Probability Theory (Undergraduate)

Yiqiao YIN
Columbia University

October 12, 2021

Abstract

This document is prepared for students in Probability Theory (Undergraduate) course offered at Columbia University in the Department of Statistics. The course instructor is Professor Shaw-Hwa Lo. The document serves students by providing lecture notes as well as homework and exam guidance. I am grateful for Professor Shaw-Hwa Lo for providing materials. It has been extremely helpful for the students in the class as well as producing this notes. In addition, I want to thank the graders for providing comments of the homework assignments. I am the TA for this class. Please email me yy2502columbia.edu if you have any questions.

Contents

1 Combinatorial Analysis

Go back to Table of Contents. Please click TOC

1.1 Introduction

Let us start with an example about an event with multiple possible outcomes. An experiment, for example, can lead to n possible outcomes:

$$a_1, a_2, ..., a_n$$

For each a_i there are m possible outcomes from another experiment, then together there are nm possible outcomes. For example, how many outcomes one possible by tossing a coin twice? We can approach to answer this question from frequentist point of view, which is objective. One can start with a fair coin. Tossing the coin for the first time, one will observe either head or tail. By tossing the coin for the second time, one will observe, assuming using a fair coin, head or tail. One can continue this experiment and will observe, assuming tossing coin twice, the following $\{HH, HT, TH, TT\}$, using "H" for heads and "T" for tails.

It is not always the case that an experiment can be repeated. We collect previous data, called prior. We can observe new data as time moves on and we can use it as new information. We update our prior data with new information and we will arrive a more sophisticated analysis, which is called posterior. This school of thoughts, which is called Bayesian, are usually more subjective due to the fact that there is a prior before analysis starts.

Frequentist and Bayesian are two schools of thoughts in the field of statistics. Frequentist dominated the field in the 60's and 70's. Starting since the past 20 to 30 years, there are a good amount of Bayesian approach emerged.

For this course, we will mostly be dealing with repeatable experiments. We may have different outcomes, but the experiments we will be discussing can be replicated under the same condition.

1.2 Permutation and Combination

Suppose that two experiments are to be performed. Then if experiment 1 can be result in any one of m possible outcomes and if, for each outcome of the experiment 1, there are n possible outcomes of experiment 2, then together there are mn possible outcomes of the two experiments.

Example 1.2.1. A small community consists of 10 women each of whom has 3 children. If one woman and one of her children are to be chosen as mother and child of the year, how many different choices are possible?

Answer. We compute $10 \times 3 = 30$ total possible choices. □

This motivation example leads to the first conclusion. If r experiments that are to be performed are such that the first one may result in any of n_1 possible outcomes; and if, for each of these n_1 possible outcomes, there are n_2 possible outcomes of the second experiment; and if ..., then there is a total of $n_1 \cdot n_2 \dots n_r$ possible outcomes of r experiments.

Let us look at the following example. Consider the word "statistics". How many different letter arrangements would you have? Let us separate these letters and count that there are three s's, three t's, two i's, one a, and one c. In this case we have permutation

$$10! = 10 \times 9 \times \cdots \times 2 \times 1$$

but we need to consider the cases that if you switch two's, example, the results will be the same in the word "statistics". Hence, we need the following

$$3!3!2!1!$$

and together we have

$$\frac{n!}{n_1!\dots n_k!}$$

different ways to arrange n objects, of which $n_1, n_2, ..., n_k$ are alike.

Definition 1.2.2. Suppose now that we have n objects. Reasoning similar to that we have just used for the letters example in lecture then shows that there are

$$n(n-1)(n-2)\dots(3)(2)(1) = n!$$

different permutations of the n objects.

Definition 1.2.3. In general, the same reasoning used in lecture shows that there are

$$\frac{n!}{n_1!n_2!\dots n_r!}$$

different permutations of n objects, of which n_1 are alike, n_2 are alike, ..., etc.

Example 1.2.4. Let us look at another example. How many different groups of 3 can be related from 5 items, A, B, C, D, and E? In this case, There are $\binom{5}{3} = 10$ number of different groups of 3.

Definition 1.2.5. We define $\binom{n}{r}$, for $r \leq n$, by

$$\binom{n}{r} = \frac{n!}{(n-r)!r!}$$

and say that $\binom{n}{r}$ represents the number of possible combinations of n objects taken r at a time.

Remark 1.2.6. By convention, 0! is 1. Thus, $\binom{n}{0} = \binom{n}{n} = 1$. We also take $\binom{n}{i}$ to be equal to 0 when either $i < 0$ or $i > n$.

Example 1.2.7. A committee of 3 is to be formed from a group of 20 people. How many different committees are possible?

Answer. There are $\binom{20}{3} = \frac{20 \cdot 19 \cdot 18}{3 \cdot 2} = 1140$ total possibilities. □

Another interesting example is the following.

Example 1.2.8. For a class of 20, 12 boys and 8 girls. How many different groups consisting of 3 boys and 2 girls can be formed? What if 2 of the boys refuse to be in the same group together?

A useful formula is in the following

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

and the famous binomial formula is

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

Proof. Consider the following

$$\begin{aligned}(x+y)^n &= \underbrace{(x+y)\dots(x+y)}_{n \text{ times}} \\ \text{if } n=2: & \\ (x+y)^2 &= x^2+xy+yx+y^2, \text{(each term contributes once)}\end{aligned}$$

while each x and y contributes $x^k y^{n-k}$. There are n choose k, i.e. $\binom{n}{k}$, number of these contributions and that's why it leads to $\binom{n}{k}x^k y^{n-k}$. □

Example 1.2.9. Ten balls marked 1 to 10 are put into 3 bags A, B, and C with 3 in A, 3 in B, and 4 in bag C. How many ways? Bags are all assumed to be distinct. We have 3! in the first bag, 3! in the second bag, and 4! in the third bag. In total, there are 10! possible outcome for all 10 balls. Thus, the answer is

$$\frac{10!}{3!3!4!}$$

Let us formally introduces binomial theorem.

Theorem 1.2.10. *IMPORTANT The binomial theorem states the following*

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

Proof. When $n=1$, we have

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

Assume the above results hold for $n-1$. Now, consider

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

Letting $i=k+1$ for the first term and $i=k$ for the second term, we have

$$\begin{aligned}(x+y)^n &= \textstyle\sum_{i=1}^{n}\binom{n-1}{i-1}x^i y^{n-i} + \sum_{i=1}^{n}\binom{n-1}{i}x^i y^{n-i} \\ &= x^n + \textstyle\sum_{i=1}^{n-1}\left[\binom{n-1}{i-1}+\binom{n-1}{i}\right]x^i y^{n-i} + y^n \\ &= \textstyle\sum_{i=0}^{n}\binom{n}{i}x^i y^{n-i}\end{aligned}$$

and we are done. □

Example 1.2.11. Expand $(x+y)^3$.

Answer.

$$\begin{aligned}(x+y)^3 &= \textstyle\binom{3}{0}x^0y^3 + \binom{3}{1}x^1y^2 + \binom{3}{2}x^2y^1 + \binom{3}{3}x^3y^0 \\ &= y^3+3xy^2+3x^2x+x^3\end{aligned}$$

□

Example 1.2.12. How many subsets are there of a set consisting of n elements?

Answer. Since there are $\binom{n}{k}$ subsets of size k, the answer is

$$\sum_{k=0}^{n} \binom{n}{k} = (1+1)^n = 2^n$$

□

The multinomial theorem is stated in the following.

