_3$ and $F = S_1$ to write

$$\mathbb{P}(S_2 \cap S_3) = \mathbb{P}(S_1 \cap S_2 \cap S_3) + \mathbb{P}(S_1^c \cap S_2 \cap S_3) = \frac{13 \cdot 12 \cdot 11}{52 \cdot 51 \cdot 50} + \frac{39 \cdot 13 \cdot 12}{52 \cdot 51 \cdot 50} = a + b.$$

This gives,

$$\mathbb{P}(S_1|S_2 \cap S_3) = \frac{a}{a+b} = \frac{1}{1+\frac{b}{a}} = \frac{1}{1+\frac{39}{11}} = \frac{11}{50}$$

4.1.2 Examples, working with the reduced sample space

Example 4.1.5. *A fair coin is flipped twice. Let the sample space be $S = \{(H,H), (H,T), (T,H), (T,T)\}$. What is the conditional probability that both flips land on heads given that (a) the first flip lands on heads? (b) at least one flip lands on heads?*

Solution. *Let $E = \{(H,H)\}$ be the event that both flips land on heads and let $F = \{(H,H), (H,T)\}$ be the event that the first flip lands on heads and let $A = \{(H,H), (H,T), (T,H)\}$ be the event that at least one flip lands on heads. We have*

(a)

$$\mathbb{P}(E|F) = \frac{\mathbb{P}(E \cap F)}{\mathbb{P}(F)} = \frac{\mathbb{P}(\{(H,H)\}}{\mathbb{P}(\{(H,H),(H,T)\})} = \frac{1/4}{2/4} = \frac{1}{2}$$

and

(b)

$$\mathbb{P}(E|A) = \frac{\mathbb{P}(E \cap A)}{\mathbb{P}(A)} = \frac{\mathbb{P}(\{(H,H)\}}{\mathbb{P}(\{(H,H),(H,T),(T,H)\})} = \frac{1/4}{3/4} = \frac{1}{3}$$

Definition 4.1.2: The Reduce Sample Space

Let us recall that in an experiment with all events equally likely (uniform distribution), for any event E, we have the formula

$$\mathbb{P}(F) = \frac{|F|}{|S|}$$

where F denotes the #elements in F, and similarly for $|S|$. Thus we can write

$$\mathbb{P}(E|F) = \frac{\mathbb{P}(E \cap F)}{\mathbb{P}(F)} = \frac{\frac{|E\cap F|}{|S|}}{\frac{|F|}{|S|}} = \frac{|E \cap F|}{|F|}.$$

This means that we can consider the conditional probability as a probability where the sample space has been "reduced" to F. That is, F has become the "new S'." Of course, the event $E' = E \cap F$ is in the new sample space S'. This new sample space is often called the **Reduced sample space.** With this remark, taking a second look at the example above gives (for (a)) the sample space $S' = F = \{(H,H),(H,T)\}$ and $E' = E \cap F = \{(H,H)\}$ and we get immediately 1/2 for our answer.

Example 4.1.6. *Suppose a box has 3 red marbles and 2 black ones. We select 2 marbles. What is the probability that second marble is red given that the first one is red?*

Solution. *Let $R_1 = \{First\ marble\ is\ is\ red\ \}$ and $R_2 = \{Second\ marble\ is\ red\}$, then*

$$\begin{aligned}
\mathbb{P}(R_2 \mid R_1) &= \frac{\mathbb{P}(R_1 \cap R_2)}{\mathbb{P}(R_1)} \\
&= \frac{(2\ red)\,(0\ \ black)\,/\begin{pmatrix}5\\2\end{pmatrix}}{3/5} \\
&= \frac{\begin{pmatrix}3\\2\end{pmatrix}\begin{pmatrix}2\\0\end{pmatrix}/\begin{pmatrix}5\\2\end{pmatrix}}{3/5} \\
&= \frac{3/10}{3/5} = \frac{1}{2}.
\end{aligned}$$

4.1.3: Remark

Here the reduced sample space is $S' = \{R, R, B, B\}$ and thus $\mathbb{P}\left(R_2 \mid R_1\right) = \frac{2}{4} = \frac{1}{2}$

Example 4.1.7. *Suppose we roll a die. Let B be the event that the outcome is odd. Find (a) $\mathbb{P}(\{3, 5\} \mid B)$, (b) $\mathbb{P}\{1, 3, 5\}|B)$ and (c) $\mathbb{P}(\{4, 6\}|B)$.*

Solution. *The original sample space is $S = \{1, 2, 3, 4, 5, 6\}$. Given the information we have, our new sample space is $S' = \{1, 3, 5\}$. Thus*

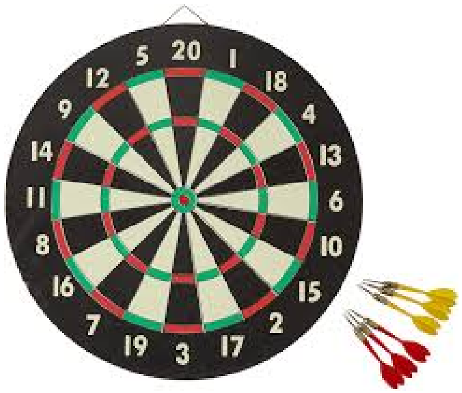
$$\mathbb{P}(\{3, 5\} \mid B) = \mathbb{P}'(\{3, 5\}) = 2/3$$

$$\mathbb{P}(\{1, 3, 5\} \mid B) = \mathbb{P}'(\{1, 3, 5\}) = 1$$

and

$$\mathbb{P}(\{4, 6\} \mid B) = \mathbb{P}'\{4, 6\} = 0$$

Example 4.1.8. *On a dartboard, let $A = \{9, 12, 5, 20, 1, 18, 4\}$ be the event that a dart lands in this upper portion of the board. Let $B = \{9, 12, 5\}$. What is $\mathbb{P}(B|A)$?*



Solution. *The original sample space is $S = \{1, 2, dots, 20\}$. The reduced sample space is $S' = \{9, 12, 5, 20, 1, 18, 4\}$. $\mathbb{P}(B \mid A) = 3/7$.*

4.2 Multiplication rule, probability of intersections

For this section we assume that whenever we divide by a probability it is always positive. Let us take another look at our definition of conditional probability. We will use notation $A_1, A_2, \ldots, A_n$ for events instead of $E, F, G, \ldots$. By the definition of conditional expectation we know:

Properties 4.2.1

$$\mathbb{P}(A_1 \cap A_2) = \mathbb{P}(A_1)\mathbb{P}(A_2 \mid A_1) \tag{4.2.1}$$

Here is our conclusion from this formula: In order for A_1 and A_2 to occur we need:

- for A_1 to occur and then,
- given that A_1 occurs we need A_2 to occur.

Iterate the formula: (Take $E = A_3$ and $F = A_1 \cap A_2$.)

$$\frac{\mathbb{P}(A_1 \cap A_2 \cap A_3)}{\mathbb{P}(A_1 \cap A_2)} = \mathbb{P}(A_3 \mid A_1 \cap A_2)$$

to get

Properties 4.2.2

$$\begin{aligned} \mathbb{P}(A_1 \cap A_2 \cap A_3) &= \mathbb{P}(A_1 \cap A_2)\mathbb{P}(A_3 \mid A_1 \cap A_2) \qquad (4.2.2) \\ &= \mathbb{P}(A_1)\mathbb{P}(A_2 \mid A_1)\mathbb{P}(A_3 \mid A_1 \cap A_2), \end{aligned}$$

where we used (4.2.1) to get the second equality. Here is our conclusion from this formula: In order for A_1 and A_2 and A_3 to occur we need:

- for A_1 to occur and then,
- given that A_1 occurs we, need A_2 to occur too and then,
- given that A_1 and A_2 occurs, we need A_3 to occur too.

The general formula is:

Properties 4.2.3: Multiplication Rule

Let $A_1, A_2, \ldots, A_n$ be events with $A_1 \cap A_2 \cap \cdots \cap A_n \neq \varnothing$ Then

$$\begin{aligned} \mathbb{P}(A_1 \cap A_2 \cap \cdots \cap A_n) &= \mathbb{P}(A_1)\mathbb{P}(A_2 \mid A_1)\mathbb{P}(A_3 \mid A_1 \cap A_2) \times \\ &\quad \mathbb{P}(A_4 \mid A_1 \cap A_2 \cup A_3) \times \\ &\quad \cdots \times \mathbb{P}(A_n \mid A_1 \cap A_2 \cap \cdots \cap A_{n-1}) \end{aligned}$$

4.2.1 Examples

Example 4.2.1. *Nancy has a play lis that has 10 rock songs and 12 country songs.She chooses three songs from the play list at random with the selections equally likely and no repeats. What is the probability that all three are country songs?*

Solution. *Let C_1, C_2, C_3 be the events that the 1st, 2nd and 3rd song is a country song, respectively. We want the probability of the interaction of these events which is given by*

$$\mathbb{P}(C_1 \cap C_2 \cap C_3) = \mathbb{P}(C_1)\mathbb{P}(C_2 \mid C_1)\mathbb{P}(C_3 \mid (C_1 \cap C_2)$$

We have $\mathbb{P}(C_1) = 12/22$, $\mathbb{P}(C_2 \mid C_1) = 11/21$ and $P(C_3 \mid C_1 \cap C_2) = 10/20$. The product of these is $1/7$.

Example 4.2.2. *An urn contains 6 white and 9 black balls. If 4 balls are to be randomly selected without replacement, what is the probability that the first 2 selected are white and the last 2 black?*

Solution. *Let W_1, W_2, W_3, W_4 be the events that respectively the first, the second, the third and the fourth ball selected is white. W^c denotes the event that the i-th ball is black, $i \in \{1, 2, 3, 4\}$). The desired probability is given by the multiplication rule:*

$$\begin{aligned} & \mathbb{P}(W_1 \cap W_2 \cap W_3^c \cap W_4^c) \\ = \ & \mathbb{P}(W_1)\mathbb{P}(W_2 \mid W_1)\mathbb{P}(W_3^c \mid W_1 \cap W_2)\mathbb{P}(W_4^c \mid W_1 \cap W_2 \cap W_3^c) \\ = \ & \frac{6}{15}\frac{5}{14}\frac{9}{13}\frac{8}{12} \end{aligned}$$

4.2.1: Remark

The above examples are of the following type: You have a container with N objects of one type and M objects of another type and you one to select n items from it. All items are equally likely. What is the probability that all items are of the second type? The answer is: Let A_i, $i \in \{1, 2, \ldots, n\}$ be the event that the ith item is of the second type. We want the probability of the intersection of the $A_i's$. We have:

$$\mathbb{P}(A_1) = \frac{M}{N+M}, \quad \mathbb{P}(A_2 \mid A_1) = \frac{M-1}{N+M-1}, \quad \mathbb{P}(A_3 \mid A_2 \cap A_1) = \frac{M-2}{N+M-2}$$

and in general

$$\mathbb{P}(A_n | A_1 \cap A_2 \cap \cdots, A_{n-1}) = \frac{M-(n-1)}{N+M-(n-1)}$$

Multiplying these probabilities gives:

$$\mathbb{P}(A_1 \cap A_2 \cdots \cap A_n) = \frac{M(M-1)\cdots(M-(n-1))}{(N+M)(N+M-1)\cdots(N+M-(n-1))}$$

4.3 The law of total probability

Sometimes it's easy (or easier) to compute a probability once we know something has or has not happened. The next Theorem is very useful many such cases.

Properties 4.3.1: (Special case of "Law of Total Probability")

For any two events E and F, we have

$$\mathbb{P}(E) = \mathbb{P}(E \mid F)\,\mathbb{P}(F) + \mathbb{P}(E \mid F^c)\,\mathbb{P}(F^c)$$

Proof. Writing E as the disjoint union

$$E = (E \cap F) \cup (E \cap F^c),$$

we have

$$\begin{aligned} \mathbb{P}(E) &= \mathbb{P}(E \cap F) + \mathbb{P}(E \cap F^c) \\ &= \mathbb{P}(E \mid F)\,\mathbb{P}(F) + \mathbb{P}(E \mid F^c)\,\mathbb{P}(F^c) \end{aligned}$$

■

Notice that since $\mathbb{P}(F^c) = 1 - \mathbb{P}(F)$ we can write the low of total probability as

Properties 4.3.2: (Same as Property 1.2)

$$\mathbb{P}(E) = \mathbb{P}(E \mid F)\,\mathbb{P}(F) + \mathbb{P}(E \mid F^c)\mathbb{P}(F^c)\,(1 - \mathbb{P}(F)) \tag{4.3.1}$$

The Law of Total Probability extends to many events. That is, the only thing we really sued above is that the whole sample space $S = F \cup F^c$ and that this is a disjoint union. Suppose we have a collection of pairwise disjoint events $F_1, \ldots, F_n$ such that

$$S = \bigcup_{j=1}^{n} F_j$$

The we can write $E = \bigcup_{j=1}^{n}(E \cap F)$ and this union is pairwise disjoint. (Above we have only two sets: $F_1 = F$ and $F_2 = F^c$.) By our property of $\mathbb{P}$ this gives

$$\mathbb{P}(E) = \sum_{j=1}^{n} P(E \cap F_j) = \sum_{j=1}^{n} \mathbb{P}(E \mid F_j)\mathbb{P}(F_j).$$

We summarize this in the following

Theorem 4.3.1: The Law of Total Probability (LTP)

Suppose $S = \cup_{j=1}^{n} F_j$ where the events $F_1, \dots, F_n$ are pairwise disjoint. Then for any event E,

$$\mathbb{P}(E) = \sum_{j=1}^{n} P(E \cap F_j) = \sum_{j=1}^{n} \mathbb{P}(E \mid F_j)\mathbb{P}(F_j).$$

4.4 Bayes' formula

Let us now take any $k \in \{1, 2, \dots, n\}$. We have

$$\begin{aligned}
\mathbb{P}(F_k \mid E) &= \frac{\mathbb{P}(F_k \cap E)}{\mathbb{P}(E)} \quad \text{(def of cond. exp)} && (4.4.1)\\
&= \frac{\mathbb{P}(F_k \cap E)}{\sum_{j=1}^{n} \mathbb{P}(E \mid F_j)\mathbb{P}(F_j)} \quad \text{(LTP)} && (4.4.2)\\
&= \frac{\mathbb{P}(E \mid F_k)\mathbb{P}(F_k)}{\sum_{j=1}^{n} \mathbb{P}(E \mid F_j)\mathbb{P}(F_j)} && (4.4.3)
\end{aligned}$$

where in the last inequality we used again the definition of conditional expectation to write

$$\mathbb{P}(F_k \cap E) = \mathbb{P}(E \mid F_k)\mathbb{P}(F_k)$$

4.4.1: Remark

The fact that the left hand side of (2.1) is equal to the right hand side of (2.3) goes by the name Bayes' formula (rule). Here are some facts you can easily find about Bayes with Google: (1) English philosopher, statistician and Presbyterian minister. (2) Wrote two books but his famous formula was unpublished.

Thomas Bayes (1701-1760)

We summarize this in

Theorem 4.4.1: Bayes' Formula

Suppose $S = \cup_{j=1}^{n} F_j$, F_j mutually exclusive. For any event E in S and any any $k \in \{1, 2, \ldots, n\}$,

$$\mathbb{P}(F_k \mid E) = \frac{\mathbb{P}(E \mid F_k)\mathbb{P}(F_k)}{\sum_{j=1}^{n} \mathbb{P}(E \mid F_j)\mathbb{P}(F_j)}$$

If we have infinitely many disjoint sets with $S = \cup_{j=1}^{\infty} F_j$, then for any fixed k we have

$$\mathbb{P}(F_k \mid E) = \frac{\mathbb{P}(E \mid F_k)\mathbb{P}(F_k)}{\sum_{j=1}^{\infty} \mathbb{P}(E \mid F_j)\mathbb{P}(F_j)}$$

Remark 4.4.1. *In the case of one set F where $F_1 = F$ and $F_2 = F^c$ and $S = F \cup F^c$, we can write Bays' formula (with $k = 1$) as*

Properties 4.4.1: (Special case of Bayes' formula)

$$\mathbb{P}(F \mid E) = \frac{\mathbb{P}(E \mid F)\,\mathbb{P}(F)}{\mathbb{P}\,(E \mid F)\,\mathbb{P}\,(F) + \mathbb{P}\,(E \mid F^c)\,\mathbb{P}(F^c)}$$

We will often apply this special case in this class and simply refer to it as Bayes' formula or even Bayes's rule.

4.4.2: Remark

VERY USEFUL to note that the denominator in Bayes's formula is always the term in the formula for the Total Law of Probability.

4.4.1 Examples

Example 4.4.1. *A box contains 3 different types of disposable flashlights. The probability that a type 1 flashlight will give over 100 hours of use is 0.7, with the corresponding probabilities for type 2 and type 3 flashlights being 0.4 and 0.3, respectively. Suppose that 20 percent of the flashlights in the box are type 1, 30 percent are type 2, and 50 percent are type 3. What is the probability that a randomly chosen flashlight will give more than 100 hours of use?*

Solution. *Let A denote the event that the flashlight chosen will give over 100 hours of use, and let F_j be the event that a type j flashlight is chosen, $j = 1, 2, 3$. By the LTP*

$$\begin{aligned}\mathbb{P}(A) &= \mathbb{P}(A \mid F_1)\mathbb{P}(F_1) + \mathbb{P}(A \mid F_2)\mathbb{P}(F_2) + \mathbb{P}(A \mid F_3)\mathbb{P}(F_3) \\ &= (0.7)(0.2) + (0.4)(0.3) + (0.3)(0.5) = 0.41\end{aligned}$$

Example 4.4.2. *By collecting data for several years, and insurance company has concluded that:*

(i) The probability that "an accident prone person" has an accident within a year is 0.4.

(ii) The probability that "Non-accident prone person" has an accident within a year is 0.2.

(iii) 30% *of the population is "accident prone".*

(a) What is the probability that a new policy holder will have an accident within a year. (b) Suppose new policyholder has an accident within one year. What is the probability that they are accident prone?

Solution. *(a) Let E be the event that the new policy holder has an accident within a year and F be the event that the policy holder is accident prone. By the LTP,*

$$\begin{aligned}\mathbb{P}(E) &= \mathbb{P}(E \mid F)\,\mathbb{P}(F) + \mathbb{P}(E \mid F^c)\,\mathbb{P}(F^c) \\ &= \mathbb{P}(E \mid F)\,\mathbb{P}(F) + \mathbb{P}(E \mid F^c)\,\mathbb{P}(F^c)\,(1 - \mathbb{P}(F)) \\ &= 0.4\,(0.3) + 0.2\,(1 - 0.3) \\ &= 0.26\end{aligned}$$

For (b) we have

$$\begin{aligned}\mathbb{P}(F \mid E) &= \frac{\mathbb{P}(F \cap E)}{\mathbb{P}(E)} \\ &= \frac{\mathbb{P}(F)\,\mathbb{P}(E \mid F)}{0.26} \\ &= \frac{(0.3)\,(0.4)}{0.26} = \frac{6}{13}.\end{aligned}$$

Example 4.4.3. *Suppose a certain test for Covid-19 is 95% accurate in both directions, positive and negative. Suppose 3% of the population is positive. If someone tests positive, what is the probability they actually have Covid-19? (Please note. These are made up numbers for this exercise although good estimates exist for different countries.)*

Solution. *Let us consider the following four events: Let $T_+ = \{\text{tests positive}\}$, $T_- = \{\text{tests negative}\}$, $P = \{\text{actually Covid positive,}\}$, $N = \{\text{actually Covid negative}\}$. Note that $P^c = N$. From the assumptions $\mathbb{P}(T_+ \mid P) = 0.95$ and $\mathbb{P}(T_+ \mid N) = 0.05$.*

Want

$$\begin{aligned}\mathbb{P}(P \mid T_+) &= \frac{\mathbb{P}(P \cap T_+)}{\mathbb{P}(T_+)} \\ &= \frac{\mathbb{P}(T_+ \mid P)\,\mathbb{P}(P)}{\mathbb{P}(T_+ \mid P)\,\mathbb{P}(P) + \mathbb{P}(T_+ \mid N)\,\mathbb{P}(N)}, \text{(by Bayes' formula)} \\ &= \frac{(0.95)\,(0.03)}{(0.95)\,(0.03) + .05\,(0.97)} \\ &\approx 0.6298\end{aligned}$$

That is, approximately 62.98%. *(Warning: always check my arithmetic–I do it by hand)*

Example 4.4.4. *This is an example of the General Bayes' formula. Here's one with more than 3 possibilities: Suppose you have a factory with machines I,II,III that producing iPhones. The machines I,II,III produce 2%,1%, and 3% defective iPhones, respectively. Out of total production, machines I makes* 35% *of all iPhones, machine* II*-25% and machine* III*-* 40%. *(a) Suppose you select one iPhone at random from the factory, for inspection:*

(i) What is the probability that the iPhone selected is defective?

(ii) What is the conditional prob that if the iPhone is defective, it was produced by machine III?

Solution. *First note that using* I, II, III *as the events that a iPhone is produced by machine* I, II, III, *respectively, these are mutually exclusive and* $S = I \cup II \cup III$. *We let* D *be the event of defective iPhone. For part (i) we have:*

$$\begin{aligned}\mathbb{P}(D) &= P(I)\,\mathbb{P}(D \mid I) + P(II)\,\mathbb{P}(D \mid II) + P(III)\,\mathbb{P}(D \mid III)\\ &= (.35)(.02) + (.25)(.01) + (.4)(.03)\\ &= \frac{215}{10,000}.\end{aligned}$$

For part (ii), we have:

$$\begin{aligned}\mathbb{P}(III \mid D) &= \frac{\mathbb{P}(III)\,\mathbb{P}(D \mid III)}{\mathbb{P}(D)}\\ &= \frac{(.4)(.03)}{215/10,000} = \frac{120}{215}.\end{aligned}$$

Example 4.4.5. *In a Multiple Choice Test, students either knows the answer or randomly guesses the answer to a question. Let* m *=number of choices in a question and let* $p =$ *be the probability that the students knows the answer to a question. What is the probability that the student actually knew the answer, given that the student answered correctly.*

Solution. *Let* $K = \{Knows\ the\ answer\}$ *and* $C = \{Answered\ correctly\}$. *Then*

$$\begin{aligned}\mathbb{P}(K \mid C) &= \frac{\mathbb{P}(C \mid K)\,\mathbb{P}(K)}{\mathbb{P}(C \mid K)\,\mathbb{P}(K) + \mathbb{P}(C \mid K^c)\,\mathbb{P}(K^c)}\\ &= \frac{1 \cdot p}{1 \cdot p + \frac{1}{m}(1-p)} = \frac{mp}{1 + (m-1)p}.\end{aligned}$$

Example 4.4.6. *Suppose that we have 3 cards that are identical in form, except that both sides of the first card are red, both sides of the second are black, and one side of the third card is red and the other side is black. The 3 cards are mixed up in a hat and 1 card is randomly selected and put on a table. If the upper side of the chosen card is red, what is the probability that the other side is black?*

Solution. *Let RR, and RB be the events that the chosen card is all red, all black or the red/black card.* $S = RR \cup BB \cup RB$. *Let R be the event that the upturned side of the chosen card is red. Then the desired probability is obtained by. We want to find* $\mathbb{P}(RB \mid R)$. *From Bayes' Formula we have:*

$$\begin{aligned} \mathbb{P}(RB \mid R) &= \frac{\mathbb{P}(RB \cap R)}{\mathbb{P}(R)} \\ &= \frac{\mathbb{P}(R \mid RB)\mathbb{P}(RB)}{\mathbb{P}(R \mid RR)\mathbb{P}(RR) + \mathbb{P}(R \mid BB)\mathbb{P}(BB) + \mathbb{P}(R \mid RB)\mathbb{P}(RB)} \\ &= \frac{1/2 \cdot 1/3}{1 \cdot 1/3 + 0 \cdot 1/3 + 1/2 \cdot 1/3} = 1/3 \end{aligned}$$

4.4.3: Remark

Again, remember that the denominator in Bayes' formula is given by the Law of Total Probabilities. So we could have jus computed the $\mathbb{P}(R)$ *using the LTP and then used the definition of conditional probability to write the desired probability as in first equality and compute again* $\mathbb{P}(RB \cap R) = P(R \mid RB)\mathbb{P}(RB)$.

Example 4.4.7. *At a certain stage of a criminal investigation, the detective in charge is 60 percent convinced of the guilt of a certain suspect. Suppose, however, that a new piece of evidence which shows that the criminal has a certain characteristic (such as left-handedness) is uncovered. If 20 percent of the population possesses this characteristic, how certain of the guilt of the suspect should the inspector now be if it turns out that the suspect has the characteristic?*

Solution. *Letting G denote the event that the suspect is guilty and C the event that they possess the characteristic of the criminal. Then*

$$\begin{aligned} \mathbb{P}(G \mid C) &= \frac{\mathbb{P}(G \cup C)}{\mathbb{P}(C)} \\ &= \frac{\mathbb{P}(C \mid G)\mathbb{P}(G)}{\mathbb{P}(C \mid G)\mathbb{P}(G) + \mathbb{P}(C \mid G^c)\mathbb{P}(G^c)} \\ &= \frac{1(0.6)}{1(0.6) + (0.2)(0.4)} \approx 0.882 \end{aligned}$$

Example 4.4.8. *If a student chooses a playlist created by her roommate, she finds a song of her favorite type 62% of the time. On the other hand, if she chooses a playlist created by her friend, she find a song of her favorite type 88% of the time. These are the only choices she has for playlists. She chooses her roommate's playlist 40% of the time. If the student is listening to his favorite type of music, what is the probability she is listing to a song from her roommate's playlist?*

Solution. *Let A be the event the song is from her roommate's playlist and B be the event the song is one of her favorite type. We Know: $P\mathbb{P}(A) = 0.4$, $P(A^c) = 0.6$, $P(B \mid A) = 0.62$ and $\mathbb{P}(B \mid A^c) = 0.88$. We need $\mathbb{P}(A \mid B)$. Bayes's Formula gives:*

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A)\mathbb{P}(B \mid A)}{\mathbb{P}(A)\mathbb{P}(B \mid A) + \mathbb{P}(A^c)\mathbb{P}(B \mid A^c)} = \frac{(0.40)(0.62)}{(0.40)(0.62) + (0.6)(0.88)} = 0.32$$

4.5 Conditional probability as just a probability

Let us recall the axioms of probability from Lecture Notes 2:

Definition 4.5.1: Axioms of Probability

A probability $\mathbb{P}$ is a rule (a "function") which assigns a number (chances to occur) to any event E in a Sample Space S that satisfies the following **axioms:**

(i) For any $E \subset S$,

$$0 \leq \mathbb{P}(E) \leq 1,$$

(ii)

$$\mathbb{P}(S) = 1, \text{ The probability if the whole space is 1,}$$

(iii) For any sequence of events E_j, $j \geq 1$ which are pairwise disjoint (mutually exclusive), that is $E_i \cap E_j = \varnothing$,

$$\mathbb{P}\left(\bigcup_{j=1}^{\infty} E_j\right) = \sum_{j=1}^{\infty} \mathbb{P}(E_j)$$

Think of $\mathbb{P} : \mathcal{F} \to \mathbb{R}$ as a "function" with "domain" $\mathcal{F}$ and "range" in the closed interval $[0, 1]$ satisfying (ii) and (iii). The triple $(S, \mathcal{F}, \mathbb{P})$ is called a **probability space.**

With this we have:

Theorem 4.5.1

Fix an event F with $\mathbb{P}(F) > 0$. For any event E in the sample space S, we define

$$\mathbb{P}'(E) = \frac{\mathbb{P}(E \cap F)}{\mathbb{P}(F)}.$$

Then $\mathbb{P}'$ is a probability.

Proof. In order to show that $\mathbb{P}'$ is a probability, we must verify (i)–(iii) in the axioms of probability. For (i) observe that since $E \cap F \subset F$ we have $\mathbb{P}(E \cap F) \leq \mathbb{P}(F)$ or what is the same from our definition of $\mathbb{P}'$,

$$0 \le \mathbb{P}'(E) = \frac{\mathbb{P}(E \cap F)}{\mathbb{P}(F)} \le 1$$

which gives (i) for $\mathbb{P}'$.

For (ii) we apply the definition with $E = S$. Then

$$\mathbb{P}'(S) = \frac{\mathbb{P}(S \cap F)}{\mathbb{P}(F)} = \frac{\mathbb{P}(F)}{\mathbb{P}(F)} = 1$$

and this shows (ii).

For (iii), we just use the definition of $\mathbb{P}'$ in terms of $\mathbb{P}$ and the fact that $\mathbb{P}$ is a probability. Let E_j, $j \ge 1$ be disjoint, that is $E_i \cap E_j = \varnothing$, $i \ne j$. Then the sets $E_j \cap F$ are also disjoint. The distribute law for unions and intersections and Axiom (iii) applied to $\mathbb{P}$ gives

$$\mathbb{P}\left(\left(\bigcup_{j=1}^{\infty} E_j\right) \cap F\right) = \mathbb{P}\left(\bigcup_{j=1}^{\infty} (E_j \cap F)\right) = \sum_{j=1}^{\infty} \mathbb{P}(E_j \cap F).$$

Thus for $\mathbb{P}'$ we have

$$\begin{aligned} \mathbb{P}'\left(\bigcup_{j=1}^{\infty} E_j\right) &= \frac{\mathbb{P}\left(\left(\bigcup_{j=1}^{\infty} E_j\right) \cap F\right)}{\mathbb{P}(F)} \\ &= \frac{1}{\mathbb{P}(F)} \sum_{j=1}^{\infty} \mathbb{P}(E_j \cap F) \\ &= \sum_{j=1}^{\infty} \left(\frac{\mathbb{P}(E_j \cap F)}{\mathbb{P}(F)}\right) \\ &= \sum_{j=1}^{\infty} \mathbb{P}'(E_j). \end{aligned}$$

■

Example 4.5.1. *This is from Example 1.4 above: Three cards are randomly selected, without replacement from an ordinary deck of 52 playing cards. What is the conditional probability that the first card selected is not a spade given that the second and third cards are spades.*

Solution. *With the notation of Example 1.4 we already know that*

$$\mathbb{P}(S_1 \mid S_2 \cap S_3) = 11/50$$

and so since $\mathbb{P}'$ is a probability,

$$\mathbb{P}(S_1^c \mid S_2 \cap S_3) = 1 - \mathbb{P}(S_1 \mid S_2 \cap S_3) = 1 - 11/50 = 39/50.$$

Chapter 5

Week 5: Discrete Random Variables

5.1 Random variables

When we perform an experiment, we are interested in some function of the outcomes, instead of the actual outcome. Want to attach a numerical value. This leads is to a definition of random variables.

Definition 5.1.1: Random Variables, r.v.

Suppose we are given a probability space $(S, \mathcal{F}, \mathbb{P})$ That is, S is a sample space, $\mathcal{F}$ is the collection of all events and $\mathbb{P}$ is a probability on $\mathcal{F}$. A random variable is a function X from the sample space S into the real numbers $\mathbb{R}$. That is, $X : S \to \mathbb{R}$.

Remark 5.1.1. *Random variables are usually denoted by capital letters such as X, Y, Z. The short notation r.v. is used for "random variables". If $x \in \mathbb{R}$ the set $X^{-1}(x) = \{s \in S : X(s) = x\}$. If X does not take the value x then this set is $\varnothing$. We can think of X as a numerical value that is random, like as if X is a random number.*

Example 5.1.1. *Toss a coin. Let X be 1 if heads and $X = 0$ if tails. That is, with $\Omega = S = \{H, T\}$, $X : S \to \mathbb{R}$ by $X(H) = 1$ and $X(T) = 0$. We can do calculus on real numbers but not on $\Omega = S = \{H, T\}$.*

Example 5.1.2. *Roll a die. Let X denote the outcome. That is $S = \{1, 2, 3, 4, 5, 6\}$ and $X : S \to \mathbb{R}$ by $X(1) = 1, X(2) = 2, \ldots X(6) = 6$.*

Example 5.1.3. *Roll a die, define*

$$Y = \begin{cases} 1 & \text{outomce= odd} \\ 0 & \text{outomce= even} \end{cases}$$

That is,

$$Y(s) = \begin{cases} 1 & s = odd \\ 0 & s = even \end{cases}.$$

and even more explicitly: $Y(2) = Y(4) = Y(6) = 0$ *and* $Y(1) = Y(3) = Y(5) = 1$.

A common question: **What values can X attain?**. That is, what is the range of X? In the previous example the range is $\{0, 1\}$, two points.

Example 5.1.4. *Toss a coin 10 times. Let X be the number of heads showing What random values can X be? The answer is clear:* $0, 1, 2, \ldots, 10$.

Example 5.1.5. *Toss a coin* 3 *times and again let X be the number of heads. We may assign probabilities to the different values of the random variable:*

$$\begin{aligned}
\mathbb{P}(X = 0) &= \mathbb{P}((T, T, T)) = \frac{1}{2^3} = \\
\mathbb{P}(X = 1) &= \mathbb{P}((T, T, H), (T, H, T), (H, T, T)) = \frac{3}{8} \\
\mathbb{P}(X = 2) &= \mathbb{P}((T, H, H), (H, H, T), (H, T, H)) = \frac{3}{8} \\
\mathbb{P}(X = 3) &= \mathbb{P}((H, H, H)) = \frac{1}{8}.
\end{aligned}$$

Note that since X must take the values of 0 *through* 3 *then*

$$1 = \mathbb{P}\left(\bigcup_{i=0}^{3} \{X = i\}\right) = \sum_{i=0}^{3} \mathbb{P}(X = i),$$

which makes sense from our previous calculation.

Example 5.1.6. *Let X be the amount of liability (damages) a driver incurs in a year. Then* $X : S \to [0, \infty)$.

5.2 Discrete random variables

All our examples above are examples of discrete random variables. That is, the random variables take only finite or countably many points. Here is the formal definition.

Definition 5.2.1

A random variable that can take on at most countable number of possible values is called a discrete random variable.

Definition 5.2.2

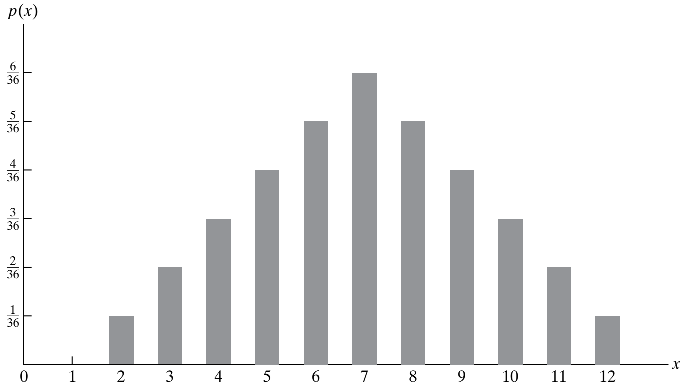
For a discrete random variable on a probability space $(S, \mathcal{F}, \mathbb{P})$ we can define the **probability mass function** (pmf) of X by $p_X(x) = \mathbb{P}\left(s \in S : X(s) = x\right) = \mathbb{P}(X^{-1}(x))$.
Note: When there is no danger of confusion we may drop the X in the pmf and simply write $p(x)$ for $p_X(x)$.

If X takes the values $x_1, x_2, x_3 \dots,$ that is, $X : S \to \{x_1, x_2, \dots\} \subset \mathbb{R}$, then we must have $0 < p_X(x_i) \leq 1$, for $i = 1, 2, \dots.$ and $p\left(x\right) = 0$ for all other values of x not attained. Recall that if X does not take the value of x then $X^{-}(x) = \varnothing$ and hence $P(X^{-}(x)) = \mathbb{P}(s \in S : X(s) = x) = 0$ for such $x's$. So, we really only need to concentrate on those values for which $p_X(x) > 0$. Also must have

$$\sum_{i=1}^{\infty} p_X(x_i) = \mathbb{P}(S) = 1. \tag{5.2.1}$$

5.2.1 Examples

Example 5.2.1. *The pmf function $p(x) = p_X(x) = \mathbb{P}(s \in S : X(s) = x)$ of a random variable is often represent graphically as in the following example: Suppose we roll two dice and X is the sum of the outcomes. Then*



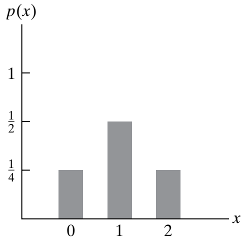
graph of pmf

This represent the probability mass function of the r.v. which gives the sum when two dice are rolled.

Example 5.2.2. *Suppose $p_X(x)$ is given by*

$$p_x(x) = \begin{cases} \frac{1}{4} & x = 0 \\ \frac{1}{2} & x = 1, \\ \frac{1}{4} & x = 2, \\ 0 & \textit{otherwise} \end{cases}$$

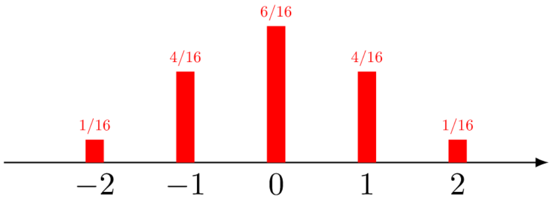
That is, X takes the three values $0, 1, 2$ with probability 1/4, 1/2. 1/4, respectively. Its graph is



graph of pmf

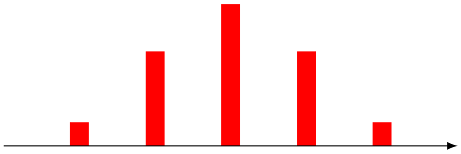
Example 5.2.3. *In the literature or different books you find other graphical (but similar) representations of pmf's. Here is another example represented differently:*

$$p_X(x) = \begin{cases} \frac{1}{16} & x = -2 \\ \frac{4}{16} & x = -1, \\ \frac{6}{16} & x = 0, \\ \frac{4}{16} & x = 1, \\ \frac{1}{16} & x = 2 \\ 0 & \textit{otherwise} \end{cases}$$



graph of pmf

or even just

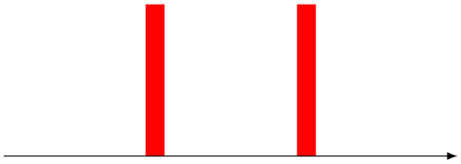


graph of pmf

However, the last graph is not as useful without some additional information on the height of the bars.

Example 5.2.4. *If we toss a coin and let $X(H) = 1$ and $X(T) = -1$, then*

$$p_X(x) = \begin{cases} \frac{1}{2} & x = -1 \\ \frac{1}{2} & x = 1, \\ 0 & \text{otherwise} \end{cases}$$



graph of pmf

Example 5.2.5. *Three balls are randomly selected without replacement from an urn containing 20 balls numbered 1 through 20. If we bet that at least one of the balls that are drawn has a number greater or equal to 17, what is the probability that we win the bet?*

Solution. *Let X denote the largest value selected. X is a random variable on the sample space S which contains $\begin{pmatrix} 20 \\ 3 \end{pmatrix}$ outcomes. Since 3 is the smallest value in the selection (remember no replacement), X takes the values in $\{3, 4, 5, \ldots, 19, 20\}$. What is $p_X(i)$? Since the number of selections that give the event $\{X = i\}$ is just the number of selections that result in the ball numbered i and any two of the balls numbered 1 through i-1 we have*

$$\mathbb{P}(X = i) = \frac{\begin{pmatrix} 1 \\ 1 \end{pmatrix} \begin{pmatrix} i-1 \\ 2 \end{pmatrix}}{\begin{pmatrix} 20 \\ 3 \end{pmatrix}}$$

The probability that we will win is given by

$$\mathbb{P}(X \geq 17) = \sum_{i=17}^{20} \mathbb{P}(X = i) = \frac{1}{1140} \sum_{i=17}^{20} \begin{pmatrix} i-1 \\ 2 \end{pmatrix} \approx 0.5$$

There are many *pmf* that arise in applications. In the next chapter we will study many of these and their properties. Here is one.

Example 5.2.6. *Suppose the p.m.f. of a r.v. X is given by $p(i) = c\frac{\lambda^i}{i!}$, for $i = 0, 1, 2, \ldots$ where $\lambda > 0$ is a parameter. (a) What is the value of c? (b) What is $\mathbb{P}(X > 2)$?*

Solution. *(a) By (6.6.1) we must have*

$$\sum_{i=0}^{\infty} \mathbb{P}(X=i) = c\sum_{i=0}^{\infty} \frac{\lambda^i}{i!} = 1$$

Since $\sum_{i=0}^{\infty} \frac{\lambda^i}{i!} = e^{\lambda}$, we must have that $c = e^{-\lambda}$. In fact, we see that $\mathbb{P}(X=0) = p_X(0) = e^{-\lambda}\frac{\lambda^0}{0!} = e^{-\lambda}$ and

$$\begin{aligned} \mathbb{P}(X>2) &= 1-\mathbb{P}(X\leq 2) \\ &= 1-\mathbb{P}(X=0)-\mathbb{P}(X=1)-\mathbb{P}(X=2) \\ &= 1-p_X(0)-p_X(1)-p_X(2) \\ &= 1-e^{-\lambda}-\lambda e^{-\lambda}-\frac{\lambda^2 e^{-\lambda}}{2}. \end{aligned}$$

5.3 Expected value of random variables

We see that a random variable X is a "function" that maps the sample space into some subset of the real line. That is, $X : S \to \mathbb{R}$, but the "function" is random. For discrete r.v. X, the values that it takes can be listed as x_1, x_2, etc., and these can be finite or infinite. It is reasonable to set the "expect value" of the r.v. to be "the values that it takes with positive probability times the probability that it takes those values." The formal definition is:

Definition 5.3.1

Let X have a *pmf* $p_X(x)$. We define the **expectation**, or **expected value** of X to be

$$\mathbb{E}[X] = \sum_{\{x:p(x)>0\}} x p_X(x). \tag{5.3.1}$$

If X takes the values $x_1, x_2, x_3 \dots$ with positive probability we simply write

$$\mathbb{E}[X] = \sum_{i=1}^{\infty} x_i p_X(x_i) = \sum_{i=1}^{\infty} x_i \mathbb{P}(s\in S : X(s)=x_i).$$

Note: If X takes on negative values, we need to be careful since the series may not converge.

5.3.1: Important Remark

As you can see, the expectation of the r.v. X is the weighted average (mean) of the values it takes with the probability that it takes them. That is, if the r.v. takes the points $\{x_1, x_2, \ldots, x_n\}$ instead of taking the average of these numbers (multiply each vale x_i by 1/n and sum them up), that is,

$$mean = \sum_{j=1}^{n} \frac{x_j}{n} = \frac{1}{n} \sum_{j=1}^{n} x_j,$$

we multiply each value x_i by $p(x_i)$ and sum them up to get

$$EX = \sum_{j=1}^{n} x_j p(x_j).$$

For this reason, $\mathbb{E}(X)$ is often (in fact quite often) referred to as the mean of X or the average of X. In this course, we use all three names.

5.3.1 Examples

Example 5.3.1. *Consider the coin flips. (a) Let $X(H) = 0$ and $X(T) = 1$. (b) Let $X(H) = -1$ and $X(T) = 1$. What is $\mathbb{E}X$?*

Solution. *For (a) we have*

$$\mathbb{E}X = 0 \cdot \frac{1}{2} + 1 \cdot \frac{1}{2} = \frac{1}{2}.$$

For (b) we have

$$\mathbb{E}X = -1 \cdot \frac{1}{2} + 1 \cdot \frac{1}{2} = 0.$$

Example 5.3.2. *Let X be the outcome when we roll a fair die. What is $\mathbb{E}X$?*

Solution.

$$\begin{aligned} \mathbb{E}X &= 1\left(\frac{1}{6}\right) + 2\left(\frac{1}{6}\right) + \cdots + 6\frac{1}{6} \\ &= \frac{1}{6}(1+2+3+4+5+6) = \frac{21}{6} = \frac{7}{2} = 3.5 \end{aligned}$$

Note that X is never 3.5 so the r.v. never takes the value $\mathbb{E}(X)$.

Example 5.3.3. *Let X be the number or tornados in Indiana per year. Meaning that the random variable X can be any number $X = 0, 1, 2, 3, \ldots$. Suppose the state did some analysis and found out that*

$$\mathbb{P}(X = i) = \frac{1}{2^{i+1}}.$$

What is $\mathbb{E}X$? That is, what is the expected number of tornados per year in Indiana?

Solution. *Note that X is infinite, but still countable, hence still discrete. Note that he pmf is given by*

$$p_X(i) = \begin{cases} \frac{1}{2} & i = 0, \\ \frac{1}{4} & i = 1, \\ \frac{1}{8} & i = 2, \\ \vdots & \vdots \\ \frac{1}{2^{n+1}} & i = n. \end{cases}$$

We have that

$$\begin{aligned} \mathbb{E}X &= 0 \cdot p_x(0) + 1 \cdot p_X(1) + 2 \cdot p_X(2) + \cdots \\ &= 0 \cdot \frac{1}{2} + 1\frac{1}{2^2} + 2\frac{1}{2^3} + 3\frac{1}{2^4} + \cdots \\ &= \frac{1}{2^2}\left(1 + 2\frac{1}{2} + 3\frac{1}{2^2} + \cdots\right) \\ &= \frac{1}{4}\left(1 + 2x + 3x^2 + \cdots\right), \textit{with } x = \frac{1}{2} \\ &= \frac{1}{4}\frac{1}{(1-x)^2} = \frac{1}{4\left(1-\frac{1}{2}\right)^2} = 1. \end{aligned}$$

Example 5.3.4. *Consider the standard deck of 52 cards. From it you draw a card and if it is the the ace of hearts stop. If not, you replace the card, shuffle the deck again, and make a new selection. What is the expected number of draws until the ace of hearts appears for the first time?*

Solution. *Let X be the number of cards drawn until the ace appears for the first time. Then $\mathbb{P}(X = n) = (51/52)^{n-1}(1/52)$, for $n \geq 1$. And*

$$\begin{aligned} \mathbb{E}X &= 1 \cdot \mathbb{P}(X = 1) + 2\mathbb{P}(X = 2) + 3\mathbb{P}(X = 3) + \cdots \\ &= \frac{1}{52}\sum_{n=1}^{\infty} n(51/52)^{n-1} \end{aligned}$$

Question: What is

$$\sum_{n=1}^{\infty} n(51/52)^{n-1} \quad ?$$

Note that this is equal to $\sum_{n=1}^{\infty} nx^{n-1}$, with $x = 51/52$. Can you compute this quantity?

Hint:

$$\sum_{n=1}^{\infty} x^n = x\sum_{n=0}^{\infty} x^n = \frac{x}{1-x}$$

for $0 \leq x < 1$.

5.4 Properties of expected value

We start out by given an alternate definition of the expected value of a discrete random variable.

Definition 5.4.1

Let S be the sample space which consists of countably many outcomes. That is, $S = \{s_1, s_2, \dots\}$. Define

$$\mathbb{E}X = \sum_{s\in S} X(s)\mathbb{P}\left(\{s\}\right) = \sum_{i=1}^{\infty} X(s_i)\mathbb{P}(\{s_i\}) \tag{5.4.1}$$

Example 5.4.1. . *Let $S = \{1, 2, 3, 4, 5, 6\}$ and $X(1) = X(2) = 1$ and $X(3) = X(4) = 3$ and $X(5) = X(6) = 5$*

We know $X = 1, 3, 5$ with $p(1) = p(3) = p(5) = \frac{1}{3}$. Our first definition of $\mathbb{E}X$ gives that

$$\mathbb{E}X = 1 \cdot \frac{1}{3} + 3\frac{1}{3} + 5\frac{1}{3} = \frac{9}{3} = 3$$

Our second definition gives, again with $S = \{1, 2, 3, 4, 5, 6\}$ that

$$\begin{aligned} \mathbb{E}X &= X(1)\mathbb{P}\left(\{1\}\right) + \cdots + X(6) \cdot \mathbb{P}\left(\{6\}\right) \\ &= 1\frac{1}{6} + 1\frac{1}{6} + 3\frac{1}{6} + 3\frac{1}{6} + 5\frac{1}{6} + 5\frac{1}{6} = 3. \end{aligned}$$

Thus the two definitions of of $E(X)$ coincide.

Remark 5.4.1. *What is the difference between the two definition? Answer:*

- *Definition 1 sums over all the values that X can attain and only cares about those. (The Range of X)*
- *Definition 2 sums over all possible outcomes outcomes of the experiment, all element in S. (The Domain of X)*

Proposition 5.4.1

If X is a discrete random variable on $S = \{s_1, s_2, \dots\}$ is then the two definitions (5.3.1), (5.4.1) give the same value.
That is,

$$\mathbb{E}\left[X\right] = \sum_{\{x:p(x)>0\}} x p_X(x) = \sum_{\{s \in S\}} X(s)\mathbb{P}(\{s\}) \tag{5.4.2}$$

Remark 5.4.2. *Note that if X is a constant, that is, if X takes only one value, say c, then $\mathbb{E}X = c$ and this follows trivially from either definition but the second makes it very clear. Of course, this uses the fact that $\sum_{\{s \in S\}} \mathbb{P}(\{s\}) = \mathbb{P}(S) = 1$.*

Proof. We start with the first definition. Let $X = x_1, x_2, \dots$

$$\begin{aligned}
\mathbb{E}X &= \sum_{x_i} x_i p(x_i) \\
&= \sum_{x_i} x_i \mathbb{P}\left(X = x_i\right) \\
&= \sum_{x_i} x_i \sum_{s \in \{s : X(s) = x_i\}} \mathbb{P}\left(s\right) \\
&= \sum_{x_i} \sum_{s \in \{s : X(s) = x_i\}} x_i \mathbb{P}\left(s\right) \\
&= \sum_{x_i} \sum_{s \in \{s : X(s) = x_i\}} X(s) \mathbb{P}\left(s\right) \\
&= \sum_{s \in S} X(s) \mathbb{P}\left(s\right),
\end{aligned}$$

where we used that each $E_i = \{s : X(s) = x_i\}$ are mutually exclusive events that union up to S. ■

Using our second definition, we can prove linearity of the expectation.