Definition 1.2.13. We have the following formula that can be accepted as an identity

$$(x_1 + x_2 + \cdots + x_r)^n = \sum_{(n_1,\ldots,n_r)} \binom{n}{n_1, n_2, ..., n_r} x_1^{n_1} x_2^{n_2} \ldots n_r^{n_r}$$

That is, the sum is over all nonnegative integer-valued vectors $(n_1, ..., n_r)$ such that $\sum_{i=1}^{r} x_i = n$.

Let us introduce the following propositions that may be helpful in this topic.

Proposition 1.2.14. *There are* $\binom{n-1}{r-1}$ *distinct positive integer-valued vectors* $(x_1, ..., x_r)$ *satisfying the equation*

$$x_1 + x_2 + \cdots + x_r = n, x_i > 0, i = 1, ..., r$$

Proposition 1.2.15. *There are* $\binom{n+r-1}{r-1}$ *distinct nonnegative integer-valued vectors* $(x_1, ..., x_r)$ *satisfying*

$$x_1 + x_2 + \cdots + x_r = n$$

Example 1.2.16. How many distinct nonnegative integer-valued solutions of $x_1 + x_2 = 3$ are possible?

Answer. There are $\binom{3+2-1}{2-1} = 4$ such solutions: (0,3), (1,2), (2,1), (3,0). □

Example 1.2.17. An investor has \$10 million dollars to invest in 5 lands. Each land must be in units of \$1 million dollars. (i) If the investor has to invest in all 5 lands (no land get left out), how many different strategies are possible? (ii) What if not all the lands need to be invested?

Answer. Let x_i be each of the investment and here i is the running index taking values $\{1, 2, 3, 4, 5\}$. Since each investment must be in unit of \$1 million, we have the following equation

$$x_1 + x_2 + x_3 + x_4 + x_5 = 10$$

In scenario (i), we must invest in all the lands, so we assume $x_i > 0$. In this case, we use the formula $\binom{n-1}{r-1} = \binom{10-1}{5-1}$. In scenario (ii), we do not have to invest all of the lands. That means we can relax the assumption to $x_i \geq 0$. In this case, we use the formula $\binom{n+r-1}{r-1} = \binom{10+5-1}{5-1}$. □

2 Axioms of Probability

Go back to Table of Contents. Please click TOC

This section introduces the concepts of the probability of an event and then show how probabilities can be computed in certain situations.

2.1 Sample Space and Events

We define event as a specific situation. It is a subset of an outcome defined in an experiment. For example, tossing a fair coin once can result in head or tail. The event can be head or tail. Sample space is the superset of all the possible outcomes. For example, tossing two fair coins together, the sample space consists of $\{HH, HT, TH, TT\}$.

Example 2.1.1. Two draws are made from the following box with 3 balls, call them G, Y, and B.

1. Consider all possible arrangements with replacement. We have $3 \times 3 = 9$. That is, we have

$$S = \{(G,G),(B,B),(Y,Y),(G,B),(G,Y),(B,G),(B,Y),(Y,G),(Y,B)\}$$

It can be anywhere in the following matrix.

1/2	G	Y	B
G	(G,G)	()	(G,B)
Y	()	(Y,Y)	()
B	()	()	(B,B)

Same question without replacement. Then we have $3 \times 2 = 6$. In matrix form, we do not count the diagonal because without replacement means once G is drawn it cannot appear for a second time.

Definition 2.1.2. Event E and event F are said to be mutually exclusive if $E \cap F =$, that is, there is no element that simultaneously exists in E and F.

The operations forming unions, intersections, and complements of events obey certain rules similar to the rules of algebra.

Proposition 2.1.3. *We have the following rules*

1. *Commutative laws:* $E \bigcup F = F \bigcup E$, *or* $EF = FE$
2. *Associative laws:* $(E \bigcup F) \bigcup G = E \bigcup (F \bigcup G)$, *or* $(EF)G = \mathbb{E}(FG)$
3. *Distribution laws:* $(E \bigcup F)G = EG \bigcup FG$ *or* $EF \bigcup G = (E \bigcup G)(F \bigcup G)$

Theorem 2.1.4. *DeMorgan's Law states that*

$$\left(\bigcup_{i=1}^{n} E_i \right)^c = \bigcup_{i=1}^{n} E_i^2$$

$$\left(\bigcap_{i=1}^{n} E_i \right)^c = \bigcap_{i=1}^{n} E_i^2$$

Example 2.1.5. A famous example is to consider event E and F, by DeMorgan's law, we have

$$(E \bigcup F)^c = E^c F^c \text{ and } (EF)^c = E^c \bigcup F^c$$

Proposition 2.1.6. *Let us introduce the following proposition, called inclusion-exclusion identity.*

$$\begin{aligned}\mathbb{P}(E_1 \bigcup E_2 \bigcup \cdots \bigcup E_n) &= \sum_{i=1}^{n} \mathbb{P}(E_i) - \sum_{i_1<i_2} \mathbb{P}(E_{i_1}E_{i_2}) + \dots \\ &+(-1)^{r+1} \sum_{i_1<i_2<\cdots<i_r} \mathbb{P}(E_{i_1}E_{i_2}\dots E_{i_r}) \\ &+ \cdots + (-1)^{n+1}\mathbb{P}(E_1E_2\dots E_n)\end{aligned}$$

Example 2.1.7. This problem is from text[1] page 34. 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?

Solution. Since all of the balls are treated in an identical manner, it follows that the set of k balls selected is equally likely to be any of the $\binom{n}{k}$ sets of k balls. Therefore,

$$\mathbb{P}(\text{special ball is selected}) = \frac{\binom{1}{1}\binom{n-1}{k-1}}{\binom{n}{k}} = \frac{k}{n}$$

We could also have obtained this result by letting A_i denote the event that the special ball is the ith ball to be chosen, $i = 1, ..., k$. Then, since each one of the n balls is equally likely to be the ith ball chosen, it follows that $\mathbb{P}(A_i) = 1/n$. Hence, because these events are clearly mutually exclusive, we have

$$\mathbb{P}(\text{special ball is selected}) = \mathbb{P}\left(\bigcup_{i=1}^{k} A_i\right) = \sum_{i=1}^{k} \mathbb{P}(A_i) = \frac{k}{n}$$

We could also have argued that $\mathbb{P}(A_i) = 1/n$, by noting that there are $n(n-1)\dots(n-k+1) = n!/(n-k)!$ equally likely outcomes of the experiment, of which $(n-1)!/(n-k)!$ result in the special ball being the ith one chosen. From this reasoning, it follows that

$$\mathbb{P}(A_i) = \frac{(n-1)!}{n!} = \frac{1}{n}$$

□

Let us introduce some simple properties (notes from class).

1. If $E \subset F$, then $\mathbb{P}(E) \leq \mathbb{P}(F)$ and $F = E \bigcup (E^c \cap F) \Rightarrow \mathbb{P}(F) \geq \mathbb{P}(E)$
2. $\mathbb{P}(E \cup F) = \mathbb{P}(E) + \mathbb{P}(F) - \mathbb{P}(EF)$ and also $\mathbb{P}(E \bigcup F) = \mathbb{P}(E) + \mathbb{P}(E^cF)$ and $\mathbb{P}(E^cF) = \mathbb{P}(F) - \mathbb{P}(EF)$
3. $$\begin{aligned}\mathbb{P}(\textstyle\bigcup_{i=1}^{n} E_i) &= \sum_{i=1}^{n} \mathbb{P}(E_i) - \sum_{i<j} \mathbb{P}(E_iE_j) + \dots \\ &+(-1)^{k+1} \sum_{i_1<i_2<\cdots<i_k} \mathbb{P}(E_{i1}E_{i2}\dots E_{ik}) \\ &+(-1)^{n+1}\mathbb{P}(E_1E_2\dots E_n)\end{aligned}$$

Sample space with equal chances for all outcome. If S is such a sample space $|S| = n$, the size of S for $w \in S$. This gives us $\mathbb{P}(w) = \frac{1}{n}$. For all $E \subset S$, an event, we have $\mathbb{P}(E) = \frac{|E|}{|S|} = \frac{m}{n}$ if $|E| = m$ and for $m \leq n$.

2.2 Axioms of Probability

Consider an experiment whose sample space is S. For each event E of the sample space S, we assume that a number $\mathbb{P}(E)$ is defined and satisfies the following three axioms

Proposition 2.2.1. *The three axioms of probability*

1. *Axiom 1:* $0 \leq \mathbb{P}(E) \leq 1$
2. *Axiom 2:* $\mathbb{P}(S) = 1$
3. *Axiom 3: For any sequence of mutually exclusive events* $E_1, E_2, ...$ *(that is, events of for which* $E_i E_j = \emptyset$ *when* $i \neq j$*),*

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

We refer to $\mathbb{P}(E)$ *as the probability of the event E.*

Example 2.2.2. If our experiment consists of tossing a coin and if we assume that a head is as likely to appear as a tail, then we have

$$\mathbb{P}(\{H\}) = \frac{1}{2} \text{ and } \mathbb{P}(\{T\}) = \frac{1}{3}$$

However, if the coin were biased and we believed that a head were twice as likely to appear as a tail, we should have

$$\mathbb{P}(\{H\}) = \frac{2}{3} \text{ and } \mathbb{P}(\{T\}) = \frac{1}{3}$$

Let us elaborate the experiment a little in the following example.

Example 2.2.3. If a die is rolled and suppose that all six sides are equally likely to appear, then we have $\mathbb{P}(\{1\}) = \mathbb{P}(\{2\}) = \mathbb{P}(\{3\}) = \mathbb{P}(\{5\}) = \mathbb{P}(\{6\}) = \frac{1}{6}$. From Axiom 3, we can compute the probability of rolling an even number to be

$$\mathbb{P}(\{2, 4, 6\}) = \mathbb{P}(\{2\}) + \mathbb{P}(\{4\}) + \mathbb{P}(\{6\}) = \frac{1}{2}$$

Let us introduce more properties.

Proposition 2.2.4.

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

Proposition 2.2.5. *If* $E \subset F$*, then* $\mathbb{P}(E) \leq \mathbb{P}(F)$*.*

Proposition 2.2.6.

$$\mathbb{P}(E \cup F) = \mathbb{P}(E) + \mathbb{P}(F) - \mathbb{P}(EF)$$

Example 2.2.7. J is taking two books along on her holiday 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?

Answer. Let B_i denote the event that J likes book i, $i = 1, 2$. Then the probability that she likes at least one of the books is

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

Because the event that J likes neither book is the complement of the event that she likes at least one of them, we obtain the result

$$\mathbb{P}(B_1^C B_2^C) = \mathbb{P}((B_1 \cup B_2)^C) = 1 - \mathbb{P}(B_1 \cup B_2) = 0.4$$

We may also calculate the probability that any of of the three events E, F, and G occurs, namely,

$$\mathbb{P}(E \cup F \cup G) = \mathbb{P}((C \cup F) \cup G)$$

which equals

$$\mathbb{P}(E \cup F) + \mathbb{P}(G) - \mathbb{P}((E \cup F) \cap G)$$

Now it follows from the distributive law that the events $(E \cup F)G)$ and $EG \cup FG$ are equivalent; hence, from the preceding equations, we obtain

$$\begin{aligned} \mathbb{P}(E \cup F \cup G) &= \mathbb{P}(E) + \mathbb{P}(F) - \mathbb{P}(EF) + \mathbb{P}(G) - \mathbb{P}(EG \cup FG) \\ &= \mathbb{P}(E) + \mathbb{P}(F) - \mathbb{P}(EF) + \mathbb{P}(G) - \mathbb{P}(EG) - \mathbb{P}(FG) + \mathbb{P}(EGFG) \\ &= \mathbb{P}(E) + \mathbb{P}(F) + \mathbb{P}(G) - \mathbb{P}(EF) - \mathbb{P}(EG) - \mathbb{P}(FG) + \mathbb{P}(EFG) \end{aligned}$$

□

Proposition 2.2.8. *This proposition states that the probability of the union of n events equals the sum of the probabilities of these events taken one at a time, minus the sumo f the probabilities of these events taken two at a time, plus the sum of the probabilities of these events taken three at a time, and so on. In mathematical formula, the proposition states the following identity*

$$\begin{aligned} \mathbb{P}(E_1 \cup E_2 \cup \cdots \cup E_n) &= \textstyle\sum_{i=1}^{n} \mathbb{P}(E_i) - \sum_{i_1 < i_2} \mathbb{P}(E_{i_1} E_{i_2}) \\ &\quad + \cdots + (-1)^{r+1} \textstyle\sum_{i_1 < \ldots i_r} \mathbb{P}(E_{i_1} E_{i_2} \ldots E_{i_r}) \\ &\quad + \cdots + (-1)^{n+1} \mathbb{P}(E_1 E_2 \ldots E_n) \end{aligned}$$

An alternative way is to write as the following:

$$\begin{aligned} & \mathbb{P}(E_1 \cup E_2 \cup \cdots \cup E_n) \\ = \; & \textstyle\sum_{i=1}^{n} \mathbb{P}(E_i) - \sum_{1 \le i < j \le n} \mathbb{P}(E_i \cap E_j) + \sum_{1 \le i < j < k \le n} \mathbb{P}(E_i \cap E_j \cap E_k) + \ldots \\ & + (-1)^n \textstyle\sum_{1 \le j \le n} \mathbb{P}(\bigcap_{i=1, i \ne j}^{n} E_i) + (-1)^{n+1} \mathbb{P}(E_1 \cap E_2 \cap \cdots \cap E_n), \text{ denote as } \triangle \end{aligned}$$

For simplicity, we will refer the above formula as $\triangle$.

Proof. We prove the alternative form, i.e. the equation $\triangle$. We use mathematical induction. We assume that the formula holds for n and we show that it holds for $n+1$. Notation: when we write E_1E_2, we assume the intersection, i.e. $E_1E_2 = E_1 \cap E_2$.

First, let us check the following

$$\begin{aligned} \mathbb{P}(E_1 \cup E_2) &= \mathbb{P}(E_1) + \mathbb{P}(E_2) - \mathbb{P}(E_1E_2) \\ \mathbb{P}(E_1 \cup E_2 \cup E_3) &= \mathbb{P}(E_1) + \mathbb{P}(E_2) + \mathbb{P}(E_3) \\ & \quad - \mathbb{P}(E_1E_2) - \mathbb{P}(E_1E_3) - \mathbb{P}(E_2E_3) + \mathbb{P}(E_1E_2E_3) \end{aligned}$$

and thus we have shown at $n = 2$ and $n = 3$ the formula holds true.