Theorem 5.4.1

(Linearity) If X and Y are discrete random variables and $a, b \in \mathbb{R}$ then
(a) $\mathbb{E}\left[aX + bY\right] = a\mathbb{E}X + b\mathbb{E}Y$.
(b) More generally, if $X_1, X_2, \dots, X_n$ are n random variables and $a_k \in \mathbb{R}$, then

$$\mathbb{E} \sum_{k=1}^{n} a_k X_k = \sum_{k=1}^{n} a_k \mathbb{E} X_k$$

Proof. We have that

$$\begin{aligned}\mathbb{E}\left[aX+bY\right] &= \sum_{\omega\in S}\left(aX(\omega)+bY(\omega)\right)\mathbb{P}\left(\omega\right)\\ &= \sum_{\omega\in S}\left(aX(\omega)\mathbb{P}\left(\omega\right)+bY(\omega)\mathbb{P}\left(\omega\right)\right)\\ &= a\sum_{\omega\in S}X(\omega)\mathbb{P}\left(\omega\right)+b\sum_{\omega\in S}Y(\omega)\mathbb{P}\left(\omega\right)\\ &= a\mathbb{E}X+b\mathbb{E}Y.\end{aligned}$$

■

5.5 Expectation of functions of r.v.'s

Suppose we have a function $f:\mathbb{R}\to\mathbb{R}$, like $f(x)=x^2$, $f(x)=x^{20}$ (or more generally $f(x)=x^n$, $n=1,2,\ldots$, $f(x)=e^x$. How do we find find the expectation of $f(X)$? That is, what is, for example, $\mathbb{E}X^2$ or $\mathbb{E}e^X$ or $\mathbb{E}\sin(X)$? One thing we must be aware of is that this is not the same as finding the expectation of X and evaluating the function at this value. That is, in general, for example, $(\mathbb{E}X)^2\neq\mathbb{E}X^2$ and similarly for other functions. Here is an example:

Example 5.5.1. *Let X denote a random variable with*

$$\begin{aligned}\mathbb{P}\left(X=-1\right) &= 0.2,\\ \mathbb{P}\left(X=0\right) &= 0.5\\ \mathbb{P}\left(X=1\right) &= 0.3\end{aligned}$$

Let $Y=X^2$. Then Y takes the values $\{0,1\}$. The pmf is given by

$$p_X(1)=\mathbb{P}(\{X=1\,or-1\})=0.2+0.3=0.5$$

and

$$p_Y(0)=0.5.$$

Thus

$$\mathbb{E}Y=0\cdot 0.5+1\cdot 0.5=0.5$$

On the other hand, $\mathbb{E}X=0.3-0.2=0.1$ and

$$\left(\mathbb{E}X\right)^2=0.01$$

Theorem 5.5.1

If X is a discrete random variable that takes values in $\{x_1, x_2, x_3, \ldots\}$ with respective probability mass function $p_X(x_i)$, then for any real valued function $f : \mathbb{R} \to \mathbb{R}$ we have that

$$\mathbb{E}\left[f\left(X\right)\right] = \sum_{i=1}^{\infty} f\left(x_i\right) p_X(x_i).$$

Proof. We consider the r.v. $Y = f(X)$ with $\{y_1, y_2, \ldots$. But we know that

$$y_j = f(x_i).$$

But we need to be careful as there can be more than one value x_i such that $y_j = f(x_i)$. Using the definition of expectation we have that

$$\begin{aligned}
\mathbb{E}\left[Y\right] &= \sum_j y_j \mathbb{P}\left(Y = y_j\right) \\
&= \sum_j y_j \mathbb{P}\left(f(X) = y_j\right) \\
&= (\star).
\end{aligned}$$

Now, we will group the sum as follows:

$$\begin{aligned}
\mathbb{P}\left(f(X) = y_j\right) &= \mathbb{P}\left(\bigcup_{i:f(x_i)=y_j} \left(f(x_i) = y_j\right)\right) \\
&= \sum_{\{i:f(x_i)=y_j\}} p(x_i).
\end{aligned}$$

Thus plugging this back into $(\star)$ we have that

$$\begin{aligned}
\mathbb{E}\left[Y\right] &= \sum_j y_j \sum_{i:f(x_i)=y_j} p(x_i) \\
&= \sum_j \sum_{\{i:f(x_i)=y_j\}} y_j p(x_i) \\
&= \sum_j \sum_{\{i:f(x_i)=y_j\}} f(x_i)p(x_i) \\
&= \sum_{i=1}^{\infty} f(x_i)p(x_i),
\end{aligned}$$

as needed. ■

5.5.1: Warning, Warning!

Be careful with the properties of expectations. The following are **not true** in general:

- $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$ (This does hold under the assumption of independence–next week Lecture Notes!)
- $\mathbb{E}[X^2] = [\mathbb{E}X]^2$ (Example above)
- $\mathbb{E}[X/Y] = \mathbb{E}[X]/\mathbb{E}[Y]$
- $\mathbb{E}[f(X)] = f(\mathbb{E}X)$

5.6 Variance

The expectation of a r.v. is quantity that help us differentiate different r.v.'s, but it doesn't tell us how spread out values are. Given a random variable X, its variance measures its squared deviation from its mean. In other words, the variance measures how spread out the values of X are.

Example 5.6.1. *Let X, Y, Z be r.v.'s define by*

$$\begin{aligned} X &= 0 \textit{ with probability 1} \\ Y &= \begin{cases} -1 & p = \frac{1}{2} \\ 1 & p = \frac{1}{2} \end{cases} \\ Z &= \begin{cases} -100 & p = \frac{1}{2} \\ 100 & p = \frac{1}{2} \end{cases}. \end{aligned}$$

The expectation of these r.v.'s are all 0*.*

But there is much greater spread in Z than Y and in Y than X. Thus expectation is not enough to detect spread, or variation.

Definition 5.6.1

If X is a r.v with mean $\mu = \mathbb{E}X$, the variance of X, denoted by $\mathrm{Var}(X)$, is defined by

$$\mathrm{Var}\,(X) = \mathbb{E}\left[(X-\mu)^2\right].$$

The variance is also often denoted by σ^2.

Remark 5.6.1.

We prove an alternate formula for the variance. (The technique of using linearity is important here!)

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

Theorem 5.6.1

We have that
$$\mathrm{Var}\left(X\right) = \mathbb{E}\left[X^2\right] - \left(\mathbb{E}\left[X\right]\right)^2.$$

Example 5.6.2. *Let X be the outcome of rolling a die. What is Var(X)? Previously we calculated that $\mathbb{E}X = \frac{7}{2}$. Thus we only need to calculate the second moment:*

$$\begin{aligned}
\mathbb{E}X^2 &= 1^2\left(\frac{1}{6}\right) + \cdots + 6^2\frac{1}{6} \\
&= \frac{91}{6}.
\end{aligned}$$

Using our formula we have that

$$Var(X) = \mathbb{E}\left[X^2\right] - \left(\mathbb{E}\left[X\right]\right)^2 = \frac{91}{6} - \left(\frac{7}{2}\right)^2 = \frac{35}{12}.$$

Here is a useful formula:

Proposition 5.6.1

For constants a, b we have that $\mathrm{Var}\left(aX+b\right) = a^2\mathrm{Var}\left(X\right)$.

Proof. We compute

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

■

Definition 5.6.2

We define

$$\sigma = SD\,(X) = \sqrt{\text{Var}(X)}$$

to be the standard deviation of X.

Example 5.6.3. *A game of dice consists of rolling a die three times and counting the number of times a 6 appears. You pay \$3 to play but you win \$4 for every 6 you get. If X represents the total number of 6's you get, (a) what are are your total winnings, (b) how much do you expect to win and (c) what is the variance of your winnings?*

Solution. *(a) Your winnings are $4X - 3$ since it costs you \$3 to play the game.*

(b) Your expected winnings are

$$\mathbb{E}(4X - 3) = 4\mathbb{E}X - 3 = 4(1/2) - 3 = -1$$

since $\mathbb{E}X = 1/2$. Why and how?

(c) From above we now with $a = 4$ and $b = -1$ we have

$$Var(4X - 3) = 4^2\,Var(X) = 13[\mathbb{E}X^2 - 1/4] = 20/3,$$

where we use the fact that $\mathbb{E}X^2 = 2/3$. Why and how?

5.6.1: Summary

(i) Expected value: For a r.v X, $E[X]$ represents the mean value of X.

(ii) Variance: For a r.v X, $Var(X)$ represents the deviation from the mean (the dispersion of X with respect its mean value). A greater $Var(X)$ means

- The "system" (experiment) represented by X has a lot of randomness–you will not expect to find the values of X near its mean.
- This system is more "unpredictable."

5.7 Summary of basic properties

Let us state the basic properties of r.v. discussed above.

Properties 5.7.1

(i) Definition: The expectation of a discrete random variable X is

$$\mathbb{E}X = \sum_x x\mathbb{P}(X = x)$$

It is the average value of X relative to the weights given by the mass distribution function $p_X(x) = P(X = x)$

(ii) Expectation of Indicators: For any event A in the sample space S, we define the random variable $\mathbb{1}_A(s)$, called the indicator of A, to be

$$\mathbb{1}_A(s) = \begin{cases} 1 & s \in B \\ 0 & s \notin B \end{cases}$$

Then

$$\mathbb{E}\mathbb{1}_A = \mathbb{P}(A)$$

(iii) Constants: The expectation of a constant random variable c is just

$$\mathbb{E}c = c$$

and

$$\mathbb{E}(cX) = c\mathbb{E}X$$

(iv) Functions of random variables:

$$\mathbb{E}g(X) = \sum_x g(x)\mathbb{P}(X = x)$$

(v) Sums: Expectation of sums equal sums of expectations, that is

$$\mathbb{E}(X + Y) = \mathbb{E}X + bEY$$

(vi) Products: In general

$$\mathbb{E}(XY) \neq (\mathbb{E}X)(\mathbb{E}Y)$$

However, if the random variables X and Y are independent, then

$$\mathbb{E}(XY) = (\mathbb{E}X)(\mathbb{E}Y)$$

5.8 Cumulative distribution function

Definition 5.8.1

Given a random variable X we define the function $F : \mathbb{R} \to [0,1]$ by

$$F_X(x_0) = \mathbb{P}\left(X \leq x_0\right) = \sum_{x \leq x_0} p_X(x), \text{ for any } -\infty < x < \infty.$$

The function is called the cumulative distribution function or distribution function in short of X. The notation CDF or c.d.f is often used. As with $p_X(x)$, the subindex X will be often dropped and we will just write $F(x)$.

5.8.1 Examples and properties

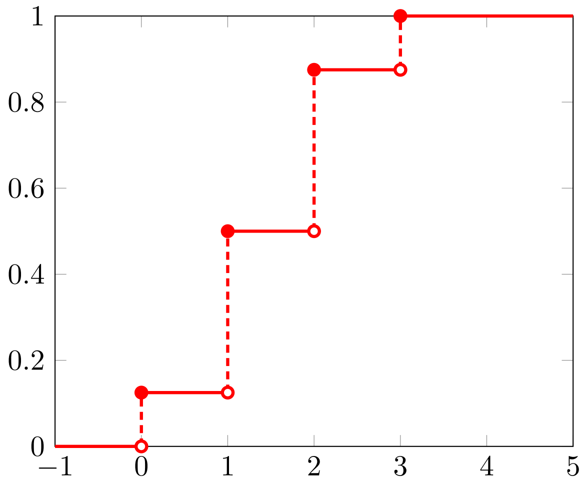
Example 5.8.1. *If we take the Example in 5.1.1 flipping a coin three times, we have the p.m.f.*

$$\begin{aligned} p_X(0) = \mathbb{P}\left(X = 0\right) &= \frac{1}{8} \\ p_X(1) = \mathbb{P}\left(X = 1\right) &= \frac{3}{8} \\ p_X(2) = \mathbb{P}\left(X = 2\right) &= \frac{3}{8} \\ p_X(3) = \mathbb{P}\left(X = 3\right) &= \frac{1}{8}. \end{aligned}$$

and the CDF

$$F_X(x) = \begin{cases} 0 & -\infty < x < 0 \\ \frac{1}{8} & 0 \leq x < 1 \\ \frac{4}{8} & 1 \leq x < 2 \\ \frac{7}{8} & 2 \leq x < 3 \\ 1 & 3 \leq x < \infty \end{cases}.$$

Its graph is given by



Graph of F_X

Recalling the definitions from Calculus (Math 161), we see that the c.d.f (CDF) is right continuous has left limit.

Remark 5.8.1. *Here are some simple observations about CDF. from the above graph.*

(i) It is a step function.

(ii This function has jumps, and not continuous everywhere. But it is always right continuous.

(iv) F is nondecreasing. That is if $x \le y$, then $F_X(x) \le F_X(y)$

(v)

$$\lim_{x\to\infty} F_X(x) = 1$$

(vi)

$$\lim_{x\to-\infty} F_X(x) = 0$$

(vii) F_X is right continuous at every x. That is, $\lim_{y\to x^+} F_X(y) = F_X(x)$. (If y approaches x from the right, $F(y)$ approaches $F_X(x)$, exactly as in calculus.)

Proposition 5.8.1

Let $F_X(x)$ be the CDF for some random variable X. Then the following holds:
(a) For any $a \in \mathbb{R}$, we have

$$\mathbb{P}\left(X < a\right) = \lim_{x \to a^-} F_X(x).$$

(b) For any $a \in \mathbb{R}$, we have

$$\mathbb{P}\left(X = a\right) = F_X(a) - \lim_{x \to a^-} F_X(x).$$

Thus $\mathbb{P}(X = a)$ measures the jump of F_X at the point a.

Proof. For part (a) we, set $A_n = \left(X \leq a - \frac{1}{n}\right)$. If $X < a$ then $X \leq a - \frac{1}{n}$, for some $n \geq 1$ and we have

$$(X < a) = \bigcup_{n=1}^{\infty} A_n$$

Since $a - \frac{1}{n} < a - \frac{1}{n+1}$, we see that the sets A_n are increasing. By the "Probability as a continuous set function" property presented in §7 of Lectures Notes Week 2, we have

$$\mathbb{P}(X < a) = \lim_{n \to \infty} \mathbb{P}(A_n)$$

Now you can replace the sequence $a_n = a - \frac{1}{n}$ with any sequence a_n that is increasing towards a, and we get the similar result,

$$\lim_{n \to \infty} F_X\left(a_n\right) = \mathbb{P}\left(X < a\right),$$

since this holds for all increasing sequences a_n towards a, then we've shown that

$$\lim_{x \to a^-} F_X(x) = \mathbb{P}\left(X < a\right).$$

For part (b). We use part (a) and get

$$\begin{aligned}\mathbb{P}\left(X = a\right) &= \mathbb{P}\left(X \leq a\right) - \mathbb{P}\left(X < a\right)\\ &= F_X(a) - \lim_{x \to a^-} F_X(x).\end{aligned}$$

This is what we mean by "$\mathbb{P}(X = a)$ is the size of the jump of the Cumulative Distribution Function (CDF) of the r.v. at a." Equivalently, the pmf (or p.m.f.) of X at the point x (any x) is the is the size of the jump at x of the CDF ■

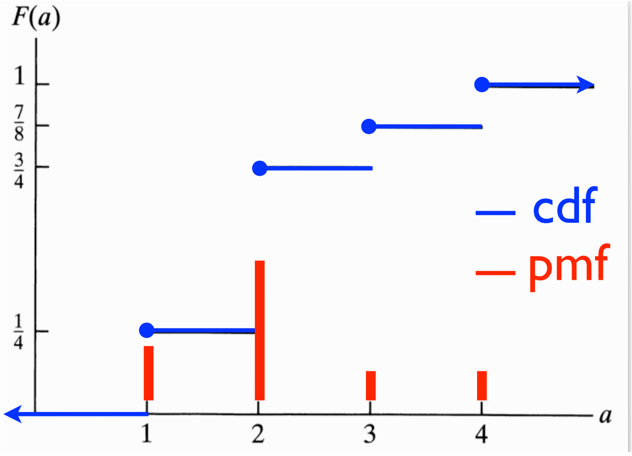
Example 5.8.2. *Let X be a random variable with p.m.f.*

$$\mathbb{P}(1) = \frac{1}{4}, \ \mathbb{P}(2) = \frac{1}{2}, \ \ \mathbb{P}(3) = \frac{1}{8}, \ P(4) = \frac{1}{8}.$$

Then the CDF (cdf) is

$$F(a) = \begin{cases} 0 & -\infty < a < 1 \\ \frac{1}{4} & 1 \leq a < 2 \\ \frac{3}{4} & 2 \leq a < 3 \\ \frac{7}{8} & 3 \leq a < 4 \\ 1 & 4 \leq a < \infty \end{cases} .$$

and a graph of both the p.m.f. and cdf, for comparison, are here.



Chapter 6

Week 6: The Classical Discrete Distributions

6.1 Independent and identically distributed r.v.'s

Definition 6.1.1

Two random discrete random variables X and Y are said to be

(i) identically distributed if their pmf are equal.

(ii) independent if for all x and y,

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

here the "," means "and. Equivalently, if for all x and y the events $A = (X = x)$ and $B = (Y = y)$ are independent

Theorem 6.1.1

(i) If two random variables X and Y are independent then

$$\mathbb{E}[f(X)g(Y)] = \mathbb{E}f(X)\mathbb{E}g(Y)$$

In particular, if X and Y are independent

$$\mathbb{E}[XY] = [\mathbb{E}X][\mathbb{E}Y]$$

(ii) If two random variables X and Y are identically distributed then

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

Proof. (i)

$$\begin{aligned} \mathbb{E}[f(X)g(Y) &= \sum_x \sum_y f(x)g(y)\mathbb{P}(X = x, Y = y) \\ &= \sum_x \sum_y f(x)g(y)\mathbb{P}(X = x)\mathbb{P}(Y = y) \\ &= \sum_x f(x)\mathbb{P}(X = x) \sum_y g(y)\mathbb{P}(Y = y) \\ &= \mathbb{E}f(X)\mathbb{E}g(Y) \end{aligned}$$

(ii)

$$\begin{aligned} \mathbb{E}[f(X)] &= \sum_x f(x)\mathbb{P}(X = x) \\ &= \sum_y f(y)\mathbb{P}(X = y) \\ &= \mathbb{E}[f(Y)] \end{aligned}$$

■

6.2 Well known and widely used discrete distributions

This week we will study various examples of well known discrete distributions that arise in many applications. Our goal is to understand some of the basic properties such as their p.m.f., CDF, expectations, variance, moments, etc. The list includes: (a) Bernoulli, (b) Binomial, (c) Geometric, (d) Negative Binomial, (e) Hypergeometric, (f) Poison and (g) Discrete Uniform. For those in actuarial science, some of these may be useful as you prepare for the exams.

6.3 Bernoulli distribution

Experiment: Success or failure. For instance, you ask one person a yes/no question, where yes is a "success" and no is a "failure". You can flip a coin (heads vs tails), seeing if one part of something is defective (yes or no). The variable X takes two values, success with probability p and failure with probability $1-p$. $\mathbb{P}(X=1)=p$ and $\mathbb{P}(X=0)=1-p$. The probability mass function (pmf) of a Bernoulli is given by

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

The CDF is

$$F_X(x)=\begin{cases}0 & -\infty<x<0\\ 1-p & 0\leq x<1\\ 1 & 1\leq x<\infty\end{cases}.$$

Expected value $\mathbb{E}X=p$ and $\mathbb{E}X^2=p$ also. So, Variance of X, $\text{Var}(X)=p(1-p)$

The notation X ∼Ber(p) is used. We summarize:

6.3.1: Mean and Variance of Ber(p)

If $X\sim$ Ber(p) , then

- $$\mu=\mathbb{E}X=p$$
- $$\mathbb{E}X^2=p$$
- $$\sigma^2=Var(X)=\mathbb{E}X^2-(\mu)^2=p(1-p)$$

6.3.1 Examples

Example 6.3.1. *Suppose 49% of people like broccoli. You select a person at random and ask them if they like broccoli. This is a clear example of a Bernoulli distribution,* $p=0.49$.

Example 6.3.2. *At a certain point in a die game, a roll of 5 or 6 is needed to win, and any other roll at that point will cause the player to lose. At that point, the player earns $15 if she wins pays $9 dollars if she loses. What is the player's expected gain or loss at this point? What is the Variance?*

Solution. *Let X be a Bernoulli random variable that indicates whether the player wins or loses. The probability of winning is $p=2/6=1/3$. $\mathbb{E}X=1/3$. Now, let Y represent the winnings. Then* *when* $X = 1$, $Y = 15$ *and when* $X = 0$, $Y = -9$. *That is,* $Y = f(X)$*and it takes two values* -9 *and* 15 *with probability* $2/3$ *and* $1/3$ *respectively. So,* $\mathbb{E}Y = -9(2/3) + 15(1/3) = -1$

6.4 Binomial distribution

The Bernoulli distributions (r.v.'s) are the building blocks for the Binomial distribution (r.v.'s) Random Variables. Recall from our earlier discussion (week 3) that the Binomial distribution is the number of successes in n independent trials. Equivalently, a Binomial is the sum of n independent Bernoulli random variables with the same probability of success p:

$$X = X_1 + X_2 + \cdots + X_n \tag{6.4.1}$$

We use the notation $X \sim$Bin(n, p). Thus note that if $n = 1$, then Binomial(n, p)=Bernoulli(p)

We say X has a **binomial distribution** with parameters n and p. Let us recall the definition from Lecture Notes Week 3:

Definition 6.4.1: The Binomial distribution

If we have an experiment where the probability of some outcome occurring (we say this is a success) is $0 < p < 1$ and the probability of not occurring (we say failure) is $1 - p$. The experiment is repeated independently n times (n independent trials), the the probability of exactly k success in the n-trials is given by

$$\binom{n}{k} p^k (1-p)^{n-k}$$

and we have

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

That is, X is the number of successes in n *indepedent* trials and we write $X \sim$ Binomial(n,p) or $X \sim bin(n, p)$.

As we showed earlier, the binomial formula gives that $\sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k} = (x+y)^n$. and it follows that

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

Thus, indeed, $p_X(k)$ is a p.m.f. The following Proposition can be iterated to compute higher moments of binomial r.v.'s and also their variance.

Proposition 6.4.1: Moments of Binomial distrobution

Let $X \sim$ Bin(n,p). Then for any $k = 1, 2, \dots$ we have

$$\mathbb{E}X^k = np\mathbb{E}(Y+1)^{k-1}$$

where $Y \sim$ Bin(n-1, p)

Proof. Applying the formula for the expectation of a function of a r.v. from week 5 with $f(x) = x^k$ we have

$$\begin{aligned} \mathbb{E}X^k &= \sum_{i=0}^{n} i^k \binom{n}{i} p^i (1-p)^{n-i} \\ &= \sum_{i=1}^{n} i^k \binom{n}{i} p^i (1-p)^{n-i} \\ &= p \sum_{i=1}^{n} i^k \binom{n}{i} p^{i-1} (1-p)^{n-i} \end{aligned}$$

Substituting the identity

$$i \binom{n}{i} = n \binom{n-1}{i-1}$$

gives

$$\begin{aligned} \mathbb{E}X^k &= pn \sum_{i=1}^{n} i^{k-1} \binom{n-1}{i-1} p^{i-1} (1-p)^{n-i} \\ &= pn \sum_{j=0}^{n-1} (j+1)^{k-1} \binom{n-1}{j} p^j (1-p)^{(n-1)-j} \\ &= np\mathbb{E}(Y+1)^{k-1} \end{aligned}$$

which is exactly want we needed to proof. ■

With $k = 1$, we obtain $\mathbb{E}X = np$. With $k = 2$, we have (applying this with Y and $(n-1)p$)

$$\mathbb{E}X^2 = np\mathbb{E}(Y+1) = np\left(\mathbb{E}Y + 1\right) = np[(n-1)p+1]$$

Since

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

We summarize this in

6.4.1: Mean and Variance of Binomial(n,p)

If $X \sim$ Bin(n,p) , then

- $$\mu = \mathbb{E}X = np$$
- $$\mathbb{E}X^2 = np\mathbb{E}(Y+1) = np\left(\mathbb{E}Y+1\right) = np[(n-1)p+1]$$
- $$\sigma^2 = Var(X) = \mathbb{E}X^2 - (\mu)^2 = np(1-p)$$

An alternative way to compute the mean and variance of a Bin(n, p) r.v. X to use (6.4.1) and write

$$S_n = X_1 + X_2 + \cdots + X_n,$$

where the r.v.'s $X_j \sim$ Ber(p) and independent. Then by the linearity of the expectation we get

$$\mu = \mathbb{E}S_n = \mathbb{E}X_1 + \cdots + \mathbb{E}X_n = np.$$

On the other hand,

$$S_n^2 = \sum_{j=1}^{n} X_j^2 + \sum_{\{j,k,j\neq k\}} X_jX_k$$

and taking expectation and using the independence we get (from Theorem 6.1(ii)) that

$$\mathbb{E}S_n^2 = np + p^2(n^2-n)$$

and this gives the variance.

This exact same proof gives the following

Corollary 6.4.1

Let $X_1, X_2, \cdots X_n$ be independent r.v's Set $S_n = \sum_{j=1}^{n} X_j$. Then

$$\text{Var}(S_n) = \sum_{j=1}^{n} \text{Var}(X_j)$$

If the $X_j's$ are independent and equally distributed (have the same distribution) then

$$\text{Var}(S_n) = n\text{Var}(X_1)$$

6.4.1 Examples

Example 6.4.1. *Flip a coin three times. Let X be the number of heads. Find $Var(X)$.*

Solution. *Let X_j indicate whether the jth flip is heads. Then $X = X_1 + X_2 + X_3$. We know $Var(X_j) = 1/4$ (it is a Bernoulli(1/2)). So, $Var(X) = 3Var(X_1) = 3/4$.*

Problem 6.4.1. *Suppose a couple decides to have 4 children. Assume the probability is the same for having a boy or a girl. Let X denote the number of girls that the couple has. (a) What is the probability the couple has at least one girl? (b). How many children would the couple have to have for there to be at least a 99% chance that at least one of their children is a girl?*

Example 6.4.2. *Roll a die n times. Let X denote the total number of $4's$ or $5's$ that appear. What is the expected number of 4's or 5's and the Variance?*

Solution. . *Here X is a Binomial random variable, with each outcome of 4 or 5 treated as a success. So $p = 2/6 = 1/3$ Then $\mathbb{E}X = n/3$ and Var$(X) = (n/3)(2/3) = 2n/9$*

6.5 The Geometric distribution

A Geometric random variable is the number of independent trials needed until the first success occurs. In other words, a Geometric random variable is the number of independent Bernoulli random variables we need to check until the first that indicates success. More precisely, let X_1, $X_2, \dots$ be independent Bernoulli(p) random variables. Then a Geometric random variable $X =$ smallest value of n so that $X_n = 1$. (Which means $X_1 = X_2 = \cdots X_{n-1} = 0$, i.e., all failures up to $n-1$.) Trials are stopped once success happens. Let p be the probability of having a success in each trial. Then $X =$ "number of trials required until first success occurs". Thus $X \in \{1, 2, 3, 4, \dots\}$.

$$p_X(n) = \mathbb{P}(X = n) = (1-p)^{n-1} p \text{ for } n = 1, 2, 3, 4 \dots.$$

Thus for example, if a Geometric random variable X has the value 10, we can think of this as 9 independent Bernoulli random variables that indicate failure, i.e., $X_1 = X_2 = \dots X_9$ followed by an independent Bernoulli that indicates success $X_{10} = 1$.

Note that

$$\sum_{n=1}^{\infty} p_X(n) = p \sum_{n=1}^{\infty} (1-p)^{n-1} = \frac{p}{1-(1-p)} = 1$$

Thus we do have a probability mass function. We say $X \sim \text{Geo}(p)$.

Next, recall from calculus that by differentiation of the geometric series,

$$\sum_{n=1}^{\infty} n x^{n-1} = \frac{1}{(1-x)^2}.$$

Thus,

$$\begin{aligned} \mathbb{E}X &= \sum_{n=1}^{\infty} n\mathbb{P}\left(X=n\right) \\ &= \sum_{n=1}^{\infty} n\left(1-p\right)^{n-1} p \\ &= \frac{p}{(1-(1-p))^2} = \frac{1}{p}. \end{aligned}$$

In order to compute the variance we differentiate

$$\sum_{n=1}^{\infty} nx^{n-1} = \frac{1}{(1-x)^2}$$

again to get

$$\sum_{n=2}^{\infty} n\left(n-1\right)x^{n-2} = \frac{2}{\left(1-x\right)^3}$$

Thus,

$$\begin{aligned} \mathbb{E}X^2 &= \sum_{n=1}^{\infty} n^2\left(1-p\right)^{n-1} p \\ &= \sum_{n=1}^{\infty} n(n-1)\left(1-p\right)^{n-1} p + \sum_{n=1}^{\infty} n\left(1-p\right)^{n-1} p \\ &= p(1-p)\sum_{n=2}^{\infty} n(n-1)\left(1-p\right)^{n-2} + \frac{1}{p} \\ &= p(1-p)\frac{2}{p^3} + \frac{1}{p} = \frac{2-p}{p} \end{aligned}$$

Thus,

$$\begin{aligned} \operatorname{Var}(X) &= \mathbb{E}X^2 - \left(\mathbb{E}X\right)^2 = \frac{2-p}{p^2} - \frac{1}{p^2} \\ &= \frac{(1-p)}{p^2} \end{aligned}$$

We summarize this.

6.5.1: Mean and Variance of Geo(p)

Set $q = 1 - p$. If $X \sim \text{Geometric}(p)$, then

- $$\mu = \mathbb{E} = \frac{1}{p}$$
- $$\mathbb{E}X^2 = \frac{2-p}{p}$$
- $$\sigma^2 = \text{Var}(X) = \frac{q}{p^2}$$

6.5.1 Examples

Example 6.5.1. *An urn contains N white and M blue balls. Balls are randomly selected, one at a time, until a blue one is obtained. If we assume that each ball selected is replaced before the next one is drawn, what is the probability that (a) exactly n draws are needed? (b) at least k draws are needed?*

Solution. *(a) If we let X denote the number of draws needed to select a blue ball. Then $X \sim$ geometric(p) with $p = \frac{M}{(M+N)}$. Hence,*

$$\mathbb{P}(X = n) = \left(\frac{M}{(M+N)}\right)^p \frac{M}{(M+N)} = \frac{MN^{n-1}}{(M+N)^n}$$

(b) Here we want

$$\begin{aligned} \mathbb{P}(X \geq k) &= p\sum_{n=k}^{\infty}(1-p)^{k-1} = \frac{p(1-p)^{k-1}}{1-(1-p)} \\ &= \left(\frac{N}{M+N}\right)^{k-1} \end{aligned}$$

Example 6.5.2. *Suppose that a person wins a game with probability 0.40, and loses otherwise. If they win, they earn \$5 dollars, and if they lose, then they loses \$4 dollars. The game is played until they they win for the first time, and then it stops. Assume that the games are independent of each other. Let X denote the number of games a person must play until (and including) the first win.*

(a) How many games does a person expect to play until (and including) the first win?

(b) What is the variance of the number of games until (and including) the first win?

(c) *What is the probability that the person plays 4 or more games altogether?*

Solution. *We apply the properties of the Geometric distribution.*

(a) *Since X is Geometric with probability of success 0.40, the person expects to play* $E(X) = \frac{1}{0.40} = 2.5$ *games.*

(b) *Since X is Geometric with probability of success 0.40, the variance of the number of games he plays is* $Var(X) = \frac{0.60}{(0.40)^2} = 3.75$.

(c) *The probability that a person plays 4 or more games is equal to the probability that the first three games are all losses, i.e.,* $(0.6)^3 = 0.216$.

Example 6.5.3. *Continue to as in the previous example and as above let X denote the number of games that a person must play until (and including) the first win.*

(a) *Find a formula for the gain or loss, in terms of X. That is, if Y denotes the gain or loss in dollars, what is Y in terms of X.*

(b) *What is his expected gain or loss (altogether) during the X games? That is, what is $\mathbb{E}Y$?*

(c) *What is the variance of his gain or loss (altogether) during the X games? That is, what is Var(Y)*

Solution. (a) *The gains or loss are*

$$Y = 5 + (-4)(X-1) = 9 - 4X,$$

since they win 1 game and lose $x - 1$ games.

(b)

$$\mathbb{E}(Y) = \mathbb{E}(9-4X) = 9 - 4\mathbb{E}X = 9 - 4(2.5) = -1$$

(c) *The variance of the gains or losses is* $Var(Y) = Var(9-4X) = 4^2Var(X) = 16(3.76) = 60$

Example 6.5.4. *One out of every eight calls to your house is a telemarketer. Assume that the likelihood of telemarketers is independent from call to call. Let X denote the number of callers until (and including) the next call by a telemarketer. What is $\mathbb{P}(X > n)$?*

Solution. *Here X is Geometric with probability of success $\frac{1}{8}$, becasue a success denotes a call from a telemarketer. So $X \geq n$ if the first n calls are unsuccessful. That is, if the first n calls are not telemarketers. So* $\mathbb{P}(X > n) = (7/8)^n$.

Problem 6.5.1. *Draw a card from a well-shuffled deck until the ace of spades appears. If a draw is unsuccessful, then replace and reshuffle the deck before making the next selection. Let X be the number of draws needed until the ace of spades appears for the first time. Find its mean and variance.*

Problem 6.5.2. *People are selected randomly and independently for a drug test. Each person passes the test 98% of the time. What is the expected number of people who take the test until the first person fails?*

Problem 6.5.3. *Given that more than j trials are needed to get the first success, what is the probability that more than k trials are needed?*

Solution. *We want to calculate*

$$\mathbb{P}(X > k | X > j)$$

Notice that this is zero if $k < j$*. So, assume* $k \geq j$*. Observe first that* $\mathbb{P}(X > n) = \mathbb{P}(\textit{1st } n \textit{ are failures}) = (1-p)^n$.

$$\begin{aligned}\mathbb{P}(X > k | X > j) &= \frac{\mathbb{P}(X > k \textit{ and } X > j)}{\mathbb{P}(X > j)} \\ &= \frac{(1-p)^k}{(1-p)^j} = (1-p)^{k-j} = \mathbb{P}(X > k - j)\end{aligned}$$

6.6 The Negative binomial distribution

A Negative Binomial random variable is the number of independent trials required until a certain number of successes have occurred. For instance, a Negative Binomial random variable could be the number of independent trials until the 3rd success occurs. The successes do not have to be consecutive (they usually are not). A Negative Binomial random variable can be interpreted as the sum of several independent Geometric random variables. For example, if X, Y, Z are independent Geometric random variables with the same parameter, then $X + Y + Z$ is a Negative Binomial random variable for the number of trials until the third success.

Here is the formal definition. We run an experiment: Suppose that independent trials are held with probability p of having a success. The trials are performed until a total of r successes are accumulated. Let X equal the number of trials required to obtain r successes. Here we have

$$\mathbb{P}(X = n) = \left(\begin{array}{c} n-1 \\ r-1 \end{array}\right) p^r (1-p)^{n-r} \text{ for } n = r, r+1, \ldots.$$

We say $X \sim NegativeBinomial(r, p) = NB(r, p)$.

The formula for the pmf follows because in order for the rth success to occur at the nth trial there must be $r - 1$ successes in the first $n - 1$ trials and the nth trial must result in a success. The probability of the first event is

$$\binom{n-1}{r-1} p^{r-1}(1-p)^{n-r}$$

and the probability of the second is p. Thus, Negative Binomial random variables have two parameters: p, the probability of success of the independent trials, and r, the desired number of successes.

The probability of success p must stay the same on each trial. Negative Binomial random variables always take values r or larger, because it takes at least r trials to have r successes.

The value of a Negative Binomial random variable is j if exactly $r-1$ of $X_1, \ldots, X_{n-1}$ are 1 and also $X_n = 1$ To show that this is a pmf, we need to show that

$$\sum_{n=r}^{\infty} \mathbb{P}\left(X = n\right) = 1 \tag{6.6.1}$$

we can observe that

$$\mathbb{P}(X < \infty) = \sum_{n=r}^{\infty} \mathbb{P}\left(X = n\right)$$

and we need to show that the probability on the left is in fact 1. But $X = X_1 + X_2 + \cdots X_r$ here X_1 is the number of trials until the firsdt success and X_2 is the number of trials after the first success until the second success, and so on, we see that all these r.v.'s are finite with probability 1 and so is their sum. This gives (6.6.1).

Let $q = (1-p)$. Using the identity

$$i\binom{n}{i} = n\binom{n-1}{i-1}$$

we have for each $k = 1, 2, \ldots$

$$\begin{aligned}
\mathbb{E}X^k &= \sum_{n=r}^{\infty} x^k \binom{n-1}{r-1} p^r q^{n-r} \\
&= \frac{r}{p} \sum_{m=r+1}^{\infty} (m-1)^{k-1} \binom{m-1}{r} p^{r+1} q^{(m-(r+1)}, \text{ with } m = n+1 \\
&= \frac{r}{p} \mathbb{E}(Y-1)^{k-1},
\end{aligned}$$

where $Y \sim NB(r+1, p)$.

With $k = 1$ we have $\mathbb{E}X = \frac{r}{p}$ and then

$$\mathbb{E}X^2 = \frac{r}{p}\mathbb{E}(Y-1) = \frac{r}{p}[\mathbb{E}(Y) - 1] = \frac{r}{p}\left(\frac{r+1}{p} - 1\right)$$

From this we have:

$$\operatorname{Var}(X) = \frac{r}{p}\left(\frac{r+1}{p} - 1\right) - \left(\frac{r}{p}\right)^2 = \frac{rq}{p^2}$$

We could also compute the mean and variance of a Negative Binomial as follows: If $X_1, X_2, \ldots, X_r$ are independent Geometric R.v.'s each with parameter p, then

$$X = X_1 + X_2 + \cdots + X_r$$

is NB(r, p). Therefore we have

$$\mathbb{E}(X) = \mathbb{E}(X_1) + \mathbb{E}(X_1) + \cdots + \mathbb{E}(X_r) = \frac{r}{p}$$

and

$$\mathrm{Var}(X) = \mathrm{Var}(X_1) + \mathrm{Var}(X_2) + \cdots + \mathrm{Var}(X_r) = \frac{r(1-p)}{p^2}$$

We summarize this.

6.6.1: Mean and Variance of a $NB(r, p)$

Let $q = 1 - p$. If $X \sim NB(r, p)$, then

- $$\mu = \mathbb{E} = \frac{r}{p}$$
- $$\mathbb{E}X^2 = \frac{r}{p}\mathbb{E}(Y-1) = \frac{r}{p}[\mathbb{E}(Y) - 1] = \frac{r}{p}\left(\frac{r+1}{p} - 1\right)$$
- $$\sigma^2 = \mathrm{Var}(X) = \frac{r}{p}\left(\frac{r+1}{p} - 1\right) - \left(\frac{r}{p}\right)^2 = \frac{rq}{p^2}$$

6.6.1 Examples

Example 6.6.1. *A coach needs to build a team of 5 basketball players quickly for a campus league. She decides to ask randomly selected people whether they played basketball in high school. Assume that 12% of the people have played basketball in high school. Let X be the number of people she has to ask until finding the 5th member of the team. (The values of X are 5, 6, ...) What is the the expected number of people she needs to ask? What is the standard deviation in the number of people she will need to ask in order to form the team?*

Solution.

$$\mu = \mathbb{E}X = \frac{5}{0.12} \approx 41.6$$

$$\sigma^2 = Var(X) = \frac{5(1-0.12)}{(0.12)^2} \approx 17.4$$

$$SD(X) = \sqrt{17.4}$$

Example 6.6.2. *Tom is frustrated because he is extremely good at spells but he is struggling to learn how to fly on his broomstick. He repeatedly tries to fly on the broomstick. Assume that his* *trials are independent, and a trial succeeds with probability 0.15. He does trials until his 4th success, and then stops. Let X denote the number of trials that are required until (and including) his 4th success with the broomstick.*

(a) What is the pmf of X?

(b) What is the probability that it takes Tom exactly 12 trials until his 4th success?

(c) What is the expected number of trials until the 4th success?

Solution. *(a) Since X is Negative Binomial with $r = 4$ successes and with $p = 0.15$ on each trial, then the pmf of X is*

$$p_X(n) = \binom{n-1}{3}(0.85)^{n-4}(0.15)^4$$

(b)

$$p_X(12) = \binom{11}{3}(.85)^8(.15)^4$$

(c) The expected value is

$$\mathbb{E}(X) = \frac{r}{p} = 4/0.15 = 26.67$$

Example 6.6.3. *Harry Potter needs to find 7 Horcruxes to defeat You-Know-Who. Harry makes repeated attempts to guess which objects are Horcruxes. Assume that his guesses about Horcruxes are independent. Each time he guesses about a Horcrux, he is correct only 1/3 of the time. Let X be the total number of times that he makes guesses until he finds all 7 Horcruxes.*

(a) What is the expected value of X?

(b) What is the variance of X?

(c) What is the probability that Harry finishes his quest to find all 7 Horcruxes on his 9th guess?

Solution. *With $p = 1/3$, $r = 7$, we have*

(a) $\mathbb{E}X = \frac{r}{p} = 7/(1/3) = 21$

(b) As before, using Negative Binomial parameters, $Var(X) = \frac{(1-p)r}{p^2} = \frac{(2/3)7}{(1/3)^2} = 42.$

(c)

$$P_x(x) = \binom{x-1}{6}(2/3)^{x-7}(1/3)^x,$$

$x = 7, 8, 9, \ldots$ and $p_X(x) = 0$, otherwise. So

$$p_X(9) = \binom{8}{6}(2/3)^7(1/3)^7 = 112/19683 = 0.00569$$

Example 6.6.4. *Suppose that you, Emely, and Jennie are working to master a task and you each have probability of success 30% when you attempt it. Suppose all the attempts by all three of you are independent. You attempt the task until your 5th success. Jennie attempts it until her 3rd success. Emely wants to be perfect and so she tries the task until her 20th success. Let X denote the total number of attempts by all three of you altogether (i.e., the sum of the number of the attempts).*

(a) Find the expected value of X.

(b) Find the variance of X.

Solution. *The r.v. X is a Negative Binomial random variable with $r = 28$ and $p = 0.3$*

(a)

$$\mathbb{E}X = \frac{r}{p} = 28/0.3 = 93.33...$$

(b) We have

$$Var(X) = \frac{r(1-p)}{p^2} = \frac{28(0.7)}{(0.3)^2} \approx 217.78$$

Example 6.6.5. *Statisticians make many predictions, but only 12% of them come true. The accuracy of their predictions are independent. You make a bet with your friend. You gets \$100 when the statistician prediction is correct. You lose \$15 when the statistician is wrong. You play the game until the statistician has their 5th success, and then you stop. Let X be the number of trials until (and including) the statistician 5th success.*

(a) Find a formula a formula for your earnings, Y, in terms of X.

(b) Find your expected earnings, that is, find $\mathbb{E}Y$.

(c) Find the variance of your expected earnings, that is, find $Var(Y)$

Solution. *(a) Your earns $Y = 100 \cdot 5 - 15(X - 5) = 575 - 15X$*

(b) Your expected earnings:

$$\mathbb{E}Y = 575 - 15\mathbb{E}X = 575 - \frac{15 \cdot 5}{0.12} = -50$$

(c) Variance of your earnings:

$$Var(Y) = Var(575 - 15X) = 15^2\, Var(X) = 15^2 \frac{(0.88)5}{(0.12)^2} = 68750$$

6.7 The hypergeometric distribution

Experiment: Suppose that a sample of size n is to be chosen randomly (without replacement) from an urn containing N balls, of which m are red and $N-m$ are blue. Let X equal the number of red balls selected. Then

$$\mathbb{P}\left(X=i\right)=\frac{\left(\begin{array}{c} m \\ i \end{array}\right)\left(\begin{array}{c} N-m \\ n-i \end{array}\right)}{\left(\begin{array}{c} N \\ n \end{array}\right)} \text{ for } i=0,1,\ldots,n.$$

We say $X \sim Hyp(N,m,n)$.

As before, we need to calculate $\mathbb{E}X$ and $\mathbb{E}X^2$ to find the its mean and variance. From the identities:

$$i\binom{m}{i}=m\binom{m-1}{i-1}, \quad n\binom{N}{n}=N\binom{N-1}{n-1}$$

one can show, exactly as the previous case.

Proposition 6.7.1

$$\begin{aligned}
\mathbb{E}X^k &= \sum_{i=1}^{n} i^k \mathbb{P}(X=i) \\
&= \sum_{i=1}^{n} i^k \frac{\binom{m}{i}\binom{N-m}{n-i}}{\binom{N}{n}} \\
&= \frac{nm}{N}\mathbb{E}(Y+1)^{k-1}
\end{aligned}$$

where Y is a hypergeometric r.v. with parameter $N-1$ and $m-1$, $n-1$.

From this, with $k=1$,

$$\mathbb{E}X=\frac{nm}{N}.$$

If $k=2$, applying the formula with $n-1$, $N-1$ and $m-1$ gives

$$\mathbb{E}X^2=\frac{nm}{N}\left(\mathbb{E}Y+1\right)=\frac{nm}{N}\left(\frac{(n-1)(m-1)}{N-1}+1\right)$$

$$\begin{aligned}\mathrm{Var}(X) = \mathbb{E}X^2 - (\mathbb{E}X)^2 &= \frac{nm}{N}\left(\frac{(n-1)(m-1)}{N-1}+1\right) - \left(\frac{nm}{N}\right)^2 \\ &= \frac{nm}{N}\left(\frac{(n-1)(m-1)}{N-1}+1-\frac{nm}{N}\right)\end{aligned}$$

If we set $= \frac{m}{N}$ (the proportion of white balls in the urn), then as

$$\mathrm{Var}(X) = np(1-p)\left(1-\frac{n-1}{N-1}\right).$$

We summarize this here:

6.7.1: Mean and Variance of a $X \sim Hyp(N,m,n)$.

Set $p = \frac{m}{N}$. If $X \sim Hyp(N,m,n)$, then

- $$\mu = \mathbb{E}X = \frac{nm}{N}.$$
- $$\mathbb{E}X^2 = \frac{nm}{N}(\mathbb{E}Y+1) = \frac{nm}{N}\left(\frac{(n-1)(m-1)}{N-1}+1\right)$$
- $$\sigma^2 = \mathrm{Var}(X) = np(1-p)\left(1-\frac{n-1}{N-1}\right)$$

6.7.1 Examples

Example 6.7.1. *Suppose that X is a Hypergeometric random variable with parameters $N = 50,000$, $m = 15,000$, and $n = 10$. What is $\mathbb{P}(X = 4)$. You do not need to evaluate the expression.*

Solution.

$$\mathbb{P}(X=4) = \frac{\binom{15000}{4}\binom{35000}{6}}{\binom{50000}{10}} \approx 0.2001...$$

Example 6.7.2. *Suppose that X and Y are independent Hypergeometric random variables that each have parameters $N = 6$, $m = 3$, and $n = 2$. What is the probability that X and Y are equal?. That is, are what is $\mathbb{P}(X = Y)$?*

Solution.

$$\begin{aligned}\mathbb{P}(X=Y) &= \mathbb{P}(X=Y=0)+\mathbb{P}(X=Y=1)+\mathbb{P}(X=Y=2) \\ &= (1/5)^2+(3/5)^2+(1/5)^2 = 11/25.\end{aligned}$$

Example 6.7.3. *When rolling a die, a high value is a 5 or 6. Roll 10 dice. Suppose that 8 of them are high value and 2 of them are not high value. If we examine four dice in this collection (without replacement), chosen at random, find the probability that two or fewer (of these four chosen dice) are high value. high value.*

Solution. *With the given conditions, if X is the number of high value among the 4 chosen dice, then X is a hypergeometric random variable with $N = 10$, $m = 8$, and $n = 4$, so*

$$\mathbb{P}(X \leq 2) = \mathbb{P}(X = 0) + \mathbb{P}(X = 1) + \mathbb{P}(X = 2)$$

but

$$P(X = 0) = P(X = 1) = 0 \quad \textit{(why?)}$$

$$\mathbb{P}(X \leq 2) = \frac{\binom{8}{2}\binom{2}{2}}{\binom{10}{4}} = 2/15$$

Example 6.7.4. *A seminar leader estimates that among 15 students in a seminar, 5 of them enjoyed the seminar and the other 10 did not. She interviews 3 of the people in the class, selected at random and without replacement. Let X denote the number of people (among these 3) who enjoyed the seminar.*

(a) What is the mass of X?

(b) What is $\mathbb{E}(X)$?

(c) What is $Var(X)$?

Solution. *Since X is a hypergeometric with $N = 15$, $m = 5$, and $n = 3$, then, (a)*

$$p_X(i) = \frac{\binom{5}{i}\binom{10}{3-i}}{\binom{15}{3}}$$

(b) The expected value of X is $\mathbb{E}(X) = \frac{(3)(5)}{15} = 1$
(c) The variance of X is

$$Var(X) = 3\frac{5}{15}\left(1 - \frac{5}{15}\right)\frac{(15-3)}{(15-1)} = 4/7$$

Example 6.7.5. *A chemistry teacher chooses 4 chemicals (without replacement) from the cabinet. There are 12 chemicals present: 9 of them are toxic, and the other 3 are non-toxic. Let* X_j *indicate whether the jth choice is toxic, for* $j = 1, 2, 3, 4$. *That is,* $X_j = 1$ *if the jth chemical chosen is toxic and* $X_j = 0$ *otherwise. Let* $X = X_1 + X_2 + X_3 + X_4$.

(a) Using the fact that $X^2 = (X_1 + X_2 + X_3 + X_4)^2$, *find* $\mathbb{E}(X^2)$ *by expanding the 16-term sum.*

(b) Find the p.m.f. of X.

(c) Use your answer from (b) to directly verify the answer of (a) directly.

(d) Find $\mathbb{E}(X)$.

(e) Find $Var(X)$.