Next, let us assume at n, the formula holds. It suffices to show that the formula also holds true at $n + 1$. Consider the following

$$\begin{aligned} & \mathbb{P}(E_1 \cup E_2 \cup \cdots \cup E_n \cup E_{n+1}) \\ = \; & \mathbb{P}(E_1 \cup E_2 \cup \cdots \cup E_n) + \mathbb{P}(E_{n+1}) \\ & - \big\{ \mathbb{P}((E_1E_{n+1}) \cup (E_2E_{n+1}) \cup \ldots (E_nE_{n+1})) \big\} \end{aligned}$$

where we used the fact that at $n = 2$ the formula holds. To complete the proof, we apply the $\triangle$ twice.

$$
\begin{aligned}
& \mathbb{P}(E_1 \cup E_2 \cup \cdots \cup E_n \cup E_{n+1}) \\
= \; & \underbrace{\mathbb{P}(E_1 \cup E_2 \cup \cdots \cup E_n)}_{\text{apply } \triangle} + \mathbb{P}(E_{n+1}) \\
& \underbrace{-\big\{\mathbb{P}((E_1E_{n+1}) \cup (E_2E_{n+1}) \cup \ldots (E_nE_{n+1}))\big\}}_{\text{apply } \triangle} \\
= \; & \textstyle\sum_{i=1}^{n} \mathbb{P}(E_i) - \sum_{1\le i<j\le n} \mathbb{P}(E_i \cap E_j) + \sum_{1\le i<j<k\le n} \mathbb{P}(E_i \cap E_j \cap E_k) \\
& + \cdots + (-1)^{n+1}\mathbb{P}(E_1 \cap E_2 \cap \cdots \cap E_n) \text{ 1st time applying } \triangle \\
& + \mathbb{P}(E_{n+1}) \\
& - \big\{ \textstyle\sum_{i=1}^{n} \mathbb{P}(E_i \cap E_{n+1}) - \sum_{1\le i<j\le n} \mathbb{P}(E_i \cap E_j \cap E_{n+1}) + \ldots \\
& + (-1)^n \textstyle\sum_{1\le j\le n} \mathbb{P}(\bigcap_{i=1, i\neq j}^{n} E_i \cap E_{n+1} \\
& + (-1)^{n+1}\mathbb{P}(E_1 \cap E_2 \cap \cdots \cap E_n \cap E_{n+1})\big\} \text{ 2nd time applying } \triangle \\
= \; & \textstyle\sum_{i=1}^{n+1} \mathbb{P}(E_i) - \sum_{1\le i<j\le n+1} \mathbb{P}(E_i \cap E_j) + \\
& \textstyle\sum_{1\le i<j<k\le n+1} \mathbb{P}(E_i \cap E_j \cap E_k) \ldots \\
& + (-1)^{n+1} \textstyle\sum_{1\le j\le n+1} \mathbb{P}(\bigcap_{i=1, i\neq j}^{n+1} E_i) \\
& + (-1)^{n+2}\mathbb{P}(E_1 \cap E_2 \cap \cdots \cap E_n \cap E_{n+1}
\end{aligned}
$$

which is the same formula as $\triangle$ for $n + 1$. Hence, the proof is complete. □

Example 2.2.9. 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?

Answer. Because each of the $\binom{15}{5}$ possible committees is equally likely, we know bottom of the fraction is $\binom{15}{5}$. Then we only need to find top of the fraction, which is the number of possible choices for men times the number of possible choices for women, e.g. $\binom{6}{3}\binom{9}{2}$. Hence, the final answer is

$$\frac{\binom{6}{3}\binom{9}{2}}{\binom{15}{5}} = \frac{240}{1001}$$

□

Example 2.2.10. If two dice are rolled, what is the probability that the sum of the upturned faces will equal 7?

Answer. To solve this problem, we assume that all of the 36 possible outcomes are equally likely. Since there are 6 possible outcomes - namely, (1,6), (2,5), (3,4), (4,3), (5,2), (6,1) - that result in the sum of the dice being equal to 7. Thus, the desired probability is $\frac{6}{36} = 1/6$. □

Example 2.2.11. If 3 balls are "randomly drawn" from a bowl containing 6 white and 5 black balls, what is the probability that one of the balls is white and the other two black?

Answer. Solution 1: If we regard the balls as being distinguishable and the order in which they are selected as being relevant, then the sample space consists of $11 \times 10 \times 9 = 990$ total possible outcomes. Furthermore, there are $6 \times 5 \times 4 = 120$ outcomes in which the first ball selected is white and the other two are black; $5 \times 6 \times 4 = 120$ outcomes in which the first is black, the second is white, and the third is black; and $5 \times 4 \times 6 = 120$ in which the first two are black and the third is white. Hence, assuming that "randomly

drawn" means that each outcome in the sample space is equally likely to occur, we see that the desired probability is

$$\frac{120+120+120}{990} = \frac{4}{11}$$

Solution 2: There are total $\binom{11}{3} = 165$ total possible outcomes in the sample space. Now among the 3 balls drawn, we want one of them to be white so this means we have $\binom{6}{1}$ total possible outcomes choosing one white balls out of six white balls and then, the rest two balls, we allow them to be black, i.e. $\binom{5}{2}$. Hence,

$$\frac{\binom{6}{1}\binom{5}{2}}{\binom{11}{3}} = \frac{4}{11}$$

□

Proposition 2.2.12. *If $\{E_n, n \geq 1\}$ is either an increasing or a decreasing sequence of events, then*

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

Proof. Suppose, first, that $\{E_n, n \geq 1\}$ is an increasing sequence, and define the events $F_n, n \geq 1$, by

$$\begin{aligned} F_1 &= E_1 \\ F_n &= E_n\left(\bigcup_{i=1}^{n-1} E_i\right)^c = E_n E_{n-1}^c \text{ for } n > 1 \end{aligned}$$

where we have used the fact that $\bigcup_1^{n-1} E_i = E_{n-1}$, since the events are increasing. In words, F_n consists of those outcomes in E_n that are not in any of the earlier E_i, $i < n$. It is easy to verify that the F_n are mutually exclusive events such that

$$\bigcup_{i=1}^{\infty} F_i = \bigcup_{i=1}^{\infty} E_i \text{ and } \bigcup_{i=1}^{n} F_i = \bigcup_{i=1}^{n} E_i \text{ for all } n \geq 1$$

Thus,

$$\begin{aligned} \mathbb{P}\left(\bigcup_{i=1}^{\infty}\right) &= \mathbb{P}\left(\bigcup_{i=1}^{\infty} F_i\right) \\ &= \sum_1^{\infty} \mathbb{P}(F_i) \text{ by Axiom (3)} \\ &= \lim_{n\to\infty} \sum_1^n \mathbb{P}(F_i) \\ &= \lim_{n\to\infty} \mathbb{P}\left(\bigcup_{i=1}^n F_i\right) \\ &= \lim_{n\to\infty} \mathbb{P}\left(\bigcup_{i=1}^n E_i\right) \\ &= \lim_{n\to\infty} \mathbb{P}(E_n) \end{aligned}$$

which proves the result when $\{E_n, n \geq 1\}$ is increasing.

If $\{E_n, n \geq 1\}$ is a decreasing sequence, then $\{E_n^c, n \geq 1\}$ is an increasing sequence; hence, from the preceding equations,

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} E_i^c\right) = \lim_{n\to\infty} \mathbb{P}(E_n^c)$$

However, because $\bigcup_{i=1}^{\infty} E_i^c = \left(\bigcap_1^{\infty} E_i \right)^c$, it follows that

$$\mathbb{P}\left(\left(\bigcap_1^{\infty} E_i \right)^c \right) = \lim_{n \to \infty} \mathbb{P}(E_n^c)$$

or, equivalently,

$$1 - \mathbb{P}\left(\bigcap_1^{\infty} E_i \right) = \lim_{n \to \infty} [1 - \mathbb{P}(E_n)] = 1 - \lim_{n \to \infty} \mathbb{P}(E_n)$$

which ends the proof. □

3 Conditional Probability and Independence

Go back to Table of Contents. Please click TOC

Let E and F denote, respectively, the event that the sum of the dice is 8 and the event that the first die is a 3, then the probability just obtained is called conditional probability that E occurs given that F has occurred. This term is denoted by

$$\mathbb{P}(E|F)$$

A general formula for $\mathbb{P}(E|F)$ that is valid for all events E and F is derived in the same manner: If the event F occurs, in order for E to occur, it is necessary that the actual occurrence be a point in both E and F; that is, in EF.