Solution. *(a) We have*

$$\mathbb{E}(X^2) = 4\mathbb{E}(X_1^2) + 12\mathbb{E}(X_1X_2) + (4)(9/12) + (12)(9/12)(8/11) = 105/11.$$

(b) Since X *∼Hyp(N,m, n) with* $N = 12$, $m = 9$ *and* $n = 4$, *we have,*

$$p_X(i) = \frac{\begin{pmatrix} 9 \\ i \end{pmatrix} \begin{pmatrix} 3 \\ 4-i \end{pmatrix}}{\begin{pmatrix} 12 \\ 4 \end{pmatrix}}$$

(c)

$$\mathbb{E}X^2 = \sum_{i=0}^{4} i^2 \frac{\begin{pmatrix} 9 \\ i \end{pmatrix} \begin{pmatrix} 3 \\ 4-i \end{pmatrix}}{\begin{pmatrix} 12 \\ 4 \end{pmatrix}} = \frac{105}{11}$$

(d) Again, since X *∼Hyp(N,m, n) with* $N = 12$, $m = 9$ *and* $n = 4$, *we have,*

$$\mathbb{E}(X) = \frac{nm}{N} = 3$$

(e)

$$Var(X) = \mathbb{E}X^2 - (\mathbb{E}X)^2 = 105/11 - 32 = 6/11.$$

6.8 The Poisson distribution

We say that X taking values in $\{0,1,2,\dots\}$ is Poisson with parameter $\lambda>0$ if

$$p_X(i)=\mathbb{P}\left(X=i\right)=e^{-\lambda}\frac{\lambda^i}{i!},\quad \text{for } i=0,1,2,3,\dots.$$

We write this as $X\sim \text{Pois}(\lambda)$. Poisson random variables arise in many situations, for example. Suppose success happens λ times on average in a given period (per year, per month etc). Then $X=$ number of times success happens in that given period. Possion is like binomial, except the values of X are infinitely countable!

Here are some specific examples that obey the Poisson distribution.

1 The number of misprints on a page of a book

2 The number of people in community that survive to age 100

3 The number of telephone numbers that are dialed in a day.

4 The number of customers entering post office on a day.

Let us recall from calculus that $\sum_{n=0}^{\infty}\frac{x^n}{n!}=e^x$. Let $X\sim Pois(\lambda)$ First we check that $p_X(i)$ is indeed a pmf. First it is obvious that $p_X(i)\geq 0$ since $\lambda>0$. We to need to check that all the probabilities add up to one. That is,

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

Next observe that

$$\begin{aligned}\mathbb{E}X &= \sum_{i=0}^{\infty}ie^{-\lambda}\frac{\lambda^i}{i!}=e^{-\lambda}\lambda\sum_{i=1}^{\infty}\frac{\lambda^{i-1}}{(i-1)!}\\ &= e^{-\lambda}\lambda e^{\lambda}=\lambda.\end{aligned}$$

Also, We first have

$$\begin{aligned}
\mathbb{E}X^2 &= \sum_{i=0}^{\infty} i^2 \frac{e^{-\lambda}\lambda^i}{i!} \\
&= \lambda \sum_{i=0}^{\infty} i \frac{e^{-\lambda}\lambda^{i-1}}{(i-1)!} \\
&= \lambda \sum_{j=0}^{\infty} (j+1) \frac{e^{-\lambda}\lambda^j}{j!}, \text{let } j = i-1 \\
&= \lambda \left[\sum_{j=0}^{\infty} j \frac{e^{-\lambda}\lambda^j}{j!} + \sum_{j=0}^{\infty} \frac{e^{-\lambda}\lambda^j}{j!} \right] \\
&= \lambda \left[\lambda + e^{-\lambda} e^{\lambda} \right] \\
&= \lambda (\lambda + 1).
\end{aligned}$$

From this we can calculate the variance to be

$$\mathrm{Var} X = \lambda (\lambda + 1) - \lambda^2 = \lambda.$$

We summarize this.

6.8.1: Mean and Variance of Pois(λ)

If $X \sim$ Pois(λ) , then

- $$\mu = \mathbb{E}X = \lambda$$
- $$\mathbb{E}X^2 = \lambda (\lambda + 1)$$
- $$\sigma^2 = Var(X) = \lambda$$

The following is useful in applications:

Theorem 6.8.1: Sums of independent Poisson are Poisson

Let

$$S_n = X_1 + X_2 + \cdots + X_n,$$

be independent with $X_j \sim$ Pois(λ_j) for all $j = 1, 2, \ldots, n$. Then $S_n \sim$Pois(λ) where

$$\lambda = \lambda_1 + \lambda_2 + \cdots + \lambda_n$$

Proof. To prove the theorem all we need to do is prove the case $n = 2$. Once this is done, we just iterate. Observe that of both X_1 and X_2 only take integer values, so does their $X = X_1 + X_2$. But then if $X = j$ we need $X_1 = i$ for some $0 \leq i \leq j$ we must have $X_2 = j - i$ Thus

$$\begin{aligned}
p_X(j) &= \sum_{i=0}^{j} \mathbb{P}(X_1 = i \,\text{and}\, X_2 = j - i) \\
&= \sum_{i=0}^{j} \mathbb{P}(X_1 = i)\mathbb{P}(X_2 = j - i) \quad \text{(by independence)} \\
&= \sum_{i=0}^{j} e^{-\lambda_1} \frac{\lambda_1^i}{i!} e^{-\lambda_2} \frac{\lambda_2^{j-i}}{(j-i)!} \\
&= e^{-(\lambda_1+\lambda_2)} \sum_{i=0}^{j} \frac{\lambda_1^i}{i!} \frac{\lambda_2^{j-i}}{(j-i)!} \\
&= \frac{e^{-(\lambda_1+\lambda_2)}}{j!} \sum_{i=0}^{j} \binom{j}{i} \lambda_1^i \lambda_2^{j-i} \\
&= e^{-(\lambda_1+\lambda_2)} \frac{(\lambda_1 + \lambda_2)^j}{j}
\end{aligned}$$

where in the last line we used the Binomial theorem applied with $x = \lambda_1$, $y = \lambda_2$ and $n = j$. This completes the proof. ■

In many numerical calculations or computer simulations, it's easier to calculate concrete concrete finite sums as those arising in the Binomial distribution. For this reason (not the only one by all means), the following theorem is useful

Theorem 6.8.2: Possion is the limit of Binomial

Let $X \sim$Pois(λ) and $X_n \sim$Bin$(n,\, p_n)$, where $np_n \to \lambda$ as $n \to \infty$. Then the p.m.f of X_n converges to the p.m.f. of X as $n \to \infty$. That is, for every $j = 0, 1, 2, \cdots$,

$$\lim_{n\to\infty} \mathbb{P}(X_n = j) = \mathbb{P}(X = j)$$

In other words, for very large n,

$$\mathbb{P}(X_n = j) \approx \mathbb{P}(X = j) = e^{-\lambda} \frac{\lambda^j}{j!}.$$

and so we can approximate the Poisson distribution by the Binomial distribution. In fact, given λ we can just take $p_n = \lambda/n$ ($n > \lambda$ to make $p < 1$.

Proof. Let us suppose $p_n = \lambda/n$. We write

$$\begin{aligned}\mathbb{P}(X_n = j) &= \frac{n!}{i!(n-i)!} p_n^j (1-p_n)^{n-j} \\ &= \frac{n(n-1)\cdots(n-j+1)}{j!}\Big(\frac{\lambda}{n}\Big)^j\Big(1-\frac{\lambda}{n}\Big)^{n-j} \\ &= \frac{n(n-1)\cdots(n-j+1)}{n^j}\frac{\lambda^j}{j!}\frac{(1-\lambda/n)^n}{(1-\lambda/n)^j}.\end{aligned}$$

The first factor tends to 1 as $n \to \infty$. $(1-\lambda/n)^j \to 1$ as $n \to \infty$ and $(1-\lambda/n)^n \to e^{-\lambda}$ as $n \to \infty$. These facts are from Math 161 (first semester calculus at Purdue).

The same argument works for the general case but we need the fact that if if a_n is a sequence such that $a_n \to \lambda$, then

$$\Big(1 - \frac{a_n}{n}\Big)^n \to e^{-\lambda}.$$

■

6.8.1 Examples

Example 6.8.1. *Suppose on average there are 6 bicycle accidents on the Purdue campus per month. What is the probability there will be at most 2 bike accidents in a given month?*

Solution. *If X is the number of accidents, we are given that $EX = 6$. Since the expectation for a Poisson is $\lambda = 6$, we have that $p_X(i) = e^{-6}\frac{(6)^i}{i!}$, Therefore*

$$\mathbb{P}(X \le 2) = \mathbb{P}(X=0) + \mathbb{P}(X=1) + \mathbb{P}(X=2) = e^{-6} + 6e^{-6} + 18e^{-6} = \frac{25}{e^6}.$$

Example 6.8.2. *Suppose on average there is one large earthquake per year in California. What's the probability that next year there will be exactly 3 large earthquakes?*

Solution. $\lambda = EX = 1$, *so* $\mathbb{P}(X = 3) = \frac{1}{6e}$.

Example 6.8.3. *Customers arrive at a checkout point at an average of 8 per hour.*

(a) What is the probability that exactly 2 customers arrive in the next hour?

(b) What is the probability that no more than 3 customers arrive in the next hour?

(c) Given that at least one customer arrives in the next hour, what is the probability that more than 3 will arrive?

Solution. *The information provided gives a Pois(8) distribution. Thus*

(a)

$$\mathbb{P}(X = 2) = \frac{8^2 e^{-8}}{2} \approx 0.01073$$

(b)

$$\begin{aligned}\mathbb{P}(X \leq 3) &= \mathbb{P}(X=0)+\mathbb{P}(X=1)+\mathbb{P}(X=2)+\mathbb{P}(X=3)\\ &= e^{-8}\left(\frac{8^0}{0!}+\frac{8}{1!}+\frac{8^2}{2!}+\frac{8^3}{3!}\right)\\ &= \approx 0.04238\end{aligned}$$

(c)

$$\begin{aligned}\mathbb{P}(X>3 \mid X>1) &= \frac{\mathbb{P}(X>3\, and\, X>1)}{\mathbb{P}(X>1)}\\ &= \frac{\mathbb{P}(X>3)}{\mathbb{P}(X\geq 1)}\\ &= \frac{1-\mathbb{P}(X\leq 3)}{1-\mathbb{P}(X<1)}\\ &- \frac{1-0.42328}{1-0.0003355}\approx 0.9579\end{aligned}$$

Example 6.8.4. *(Continuation of previous example)*

(a) How many customers should the server expect in the next 3 hours?

(b) What is the probability that exactly 21 customers will arrive at the checkout point in the next 3 hours.?

(c) What is the probability that exactly 6 customers will arrive at the checkout point in the next 15 minutes.

Solution. *(a) The time interval of 1 hour has been changed to 3. Thus the new $\lambda = 24$ we have a r.v. with X ∼Pois(24). Thus $\mathbb{E}X = 24$.*

(b) For this new r.v. we have

$$\mathbb{P}(X=21)=\frac{(24)^{21}}{21!}e^{-24}$$

(c) For this, 15 minutes is 1/4 of an hour and have a X ∼Pois(2). This time $\mathbb{E}X = 2$ and

$$\mathbb{P}(6)=\frac{2^6e^{-2}}{6!}$$

6.9 The discrete uniform distribution

We say X is *discrete uniform*, and write this as $X \sim D.Uniform(n)$, if $X \in \{1, 2, \ldots, n\}$ and

$$p_X(i) = \mathbb{P}(X = i) = \frac{1}{n} \quad \text{for } i = 1, 2, \ldots, n.$$

Thus, every outcome is equally likely. The classical examples that every child knows is that of flipping a coin or rolling a die.

Example 6.9.1. *For this we have*

$$\mathbb{E}X = \sum_{i=1}^{n} i\frac{1}{n} = \frac{1}{n}\sum_{i=1}^{n} i = \frac{1}{n}\frac{n(n+1)}{2} = \frac{n+1}{2}$$

and

$$\mathbb{E}X^2 = \frac{1}{n}\sum_{i=1}^{n} i^2 = \frac{1}{n}\frac{n(n+1)(2n+1)}{6} = \frac{(n+1)(2n+1)}{6}.$$

Hence

$$\begin{aligned} VarX &= \mathbb{E}X^2 - (\mathbb{E}X)^2 = \frac{(n+1)(2n+1)}{6} - \frac{(n+1)^2}{4} \\ &= \frac{n^2+1}{12} \end{aligned}$$

6.10 Summary

(i) Find individual probabilities, $\mathbb{P}(X = x)$.

- Graph the probability mass function.
- Calculate expected value (average), variance, and standard deviation for X and functions of X.
- Calculate the cumulative distribution function by adding up each individual probability stepwise.
- Remember that it is important to check that you have a valid probability mass function first.

(ii) How do you know if your probability mass function is valid?

- The probability mass function table must include every individual value that the random variable can take.
- Each individual probability must be between 0 and 1.
- $0 \le \mathbb{P}(X = x) \le 1$ for all x.
- The sum of the probabilities over all the possible values of the random variable must sum to 1.

$$\sum_x \mathbb{P}(X = x) = 1$$

How to distinguish between the various discrete random variables? Although there is no recipe for this other than doing a lot of problems to practice, there are some markers that we can keep in mind when working on problems.

(i) Bernoulli vs. Geometric vs. Binomial

- In Bernoulli we do a single yes/no trial. We use $X = 1$ if you get a success, or $X = 0$ if you get a failure. Bernoulli random variables are the building blocks for many other random variables.
- In Geometric we continuing to do more independent Bernoulli trials until you get your 1st success. The X is the number of trials you have to do.
- In Binomial we do a pre-set number of independent Bernoulli trials and then we count up the number of successes. The r.v. X is the number of successes.

(ii) Geometric vs. Negative Binomial

- In Geometric we do trials until we get our 1st success.
- In Negative Binomial we do trials until we get our 2nd or 4th or 15th or rth success. A Negative Binomial is the sum of Geometrics. Binomial vs. Negative Binomial
- In Binomial the number of trials set in advance, and we wait until all the trials are over to count our successes. Successes can come in any order.
- In Negative Binomial the number of successes is set in advance, and we do as many trials as necessary in order to get that last success. The final success has to be on the last trial.

(iii) Binomial vs. Hypergeometric

- The X is the number of successes in n trials for both.
- Binomial is sampling with replacement. You will know the sample size and the probability of success on a single trial. The probability of success will be the same for each trial, and the trials are independent.
- Hypergeometric is sampling without replacement. You will know the population size, the sample size, and the population number of successes. The probability of success will not be the same for each trial because for each trial, the population size will be decreased by 1.

(iv) Binomial vs. Poisson

- The X is the number of successes (or arrivals or counts) for both.
- Binomial is used when you know the number of trials n and the probability of success on each trial p. The probability of success will be the same for each trial, and the trials are independent from each other.
- Poisson is used when you know the average rate of arrival for the counts λ. You have a set interval instead of a set number of trials.
- If your sample size n is really big and your probability of success on a single trial p is really small, you may want to use the Poisson approximation to the Binomial. To make this switch, find the average for the Binomial $\mathbb{E}X = np$, and set that equal to λ, The X stays the same.

(v) Bernoulli vs. Discrete Uniform

- Bernoulli is a single yes/no trial with a defined probability of success. The probability of success and the probability of failure do not have to be the same, but they do need to add up to 1.
- Discrete Uniform has one or more possible outcomes, and all of the outcomes are equally likely.
- A Bernoulli trial with a probability of success p = 0.5, like a coin toss, is similar to a Discrete Uniform distribution with N = 2. The only thing different is that the outcomes are shifted to 0, 1 for the Bernoulli, instead of the 1,2 in a Discrete Uniform. So the expected value is different (because of the shift), but the variance does not change.

(vi) Poisson vs. Discrete Uniform

- Poisson counts up the number of successes in an interval when you know the average rate of arrival.
- Discrete Uniform has only one success, and it must come from one of the N possible outcomes.

Summary Table

Name & Range	Abbrev.	Parameters	pmf (p.m.f.)	$\mathbb{E}[X]$	Var(X)
Bernoulli	Ber(p)	$p \in [0,1]$	$p_X(1) = p$, $p_X(0) = 1 - p$	p	$p(1-p)$
Binomial	Bin(n,p)	n, p	$p_X(k) = \binom{n}{k} p^k (1-p)^{n-k}$	np	$np(1-p)$
Geometric	Geo(p)	p	$p_X(k) = (1-p)^{k-1} p,\ k \geq 1$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
Negative Binomial	NBin(r,p)	r, p	$p_X(k) = \binom{k-1}{r-1} p^r (1-p)^{k-r}, k \geq r$	$\frac{r}{p}$	$\frac{r(1-p)}{p^2}$
Hyper--geometric	Hyp(N,m,n)	N, m, n	$p_X(k) = \frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}}$	$\frac{nm}{N}$	$\frac{nm(N-n)}{N(N-1)}(1-\frac{m}{N})$
Poisson	Pois(λ)	$\lambda > 0$	$p_X(k) = e^{-\lambda} \frac{\lambda^k}{k!}$	λ	λ
Discrete Uniform	U.Uniform	$N \geq 1$	$p_X(k) = 1/N$	$\frac{(N+1)}{2}$	$\frac{(N^2-1)}{12}$

6.10.1: shorter summary

(i) Ber(p): What is X? $X = 0$ or $X = 1$ (yes or no). When to use? 1 success or failure.

(ii) Bim(n, p): What is X? $X = 0, 1, \ldots, n$. When to use? Success in n trials.

(iii) Geo(p): What is X? $X = 1, 2, 3, \ldots$ (trials). When to use? # trials to 1st success.

(iv) NBin(r,p): What is X? $X = r, r + 1, \ldots$ (trials). When to use? # trials to r success.

(v) Hyp(N, m, n): What is X? $X = 1, 2, \ldots, n$. When to use? “red/blue” # red selected

(vi) Pois(λ): What is X? $X = 0, 1, \ldots, n$ (events). When to use? # events in one period.

(vii) Discrete Uniform: What is X? $X = 1, 2, \ldots$. When to use? Equality likely

Chapter 7

Week 7: Continuous random variables

7.1 Continuous random variables, their CDF's and PDF's

In the previous week we considered discrete random variable. That is, random variables whose set of possible values is either finite or countably infinite. However, there also exist random variables whose set of possible values is uncountable. It takes on values within an interval of finite or infinite length. For instance, X could be a waiting time, such as the time until the phone rings or the next email arrives, the time it takes for a student to finish a probability exam, the value of the stock price of a company at the end of the day, the time until you win a lottery game. X could take on nonnegative real number. As another example, X could be the height of a randomly chosen person, in which case X could be any real number from (say) 0 to 9 (height given in feet). Another example is that X is the horizontal distance to the left or right of the location where a dart lands, in comparison to the center of a draft board. And many, many more! Our goal for the week is to

(i) Distinguish between discrete and continuous random variables when reading a story.

(ii) Calculate probabilities for continuous random variables.

(iii) Calculate and graph a density (i.e., probability density function, PDF).

(iv) Calculate and graph a CDF (i.e., a cumulative distribution function)

(v) Calculate the mean, variance, standard deviation, and median of the continuous random variable.

Definition 7.1.1: Continuous Random Variable

A random variable X is said to have a **continuous distribution** if there exists a nonnegative function f such that

$$\mathbb{P}\left(a \leq X \leq b\right) = \int_a^b f(x)dx$$

for every a and b. And more generally, for any subset B of $\mathbb{R}$ (that is, $B \subset \mathbb{R}$),

$$\mathbb{P}(X \in B) = \int_B f(x)dx$$

The function f is called the probability density function (PDF or pdf or p.d.f.) of the random variable X.

7.1.1: Remark

Given any function f satisfying the following two properties is called a density, and could be considered a pdf of some random variable X:

(i) $f(x) \geq 0$ for all x,

(ii) $\int_{-\infty}^{\infty} f(x)dx = 1$.

Here are some simple properties of continuous r.v.'s which follow from our knowledge of integrals.

Properties 7.1.1

(i) Since the r.v. takes values in the real line,

$$1 = \mathbb{P}\left(-\infty < X < \infty\right) = \int_{-\infty}^{\infty} f(x)dx$$

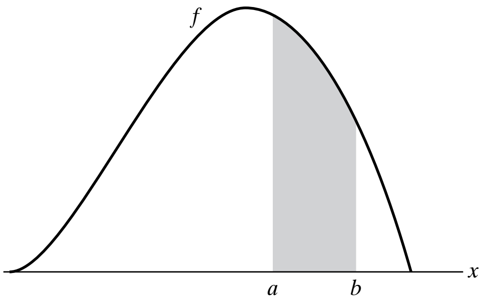
(ii) For all $a \in \mathbb{R}$,

$$\mathbb{P}\left(X = a\right) = \int_a^a f(x)dx = 0.$$

(iii) From (ii) we see that in fact

$$\begin{aligned}\mathbb{P}\left(a < X < b\right) &= \mathbb{P}(a \leq X \leq b) = \mathbb{P}(a \leq X < b) = P(a < X \leq b) \\ &= \int_a^b f(x)dx \\ &= \text{the area under the graph of f from from a to b}\end{aligned}$$

See figure below.



$P(a \leq X \leq b)$ = area of shaded region

Definition 7.1.2: Cumulative distribution function

The CDF (cumulative distribution function) of the continuous r.v. X, $F_X(x) = F(x)$ is defined by

$$F_X(a) = \mathbb{P}\left(X < a\right) = \mathbb{P}\left(X \leq a\right) = \int_{-\infty}^{a} f(x)dx$$

It is just the area under the graph of the PDF f on the interval $(-\infty, a]$; same as under f on interval $(-\infty, a)$.

Proposition 7.1.1

Replacing a by x and using t for the "dummy" variable of integration (as in Math 161) we have

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

and it follows from the Fundamental Theorem of Calculus (again Math 161) that

$$\frac{dF_X(x)}{dx} = f(x)$$

or simply

$$F_X'(x) = f(x).$$

That is, for a continuous r.v. X, the the derivative of the cumulative distribution function (CDF) is given by the probability density function (p.d.f.).

7.1.1 Examples

Example 7.1.1. *Let X have PDF f_X and set $Y = kX, k > 0$. What is the PDF for Y?*

Solution.

$$F_Y(a) = \mathbb{P}(Y \leq a) = \mathbb{P}(X \leq a/k) = F_X(a/k)$$

Differentiating gives

$$f_Y(a) = \frac{1}{k} f_X(a/k)$$

Example 7.1.2. *Suppose that X is a continuous random variable whose probability density function is given by*

$$f(x) = \begin{cases} C(4x - 2x^2) & 0 < x < 2, \\ 0 & \textit{otherwise} \end{cases}$$

(a) What is the value of C? What is $\mathbb{P}(X > 1)$?

Solution. *(a)*

$$1 = \int_{-\infty}^{\infty} f(x)dx = C\int_0^2 (4x - 2x^2)dx = C\left[2x^2 - \frac{2x^3}{3}\right]\Big|_0^2 = C\frac{8}{3}$$

and so $C = \frac{3}{8}$

(b)

$$\mathbb{P}(X > 1) = \int_1^{\infty} f(x)dx = \frac{3}{8}\int_1^2 (4x - 2x^2)dx = \frac{1}{2}$$

7.1.2: Notation

For any subset $B \subset \mathbb{R}$, its **indicator function** is written is given simply by

$$\mathbb{1}_B(x) = \begin{cases} 1 & x \in B \\ 0 & x \notin B \end{cases}$$

With this we may write, for example, simply as

$$f(x) = \begin{cases} C(4x - 2x^2) & 0 < x < 2, \\ 0 & \text{otherwise} \end{cases} = C(4x - 2x^2)\mathbb{1}_{(0,\,2)}(x)$$

can also be written simply as $f(x)$

Example 7.1.3. *Suppose we are given*

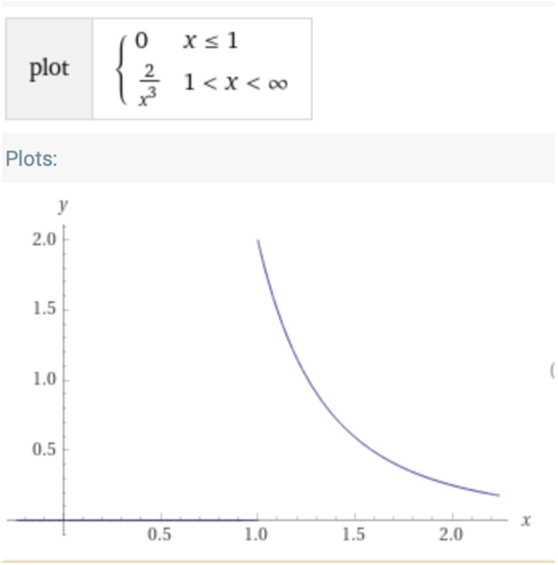
$$f(x) = \frac{c}{x^3}\mathbb{1}_{(1,\,\infty)}(x)$$

is the pdf of X. (a) What must the value of c be? (b) What is $F_X(x)$? (c) What is $\mathbb{P}(3 < X < 5)$?

Solution. *(a) We have*

$$1 = \int_{-\infty}^{\infty} f(x)dx = c\int_1^{\infty} \frac{1}{x^3}dx = \frac{C}{2},$$

thus $C = 2$.



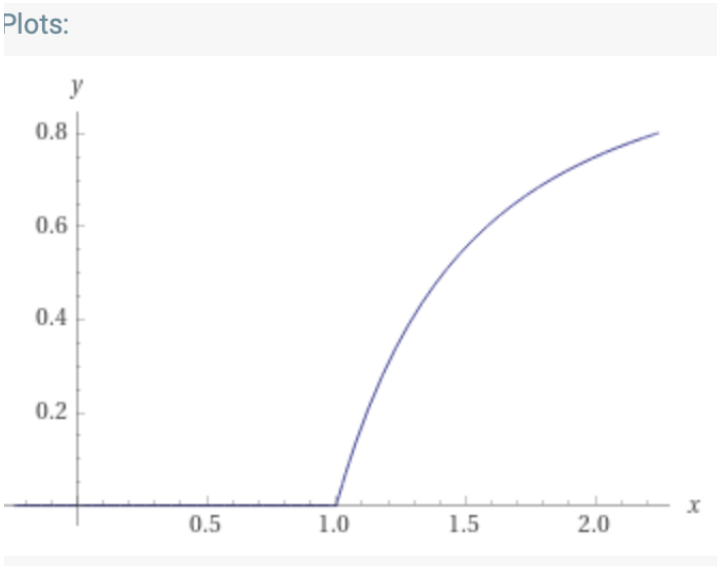
plot $\begin{cases} 0 & x \leq 1 \\ \frac{2}{x^3} & 1 < x < \infty \end{cases}$

Plots:

Graph of f_X.

(b) $F_X(x) = 0$ *if* $x \leqslant 1$. *For* $x > 1$,

$$F_X(x) = \mathbb{P}(X \leq x) = \int_1^x \frac{2}{y^3} dy = 1 - \frac{1}{x^2}$$



Plots:

Graph of F_X

(c)

$$\begin{aligned}\mathbb{P}(3 < X < 5) &= P(X < 5) - P(X \leq 3) \\ &= F(5) - F(3) \\ &= (1 - 1/25) - (1 - 1/9) \\ &= (1/9 - 1/25) = 16/225.\end{aligned}$$

Example 7.1.4. *The amount of time in hours that a computer functions before breaking down is a continuous random variable with probability density function given by*

$$f(x) = \lambda e^{-\frac{x}{100}} \mathbb{1}_{(0,\ \infty)}(x)$$

(a) What is the probability that a computer will function between 50 and 150 hours before breaking down?
(b) What is the probability that it will function for fewer than 100 hours?

Solution. *(a) We first need to find λ. We have*

$$\begin{aligned}1 = \lambda \int_{-\infty}^{\infty} f(x)dx &= \lambda \int_0^{\infty} e^{-\frac{x}{100}} dx \\ &= -\lambda(100)e^{-x/100}\Big|_0^{\infty} = 100\lambda\end{aligned}$$

and so $\lambda = 1/100$. With this,

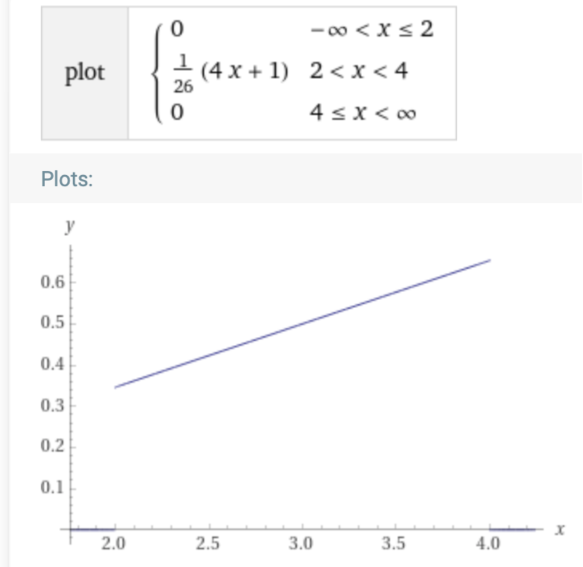
$$\begin{aligned}\mathbb{P}(50 < X < 150) &= \frac{1}{100}\int_{50}^{150} e^{-\frac{x}{100}} dx \\ &= -e^{-x/100}\Big|_{50}^{150} = e^{-1/2} - e^{-3/2} \approx 0.384\end{aligned}$$

(b) In the same way,

$$\mathbb{P}(X < 100) = \frac{1}{100}\int_0^{100} e^{-x/100}dx = 1 - e^{-1} \approx 0.633$$

Example 7.1.5. *Suppose X is a r.v. with p.d.f. given by*

$$f(x) = \frac{1}{26}(4x + 1)\mathbb{1}_{(2,\ 4)}(x).$$



plot $\begin{cases} 0 & -\infty < x \leq 2 \\ \frac{1}{26}(4x+1) & 2 < x < 4 \\ 0 & 4 \leq x < \infty \end{cases}$

Plots:

What is the CDF?

Solution. *With*

$$F_X(x) = \mathbb{P}(X \leq x)$$

we see that $F_X(x) = 0$ for $x < 2$ and $F_X(x) = 1$, for $x \geq 4$. For $2 < x < 4$ we have

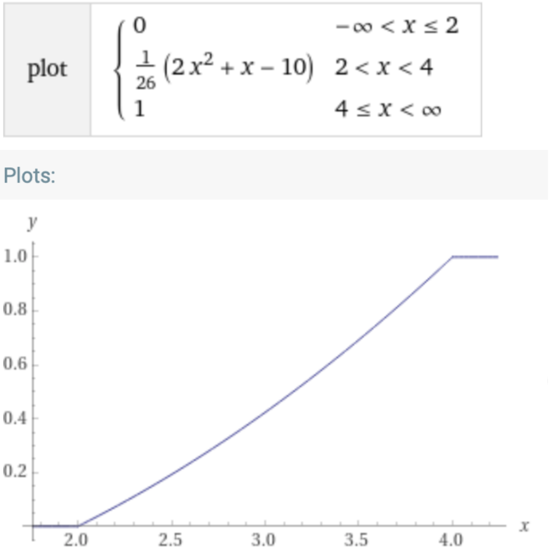
$$\begin{aligned} F_X(x) &= \frac{1}{26}\int_2^x (4t+1)dt \\ &= \frac{1}{26}(2t^2+t)\bigg|_2^x = \frac{1}{26}[(2x^2+x)-10] = \frac{1}{26}(2x^2+x-10) \end{aligned}$$

Thus

$$F_X(x) = \begin{cases} 0, & x < 2 \\ \frac{1}{26}(2x^2+x-10), & 2 \leq x \leq 4 \\ 1, & 4 < x \end{cases}$$

Using the indicator nation we can write this as

$$F_X(x) = \frac{1}{26}(2x^2+x-10)\mathbb{1}_{[2,\,4]}(x) + \mathbb{1}_{(4,\,\infty)}(x)$$



plot $\begin{cases} 0 & -\infty < x \le 2 \\ \frac{1}{26}\left(2x^2 + x - 10\right) & 2 < x < 4 \\ 1 & 4 \le x < \infty \end{cases}$

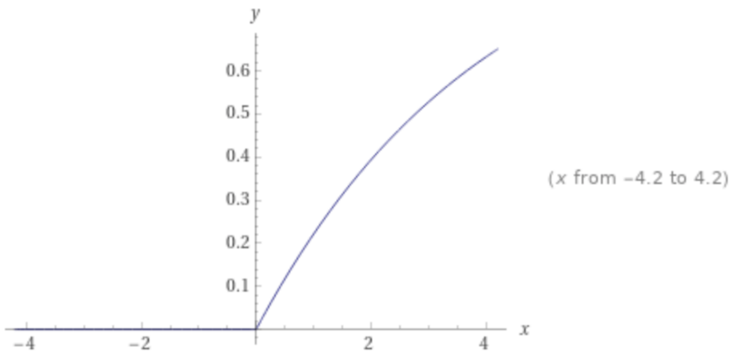
Plots:

Example 7.1.6. *Suppose X has CDF given by*

$$F_X(x) = \begin{cases} 1 - e^{-x/4}, & x > 0 \\ 0 & otherwsie \end{cases}$$

Equivalently,

$$F_X(x) = (1 - e^{-x/4})\mathbb{1}_{(0,\infty)}(x).$$



Graph F_X

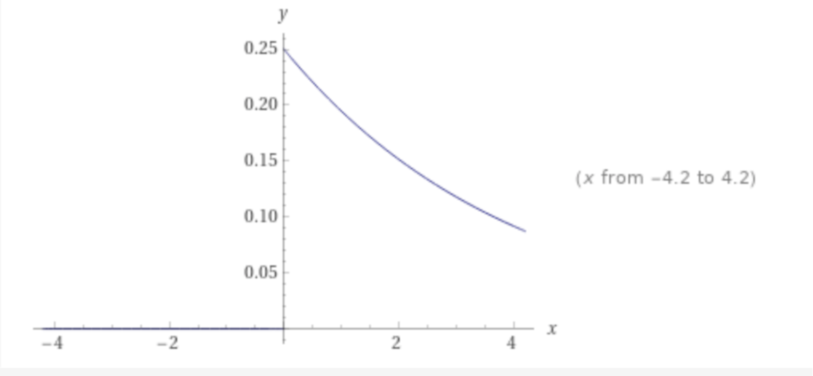
What is the p.d.f., f?

Solution. *Differentiating $F_X(x)$ gives*

$$f(x) = \begin{cases} \frac{1}{4}e^{-x/4}, & x > 0 \\ 0 & otherwise \end{cases}$$

Equivalently,

$$f(x) = \frac{1}{4} e^{-x/4} \mathbb{1}_{(0,\,\infty)}(x).$$



Graph f_X

Problem 7.1.1. *Suppose that X is a random variable whose density is*

$$f(x) = \frac{1}{2(1+|x|)^2}$$

- *Draw the graph of $f(x)$*
- *Find $\mathbb{P}(-1 < X < 2)$*
- *Find $\mathbb{P}(|X| > 1)$*
- *Is $\mathbb{E}(X)$ defined? That is, do we have $\mathbb{E}|X| < \infty$?*

Example 7.1.7. *Suppose the average lifetime of a particular kind of transistor is 100 working hours, and that the lifetime distribution is approximately exponential. Estimate the probability that the transistor will work for at least 50 hours.*

Solution. *Since the mean of the exponential distribution is $\frac{1}{\lambda}$, we have $\lambda = 1/100 = 0.01$.*

$$\mathbb{P}(X > 50) = e^{-\lambda 50} = e^{-0.5} \approx 0.606$$

7.2 Expectation of continuous random variables

In this section we give the definition of the expectation of a continuous r.v. and derive several of its properties. Let us recall first the definition of the expectation of a discrete r.v. X with pmf $p_X(x)$. It is given by

$$\mathbb{E}X = \sum_{x_i} x_i p_X(x_i) = \sum_{x} x\mathbb{P}(X = x_i)$$

where the sum ranges over all (discrete) values of x for which $p_X(x_i) > 0$.

Now, let us consider a continuous r.v. with pdf f. Recall that

$$\frac{dF_X}{dx}(x) = f(x)$$

or

$$f(x)dx = dF_X(x)$$

That is,

$$f(x)dx \approx F_X(dx + x) - F_X(x),$$

with the differential notation of Math 161. That is, if we take (as in calculus), $h = dx$ small, than we know that

$$f(x) \approx \frac{F_X(x+h) - F(x)}{h}$$

from the definition of the derivative of F_X as the limit of the right had side as $h \to 0$. But since

$$F_X(x + dx) - F_X(x) = \mathbb{P}(x \leq X \leq x + dx)$$

or

$$f(x)dx \approx \mathbb{P}(x \leq X \leq x + dx)$$

With this we see that

$$\mathbb{E}X \approx \sum_{x_i} x_i \mathbb{P}(x_i \leq X \leq x_i + dx) \approx \int_{-\infty}^{\infty} x f(x) dx$$

Thus we make the following definition.

Definition 7.2.1: Expectation of continuous random variables

Let $\mathbb{P}$ be a probability on a sample space Ω. Let $X : \Omega \to \mathbb{R}$ be a continuous random variable with pdf f. The expected value of X is defined by

$$\mathbb{E}X = \int_{-\infty}^{\infty} x f(x) dx$$

The above motivation says that the expectation of a continuous random variable can be obtained as the limit of the expectation of a sequence of discrete random variables. The following makes this precise.

Theorem 7.2.1

Suppose X is a nonnegative continuous random variable with finite expectation and PDF f. There exists a sequence of discrete random variables X_n such that

$$\lim_{n\to\infty} \mathbb{E}X_n = \mathbb{E}X = \int_0^{\infty} x f(x) dx$$

(The integral here is only on $(0, \infty)$ since the r.v. does not take negative values. So, $P(X < 0) = \int_{-\infty}^0 f(x)dx = 0$ and since the PDF function f is not nennegative, this means that $f(x) = 0$ if $x \leq 0$.
)

Proof. Fix n. Since $X(\omega) \geq 0$, we can define

$$X_n(\omega) = \frac{k}{2^n}$$

if

$$k/2^n \leq X(\omega) < (k+1)/2^n)$$

for some $k = 0, 1, \ldots$

By the definition of expectation for discrete r.v.'s and the fact that $\mathbb{P}(a \leq X < b) = \int_a^b f(x)dx$, we have

$$\begin{aligned} \mathbb{E}X_n &= \sum_{k=0}^{\infty} \frac{k}{2^n} \mathbb{P}(k/2^n \leq X(\omega) < (k+1)/2^n)(\omega)) \\ &= \sum_{k=0}^{\infty} \frac{k}{2^n} \int_{\frac{k}{2^n}}^{\frac{k+1}{2^n}} f(x)dx = \sum_{k=0}^{\infty} \int_{\frac{k}{2^n}}^{\frac{k+1}{2^n}} \frac{k}{2^n} f(x)dx \end{aligned}$$

Next, observe that for any nonnegative function $g(x)$, the additivity properties of the integral (Math 161 again: $b \in (a,c)$ implies $\int_a^b = \int_a^c + \int_c^b$) gives that

$$\sum_{k=0}^{\infty} \int_{\frac{k}{2^n}}^{\frac{k+1}{2^n}} g(x)dx = \int_0^{\infty} g(x)dx$$

and that for any $x \in [k/2^n, (k+1)2^n]$, $(x - k/2^n) \leq 1/2^n$. Applying the first observation with $g(x) = xf(x)$ and using the second we have

$$\begin{aligned} 0 \leq \mathbb{E}X - \mathbb{E}X_n &= \sum_{k=0}^{\infty} \int_{\frac{k}{2^n}}^{\frac{k+1}{2^n}} xf(x) - \sum_{k=0}^{\infty} \int_{\frac{k}{2^n}}^{\frac{k+1}{2^n}} \frac{k}{2^n} f(x)dx \\ &= \sum_{k=0}^{\infty} \int_{\frac{k}{2^n}}^{\frac{k+1}{2^n}} \left(x - \frac{k}{2^n}\right) f(x)dx \\ &\leq \frac{1}{2^n} \sum_{k=0}^{\infty} \int_{\frac{k}{2^n}}^{\frac{k+1}{2^n}} f(x)dx = \frac{1}{2^n} \int_0^{\infty} f(x)dx \\ &= \frac{1}{2^n} \to 0, \quad \text{as } n \to \infty \end{aligned}$$

This shows that the sequence $\mathbb{E}X_n$ converges to $\mathbb{E}X$ as $n \to \infty$. ■

In a similar way one can show that if X and Y are two nonnegative r.v.'s there is always an increasing sequence of discrete r.v.'s X_n and Y_n such that $X_n \to X$, $Y_n \to Y$ and $X_n + Y_n$ increasing to $X + Y$.

Corollary 7.2.1: Linearity of expectation for continuous r.v.'s

Let X and Y be two nonnegative continuous random variables. Then,

$$\mathbb{E}(X+Y) = \mathbb{E}X + \mathbb{E}Y.$$

By writing $X = X^+ - X^-$ and $Y = Y^+ - Y^-$ where $X^+ = X\mathbb{1}_{(X\geq 0)}$ and $X^- = -X\mathbb{1}_{(X\leq 0)}$ and similarly for Y^+ and Y^-, and applying the above to each part, the linearity of the of expectation follows for all continuous r.v.'s for which $\mathbb{E}|X| < \infty$ follows. We list it here for convience.

Corollary 7.2.2

Suppose X and Y are two contiouous random variables with PDF f_X and f_Y, respectively, for which

$$\int_{-\infty}^{\infty} |x| f_X(x)dx < \infty$$

and

$$\int_{-\infty}^{\infty} |x| f_Y(x)dx < \infty.$$

and similarly for Y.) Then,

$$\mathbb{E}(X+Y) = \mathbb{E}X + \mathbb{E}Y.$$

Let us recall that for discrete r.v., we showed that $\mathbb{E}g(X) = \sum_x g(x)\mathbb{P}(X = x)$ and this shows that the expectation of a function of a r.v. only depends on the distribution of the r.v. (and the function of course). The usefulness of this property comes from the fact that it implies that whenever two r.v.'s X and Y are equally distributes they have $\mathbb{E}(g(X)) = \mathbb{E}(g(Y))$. The same holds for continuous r.v.'s.

Theorem 7.2.2

Suppose X is a continuous random variable with PDF f. Then for $g : \mathbb{R} \to \mathbb{R}$,

$$\mathbb{E}(g(X)) = \int_{-\infty}^{\infty} g(x)f(x)dx$$

Note that in particular,

$$E|X| = \int_{-\infty}^{\infty} |x|f(x)dx$$

Corollary 7.2.3

If X and Y are equally distributed, that is $\mathbb{P}(X < a) = \mathbb{P}(Y < a)$, for all $-\infty < a < \infty$, then

$$\mathbb{E}(g(X)) = \mathbb{E}(g(Y)$$

The Theorem follows from the following

Proposition 7.2.1

Suppose X is a nonnegative continuous r.v. Then

$$\mathbb{E}X = \int_0^{\infty} \mathbb{P}(X > a)\, da$$

Proof. . Let f be the PDF of X. Recall that

$$\mathbb{P}(X > a) = 1 - \mathbb{P}(X \le a) = 1 - F_X(a) = \int_0^\infty f(x)dx - \int_0^a f(x)dx = \int_a^\infty f(x)\,dx$$

From this we have

$$\begin{aligned}
\int_0^\infty \mathbb{P}(X > a)\,da &= \int_0^\infty \left(\int_a^\infty f(x)\,dx\right) da \\
&= \int_0^\infty \left(\int_0^\infty f(x)\mathbb{1}_{(a,\,\infty)}(x)\,dx\right) da \\
&= \int_0^\infty f(x)\left(\int_0^\infty \mathbb{1}_{(a,\,\infty)}(x)\,da\right) dx
\end{aligned}$$

Now, notice that for every fix x and variable a, $\mathbb{1}_{(a,\,\infty)}(x) = \mathbb{1}_{(0,\,x)}(a)$ and the right-hand side of the above equation is equal to

$$\begin{aligned}
\int_0^\infty f(x)\left(\int_0^\infty \mathbb{1}_{(0,\,x)}(a)\,da\right) dx &= \int_0^\infty f(x)\left(\int_0^x 1\,da\right) dx \\
&= \int_0^\infty xf(x)\,dx = \mathbb{E}X
\end{aligned}$$

Apply the Proposition now to the r.v. $g(X)$. (We assume $g(x) \ge 0$ for all x so the r.v. is nonnegative.) Then as above,

$$\begin{aligned}
\mathbb{E}(g(x)) &= \int_0^\infty \mathbb{P}(g(X) > a)da \\
&= \int_0^\infty \left(\int_{\{g(x)>a\}} f(x)\right) dxda \\
&= \int_0^\infty f(x)\left(\int_0^{g(x)} da\right) dx \\
&= \int_0^\infty g(x)f(x)dx
\end{aligned}$$

■

7.3 The Exponential Distribution and its lack of memory property

A continuous random viable X with PDF given by

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & if\ x \ge 0 \\ 0 & if\ x < 0 \end{cases}$$

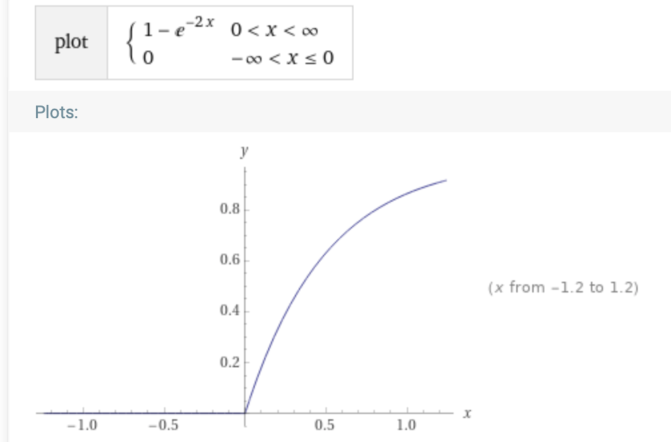
where $\lambda > 0$ is called an exponential random variable or is said to have an exponential distribution. For any $a \ge 0$, its distribution function is

$$F(a) = \mathbb{P}(X \le a) = \int_0^a \lambda e^{-\lambda x}dx = -e^{-\lambda x}\Big|_0^a = 1 - e^{-\lambda a}$$

and otherwise $F(a) = 0$. That is,

$$F(a) = \begin{cases} 1 - e^{-\lambda a} & a \geq 0 \\ 0 & a < 0 \end{cases}$$

The notation X ∼Exp(λ) is used and we say that X *is exponentially distributed with parameter* λ.

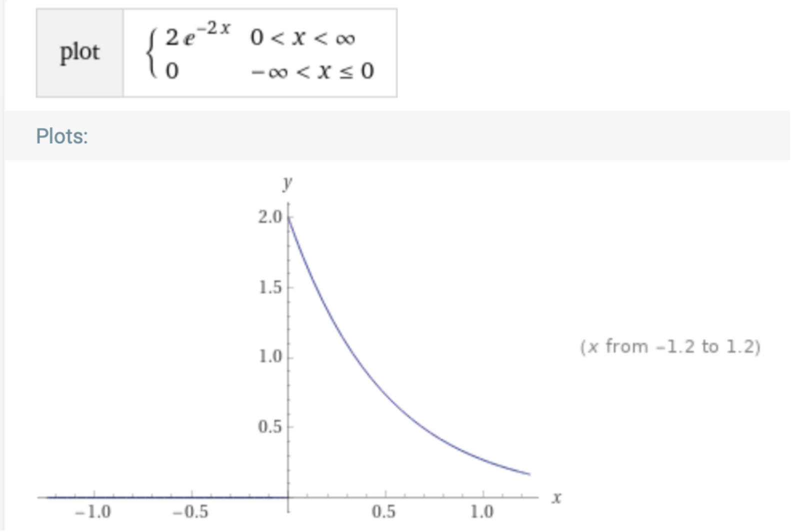


plot $\begin{cases} 1 - e^{-2x} & 0 < x < \infty \\ 0 & -\infty < x \leq 0 \end{cases}$

Plots:

(x from −1.2 to 1.2)

CDF graph of Exp(2)



plot $\begin{cases} 2\,e^{-2x} & 0 < x < \infty \\ 0 & -\infty < x \leq 0 \end{cases}$

Plots:

(x from −1.2 to 1.2)

PDF graph of Exp(2)

The exponential distribution arises as the distribution of the amount of time until some specific event occurs. For instance, the amount of time (starting from now) until the next storm arrives, or until the next earthquake occurs, or or your next phone call, email or text message arrive.

One use integration by parts to find both $\mathbb{E}(X)$ and $\mathbb{E}(X^2)$ and hence the variance of X. But more generally, we can also proceed as follows. Let $r > 0$ (any real number). Then integration by parts with

$u = x^r$ and $dv = \lambda e^{-\lambda x}$ gives

$$\begin{aligned}
\mathbb{E}(X^r) &= \int_0^\infty \lambda x^r e^{-\lambda x} dx \\
&= -x^r e^{-\lambda x}\Big|_0^\infty + r\int_0^\infty x^{r-1} e^{-\lambda x} dx \qquad (7.3.1)\\
&= \frac{r}{\lambda}\int_0^\infty x^{r-1}\lambda e^{-\lambda x} dx = \frac{r}{\lambda}\mathbb{E}(X^{r-1}).
\end{aligned}$$

Applying this with $r = 1$ gives

$$\mathbb{E}X = \frac{1}{\lambda} \qquad (7.3.2)$$

and with $r = 2$ gives

$$\mathbb{E}X^2 = \frac{2}{\lambda}\mathbb{E}X = \frac{2}{\lambda^2} \qquad (7.3.3)$$

Note that if $r = n$ a positive integer then

$$\mathbb{E}X^n = \frac{n!}{\lambda^n} \qquad (7.3.4)$$

From these calculations,

$$\mathrm{Var}(X) = \mathbb{E}X^2 - (\mathbb{E}X)^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2} \qquad (7.3.5)$$

Example 7.3.1. *Suppose that the length of time to service a customer at the postoffice in minutes is an exponential random variable with parameter $\lambda = \frac{1}{10}$. If someone arrives immediately ahead of you, find the probability that you will have to wait*

(i) *more than10 minutes,*

(ii) *between 10 and 20 minutes.*

Solution. *Let X denote the length of time the customer is at the service desk. Then,*

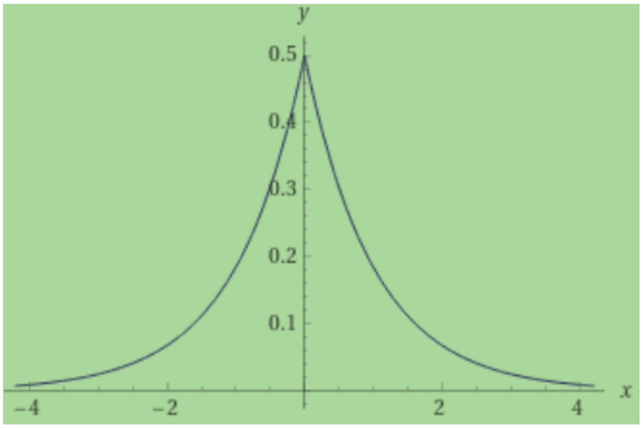
(i)

$$\mathbb{P}(X > 10) = 1 - F(10) = e^{-1} \approx 0.369.$$

$$\mathbb{P}(10 < X < 20) = F(20) - F(10) = e^{-1} - e^{-2}$$

A double exponential random variable is a r.v. X with PDF given by

$$f(x) = \frac{1}{2}\lambda e^{\lambda|x|}, \ -\infty < x < \infty.$$



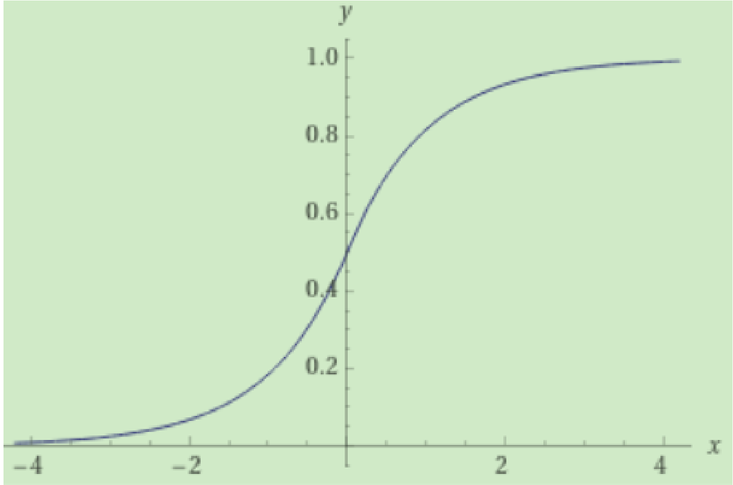
Graph PDF with $\lambda = 1$

The 1/2 is needed here since

$$\int_{-\infty}^{\infty} \frac{1}{2}\lambda e^{-\lambda|x|}dx = 2\int_{0}^{\infty} \frac{1}{2}\lambda e^{-\lambda|x|}dx = 1$$

The CDF is given by

$$\begin{aligned} F(a) &= \begin{cases} \frac{1}{2}\int_{-\infty}^{a} \lambda e^{\lambda x}dx, & a < 0 \\ \frac{1}{2}\int_{\infty}^{0} \lambda e^{\lambda a}dx + \frac{1}{2}\lambda e^{-\lambda x}dx, & \geq 0 \end{cases} \\ &= \begin{cases} \frac{1}{2}e^{\lambda a}, & a < 0 \\ 1 - \frac{1}{2}e^{-\lambda a}, & a \geq 0 \end{cases} \end{aligned}$$



Graph CDF with $\lambda = 1$

Definition 7.3.1: Lack of memory property

We say that a nonnegative random variable X has the lack of memory property if

$$\mathbb{P}(X > s + t | X > t) = \mathbb{P}(X > s), \quad \text{for all } t, s \geq 0.$$

If we think of X as being the lifetime of some instrument, the equation states that the probability that the instrument survives for at least $s+t$ hours, given that it has survived t hours, is the same as the initial probability that it survives for at least s hours. In other words, if the instrument is alive at age t, the distribution of the remaining amount of time that it survives is the same as the original lifetime distribution. (That is, it is as if the instrument does not "remember" that it has already been in use for time t–**lack of memory**.)

Properties 7.3.1

The exponential distribution has the lack of memory property. That is,

$$\mathbb{P}(X > s + t | X > t) = \mathbb{P}(X > s), \quad \text{for all } t, s \geq 0.$$

To see this, observe that (since $(X > s+t) \subset (X > t)$)

$$\begin{aligned} \mathbb{P}(X > s + t | X > t) &= \frac{\mathbb{P}(X > s + t, X > t)}{\mathbb{P}(X > t)} \\ &= \frac{\mathbb{P}(X > s + t)}{\mathbb{P}(X > t)} \\ &= \frac{e^{-\lambda(s+t)}}{e^{-\lambda t}} = e^{-\lambda s} = \mathbb{P}(X > s). \end{aligned}$$

7.4 The uniform distribution

For any $-\infty < a < b < \infty$, a continuous random variable X is said to be uniformly distributed on the interval $[a, b]$ if its PDF

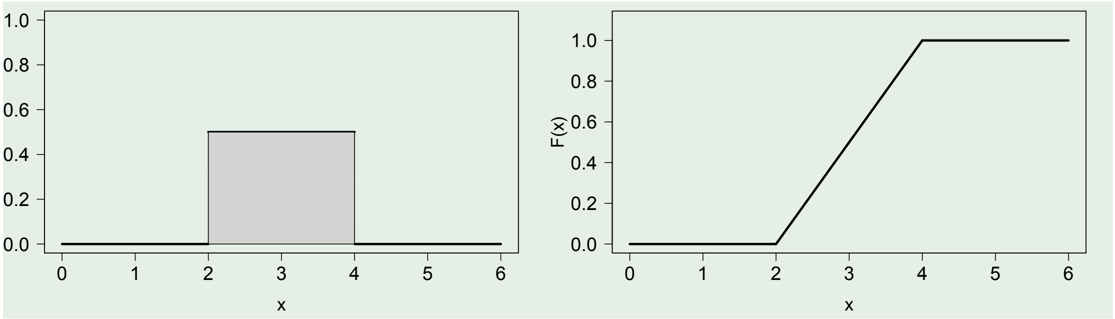
$$f(x) = \begin{cases} \frac{1}{b-a}, & a \leq x \leq b \\ 0, & \text{otherwise} \end{cases} = \frac{1}{b-a} \mathbb{1}_{[a,b]}(x).$$

This normalization is needed so that the area under the curve is 1.

Thus, X can only attain values in $X \in [a, b]$. We write this as $X \sim Uniform(a, b)$. The CDF is

$$F_X(x) = \int_{-\infty}^{x} f(t)dt = \begin{cases} 0 & x < a \\ \frac{x-a}{b-a} & a \leq x \leq b \\ 1 & x > b \end{cases}.$$

Example of uniform on $[2, 4]$, PDF and CDF:



Uniform(2, 4), PDF (left) CDF (right)

Observe that

$$\begin{aligned}\mathbb{E}X &= \int_{-\infty}^{\infty} x f_X(x)dx = \int_a^b x\frac{1}{b-a}dx \\ &= \frac{1}{b-a}\left(\frac{b^2}{2}-\frac{a^2}{2}\right) = \frac{a+b}{2}.\end{aligned}$$

Which makes sense, right? It should be the midpoint of the interval $[a, b]$.
We compute first the second moment

$$\begin{aligned}\mathbb{E}X^2 &= \int_a^b x^2\frac{1}{b-a}dx = \frac{1}{b-a}\left(\frac{b^3}{3}-\frac{a^3}{3}\right) \\ &= \frac{1}{3}\frac{1}{b-a}(b-a)\left(a^2+ab+b^2\right) \\ &= \frac{a^2+ab+b^2}{3}.\end{aligned}$$

Thus after some algebra

$$\text{Var}X = \frac{a^2+ab+b^2}{3} - \left(\frac{a+b}{2}\right)^2 = \frac{(b-a)^2}{12}.$$

7.4.1 Examples

Example 7.4.1. *X and Y be equally distributed. That is, they both have the same PDF, $f_X = f_Y$. Then*

$$\mathbb{E}(g(X)) = \mathbb{E}(g(Y)).$$

Example 7.4.2. *Let X be a r.v. with mean $\mu_X = \mathbb{E}X$. Then its variance is given by*

$$Var(X) = \int_{-\infty}^{\infty}(x-\mu_X)^2 f(x)dx$$

Example 7.4.3. *A stick of length 1 is split at a point U that is uniformly distributed over (0, 1). Determine the expected length of the piece that contains the point $0 \le x \le 1$.*

Solution. *Let $L_x(U)$ be the length of the substick that contains the point x. Then*

$$L_x(U) = (1-U)\mathbb{1}_{(0,x)}(U) + U\mathbb{1}_{(x,1)}(U) = \begin{cases} (1-U), & 0 < U < x \\ U, & x < U < 1 \end{cases}$$

Then

$$\mathbb{E}(L_x(U)) = \int_0^1 L_x(u)du = \int_0^x (1-u)du + \int_x^1 udu = \frac{1}{2} + x(1-x)$$

Example 7.4.4. *Suppose that if you are s minutes early for an appointment, then you incur the cost $a \cdot s$ and if you are s minutes late, then you incur the cost $b \cdot s$. Suppose also that the travel time from where you presently are to the location of your appointment is a continuous random variable X having a pdf f and CDF F. Determine the time at which you should depart if you want to minimize your expected cost.*

Solution. *Let X be your time traveled. Let $C_t(X)$ be your cost if you leave t minutes before your appointment. Then*

$$C_t(X) \quad = \quad a(t-X)\mathbb{1}_{[0,\,t]}(X) + b(X-t)\mathbb{1}_{[t,\,\infty)}(X)$$

Set

$$\begin{aligned} h(t) = \mathbb{E}[C_t(X)] &= \int_0^\infty C_t(x)f(x)dx = a\int_0^t (t-x)f(x)dx + b\int_t^\infty (x-t)f(x)dx \\ &= at\int_0^t f(x)dx - a\int_0^t xf(x)dx + b\int_t^\infty xf(x)dx - bt\int_t^\infty f(x)dx \end{aligned}$$

Differentiating with respect to t (using the Fundamental Theorem of Calculus, product rule and formula for the CDF) Gives

$$\begin{aligned} h'(t) &= aF(t) + atf(t) - atf(t) - btf(t) - b[1-F(t)] + btf(t) \\ &= (a+b)F(t) - b \end{aligned}$$

Thus setting this equal to zero show that the minimizing time, call it t_m, is such that

$$F(t_m) = \frac{b}{a+b},$$

or equivalently, for which,

$$\mathbb{P}(X \leq t_m) = \frac{b}{a+b}.$$

Problem 7.4.1. *A point is chosen at random on a line segment of length L. Interpret this statement, and find the probability that the ratio of the shorter to the longer segment is less than 1/4*

Problem 7.4.2. *You arrive at a bus stop at 10 o'clock, knowing that the bus will arrive at some time uniformly distributed between 10 and 10:30. (a) What is the probability that you will have to wait longer than 10 minutes?*

(b) If, at 10:15, the bus has not yet arrived, what is the probability that you will have to wait at least an additional 10 minutes?

Problem 7.4.3. *A product that is sold seasonally yields a net profit of b dollars for each unit sold and a net loss of ℓ dollars for each unit left unsold when the season ends. Suppose that the seasonal demand is a continuous random variable having probability density function f. The number of units of the product that are ordered at a specific department store during any season is a counties random variable with PDF f. Show that the optimal amount to stock is the value s^* that satisfies*

$$F(s^*) = \frac{b}{b+\ell}$$

Hint: Work as in the above example with the profit

$$P(s) = \begin{cases} bX - (s-X)\ell & X \leq s \\ sb & X > s \end{cases}$$

7.5 Summary: discrete vs continuous random variables

Properties 7.5.1: Discrete r.v.'s notation

(i) X can take on only distinct "discrete" values in a set. e.g. $X \in \{0, 1, \cdots\}$

(ii) Probability mass function (p.m.f) $p_x(x) = \mathbb{P}(X = x)$

(iii) Cumulative distribution function (CDF)

$$F_X(x) = F(x) = \mathbb{P}(X \leq x) = \sum_{x_i \leq x} p_X(x_i)$$

(iv) Mean

$$\mu_X = \mu = \mathbb{E}X = \sum_x x p_X(x)$$

(v) Variance:

$$\begin{aligned} \sigma_X^2 &= \sigma^2 = \mathrm{Var}(X) = \mathbb{E}(X-\mu)^2 \\ &= \sum_x (x-\mu)^2 p_X(x) \\ &= \mathbb{E}(X^2) - \mu^2 \\ &= \sum_x x^2 p_X(x) - \mu^2 \end{aligned}$$

Properties 7.5.2: Continuous r.v.'s notation

(i) X can take on all possible values in an interval of real numbers, e.g., $X \in [a, b]$ for any $-\infty < a < b < \infty$.

(ii) Probability density function (PDF) $f_X(x) = f(x)$

(iii) Cumulative distribution function (CDF)

$$F_X(x) = F(x) = \int_{-\infty}^{x} f(t)dt$$

(iv) Mean

$$\mu_X = \mu = \mathbb{E}X = \int_{-\infty}^{\infty} xf(x)dx$$

(v) Variance:

$$\begin{aligned} \sigma_X^2 &= \sigma^2 = \mathrm{Var}(X) = \mathbb{E}(X-\mu)^2 \\ &= \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx \\ &= \mathbb{E}(X^2) - \mu^2 \\ &= \int_{-\infty}^{\infty} x^2 f(x)dx - \mu^2 \end{aligned}$$

Chapter 8

Weeks 8 and 9: The Classical Continuous Distributions

As in the case of discrete random variables, there are many important continuous distributions that naturally arise in many applications in different fields. In this chapter we will introduce several of these distributions and derive their basic properties. For all, it is important to know how to compute their means and variances. We begin with the normal distribution which is central in in many branches of mathematics, statistics and applications.

8.1 The Normal Distribution

The Normal Distribution is the queen of all distribution. We say that X is a normal (Gaussian) random variable, or X is normally distributed with parameters μ and σ^2, if the density of X is given by

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

This is one of the most widely used distribution in many applications and modeling. It arises naturally in statistics, mathematics, physical and biological phenomena, in social sciences, in economics and engineering problems. Many things that are clustered near their expected value are normally distributed. The graph of the density of the normal distribution is the well known "shaped bell curve." In fact, the "bell curve" looks like a bell because most of the density is concentrated around its mean. Many things that are clustered near their expected value are normally distributed. Only a relatively small amount of the density is concentrated far from the center. It is a "universal" distribution in the sense that "all" others, properly normalized, converge to it in the appropriate mathematical sense. The latter is the content of the famous *Central Limit Theorem.* This universality gives the Normal Distribution its special place in science and beyond. Our goal in this lecture is to give its very basic properties and a few basic examples. Further properties and more examples will be given in subsequent weeks.

(i) Two parameters μ and σ^2 are use to define the normal probability density. μ is its mean (or center of the distribution)

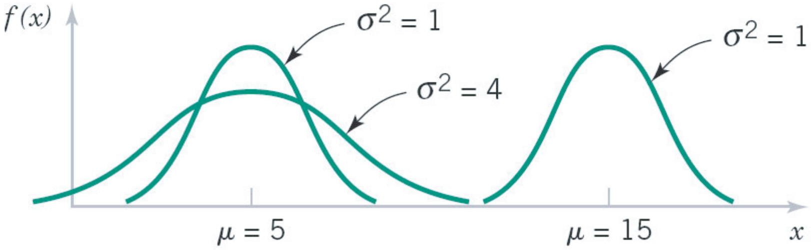
(ii) σ^2 is its variance.

(iii) The distribution is symmetric about μ.

(iv) A normal distribution can occur (is none zero) on any interval of the real line.

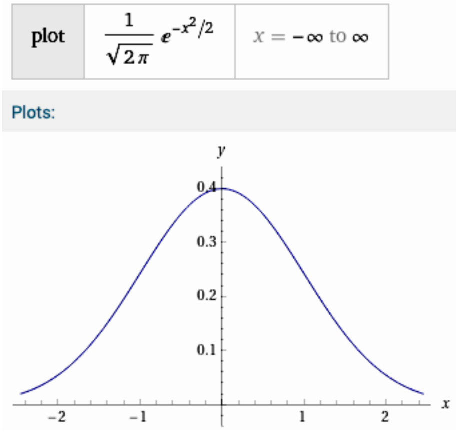
(v) It always has a bell-shape.

(vi) The parameter μ tells us where it is centered and where there is a high probability of X occurring.



$X \sim \mathcal{N}(5,1)$, $X \sim \mathcal{N}(5,4)$, $X \sim \mathcal{N}(15,1)$

Many properties of the general normal $\mathcal{N}\left(\mu, \sigma^2\right)$ follow from the case $\mathcal{N}(0,1)$ which we will refer to as the **"standard normal distribution."**



plot $\frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ $x = -\infty$ to ∞

Graph of Standard normal $Z \sim \mathcal{N}(0, 1)$

Properties 8.1.1

It is important to check that $f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ is indeed a PDF. Of course, it is clearly nonnegative. But we must also have

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx = 1.$$

To show this we use polar coordinates. Let $I = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-x^2/2} dx$. The trick is to write

$$\begin{aligned} I^2 &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-x^2/2} e^{-y^2/2} dx dy \\ &= \frac{1}{2\pi} \int_0^{2\pi} \left(\int_0^{\infty} r e^{-r^2/2} dr \right) d\theta = 1. \end{aligned}$$

Properties 8.1.2

Let us in fact check that if $Z \sim \mathcal{N}(0,1)$, then in fact

$$\mathbb{E}Z = 0 \text{ and } \operatorname{Var}(Z) = 1.$$

To show this, u substitution (with $u = x^2/2$) immediately gives

$$\mathbb{E}Z = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x e^{-x^2/2} dx = -\frac{1}{\sqrt{2\pi}} e^{-x^2/2} \Big|_{-\infty}^{\infty} = 0.$$

integration by parts gives

$$\begin{aligned} \operatorname{Var}(Z) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-\infty} x^2 e^{-x^2/2} dx \\ &= \frac{1}{\sqrt{2\pi}} \left(-x e^{-x^2/2} \Big|_{-\infty}^{\infty} + \int_{-\infty}^{\infty} e^{-x^2} dx \right) = 1 \end{aligned}$$

Properties 8.1.3

To help us compute the mean and variance of X is not too hard to show $X \sim \mathcal{N}\left(\mu, \sigma^2\right)$ if and only if

$$\frac{X - \mu}{\sigma} = Z \text{ where } Z \sim \mathcal{N}(0,1).$$

Proof. Suppose Z is a standard normal. Then

$$\begin{aligned} F_X(x) &= \mathbb{P}\left(X \le x\right) = \mathbb{P}\left(\sigma Z + \mu \le x\right) \\ &= \mathbb{P}\left(Z \le \frac{x-\mu}{\sigma}\right) \\ &= F_Z\left(\frac{x-\mu}{\sigma}\right) \end{aligned}$$

for $\sigma > 0$. Similar for $\sigma < 0$. By the chain rule

$$\begin{aligned} f_X(x) &= F'_X(x) \\ &= F'_Z\left(\frac{x-\mu}{\sigma}\right)\frac{1}{\sigma} \\ &= \frac{f_Z\left(\frac{x-\mu}{\sigma}\right)}{\sigma} \\ &= \frac{1}{\sigma}\frac{1}{\sqrt{2\pi}}e^{-(x-\mu)^2/(2\sigma^2)}. \end{aligned}$$

On the other direction, if $X \sim \mathcal{N}\left(\mu, \sigma^2\right)$ and we set $Z = \frac{X-\mu}{\sigma}$ we have

$$\begin{aligned} F_Z(x) &= \mathbb{P}(Z \le x) \\ &= \mathbb{P}(X \le \sigma x + \mu) \\ &= F_X(\sigma x + \mu), \ \ \text{(remember } \ \sigma > 0) \end{aligned}$$

Differentiating both sides gives

$$\begin{aligned} f_Z(x) &= F'_Z(x) = \sigma F'_X(\sigma x + \mu) \\ &= \sigma f_X(\sigma x + \mu) = \sigma\left(\frac{1}{\sigma}\frac{1}{\sqrt{2\pi}}e^{-(\sigma x + \mu - \mu)^2/(2\sigma^2)}\right) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2} \end{aligned}$$

■

Properties 8.1.4: Mean and Variance of $X \sim \mathcal{N}\left(\mu, \sigma^2\right)$

Since $X = \mu + \sigma Z$ we have that

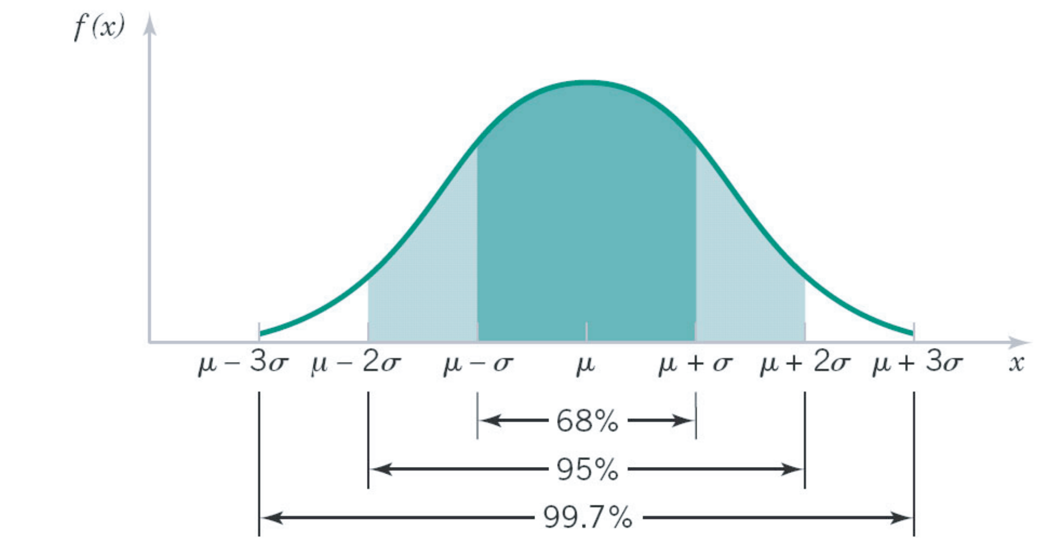
$$\mathbb{E}X = \mu + \sigma \mathbb{E}Z = \mu$$

Using our formula that $\mathrm{Var}(aY + b) = a^2 \mathrm{Var}(Y)$, for any r.v. Y, we have

$$\mathrm{Var}(X) = \sigma^2 \mathrm{Var}(Z) = \sigma^2$$

Properties 8.1.5: $X \sim \mathcal{N}\left(\mu, \sigma^2\right)$

"The 68 − 95 − 99.7 Rule"



Properties of the normal distributions relative to the mean.

Recall that the standard devision of a r.v. is SD $= \sigma = \sqrt{\mathrm{Var}(X)}$

(i) 68% of the observations lie within 1 SD of the mean.

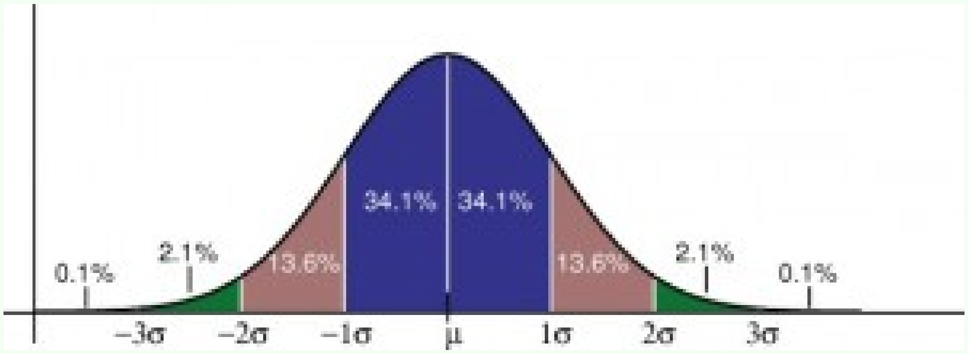
(ii) 95% of the observations lie within 2 SD's of the mean.

(iii) 99.7% of the observations lie within 3 SD's of the mean.

(iv) There is a very small probability of observing and outcome of of normal r.v. outside of 3σ away from the mean.

Properties 8.1.6: “The $68 - 95 - 99.7$ Rule.” or “Grading on the curve”

$X \sim \mathcal{N}\left(\mu, \sigma^2\right)$



“The 68 − 95 − 99.7 Rule”

Properties 8.1.7: Notation for the CDF of the standard normal

Let $Z \sim \mathcal{N}(0,1)$ be the standard normal as above. That is, Z has PDF

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

In this case It is customary to denote the cumulative distribution function of a standard normal as

$$F_Z(x) = \Phi(x) = \mathbb{P}(X \leq x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-y^2/2} dy$$

Due to the symmetry of the graph $\frac{1}{\sqrt{2\pi}} e^{-x^2/2}$, Φ satisfies the property that

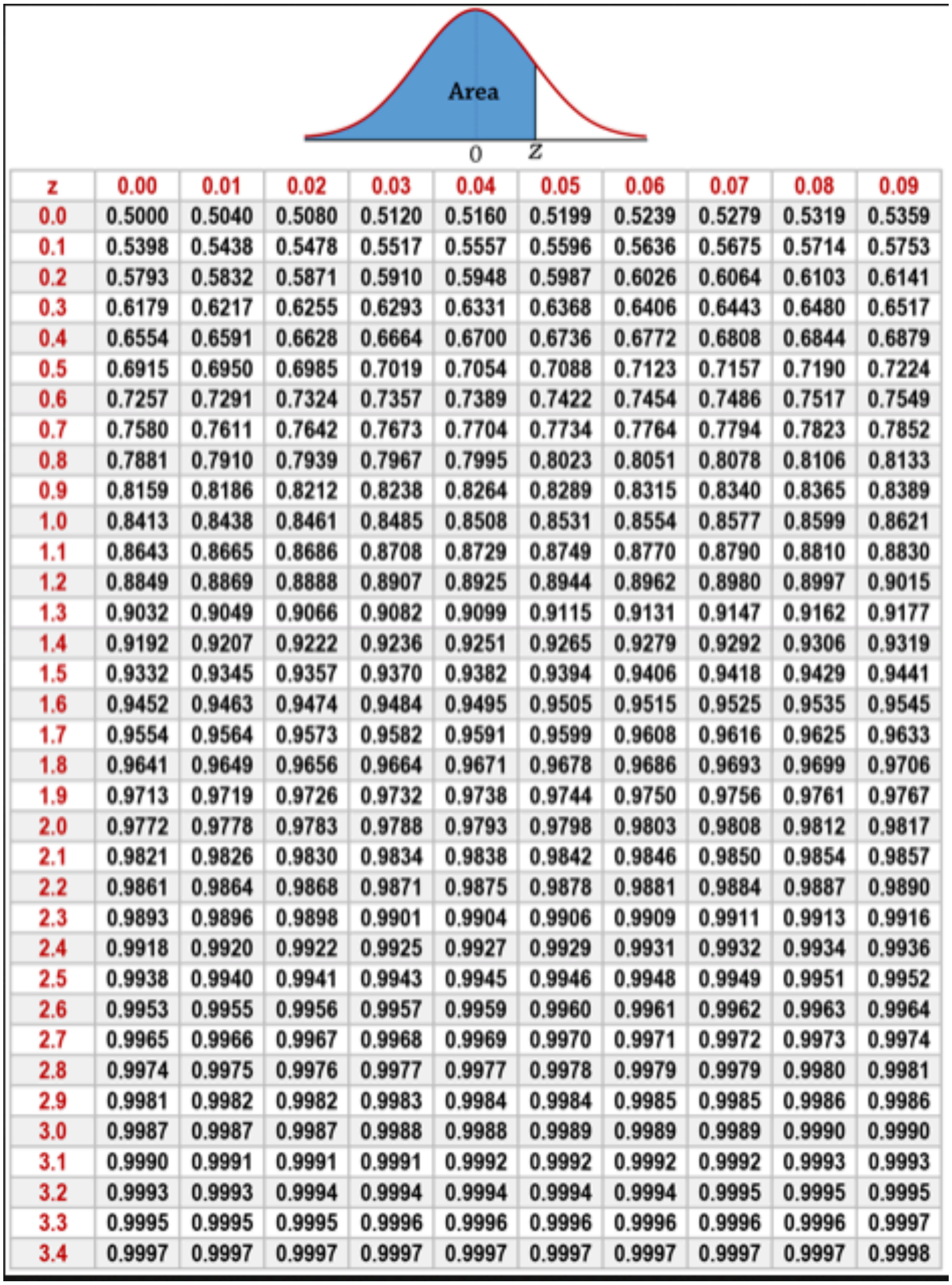
$$\Phi(-x) = 1 - \Phi(x), \text{ for any } \ -\infty < x < \infty.$$

This formula says that Φ for negative $x's$ is the same as $1 - \Phi$ for positive $x's$. For example,

$$\Phi(-3) = 1 - \Phi(3).$$

The equation follows from. Suppose $x > 0$. Then

$$\begin{aligned} \Phi(-x) &= \mathbb{P}(Z \leq -x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-x} e^{-y^2/2} dy \\ &= \frac{1}{\sqrt{2\pi}} \int_{x}^{\infty} e^{-y^2/2} dx \\ &= \mathbb{P}(Z \geq x) = 1 - \mathbb{P}(Z \leq x) \\ &= 1 - \Phi(x). \end{aligned}$$



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

8.1.1 Examples

Example 8.1.1. *Suppose $X \sim \mathcal{N}\left(\mu, \sigma^2\right)$ and $\mathbb{P}(X \leq 0) = 1/3$, $\mathbb{P}(X \leq 1) = 2/3$. (a) What are the values of μ and σ? (b) What if instead, $\mathbb{P}(X \leq 1) = 3/4$?*

Example 8.1.2. *Suppose the distribution of height over a large population of individuals is approximately normal. Ten percent of individuals in the population are over 6 feet tall, while the average height is 5 feet 10 inches. What, approximately, is the probability that in a group of 100 people picked at random from this population there will be two or more individuals over 6 feet 2 inches tall?*

Example 8.1.3. *Find $\mathbb{P}(1 \leq X \leq 4)$ if $X \sim \mathcal{N}(2, 25)$.*

Solution. *Then $\mu = 2$ and $\sigma^2 = 25$ thus $\frac{X-2}{5} = Z$ so that*

$$\begin{aligned}
\mathbb{P}(1 \leq X \leq 4) &= \mathbb{P}\left(\frac{1-2}{5} \leq \frac{X-2}{5} \leq \frac{4-2}{5}\right) \\
&= \mathbb{P}(-0.2 \leq Z \leq 0.4) \\
&= \mathbb{P}(Z \leq .4) - \mathbb{P}(Z \leq -.2) \\
&= \Phi(0.4) - \Phi(-0.2) \\
&= .6554 - (1 - \Phi(0.2)) \\
&= .6554 - (1 - .5793).
\end{aligned}$$

Example 8.1.4. *Suppose X is normal with mean 6. If $\mathbb{P}(X > 16) = .0228$, then what is the standard deviation of X?*

Solution. *We apply the properties from above: that says $\frac{X-\mu}{\sigma} = Z$ is $\mathcal{N}(0,1)$ and get*

$$\begin{aligned}
\mathbb{P}(X > 16) = .0228 &\iff \mathbb{P}\left(\frac{X-6}{\sigma} > \frac{16-6}{\sigma}\right) = .0228 \\
&\iff \mathbb{P}\left(Z > \frac{10}{\sigma}\right) = .0228 \\
&\iff 1 - \mathbb{P}\left(Z \leq \frac{10}{\sigma}\right) = .0228 \\
&\iff 1 - \Phi\left(\frac{10}{\sigma}\right) = .0228 \\
&\iff \Phi\left(\frac{10}{\sigma}\right) = .9772.
\end{aligned}$$

Using the standard normal table we see that $\Phi(2) = .9772$, thus we must have that

$$2 = \frac{10}{\sigma}$$

and hence $\sigma = 5$.

Example 8.1.5. *Suppose $X \sim \mathcal{N}(3, 9)$ find $\mathbb{P}(|X - 3| > 6)$.*

Solution.

$$\begin{aligned}
\mathbb{P}\left(|X-3|>6\right) &= \mathbb{P}\left(X-3>6\right)+\mathbb{P}\left(-\left(X-3\right)>6\right)\\
&= \mathbb{P}\left(X>9\right)+\mathbb{P}\left(X<-3\right)\\
&= \mathbb{P}\left(Z>2\right)+\mathbb{P}\left(Z<-2\right)\\
&= 1-\Phi(2)+\Phi(-2)\\
&= 2\left(1-\Phi(2)\right)\\
&\approx .0456.
\end{aligned}$$

Example 8.1.6. **Grading on the curve:** *In a perfect world where exam scores can be modeled with the Normal distribution, instructors often use the test scores to estimate the normal parameters μ and σ^2 and then assign the letter grade A to those whose test scores are greater than $\mu+\sigma$, B to those whose scores are between μ and $\mu+\sigma$, C to those whose scores are between $\mu-\sigma$ and μ, and D to those whose scores are between $\mu-2\sigma$ and $\mu-\sigma$ and F to those getting a score below $\mu-2\sigma$. With this in mind, what are the expected percentages of A's, B's, C's, D's, F's?*

Solution. *#A's*

$$\mathbb{P}(X>\mu+\sigma)=\mathbb{P}(\frac{X-\mu}{\sigma}>1)=1-\Phi(1)\approx 0.1587$$

About 16% *A's.*

#B's

$$\mathbb{P}(\mu<X<\mu+\sigma)=\mathbb{P}(0<\frac{X-\mu}{\sigma}<1)=\Phi(1)-\Phi(0)=0.3413$$

About 34% *B's.*

#C's

$$\mathbb{P}(\mu-\sigma<X<\mu)=\mathbb{P}(-1<\frac{X-\mu}{\sigma}<0)=\Phi(0)-\Phi(-1)=0.3413$$

About 34% C's.

#D's

$$\mathbb{P}(\mu-2\sigma<X<\mu-\sigma)=\mathbb{P}(-2<\frac{X-\mu}{\sigma}<-1)=\Phi(-1)-\Phi(-2)=\Phi(2)-\Phi(-1)\approx 0.1359$$

About 13.5% *D's*

#F's

$$\mathbb{P}(X<\mu-\sigma)=\mathbb{P}(\frac{X-\mu}{\sigma}<-2)=\Phi(-2)=1-\Phi(2)\approx 0.228$$

About 2% *F's*

Remark 8.1.1. *This numbers can also be read from the graph given in Property 1.6 above. But of course, the way those numbers were found is exactly by doing the calculations we just did. "If you know the math no need to remember formulas (formulae)." Keep this in mind as you proceed on your career. Memorizing formulas can (is) useful for taking exams but "useless" in the long term as one will soon forget them. Learning the concepts and theory behind them is "priceless" as these will remain with you for a long time and can be use to discover new applications and even new formulas!*

Example 8.1.7. *The annual rainfall in inches in the midwest is normally distributed with $\mu = 40$ and $\sigma = 4$. What is the probability that, starting with this year, it will take over 10 years before a year occurs having a rainfall of over 50 inches?*

Solution.

$$\mathbb{P}(X > 50) = \mathbb{P}(\frac{X-40}{4} > \frac{10}{4}) = 1 - \Phi(2.5) \approx 1 - 0.9938$$

or

$$\mathbb{P}(X < 50) = 0.9938.$$

Assuming independence of rain from year to year, the probability that it will take over 10 years before a year occurs having a rainfall of over 50 inches is approximately $(0.9938)^{10}$.

8.2 The normal approximates the binomial

Theorem 8.2.1

If S_n is a binomial with parameter n and p, then

$$\mathbb{P}\left(a \leq \frac{S_n - np}{\sqrt{np\,(1-p)}} \leq b\right) \to \mathbb{P}\left(a \leq Z \leq b\right)$$

as $n \to \infty$ where $Z \sim \mathcal{N}(0,1)$.

Or, if S_n denotes the number of successes that occur when n independent trials, each resulting in a success with probability p, are performed, then, for any $a < b$,

$$\mathbb{P}\left(a \leq \frac{S_n - np}{\sqrt{np\,(1-p)}} \leq b\right) \to \mathbb{P}\left(a \leq Z \leq b\right) = \Phi(b) - \Phi(a).$$

Recall that if $S_n \sim Bin\,(n,p)$ then its mean is $\mu = np$ and standard deviation is $\sigma = \sqrt{np(1-p)}$. Thus what the theorem says is that we can compute $\mathbb{P}\left(a \leq S_n \leq b\right)$ by approximating it with a normal. That is, for large n,

$$\frac{S_n - np}{\sqrt{np\,(1-p)}} \approx Z$$

which is the same as

$$\frac{S_n - \mu}{\sigma} \approx Z.$$

We can then write (again for large n)

$$\begin{aligned}\mathbb{P}\left(a \leq S_n \leq b\right) &= \mathbb{P}\left(\frac{a-\mu}{\sigma} \leq \frac{S_n-\mu}{\sigma} \leq \frac{b-\mu}{\sigma}\right) \\ &\approx \mathbb{P}\left(\frac{a-\mu}{\sigma} \leq Z \leq \frac{b-\mu}{\sigma}\right).\end{aligned}$$

It is important to emphasize here using the above we are not approximating the discrete probability mass function (pmf) of S_n by the continuous probability density function (PDF) of Z. If we try to do that we will get, for every value of i taken by the discrete r.v., S_n that

$$\begin{aligned}\mathbb{P}\left(S_n = i\right) &= \mathbb{P}\left(\frac{S_n-\mu}{\sigma} = \frac{i-\mu}{\sigma}\right) \\ &\approx \mathbb{P}\left(Z = \frac{i-\mu}{\sigma}\right) = 0\end{aligned}$$

since for a continuous random variables X $\mathbb{P}\left(X = a\right) = 0$ for every $\in \mathbb{R}$. Hence we need inequalities if we want to estimate a discrete random variable using a continuous random variable. To do this we observe that

$$\mathbb{P}(S_n = i) = \mathbb{P}\left(i - \frac{1}{2} < S_n < i + \frac{1}{2}\right)$$

because S_n can only be integers, so we are not hurting anything by saying " $i - \frac{1}{2} < S_n < i + \frac{1}{2}$" as we know that S_n can only be i in that interval anyways. This is called the "$\pm\frac{1}{2}$ continuity correction." We will elaborate on this further in week 14 when we will see this result as a special case of the Central Limit Theorem.

Example 8.2.1. *Why is this approximation useful in computations? Suppose a fair coin is tossed 100 times. What is the probability there will be more than 60 heads?*

Solution. *If we let S_{100} be the numbers of heads in 100 coin tosses, then $S_{100} \sim Bin\left(100, \frac{1}{2}\right)$ and we know that*

$$\begin{aligned}\mathbb{P}\left(S_{100} > 60\right) &= \sum_{i=61}^{100} \mathbb{P}\left(S_{100} = i\right) \\ &= \sum_{i=61}^{100} \left(\begin{array}{c}100\\ i\end{array}\right) \left(\frac{1}{2}\right)^i \left(\frac{1}{2}\right)^{100-i} \\ &= \sum_{i=61}^{100} \left(\begin{array}{c}100\\ i\end{array}\right) \left(\frac{1}{2}\right)^{100}.\end{aligned}$$

But this would be almost impossible to do by hand. On the other hand, if we set $n = 100$, $\mu = np = 50$ and $\sigma = \sqrt{np(1-p)} = \sqrt{50\frac{1}{2}} = 5$. We want more than 60, so approximate using $60 + \frac{1}{5}$:

$$\begin{aligned}\mathbb{P}\left(S_{100} > 60\right) = \mathbb{P}\left(S_{100} \geq 60.5\right) &= \mathbb{P}\left(\frac{S_{100}-50}{5} \geq \frac{60.5-50}{5}\right) \\ &\approx \mathbb{P}\left(Z \geq 2.1\right) \\ &= 1 - \Phi(2.1) \\ &\approx .0179\end{aligned}$$

Example 8.2.2. *Suppose instead we want to estimate the probability of getting exactly* 60 *heads. Then*

$$\begin{aligned}\mathbb{P}\left(S_n = 60\right) &= \mathbb{P}\left(59.5 \leq S_n \leq 60.5\right) \\ &\approx \mathbb{P}\left(1.9 \leq Z \leq 2.1\right) \\ &= \Phi(2.1) - \Phi(1.9).\end{aligned}$$

and we could also find this on the table given above. Notice that we could write this expressions as

$$\mathbb{P}\left(S_n = 60\right) \approx \frac{1}{\sqrt{2\pi}} \int_{1.9}^{2.1} e^{-x^2/2} dx.$$

Earlier (see Theorem 8.2 in Lecture Notes for week 6), we saw that a Poisson random variable is the limit of Binomial random variables. Thus it stands to reason that if the Normal distribution approximates the Binomial, it should also approximate the Poisson. We state it here in an "informal" way.

Properties 8.2.1: Approximation of Poisson by Normal

Let $X \sim Pois(\lambda)$. Recall that its mean $\mu = \lambda$ and its variance $\sigma^2 = \lambda$ as well. Then if λ is sufficiently large,

$$\mathbb{P}(a \leq \frac{X-\lambda}{\sqrt{\lambda}} \leq b) \approx \mathbb{P}(a \leq Z \leq b) = \Phi(b) - \Phi(a).$$

8.3 The Exponential Distribution

We have already discuses this in week 7. This is repeated from there.

A continuous random viable X with PDF given by

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & if\ x \geq 0 \\ 0 & if\ x < 0 \end{cases}$$

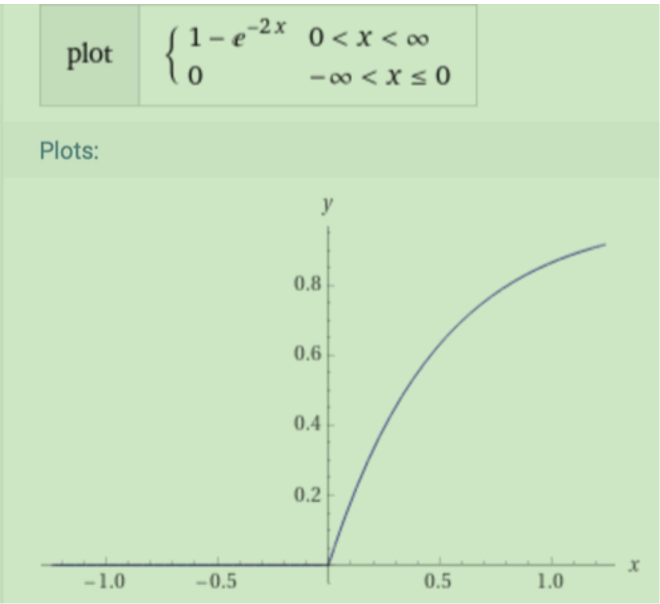
where $\lambda > 0$ is called an exponential random variable or is said to have an exponential distribution. For any $a \geq 0$, its distribution function is

$$F(a) = \mathbb{P}(X \leq a) = \int_0^a \lambda e^{-\lambda x} dx = -e^{-\lambda x}\Big|_0^a = 1 - e^{-\lambda a}$$

and otherwise $F(a) = 0$. That is,

$$F(a) = \begin{cases} 1 - e^{-\lambda a} & a \geq 0 \\ 0 & a < 0 \end{cases}$$

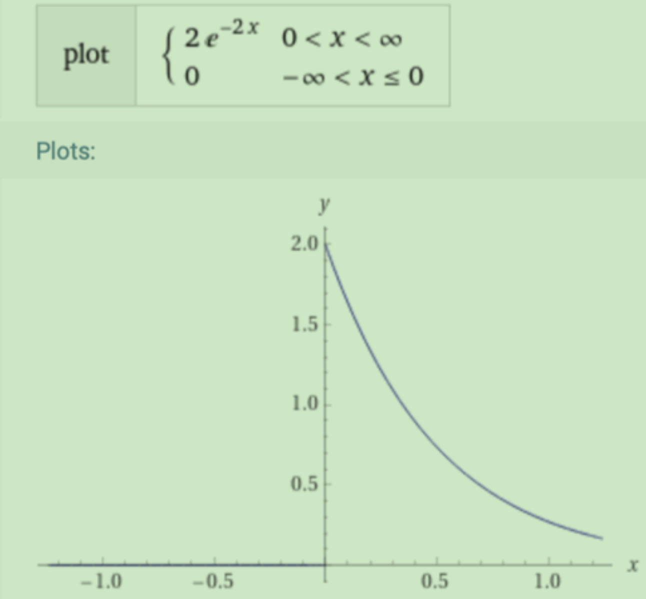
The notation $X \sim$Exp(λ) is used and we say that X *is exponentially distributed with parameter λ.*



plot $\begin{cases} 1 - e^{-2x} & 0 < x < \infty \\ 0 & -\infty < x \leq 0 \end{cases}$

Plots:

CDF graph of Exp(2)



plot $\begin{cases} 2\,e^{-2x} & 0 < x < \infty \\ 0 & -\infty < x \leq 0 \end{cases}$

Plots:

PDF graph of Exp(2)

The exponential distribution arises as the distribution of the amount of time until some specific event occurs. For instance, the amount of time (starting from now) until the next storm arrives, or until the next earthquake occurs, or or your next phone call, email or text message arrive.

One use integration by parts to find both $\mathbb{E}(X)$ and $\mathbb{E}(X^2)$ and hence the variance of X. But more generally, we can also proceed as follows. Let $r > 0$ (any real number). Then integration by parts with

$u = x^r$ and $dv = \lambda e^{-\lambda x}$ gives

$$\begin{aligned}\mathbb{E}(X^r) &= \int_0^\infty \lambda x^r e^{-\lambda x} dx \\ &= -x^r e^{-\lambda x}\Big|_0^\infty + r\int_0^\infty x^{r-1}e^{-\lambda x}dx \\ &= \frac{r}{\lambda}\int_0^\infty x^{r-1}\lambda e^{-\lambda x}dx = \frac{r}{\lambda}\mathbb{E}(X^{r-1}).\end{aligned} \tag{8.3.1}$$

Applying this with $r = 1$ gives

$$\mathbb{E}X = \frac{1}{\lambda} \tag{8.3.2}$$

and with $r = 2$ gives

$$\mathbb{E}X^2 = \frac{2}{\lambda}\mathbb{E}X = \frac{2}{\lambda^2} \tag{8.3.3}$$

Note that if $r = n$ a positive integer then

$$\mathbb{E}X^n = \frac{n!}{\lambda^n} \tag{8.3.4}$$

From these calculations,

$$\text{Var}(X) = \mathbb{E}X^2 - (\mathbb{E}X)^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2} \tag{8.3.5}$$

Definition 8.3.1: Lack of memory property (or memoryless)

We say that a nonnegative random variable X has the lack of memory property if

$$\mathbb{P}(X > s + t | X > t) = \mathbb{P}(X > s), \quad \text{for all } t, s \geq 0.$$

If we think of X as being the lifetime of some instrument, the equation states that the probability that the instrument survives for at least $s+t$ hours, given that it has survived t hours, is the same as the initial probability that it survives for at least s hours. In other words, if the instrument is alive at age t, the distribution of the remaining amount of time that it survives is the same as the original lifetime distribution. It is as if the instrument does not remember that it has already been in use for time t,

Properties 8.3.1

The exponential distribution has the lack of memory property. That is,

$$\mathbb{P}(X > s + t | X > t) = \mathbb{P}(X > s), \quad \text{for all } t, s \geq 0.$$

To see this, observe that (since $(X > s + t) \subset (X > t)$)

$$\begin{aligned} \mathbb{P}(X > s + t | X > t) &= \frac{\mathbb{P}(X > s + t, X > t)}{\mathbb{P}(X > t)} \\ &= \frac{\mathbb{P}(X > s + t)}{\mathbb{P}(X > t)} \\ &= \frac{e^{-\lambda(s+t)}}{e^{-\lambda t}} = e^{-\lambda s} = \mathbb{P}(X > s). \end{aligned}$$

8.3.1 Examples

Example 8.3.1. *Suppose that the length of time to service a customer at the postoffice in minutes is an exponential random variable with parameter $\lambda = \frac{1}{10}$. If someone arrives immediately ahead of you, find the probability that you will have to wait*

(i) *more than10 minutes,*

(ii) *between 10 and 20 minutes.*

Solution. *Let X denote the length of time the customer is at the service desk. Then,*

(i)

$$\mathbb{P}(X > 10) = 1 - F(10) = e^{-1} \approx 0.369.$$

$$\mathbb{P}(10 < X < 20) = F(20) - F(10) = e^{-1} - e^{-2}$$

Example 8.3.2. *Suppose the life of an Iphone has exponential distribution with mean life of* 4 *years*

(a) What is the probability the phone lasts more than 5 years?