Definition 3.0.1. If $\mathbb{P}(F) > 0$, then

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

Proposition 3.0.2. *The multiplication rule.*

$$\mathbb{P}(E_1E_2E_3 \dots E_n) = \mathbb{P}(E_1)\mathbb{P}(E_2|E_1)\mathbb{P}(E_3|E_1E_2) \dots \mathbb{P}(E_n|E_1 \dots E_{n-1})$$

Example 3.0.3. Let us discuss an example. Toss a die twice n and toss two dice. There are 36 outcomes. Assume all outcomes are equally likely (fair dice). There is 1/36 chance for each outcome. Suppose that we observe that the first die is a 4. Given this information, what is the chance the sum of the two dice is no bigger than 7?

In this case, we have (4,1), (4,2), (4,3), (4,4), (4,5), (4,6). Then the chance 3/6 = 1/2.

Suppose we observe one of two dice is a 4, then the probability becomes what? The answer is 6/11.

Remark 3.0.4. Let E, F be two events and let $\mathbb{P}(F) > 0$, and we have

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

Given F already occurred, the chance of E will occur is $\mathbb{P}(E|F)$

3.1 Bayes' Formula

Let E and F be events. We may express E as

$$E = EF \cup EF^c$$

for, in order for an outcome to be in E, it must either be in both E and F or be in E but not in F. As EF and EF^c are clearly mutually exclusive, we have, by Axiom 3, we have

$$\begin{aligned} \mathbb{P}(E) &= \mathbb{P}(EF) + \mathbb{P}(EF^c) \\ &= \mathbb{P}(E|F)\mathbb{P}(F) + \mathbb{P}(E|F^c)\mathbb{P}(F^c) \\ &= \mathbb{P}(E|F)\mathbb{P}(F) + \mathbb{P}(E|F^c)[1 - \mathbb{P}(F)] \end{aligned}$$

Definition 3.1.1. The odds of an event A are defined by

$$\frac{\mathbb{P}(A)}{\mathbb{P}(A^c)} = \frac{\mathbb{P}(A)}{1 - \mathbb{P}(A)}$$

That is, the odds of an event A tell how much more likely it is that the event A occurs than it is does not occur. For instance, $\mathbb{P}(A) = \frac{2}{3}$, then $\mathbb{P}(A) = 2\mathbb{P}(A^c)$, so the odds are 2. If the odds are equal to α, then it is common to say that the odds are "α to 1" in favor of the hypothesis.

Consider now a hypothesis H that is true with probability $\mathbb{P}(H)$, and suppose that new evidence E is introduced. Then, the conditional probabilities, given evidence E, that H is that and that H is not true are respectively given by

$$\mathbb{P}(H|E) = \frac{\mathbb{P}(E|H)\mathbb{P}(H)}{\mathbb{P}(E)} \text{ and } \mathbb{P}(H^c|E) = \frac{\mathbb{P}(E|H^c)\mathbb{P}(H^c)}{\mathbb{P}(E)}$$

Therefore, the new odds after the evidence E has been introduced are

$$\frac{\mathbb{P}(H|E)}{\mathbb{P}(H^c|E)} = \frac{\mathbb{P}(H)}{\mathbb{P}(H^c)} \frac{\mathbb{P}(E|H)}{\mathbb{P}(E|H^c)}$$

We can further generalize: Suppose that $F_1, ..., F_n$ are mutually exclusive events such that

$$\bigcup_{i=1}^{n} F_i = S$$

In other words, exactly one of the events $F_1, ..., F_n$ must occur. By writing,

$$E = \bigcup_{i=1}^{n} EF_i$$

and using the fact that the events EF_i, for $i = 1, ..., n$ are mutually exclusive, we obtain

$$\mathbb{P}(E) = \sum_{i=1}^{n} \mathbb{P}(EF_i) = \sum_{i=1}^{n} \mathbb{P}(E|F_i)\mathbb{P}(F_i)$$

Let $F_1, ..., F_n$ be a set of mutually exclusive and exhaustive events (meaning that exactly one of these events must occur). Suppose now that E has occurred and we are interested in determining which one of the F_j also occurred. Then we have

Proposition 3.1.2. *IMPORTANT* *We have the following proposition:*

$$\begin{aligned} \mathbb{P}(F_j|E) &= \frac{\mathbb{P}(EF_j)}{\mathbb{P}(E)} \\ &= \frac{\mathbb{P}(E|F_j)\mathbb{P}(F_j)}{\sum_{i=1}^{n} \mathbb{P}(E|F_i)\mathbb{P}(F_i)} \end{aligned}$$

which is known as Bayes' formula.

Example 3.1.3. Discuss an example from class. There is a fair deck of cards (a fair deck has 52 cards and each is drawn equally likely). Deal the cards to 4 players (each has 1 card), say E, W, N, and S. If North has 6 spades, what is the chance that East has 3 spades?

Use reduced sample space: Given N has 6 spades and other 7 (non-spade), E, W, S will share other 7 spades among $13 \times 3 = 39$ cards. Hence,

$$\frac{\binom{7}{3}\binom{32}{10}}{\binom{39}{13}}$$

3.2 Independent Events

From the idea of Bayes' rule, we can discuss the notion of independent events.

Definition 3.2.1. Consider two events E and F. They are independent if equation

$$\mathbb{P}(EF) = \mathbb{P}(E)\mathbb{P}(F)$$

holds. If they are not independent, we say they are dependent.

Proposition 3.2.2. *If E and F are independent, then so are E and F^c.*

Definition 3.2.3. Three events E, F, and G are said to be independent if

$$\begin{aligned}\mathbb{P}(EFG) &= \mathbb{P}(E)\mathbb{P}(F)\mathbb{P}(G)\\ \mathbb{P}(EF) &= \mathbb{P}(E)\mathbb{P}(F)\\ \mathbb{P}(EG) &= \mathbb{P}(E)\mathbb{P}(G)\\ \mathbb{P}(FG) &= \mathbb{P}(F)\mathbb{P}(G)\end{aligned}$$

Example 3.2.4. A famous example is Gambler's Ruin. Please see [1] page 84.

Proposition 3.2.5. *We have the following properties*

1. $0 \leq \mathbb{P}(E|F) \leq 1$
2. $\mathbb{P}(S|F) = 1$
3. *If E_i, for $i = 1, 2, ...$ are mutually exclusive events, then*

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} E_i|F\right) = \sum_{i=1}^{\infty} \mathbb{P}(E_i|F)$$

Let us discuss a birth problem as this problem can be related many problems in conditional probabilities.

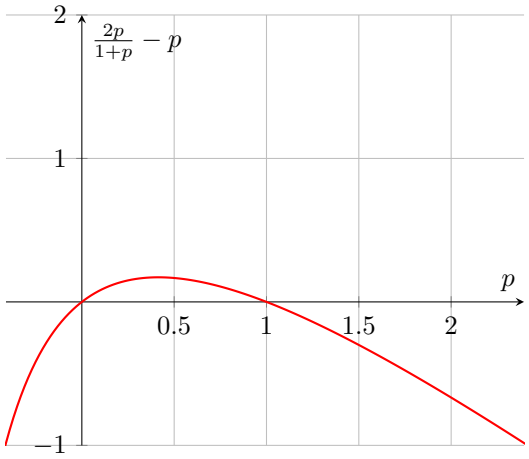
Example 3.2.6. Female chimp gave birth. It is not certain which of two male chimps is the father. Before genetic analysis, it is believed that the probability that male number 1 is the father is p and the probability that male number 2 is the father is $1 - p$. DNA obtained from the mother, male number 1, and male number 2 indicates that on one specific location of the genome, the mother has the gene pair (A,A), male number 1 has gene pair (a,a), and male number 2 has the gene pair (A,a). If a DNA test shows that the baby chimp has the gene pair (A,a), what is the probability that male number 1 is the father?