(b) Given that the iPhone has already lasted 3 *years, what is the probability that it will last another* 5 *more years?*

Solution. *Let X denote the life of an iPhone. Note that $X \sim$ exponential$(\frac{1}{4})$*

(a) since $\lambda = \frac{1}{\mu} = \frac{1}{4}$. Then

$$\mathbb{P}\left(X > 5\right) = e^{-\frac{1}{4}\cdot 5}.$$

(b) We compute

$$\begin{aligned}
\mathbb{P}(X > 5+3 \mid X > 3) &= \frac{\mathbb{P}((X > 8) \cap (X > 3))}{\mathbb{P}(X > 3)} \\
&= \frac{\mathbb{P}(X > 8)}{\mathbb{P}(X > 3)} \\
&= \frac{e^{-\frac{1}{4}\cdot 8}}{e^{-\frac{1}{4}\cdot 3}} \\
&= e^{-\frac{1}{4}\cdot 5}.
\end{aligned}$$

Memoryless: Note that the probability of lasting 5 *more years, is the same as if it started* 5 *years from anew!!*

Example 8.3.3. *The number of days from beginning of a calendar year until accident for a bad driver is exponentially distributed. An insurance company expects 30% of bad drivers will have an accident during first 50 days. What is the probability that a bad driver will have an accident during first 80 days?*

Solution. *Let $X \sim exp(\lambda)$ number of days until accident. We know*

$$.3 = \mathbb{P}(X \leq 50) = \int_0^{50} \lambda e^{-\lambda x} dx = -e^{-\lambda t} \mid_0^{50} = 1 - e^{-50\lambda}.$$

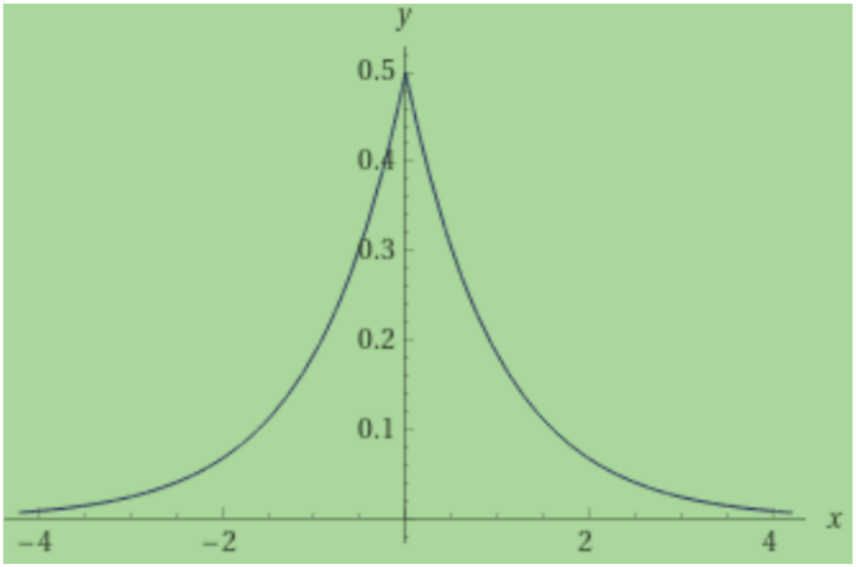
Solve for λ and get $\lambda = -\frac{1}{50} \ln .7$. Then compute

$$\begin{aligned}
\mathbb{P}(X \leq 80) &= \int_0^{80} \lambda e^{-\lambda x} dx = -e^{-\lambda t} \mid_0^{80} = 1 - e^{-80\lambda} \\
&= 1 - e^{\left(\frac{80}{50}\right) \ln .7} = .435.
\end{aligned}$$

8.3.2 Double Exponential

A double exponential random variable is a r.v. X with PDF given by

$$f(x) = \frac{1}{2} \lambda e^{\lambda |x|}, \ -\infty < x < \infty.$$



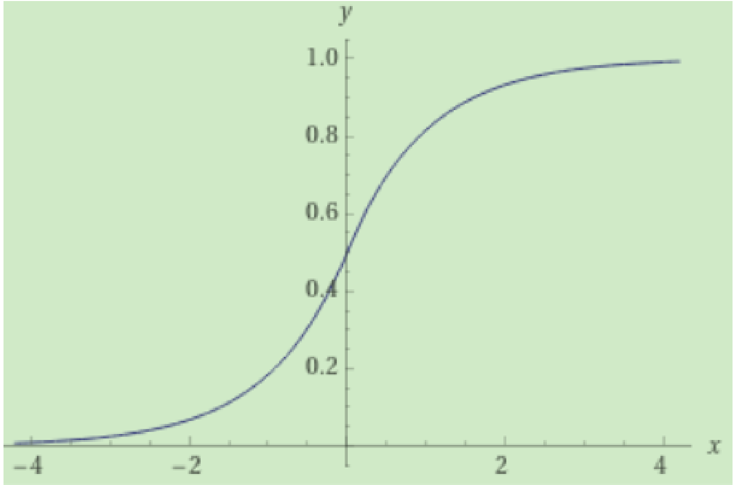
Graph PDF with $\lambda = 1$

The 1/2 is needed here since

$$\int_{-\infty}^{\infty} \frac{1}{2}\lambda e^{-\lambda|x|}dx = 2\int_{0}^{\infty} \frac{1}{2}\lambda e^{-\lambda|x|}dx = 1$$

The CDF is given by

$$\begin{aligned} F(a) &= \begin{cases} \frac{1}{2}\int_{-\infty}^{a} \lambda e^{\lambda x}dx, & a < 0 \\ \frac{1}{2}\int_{\infty}^{0} \lambda e^{\lambda a}dx + \frac{1}{2}\lambda e^{-\lambda x}dx, & \geq 0 \end{cases} \\ &= \begin{cases} \frac{1}{2}e^{\lambda a}, & a < 0 \\ 1 - \frac{1}{2}e^{-\lambda a}, & a \geq 0 \end{cases} \end{aligned}$$



Graph CDF with $\lambda = 1$

8.4 The Uniform Distribution

For any $-\infty < a < b < \infty$, a continuous random variable X is said to be uniformly distributed on the interval $[a, b]$ if its PDF

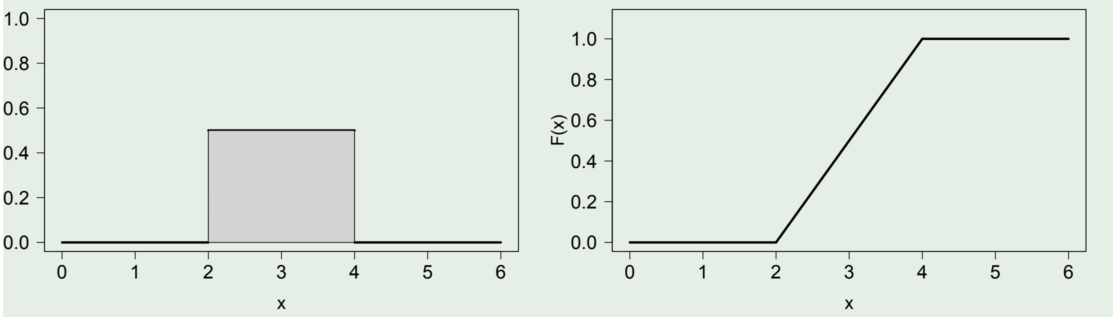
$$f(x) = \begin{cases} \frac{1}{b-a}, & a \le x \le b \\ 0, & \text{otherwise} \end{cases} = \frac{1}{b-a} \mathbb{1}_{[a,b]}(x).$$

This normalization is needed so that the area under the curve is 1.

Thus, X can only attain values in $X \in [a, b]$. We write this as $X \sim Uniform(a, b)$. The CDF is

$$F_X(x) = \int_{-\infty}^{x} f(t)dt = \begin{cases} 0 & x < a \\ \frac{x-a}{b-a} & a \le x \le b \\ 1 & x > b \end{cases}.$$

Example of uniform on $[2, 4]$, PDF and CDF:



Uniform(2, 4), PDF (left) CDF (right)

Observe that

$$\begin{aligned} \mathbb{E}X &= \int_{-\infty}^{\infty} x f_X(x) dx = \int_a^b x \frac{1}{b-a} dx \\ &= \frac{1}{b-a} \left(\frac{b^2}{2} - \frac{a^2}{2} \right) = \frac{a+b}{2}. \end{aligned}$$

Which makes sense, right? It should be the midpoint of the interval $[a, b]$.

We compute first the second moment

$$\begin{aligned} \mathbb{E}X^2 &= \int_a^b x^2 \frac{1}{b-a} dx = \frac{1}{b-a} \left(\frac{b^3}{3} - \frac{a^3}{3} \right) \\ &= \frac{1}{3} \frac{1}{b-a} (b-a) \left(a^2 + ab + b^2 \right) \\ &= \frac{a^2 + ab + b^2}{3}. \end{aligned}$$

Thus after some algebra

$$\mathrm{Var}X = \frac{a^2+ab+b^2}{3} - \left(\frac{a+b}{2}\right)^2 = \frac{(b-a)^2}{12}.$$

8.4.1 Examples

Example 8.4.1. *X and Y be equally distributed. That is, they both have the same PDF, $f_X = f_Y$. Then*

$$\mathbb{E}(g(X)) = \mathbb{E}(g(Y)).$$

Example 8.4.2. *Let X be a r.v. with mean $\mu_X = \mathbb{E}X$. Then its variance is given by*

$$Var(X) = \int_{-\infty}^{\infty} (x-\mu_X)^2 f(x)dx$$

Example 8.4.3. *A stick of length 1 is split at a point U that is uniformly distributed over (0, 1). Determine the expected length of the piece that contains the point $0 \le x \le 1$.*

Solution. *Let $L_x(U)$ be the length of the substick that contains the point x. Then*

$$L_x(U) = (1-U)\mathbb{1}_{(0,x)}(U) + U\mathbb{1}_{(x,1)}(U) = \begin{cases} (1-U), & 0 < U < x \\ U, & x < U < 1 \end{cases}$$

Then

$$\mathbb{E}(L_x(U)) = \int_0^1 L_x(u)du = \int_0^x (1-u)du + \int_x^1 udu = \frac{1}{2} + x(1-x)$$

Example 8.4.4. *Suppose that if you are s minutes early for an appointment, then you incur the cost $a \cdot s$ and if you are s minutes late, then you incur the cost $b \cdot s$. Suppose also that the travel time from where you presently are to the location of your appointment is a continuous random variable X having a pdf f and CDF F. Determine the time at which you should depart if you want to minimize your expected cost.*

Solution. *Let X be your time traveled. Let $C_t(X)$ be your cost if you leave t minutes before your appointment. Then*

$$C_t(X) \quad = \quad a(t-X)\mathbb{1}_{[0,\,t]}(X) + b(X-t)\mathbb{1}_{[t,\,\infty)}(X)$$

Set

$$\begin{aligned} h(t) = \mathbb{E}[C_t(X)] &= \int_0^\infty C_t(x)f(x)dx = a\int_0^t (t-x)f(x)dx + b\int_t^\infty (x-t)f(x)dx \\ &= at\int_0^t f(x)dx - a\int_0^t xf(x)dx + b\int_t^\infty xf(x)dx - bt\int_t^\infty f(x)dx \end{aligned}$$

Differentiating with respect to t (using the Fundamental Theorem of Calculus, product rule and formula for the CDF) Gives

$$\begin{aligned} h'(t) &= aF(t) + atf(t) - atf(t) - btf(t) - b[1-F(t)] + btf(t) \\ &= (a+b)F(t) - b \end{aligned}$$

Thus setting this equal to zero show that the minimizing time, call it t_m, is such that

$$F(t_m) = \frac{b}{a+b},$$

or equivalently, for which,

$$\mathbb{P}(X \leq t_m) = \frac{b}{a+b}.$$

Problem 8.4.1. *A point is chosen at random on a line segment of length L. Interpret this statement, and find the probability that the ratio of the shorter to the longer segment is less than* $1/4$

Problem 8.4.2. *You arrive at a bus stop at 10 o'clock, knowing that the bus will arrive at some time uniformly distributed between 10 and 10:30. (a) What is the probability that you will have to wait longer than 10 minutes? (b) If, at 10:15, the bus has not yet arrived, what is the probability that you will have to wait at least an additional 10 minutes?*

Problem 8.4.3. *A product that is sold seasonally yields a net profit of b dollars for each unit sold and a net loss of ℓ dollars for each unit left unsold when the season ends. Suppose that the seasonal demand is a continuous random variable having probability density function f. The number of units of the product that are ordered at a specific department store during any season is a counties random variable with PDF f. Show that the optimal amount to stock is the value s^* that satisfies*

$$F(s^*) = \frac{b}{b+\ell}$$

Hint: Work as in the above example with the profit

$$P(s) = \begin{cases} bX - (s-X)\ell & X \leq s \\ sb & X > s \end{cases}$$

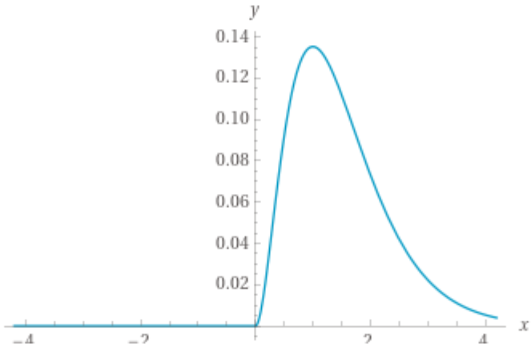
8.5 The Gamma Distribution

We say X has a Gamma distribution and write $X \sim Gamma\,(\alpha, \lambda)$, with parameter $\alpha, \lambda > 0$, if its density is given by

$$f(x) = \begin{cases} \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

where $\Gamma(\alpha)$ is the Gamma function

$$\Gamma(\alpha) = \int_0^\infty e^{-y} y^{\alpha-1} dy.$$



Gamma(3, 2)

The Gamma function satisfies the identity

$$\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1). \tag{8.5.1}$$

To see this, observe that integration by parts with $u = y^{\alpha-1}$ and $dv = e^{-y}$ gives

$$\begin{aligned}\Gamma(\alpha) &= -e^y y^{\alpha-1}\Big|_0^\infty + \int_0^\infty e^{-y}(\alpha-1)y^{\alpha-2}dy \\ &= (\alpha-1)\int_0^\infty e^{-y}y^{\alpha-1}dy = (\alpha-1)\Gamma(\alpha-2).\end{aligned}$$

Proposition 8.5.1: Moments of the Gamma Distribution

Let X ~Gamma(α, λ). Then for any positive integer $k = 1, 2, \ldots$

$$\mathbb{E}X^k = \frac{\Gamma(\alpha+k)}{\lambda^k\Gamma(\alpha)}$$

Proof.

$$\begin{aligned}\mathbb{E}X^k &= \int_0^\infty x^k f(x)dx = \frac{1}{\Gamma(\alpha)}\int_0^\infty \lambda^\alpha x^{(\alpha+k-1)}e^{-\lambda x}dx \\ &= \left(\frac{\Gamma(\alpha+k)}{\lambda^k\Gamma(\alpha)}\right)\left(\frac{1}{\Gamma(\alpha+k)}\right)\int_0^\infty \lambda^{\alpha+k}x^{(\alpha+k-1)}e^{-\lambda x}dx \\ &= \frac{\Gamma(\alpha+k)}{\lambda^k\Gamma(\alpha)}\end{aligned}$$

where we used the fact that

$$\frac{1}{\Gamma(\alpha+k)}\int_0^\infty \lambda^{\alpha+k}x^{(\alpha+k-1)}e^{-\lambda x}dx = 1$$

■

If we now apply this with $k = 1$ and $k = 2$ we can compute the mean and variance of the Gamma distribution.

Properties 8.5.1: Mean and Variance Gamma

Let X ~Gamma(α, λ). Then using (8.5.1) we have

$$\mu = \mathbb{E}X = \frac{\Gamma(\alpha+1)}{\lambda\Gamma(\alpha)} = \frac{\alpha}{\lambda}$$

Applying (8.5.1) twice gives

$$\begin{aligned}
\mathbb{E}X^2 &= \frac{\Gamma(\alpha+2)}{\lambda^2\Gamma(\alpha)} = \frac{(\alpha+1)\Gamma(\alpha+1)}{\lambda^2\Gamma(\alpha)} \\
&= \frac{(\alpha+1)\alpha\Gamma(\alpha)}{\lambda^2\Gamma(\alpha)} \\
&= \frac{(\alpha+1)\alpha}{\lambda^2} \\
&= \left(\frac{\alpha}{\lambda}\right)^2 + \frac{\alpha}{\lambda^2}.
\end{aligned}$$

Thus

$$\sigma^2 = \text{Var}(X) = \mathbb{E}X^2 - (\mathbb{E}X)^2 = \frac{\alpha}{\lambda^2}$$

It follows from (8.5.1) that when $\alpha = n = 1, 2, 3, \ldots, \Gamma(n) = (n-1)!$ In this case the Gamma distribution arises as the distribution of the sum of n independent exponential random variables. If an exponential random variable is viewed as the waiting time until the first event occurs, a Gamma random variable is the waiting time until the nth event occurs. Thus, if $X_1, X_2 \cdots X_n$ are independent exponential random variables with parameter λ and se set

$$X = X_1 + X_2 + \cdots + X_n$$

then X is a Gamma random variable, $X \sim \textit{Gamma}\,(n, \lambda)$ with density

$$f(x) = \begin{cases} \frac{\lambda^n}{(n-1)!}x^{n-1}e^{-\lambda x} & x \geq 0 \\ 0 & x < 0. \end{cases} \tag{8.5.2}$$

Definition 8.5.1: The n-Erlang distribution

The case *Gamma*(n, λ) is often called the "n-Erlang distribution." It arises in queueing theory often used in problems in engineering that consider waiting times in queueing systems. Its PDF is given by (8.5.2). Its CDF can be easily computed to be

$$F(x) = \sum_{k=n}^{\infty} \frac{e^{-\lambda x}(\lambda x)^k}{k!}$$

Example 8.5.1. *Consider 300 students who are waiting for service at the registrar's office. Assume that their waiting times are independent exponential random variables, and each waiting time has density* $f_X(x) = 2e^{-2x}, x > 0$ *and* $f(x) = 0$ *otherwise. Find the probability that the 300 students collectively (i.e., altogether) spend between 145 and 152 hours waiting for their appointments.*

Solution. *From our discussion above, this is the same as computing the probability that Gamma(300,2) random variable Y is between 145 and 152. This is*

$$\mathbb{P}(145 < Y < 152) = \frac{1}{299!}\int_{145}^{152} 2^{300} x e^{-2x} dx$$

Example 8.5.2. *A student waits for a bus 10 times during a week. Let X be the number of hours that the student waits. Assume that the waiting times are independent Exponential random variables, each with average 30 minutes. a. What is his expected time spent waiting for the 10 buses altogether? b. What is the standard deviation of his waiting time for the 10 buses altogether?*

Example 8.5.3. *A student estimates that the time needed to solve each homework problem is Exponentially distributed and is independent of all the other homework problems. Each problem takes, on average, 15 minutes to solve. a. What is the expected time spent on a 6-question homework assignment? b. What is the variance of the time spent on a 6-question homework assignment?*

Example 8.5.4. *Suppose X ∼Gamma(3, 5). Find the probability that X is larger that $\mathbb{E}X$*

8.6 The Chi-square Distribution

If Y ∼Gamma$\left(\frac{n}{2}, \frac{1}{2}\right) = \chi^2_n$ this is called the *Chi-squared distribution* with n degrees of freedom. The Chi-Square distribution is used a lot in statistics, especially in hypothesis testing. It arises as the sum of n independent standard normal random variables. (We shall see this later.) From our knowledge of the mean and variance of the Gamma distribution we see that the mean and variance of the chi-square $\mathbb{E}X = n$ and Var$X = 2n$. If the random variable $X \sim Chi - square$ we write this as $X \sim \chi^2(n)$. We summarize:

Properties 8.6.1: The Chi-Square Distribution

Suppose $X \sim \chi^2(n)$. Then

(a) X has PDF

$$f_X(x) = \frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})} x^{(n/2-1)} e^{x/2}$$

(b) X has CDF

$$F_X(x) = \frac{1}{\Gamma(n/2)}\gamma(n/2, x/2) = \frac{1}{\Gamma(n/2)}\int_0^{x/2} t^{n/2-1}e^{-t}dt$$

(c) X has mean

$$\mu = \mathbb{E}X = n.$$

(d) X has Var(X)

$$\sigma^2 = \text{Var}(X) = 2n$$

8.7 The Weibull Distribution

The Weibull distribution is widely used in engineering. It was originally proposed for the interpretation of fatigue data. In particular, it is widely used in the field of life phenomena as the distribution of the lifetime of some object, especially when the "weakest link" model is appropriate for the object. That is, consider an object consisting of many parts, and suppose that the object experiences failure when any of its parts fail. It has been shown (both theoretically and empirically) that under these conditions a Weibull distribution provides a close approximation to the distribution of the lifetime of the the item. The Weibull CDF is given by

$$F_X(x) = \begin{cases} 1 - e^{-(x/\alpha)^\beta}, & x > 0 \\ 0, & x \leq 0 \end{cases}$$

where α and β are positive. If the random variable X has the Weibull distribution with parameters α, β we write $X \sim Weibull(\alpha, \beta)$. Differentiating F_X we find that its PDF to be

$$f_X(x) = \begin{cases} \frac{\beta}{\alpha}\left(\frac{x}{\alpha}\right)^{\beta-1} e^{-(x/\alpha)^\beta}, \; x > 0 \\ 0, \; x \leq 0 \end{cases}$$

The substitution $u = (x/\alpha)^\beta$ immediately gives that

$$\mathbb{E}X = \beta \int_0^\infty \left(\frac{x}{\alpha}\right)^\beta e^{-(x/\alpha)^\beta} dx = \alpha \int_0^\infty u^{1/\beta} e^{-u} du = \alpha\Gamma(1/\beta + 1).$$

In the same way substitution

$$\mathbb{E}X^2 = \beta \int_0^\infty x \left(\frac{x}{\alpha}\right)^\beta e^{-(x/\alpha)^\beta} dx = \alpha^2 \int_0^\infty u^{2/\beta} e^{-u} du = \alpha^2\Gamma(2/\beta/2 + 1).$$

This gives the variance

$$\mathrm{Var}(X) = \alpha^2 \left[\Gamma(\beta/2 + 1) + (\Gamma(1/\beta + 1))^2\right]$$

Summary:

Properties 8.7.1: The Weibull Distribution

Suppose $X \sim Weibull(\alpha, \beta)$. Then

(a) X has CDF

$$F_X(x) = \begin{cases} 1 - e^{-(x/\alpha)^\beta}, & x > 0 \\ 0, & x \leq 0 \end{cases}$$

(b) X has PDF

$$f_X(x) = \begin{cases} \frac{\beta}{\alpha}\left(\frac{x}{\alpha}\right)^{\beta-1} e^{-(x/\alpha)^\beta}, \; x > 0 \\ \\ 0, \; x \leq 0 \end{cases}$$

(c) X has mean

$$\mu = \mathbb{E}X = \alpha\Gamma(1/\beta + 1).$$

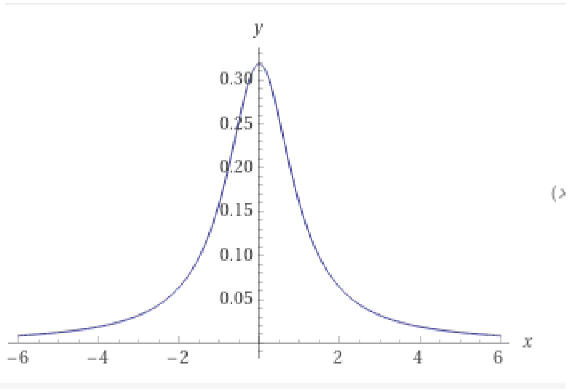
(d) X has $\mathrm{Var}(X)$

$$\sigma^2 = \mathrm{Var}(X) = \alpha^2 \left[\Gamma(\beta/2 + 1) + (\Gamma(1/\beta + 1))^2\right]$$

8.8 The Cauchy Distribution

The Cauchy distribution arises in many applications of mathematics to physics, especially in the theory of waves and heat propagation. It arises also as transformation of other distributions and in particular the normal distribution. For any $z \in \mathbb{R}$ and $t > 0$, we say that X is a Cauchy random variable with parameters t, z and write $X \sim$Cauchy(t, z) if its PDF function is given by

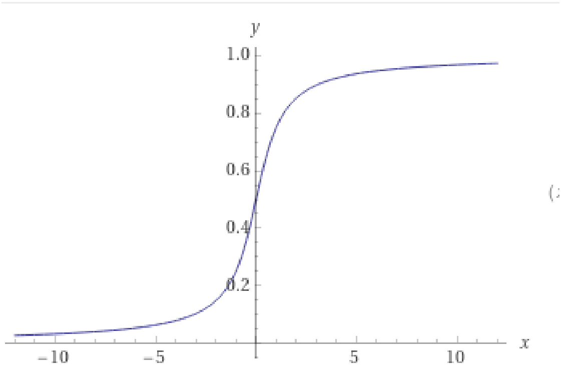
$$f_X(x) = \frac{t}{\pi(t^2 + |x - z|^2)} = \frac{1}{\pi t(1 + |\frac{x-z}{t}|^2)}$$



PDF: $t = 1$, $z = 0$

The CDF can be obtained from the fact that the the indefinite integral of $\frac{1}{1+|x|^2}$ is the arctan(x) (or $\tan^{-1}(x)$) to obtain the nice form

$$F_X(x) = \frac{1}{2} + \frac{1}{\pi} \arctan\left(\frac{x-z}{t}\right) = \frac{1}{2} + \frac{1}{\pi} \tan^{-1}\left(\frac{x-z}{t}\right)$$



CDF: $t = 1$, $z = 0$

It is important to note here that the Cauchy distribution does not have a mean. That is, That is $\mathbb{E}X$

does not exists. To see this, take $t = 1$ $z = 0$

$$\begin{aligned}\mathbb{E}X &= \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{x}{1+x^2}dx \\ &\sim \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{1}{x}dx \\ &\sim \lim_{x\to\infty} \ln|x| - \lim_{x\to-\infty} \ln|x|\end{aligned}$$

which is not defined. On the other hand, $\mathbb{E}|X|$ "does exists" but its infinite. That is,

$$\mathbb{E}|X| = \infty.$$

To see this, observe that

$$\mathbb{E}|X| = \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{|x|}{1+|x|^2}dx = \frac{2}{\pi}\int_{0}^{\infty} \frac{|x|}{1+|x|^2}dx = \infty.$$

Often in the literature this density is written as $P_t(x,z)$ and called the "Poisson kernel." To check that this is indeed a density observe that

$$\begin{aligned} t\int_{-\infty}^{\infty} \frac{dx}{\pi(t^2+|x-z|^2)}dx &= t\int_{-\infty}^{\infty} \frac{dx}{\pi(t^2+|x|^2)} \\ &= 2t\int_{0}^{\infty} \frac{dx}{\pi(t^2+|x|^2)} \\ &= \frac{2t}{\pi}\int_{0}^{\infty} \frac{dx}{(1+|\frac{x}{t}|^2)} \\ &= \frac{2}{\pi}\int_{0}^{\infty} \frac{dx}{(1+|x|^2)}, \text{ (u substitution with } u = \frac{x}{t}) \\ &= 1, \end{aligned}$$

where the last equality follows from Math 161 (the integral is the $\arctan(x)$. evaluated at the limits. As an example,

Problem 8.8.1. *Let $g : \mathbb{R} \to \mathbb{R}$ be a function and let $X \sim$Cauchy(t,z). Recall that*

$$\mathbb{E}(g(X)) = \int_{-\infty}^{\infty} g(x)f_X(x)dx = \frac{t}{\pi}\int_{-\infty}^{\infty} \frac{g(x)}{(t^2+|x-z|^2)}dx$$

Thus, we may think of the expectation as a function of the two variables $t > 0$ and $-\infty < z < \infty$. Set

$$U_g(t,z) = \mathbb{E}(g(X)).$$

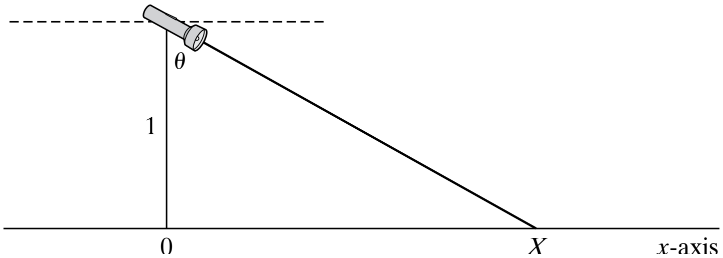
Show that this function satisfies the equation.

$$\frac{\partial^2 U_g(t,z)}{\partial^2 t} + \frac{\partial^2 U_g(t,z)}{\partial^2 z} = 0$$

Hint: You can do this by bring the derivatives inside the integral.

Remark 8.8.2. *The above equation is called Laplace's equation (named after Pierre-Simon Laplace (1749-1827) used in the theory of sound waves heat propagation. Functions satisfying this equation are called "Harmonic Functions." In fact the function $U_g(t, z)$ also satisfies the "initial condition" that $U_g(0, z) = g(z)$. That is, if you start with an initial wave at a point z given by $g(z)$ and let it evolve with time, the the wave at time t and point z is given by the expectation of the the random variable $g(X)$ where $X \sim$Cauchy(t, z). This gives a beautiful and extremely useful connection to probability.*

Example 8.8.1. *Suppose that a narrow-beam flashlight is spun around its center, which is located a unit distance from the x-axis as shown in the picture below. Consider the point X at which the beam intersects the x-axis when the flashlight has stopped spinning. (If the beam is not pointing toward the x-axis, repeat the experiment.)*



The angle θ is uniformly distributed on $[-\pi/2, \pi/2]$ and

$$\mathbb{P}(\theta < a) = \frac{a + \pi/2}{\pi} = \frac{1}{2} + \frac{a}{\pi}.$$

Thus,

$$\mathbb{P}(X \leq x) = \mathbb{P}(\tan(\theta) \leq x) = \mathbb{P}(\theta \leq \tan^{-1}(x) = \frac{1}{2} + \frac{1}{\pi}\tan^{-1}(x)$$

Differentiating this gives PDF of the Cauchy with $t = 1$ and $z = 0$.

8.9 The Beta Distribution

The Beta distribution arises in many examples of modeling dealing with with percents, proportions, and fractions. We say that the random variable X has the beta distribution with parameter a and b, and as usual write $X \sim Beta(a, b)$ if its density is

$$f_X(x_= \begin{cases} \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1}, \ 0 < x < 1 \\ 0, \ otherwise. \end{cases}$$

Here,

$$B(a, b) = \text{Beta(a, b)} = \int_0^1 x^{a-1}(1-x)^{b-1}dx = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

Notice that when $a = 1$, $b = 1$, the Beta distribution in just the Uniform distribution on $[0, 1]$. Notice that the formula for Beta(a, b) immediately gives the fact that the integral of f_X on $[0, 1]$ is 1. Hence, f_X is indeed a PDF. However, the formula for the CDF is not as easy to write explicitly. On the other hand, it is easy obtain its mean and variance. In fact, using the formula for $B(a,b)$ in terms of the Γ function and the fact that for any α, $\Gamma(\alpha+1)=\alpha\Gamma(\alpha)$ as in (8.5.1),

$$\begin{aligned}
\mu &= \mathbb{E}X = \frac{1}{B(a,b)}\int_0^1 x x^{a-1}(1-x)^{b-1}dx \\
&= \frac{1}{B(a,b)}B(a+1,b) \\
&= \frac{1}{B(a,b)}\frac{\Gamma(a+1)\Gamma(b)}{\Gamma(a+b+1)} \\
&= \left(\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\right)\left(\frac{a\Gamma(a)\Gamma(b)}{(a+b)\Gamma(a+b)}\right) \\
&= \frac{a}{a+b}
\end{aligned}$$

Similarly,

$$\begin{aligned}
\mathbb{E}X^2 &= \frac{1}{B(a,b)}\int_0^1 x^{a+1}(1-x)^{b-1}dx \\
&= \frac{1}{B(a,b)}B(a+2,b) \\
&= \left(\frac{1}{B(a,b)}\right)\left(\frac{\Gamma(a+2)\Gamma(b)}{\Gamma(a+b+2)}\right) \\
&= \left(\frac{1}{B(a,b)}\right)\left(\frac{(a+1)\Gamma(a+1)\Gamma(b)}{(a+b+1)\Gamma(a+b+1)}\right) \\
&= \frac{(a+1)}{(a+b+1)}\left(\frac{1}{B(a,b)}\right)\left(\frac{\Gamma(a+1)\Gamma(b)}{\Gamma(a+b+1)}\right) \\
&= \left(\frac{(a+1)}{(a+b+1)}\right)\left(\frac{a}{a+b}\right).
\end{aligned}$$

From this,

$$\sigma^2 = \mathbb{E}X^2 - (\mathbb{E}X)^2 = \mathrm{Var}(X) = \frac{ab}{(a+b)^2(a+b+1)}$$

Example 8.9.1. *A candy company that delivers candy to local stores needs to figure out the best schedule to deliver their products. From data collected at the store, they have concluded that store?s candy inventory can be modeled by a Beta distribution with $a=2$ and $b=5$ where X some proportion of the candy that is purchased.*

(a) What is the PDF of X.?

(b) What is the CDF of X?

(c) What is the expected proportion of candy that will be purchased and the standard deviation?

(d) What is the probability that more than 3/4 of the candy the company delivers will be purchased?

(a) From the formula we have with $a=2$ and $b=3$, that (recall $\Gamma(n)=(n-1)!$)

$$f(x) = 30x(1-x)^4, \;\; 0 \le x \le 1$$

and 0 otherwise.

(b) The CDF, $F(a)$ is zero for $a < 0$ and 1 for $a \geq 1$. For $0 < a < 1$ we compute the integral to find that

$$\begin{aligned} F(a) &= 30\int_0^a x(1-x)^4 dx \\ &= 30\int_{1-a}^1 (1-u)u^4 du \\ &= 30\int_{1-a}^1 (u^4-u^5) du \\ &= 1-6(1-a)^5+5(1-a)^6 \end{aligned}$$

c) The formula for the mean and variance given above gives that

$$\mathbb{E}X = \frac{2}{2+5} = 2/7.$$

$$SD = \sigma = \sqrt{\frac{2\cdot 5}{(2+5)^2(2+5+1)}} = \sqrt{5/196}.$$

(d)

$$\mathbb{P}(X > 3/4) = 1-\mathbb{P}(X \leq 3/4) = 1-(1-6(1-3/4)^5-5(1-3/4)^6) = \frac{19}{4096}$$

8.10 Finding the distribution of a function of a Random variable

We have already done several examples when we are given a random variable X and asked to compute the distribution of the random variable $Y = g(X)$, where g is some specific function. In this section we will give a general formula which works for quite general class of functions g. Recall that since

$$F(x) = \mathbb{P}(X \leq x) = \int_{-\infty}^{x} f(y)dy$$

then by the fundamental theorem of calculus we have

$$f(x) = F'(x)$$

and this is the fundamental relationship between the CDF of X and the PDF of X that we will use below. Let's start with some examples, some which we have already discussed before.

Example 8.10.1. *If X is continuous with distribution function F_X and density function f_X, find a formula for the density function of the random variable $Y = 2X$.*

Solution. *Let us emphasize the steps to do this:*
Step 1: *First you start with the distribution of Y: First start by writing the CDF of Y and in terms of the CDF of F_X:*

$$\begin{aligned} F_Y(x) &= \mathbb{P}(Y \leq x) \\ &= \mathbb{P}(2X \leq x) \\ &= \mathbb{P}\left(X \leq \frac{x}{2}\right) \\ &= F_X\left(\frac{x}{2}\right). \end{aligned}$$

Once the relationship between F_X and F_Y has been found, we can differentiate.
Step 2: *Then use the relation $f_Y(y) = F_Y'(y)$ and take a derivative of both sides to get*

$$\begin{aligned} F_Y'(x) &= \frac{d}{dx}\left[F_X\left(\frac{x}{2}\right)\right], \\ F_Y'(x) &= F_X'\left(\frac{x}{2}\right)\cdot\left(\frac{x}{2}\right)', \text{ by chain rule on RHS} \end{aligned}$$

Thus,

$$f_Y(x) = f_X\left(\frac{x}{2}\right)\frac{1}{2}.$$

Our general goal is to do the same for $Y = g(X)$ where $g : \mathbb{R} \to \mathbb{R}$ is a function given that we know the CDF and PDF for X. Why is this useful?

Example 8.10.2. *Suppose X represent the income for a random US worker. And let $Y = g(X)$ be the amount of taxes a US worker pays per year. Note that taxes Y is dependent on the random variable X. So if we only care about the random variable Y then finding its PDF and CDF can help us find out everything we need to know about Y given we can find the PDF. Recall that any probability and expected value can be found using the PDF*

Example 8.10.3. *Let us repeat the example above with a more concrete problem. Suppose you go to a picture frame store and randomly pick a frame. Let X be the side of a randomly chosen frame (in feet) and suppose you know the size of the frame has the following PDF:*

$$f_X(x) = \begin{cases} \frac{1}{9}x^2 & 0 < x < 3 \\ 0 & \textit{otherwise} \end{cases}$$

Let Y be the perimeter of the randomly chosen box. Find the PDF of Y.

Solution. *Solution: We want to find the PDF of $Y = 4X$. First start by writing the CDF of Y and in terms of F_X:*

$$\begin{aligned} F_Y(x) &= \mathbb{P}(Y \leq x) \\ &= \mathbb{P}(2X \leq x) \\ &= \mathbb{P}\left(X \leq \frac{x}{2}\right) \\ &= F_X\left(\frac{x}{2}\right). \end{aligned}$$

Then use the relation $f_Y(y) = F_Y'(y)$ and take a derivative of both sides:

$$\begin{aligned} F_Y'(x) &= \frac{d}{dx}\left[F_X\left(\frac{x}{2}\right)\right], \\ F_Y'(x) &= F_X'\left(\frac{x}{2}\right)\cdot\left(\frac{x}{2}\right)', \text{ by chain rule on RHS} \\ f_Y(x) &= f_X\left(\frac{x}{2}\right)\frac{1}{2} \\ &= \begin{cases} \frac{1}{9}\left(\frac{x}{2}\right)^2\cdot\frac{1}{2} & 0 < \frac{x}{2} < 3 \\ 0 & \textit{otherwise} \end{cases} \\ &= \begin{cases} \frac{x^2}{72} & 0 < x < 6 \\ 0 & \textit{otherwise} \end{cases}. \end{aligned}$$

Example 8.10.4. *Let $X \sim Uniform\,((0,10))$ and $Y = e^{3X}$. Find the PDF f_Y of Y.*

Solution. *Recall that since $X \sim Uniform\,((0,1))$ then*

$$f_X(x) = \begin{cases} \frac{1}{10} & 0 < x < 10 \\ 0 & otherwise \end{cases}.$$

Step1: *First start by writing the CDF of Y and in terms of F_X:*

$$\begin{aligned} F_Y(y) &= \mathbb{P}(Y \leq y) \\ &= \mathbb{P}\left(e^{3X} \leq y\right), \text{ then solve for } X \\ &= \mathbb{P}(3X \leq \ln y) \\ &= \mathbb{P}\left(X \leq \frac{1}{3}\ln y\right) \\ &= F_X\left(\frac{1}{3}\ln y\right). \end{aligned}$$

Step 2: *Then use the relation $f_Y(y) = F_Y'(y)$ and take a derivative*

$$\begin{aligned} f_Y(y) &= F_Y'(y) \\ &= \frac{d}{dy}\left[F_X\left(\frac{1}{3}\ln y\right)\right], \text{ use chain rule} \\ &= F_X'\left(\frac{1}{3}\ln y\right)\frac{1}{3y} \\ &= f_X\left(\frac{1}{3}\ln y\right)\frac{1}{3y}, \text{ since } F_X' = f_X \\ &= \begin{cases} \frac{1}{10}\cdot\frac{1}{3y} & 0 < \frac{1}{3}\ln y < 10 \\ 0 & otherwise \end{cases} \end{aligned}$$

but since

$$\begin{aligned} 0 < \frac{1}{3}\ln y < 10 &\iff 0 < \ln y < 30 \\ &\iff e^0 < y < e^{30} \\ &\iff 1 < y < e^{30}. \end{aligned}$$

then

$$f_Y(y) = \begin{cases} \frac{1}{30y} & 1 < y < e^{30} \\ 0 & otherwise \end{cases}.$$

Example 8.10.5. *A survey was conducted of all US households who have square swimming pools. It was found that a randomly selected square swimming pool in the US has side length X in meters with PDF given by*

$$f_X(x) = \begin{cases} \frac{1}{4}x & 1 < x < 3 \\ 0 & otherwise \end{cases}.$$

Let Y be the area of the bottom of the pool. Find the PDF of Y. What known distribution is Y?

Solution. *We want to find the PDF of $Y = X^2$.*

Step 1: *First start by writing the CDF of Y and in terms of F_X. We have*

$$\begin{aligned} F_Y(y) &= \mathbb{P}(Y \leq y) \\ &= \mathbb{P}\left(X^2 \leq y\right), \text{ then solve for } X \\ &= \mathbb{P}(X \leq \sqrt{y}) \\ &= F_X(\sqrt{y}). \end{aligned}$$

Step 2: *Then use the relation $f_Y(y) = F'_Y(y)$ and take a derivative of both sides*

$$\begin{aligned} f_Y(y) &= F'_Y(y) \\ &= \frac{d}{dy}\left[F_X(\sqrt{y})\right], \text{ use chain rule} \\ &= F'_X(\sqrt{y})\frac{1}{2\sqrt{y}} \\ &= f_X(\sqrt{y})\frac{1}{2\sqrt{y}}, \text{ since } F'_X = f_X \\ &= \begin{cases} \frac{1}{4}\sqrt{y} \cdot \frac{1}{2\sqrt{y}} & 1 < \sqrt{y} < 3 \\ 0 & otherwise \end{cases} \\ &= \begin{cases} \frac{1}{8} & 1 < y < 9 \\ 0 & otherwise \end{cases}. \end{aligned}$$

So it turn out

$$Y \sim U[1,9].$$

Problem 8.10.1. *Show that if $Z \sim \mathcal{N}(0,1)$ then $Y = Z^2$ is a Gamma with parameters $(\frac{1}{2}, \frac{1}{2})$.*

Example 8.10.6. *Let $X \sim Uniform((0,1])$ $Y = -ln(X)$. Find the pdf of Y? What distribution is it?*

Solution. *Recall that*

$$f_X(x) = \begin{cases} 1 & 0 < x < 1 \\ 0 & otherwise \end{cases}.$$

Step 1: *First start with the CDF Y and write it terms of F_X*

$$\begin{aligned} F_Y(x) &= \mathbb{P}(Y \leq x) \\ &= \mathbb{P}(-lnX \leq x) \\ &= \mathbb{P}(lnX > -x) \\ &= \mathbb{P}\left(X > e^{-x}\right) \\ &= 1 - \mathbb{P}\left(X \leq e^{-x}\right) \\ &= 1 - F_X\left(e^{-x}\right). \end{aligned}$$

Step 2: *Then take a derivative*

$$\begin{aligned} f_Y(x) &= F_Y'(x) \\ &= 1 - \frac{d}{dx} F_X\left(e^{-x}\right) \\ &= -F_X'(e^{-x}) \cdot \left(-e^{-x}\right) \\ &= -f_X(e^{-x}) \cdot \left(-e^{-x}\right) \\ &= f_X(e^{-x}) \cdot e^{-x} \\ &= \begin{cases} 1 \cdot e^{-x} & 0 < e^{-x} < 1 \\ 0 & \textit{otherwise} \end{cases} \\ &= \begin{cases} e^{-x} & -\infty < -x < 0 \\ 0 & \textit{otherwise} \end{cases} \\ &= \begin{cases} e^{-x} & 0 < x < \infty \\ 0 & \textit{otherwise} \end{cases} \end{aligned}$$

Thus $Y \sim$ exponential(1).

Example 8.10.7. *Suppose X is uniform on $\left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$ and $Y = \tan X$. Find the density of Y and what known distribution is it?*

Solution. *Step1: Find the cdf of Y and write in terms of F_X*

$$\begin{aligned} F_Y(x) &= \mathbb{P}(\tan X \leq x) \\ &= \mathbb{P}\left(X \leq \tan^{-1} x\right) \\ &= F_X(\tan^{-1} x) \end{aligned}$$

Step2: Take a derivative and recall that since $\frac{1}{\frac{\pi}{2}+\frac{\pi}{2}} = \frac{1}{\pi}$ then

$$f_X(x) = \begin{cases} \frac{1}{\pi} & -\frac{\pi}{2} < x < \frac{\pi}{2} \\ 0 & \textit{otherwise.} \end{cases}$$

Thus

$$\begin{aligned} f_Y(x) &= F_Y'(x) \\ &= \frac{d}{dx} F_X(\tan^{-1} x) \\ &= F_X'\left(\tan^{-1} x\right)\left(\tan^{-1} x\right)' \\ &= F_X'\left(\tan^{-1} x\right) \frac{1}{1+x^2} \\ &= \begin{cases} \frac{1}{\pi}\frac{1}{1+x^2} & -\frac{\pi}{2} < \tan^{-1} x < \frac{\pi}{2} \\ 0 & \textit{otherwise.} \end{cases} \\ &= \begin{cases} \frac{1}{\pi}\frac{1}{1+x^2} & -\infty < x < \infty \\ 0 & \textit{otherwise.} \end{cases} \end{aligned}$$

Thus Y is $Cauchy(1,0)$, as we had already shown before.

Example 8.10.8. *The time, T, that a manufacturing system is out of operation has cumulative distribution function*

$$F(t) = \begin{cases} 1 - \left(\frac{2}{t}\right)^2 & ,t > 2 \\ 0 & otherwise \end{cases}.$$

The resulting cost to the company is $Y = T^2$. Let f_Y be the density function for Y. Determine $f_Y(y)$, for $y > 4$.

Solution. Step 1: *Find the cdf of Y is and*

$$\begin{aligned} F_Y(y) &= \mathbb{P}\left(T^2 \leq y\right) \\ &= \mathbb{P}\left(T \leq \sqrt{y}\right) \\ &= F\left(\sqrt{y}\right) \\ &= 1 - \frac{4}{y} \end{aligned}$$

for $y > 4$.

Step 2: *Take a derivative*

$$\begin{aligned} f_Y(y) &= F_Y'(y) \\ &= \frac{4}{y^2}. \end{aligned}$$

In order to derive a formula for a large class of functions of a random variable, we will use some basic facts from Calculus (Math 161) about increasing or decreasing functions. Here we will state them just for increasing. One thing to note, is that we've been using the following useful property:

Properties 8.10.1: Derivative of Monotone functions

The function $g : \mathbb{R} \to \mathbb{R}$ is strictly increasing increasing if $g(x) < g(y)$ whenever $x < y$ and it is strictly deceasing if $g(y) < g(x)$ whenever $x < y$. Such functions are called strictly monotone. Suppose $g : \mathbb{R} \to \mathbb{R}$ is a strictly monotone function. Then the inverse function $g^{-1}(x)$ and satisfies

$$g(g^{-1}(x)) = x$$

and

$$g^{-1}(g((x)) = x$$

Using the first identity and the chain rule we have

$$g'(g^{-1}(x))\left(g^{-1}\right)'(x) = 1 \tag{8.10.1}$$

from which it follows that

$$\left(g^{-1}\right)'(x) = \frac{1}{g'(g^{-1}(x))}$$

In words, in order to find the derivative of the inverse function or g at the point x, you find the derivate of the function g, evaluate it at the point $g^{-1}(x)$ and take the reciprocal.

Proposition 8.10.1: Distribution of g(X)

Suppose g is a strictly monotone function and X is a random variable with PDF function $f_X(x)$. Then $y = g(X)$ has a PDF given by

$$f_Y(y) = \begin{cases} \frac{f_X(g^{-1}(y)}{|g'(g^{-1}(y))|}, & \text{if y is in the range of g} \\ 0, & \text{if y is not in the range of g} \end{cases}$$

If y is in the range of g and we let we let x be such that $g(x) = y$, or $x = g^{-1}(y)$, we can write this as

$$f_Y(y) = \frac{f_X(x)}{|g'(x)|},\ x = g^{-1}(y).$$

Proof. Assume g is strictly increasing This follows from:

$$F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(g(X) \leq y) = \mathbb{P}(X \leq g^{-1}(y)) = F_X(g^{-1}(y))$$

Now apply the chain rule as above, (8.10.1). Note that if y is not in the range of g then $\mathbb{P}(g(X) \leq y) = 0$.

If g is strictly decreasing then,

$$F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(g(X) \leq y) = \mathbb{P}(g^{-1}(y) \leq X) = 1 - F_X(g^{-1}(y))$$

and again apply the chain rule. Observe, however, that in the case when the function is decreasing we seem to get minus what we got in the case when it is increasing. But recall that if g is increasing, the derivative is positive and if is decreasing, the derivative is negative. Hence the reason for the absolute value in the denominator. ■

Chapter 9

Weeks 10 and 11: Multivariate Distributions

9.1 Joint distribution functions

Up to this point our discussion has ben focused probabilities notions related to one random variable, its pmf, CDF, PDF, expectation, variance, etc. For many reasons, both theoretical and for applications, it is important to define these notions for random vectors, that is, random variables that map the probability sample space not just to $\mathbb{R}$ but to $\mathbb{R}^2$, or even $\mathbb{R}^n$, for any $n = 1, 2, \ldots$. For the most part, our discussion in this class will be limited to $n = 2$. However, this is no limitation. All the notions easily extend to any n once we develop all the tools for $n = 2$.

Definition 9.1.1: Joint cumulative probability distribution function

Let X and Y be two random variables defined on the same probability space (sample space). The function

$$F_{X,Y}(x, y) = \mathbb{P}(X \leq x, Y \leq y), \quad -\infty < x, y < \infty$$

is the joint distribution function of the random vector (X, Y). Let's make sure we understand the notation. By $-\infty < x, y < \infty$ we mean that x and y are any two points on the real line (no order). By $(X \leq x, Y \leq y)$, we mean the intersection of the sets $(X \leq x)$ and $(Y \leq y)$. Thus the comma "," will always mean in such notion "and".

Using the fact that $(Y < \infty) = S$ (the entire sample space) and the continuity property of probability, we see that

$$\begin{aligned}
F_X(x) &= \mathbb{P}(X \leq x) = P(X < x, Y < \infty) \\
&= \mathbb{P}(\lim_{y \to \infty} \{X \leq x, Y \leq y\}) \\
&= \lim_{y \to \infty} \mathbb{P}(X \leq x, Y \leq y) \\
&= F_{X,Y}(x, \infty).
\end{aligned}$$

In the same way,

$$F_Y = F_{X,Y}(\infty, y)$$

The functions F_X and F_Y are the “marginal” distribution functions of the (X, Y).

Let us recall that for a a single random variable we have formula that we use all the time

$$F_X(x) = 1 - \mathbb{P}(X > x).$$

or

$$\mathbb{P}(X > x) = 1 - F_X(x).$$

This formula comes from the general fact that $\mathbb{P}(A) = 1 - \mathbb{P}(A^c)$ for any event A in our sample space. However, now the function $F_{X,Y}$ is defined in terms of the intersection of two sets and in $A \cap B$ and when taking the complement we obtain $\mathbb{P}((A \cap B)^c) = \mathbb{P}(A^c \cup B^c)$ (by De morgan’s law). This gives that

$$\begin{aligned} P(X > x, Y > y) &= \mathbb{P}((X > x) \cap (Y > y)) \\ &= 1 - \mathbb{P}(((X > x) \cap (Y > y))^c)) \\ &= 1 - \mathbb{P}((X > x)^c \cup (Y > y)^c) \\ &= 1 - \mathbb{P}((X \leq x) \cup (Y \leq y)) \end{aligned}$$

With $A = (X > x)$ and $B = (Y > y)$, the “Inclusion-Exclusion” rule that $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$ gives that the last line of the previous string of inequalities is

$$\begin{aligned} \mathbb{P}(X > x, Y > y) &= 1 - [\mathbb{P}(X \leq x) + \mathbb{P}(Y \leq y) - \mathbb{P}(X \leq x, Y \leq y)] \\ &= 1 - F_X(x) - F_Y(y) + F_{X,Y}(x, y)] \end{aligned}$$

We summarize this

Properties 9.1.1: Properties of $F_{X,Y}$

$$\begin{aligned} \mathbb{P}(X > x, Y > y) &= 1 - [\mathbb{P}(X \leq x) + \mathbb{P}(Y \leq y) - \mathbb{P}(X \leq x, Y \leq y)] \\ &= 1 - F_X(x) - F_Y(y) + F_{X,Y}(x, y)] \end{aligned}$$

and in general For any $x_1 < x_2, y_1 < y_2$, we have

$$\begin{aligned} \mathbb{P}(x_1 < X \leq x_2,\ y_1 < Y \leq y_2) &= F_{X,Y}(x_2, y_2) + F_{X,Y}(x_1, y_1) \\ &- F_{X,Y}(x_1, y_2) - F_{X,Y}(x_2, y_1) \end{aligned}$$

9.2 Discrete random variables

The above formulas and definitions apply to both continuous and discrete random variables. In case of discrete random variables we also have:

Definition 9.2.1: Joint probability mass function

If we have two discrete random variables X and Y, we define the joint probability mass function by

$$p(x_i, y_j) = \mathbb{P}\left(X = x_i, Y = y_j\right).$$

(i) The joint probability mass function has the property that

$$0 \leq p(x_i, y_j) \leq 1, \text{ for all } x_i, y_j$$

and

(ii)

$$\sum_i \sum_j p(x_i, y_j) = 1$$

(iii)

$$p_X(x) = \mathbb{P}(X = x) = \sum_{y_j} p(x, y_j)$$

(iv)

$$p_Y(y) = \mathbb{P}(Y = y) = \sum_{x_i} p(x_i, y)$$

Example 9.2.1. *Suppose you roll two 3-sided dice. Mark one of the die red and the other blue. Let X be the largest value obtained on any of the two dice . Let Y be the sum of the two dice. Find (a) the joint pmf of X and Y and (b) $\mathbb{P}(X = 2 \mid Y = 4)$.*

Solution. *(a) First need to find the values of $X = 1, 2, 3$ and $Y = 2, 3, 4, 5, 6$. The table for possible outcomes of the red and blue die are given in the first column and first row, respectively, and the associated values (X, Y) are the other nine entries.*

Outcome of dice	**1**	**2**	**3**
1	$(X = 1, Y = 2) = (1, 2)$	$(2, 3)$	$(3, 4)$
2	$(2, 3)$	$(2, 4)$	$(3, 5)$
3	$(3, 4)$	$(3, 5)$	$(3, 6)$

Using this table we have that the p.m.f. is given by:

$X \backslash Y$	*2*	*3*	*4*	*5*	*6*	*Raw sum* $\mathbb{P}(X = i)$
1	$\mathbb{P}(X = 1, Y = 2) = \frac{1}{9}$	0	0	0	0	$\frac{1}{9}$
2	0	$\frac{2}{9}$	$\frac{1}{9}$	0	0	$\frac{3}{9}$
3	0	0	$\frac{2}{9}$	$\frac{2}{9}$	$\frac{1}{9}$	$\frac{5}{9}$
Column sum $\mathbb{P}(Y = j)$	$\frac{1}{9}$	$\frac{2}{9}$	$\frac{3}{9}$	$\frac{2}{9}$	$\frac{1}{9}$	*1*

For (b) we have that $\mathbb{P}(X = 2 \mid Y = 4) = \frac{1/9}{3/9} = \frac{1}{3}$.

Example 9.2.2. *Suppose that 3 balls are randomly selected from container of 3 red, 4 blue, and 5 white balls. Let R and B denote, respectively, the number of red and blue balls chosen. What is their joint probability mass function?*

Solution. *Let $p(i,j) = P(R = i, B = j)$, $i, j = 0, 1, 2, 3$. Then we have*

$$
\begin{aligned}
p(0,0) &= \frac{\binom{5}{3}}{\binom{12}{3}} = \frac{10}{220}, \quad p(0,1) = \frac{\binom{4}{1}\binom{5}{2}}{\binom{12}{3}} = \frac{40}{220} \\
p(0,2) &= \frac{\binom{4}{2}\binom{5}{1}}{\binom{12}{3}} = \frac{30}{220}, \quad p(0,2) = \frac{\binom{4}{3}}{\binom{12}{3}} = \frac{4}{220},
\end{aligned}
$$

In the same way, we can compute

$$
\begin{aligned}
p(1,0) &= \frac{30}{220}, \quad p(1,1) = \frac{60}{220}, \quad p(1,2) = \frac{18}{220} \\
p(2,0) &= \frac{15}{220}, \quad p(2,1) = \frac{12}{220} \\
p(3,0) &= \frac{1}{220}, \quad p(3,1) = p(3,2) = p(3,3) = 0
\end{aligned}
$$

$$\mathbb{P}(X = i, Y = j)$$

$\mathbf{R}\backslash\mathbf{B}$	$\mathbf{0}$	$\mathbf{1}$	$\mathbf{2}$	$\mathbf{3}$	*Raw sum* $\mathbb{P}(\mathbf{R} = i)$
$\mathbf{0}$	$\frac{10}{220}$	$\frac{40}{220}$	$\frac{30}{220}$	$\frac{4}{220}$	$\frac{84}{220}$
$\mathbf{1}$	$\frac{30}{220}$	$\frac{60}{220}$	$\frac{18}{220}$	*0*	$\frac{108}{220}$
$\mathbf{2}$	$\frac{15}{220}$	$\frac{12}{220}$	0	0	$\frac{27}{220}$
$\mathbf{3}$	$\frac{1}{220}$	0	0	0	$\frac{1}{220}$
Column sum $\mathbb{P}(\mathbf{B} = j)$	$\frac{56}{220}$	$\frac{112}{220}$	$\frac{48}{220}$	$\frac{4}{220}$	

9.3 Continuous random variables

The random variables X, Y are said to be jointly continuous if they have a joint probability density function $f_{X,Y}(x,y)$. That is, if there is a nonnegative function $f_{X,Y}(x,y)$ defined for all x, y such that for any a, b, c, d we have

$$P(a \le X \le b, c \le Y \le d) = \int_c^d \int_a^b f_{X,Y}(x,y)dxdy$$

or more generally, for any set D in the plane

$$\mathbb{P}((X,Y) \in D) = \iint_D f_{X,Y}(x,y)dxdy$$

In particular, for any sets $A \subset \mathbb{R}$ and $B \subset \mathbb{R}$ we have

$$P(X \in A, Y \in B) = \int_B \int_A f_{X,Y}(x,y)dxdy.$$

Applying this to $A = (-\infty, a]$ and $B = (-\infty, \infty)$ gives that

$$\begin{aligned} F_X(a) = \mathbb{P}(X \leq a) &= \mathbb{P}(X \in A, Y \in (-\infty, \infty)) \\ &= \int_{-\infty}^{a} \int_{-\infty}^{\infty} f_{X,Y}(x,y)dydx \\ &= \int_{-\infty}^{a} \left(\int_{-\infty}^{\infty} f_{X,Y}(x,y)dy \right) dx \end{aligned}$$

and we see that the PDF of X is given by

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

In the same way, the PDF for Y is

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

Also,

$$F_{X,Y}(a,b) = \mathbb{P}(X \leq a, Y \leq b) = \int_{-\infty}^{b} \int_{-\infty}^{a} f_{X,Y}(x,y)dxdy$$

and as before differentiating and using the fundamental theorem of calculus (twice) we get

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

Notice that since since $\mathbb{P}(X \in (-\infty, \infty), Y \in (-\infty, \infty) = 1$, we must have

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x,y)dxdy = 1$$

Properties 9.3.1: Summary

The jointly continuous random variables X, Y have a joint probability density function $f_{X,Y}(x,y)$ satisfying:

(a) For any sets $A \subset \mathbb{R}$ and $B \subset \mathbb{R}$ we have

$$P(X \in A, Y \in B) = \int_B \int_A f_{X,Y}(x,y)dxdy.$$

(b)

$$F_{X,Y}(a,b) = \mathbb{P}(X \leq a, Y \leq b) = \int_{-\infty}^{b}\int_{-\infty}^{a} f(x,y)dxdy$$

(c)

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

(d)

$$F_X(a) = \int_{-\infty}^{a}\left(\int_{-\infty}^{\infty} f_{X,Y}(x,y)dy\right)dx \text{ and } F_Y(a) = \int_{-\infty}^{a}\left(\int_{-\infty}^{\infty} f_{X,Y}(x,y)dx\right)dy$$

(e)

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy \quad \text{and} \quad f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dx$$

9.4 Examples

Example 9.4.1. *Suppose X, Y have the joint PDF*

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

(a) What is the vale of c?

(b) Find $\mathbb{P}(X < Y)$

(c) Find $\mathbb{P}(X > 2, Y < 5)$

(d) Find the marginal $f_X(x)$

(e) Find $\mathbb{E}X$.

Solution. *For (a) we must have*

(a)

$$\begin{aligned} 1 &= \int_0^\infty \int_0^\infty ce^{-x}e^{-2y}dxdy = c\int_0^\infty e^{-2y}\left[-e^{-x}\right]_{x=0}^{x=\infty} dy \\ &= c\int_0^\infty e^{-2y}dy = c\left[-\frac{1}{2}e^{-2y}\right]_0^\infty = c\frac{1}{2}. \end{aligned}$$

Thus $c = 2$.

(b) For (b), set $D = \{(x, y) : 0 < x < y, 0 < y < \infty\}$. *Then*

$$\begin{aligned}
\mathbb{P}(X < Y) &= \int\int_D f_{X,Y}(x, y)dxdy \\
&= \int_0^\infty \int_0^y 2e^{-x}e^{-2y}dxdy \\
&= 2\int_0^\infty e^{-2y}(\left(\int_0^y e^{-x}dx\right)dy \\
&= 2\int_0^\infty e^{-2y}[1 - e^{-y}]dy \\
&= 2\int_0^\infty (e^{-2y} - e^{-3y})dy \\
&= 2\left(-\frac{1}{2}e^{-2y}\Big|_0^\infty + \frac{1}{3}e^{-3x}\Big|_0^\infty\right) \\
&= \frac{1}{3}
\end{aligned}$$

(c)

$$\begin{aligned}
\mathbb{P}(X > 2, Y < 5) &= 2\int_2^\infty e^{-x}\int_0^5 e^{-2y}dydx \\
&= 2\left[-e^{-x}\Big|_2^\infty\right]\left[-\frac{1}{2}e^{-2y}\Big|_0^5\right] \\
&= e^{-2} - e^{-12}.
\end{aligned}$$

(d)

$$f_X(x) = 2e^{-x}\int_0^\infty e^{-2y}dy = 2e^{-x}\left[-\frac{1}{2}e^{-2y}\Big|_0^\infty\right] = e^{-x}$$

(e) $\mathbb{E}X = \int_0^\infty xe^{-x}dx = 1$

9.5 Uniform Distribution in the plane

Let us recall that the CDF of of a uniform random variable X *on the interval* $I = (a, b)$ *is given by*

$$F(x) = \begin{cases} \frac{x-a}{b-a}, & x \in (a, b) \\ 0, & \textit{otherwise} \end{cases}$$

In particular, for any subinterval $J = (c, d) \subset I = (a, b)$ $(a < c < d < b)$ *we have*

$$\mathbb{P}(X \in J) = \frac{(c - d)}{(b - a)} = \frac{\textit{Length(J)}}{\textit{Length(I)}}$$

Thus the probability of the random variable being on an interval J only depends on its length and not on its location. That is, is, if two subintervals J_1 and J_2 of I have the same length, $\mathbb{P}(X \in J_1) = \mathbb{P}(X \in J_2)$ regardless of where the intervals are located in I. For example, for a uniform on $(0, 1)$, $\mathbb{P}(X \in (0, \frac{1}{4})) = \mathbb{P}(X \in (\frac{3}{4}, 1))$. In the same way we have

Definition 9.5.1: Uniform Distribution in the plane

Fox a region D in the plane with finite area. The pair (X, Y) has the uniform the plane has uniform distribution on D if for any subregion A contained in D,

$$\mathbb{P}((X, Y) \in A) = \frac{\text{Area(A)}}{\text{Area(D)}}$$

and

$$\mathbb{P}((X, Y) \in D) = 1.$$

Thus,

$$f_{X,Y}(x, y) = \begin{cases} \frac{1}{\text{Area(D)}}, & (x, y) \in D \\ 0, & otherwise \end{cases} = \frac{1}{\text{Area(D)}} \mathbb{1}_D(x, y)$$

Example 9.5.1. *Consider D a circle centered at $(0, 0)$ and radius r. Let (X, Y) be a point in D uniformly distributed.*

(a) Find the marginal density functions for X and Y.

(b) Let $Z = \sqrt{X^2 + Y^2}$ be the distance from the point (X, Y) to the center of the circle. Find the CDF of Z.

(c) Find $\mathbb{E}Z$.

Solution. *(a) From formula on marginal functions we have and the density of the (X, Y) we have*

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y) dy = \frac{1}{\pi r^2} \int_{-\sqrt{r^2-x^2}}^{\sqrt{r^2-x^2}} dy = \frac{2}{\pi r^2} \sqrt{r^2 - x^2}$$

Similarly,

$$f_Y(y) = \frac{2}{\pi r^2} \sqrt{r^2 - y^2}$$

(b) For any $a \leq 0$, $F_Z(a) = 0$. For $0 \leq a \leq r$, let $D(a)$ be the circle centered at $(0, 0)$ and radius a. Then

$$F_Z(a) = \mathbb{P}(\sqrt{X^2 + Y^2} \leq a) = \mathbb{P}(X^2 + Y^2 \leq a^2) = \frac{\textit{Area(D(a))}}{\pi r^2} = \frac{a^2}{r^2}$$

For $a \geq r$, $F_Z(a) = 1$. That is, the CDF of Z is

$$F_Z(a) = \begin{cases} 0, & a \leq 0 \\ \frac{a^2}{r^2}, & 0 < a < r \\ 1, & 1 \leq a < \infty \end{cases}$$

and the PDF is

$$f_Z(a) = \begin{cases} 0, & a \leq 0 \\ \frac{2a}{r^2}, & 0 < a < r \\ 0, & 1 \leq a < \infty \end{cases}$$

(c)

$$\mathbb{E}Z = \frac{2}{r^2}\int_0^r a^2 da = \frac{2r}{3}$$

Example 9.5.2. *Suppose the joint PDF of (X, Y) is given by*

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

What is the PDF of X/Y

Solution. *The CDF is 0 for $a \leq 0$. For $a > 0$ we have*

$$\begin{aligned} F_{X/Y}(a) &= \mathbb{P}(X/Y \leq a) = \int_0^\infty e^{-y}\left(\int_0^{ay} e^{-x}dx\right)dy \\ &= \int_0^\infty (1 - e^{-ay})e^{-y}dy \\ &= (-e^{-y} + \frac{e^{-(1+a)y}}{1+a})\Big|_0^\infty = 1 - \frac{1}{1+a}. \end{aligned}$$

Differentiating this in a gives

$$f_{X/Y}(a) = \frac{1}{(1+a)^2}.$$

Problem 9.5.1. *What is $\mathbb{E}(X/Y)$?*

9.6 The Multinomial Distribution

The joint CDF distribution function for a vector $(X_1, X_2, \cdots X_n)$ of random variables, where $n = 2, 3, 4 \ldots$ is any integer, can be defined in the same way, both for discrete r.v. and for continuous r.v. In the discrete setting we have

$$F_{X,X_2,\ldots X_n}(a_1, a_2, \ldots a_n) = \mathbb{P}(X_1 \leq a_1, X_2 \leq a_2, \ldots, X_n \leq a_n)$$

In the continuous case we say that a the nonnegative function $f(x_1, x_2, \ldots, x_2)$ is the joint PDF of the vector $(X_1, X_2, \cdots X_n)$ if for any sets $A_1, A_2, \ldots, A_n$, with $A_i \in \mathbb{R}$ we have

$$\mathbb{P}(X_1 \in A_1, X_2 \in A_2, \ldots, X_n \in A_n) = \int_{A_n}\int_{A_{n-1}} \cdots \int_{A_1} f(x_1, x_2, \ldots, x_2)dx_1dx_2\cdots dx_n$$

In particular, with $A_i = (-\infty, a_i]$ we this gives

$$\begin{aligned} F_{X,X_2,\ldots X_n}(a_1, a_2, \ldots a_n) &= \mathbb{P}(X_1 \leq a_1, X_2 \leq a_2, \ldots, X_n \leq a_n) \\ &= \int_{-\infty}^{a_n}\int_{-\infty}^{a_{n-1}} \cdots \int_{-\infty}^{a_1} f(x_1, x_2, \ldots, x_2)dx_1dx_2\cdots dx_n \end{aligned}$$

An important example of a multidimensional distribution is the *multinomial distribution.* which we have already introduced during week 1. Suppose we have when a sequence of n independent and identical experiments resulting on r outmodes.

Definition 9.6.1: The multinomial distribution

If we let X_k denote the number of the experiments resulting on outcome number k. Suppose we have $p_1, p_2, \ldots, p_n$ with $\sum_{k=1}^{r} p_k = 1$. If $\sum_{k=1}^{r} n_k = n$, then

$$\mathbb{P}(X_1 = n_1, X_2 = n_2, \ldots, X_r = n_r) = \frac{n!}{n_1!\, n_2! \cdots n_r!}\, p_1^{n_1}\, p_2^{n_2} \cdots p_r^{n_r}$$

Notice that if we have $0 < p < 1$ and let $p_1 = p$ and $p_n = (1-p)$, $n_1 = k$ and $n_n = (n-k)$ we get back the Binomial distribution with k–successes and $(n-1)$–failures:

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

9.7 Independence

During the lectures for week 6 on discrete random variables and later for continuous random variables we have already what it means for two random variables to be independent. In this section we want to derive a few of the important properties that follow from independence, especially as they relate to joint probability functions for continuous random variables. We combine the definitions.

9.7.1 Definitions and basic properties

Definition 9.7.1: Independent random variables

Two random discrete random variables X and Y are said to be independent if for all sets A and B in $\mathbb{R}$ we have

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

That is, the random variables X and Y are independent if whenever the joint distribution is the product of their distributions. With $A = (\infty, a]$ and $B = (-\infty, b]$, we have

$$F_{X,Y}(a, b) = F_X(a) F_Y(b)$$

We list the following properties that follow from the definition (as already listed in Lecture 6). The first, (9.7.1) follows by taking $A = \{x\}$ and $B = \{y\}$ and the second, (9.7.2), by taking partial derivatives of the CDF.

Properties 9.7.1: Joint pmf and PDF

- Two discrete r.v.'s X and Y are independent iff their joint probability mass function satisfies:

$$p_{X,Y}(x,y) = p_X(x)p_Y(y) \tag{9.7.1}$$

- Two jointly continuous r.v.'s X and Y are independent iff their joint probability distribution function satisfies:

$$f_{X,Y}(x,y) = f_X(x)f_Y(y) \tag{9.7.2}$$

9.8 Sums of Independent random variables

Recall 9.8.1: Sums of Poissons is a Poisson

In Theorem 8.1 for week 6 we showed that if

$$S_n = X_1 + X_2 + \cdots + X_n,$$

where are independent with $X_j \sim \text{Pois}(\lambda_j)$ for all $j = 1, 2, \ldots, n$. Then $S_n \sim$Pois(λ) where

$$\lambda = \lambda_1 + \lambda_2 + \cdots + \lambda_n$$

The proof was the observation that for any two discrete r.v.'s X_1 and X_2 (whatever they are),

$$\begin{aligned}\mathbb{P}(X_1 + X_2 = k) &= \sum_{j=0}^{k} \mathbb{P}(X_1 = j,\ X_2 = k - j) \\ &= \sum_{j=0}^{k} \mathbb{P}(X_1 = j)\mathbb{P}(X_2 = k - j)\end{aligned}$$

together with the explicit expression for the pmf of the Poisson distribution.

This formula holds for any two independent discrete r.v. and is useful to record it for future uses.

Properties 9.8.1: The pmf of sum of discrete r.v.

Let X and Y be two independent random variables with pmf p_X and p_Y, respectively. Then

$$p_{X+Y}(k) = \sum_{j=0}^{k} p_X(j)p_Y(k - j)$$

This is called the discrete convolution of the two probability mass functions p_X and p_Y and is written as $p_X * p_Y$. That is, $p_{X+Y}(k) = p_X * p_Y(k)$ where $p_X * p_Y$ is the right hand side as above. The fact that $X + Y = Y + X$ means that $p_X * p_Y = p_Y * p_X$.

For continuous random variables we have a similar expression with sums replaced by integrals.

Properties 9.8.2: CDF and PDF of sum of continuous independent r.v.

If X and Y are independent with joint PDF $f_X(x)f_Y(y)$, we see that

$$\begin{aligned} F_{X+Y}(a) = \mathbb{P}(X+Y \leq a) &= \iint_{\{X+Y\leq a\}} f_X(x)f_Y(y)dxdy \\ &= \int_{-\infty}^{\infty}\left(\int_{-\infty}^{a-y} f_X(x)dx\right) f_Y(y)dy \\ &= \int_{-\infty}^{\infty} F_X(a-y)f_Y(y)dy \end{aligned} \tag{9.8.1}$$

Differentiating gives

$$f_{X+Y}(a) = \int_{-\infty}^{\infty} \frac{d}{d\,a} F_X(a-y)f_Y(y)dy = \int_{-\infty}^{\infty} f_X(a-y)f_Y(y)dy = f_X * f_Y(a). \tag{9.8.2}$$

That is, the PDF of two independent random variables is the "convolution" of f_X with f_Y, denoted by $f_X * f_Y$. Bi either changing variables in the the integral or simply noting that $X+Y = Y+X$, we see that $f_X * f_Y{=}f_Y * f_X$. So, if you forget the order, it makes no difference.

9.9 Sums of uniform random variables

9.9.1 Examples

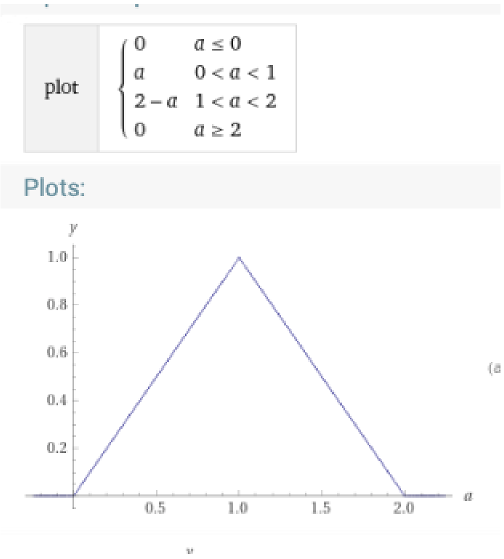
Example 9.9.1. *Suppose X and Y are independent uniform r.v.'s in $(0,1)$. Find f_{X+Y}.*

Solution. *From*

$$f_X(a) = f_Y(a) = \begin{cases} 1, & 0 < a < 1 \\ 0, & otherwise \end{cases}$$

we obtain

$$\begin{aligned} f_{X+Y}(a) = &= \int_0^1 f_X(a-y)dy \\ &= \int_{a-1}^{a} f(u)du = \begin{cases} 0, & a \leq 0 \\ a, & 0 \leq a \leq 1 \\ 2-a, & 1 \leq a \leq 2 \\ 0, & a \geq 2 \end{cases} \end{aligned}$$



plot $\begin{cases} 0 & a \le 0 \\ a & 0 < a < 1 \\ 2-a & 1 < a < 2 \\ 0 & a \ge 2 \end{cases}$

Example 9.9.2. *Let X and Y be independent uniformly distributed random variables on $I = (0, a)$ and $J = (0, b)$, respectively. Then (X, Y) is uniformly distributed on the rectangle $I \times J = \{(x, y) : 0 < x < a, 0 < y < b\}$*

Solution. *For any set $A \subset I$ and any $B \subset J$*

$$\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B) = \frac{\textit{Length(A)}}{a}\frac{\textit{Length(B)}}{b} = \frac{\textit{Area}(A \times B)}{\textit{Area}(I \times J)}$$

Example 9.9.3. *Suppose X and Y are uniform in $(0, 1)$. Find*

(i) $\mathbb{P}(X^2 + Y^2 \le 1)$

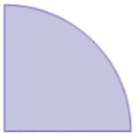
(ii) $\mathbb{P}(|X - Y| \le \frac{1}{2})$

(iii) $\mathbb{P}(Y \le X^2)$

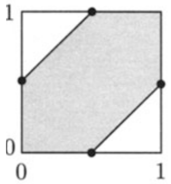
Solution. *Notice that since the area X and Y are uniform on $(0, 1)$ $X + Y$ is uniform in the square $Q = \{(x, y) : 0 < x < 1, 0 < y < 1\}$ and this has area 1. To compute the probabilities we only ned to find the are of the regions $R_1 = \{(x, y) \in Q : x^2 + y^2 \le 1\}$, $R_2 = \{(x, y) : |x - y| \le \frac{1}{2}\}$ and $R_3 = \{(x, y) \in Q : y \le x^2\}$*

(i) *The picture below gives the region R_1. It is a quarter of the circle and hence it has area $\frac{\pi}{4}$. Thus*

$$\mathbb{P}(X^2 + Y^2 \le 1) = \frac{\pi}{4}$$



The region R_2 is given by



Notice that when $x = 0$, $y = \frac{1}{2}$ and when $y = 0$, $x = \frac{1}{2}$. Thus each of the unshaded triangle has area ($\frac{1}{2}b \times h$) $\frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{8}$ and together their area is $\frac{1}{4}$. This gives

(ii)
$$\mathbb{P}(|X - Y| \leq \frac{1}{2}) = 1 - \frac{1}{4} = \frac{3}{4}$$

R_3 represents the region below the parabola $y = x^2$ from $x = 0$ to $x = 1$. Thus

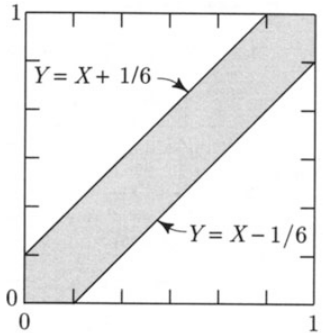
$$Area(R_3) = \int_0^1 x^2 dx = \frac{1}{3}$$

and we have

(iii)
$$\mathbb{P}(Y \leq X^2) = \frac{1}{3}$$

Example 9.9.4. *Two people try to meet at a certain place between 5:00 pm and 5:30 pm. Suppose that each person arrives at a time distributed uniformly at random in this time interval, independent of the other, and waits for the other at most 5 minutes. What is the probability that they meet?*

Solution. *There are 6 5-minute intervals in 30 minutes. Let X and Y be the arrival times measured as fractions of the 30 minute interval, staring from 5:00 P.M. Then X and Y are independent uniform $(0,1)$ random variables. The two people meet if and only if $|X - Y| \leq \frac{1}{6}$*



The are of each triangle is $\frac{1}{2}(\frac{5}{6})^2$. This gives

$$\mathbb{P}(|X - Y| \leq \frac{1}{6}) = 1 - \left(\frac{5}{6}\right)^2 = \frac{11}{36}$$

We can also compute this probability (area) using integrals as follows: (Explain it to yourself!)

$$\mathbb{P}(|X-Y| \leq \frac{1}{6}) = 2\int_0^{\frac{1}{6}} (x+\frac{1}{6})dx + \int_{\frac{1}{6}}^{\frac{5}{6}} \left(\int_{x-\frac{1}{6}}^{x+\frac{1}{6}} dy \right) dx = \frac{3}{36} + \frac{8}{36} = \frac{11}{36}.$$

9.10 Sums of gamma random variables

Recall that X ∼Gamma(α, λ) if it has the PDF given by

$$f(x) = \begin{cases} \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

where $\Gamma(\alpha)$ is the Gamma function

$$\Gamma(\alpha) = \int_0^\infty e^{-y} y^{\alpha-1} dy.$$

Proposition 9.10.1: Sums of gammas is gamma

Let X_i, $i = 1, 2, \ldots n$ be independent random variables with X_i ∼Gamma(α_i, λ). Then $\sum_{i=1}^n X_i \sim$ Gamma(α, λ), where $\alpha = \sum_{i=1}^n \alpha_i$

Proof. It is enough to prove this for two random variables X_1 and X_2. Once we do this we reapply the argument to $Z = X_1 + X_2$ and $Y = X_3$ to get it for three random variables and continue. From the convolution formula we have

$$\begin{aligned} f_{X_1+X_2}(a) &= \frac{\lambda^{\alpha_1}\lambda^{\alpha_2}}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \int_0^a e^{-\lambda(a-y)}(a-y)^{\alpha_1-1} e^{-\lambda y} y^{\alpha_2-1} dy \\ &= e^{-\lambda a} \frac{\lambda^{\alpha_1+\alpha_2}}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \int_0^a (a-y)^{\alpha_1-1} y^{\alpha_2-1} dy, \ \text{(letting, } u = \frac{y}{a}) \\ &= a^{\alpha_1+\alpha_2} e^{-\lambda a} \frac{\lambda^{\alpha_1+\alpha_2}}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \int_0^1 (1-u)^{\alpha_1-1} u^{\alpha_2-1} du \\ &= a^{\alpha_1+\alpha_2} e^{-\lambda a} \frac{\lambda^{\alpha_1+\alpha_2}}{\Gamma(\alpha_1)\Gamma(\alpha_2)} Beta(\alpha_1, \alpha_2) \\ &= a^{\alpha_1+\alpha_2} e^{-\lambda a} \frac{\lambda^{\alpha_1+\alpha_2}}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \frac{\Gamma(\alpha_1)\Gamma(\alpha_2)}{\Gamma(\alpha_1+\alpha_2)}, \ \text{(Formula for Beta(a, b) in terms Gamma)} \\ &= \frac{\lambda^{\alpha_1+\alpha_2}}{\Gamma(\alpha_1+\alpha_2)} a^{\alpha_1+\alpha_2} e^{-\lambda a} \end{aligned}$$

■

Example 9.10.1. *In our discussion of the gamma distribution in the lecture notes for week 8 we mentioned that it arises as a sum of exponential random variables. We now verify this fact. Let us recall that an $X \sim$ Exp(λ) is in equivalent to $X \sim$ Gamma(1, λ). From our proposition above it follows that if $X_1, X_2, \ldots, X_n$ are Exp(λ) and independent then*

$$X = X_1 + X_2 \cdots + X_n$$

is Gamma(n, λ). Thus, as stated earlier, the gamma random variables can be thought as the time until an nth event occurs.

Example 9.10.2. *Let $Z_1, Z_2, \ldots, Z_n$ be independent standard normal. What is the distribution of $\sum_{j=1}^{n} Z_j^2$?*

Solution. *Let us take the case of a single random variable, call it Z.*

$$F_{Z^2}(a) = \mathbb{P}(Z^2 \leq a) = \mathbb{P}(-\sqrt{a} \leq Z \leq \sqrt{a}) = F_Z(\sqrt{a}) - F_Z(-\sqrt{a})$$

Differentiating this gives

$$f_{Z^2}(a) = \frac{1}{2\sqrt{a}}[\big(f_Z(\sqrt{a}) + f_Z(-\sqrt{a})\big) = \frac{\frac{1}{2}^{\frac{1}{2}}}{\Gamma(\frac{1}{2})} x^{-\frac{1}{2}} e^{-\frac{x}{2}}$$

which is Gamma($\frac{1}{2}, \frac{1}{2}$). Thus from our Proposition we have that

$$X = \sum_{j=1}^{n} X_k \sim \textit{Gamma}(\frac{n}{2}, \frac{1}{2})$$

which is a Chi-squared with n degrees of freedom.
Note: In the above computation we use the fact that $\Gamma(1/2) = \sqrt{\pi}$.

9.11 Sums of normal random variables

Properties 9.11.1: Sums of Normal is Normal

Suppose $X_i \sim$ are normal with parameters (μ_i, σ_i^2), $i = 1, 2, \ldots n$, respectively. Then $\sum_{i=1}^{n} X_i$ is normal with parameter $\mu = \sum_{i=1}^{n} \mu_i$ and $\sigma^2 = \sum_{i=1}^{n}$.

Proof. As in the previous situations, we can assume we Normal random variables X and Y. We may also assume (as we shall see shortly) that we can take $X \sim N(0, \sigma^2)$ and $Y = Z \sim N(0, 1)$, the standard normal. Set

$$c = \frac{1 + \sigma^2}{2\sigma^2}$$

and

$$\sqrt{c}\,\sigma = \sqrt{1 + \sigma^2}.$$

Also note that

$$\begin{aligned} c\left(y^2 - 2y\frac{a}{1 + \sigma^2}\right) &= c\left\{\left(y - \frac{a}{1 + \sigma^2}\right)^2 - \frac{a^2}{(1 + \sigma^2)^2}\right\} \\ &= c\left(y - \frac{a}{1 + \sigma^2}\right)^2 - \frac{a^2}{2\sigma^2(1 + \sigma^2)} \end{aligned}$$

and write

$$\begin{aligned} f_X(a-y)f_Y(f) &= \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(a-y)^2}{2\sigma^2}}\frac{1}{\sqrt{2\pi}}e^{-\frac{-y^2}{2}} \\ &= \frac{1}{2\pi\sigma}e^{-\frac{a^2}{2\sigma^2}}\, e^{-c\left(y^2-2y\frac{a}{1+\sigma^2}\right)} \end{aligned}$$

Recalling the convolution formula for the sum of counties random variables we get

$$\begin{aligned} f_{X+Y}(a) &= \frac{1}{2\pi\sigma}e^{-\frac{a^2}{2\sigma^2}}\, e^{\left(\frac{a^2}{2\sigma^2(1+\sigma^2)}\right)}\int_{-\infty}^{\infty} e^{-c\left(y-\frac{a}{1+\sigma^2}\right)^2}dy, \quad \left(let\ u = y-\frac{a}{1+\sigma^2}\right) \\ &= \frac{1}{2\pi\sigma}\,e^{-\frac{a^2}{2(1+\sigma^2)}}\int_{-\infty}^{\infty} e^{-(\sqrt{c}\,u)^2}du,, \quad (let\ v = \sqrt{c}\,u) \\ &= \frac{1}{\sqrt{c}}\frac{1}{2\pi\sigma}\,e^{-\frac{a^2}{2(1+\sigma^2)}}\int_{-\infty}^{\infty} e^{-v^2}dv \\ &= \sqrt{2\pi}\frac{1}{\sqrt{c}}\frac{1}{2\pi\sigma}\,e^{-\frac{a^2}{2(1+\sigma^2)}} = \frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{1+\sigma^2}}e^{-\frac{a^2}{2(1+\sigma^2)}} \end{aligned}$$

So, $X+Y \sim N(0, 1+\sigma^2)$.

If $X \sim N(\mu_1, \sigma_1)$ and $Y \sim N(\mu_2, \sigma_2)$, then $\frac{X-\mu_1}{\sigma_2} \sim N(0, \frac{\sigma_1^2}{\sigma_2^2})$ and $\frac{Y-\mu_2}{\sigma_2} \sim N(0,1)$. Thus

$$\frac{X-\mu_1}{\sigma_2} + \frac{Y-\mu_2}{\sigma_2} \sim N(0, \frac{\sigma_1^2}{\sigma_2^2})$$

But

$$\begin{aligned} X+Y &= \sigma_2\left(\frac{X-\mu_1}{\sigma_2}+\frac{Y-\mu_2}{\sigma_2}\right)+\mu_1+\mu_2 \sim N(\mu_1+\mu_2, \sigma^2(1+\frac{\sigma_1^2}{\sigma_2^2}) \\ &= \sim N(\mu_1+\mu_2, \sigma_1+\sigma_1^2) \end{aligned}$$

■

Example 9.11.1. *A soccer team will plays a 44-game season. 26 of these games are against class A teams and 18 are against class B teams. Suppose that the team will win each game against a class A team with probability* 0.4 *and will win each game against a class B team with probability* 0.7*. Suppose also that the results of the different games are independent. What is the (approximate) probability that the team wins* 25 *games or more?*

Solution. *Let X_A and X_B be the number of games the team wins against class A and against class B teams. Note that X_A and X_B are independent binomial random variables with*

$$\begin{aligned} &\mathbb{E}X_A = 26(0.4) = 10.4, \quad \mathit{Var}(X_A) = 26(0.4)(.6) = 6.24 \\ &\mathbb{E}X_b = 18(.7) = 12.6 \quad \mathit{Var}(Y_A) = 18(0.7)(.3) = 3.78 \end{aligned}$$

By the normal approximation to the binomial X_A can be approximated by a $N(10.4, 6.24)$ and X_B by a normal $N(12.6, 3.78)$ and their sum $X_A + X_B$ by a Normal $N(23, 10.02)$. Thus

$$\begin{aligned}
\mathbb{P}(X_A + X_B \geq 25) &= \mathbb{P}(X_A + X_B \geq 24.5) \\
&= \mathbb{P}\left(\frac{X_A + X_B - 23}{\sqrt{10.02}} > \frac{24.5 - 23}{\sqrt{10.02}}\right) \\
&\approx \mathbb{P}(Z \geq \frac{1.5}{\sqrt{10.02}}) \\
&= 1 - \Phi(\frac{1.5}{\sqrt{10.02}}) \\
&\approx 0.3178
\end{aligned}$$

9.12 Distributions of maximum and minimum

Remark 9.12.1. *The definition of independence extends to any number of random and we have that the r.v's $X_1, X_2, \ldots, X_n$ are independent iff*

$$F_{X,X_2,\ldots X_n}(a_1, a_2, \ldots a_n) = F_{X_1}(a_1)F_{X_2}(a_2) \cdots F_{X_n}(a_n).$$

from which it follows by taking partial derivatives that in the counties case

$$f_{X,X_2,\ldots X_n}(x_1, x_2, \ldots x_n) = f_{X_1}(x_1)f_{X_2}(x_2) \cdots f_{X_n}(x_n).$$

Let $X_1, X_2, \ldots, X_n$ be independent random variables with CDF F_i, $i = 1, 2, \ldots, n$, respectively. Define

$$X_{max} = \max(X_1, X_2, \ldots, X_n)$$

This means that X_{max} is less than or equal to x if and only if all the $X_i's$ are less than or equal to x. Define

$$X_{min} = \min(X_1, X_2, \ldots, X_n)$$

This means that X_{min} is greater than x if and only if all the $X_i's$ are greater than x. From these we have

$$\begin{aligned}
F_{X_{max}}(a) &= \mathbb{P}(X_{max} \leq a) \\
&= \mathbb{P}(X_1 \leq a, X_2 \leq a, \ldots, X_n \leq a) \\
&= \mathbb{P}(X_1 \leq a)\mathbb{P}(X_2 \leq a) \cdots \mathbb{P}(X_n \leq a) \\
&= F_1(a)F_2(a) \cdots F_n(a)
\end{aligned}$$

Similarly,

$$\begin{aligned}
F_{X_{min}}(a) &= \mathbb{P}(X_{min} \leq a) = 1 - P(X_{min} > a) \\
&= 1 - \mathbb{P}(X_1 > a, X_2 > a, \ldots, X_n > a) \\
&= 1 - [\mathbb{P}(X_1 > a)\mathbb{P}(X_2 > a) \cdots \mathbb{P}(X_n > a)] \\
&= 1 - (1 - F_1(a))(1 - F_2(a)) \cdots (1 - F_n(a))
\end{aligned}$$

We summarize in the following

Theorem 9.12.1: Max and Min of independent r.v.

Let $X_1, X_2, \dots, X_n$ be independent random variables with CDF F_i, $i = 1, 2, \dots, n$, respectively. Then the CDF of X_{max} and X_{min} are given by

$$\begin{aligned} F_{X_{max}}(a) &= \mathbb{P}(X_{max} \le a) \\ &= \mathbb{P}(X_1 \le a, X_2 \le a, \dots, X_n \le a) \\ &= \mathbb{P}(X_1 \le a)\mathbb{P}(X_2 \le a) \cdots \mathbb{P}(X_n \le a) \\ &= F_1(a)F_2(a) \cdots F_n(a) \end{aligned}$$

$$\begin{aligned} F_{X_{min}}(a) &= \mathbb{P}(X_{min} \le a) = 1 - P(X_{min} > a) \\ &= 1 - \mathbb{P}(X_1 > a, X_2 > a, \dots, X_n > a) \\ &= 1 - [\mathbb{P}(X_1 > a)\mathbb{P}(X_2 > a) \cdots \mathbb{P}(X_n > a)] \\ &= 1 - (1 - F_1(a))(1 - F_2(a)) \cdots (1 - F_n(a)) \end{aligned}$$

If in addition, the random variables are continuous and have the same distribution, that is, they are independent, identically distributed with command CDF F and command PDF function f, we have:

$$F_{X_{max}}(a) = F(a)^n \tag{9.12.1}$$

and

$$F_{X_{min}}(a) = 1 - (1 - F(a))^n. \tag{9.12.2}$$

Differentiating these we get the PDF's

$$f_{X_{max}}(a) = nF(a)^{n-1}f(a)$$

$$f_{X_{min}}(a) = n(1 - F(a))^{n-1}f(a).$$

Example 9.12.1. *Let $X_1, X_2, \dots, X_n$ be independent exponential with parameter λ_i. Find the distribution of X_{min}.*

Solution. *For each $i = 1, 2, \dots, n$ we have*

$$F_i(a) = \begin{cases} 0, & a < 0 \\ 1 - e^{-\lambda_i a}, & a \ge 0 \end{cases}$$

From this and theorem we see that $F_{X_{min}}(a) = 0$ for $a < 0$ and otherwise

$$F_{X_{min}}(a) = 1 - e^{-(\lambda_1 + \lambda_2 + \cdots + \lambda_n)a}.$$

Thus the X_{min} is Exp($\lambda_1 + \lambda_2 + \cdots + \lambda_n$)

9.12.1: Remark

This example was already a problem in your homework.

9.13 Distributions of order statistics

Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed random variables with command CDF F (discrete or continuous). We define $X_{(1)} =$ smallest of $(X_1, X_2, \ldots X_n$. That with this notation

$$X_{(1)} = X_{min}$$

We define

$$\begin{aligned}
X_{(2)} &= \text{second smallest of } (X_1, X_2, \ldots X_n), \\
X_{(3)} &= \text{3rd smallest of } (X_1, X_2, \ldots X_n), \\
&\vdots \\
X_{(j)} &= \text{jth smallest of } (X_1, X_2, \ldots X_n) \;\text{ in general}
\end{aligned}$$

With this notation,

$$X_{(n)} = X_{max}$$

and we have

$$X_{(1)} \leq X_{(2)} \leq X_{(3)} \cdots \leq X_{(n)}$$

The random variables $X_{(j)}$ are called the *order statistics of* $X_1, X_2, \ldots, X_n$. Now assume the random-variables are continuous. Above we computed the CDF of $X_{(1)}$ and $X_{(n)}$ and also their PDF. To compute the general case of $X_{(j)}$ we observe that in n order for $X_{(j)}$ to equal a, it is necessary for $j-1$ of the n values of $X_1, X_2, \ldots, X_n$ to be less than a and $n-j$ of them to be greater than a and 1 of them to equal a. The probability density that any given set of $j-1$ of the $X_j's$ are less than a, another given set of $n-j$ are all greater than a, and the remaining value is equal to a is given by

$$F(a)^{j-1}[1-F(a)]^{n-j}f(a).$$

and since there are

$$\binom{n}{j-1, n-j, 1} = \frac{n!}{(n-j)!\,(j-1)!}$$

different partitions of the n random variables into the three groups we get that the PDF for $X_{(j)}$ is given by

$$f_{(j)}(a) = \frac{n!}{(n-j)!\,(j-1)!}F(a)^{j-1}[1-F(a)]^{n-j}f(a).$$

And the CDF

$$F_{(j)}(a) = \frac{n!}{(n-j)!\,(j-1)!}\int_{-\infty}^{a} F(y)^{j-1}[1-F(y)]^{n-j}f(y)dy. \tag{9.13.1}$$

Note that for $j = 1, n$, these agrees with

$$f_{X_{max}}(a) = nF(a)^{n-1}f(a)$$

$$f_{X_{min}}(a) = n(1-F(a))^{n-1}f(a).$$

as computed above.

Problem 9.13.1. *For $j = 1$ and $j = n$ the equation (9.13.2) seems to be different than that in (9.12.1) and (9.12.2). Show that in fact they are the same.*

We summarize:

Properties 9.13.1: CDF and PDF of order statistics

Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed with command CDF F and command PDF f. Let

$$X_{(1)} \leq X_{(2)} \leq X_{(3)} \cdots \leq X_{(n)}$$

be their order statistics. Then for every $j = 1, 2, \ldots, n$, the PDF of $X_{(j)}$ is

$$f_{(j)}(a) = \frac{n!}{(n-j)!\,(j-1)!} F(a)^{j-1}[1 - F(a)]^{n-j} f(a).$$

and its CDF is

$$F_{(j)}(a) = \frac{n!}{(n-j)!\,(j-1)!} \int_{-\infty}^{a} F(y)^{j-1}[1 - F(y)]^{n-j} f(y) dy. \tag{9.13.2}$$

It is also possible to compute the joint PDF of the order statistics when the random variables $X_1, X_2, \ldots, X_n$ are independent and have the same distribution. In this case we know that

$$f_{X_1, X_2, \ldots, X_n}(x_1, x_2, \ldots, x_n) = f_{X_1}(x_1) f_{X_2}(x_2) \cdots f_{X_n}(x_n)$$

Of course, the sequence consisting of the order statistics $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ is not independent. However, from the fact that there are $n!$ equally likely ways that $X_1, X_2, \ldots, X_n$ can be placed in order, one can show that

$$f_{X_{(1)}, X_{(2)}, \ldots, X_{(n)}}(x_1, x_2, \ldots, x_n) = n! f_{X_1}(x_1) f_{X_2}(x_2) \cdots f_{X_n}(x_n)$$

whenever $x_1 < x_2 < \cdots < x_n$ and zero otherwise

Properties 9.13.2: The joint PDF of order statistics:

Suppose $X_1, X_2, \ldots, X_n$ are independent and have the same distribution. Then

$$\begin{cases} f_{X_{(1)}, X_{(2)}, \ldots, X_{(n)}}(x_1, x_2, \ldots, x_n) = n! f_{X_1}(x_1) f_{X_2}(x_2) \cdots f_{X_n}(x_n), & x_1 < x_2 < \cdots < x_n \\ 0, & \textit{otherwise} \end{cases}$$

Example 9.13.1. *When a sample of $2m + 1$ random variables is observed, the $(m + 1)$ order statistic is called the sample median. If a sample of size 3 from a uniform distribution on $(0, 1)$ is observed, find the probability that the sample median is between $\frac{1}{4}$ and $\frac{3}{4}$*

Solution. *From the formula above we have that*

$$f_{(2)}(a) = \frac{3}{1!1!1!} a(1 - 1), \ 0 < a < 1$$

and zero else. Thus

$$\mathbb{P}\left(\frac{1}{4} < X_{(2)} < \frac{3}{4}\right) = 6 \int_{\frac{1}{4}}^{\frac{3}{4}} a(1 - a) da = \frac{11}{16}$$

9.14 Conditional random variables and expectation

Definition 9.14.1: Conditional random variables and expectations

Let X and Y be two random variables.

(i) Suppose X and Y are discrete random variables with pmf p_X, p_Y, respectively and joint pmf $p_{X,Y}$. We define the new random variable $X|Y$ with pmf

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

(ii)

(iii) Suppose X and Y are two continuous r.v.'s with PDF f_X, f_Y, respectively and joint PDF $f_{X,Y}$. We define similarly the random variable X given Y, $X|Y$, with PDF

$$f_{X|Y=y}(x,y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$$

Note that if X and Y are independent then in both the continuous and discrete cases, $X|Y$ are just X exactly as with events.

9.14.1 examples

Example 9.14.1. *Suppose the joint pmf of (X,Y) is*

$x\backslash y$	*0*	*1*
0	0.4	0.2
1	0.1	0.3

Observe that the the second column is

$$p_{X|Y}(0 \mid 1) = \frac{0.2}{0.5} = \frac{2}{5} \text{ and } p_{X|Y}(1 \mid 1) = \frac{0.3}{0.5} = \frac{3}{5}.$$

Are X and Y independent? Note that $p_X(0) = 0.4 + 0.2 = 0.6 \neq p_{X|Y}(0 \mid 1)$ an hence they are not.