Answer. Let M_i be the event that male number i is the father. Let $B_{A,a}$ be the event that baby chimp has the gene pair (A,a). Then $\mathbb{P}(M_1|B_{A,a})$ is obtained:

$$\begin{aligned}\mathbb{P}(M_1|B_{A,a}) &= \frac{\mathbb{P}(M_1 B_{A,a})}{\mathbb{P}(B_{A,a})}\\ &= \frac{\mathbb{P}(B_{A,a}|M_1)\mathbb{P}(M_1)}{\mathbb{P}(B_{A,a}|M_1)\mathbb{P}(M_1)+\mathbb{P}(B_{A,a}|M_2)\mathbb{P}(M_2)}\\ &= \frac{1\cdot p}{1\cdot p+(1/2)(1-p)}\\ &= \frac{2p}{1+p}\end{aligned}$$

Now let us compare result with p

Figure 1: The figure presents the graph of $\frac{2p}{1+p} - p$.



Hence, we arrived the inequality

$$\frac{2p}{1+p} > p$$

We conclude the information that the baby's gene pair is (A,a) increases the probability that male number 1 is the father. □

Example 3.2.7. Suppose that we have 3 cards that are identical in form, except that both sides of the first card are colored red, both sides of the second card are colored black, and one side of the third card is colored red and the other side black. The 3 cards are mixed up and 1 card is randomly selected and put down on the table. If the upper side of the chosen card is colored red, what is the probability that the other side is colored black?

Answer. Let us denote these 3 cards to be RR, BB, and RB. Here "R" means "red" and "B" means "black". The notation "RR" means the card that have both sides red, "BB" means the card that has both sides black, and the last one "RB" means the card with one side red and the other side black. Also denote "R" to be the event that the card on the table the up side is red. The desired probably is the following

$$\begin{aligned}
\mathbb{P}(\text{RB}|\text{R}) &= \frac{\mathbb{P}(\text{RB}\cap\text{R})}{\mathbb{P}(\text{R})}, \text{ by definition of conditional probability} \\
&= \frac{\mathbb{P}(\text{R}|\text{RB})\mathbb{P}(\text{RB})}{\mathbb{P}(\text{R}|\text{RR})\mathbb{P}(\text{RR})+\mathbb{P}(\text{R}|\text{RB})\mathbb{P}(\text{RB})+\mathbb{P}(\text{R}|\text{BB})\mathbb{P}(\text{BB})}, \star \\
&= \frac{(1/2)(1/3)}{(1)(1/3)+(1/2)(1/3)+0(1/3)} \\
&= 1/3
\end{aligned}$$

while $\star$ means by Bayes' theorem and Law of Total Probability. □

Example 3.2.8. A box contains two coins: a regular coin and one fake two-headed coin $\mathbb{P}(H) = 1$. I choose a coin at random and toss it twice. Define the following events.

- A: First coin toss results in an H.
- B: Second coin toss results in an H.
- C: Coin 1 (the regular coin) has been selected.

Find $\mathbb{P}(A|C)$.

Answer. 1. $\mathbb{P}(A|C) = \mathbb{P}(B|C) = 1/2$.

□

Example 3.2.9. In answering a question on a multiple-choice test, a student either knows the answer or guesses. Suppose the probability that the student knows the answer is 60% and the probability that the student guess the answer randomly to be 40%. Moreover, let us assume that the student who guesses the answer correctly to be 25%. Question: What is the chance that the student knows the correct answer given that the answer is correct?

Answer. Let us denote the following events:

- Let E_1 be the events that the student knows the answer, and from the problem statement, we have $\mathbb{P}(E_1) = 0.6$
- Let E_2 be the events that the student guess the answer randomly, and from the problem statement, we have $\mathbb{P}(E_2) = 0.4$.
- Let A be the event that the student answers a question correctly.

Next, the probability that the student answers the question correctly given that he knows the answer would be 1. Hence, this gives us $\mathbb{P}(A|E_1) = 1$. Moreover, the probability that the student answers correctly given that he guesses the problem would be $1/4$, i.e. $\mathbb{P}(A|E_2) = 1/4$ this is given from the problem statement. With all the above information, we can compute the probability that the student knows the answer given that he answers it correctly to be

$$\begin{aligned} \mathbb{P}(E_1|A) &= \frac{\mathbb{P}(E_1)\mathbb{P}(A|E_1)}{\mathbb{P}(E_1)\mathbb{P}(A|E_1)+\mathbb{P}(E_2)\mathbb{P}(A|E_2)}, \star \\ &= \frac{3/4 \cdot 1}{3/4 \cdot 1 + (1/4)\cdot(1/4)} \\ &= 12/13 \approx 0.92 \end{aligned}$$

Remark. In the step $\star$, we used Bayes' Theorem for the top of the fraction and the Law of Total Probability in the bottom of the fraction. □

4 Random Variables

Go back to Table of Contents. Please click TOC

This section we discuss random variables. We will start with definition of random variables and begin our discussion with discrete random variables. We can then discuss expectation and variance (1st moment and 2nd moment) of the random variables. Afterwards, we will move forward to discuss Bernoulli and Binomial random variables (and Poisson) as special case studies.

4.1 Random Variables

When we perform an experiment, we are often times interested in some function of the outcome. This way can generalize the situation and what will occur in future events.

Example 4.1.1. Consider an experiment of tossing 3 fair coins. Let Y denote the number of heads. Then Y is a random variable taking one of the values 0, 1, 2, and 3 with a certain probability respectively. We can write

$$\begin{aligned} \mathbb{P}(Y=0) &= 1/8 \\ \mathbb{P}(Y=1) &= 3/8 \\ \mathbb{P}(Y=2) &= 3/8 \\ \mathbb{P}(Y=3) &= 1/8 \end{aligned}$$

We notice that since Y must be one of the values from 0 through 3, then we must have

$$1 = P\bigg(\bigcup_{i=0}^{3}(Y=i)\bigg) = \sum_{i=0}^{3}\mathbb{P}(Y=i)$$

Example 4.1.2. Consider another example. Four balls are to be randomly selected, without replacement, from an urn that contains 20 balls numbered 1 through 20. (Up to here, we know there are $\binom{20}{4}$ possible outcomes.) Let X be the largest numbered ball selected, then X is a random variable that takes on one of the values 4, 5, ..., 20. The probability that X takes on each of its possible values is

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

Suppose we want to determine $\mathbb{P}(X>10)$. One way is to compute

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

We can, alternatively, compute the complement of the above event and subtract that probability from 100%. We omit the computation here. One can refer to text [1] page 113.

All of these are motivating examples that it is necessary to come up with a notion of random variable instead of paying attention to a single event.

4.2 Discrete Random Variables

A random variable that can take on at most a countable number of possible values is said to be discrete. For a discrete random variable X, we define the probability mass function $\mathbb{P}(a)$ of X by

$$\mathbb{P}(a) = \mathbb{P}(X=a)$$

The probability mass function $\mathbb{P}(a)$ is positive for at most a countable number of values of a. That is, if X must assume one of the values $x_1, ...$, then

$$\begin{aligned} \mathbb{P}(x_i) &\geq 0 \text{ for } i = 1, 2, ... \\ \mathbb{P}(x) &= 0 \text{ for all other values of } x \end{aligned}$$

Since X must take on one of the values x_i, we have

$$\sum_{i=1}^{\infty} \mathbb{P}(x_i) = 1$$

Example 4.2.1. The probability mass function of a random variable X is given by $\mathbb{P}(i) = c\lambda^i/i!$, $i = 0, 1, 2, ...$, where λ is some positive value. Find (a) $\mathbb{P}(X = 0)$ and (b) $\mathbb{P}(X > 2)$.