Example 9.14.2. *(Same without the table) Suppose that $p(x,y)$, the joint probability mass function of X and Y is given by $p(0,0) = 0.4$, $p(0,1) = 0.2$, $p(1,0) = 0.1$ and $p(1,1) = 0.3$. Compute the conditional probability mass function of X given $Y = 1$.*

Solution.

$$p_Y(1) = \sum_x p(x,1) = p(0,1) + p(1,1) = 0.5$$

$$p_{X|Y}(0 \mid 1) = \frac{p(0,1)}{p_Y(1)} = \frac{2}{5}$$

and

$$p_{X|Y}(1 \mid 1) = \frac{3}{5}$$

Example 9.14.3. *Conditioning on at most 2 heads on 4 coin tosses: Let X be the number of heads in four tosses of a fair coin. Calculate the conditional expectation of X given 2 or less heads.*

Solution. *Let $A = (X \leq 2)$. Since X has the binomial $(4, \frac{1}{2})$ distribution,*

$$p(X = k) = \frac{\begin{pmatrix} 4 \\ k \end{pmatrix}}{2^4}, \quad k = 0, 1, 2, 3, 4$$

and so

$$\mathbb{P}(X \leq 2) = \frac{1 + 4 + 6}{16} = \frac{11}{16}.$$

and

$$\mathbb{P}(X = k \mid X \leq 2) = \begin{pmatrix} 4 \\ k \end{pmatrix} /11, \quad k = 0, 1, 2$$

$$\mathbb{E}(X \mid X \leq 2) = \sum_{k=0}^{2} k \begin{pmatrix} 4 \\ k \end{pmatrix} /11 = \frac{1 \cdot 4 + 2 \cdot 6}{11} = \frac{16}{11}.$$

Example 9.14.4. *The conditional CDP of X given $Y = y$ is*

$$\begin{aligned} F_{X|Y}(a \mid y) &= \mathbb{P}\left(X \leq a \mid Y = y\right) \\ &= \int_{-\infty}^{a} f_{X|Y}\left(x \mid y\right) dx. \end{aligned}$$

Fact: *If X, Y are independent then*

$$f_{X|Y}(x \mid y) = f_X(x).$$

Example 9.14.5. *The joint PDF of X, Y is given by*

$$f_{X,Y}(x,y) = \begin{cases} \frac{12}{5} x\left(2 - x - y\right) & 0 < x < 1, 0 < y < 1 \\ 0 & \textit{otherwise} \end{cases}.$$

Compute the conditional PDF of X given that $Y = y$ where $0 < y < 1$.

Solution. *We have*

$$\begin{aligned} f_{X|Y}(x \mid y) &= \frac{f(x,y)}{f_Y(y)} = \frac{x\left(2 - x - y\right)}{\int_0^1 x\left(2 - x - y\right) dx} \\ &= \frac{x\left(2 - x - y\right)}{\frac{2}{3} - \frac{y}{2}} = \frac{6x(2 - x - 5}{4 - 3y} \end{aligned}$$

Example 9.14.6. *Let X and Y have the joint PDF*

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

Find $\mathbb{P}(X > 1 \mid Y = y)$.

The conditional PDF is given by

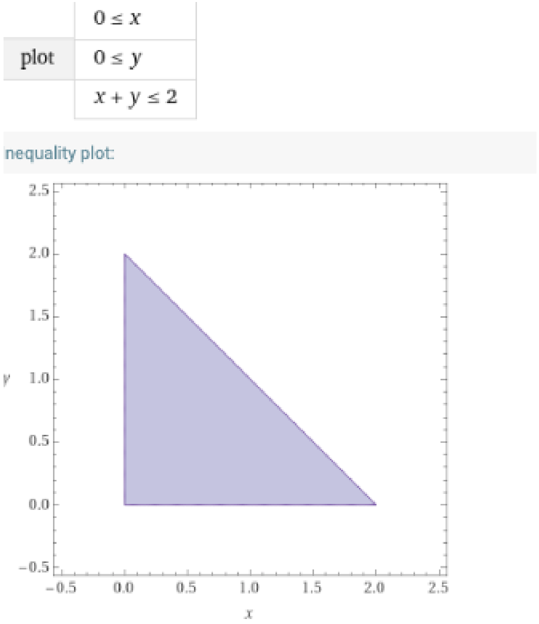
$$f_{X|Y}(x \mid y) = \frac{f_{x,Y}(x,y)}{f_Y(y)} = \frac{\frac{e^{-x/y}e^{-y}}{y}}{e^{-y}\int_0^\infty (1/y)e^{-x/y}dx} = \frac{1}{y}e^{-x/y}$$

$$\mathbb{P}(X > 1 \mid Y = y) = \int_1^\infty \frac{1}{y}e^{-x/y}dy = -e^{-x/y}\Big|_1^\infty = e^{-\frac{1}{y}}.$$

Example 9.14.7. *Suppose that a point (X, Y) is chosen uniformly at random from the triangle*

$$T = \{(x,y) : x \geq 0, y \geq 0, x + y \leq 2\}$$

Find $\mathbb{P}(Y > 1 \mid X = x)$



	0 ≤ x
plot	0 ≤ y
	x + y ≤ 2

nequality plot:

Solution.

$$f_{X,Y} = \begin{cases} \frac{1}{2}, & 0 \leq x, 0 \leq y, x + y \leq 2 \\ 0, & otherwise \end{cases}$$

So, for $0 \leq x \leq 2$,

$$f_X(x) = \int_0^\infty f(x,y)dy = \int_0^{2-x} \frac{1}{2}dy = \frac{1}{2}(2 - x)$$

and so

$$f_{Y|X=x}(x,y) = \frac{f(x,y)}{f_X(x)} = \frac{1}{2-x},\ 0 \leq y \leq 2-x$$

and 0 *else.*

$$\mathbb{P}(Y > 1 \mid X = x) = \begin{cases} \int_0^{2-x} \frac{dy}{2-x} = \frac{1-x}{2-x}, & 0 \leq x \leq 1 \\ 0 & \textit{otherwise} \end{cases}$$

9.15 Joint PDF of functions of random variables

In our lecture notes for weeks 8 and 9 we gave a a formula to compute the PDF of the random random variable $Y = g(X)$ where g is either increasing or decreasing and X is a continuous random variable. In fact, we had the following formula:

$$\boxed{f_Y(y) = \frac{f_X(x)}{|g'(x)|},\ x = g^{-1}(y)}$$

That is,

$$\boxed{f_Y(y) = \frac{f_X(g^{-1}(y))}{|g'(g^{-1}(y))|}}$$

The goal of this section is do the "same" for functions of two random variables. That is, we want to answer the following

Question 9.15.1. *Suppose $f_{X_1,X_2}(x_1,x_2)$ is the joint PDF of (X_1,X_2). Suppose g_1 and g_2 are functions of two variables. That is, $g_1 : \mathbb{R}^2 \to \mathbb{R}$ and $g_2 : \mathbb{R}^2 \to \mathbb{R}$. What is the joint PDF of (Y_1,Y_2) where $Y_1 = g_1(X_1,X_2)$ and $y_2 = g_2(X_1,X_2)$?*

The answer is given by

Theorem 9.15.1

Suppose $Y_1 = g_1(X_1,X_2)$ and $Y_2 = g_2(X_1,X_2)$. Let h_1 and h_2 be such that $X_1 = h_1(Y_1,Y_2)$ and $X_2 = h_2(Y_1,Y_2)$. Let J be the Jacobian of the mapping $(x_1,x_2) \to (g_1(x_1,x_2), g_2(x_1,x_2))$. That is,

$$J(x_1,x_2) = \begin{vmatrix} \frac{\partial g_1}{\partial x_1} & \frac{\partial g_1}{\partial x_2} \\ \frac{\partial g_2}{\partial x_1} & \frac{\partial g_2}{\partial x_2} \end{vmatrix} = \frac{\partial g_1}{\partial x_1}\frac{\partial g_2}{\partial x_2} - \frac{\partial g_1}{\partial x_2}\frac{\partial g_2}{\partial x_1}$$

Then by, provided this quantity is not zero, the change of variables theorem from multivariable calculus,

$$\boxed{f_{Y_1,Y_2}(y_1,y_2) = \frac{f_{X_1,X_2}}{|J|}}$$

which is defined on the range of the random variables (Y_1,Y_2), and is analogous to the one variable result.

Remark 9.15.2. *The pair of functions h_1 and h_2 can be though as the "inverse" of the pair g_1 and g_2 similar to the way g^{-1} which is the inverse of g. In the same way the Jacobian J plays the role of $g'(g^{-1}(y)) = g'(x)$*

Properties 9.15.1: Steps to Find f_{Y_1,Y_2}

Step 1: Find the Jacobian $J(x_1, x_2)$ as above.

Step 2: Set $y_1 = g_1(x_1, x_2)$ and $y_2 = g_2(x_1, x_2)$ Solve these for x_1 and x_2 in terms of y_1 and y_2. This will give you x_1 as a function of y_1 and y_2 and similarly for x_2. Set

$$\begin{aligned} x_1 &= h_1(y_1, y_2), \\ x_2 &= h_2(y_1, y_2). \end{aligned}$$

Step 3: The joint pdf of Y_1 and Y_2 is

$$\begin{aligned} f_{Y_1,Y_2}(y_1, y_2) &= \frac{f_{X_1,X_2}(x_1, x_2)}{|J(x_1, x_2)|} \\ &= \frac{f_{X_1,X_2}(h_1(y_1, y_2), h_2(y_1, y_2))}{|J(h_1(y_1, y_2), h_2(y_1, y_2))|} \end{aligned}$$

Example 9.15.1. *Suppose X and Y have a joint PDF $f_{X,Y}(x, y)$. Let $U = 1 - X - Y$ and $V = X - Y$. Find the joint PDF of U and V in terms of $f_{X,Y}$.*

Solution*Step 1: Find the Jacobian. With*

$$\begin{aligned} g_1(x, y) = 1 - x - y, \\ g_2(x, y) = x - y, \end{aligned}$$

we have

$$J(x, y) = \begin{vmatrix} -1 & -1 \\ 1 & - \end{vmatrix} = 2.$$

Step 2: Set

$$\begin{aligned} u &= g_1(x, y) = 1 - x - y \\ v &= g_2(x, y) = x - y. \end{aligned}$$

Subtracting one equation from the other solves for x and adding them solves for y.

$$\begin{aligned} x &= \frac{1 - u + v}{2} = h(u, v), \\ y &= \frac{1 - u - v}{2} = h(u, v). \end{aligned}$$

Step 3: The joint PDF of U and V is given

$$\begin{aligned} f_{U,V}(u,v) &= \frac{f_{X,Y}(x,y)}{|J(x,y))|} \\ &= \frac{f_{X,Y}\left(\frac{(1-u+v)}{2},\frac{(1-u-v)}{2}\right)}{2} \\ &= \frac{1}{2}f_{X,Y}\left(\frac{1-u+v}{2},\frac{1-u-v}{2}\right) \end{aligned}$$

Chapter 10

Week 12: Expectation

Let us recall from Chapters 5 and 6, the definition and basic properties of expectation for discrete and continuous random variables.

10.1 Properties

Properties 10.1.1: Expectation of discrete and continuous random variables

(i) Definition 1: The expectation of a discrete random variable X is

$$\mathbb{E}X = \sum_x x\mathbb{P}(X = x)$$

It is the average value of X relative to the weights given by the mass distribution function $p_X(x) = P(X = x)$

(ii) Definition 2: The expectation of a continuous random variable X with PDF f_X is

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

(iii) Expectation of Indicators: For any event A in the sample space S, we define the random variable $\mathbb{1}_A(s)$, called the indicator of A, to be

$$\mathbb{1}_A(s) = \begin{cases} 1 & s \in A \\ 0 & s \notin A \end{cases}$$

Then

$$\mathbb{E}\mathbb{1}_A = \mathbb{P}(A)$$

(iv) Constants: The expectation of a constant random variable c is just

$$\mathbb{E}c = c$$

and

$$\mathbb{E}(cX) = c\mathbb{E}X$$

(v) Functions of random variables, discrete case:

$$\mathbb{E}g(X) = \sum_x g(x)\mathbb{P}(X = x)$$

(vi) Functions of random variables, continuous case:

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

(vii) Sums: Expectation of sums equal sums of expectations, that is

$$\mathbb{E}(X + Y) = \mathbb{E}X + \mathbb{E}Y$$

These properties are better formulated using the joint distributions of random variables.

Theorem 10.1.1: Expectation of jointly distributed random variables

Let $g : \mathbb{R}^2 \to \mathbb{R}$.

(i) If If (X, Y) are discrete random variables with joint pmf $p_{X,Y}(x, y)$ then

$$\mathbb{E}\left[g\left(X, Y\right)\right] = \sum_y \sum_x g(x, y) p_{X,Y}(x, y).$$

(ii) If (X, Y) are jointly continuous random variables with PDF $f_{X,Y}(x, y)$, then

$$\mathbb{E}\left[g(X, Y)\right] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f_{X,Y}(x, y) dxdy.$$

(iii) If X and Y are independent and $g(x, y) = g_1(x)g_2(y)$, then

$$\begin{aligned}
\mathbb{E}(g(X, Y)) &= \mathbb{E}(g_1(X)g_2(Y)) \\
&= \sum_y \sum_x g_1(x)g_2(y)p_X(x)p_Y(y) \\
&= \sum_x g_2(x)p_X(x) \sum_y g_2(y)p_Y(y) \\
&= \mathbb{E}(g_1(X))\mathbb{E}(g_2(Y)).
\end{aligned}$$

Similarly in the continuous case.

Proof. Let us proof (ii) in the case when $g \geq 0$. In this case the random variable $Z = g(X, Y) \geq 0$. Recall that

$$\mathbb{P}(g(X, Y) > t) = \int\int_{\{g(x,y)>t\}} f_{X,Y}(x, y) dxdy = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \mathbb{1}_{\{g(x,y)>t\}}(x, y) f_{X,Y}(x, y) dxdy$$

The by our formula from Chapter 6 (already used many times including in homework and exam problems)

$$\begin{aligned}
\mathbb{E}g(X, Y) &= \int_0^{\infty} \mathbb{P}(g(X, Y) > t) dy \\
&= \int_0^{\infty} \left(\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \mathbb{1}_{\{g(x,y)>t\}}(x, y) f_{X,Y}(x, y) dxdy \right) dt \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \left(\int_0^{g(x,y)} dt \right) f_{X,Y}(x, y) dxdy \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) dxdy
\end{aligned}$$

■

From this we have the following particular cases that arise quite often and which are easy to verify.

Corollary 10.1.1: $\mathbb{E}(X+Y) = \mathbb{E}X + \mathbb{E}Y$

With the function $g(x,y) = x+y$, we have in the continuous case:

$$\begin{aligned}\mathbb{E}(X+Y) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x+y)f_{X,Y}(x,y)dxdy \\ &= \int_{-\infty}^{\infty} x\left(\int_{-\infty}^{\infty} f_{X,Y}(x,y)dy\right)dx + \int_{-\infty}^{\infty} y\left(\int_{-\infty}^{\infty} f_{X,Y}(x,y)dx\right)dy \\ &= \int_{-\infty}^{\infty} xf_X(x)dx + \int_{-\infty}^{\infty} yf_Y(y)dy = \mathbb{E}X + \mathbb{E}Y\end{aligned}$$

Similarly for the discrete case.

Corollary 10.1.2

If X is a nonnegative integer valued discrete random variable then

$$\mathbb{E}X = \sum_{j=1}^{\infty}\mathbb{P}(X \geq j)$$

Note: This is the discrete analogue of the formula for continuous random variable which states if $X \geq 0$, then

$$\mathbb{E}X = \int_0^{\infty}\mathbb{P}(X > t)dt$$

Proof. For $i \geq 1$, let $A_i = \{X \geq i\}$. Then

$$\mathbb{1}_{A_i}(s) = \begin{cases}1, & s \in A_i \\ 0, & s \notin A_i\end{cases} = \begin{cases}1, & X(s) \geq i \\ 0, & X(s) < i\end{cases}$$

We can write

$$\sum_{i=1}^{\infty}\mathbb{1}_{A_i}(s) = \sum_{i=1}^{X(s)}\mathbb{1}_{A_i}(s) + \sum_{i=X(s)+1}^{\infty}\mathbb{1}_{A_i}(s) = \sum_{i=1}^{X(s)} 1 = X(s)$$

Thus,

$$\mathbb{E}X = \sum_{i=1}^{\infty}\mathbb{E}X_i = \sum_{i=1}^{\infty}\mathbb{P}(X \geq i).$$

■

The above formulas hold for the joint distribution of a random variables when one is discrete and the other is continuous. We list the main formula only.

Properties 10.1.2: Discrete & Continuous Random Variables

Let $g : \mathbb{R}^2 \to \mathbb{R}$. Let X be a discrete random variable and Y be a continuous random variable. Suppose X and Y are independent. Let $p_X(x)$ be the pmf of X and $f_Y(y)$ be the PDF of Y. Then

$$\mathbb{E}\left[g\left(X, Y\right)\right] = \sum_{x_i} \int_{-\infty}^{\infty} g(x_i, y) p_X(x_i) f_Y(y) dy.$$

10.1.1 Examples

Example 10.1.1. *Let us recall that from the definition of properties of a probability we have that for any collection of events $\{A_i\}_{i=1}^n$ that are mutually disjoint, that is, $A_i \cap Aj = \varnothing$, we have*

$$\mathbb{P}\left(\bigcup_i^n A_i\right) = \sum_{i=1}^n \mathbb{P}(A_i).$$

and we know from the inclusion-execution formula that without the mutual disjointness the inequality may not hold. Bool's inequality says that

$$\mathbb{P}\left(\bigcup_i^n A_i\right) \leq \sum_{i=1}^n \mathbb{P}(A_i).$$

for all sequences of $A_i's$ regardless of whether they are disjoint or not.

Solution. *Set $X = \sum_{i=1}^n \mathbb{1}_{A_i}$. We have*

$$\mathbb{E}X = \sum_{i=1}^n \mathbb{E}X_i = \sum_{i=1}^n \mathbb{P}(A_i)$$

On the other hand, if we set

$$Y = \begin{cases} 1, & X \geq 1 \\ 0, & otherwise \end{cases}$$

and we see that $X \geq Y$. Thus

$$\mathbb{E}Y \leq \mathbb{E}X = \sum_{i=1}^n \mathbb{P}(A_i).$$

But what is $\mathbb{E}Y$? From the definition of Y we see that Y is nonzero if and only if at least one of the sets occurs. That is $Y = \mathbb{1}_A$ where $A = \cup_i^n A_i$. This gives that

$$\mathbb{P}\left(\bigcup_i^n A_i\right) \leq \sum_{i=1}^n \mathbb{P}(A_i).$$

Example 10.1.2. *Let $X_1, X_2, \ldots X_n$ be independent and identically distributed random variables with common CDF F and mean μ. Such a sequence of random variables is said to constitute a sample from the distribution F. The random variable*

$$\overline{X} = \sum_{j=1}^n \frac{X_j}{n}$$

is called the sample mean. Compute $\mathbb{E}\overline{X}$.

Solution.

$$\mathbb{E}\overline{X} = \frac{1}{n}\sum_{j=1}^{n} \mathbb{E}X_j = \mu$$

Thus the expected value of the sample mean is μ. In statistics, the sample means is used to estimate the mean when that is not known.

Example 10.1.3. *Suppose the joint p.m.f of X and Y is given by*

$X \diagdown Y$	0	2
0	.2	.7
1	0	.1

Find $\mathbb{E}[XY]$.

Solution. *Using the formula with the function $g(x,y) = xy$ we have*

$$\begin{aligned} \mathbb{E}[XY] &= \sum_{i,j} x_i y_j p(x_i, y_j) \\ &= 0 \cdot 0p(0,0) + 1 \cdot 0p(1,0) + 0 \cdot 2p(0,2) + 1 \cdot 2p(1,2) \\ &= 0.2 \end{aligned}$$

Example 10.1.4. *Suppose X, Y are independent exponential r.v. with parameter $\lambda = 1$. Set up a double integral that represents*

$$\mathbb{E}\left[X^2 Y\right].$$

Solution. *Since X, Y are independent then*

$$f_{X,Y}(x,y) = e^{-1x}e^{-1y} = e^{-(x+y)}. \ 0 < x, y < \infty.$$

$$\begin{aligned} \mathbb{E}\left[X^2 Y\right] &= \int_0^\infty \int_0^\infty x^2 y e^{-(x+y)} dy dx \\ &= \int_0^\infty y e^{-y} \left(\int_0^\infty x^2 e^{-x} dx\right) dy = 2 \end{aligned}$$

Example 10.1.5. *Suppose the joint PDF of X, Y is*

$$f(x,y) = \begin{cases} 10xy^2 & 0 < x < y, 0 < y < 1 \\ 0 & \textit{otherwise} \end{cases}.$$

Find $\mathbb{E}XY$ and $Var(Y)$.

Solution.

$$\begin{aligned}
\mathbb{E}XY &= \int_0^1 \int_0^y xy\left(10xy^2\right)dxdy = 10\int_0^1\int_0^y x^2y^3dxdy \\
&= \frac{10}{3}\int_0^1 y^3y^3dy = \frac{10}{3}\frac{1}{7} = \frac{10}{21}.
\end{aligned}$$

To find the Var(Y) we need to find $\mathbb{E}Y$ *and* $\mathbb{E}Y^2$ *as usual. For these we hav,*

$$\begin{aligned}
\mathbb{E}Y &= \int_0^1 \int_0^y y\left(10xy^2\right)dxdy = 10\int_0^1\int_0^y y^3xdxdy \\
&= 5\int_0^1 y^3y^2dy = \frac{5}{6}.
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}Y^2 &= \int_0^1 \int_0^y y^2\left(10xy^2\right)dxdy = 10\int_0^1\int_0^y y^4xdxdy \\
&= 5\int_0^1 y^4y^2dy = \frac{5}{7}.
\end{aligned}$$

Thus,

$$\mathit{Var}(Y) = \frac{5}{7} - \left(\frac{5}{6}\right)^2 = \frac{5}{252}.$$

Problem 10.1.1. *Suppose X and Y are independent random variables with X uniform on $(0,1)$ Y has exponential with mean 1. Find*

(i) $\mathbb{E}(X+Y)$

(ii) $\mathbb{E}(XY)$

(iii) $\mathbb{E}(X+Y)^2$

(iv) $\mathbb{E}(X^2e^{2Y})$.

Problem 10.1.2. *Suppose X_1 and X_2 are two jointly continuous, independent random variables with the same PDF, say f (and CDF, say F). Find $\mathbb{P}(X_1 < X_2)$.*

Problem 10.1.3. *Let X and Y be independent random variables with $\mathbb{E}X = 1$, $\mathbb{E}Y = 2$, Var$(X) = 3$, and Var$(Y) = 4$. (a) Find $\mathbb{E}(8X^2 + 6Y^2 - XY + 8X + 5Y - 1)$. (b) Assume they normally distributed, find $\mathbb{P}(2X > 3Y - 5)$.*

10.2 Expectation of conditional expectation

Properties 10.2.1

Applying the above formula to the conditional random variable $X|Y$ as introduced in Chapter 10 we have for any $g : \mathbb{R} \to \mathbb{R}$,

$$\mathbb{E}(g(X) \mid Y = y_j) = \sum_{x_i} g(x_i) p_{X|Y=y_j}(x_i) = \sum_{x_i} g(x_i) \frac{p_{X,Y}(x_i, y)}{p_Y(y_j)}, \tag{10.2.1}$$

for the discrete case and

$$\mathbb{E}(g(X) \mid Y = y) = \int_{-\infty}^{\infty} g(x) f_{X|Y=y}(x, y) dx = \int_{-\infty}^{\infty} g(x) \frac{f_{X,Y}(x, y)}{f_Y(y)} dx, \tag{10.2.2}$$

in the continuous case. Equivalently:

$$\mathbb{E}(g(X) \mid Y = y_j) p_Y(y_j) = \sum_{x_i} g(x_i) p_{X,Y}(x_i, y_j), \tag{10.2.3}$$

and

$$\mathbb{E}(g(X) \mid Y = y) f_Y(y) = \int_{-\infty}^{\infty} g(x) f_{X,Y}(x, y) dx, \tag{10.2.4}$$

Finally,

$$\mathbb{E}(g(X, Y) \mid Y = y) = \mathbb{E}(g(X, y) \mid Y = y)$$

For example,

$$\mathbb{E}(XY \mid Y = y) = y\mathbb{E}(X \mid Y = y), \;\; \mathbb{E}(X + Y) \mid Y = y) = \mathbb{E}(X \mid Y = y) + y.$$

Notice that in the discrete case (10.2.3) gives

$$\begin{aligned}
\sum_{y_j} \mathbb{E}(g(X) \mid Y = y_j) p_Y(y_j) &= \sum_{y_j} \sum_{x_i} g(x_i) p_{X,Y}(x_i, y_j) \\
&= \sum_{x_i} g(x_i) \sum_{y_j} p_{X,Y}(x_i, y_j) \\
&= \sum_{x_i} g(x_i) p_X(x_i) = \mathbb{E}(g(X),
\end{aligned}$$

where we used the fact that $p_X(x)$ is the marginal of $p_{X,Y}$. That is, $p_X(x) = \sum_{y_j} p_{X,Y}(x_i, y_j)$ On the other hand, the left hand,

$$\mathbb{E}(\mathbb{E}(g(X) \mid Y)) = \sum_{y_j} \mathbb{E}(g(X) \mid Y = y_j) p_Y(y_j)$$

Thus

$$\mathbb{E}(\mathbb{E}(g(X) \mid Y)) = \mathbb{E}(g(X)).$$

In the same way it follows from (10.2.4) that

$$\mathbb{E}(\mathbb{E}(g(X) \mid Y)) = \mathbb{E}(g(X)),$$

for continuous random variables.

Notice that also from the right hand side of (10.2.1) and (10.2.2) that if X and Y are independent then (using the fact that $p_{X,Y}(x_i, y_j) = p_X(x_i)p_Y(y_j)$)

$$\mathbb{E}(g(X) \mid Y = y_j) = \sum_{x_i} g(x_i)p_X(x_i) = \mathbb{E}g(X), \text{ for all } y_j, \tag{10.2.5}$$

in the discrete case and

$$\mathbb{E}(g(X) \mid Y = y) = \int_{-\infty}^{\infty} g(x)f_X(x)dx = \mathbb{E}X, \text{ for all } y \tag{10.2.6}$$

in the continuous case. In either case, we see if X and Y are independent, the random variable $\mathbb{E}(X \mid Y)$ is constant and equals $\mathbb{E}X$. We summarize:

Properties 10.2.2: The expectation conditional expectation

Let $g : \mathbb{R} \to \mathbb{R}$. Let X and Y be two random variables. We have

$$\mathbb{E}(\mathbb{E}(g(X) \mid Y)) = \mathbb{E}(g(X))$$

for both discrete and continuous random variables. In particular, if $g(x) = x$ we have

$$\mathbb{E}(\mathbb{E}(X \mid Y)) = \mathbb{E}X.$$

(i) That is, the expectation of the conditional expectation is just the expectation.

(ii) If X and Y are independent, then the random variable $\mathbb{E}(X \mid Y)$ is constant and it equals $\mathbb{E}X$. This is as "it should be." If the random variables are independent then knowing Y provides no additional information on X. Thus the expectation fo X given Y should just be the expectation of X.

10.2.1 Examples

Example 10.2.1. *Recall the Law of total probability from Chapter 4: Suppose $S = \cup_{j=1}^{n} F_j$ where the events $F_1, \dots, F_n$ are pairwise disjoint. Then for any event A,*

$$\mathbb{P}(A) = \sum_{j=1}^{n} P(A \cap F_j) = \sum_{j=1}^{n} \mathbb{P}(A \mid F_j)\mathbb{P}(F_j).$$

This is in fact a special case of the inequality above that for discrete random variables,

$$\mathbb{E}X = \sum_{y_j} \mathbb{E}(X \mid Y = y_j)p_Y(y_j) = \sum_{y_j} \mathbb{E}(X \mid Y = y_j)\mathbb{P}(Y = y_j)$$

Indeed, let $X = \mathbb{1}_A$ and define Y taking values $i \in \{1, 2, \dots, n\}$ such that $\{Y = i\} = F_i$. Then $\mathbb{E}X = \mathbb{P}(A)$, $\mathbb{E}(X \mid Y = y_i) = \mathbb{P}(A \mid F_i)$ and $\mathbb{P}(Y = i) = \mathbb{P}(F_i)$.

Example 10.2.2. *Suppose you are on the 10th floor of a building with three doors marked A, B, C. Door A leads to to stairs that will take you out of the building in 15 minutes. Door B leads to stairs that return you to the 10th floor lobby in 25 minutes. Door C leads to stairs that will return you to the 10th floor lobby in 30 minutes. Suppose you are equally likely to choose any of the three doors at anytime, what is the expected time until you get out of the building?*

Solution. *Let X be the total time until you get out of the building and let Y be the initial door you pick. Then*

$$\begin{aligned}\mathbb{E}X &= \mathbb{E}(X \mid Y = A)\mathbb{P}(Y = A) + \mathbb{E}(X \mid Y = B)\mathbb{P}(Y = B) + \mathbb{E}(X \mid Y = C)\mathbb{P}(Y = C) \\ &= \frac{1}{3}\left(\mathbb{E}(X \mid Y) + \mathbb{E}(X \mid Y = B) + \mathbb{E}(X \mid Y = C)\right)\end{aligned}$$

We are given that

$$\mathbb{E}(X \mid Y = A) = 15.$$

On the other hand, if you pick any of the other doors you expected exit time the time it takes you to get back back the 10th flot plus the expect time to get out of the building. That is, after each return, you start a fresh. So,

$$\mathbb{E}(X \mid Y = B) = 25 + \mathbb{E}X$$

and

$$\mathbb{E}(X \mid Y = B) = 30 + \mathbb{E}X.$$

This gives,

$$\mathbb{E}X = \frac{1}{3}\left(15 + 25 + \mathbb{E}X + 30 + \mathbb{E}X\right).$$

Solving gives $\mathbb{E}X = 70$ minutes.

Example 10.2.3. *Let Y be a uniform random variable on $(0, 1)$ and suppose the conditional distribution of X given $Y = p$ is binomial with parameters n and p. Find the probability mass function of X.*

Solution.

$$\begin{aligned}\mathbb{P}(X = k) &= \int_0^1 \mathbb{P}(X = k \mid Y = p) f_Y(p)dp \\ &= \int_0^1 \mathbb{P}(X = k \mid Y = p)dp \\ &= \frac{n!}{k!(n-k)!}\int_0^1 p^k(1-p)^{n-k}dp\end{aligned}$$

Now recall that from our the properties of the Beta distribution that

$$\int_0^1 x^{a-1}(1-x)^{b-1}dx = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}.$$

Apply this with $a = k+1$ and $b = n-k+1$ and recall that $\Gamma(m) = (m-1)!$ for any m. This gives that

$$\begin{aligned}\mathbb{P}(X = k) &= \frac{n!}{k!(n-k)!}\frac{\Gamma(k+1)\Gamma(n-k+1)}{\Gamma(n+2)}, \\ &= \frac{1}{n+1}\end{aligned}$$

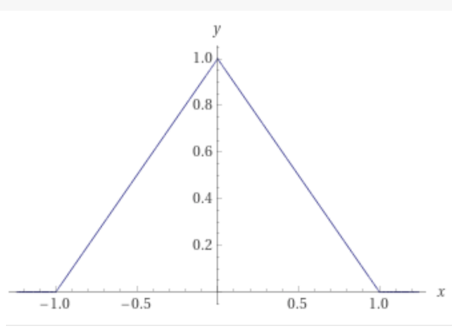
for all $k = 0, 1, \ldots n$. That is, the random variable X is has a discrete uniform distribution on $\{0, 1, \ldots n\}$

Example 10.2.4. *Suppose X and Y are independent. Find $\mathbb{E}(X+Y|X=x)$*

Solution. *From our properties listed above, $\mathbb{E}(X+Y \mid X=x) = x + \mathbb{E}(Y \mid X = x)$. Since X and Y are independent, $\mathbb{E}(Y \mid X = x) = \mathbb{E}(Y)$. Thus,*

$$\mathbb{E}(X+Y \mid X=x) = x + \mathbb{E}Y.$$

Problem 10.2.1. *Suppose (X,Y) has uniform distribution on the triangle*



Let $-1 < x < 1$. Find

(i)
$$\mathbb{P}(Y \geq \frac{1}{2} \mid X = x)$$

(ii)
$$\mathbb{P}(Y < \frac{1}{2} \mid X = x)$$

(iii)
$$\mathbb{E}(Y \mid X = x)$$

(iv)
$$Var(Y \mid X = x)$$

10.3 Covariance.

As we have already seen, the expected value and the variance of a random variable give us information about that random variable. In the same way, covariance between two random variables give us information about the relationship between the two random variables. This quantity is defied by

Definition 10.3.1: Covariance

Let X and Y be two random variables with mean $\mu_X = \mathbb{E}X$, $\mu_Y = \mathbb{E}Y$. Their covariance is defined by

$$\operatorname{Cov}(X, Y) = \mathbb{E}\left[(X - \mu_X)(Y - \mu_Y)\right] \tag{10.3.1}$$

Note that $\operatorname{Cov}(X, Y) = \operatorname{Cov}(Y, X)$ and that multiplying out gives that

$$(X = \mu_Y)(Y - \mu_Y) = XY - X\mu_Y - \mu_X Y + \mu_X\mu_Y$$

Taking expectations we see that

$$\operatorname{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}X\mathbb{E}Y \tag{10.3.2}$$

Properties 10.3.1

The covariance $\operatorname{Cov}(X, Y)$ satisfies the following properties:

(i) If $X = Y$, then

$$\operatorname{Cov}(X, X) = \operatorname{Var}(X)$$

(ii) X and Y are independent then

$$\operatorname{Cov}(X, X) = 0.$$

For any constants a and b,

(iii)

$$\operatorname{Cov}(aX, bY) = \operatorname{Cov}(aX, bY) = ab\operatorname{Cov}(X, Y).$$

These follow directly from (10.3.2).

Using the fact that $\sum_{k=1}^{n} X_k \sum_{j=1}^{m} Y_j = \sum_{k=1}^{n}\sum_{j=1}^{m} X_k Y_j$ we from (10.3.2) that

(i)

$$\operatorname{Cov}(\sum_{k=1}^{n} X_k, \sum_{j=1}^{m} Y_j) = \sum_{k=1}^{n}\sum_{j=1}^{m}\operatorname{Cov}(X_k, y_j).$$

The covariance of two random variables gives information on the relationship between the random variables. It is a measurement of how much the two random variables change together.

Properties 10.3.2

In terms of the joint pmf (discrete case) or PDF (continuous case) we have:

$$\mathrm{Cov}(X,Y) = \sum_x \sum_y xy\, p_{X,Y}(x,y) - \left(\sum_x x\, p_X(x)\right)\left(\sum_y y\, p_Y(y)\right)$$

$$\mathrm{Cov}(X,Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy\, f_{X,Y}(x,y)dxdy - \left(\int_{-\infty}^{\infty} x f_X(x)dx\right)\left(\int_{-\infty}^{\infty} y f_Y(y)dy\right)$$

We also have the following formula which gives the relationship between the variation $\mathrm{Var}(X+Y)$ and $\mathrm{Cov}(X,Y)$.

$$\begin{aligned}
\mathrm{Var}(X+Y) &= \mathbb{E}(X+Y)^2 - (\mathbb{E}(X+Y))^2 \\
&= \mathbb{E}(X)^2 + \mathbb{E}(Y)^2 + 2\mathbb{E}(XY) - (\mathbb{E}(X))^2 - (\mathbb{E}(Y))^2 - 2\left(\mathbb{E}(X)\right)\left(\mathbb{E}(Y)\right) \\
&= [\mathbb{E}(X)^2 - (\mathbb{E}(X))^2] + [\mathbb{E}(Y)^2 - (\mathbb{E}(Y))^2] + [2\mathbb{E}(XY) - 2\left(\mathbb{E}(X)\right)\left(\mathbb{E}(Y)\right)] \\
&= \mathrm{Var}X + \mathrm{Var}X + 2\mathrm{Cov}(X,Y)
\end{aligned}$$

We summarize.

Properties 10.3.3: Sum of Variation

For any two random variables X and Y we have

$$\mathrm{Var}\left(X+Y\right) = \mathrm{Var}\left(X\right) + \mathrm{Var}\left(Y\right) + 2\mathrm{Cov}\left(X,Y\right).$$

In general, for any sequence of random variables, $X_1, \ldots, X_n$,

$$\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathrm{Var}\left(X_i\right) + 2 \sum_{\{i,j=1:i<j\}}^{n} \mathrm{Cov}\left(X_i, X_j\right).$$

10.3.1 Examples

Example 10.3.1. *Let X uniformly distributed on $\{-2,-1,0,1,2\}$. and define Y such that $Y=1$ if $X=0$ and $Y=0$ otherwise. Find Cov(X,Y). Are X and Y independent?*

Solution.

$$\mathbb{E}X = \frac{1}{5}(-2-1+0+1+2) = 0$$

Sine $XY=0$. Cov$(X,Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) = 0$. But clearly X and Y are not independent since the definition of Y completely depends on X.

Example 10.3.2. *Let X have uniform distribution on $-1,0,1$ and let $Y = X^2$. Are X and Y uncorrelated? Are X and Y independent?*

Solution. *X takes the values $\{-1, 0, 1\}$ and Y takes the values $\{0, 1\}$.*

$$p(-1,0) = 0, p(-1,1) = 1/3, p(0,0) = 1/3, p(0,1) = 0, p(1,0) = 0, p(1,1) = 1/3.$$

$\mathbb{E}(X) = \mathbb{E}(XY) = 0$ and so $Cov(X,Y) = 0$. But $\mathbb{P}(X = -1, Y = 0) = 0 \neq \mathbb{P}(X = -1)\mathbb{P}(Y = 0)$ so X and Y not independent.

Remark 10.3.1. *Both Examples above are the "same" and they can be generalized considerably. Consider a random variable X with mean zero. Let $Y = 0$ whenever $X \neq 0$ and let $Y = \pi$ otherwise. Then $Cov(X,Y) = 0$ but clearly X and Y are not independent.*

Example 10.3.3. *Suppose A and B are two events in the sample space and let $X = \mathbb{1}_A$ and $Y = \mathbb{1}_B$, Find $Cov(\mathbb{1}_A, \mathbb{1}_B)$.*

Solution. *We have $\mathbb{E}\mathbb{1}_A = \mathbb{P}(A)$ and $\mathbb{E}\mathbb{1}_B = \mathbb{P}(B)$ and $\mathbb{E}(\mathbb{1}_A\mathbb{1}_B) = \mathbb{P}(A \cap B)$. Thus*

$$\begin{aligned} Cov(\mathbb{1}_A, \mathbb{1}_B) &= \mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B) \\ &= \mathbb{P}(A)[\mathbb{P}(B \mid A) - \mathbb{P}(B)] \end{aligned}$$

10.4 Correlation Coefficient

Definition 10.4.1: Correlation coefficient

The **correlation coefficient** of two random variables X and Y, denoted by $\rho(X,Y)$ is defined by

$$\rho(X,Y) = \frac{\text{Cov}(X,Y)}{\sqrt{\text{Var}(X)\,\text{Var}(Y)}}.$$

assuming both Var(X) and Var(Y) are both not zero.
The random variables variables X and Y are said to be

(i) **Positively correlated** if

$$\rho(X,Y) > 0$$

(ii) **Negatively Correlated** if

$$\rho(X,Y) < 0$$

(iii) **Uncorrelated** if

$$\rho(X,Y) = 0$$

Properties 10.4.1

For any two random variables X and Y the **correlation coefficient** satisfies

$$-1 \leq \rho(X,Y) \leq 1$$

Proof. Let $a = \sqrt{\mathrm{Var}(X)}$, $b = \sqrt{\mathrm{Var}(Y)}$. then

$$\begin{aligned}
0 \leq \mathrm{Var}\left(\frac{X}{a} + \frac{Y}{b}\right) &= \mathrm{Var}\left(\frac{X}{a}\right) + \mathrm{Var}\left(\frac{Y}{b}\right) + 2\mathrm{Cov}\left(\frac{X}{a}, \frac{Y}{b}\right) \\
&= \frac{\mathrm{Var}(X)}{a^2} + \frac{(Y)}{b^2} + 2\frac{\mathrm{Cov}\,(X,Y)}{ab} \\
&= = 1 + 1 + 2\frac{\mathrm{Cov}\,(X,Y)}{ab} \\
&= 2[1 + \frac{\mathrm{Cov}\,(X,Y)}{ab}] \\
&= 2[1 + \rho(X,Y) \qquad (10.4.1)
\end{aligned}$$

This shows that

$$-1 \leq \rho(X,Y)$$

In the same way,

$$\begin{aligned}
0 \leq \mathrm{Var}\left(\frac{X}{a} - \frac{Y}{b}\right) &= \mathrm{Var}\left(\frac{X}{a}\right) + \mathrm{Var}\left(\frac{Y}{b}\right) - 2\mathrm{Cov}\left(\frac{X}{a}, \frac{Y}{b}\right) \\
&= \frac{\mathrm{Var}(X)}{a^2} + \frac{(Y)}{b^2} - 2\frac{\mathrm{Cov}\,(X,Y)}{ab} \\
&= 2 - 2\frac{\mathrm{Cov}\,(X,Y)}{ab} \\
&= 2[1 - \frac{\mathrm{Cov}\,(X,Y)}{ab}] \\
&= 2[1 - \rho(X,Y)] \qquad (10.4.2)
\end{aligned}$$

This shows that

$$-1 \leq \rho(X,Y) \leq 1$$

■

Let us now recall that for any random variable X, if $\mathrm{Var}(X) = 0$, then $X = c$, a constant. From (10.4.2) we see that if $\rho(X,Y) = 1$, then

$$\mathrm{Var}\left(\frac{X}{a} - \frac{Y}{b}\right) = 0$$

and thus

$$\left(\frac{X}{a} - \frac{Y}{b}\right) = c$$

for some constant or

$$X = \frac{a}{b}Y + c$$

In the same way, (10.4.2) gives that

$$X = -\frac{a}{b}Y + c$$

for some constant c.

In ether case $\rho(X,Y) = \pm 1$ says that the X is a linear function of Y and visa verza. The correlation coefficient $\rho(X,Y)$ is a measurement of the linearity between X and Y. A positive value of value of ρ says that X tends to increase when Y does, whereas a negative value indicates that X tends to decrease when Y increases where as $\rho = 0$ says X and Y are uncorrelated.

10.4.1 Examples

Example 10.4.1. *Suppose X is uniform on $(0,1)$ and $Y = X^2$. Find the correlation between X and Y.*

Solution.

$$\mathbb{E}X = \int_0^1 x dx = \frac{1}{2}$$

$$\mathbb{E}Y = \mathbb{E}X^2 = \int_0^1 x^2 dx = \frac{1}{3}$$

$$\mathbb{E}Y^2 = \mathbb{E}X^4 = \int_0^1 x^4 dx = \frac{1}{5}$$

$$Var(X) = \frac{1}{3} - (\frac{1}{2})^2 = \frac{1}{12}.$$

$$Var(Y) = \frac{1}{5} - (\frac{1}{3})^2 = \frac{4}{45}.$$

$$\mathbb{E}(XY) = \mathbb{E}X^3 = \int_0^1 X^3 dx = \frac{1}{4}$$

$$Cov(X,Y) = \mathbb{E}(XY) - \mathbb{E}X\mathbb{E}Y = \frac{1}{4} - (\frac{1}{2})(\frac{1}{3}) = \frac{1}{12}$$

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)\,Var(Y)}} = \frac{\frac{1}{12}}{\sqrt{\frac{1}{12}\frac{4}{45}}} = \frac{\sqrt{15}}{4}$$

Example 10.4.2. *Roll a die. Let X be the value. Define $Y = 1$, if the die is 4, 5, or 6, and 0 otherwise. Find the correlation between X and Y.*

Example 10.4.3. *Suppose X, Y are random variables whose joint pdf is given by*

$$f(x,y) = \begin{cases} \frac{1}{y} & 0 < y < 1, 0 < x < y \\ 0 & \textit{otherwise} \end{cases}.$$

(a) Find the covariance of X and Y.

(b) Compute $Var(X)$ and $Var(Y)$.

(c) Calculate $\rho(X,Y)$.

Solution. *We start by finding the covariance of X and Y.*

(a)

$$\begin{aligned} \mathbb{E}XY &= \int_0^1 \int_0^y xy\frac{1}{y} dxdy = \int_0^1 \frac{y^2}{2} dy = \frac{1}{6} \\ \mathbb{E}X &= \int_0^1 \int_0^y x\frac{1}{y} dxdy = \int_0^1 \frac{y}{2} dy = \frac{1}{4}. \\ \mathbb{E}Y &= \int_0^1 \int_0^y y\frac{1}{y} dxdy = \int_0^1 ydy = \frac{1}{2}. \end{aligned}$$

Thus

$$\begin{aligned} Cov(X,Y) &= \mathbb{E}XY - \mathbb{E}X\mathbb{E}Y \\ &= \frac{1}{6} - \frac{1}{4}\frac{1}{2} \\ &= \frac{1}{24}. \end{aligned}$$

(b) We have that

$$\begin{aligned} \mathbb{E}X^2 &= \int_0^1 \int_0^y x^2 \frac{1}{y} dx dy = \int_0^1 \frac{y^2}{3} dy = \frac{1}{9}. \\ \mathbb{E}Y^2 &= \int_0^1 \int_0^y y^2 \frac{1}{y} dx dy = \int_0^1 y^2 dy = \frac{1}{3}. \end{aligned}$$

Thus

$$\begin{aligned} Var(X) &= \mathbb{E}X^2 - (\mathbb{E}X)^2 \\ &= \frac{1}{9} - \left(\frac{1}{4}\right)^2 = \frac{7}{144} \end{aligned}$$

And

$$\begin{aligned} Var(Y) &= \mathbb{E}Y^2 - (\mathbb{E}Y)^2 \\ &= \frac{1}{3} - \left(\frac{1}{2}\right)^2 = \frac{1}{12}. \end{aligned}$$

(c)

$$\begin{aligned} \rho(X,Y) &= \frac{Cov(X,Y)}{\sqrt{Var(X)\ Var(Y)}} \\ &= \frac{\frac{1}{24}}{\sqrt{\left(\frac{7}{144}\right)\left(\frac{1}{12}\right)}} \approx .6547. \end{aligned}$$

Chapter 11

Moment Generating Functions.

Moment generating functions provide a powerful and elegant tool to derive many important (and often nontrivial) properties of random variables.

11.1 Definition

Let us recall from Math 162 that the Taylor (or Mclaurin) series of a function f is given by

$$f(t) = \sum_{j=0}^{\infty} \frac{f^{(j)}(0)}{j!} t^j$$

where $f^{(j)}(0)$ is the jth derivative of f evaluated at 0. Here f is any function. Thus knowing all the derivatives of the function at 0 tells us what the function is at any other point t (provided of course that the series converges).

In general, given any sequence of numbers $a_0, a_1, a_2, \ldots$, we can consider the function

$$f(t) = \sum_{j=0}^{\infty} \frac{a_j}{j!} t^j = a_0 + \frac{a_1}{1!} t + \frac{a_2}{2!} t^2 + \frac{a_3}{3!} t^3 + \cdots \tag{11.1.1}$$

Differentiating the both sides gives that the jth derivative of f at 0 is just a_j. That is, $f^{(j)}(0) = a_j$, for all $j \geq 0$. In such case we say that the sequence of umbers $a_0, a_1, a_2, \ldots$ **generates** the function f.

Question 11.1.1

Suppose we are given a random variable X and use $a_j = \mathbb{E}X^j$ for $j \geq 0$, the moments of X. What function does this sequence **generate?**

Answer 11.1.1: Moment Generating Function

With $a_j = \mathbb{E}X^j$, recalling that $e^x = \sum_{j=0}^{\infty} \frac{x^j}{j!}$, the Taylor series (11.1.1) gives

$$\begin{aligned} f(t) &= \sum_{j=0}^{\infty} \frac{\mathbb{E}X^j}{j!} t^j = \mathbb{E}\left(\sum_{j=0}^{\infty} \frac{X^j}{j!} t^j\right) \\ &= \mathbb{E}\left(\sum_{j=0}^{\infty} \frac{(t\,X)^j}{j!}\right) = \mathbb{E}\, e^{tX}. \end{aligned}$$

In this particular case, to emphasize the dependence on the random variable X and the fact that the $a_j's$ are the jth moments moments of X, the function f is called the **Moment Generating Function (mgf)** of X and is denoted by $m_X(t)$ or $M_X(t)$. That is,

$$M_X(t) = m_X(t) = \mathbb{E}\, e^{tX}$$

In our class, following the book's notation, we will use mostly $m_X(t)$.
Applying the formula for expectations of functions of random variables with $g(x) = e^{tx}$ we see that

$$m_X(t) = \begin{cases} \mathbb{E}\, e^{tX} = \sum_{x_i} e^{tx_i} p_X(x_i), & X \text{ discrete} \\ \\ \mathbb{E}\, e^{tX} = \int_{-\infty}^{\infty} e^{tx} f_X(x) dx, & X \text{ continous} \end{cases} \tag{11.1.2}$$

$$m_X^{(j)}(t) = \begin{cases} \mathbb{E}\, e^{tX} = \sum_{x_i} x_i^j e^{tx_i} p_X(x_i), & X \text{ discrete} \\ \\ \mathbb{E}\, e^{tX} = \int_{-\infty}^{\infty} x^j e^{tx} f_X(x) dx, & X \text{ continous} \end{cases} \tag{11.1.3}$$

In both cases

$$m_X^{(j)}(0) = \mathbb{E}X^j, \;\; j \geq 0$$

In terms of the moment generating function, the variance of the random variable X is given by

$$\text{Var}(X) = m_X''(0) - (m_X'(0))^2$$

11.2 Examples

Examples: Bernoulli, Binomial, Poisson, Discrete uniform

- **Bernoulli:** Suppose X is a Bernoulli random variable with $\mathbb{P}(X=1)=p$ and $\mathbb{P}(X=0)=1-p$. Then

$$m_X(t)=\mathbb{E}e^{tX}=e^{t\cdot 1}p+e^{t\cdot 0}(1-p)=pe^t+(1-p).$$

- **Binomial:** Let $X\sim Bin(n,p)$. Then

$$\begin{aligned} m_X(t) &= \mathbb{E}e^{tX}=\sum_{i=0}^{n}e^{ti}\binom{n}{i}p^i(1-p)^{n-i} \\ &= \sum_{i=0}^{n}\binom{n}{i}\left(e^tp\right)^i(1-p)^{n-i}=\left(e^tp+1-p\right)^n, \end{aligned}$$

where we used the binomial theorem $\sum_{i=0}^{n}\binom{n}{i}a^ib^{n-i}=(a+b)^n$.

- **Poisson:** If $X\sim Poisson(\lambda)$ then

$$\begin{aligned} m_X(t)=\mathbb{E}e^{tX} &= \sum_{n=0}^{\infty}e^{tn}e^{-\lambda}\frac{\lambda^n}{n!} \\ &= e^{-\lambda}\sum_{n=0}^{\infty}e^{tn}\frac{\lambda^n}{n!} \\ &= e^{-\lambda}\sum_{n=0}^{\infty}\frac{\left(e^t\lambda\right)^n}{n!}. \end{aligned}$$

Applying $\exp(x)=e^x=\sum_{n=0}^{\infty}\frac{x^n}{n!}$ with $x=(e^t\lambda)$ gives

$$\begin{aligned} m_X(t) &= e^{-\lambda}e^{e^t\lambda}=e^{e^t\lambda-\lambda} \\ &= \exp\left(\lambda\left(e^t-1\right)\right) \end{aligned}$$

- **Discrete uniform** If $X\sim$ Uniform on $\{1,2,\ldots,N\}$ then

$$\begin{aligned} m_X(t) &= \sum_{k=1}^{N}\frac{e^{kt}}{N}=\frac{1}{N}\sum_{k=1}^{N}e^{kt} \\ &= \frac{e^t}{N}\sum_{k=1}^{N}e^{(k-1)t}=\frac{e^t(e^{Nt}-1)}{N(e^t-1)} \end{aligned}$$

Example: Exponential

If $X \sim exp(\lambda)$ then

$$\begin{aligned} m_X(t) &= \mathbb{E}e^{tX} \\ &= \int_0^\infty e^{tx}\lambda e^{-\lambda x}dx \\ &= \begin{cases} \frac{\lambda}{\lambda - t}, & t < \lambda \\ = \infty, & t \geq \lambda \end{cases} \end{aligned}$$

Example: Normal

Let Z be the standard normal, that is, $Z \sim N(0,1)$.

$$\begin{aligned} m_X(t) &= \mathbb{E}e^{tX} = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{tx}e^{-x^2/2}dx \\ &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-(x^2-2tx)/2}dx \\ &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-(x-t)^2/2+t^2/2}dx \\ &= e^{t^2/2}\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-(x-t)^2/2}dx \\ &= e^{t^2/2} \end{aligned}$$

If $X \sim N(\mu, \sigma^2)$ then $X = \mu + \sigma Z$ and

$$\begin{aligned} m_X(t) &= \mathbb{E}e^{tX} = \mathbb{E}e^{t\mu}e^{t\sigma Z} = e^{t\mu}\mathbb{E}e^{(t\sigma)Z} \\ &= e^{t\mu}m_X\left(t\sigma\right) = e^{t\mu}e^{(t\sigma)^2/2} = \exp\left(t\mu + \frac{t^2\sigma^2}{2}\right). \end{aligned}$$

By differentiating we have

$$m'_X(t) = (\mu + t\sigma^2)\exp\left(t\mu + \frac{t^2\sigma^2}{2}\right).$$

$$m''_X(t) = (\mu + t\sigma^2)^2\exp\left(t\mu + \frac{t^2\sigma^2}{2}\right) + \sigma^2\exp\left(t\mu + \frac{t^2\sigma^2}{2}\right)$$

which gives

$$\mathbb{E}X = m'_X(0) = \mu, \quad \text{and } \mathbb{E}X^2 = m''_X(0) = \mu^2 + \sigma^2$$

and

$$\mathrm{Var}(X) = \mathbb{E}X^2 - (\mathbb{E}X)^2 = \sigma^2$$

as we already knew.

11.3 Independence and distributions

Properties 11.3.1: Moment generating functions and Independence

Recall that if X and Y are independent, then $\mathbb{E}(g_1(X)g_2(Y)) = \mathbb{E}(g_1(X))\mathbb{E}(g_2(Y))$ for any two functions g_1 and g_2. Applying with both g_1 and g_2 equal e^{tx} gives that under independence,

$$\begin{aligned} m_{X+Y}(t) &= \mathbb{E}e^{t(X+Y)} = \mathbb{E}\left(e^{tX}e^{tY}\right) \\ &= \mathbb{E}\left(e^{tX}\right)\mathbb{E}\left(e^{tY}\right) \\ &= m_X(t)m_Y(t). \end{aligned}$$

More generally, if $X_1, X_2, \ldots, X_n$ are independent, then

$$m_{X_1+X_2+\cdots+X_n}(t) = m_{X_1} \cdot m_{X_2}(t) \cdots m_{X_n}(t)$$

Properties 11.3.2: Moment generating functions and distributions

For any two random variables X and Y.

(i) If $m_X(t) = m_Y(t) < \infty$ for all t in an interval, then X and Y have the same distribution. In words, if two random variables have the same moment generating function if and only if they have same distribution.

(ii) In the same way the distributions of a sequence of random variables X_n converges to the distribution of the random variable X (in the appropriate sense) if and only if the moment generating functions converge. That is, $m_{X_n}(t) \to m_X(t)$ if and only if $\mathbb{P}(X_n \leq a) \to \mathbb{P}(X \leq a)$. (See the case of binomials and Poisson below for an example. This will also be applied to the Central Limit Theorem.)

These will not be proved but it will be used.

Remark 11.3.1. *Recall that earlier we discussed the fact that if $X_n \sim Bin(n, p_n)$ with $np_n \to \lambda$ as $n \to \infty$, then the pmf of X_n converges to the pmf of X where $X \sim Poiss(\lambda)$. In fact, we showed this when $p_n = \lambda/n$. This also follows from the following:*

$$\begin{aligned} m_{X_n}(t) &= \left(e^t p_n + 1 - p_n\right)^n = \left(1 + p_n(e^t - 1)\right)^n \\ &= \left(1 + \frac{\lambda(e^t - 1)}{n}\right)^n \to \exp(\lambda(e^t - 1)) = m_X(t). \end{aligned}$$

where we use the calculus fact that $(1 + \frac{x}{n})^n \to e^x$, as $n \to \infty$, applied with $x = \lambda(e^t - 1)$.

Example: Geometric(p) and Negative Binomial

(i) Recall X is geometric p if its pms is given by

$$p_X(k) = (1-p)^{k-1}p, \;\; k = 1, 2, \ldots$$

With this,

$$\begin{aligned} m_X(t) &= \sum_{k=1}^{\infty} e^{kt} p(1-p)^{k-1} \\ &= pe^t \sum_{k=1}^{\infty} e^{(k-1)t} = pe^t \sum_{k=0}^{\infty} \left(e^t p(1-p)\right)^k \\ &= \frac{pe^t}{1 - e^t(1-p)} \end{aligned}$$

The quotient rule for differentiation easily give both the mean and the variance.

(ii) Recall that if $X_i \sim Geo(p)$ for $i = 1, 2, \ldots, r$, and they are independent, then $X = \sum_{i=1}^{r} X_i \sim NB(r, p)$. Thus if $X \sim NB(r, p)$,

$$m_X(t) = \left(\frac{pe^t}{1 - e^t(1-p)}\right)^r$$

In both formulas we must have $0 < 1 - e^t(1-p)$ which is the same as $e^t(1-p) < 1$ or $t < \ln(\frac{1}{1-p})$. Again, differentiation easily gives the mean μ and the variance σ^2.

Definition 11.3.1: Joint moment generating function

For two random variables X and Y their joint moment generating function is defined by

$$m_{X,Y}(t,s) = \mathbb{E}\left(e^{tX+sY}\right)$$

and in general, for any $X_1, X_2, \dots, X_n$,

$$m_{X_1,X_2,\dots,X_n}(t_1,t_2,\dots,t_n) = \mathbb{E}\left(e^{t_1X_1+t_2X_2,\dots,t_nX_n}\right)$$

The individual moment generating functions can be obtained from

$$m_{X_1,X_2,\dots,X_n}(t_1,t_2,\dots,t_n)$$

by setting all the $t_i's$ be zero except for one. That is,

$$m_{X_j}(t) = m_{X_1,X_2,\dots,X_n}(0,\dots 0,t,0\dots 0)$$

with t on the jth place.
As before the moment generating function determines the joint distribution and independence. The random variables $X_1, X_2, \dots, X_n$ are independent if and only if

$$m_{X_1,X_2,\dots,X_n}(t_1,t_2,\dots,t_n) = m_{X_1}(t_1)m_{X_2}(t_2)\cdots m_{X_n}(t_n)$$

11.3.1 Examples

Example 11.3.1. *Suppose that the mgf of X is given by $m_X(t) = e^{3(e^t-1)}$. Find $\mathbb{P}(X=0)$.*

Solution. *Given that*

$$m(t) = e^{3(e^t-1)} = e^{\lambda(e^t-1)} \quad \textit{where } \lambda = 3,$$

we see that $X \sim Poisson(3)$. Thus

$$\mathbb{P}(X=0) = e^{-\lambda}\frac{\lambda^0}{0!} = e^{-3}.$$

Example 11.3.2. *Earlier in the course we showed that if $X \sim \mathcal{N}\left(\mu_1, \sigma_1^2\right)$ and $Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$ are independent then*

$$X + Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$$

Use moment generating functions to verify this fact.

Solution. *By the fact that if two random variables have the same moment generating function, we just need to verify that $X+Y$ has the same moment generating function of the normal normal with $\mu = \mu_1 + m\mu_2$ and $\sigma^2 = \sigma_1^2 + \sigma_2^2$. To do this, since X and Y are independent, we have*

$$\begin{aligned} m_{X+Y}(t) &= m_X(t)m_Y(t) \\ &= \exp\left(t\mu_1 + \frac{t^2\sigma_1^2}{2}\right)\exp\left(t\mu_2 + \frac{t^2\sigma_2^2}{2}\right) \\ &= \exp\left(t(\mu_1+\mu_2) + \frac{t^2(\sigma_1^2+\sigma_2^2}{2})\right). \end{aligned}$$

which is the mgf of a normal with $\mu = \mu_1 + \mu_2$, $\sigma^2 = \sigma_1^2 + \sigma_2^2$.

Example 11.3.3. *In the same way we can verify that if* $X \sim bin(n, p)$ *and* $Y \sim bin(m, p)$*, and they are independent, then*

$$\begin{aligned} m_{X+Y}(t) &= m_X(t)m_Y(t) \\ &= \left(pe^t + (1-p)\right)^n \left(pe^t + (1-p)\right)^m \\ &= \left(pe^t + (1-p)\right)^{n+m}. \end{aligned}$$

Thus

$$X + Y \sim bin(m + n, p).$$

Example 11.3.4. *Let's recall that if* $X_i \sim Exp(\lambda)$ *for* $i = 1, 2, \ldots, n$ *and they are independent, then*

$$X = \sum_{i=1}^{n} X_i \sim Gamma(n, \lambda)$$

From this and the formula for the moment generating function of the exponential, we see that the moment generating function for a $Gamma(n, \lambda)$ *random variable is given by*

Example: $X \sim Gamma(n, \lambda)$

$$m_X(t) = \left(\frac{\lambda}{\lambda - t}\right)^n = \left(1 - \frac{t}{\lambda}\right)^{-n}, \ t < \lambda$$

Example 11.3.5. *Suppose* $m_X(t) = \left(1 - \frac{t}{2}\right)^{-3}$. *Find* $\mathbb{E}(X)$ *and* $Var(X)$.

Solution. *Differentiation using the chain rule gives*

$$m_X'(t) = -3\left(1 - \frac{t}{2}\right)^{-4}(-\frac{1}{2}) = \frac{3}{2}\left(1 - \frac{t}{2}\right)^{-4},$$

and

$$m_X''(t) = \frac{3}{2}(-4)\left(1 - \frac{t}{2}\right)^{-5}(-\frac{1}{2}) = 3\left(1 - \frac{t}{2}\right)^{-5}.$$

Thus, $\mathbb{E}(X) = m_X'(0) = \frac{3}{2}$, $\mathbb{E}(X^2) = m_X''(0) = 3$, *and*

$$Var(X) = m_X''(0) - \left(m_X'(0)\right)^2 = \frac{3}{4}.$$

Remark 11.3.2. *Since* X *is in fact a* $Gamma(3, 2)$ *random variable and we know that* $X \sim Gamma(\alpha, \lambda)$*, then* $\mu = \frac{\alpha}{\lambda}$ *and* $\sigma^2 = \frac{\alpha}{\lambda^2}$*, we knew how to find it already. But the point here is that this is a simple exercise once we know the form of the moment generating function.*

Example 11.3.6. *Suppose* X *is a* ***discrete random variable*** *with*

$$m_X(t) = \frac{1}{7}e^{2t} + \frac{3}{7}e^{3t} + \frac{2}{7}e^{5t} + \frac{1}{7}e^{8t}.$$

Find the pmf of X *and its* $Var(X)$

Solution. *Since we have, by definition of the mgf,*

$$m_X(t) = \mathbb{E}(e^{tX}) = \sum_{x_j} e^{tx_j} p_X(x_j) = \frac{1}{7}e^{2t} + \frac{3}{7}e^{3t} + \frac{2}{7}e^{5t} + \frac{1}{7}e^{8t}$$

we see that

$$p(2) = \frac{1}{7},\ p(3) = \frac{3}{7},\ p(5) = \frac{2}{7},\ p(8) = \frac{1}{7}.$$

Notice that the sum of this factors is indeed 1 and we do have a pmf.

$$\begin{aligned}
m'_X(t) &= \frac{1}{7}\left(2e^{2t} + 9e^{3t} + 10e^{5t} + 8e^{8t}\right) \\
\mathbb{E}X &= m'_X(0) = \frac{1}{7}(2 + 9 + 10 + 8) = \frac{29}{7} \\
m''_X(t) &= \frac{1}{7}\left(4e^{2t} + 27e^{3t} + 50e^{5t} + 64e^{8t}\right) \\
\mathbb{E}X^2 &= m''_X(0) = \frac{1}{7}(4 + 27 + 50 + 64) = \frac{145}{7}
\end{aligned}$$

$$Var(X) = \frac{145}{7} - \left(\frac{29}{7}\right)^2 \approx 3.55$$

The following example is used in Problem 11.3.3 below. Please read!

Example 11.3.7. *(Wald's Equations) Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed random variables. That is, they all have the same distribution as X_1. Let N be a nonnegative random variable that takes integer values. valued random variable and that it is independent of the sequence $X_1, X_2, \ldots$. Let $S_N = \sum_{k=1}^{N} X_k$. Then*

$$\mathbb{E}[S_N] = \mathbb{E}[N]\mathbb{E}[X_1] \tag{11.3.1}$$

and

$$\mathbb{E}[S_N^2] = \mathbb{E}[N(N-1)]\left(\mathbb{E}[X_1]\right)^2 + \mathbb{E}[N]\mathbb{E}[X_1^2] \tag{11.3.2}$$

Solution. *We find $m_{S_N}(t)$ and compute its first and second derivative.*

$$\begin{aligned}
\mathbb{E}\exp\left(t\sum_{k=1}^{N} X_k \Big| N = n\right) &= \mathbb{E}\exp\left(t\sum_{k=1}^{n} X_k \Big| N = n\right) \\
&= \mathbb{E}\exp\left(t\sum_{k=1}^{n} X_k\right) \\
&= (m_{X_1}(t))^n
\end{aligned}$$

Thus,

$$\mathbb{E}\left(e^{tS_N}\Big|N\right) = (m_{X_1}(t))^N$$

and

$$m_{S_N}(t) = \mathbb{E}[m_{X_1}(t))^N]$$

Differentiating gives

$$m'_{S_N}(t) = \mathbb{E}[N(m_{X_1}(t))^{N-1} m'_{X_1}(t)]$$

Recall that for any random variable X, $m_X(0) = 1$

Thus

$$\begin{aligned} \mathbb{E}[S_N] &= \mathbb{E}\left(N(m_{X_1}(0))^{N-1} m'_{X_1}(0)\right) \\ &= \mathbb{E}[N\mathbb{E}(X_1)] \\ &= \mathbb{E}[N]\mathbb{E}[X_1] \end{aligned} \tag{11.3.3}$$

Differentiating the one more time (using the product rule) we get

$$m''_{S_N}(t) = \mathbb{E}[N(N-1)(m_{X_1}(0))^{N-2}\left(m'_{X_1}(t)\right)^2 + N(m_{X_1}(t))^{N-1} m''_{X_1}(t)]$$

$$\begin{aligned} \mathbb{E}[S_N^2] = m''_{S_N}(0) &= \mathbb{E}[N(N-1)\left(\mathbb{E}[X_1]\right)^2] + \mathbb{E}[N\mathbb{E}(X_1^2)] \\ &= \mathbb{E}[N(N-1)]\left(\mathbb{E}[X_1]\right)^2 + \mathbb{E}[N]\mathbb{E}[X_1^2] \end{aligned} \tag{11.3.4}$$

Note that in the previous equations we used the fact that $\mathbb{E}[X_1]$ and $\mathbb{E}[X_1^2]$ are constants to take them out of the expectations.

Example 11.3.8. *Continuing as in the previous Example. Suppose $X_i's$ are independent Bernoulli random variables with*

$$\mathbb{P}(X_i = 1) = \mathbb{P}(X_i = -1) = \frac{1}{2}.$$

Then $\mathbb{E}(X_1) = 0$, $\mathbb{E}X_1^2 = 1$ and by (11.3.1) and (11.3.2)

$$\mathbb{E}[S_N] = \mathbb{E}[N]\mathbb{E}[X_1] = 0 \tag{11.3.5}$$

and

$$\mathbb{E}S_N^2 = \mathbb{E}[N(N-1)](\mathbb{E}[X_1])^2 + \mathbb{E}[N]\mathbb{E}[X_1^2] = \mathbb{E}N \tag{11.3.6}$$

Example: Simple symmetric random walk

With the $X_i's$ as in the previous example, set $S_0 = 0$ and for $n \geq 1$, define

$$S_n = \sum_{i=0}^{n} X_i.$$

This is called a "simple symmetric random walk" on the integers. We start at 0. We flip a coin and if heads we move to the right one unit and if tails we move to the left one unit. Once we are at the new location we flip a coin again and repeat. Thus S_n is our location on the integers after n flips. Let a and b be two positive integers and let N be the first time we reach either $-a$ or b. Now, N is not quite independent of X_i but as above it holds that we have (11.3.5) and (11.3.6). That is, if

$$S_N = \sum_{i=1}^{N} X_i$$

where the X_i are Bernoulli's as above and N the first n such that $S_n = -a$ or $S_n = b$, we have

$$\boxed{\mathbb{E}S_N = 0, \quad \text{and } \mathbb{E}S_N^2 = \mathbb{E}N} \tag{11.3.7}$$

Note that the random variable S_N takes on only two values, $-a$ and b,

Problem 11.3.3. *Continuing with the symmetric random walk setting, use (11.3.7) to compute*

(i) $\mathbb{P}(S_N = -a)$ *and* $\mathbb{P}(S_N = b)$.

(ii) *Using your answer for (i), compute* $\mathbb{E}N$.

(iii) *From (ii), answer the following question "A person is playing a game. The person stands at the 20th meter of a 100-meter long bridge. They walk backward or forwards one meter at a time. And for each step (of 1 meter) they flip a coin: tails means backward, heads means forwards. The person stops when they reach either the beginning or the end of the bridge. What is the expected number of steps to reach either side? Note: for this problem you may set the origins (zero) at the 20th meter mark.*

11.4 The Central Limit Theorem

The central limit theorem is central result in probability theory and the cornerstone of may of its applications to so many different areas. It says that the sum of independent random variables arising from the same distribution is approximately a normal random variable, regardless of the starting distribution. This universality is one of the reasons the normal distribution plays such an important role in probability, it is "universal" in the sense that "everything tends to be normal in the long run."

Theorem 11.4.1: Central Limit Theorem (CLT)

Let $X_1, X_2, X_3 \ldots$ be independent and identically distributed i.i.d. with mean μ and variance σ^2. Set $S_n = \sum_{k=1}^{n} X_k$. This is called the partial sum of the random variables. Then the distribution of

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} = \frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}}$$

tends to the standard normal Z as $n \to \infty$. That is, for every real number a,

$$\lim_{n\to\infty} \mathbb{P}\left(a \leq \frac{S_n - n\mu}{\sigma\sqrt{n}} \leq b\right) = \mathbb{P}\left(a \leq Z \leq b\right) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-x^2/2} dx$$

Proof. The proof of this result used the moment generating functions. It is based on the fact that distributions of a sequence of random variables converge to that of another random variable if and only if the moment generating functions converge for every t. It suffices to consider $\mu = 0, \sigma = 1$. Once this is done, we apply the theorem to $\overline{S_n} = \frac{S_n - n\mu}{\sigma}$. Then we must show that

$$\lim_{n\to\infty} m_{\frac{1}{\sqrt{n}}S_n}(t) = m_Z(t)$$

where $Z \sim N(0,1)$. First, observe that for any constant c and any random variable X,

$$m_{cX}(t) = \mathbb{E}[e^{t(cX)}] = m_X(ct).$$

We now use the independence of the X_j and the fact that they have same distribution to obtain

$$\begin{aligned} m_{\frac{1}{\sqrt{n}}S_n}(t) &= m_{\frac{1}{\sqrt{n}}(X_1 + \cdots + X_n)}(t) \\ &= m_{\frac{1}{\sqrt{n}}X_1}(t) \cdots m_{\frac{1}{\sqrt{n}}X_n}(t) \\ &= \left(m_{\frac{1}{\sqrt{n}}X_1}(t)\right)^n \\ &= \left(m_{X_1}(\frac{t}{\sqrt{n}})\right)^n \end{aligned}$$

Thus we must show that

$$\lim_{n\to\infty} \left(m_{X_1}(\frac{t}{\sqrt{n}})\right)^n = m_Z(t) = e^{t^2/2}$$

Taking natural log of both sides this is equivalent to showing that

$$\lim_{n\to\infty} n \ln\left(m_{X_1}(\frac{t}{\sqrt{n}})\right) = \lim_{n\to\infty} \frac{\ln\left(m_{X_1}(\frac{t}{\sqrt{n}})\right)}{\frac{1}{n}} = t^2/2.$$

Using $m_X(0) = 1$, we see that for any t,

$$\lim_{n\to\infty} m_{X_1}\left(\frac{t}{\sqrt{n}}\right) = m_X(0) = 1$$

which gives that

$$\lim_{n\to\infty} \ln\left(m_{X_1}(\frac{t}{\sqrt{n}}\right) = 0 \text{ and of course } \lim_{n\to\infty} \frac{1}{n} = 0.$$

So we can apply L'Hospital rule. We now use the fact that $m'_{X_1}(0) = \mathbb{E}[X] = 0$ and $m''_{X_1}(0) = \mathbb{E}[X^2] = 1$ and o apply L'Hospital. Let $y = \frac{1}{\sqrt{n}}$. Then (differentiate with respect to y as this is the variable in the application of L'Hospital!)

$$\begin{aligned}
\lim_{n\to\infty}\left(\frac{\ln\left(m_{X_1}(\frac{t}{\sqrt{n}}\right)}{\frac{1}{n}}\right) &= \lim_{y\to 0}\left(\frac{\ln(m_{X_1}(yt))}{y^2}\right) \\
&= \lim_{y\to 0}\left(\frac{\left(\frac{m'_X(yt)t}{m_X(yt)}\right)}{2y}\right), \quad \left(\text{also } \frac{0}{0}\right) \\
&= \lim_{y\to 0}\left(\frac{\left(\frac{m''_X(yt)t^2 m_X(yt)+(m'_X(yt)t)^2}{(m_X(yt))^2}\right)}{2}\right) \\
&= \left(\frac{\left(\frac{m''_X(0)t^2 m_X(0)+(m'_X(0)t)^2}{(m_X(0))^2}\right)}{2}\right) = t^2/2,
\end{aligned}$$

as we wanted. ■

Properties 11.4.1: The CLT is often used as follows:

If $X_1, X_2, \ldots$ are i.i.d. with mean μ and variance σ^2, then for large n,

$$\mathbb{P}\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \le a\right) = \mathbb{P}\left(\frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} \le a\right) \approx \mathbb{P}\left(Z \le a\right) = \Phi(a).$$

The larger the n the better the approximation.

11.4.1 Continuity correction

Used for in approximations.

Remark: Important: CLT approximation applied to discrete r.v

As pointed our earlier, when we apply the Central Limit Theorem to sums of discrete random variables we use the ±0.5 *continuity correction* as in the first example below.

11.4.2 Examples

Example 11.4.1. *If 10 fair dice are rolled, find the approximate probability that the sum obtained is between 30 and 40, inclusive. That is, let Let X_i denote the value of the ith die. Find the approximate value for*

$$\mathbb{P}\left(30 \leq \sum_{i=1}^{10} X_i \leq 40\right).$$

Solution. *Recall that*

$$\mathbb{E}\left(X_i\right) = \frac{7}{2} \quad Var(X_i) = \frac{35}{12}.$$

Take

$$X = X_1 + \cdots + X_{10}$$

Using the CLT we need with $n = 10$,

$$\begin{aligned} n\mu &= 10 \cdot \frac{7}{2} = 35 \\ \sigma\sqrt{n} &= \sqrt{\frac{350}{12}} \end{aligned}$$

thus using the continuity correction, then

$$\begin{aligned} \mathbb{P}\left(30 \leq \sum_{i=1}^{10} X_i \leq 40\right) &= \mathbb{P}\left(29.5 \leq X \leq 40.5\right) \\ &= \mathbb{P}\left(\frac{29.5 - 35}{\sqrt{\frac{350}{12}}} \leq \frac{X - 35}{\sqrt{\frac{350}{12}}} \leq \frac{40.5 - 35}{\sqrt{\frac{350}{12}}}\right) \\ &\approx \mathbb{P}\left(-1.0184 \leq Z \leq 1.0184\right) \\ &= \Phi\left(1.0184\right) - \Phi\left(-1.0184\right) \\ &= 2\Phi\left(1.0184\right) - 1 = .692. \end{aligned}$$

Remark 11.4.1

Nothing needs to be done, in terms of "continuity correction," for continuous random variables.

Example 11.4.2. *Let X_i, $i = 1, 2, \ldots 10$ be independent random variables each uniformly distributed on $(0, 1)$. Find an approximation for*

$$\mathbb{P}\left(\sum_{j=1}^{10} X_j > 6\right).$$

Solution. *We apply the CLT with $n = 10$, $\mu = \mathbb{E}(X_i) = \frac{1}{2}$ and $Var(X) = \frac{1}{12}$ to get*

$$\begin{aligned}
\mathbb{P}\left(\sum_{j=1}^{10} X_j > 6\right) &= \mathbb{P}\left(\frac{\sum_{j=1}^{10} X_j - 5}{\sqrt{10\frac{1}{12}}} > \frac{6-5}{\sqrt{10\frac{1}{12}}}\right) \\
&= \mathbb{P}\left(\frac{\sum_{j=1}^{10} X_j - 5}{\sqrt{10\frac{1}{12}}} > \frac{6-5}{\sqrt{10\frac{1}{12}}}\right) \\
&= \mathbb{P}\left(\frac{\sum_{j=1}^{10} X_j - 5}{\sqrt{\frac{5}{6}}} > \frac{1}{\sqrt{\frac{5}{6}}}\right) \\
&\approx 1 - \Phi(\sqrt{1.2}) \approx 0.1367
\end{aligned}$$

Example 11.4.3. *An instructor has 50 exams that will be graded in sequence. The times required to grade the 50 exams are independent, with a common distribution that has mean 20 minutes and standard deviation 4 minutes. Approximate the probability that the instructor will grade at least 25 of the exams in the first 450 minutes of work.*

Solution. *Let X_i be the time that it takes to grade exam i. The time that takes to grade the first 25 exams is $S_{25} = \sum_{i=1}^{25} X_i$. We need to approximate $\mathbb{P}(S_{25} \le 450)$, $n = 25$, $\mu = 20$, $\sigma^2 = 16$. $n\mu = 25 \cdot 20 = 500$, $\sigma\sqrt{n} = 4 \cdot 5 = 20$*

$$\begin{aligned}
\mathbb{P}\left(\frac{S_{25} - 500}{20} \le \frac{450 - 500}{20}\right) &\approx \mathbb{P}(Z \le -2.5) \\
&= \mathbb{P}(Z \ge 2.5) = 1 - \Phi(2.5) \approx 0.006
\end{aligned}$$

Chapter 12

Week 15: Two useful inequalities

Inequalities play a big role in all areas of mathematics where one is very often only able to estimate quantities rather than compute their exact value. For example, suppose we know the expectation of a random variable but nothing else. Can we estimate what is the largest possible value of $\mathbb{P}(|X| > \lambda)$ for $\lambda > 0$? For example, what is the probability that the r.v. deviates from its mean by a prescribed amount? Such inequalities are knows as the Markov's inequality and Chebyshev's inequality. In fact, as we shall see, the latter follows trivially from the former which itself is (almost) trivial. But despite their "trivial" nature, these inequalities are extremely useful.

12.1 Markov's and Chebyshev's inequalities

Proposition 12.1.1: Markov's inequality

Let X be any random variable. Then for any $\lambda > 0$ and $p > 0$,

$$\mathbb{P}(|X| > \lambda\} \leq \frac{1}{\lambda}\mathbb{E}|X|.$$

Remark 12.1.1: The p-version of Markov's inequality

If $p > 0$ is any positive number, then applying Markov's inequality to the random variable $|X|^p$ and the new λ^p gives that

$$\mathbb{P}(|X| > \lambda) = \mathbb{P}(|X|^p > \lambda^p) \leq \frac{1}{\lambda^p}\mathbb{E}|X|^p$$

which is very useful as we will see below.

Proof. Fix λ in the statement of the Proportion. Let $A = \{|X| > \lambda\}$ and define the random variable by

$$\mathbb{1}_A(\omega) = \begin{cases} 1, & \omega \in A \\ 0, & \omega \in A^c \end{cases}$$

Recall that $\mathbb{P}(A) = \mathbb{E}\mathbb{1}_A$ and so $\lambda\mathbb{P}(A) = \mathbb{E}\lambda\mathbb{1}_A$. However, for $\omega \in A$, $\lambda \leq |X|(\omega)$. This gives that

$$\lambda\mathbb{P}(A) = \mathbb{E}\lambda\mathbb{1}_A \leq \mathbb{E}(|X|\mathbb{1}_A) \leq \mathbb{E}|X|,$$

which is the asserted statement. ■

Remark

The above inequalities are only meaningful if the the parameter λ is large relative to the $\mathbb{E}|X|$. For if $\lambda < \mathbb{E}|X|$, then the inequality is trivial since any probability is at most 1. But the important point to be made here is that this bounds apply to all r.v. regardless of their distributions. They also apply equally well to discrete and continuous random variables.

Applying the inequality to the random variable $Y = (X - \mathbb{E}X)$ with $p = 2$ gives that

Corollary 12.1.1: Chebyshev's inequality

Let X be any random variable with variance σ^2. Then for any $\lambda > 0$ and $p > 0$,

$$\begin{aligned} \mathbb{P}(|X - \mathbb{E}X| > \lambda\} &\leq \frac{1}{\lambda^2}\mathbb{E}|X - \mathbb{E}X|^2 \\ &= \frac{1}{\lambda^2}\mathbb{E}(X - \mathbb{E}X)^2 \\ &= \frac{\sigma^2}{\lambda^2} \end{aligned}$$

Taking $\lambda = k\sigma$ gives

$$\mathbb{P}(|X - \mathbb{E}X| > k\sigma) \leq \frac{\sigma^2}{k^2\sigma^2} = \frac{1}{k^2}$$

That is, for any random variable X, the probability that X differs from its mean by more than a k SD's is no more than $1/k^2$.
Taking complements we get

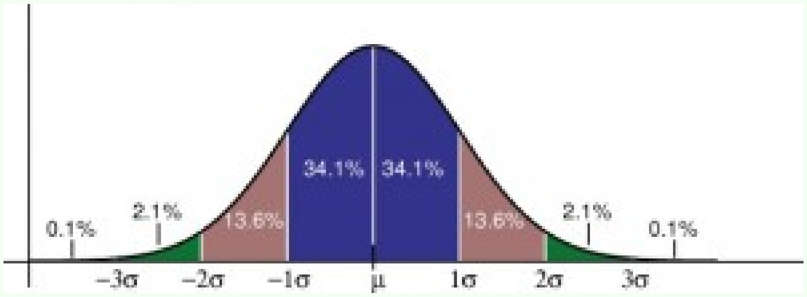
$$\mathbb{P}(|X - \mathbb{E}X| \leq k\sigma) = 1 - \mathbb{P}(|X - \mathbb{E}X| \geq k\sigma) \geq (1 - \frac{1}{k^2}) = \frac{k^2 - 1}{k^2}$$

Remark

It is important to emphasize here that these inequalities are universal in that they apply to <u>ALL</u> random variables without any assumptions about their distributions. If we have specific knowledge of the distributions we may be able to do better. For example, recall the "The $68 - 95 - 99.7$ Rule."

Properties 12.1.1

$X \sim \mathcal{N}\left(\mu, \sigma^2\right)$



"The $68 - 95 - 99.7$ Rule"

Chebyshev's inequality gives for $Z \sim N(\mu, \sigma^2)$,

$$\mathbb{P}\left(|X - \mu| > 2\sigma\right) \leq \frac{1}{4} = 0.25$$

where as the the approximate value is

$$\mathbb{P}\left(|X - \mu| > 2\sigma\right) \approx 0.044$$

or even a little more precise if you use the table to get

$$\mathbb{P}\left(\left|\frac{X - \mu}{\sigma}\right| > 2\right) = 2[1 - \Phi(2)] \approx 0.0456$$

Here is another application of Chebyshev's inequality:

Example 12.1.1. *Suppose X is a random variable with $Var(X) = \sigma^2 = 0$. Then outside an event of probability zero (which does not change the distribution of the random variable in any way), X is constant. In fact, $X = \mathbb{E}X$.*

Solution. *For every integer $n \geq 1$, we have*

$$\mathbb{P}\left(|X - \mathbb{E}X| > \frac{1}{n}\right) \leq n^2 \mathbb{E}|X - \mathbb{E}X|^2 = 0$$

Now, if $|X - \mathbb{E}X| > 0$ then there is and $n \geq 1$ such that $|X - \mathbb{E}X| > \frac{1}{n}$ and

$$(|X - \mathbb{E}X| > 0) = \bigcup_{n=1}^{\infty} \left(|X - \mathbb{E}X| > \frac{1}{n}\right)$$

and

$$\mathbb{P}\left(|X - \mathbb{E}X| > 0\right) = \sum_{n=1}^{\infty} \mathbb{P}\left(|X - \mathbb{E}X| > \frac{1}{n}\right) = 0$$

This means

$$\mathbb{P}\left(|X - \mathbb{E}X| = 0\right) = 1$$

12.2 Laws of Large Numbers

Theorem 12.2.1: Weak Law of Large numbers

Let $X_1, X_2, \dots$ be a sequence of independent and identically distributed random variables with $\mu = \mathbb{E}[X_i]$ and $\sigma^2 = \mathrm{Var}(X_i)$. Then for any number $r > 0$,

$$\mathbb{P}\left(\left|\frac{S_n}{n} - \mu\right| > r\right) \to 0, \quad \text{as } n \to \infty. \tag{12.2.1}$$

This is the same as saying that for any number $r > 0$, no matter how small,

$$\mathbb{P}\left(\left|\frac{S_n}{n} - \mu\right| \leq r\right) \to 1, \quad \text{as } n \to \infty.$$

This is read as

$$\frac{S_n}{n} = \frac{X_1 + X_2 + \cdots + X_n}{n} \to \mu, \quad \text{as } n \to \infty \tag{12.2.2}$$

in probability.

The weak law of large numbers was first proved for Bernoulli random variables by James Bernoulli. The proof was published in 1713 several years after he died 1713 by his nephew Nicholas Bernoulli. Bernoulli had to resort to a quite ingenious proof to establish the result. The general form of the weak law of large numbers presented in Theorem 2.1 was proved by the Russian mathematician Khintchine.

Proof. We give the Proof of the Weal Law since it follows easily from Chebyshev's inequality and we have all the tools to do it. We start by noticing that

$$\frac{S_n}{n} - \mu = \frac{S_n - n\mu}{n} = \frac{S_n - E[S_n]}{n}$$

Thus,

$$\begin{aligned}
\mathbb{E}\left(\frac{S_n - E[S_n]}{n}\right)^2 &= \frac{1}{n^2}\mathrm{Var}\left(S_n\right) \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) = \frac{1}{n^2}n\sigma^2, \quad \text{(from i.i.d.)} \\
&= \frac{\sigma^2}{n}
\end{aligned}$$

Applying Chebyshev's inequality we have that

$$\mathbb{P}\left(\left|\frac{S_n}{n} - \mu\right| > r\right) \leq \frac{\sigma^2}{nr^2} \to 0, \quad \text{as } \; r \to \infty.$$

This competes the proof. ■

The strong law of large numbers one of the best known and widely used results in probability. It states that the average of a sequence of independent random variables having a common distribution will, with probability 1, converge to the mean of that distribution.

Theorem 12.2.2: The strong Law of Large numbers

Let $X_1, X_2, \dots$ be a sequence of independent and identically distributed random variables (i.i.d.) with $\mu = \mathbb{E}[X_i] = \mathbb{E}[X_1]$. Then with probability 1,

$$\frac{S_n}{n} = \frac{X_1 + X_2 + \cdots + X_n}{n} \to \mu, \quad \text{as} \quad n \to \infty \tag{12.2.3}$$

Remark 12.2.1

It is very important to note here that “convergence in probability” (as in the Weak Law) is not the same as ”convergence with probability 1” (as in the Strong Law). The strong law implies the Weak Law but the Weak Law does not imply the Strong Law. In this course we will not elaborate further on the difference.

Chapter 13

Revisit to Chapter 9 after CLT

Recall that the main (in fact only) point of Chapter 9 was the following Theorem.

Theorem

If S_n is a binomial with parameter n and p, then

$$\mathbb{P}\left(a \leq \frac{S_n - np}{\sqrt{np\,(1-p)}} \leq b\right) \to \mathbb{P}\left(a \leq Z \leq b\right) \tag{13.0.1}$$

as $n \to \infty$ where $Z \sim \mathcal{N}(0,1)$.

Or, if S_n denotes the number of successes that occur when n independent trials, each resulting in a success with probability p, are performed, then, for any $a < b$,

$$\mathbb{P}\left(a \leq \frac{S_n - np}{\sqrt{np\,(1-p)}} \leq b\right) \to \mathbb{P}\left(a \leq Z \leq b\right) = \Phi(b) - \Phi(a). \tag{13.0.2}$$

Now that we have the Central Limit Theorem we can view this in a different and better light. Indeed, since a binomial random variable of parameters (n, p) is the sum of independent Bernoulli random variables with parameter p we see that $S_n = \sum_{i=1}^{n} X_i$, where the $X_i's$ are i.i.d. with Bernoulli p, $\mathbb{P}(X_i = 1) = p$, $\mathbb{P}(X_i = 0) = 1 - p$. Since the mean $\mu = \mathbb{E}[X] = p$ and $\operatorname{Var}(X) = p(1-p)$, the CLT immediately gives the above Theorem.

When approximating the binomial with the normal (and other discrete such discrete r.v.'s) we need to take the $\pm 0.5 = \pm 1/2$ *continuity correction* into account. This gives

If S_n is a binomial random variable with $S_n \sim Bin(n,p)$, then

$$\begin{aligned} \mathbb{P}\left(a \le S_n \le b\right) &= \mathbb{P}\left(a - \frac{1}{2} \le S_n \le b + \frac{1}{2}\right) \\ &= \mathbb{P}(\left(\frac{a - \frac{1}{2} - n\mu}{\sqrt{np(1-p}} \le \frac{S_n - n\mu}{\sqrt{np(1-p}} \le \frac{b + \frac{1}{2} - n\mu}{\sqrt{np(1-p}}\right) \\ &\approx \Phi\left(\frac{b + \frac{1}{2} - n\mu}{\sqrt{np(1-p}}\right) - \Phi\left(\frac{a - \frac{1}{2} - n\mu}{\sqrt{np(1-p}}\right) \end{aligned} \tag{13.0.3}$$

If we wanted to approximate $\mathbb{P}\left(a < S_n < b\right) = \mathbb{P}\left(a + 1 \le S_n \le b - 1\right)$ we will use

$$\mathbb{P}\left(a < S_n < b\right) = \mathbb{P}\left(a + \frac{1}{2} \le S_n \le b - \frac{1}{2}\right) \tag{13.0.4}$$

$$\approx \Phi\left(\frac{b - \frac{1}{2} - n\mu}{\sqrt{np(1-p}}\right) - \Phi\left(\frac{a + \frac{1}{2} - n\mu}{\sqrt{np(1-p}}\right) \tag{13.0.5}$$

Other cases, such as $P\left(a \le S_n < b\right)$, are handled similarly.

Using the fact that the sum of independent Poisson X_i with parameters λ_i, $i = 1, 2, \dots, n$, is a Poisson with parameter $\lambda_1 + \lambda_2 + \cdots + \lambda_n$, the same argument shows if $Y_\lambda = \sum_{i=1}^{\lambda} X_i$ where $X_i \sim Pois(1)$, λ a positive integers, then

Theorem

$$\mathbb{P}\left(a \le \frac{Y_\lambda - \lambda}{\sqrt{\lambda}} \le b\right) \to \mathbb{P}\left(a \le Z \le b\right), \tag{13.0.6}$$

as $\lambda \to \infty$.

This gives the approximation as with the binomial above:

If Y_λ is a Poisson random variable with parameter λ, then

$$\mathbb{P}\left(a \le Y_\lambda \le b\right) \approx \Phi\left(\frac{b + \frac{1}{2} - \lambda}{\sqrt{\lambda}}\right) - \Phi\left(\frac{a - \frac{1}{2} - \lambda}{\sqrt{\lambda}}\right) \tag{13.0.7}$$

and

$$\mathbb{P}\left(a < Y_\lambda < b\right) \approx \Phi\left(\frac{b - \frac{1}{2} - \lambda}{\sqrt{\lambda}}\right) - \Phi\left(\frac{a + \frac{1}{2} - \lambda}{\sqrt{\lambda}}\right) \tag{13.0.8}$$

Example 13.0.1. *Although many airline customers bypass ticket counters these days, from tracking data it has been determined that the number of people served at the ticket counter of American Airlines at the Indianapolis airport from 5:00 am to 12:00 noon daily is Poisson with parameter* $\lambda = 1600$*. Estimate the* *probability that on a certain day the number of customers during these hours is strictly between 1550 and 1650*

Solution. *With X denoting the number of customers (always a nonnegative integer) we have $\mathbb{E}[X] = 1600$ and also $Var(X) = 1600$. We want $\mathbb{P}(1550 < X < 1650)$ (or equivalently $\mathbb{P}(1551 \le X \le 1649)$). The continuity correction gives:*

$$\begin{aligned}
\mathbb{P}(1550 < X < 1650) &= \mathbb{P}(1550.5 \le X \le 1649.5) \\
&= \mathbb{P}\left(\frac{1550.5 - 1600}{\sqrt{1600}} \le \frac{X - 1600}{\sqrt{1600}} \le \frac{1649.5 - 1600}{\sqrt{1600}}\right) \\
&= \mathbb{P}\left(\frac{-49.5}{40} \le \frac{X - 1600}{\sqrt{1600}} \le \frac{49.5}{40}\right) \\
&\approx \Phi\left(\frac{49.5}{40}\right) - \Phi\left(\frac{-49.5}{40}\right) \\
&= \Phi(1.2375) - \Phi(-1.2375) \approx 0.782
\end{aligned}$$

13.1 Examples of applications of Markov's and Chebyshev's inequalities

Recall the inequalities

Properties 13.1.1

Let X be any random variable. For any $\lambda > 0$ and $p > 0$,

(i) Markov's inequality:

$$\boxed{\mathbb{P}(|X| > \lambda) = \mathbb{P}(|X|^p > \lambda^p) \leq \frac{1}{\lambda^p}\mathbb{E}|X|^p,}$$

The case $p = 1$ gives the general case $p > 0$ as explained in Lecture Notes for week 14 and video for today's lecture.

(ii) Chebyshev's inequality: Applying Markov with $Y = |X - \mu|$ and $p = 2$ gives

$$\mathbb{P}(|X - \mathbb{E}X| > \lambda\} \leq \frac{\sigma^2}{\lambda^2}$$

Taking $\lambda = k\sigma$, any $k > 0$, gives

$$\boxed{\mathbb{P}(|X - \mathbb{E}X| > k\sigma) \leq \frac{\sigma^2}{k^2\sigma^2} = \frac{1}{k^2}} \tag{13.1.1}$$

Taking complements gives

$$\mathbb{P}(|X - \mathbb{E}X| \leq k\sigma) = 1 - \mathbb{P}(|X - \mathbb{E}X| \geq k\sigma) \geq (1 - \frac{1}{k^2}) = \frac{k^2 - 1}{k^2}$$

That is,

$$\boxed{\mathbb{P}(|X - \mathbb{E}X| \leq k\sigma) \geq \frac{k^2 - 1}{k^2}} \tag{13.1.2}$$

13.2 Applications of Markov's inequality

The real usefulness of these inequalities is that they apply for any random variables regardless of their distributions.

Example 13.2.1. *Consider an exam in a random class with average of 74%. What is the maximum probability that a randomly chosen student's score is 80% or higher?*

Solution. *Apply Markov's inequality with $\lambda = 80$,*

$$\mathbb{P}(X > 80) \leq \frac{1}{80}\mathbb{E}[X] = \frac{74}{80} = 0.925$$

Remark 13.2.1

Notice that if we use Markov's inequality to get a bound for the probability that a student's score is greater than λ where λ is anything smaller than 74 we just get that it is less than or equal to 1. This, while correct, does not give any useful information. For example, suppose instead of %80 we pick %60. Then we obtain

$$\mathbb{P}(X > 60) \leq \frac{1}{60}\mathbb{E}[X] = \frac{74}{60} \approx 1.233$$

from which we can only conclude that the probability is less than or equal to one. Not useful information as this is always the case with all probabilities.

Example 13.2.2. *The average salary of a math professor at university A is $78,000 per year. What is an upper bound on the probability that a randomly selected professor makes $110,000?*

Solution. *Apply Markov's inequality with $\lambda = 110,000$,*

$$\mathbb{P}(X > 110,000) \leq \frac{1}{110,000}\mathbb{E}[X] = \frac{77,000}{110,000} = 0.7$$

Of course, this is an upper bound, with further information on the distribution, the probability could be found to be much lower.

Example 13.2.3. *The average speed on a section of Interstate 65 is 50 miles per hour although the speed limit is 65. Find a bound on the probability that a random driver is driving 10 miles above the speed limit.*

Solution.

$$\mathbb{P}(X > 75) \leq \frac{1}{75}\mathbb{E}[X] = \frac{50}{75} \approx 0.666$$

Example 13.2.4. *At a certain Ivy League school the average number of students in a class at 30. Without any further information what can we say about the the largest possible probability that a class selected at random will have more than 40 students?*

Solution.

$$\mathbb{P}\left(X > 40\right) \leq \frac{1}{40}\mathbb{E}[X] = \frac{30}{40} = \frac{3}{4}$$

The maximum probability will not exceed $\frac{3}{4}$.

Remark 13.2.2

In all these example we are not calculating or approximating the probabilities. We are simply getting bounds on the probabilities. Although there are more sophisticated methods to find bounds, and especially if we have some more information available, these simple calculations are often useful to rule out the "I am absolutely 100 percent certain that this or that has happened" while we have very little factual information.

13.2.1 Applications of Chebyshev's inequality

Example 13.2.5. *In Example 13.2.1 with the class average 74%, suppose we also know that the standard deviation $\sigma = 10$. Find a lower bound that the probability of a randomly chosen student's score is between 58% and 90%.*

Solution. *Apply Chebyshev's inequality* (13.1.2) *to get*

$$\begin{aligned}\mathbb{P}(58 \leq X \leq 90) &= \mathbb{P}(|X - 74| \leq 16) \\ &= \mathbb{P}(|X - 74| \leq 1.6\sigma) \\ &\geq \frac{(1.6)^2 - 1}{(1.6)^2} = \frac{1.56}{2.56} \approx 0.609\end{aligned}$$

Example 13.2.6. *In Example 13.2.2 with the average salary of a math professor at university A of $78,000 per year, assume the standard deviation is $\sigma = \$8,000$, find an upper bound on the probability that a random professor makes less than $48,000 or more than $108,000.*

Solution. *We need an upper estimate on $\mathbb{P}(X < 48,000$ or $X > 108,000)$. Apply Shebyshev's inequality* (13.1.1).

$$\begin{aligned}\mathbb{P}(\mathbb{P}(X < 48,000 \text{ or } X > 108,000) &= \mathbb{P}\left(|X - 78,000| > \frac{30,000}{8,000}\sigma\right) \\ &= \mathbb{P}(|X - 78,000| > 3.75\sigma) \\ &\leq \frac{1}{(3.75)^2} \approx 0.0711\end{aligned}$$

Example 13.2.7. *In Example 13.2.4 with the Ivy League school with average class size of 30, suppose the standard deviation of the class size is $\sigma = 6$. Calculate a lower estimate of the probability that a random class has between 16 and 44 students.*

Solution. *Apply Chebyshev's inequality* (13.1.2).

$$\begin{aligned}\mathbb{P}(16 \leq X \leq 44) &= \mathbb{P}(|X - 30| \leq 14) \\ &= \mathbb{P}\left(|X - 30| \leq \frac{7}{3}6\right) \\ &\geq 1 - \frac{1}{(7/3)^2} = \frac{40}{49} =\approx 0.816\end{aligned}$$