Answer. Since $\sum_{i=0}^{\infty} \mathbb{P}(i) = 1$, we have

$$c \sum_{i=0}^{\infty} \frac{\lambda^i}{i!} = 1$$

which, using $e^x = \sum_{i=0}^{\infty} x^i/i!$, implies that

$$ce^{\lambda} = 1 \text{ or } c = e^{-\lambda}$$

Thus, we have

1. $\mathbb{P}(X = 0) = e^{-\lambda}\lambda^0/0! = e^{-\lambda}$.
2. $\mathbb{P}(X > 2) = 1 - \mathbb{P}(X \leq 2) = 1 - e^{-\lambda} - \lambda e^{-\lambda} - \frac{\lambda^2 e^{-\lambda}}{2}$

□

The cumulative distribution function F can be expressed in terms of $\mathbb{P}(a)$ by

$$F(a) = \sum_{\text{all } x \leq a} \mathbb{P}(x)$$

If X is a discrete random variable whose possible values are $x_1, x_2, ...$, where $x_1 < x_2 < x_3 < ...$, then the distribution function F of X is a step function. That is, the value of F is constant in the intervals (x_{i-1}, x_i) and then takes a step (or jump) of size $\mathbb{P}(x_i)$ at x_i. For instance, if X has a probability mass function given by

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

then its cumulative distribution function is

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

4.3 Expected Value

One of the most important concepts in probability theory is that of the expectation of a random variable. If X is a discrete random variable having a probability mass function $\mathbb{P}(x)$, or the expected value, of X, denoted by $E[X]$, is defined by

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

The expected value of X is a weighted average of the possible values that X can take on, each value being weighted by the probability that X assumes it.

Example 4.3.1. Find $E[X]$, where X is the outcome when we roll a fair die.

Answer. Since $\mathbb{P}(1) = \mathbb{P}(2) = \mathbb{P}(3) = \mathbb{P}(4) = \mathbb{P}(5) = \mathbb{P}(6) = 1/6$, we obtain

$$E[X] = 1\big(\frac{1}{6}\big) + 2\big(\frac{1}{6}\big) + 3\big(\frac{1}{6}\big) + 4\big(\frac{1}{6}\big) + 5\big(\frac{1}{6}\big) + 6\big(\frac{1}{6}\big) = 7/2$$

□

4.4 Expectation of a Function of a Random Variable

Suppose that we are given discrete random variable along with its probability mass function and we want to compute expected value of some function of X, say, $g(X)$. We can determine $E[g(X)]$ by using definition of expected value.

Example 4.4.1. Let X denote a random variable that takes on any of the values -1, 0, and 1 with respective probabilities

$$\mathbb{P}(X = -1) = 0.2, \mathbb{P}(X = 0) = 0.5, \mathbb{P}(X = 1) = 0.3$$

Compute $E[X^2]$.

Answer. Let $Y = X^2$. Then the probability mass function of Y is given by

$$\begin{aligned} \mathbb{P}(Y = 1) &= \mathbb{P}(X = -1) + \mathbb{P}(X = 1) = 0.5 \\ \mathbb{P}(Y = 0) &= \mathbb{P}(X = 0) = 0.5 \end{aligned}$$

Hence,

$$E[X^2] = E[Y] = 1(0.5) + 0(0.5) = 0.5$$

□

Proposition 4.4.2. *If X is a discrete random variable that takes on one of the values x_i, $i \geq 1$, with respective probabilities $\mathbb{P}(x_i)$, then, for any real-valued function g,*

$$E[g(X)] = \sum_i g(x_i)\mathbb{P}(x_i)$$

Proof. Please see text [1] Page 122 for proof. □

Proposition 4.4.3. *If a and b are constants, then*

$$E[aX + b] = aE[X] + b$$

Proof. We prove

$$\begin{aligned} E[aX+b] &= \sum_{x:\mathbb{P}(x)>0}(ax+b)\mathbb{P}(x) \\ &= a\sum_{x:\mathbb{P}(x)>0} x\mathbb{P}(x) + b\sum_{x:\mathbb{P}(x)>0}\mathbb{P}(x) \\ &= aE[X]+b \end{aligned}$$

□

The expected value of a random variable X, $E[X]$, is also referred to as the mean or the first moment of X. The quantity $E[X^n]$, $n \geq 1$, is called the nth moment of X. By Proposition 4.1, we note that

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

4.5 Variance

Besides expectation, it is also imporant to measure the variation.

Definition 4.5.1. If X is a random variable with mean μ, then the variance of X, denoted by $\text{Var}(X)$, is defined by

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

Alternatively, one can derive

$$\begin{aligned} \text{Var}(X) &= E[(X-\mu)^2] \\ &= \sum_x (x-\mu)^2\mathbb{P}(x) \\ &= \sum_x (x^2 - 2\mu x + \mu^2)\mathbb{P}(x) \\ &= \sum_x x^2\mathbb{P}(x) - 2\mu\sum_x x\mathbb{P}(x) + \mu^2\sum_x \mathbb{P}(x) \\ &= E[X^2] - 2\mu^2 + \mu^2 \\ &= E[X^2] - \mu^2 \end{aligned}$$

Example 4.5.2. Calculate $\text{Var}(X)$ if X represents the outcome when a fair die is rolled.

Answer. You can easily find $E[X] = \frac{7}{2}$. Now, we find

$$\begin{aligned} E[X^2] &= 1^2(1/6) + 2^2(1/6) + 3^2(1/6) + 4^2(1/6) + 5^2(1/6) + 6^2(1/6) \\ &= (91)(1/6) \end{aligned}$$

and thus we have variance

$$\text{Var}(X) = \frac{91}{6} - \left(\frac{7}{2}\right)^2 = \frac{35}{12}$$

□

A useful identity is that for any constants a and B,

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

To prove this equality, let $\mu = E[X]$ and note that $E[aX+b] = a\mu + b$. Therefore, we have

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

Remark 4.5.3. Please note the following.

1. Analogous to the means being the center of gravity of a distribution of mass, the variance represents, in the terminology of mechanics, the moment of inertia.
2. The square root of the $\text{Var}(X)$ is called the standard deviation of X, and we denote it by $\text{SD}(X)$. That is,

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

Discrete random variables are often classified according to their probability mass functions. In the future, we may deal with probability distribution function (or probability density function) for continuous random variables.

4.6 The Bernoulli and Binomial Random Variables

Suppose that a trial, or an experiment, whose outcome can be classified as either a success or a failure is performed. If we let $X = 1$ when the outcome is a success and $X = 0$ when it is a failure, then the probability mass function of X is given by

$$\begin{aligned} \mathbb{P}(0) &= \mathbb{P}(X = 0) = 1 - p \\ \mathbb{P}(1) &= \mathbb{P}(X = 1) = p \end{aligned}$$

where p, $0 \leq p \leq 1$, is the probability that the trial is a success. A random variable X is said to be a Bernoulli random variable if its probability mass function is given by the above equation for some $p \in (0, 1)$.

Suppose now that n independent trials, each of which results in a success with probability p or in a failure with probability $1-p$, are to be performed. If X represents the number of successes that occur in the n trials, then X is said to be a binomial random variable with parameters (n, p). Thus, a Bernoulli random variable is just a binomial random variable with parameters $(1, p)$. The probability mass function of a binomial random variable having parameters (n, p) is given by

$$\mathbb{P}(i) = \binom{n}{i} p^i (1-p)^{n-i} \text{ for } i = 0, 1, ..., n$$

Example 4.6.1. It is known that screws produced by a company can be defective with probability 0.01, independently of one another. The company sells the screws in packages of 10 and offers a money-back guarantee that at most 1 of the 10 screws is defective. What proportion of packages sold must the company replace?

Answer. If X is the number of defective screws in a package, then X is a binomial random variable with parameters $(10, 0.01)$. Hence, the probability that a package will have to be replaced is

$$1 - \mathbb{P}(X = 0) - \mathbb{P}(X = 1) = 1 - \binom{10}{0}(0.01)^0(0.99)^{10} - \binom{10}{1}(0.01)^1(0.99)^9 \equiv 0.004$$

□

Example 4.6.2. Consider a jury trial in which it takes 8 of the 12 jurors to convict the defendant; that is, in order for the defendant to be convicted, at least 8 of the jurors must vote him guilty. If we assume that jurors act independently and that whether or not the defendant is guilty, each makes the right decision with probability θ, what is the probability that the jury renders a correct decision?

Answer. The problem, as stated, is incapable of an actual solution. However, we can work out an expression to model this environment. The situation is binary and we have either guilty or not guilty. The former requires 8 votes and the latter requires 5. Hence, we have

$$\begin{aligned} \text{if he is guilty:} \quad & \sum_{i=8}^{12} \binom{12}{i} \theta^i (1-\theta)^{12-i} \\ \text{if he is not Guilty:} \quad & \sum_{i=5}^{12} \binom{12}{i} \theta^i (1-\theta)^{12-i} \end{aligned}$$

and hence, by letting probability that the defendant is guilty to be α, we can write out the expression for rendering a correct decision

$$\alpha \sum_{i=8}^{12} \binom{12}{i} \theta^i (1-\theta)^{12-i} + (1-\alpha) \sum_{i=5}^{12} \binom{12}{i} \theta^i (1-\theta)^{12-i}$$

□

We can examine the properties of a binomial random variable with parameters n and p. To begin, let us compute its expected value and variance. To begin, note that

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

Using the identity

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

gives

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

where Y is a binomial random variable with parameters $n-1$, p. Setting $k=1$, we would arrive

$$E[X] = np$$

which gives us the expected number of successes that occur in n independent trials when each is a success with probability p. Setting $k=2$, we yield

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

Since $E[X] = np$, we obtain

$$\begin{aligned} E[X] &= np \\ \text{Var}(X) &= n\mathbb{P}(1-p) \end{aligned}$$

In the following, let us provide a step-by-step derivation of these two random variables: Bernoulli and Binomial.

Bernoulli Random Variable. We define X, a discrete random variable, to be Bernoulli random variable if it has the support $R_X = \{0, 1\}$ with the probability mass function to be

$$\mathbb{P}_X(x) = \begin{cases} p & \text{if } x = 1 \\ 1-p & \text{if } x = 0 \\ 0 & \text{otherwise} \end{cases}$$

First, let us validate that this is indeed a probability mass function. We write the following

$$\begin{aligned} \sum_{x\in R_x} \mathbb{P}_X(x) &= \mathbb{P}(X=1)+\mathbb{P}(X=0) \\ &= p+(1-p)=1 \end{aligned}$$

and therefore we confirm that this is indeed a probability mass function.

Next, we find the expected value (or the expectation) of a Bernoulli random variable.

$$\begin{aligned} \mathbb{E}(X) &= \sum_{x\in R_x} x\mathbb{P}(X) \\ &= 1\cdot\mathbb{P}(X=1)+0\cdot\mathbb{P}(X=0) \\ &= p+0 \\ &= p \end{aligned}$$

and to find the variance of X we first find

$$\begin{aligned} \mathbb{E}X^2 &= \sum_{x\in R_X} x^2\mathbb{P}(x) \\ &= 1^2p+0 \\ &= p \end{aligned}$$

and thus $\text{var}(X)=\mathbb{E}X^2-(\mathbb{E}X)^2=p-p^2=p(1-p)$. Last, we shall find the moment generating function for Bernoulli random variable X. It suffices to compute the following

$$\begin{aligned} M_X(t) &= \mathbb{E}(\exp(tX)) \\ &= \sum_{x\in R_X}\exp(tx)\mathbb{P}(x) \\ &= \exp(t\cdot 1)p+\exp(t\cdot 0)(1-p) \\ &= 1-p+\exp(t)p \end{aligned}$$

and we are done. The reason the moment generating function is important is because we can use the moment generating function to efficiently find any moments we desire. For instance, if we want to use the moment generating function to find the expectation, we can do the following

$$\begin{aligned} \text{Take derivative: } \tfrac{\partial}{\partial t}M_X(t) &= p\exp(t) \\ \text{Setting } t=0 \text{ and we obtain : } p\exp(0) &= p \end{aligned}$$

and in a similar fashion we can find $\mathbb{E}X^2$ (which is called the second moment)

$$\begin{aligned} \text{Take derivative: } \tfrac{\partial}{\partial t}M_X(t) &= p\exp(t) \\ \text{Taking a 2nd derivative: } \tfrac{\partial}{\partial t}p\exp(t) &= p\exp(t) \\ \text{Setting } t=0 \text{ and we obtain : } p\exp(0) &= p \end{aligned}$$

which happens to be the same with the first moment. In other words, we can generalize the computation for the kth moment to be

$$\frac{\partial^k}{(\partial t)^k}M_X(t)\Big|_{t=0}$$

Next, we can repeat the above with Binomial distribution (or Binomial random variable).

Binomial Random Variable. We define X to be a discrete random variable. Let $n\in\mathbb{N}$ and $p\in(0,1)$. Let the support of X be $R_X=\{0,1,...,n\}$. We say that X has a binomial distribution with parameters n and p if its probability mass function is

$$\mathbb{P}_X(x)=\begin{cases} \binom{n}{x}p^x(1-p)^{n-x} & \text{if } \in R_X \\ 0 & \text{if } x\notin R_X \end{cases}$$

To show that this is a probability mass function, it suffices to show that the sum of probability of X over the entire support equals to one. Hence, we write

$$\begin{aligned} \sum_{x \in R_X} \mathbb{P}(x) &= \sum_{x=0}^{n} \binom{n}{x} p^x (1-p)^{n-x} \\ &= (p + (1-p))^n \\ &= 1^n = 1 \end{aligned}$$

and this is valid because of binomial expansion formula: $(a+b)^n = \sum_{x=0}^{n} a^x b^{n-x}$. This random variable is related to Bernoulli distribution which leads to the following propositions.

Proposition 4.6.3. *If a random variable X has a binomial distribution with parameters n and p, with $n = 1$, then X has a Bernoulli distribution with parameter p.*

Proposition 4.6.4. *If a random variable X has a binomial distribution with parameters n and p, then X is a sum of n jointly independent Bernoulli random variables with parameter p.*

The above propositions are helpful because they simplify the computation of searching for expected value and variance of X if X is a Binomial random variable. To illustrate this, let us find them in the following

$$\begin{aligned} \mathbb{E}X &= \mathbb{E}\Big[\sum_{i=1}^{n} Y_i\Big] \text{ we use } Y_i \text{ to represent } n \text{ Bernoulli random variables} \\ &= \sum_{i=1}^{n} \mathbb{E}Y_i \\ &= np \text{ since Bernoulli r.v. has mean } p \end{aligned}$$

and the search for the variance follows a similar fashion

$$\begin{aligned} \text{var}(X) &= \text{var}\Big[\sum_{i=1}^{n} Y_i\Big] \\ &= \sum_{i=1}^{n} \text{var}(Y_i) \text{ here we assume they are jointly independent} \\ &= \sum_{i=1}^{n} p(1-p) \\ &= np(1-p) \end{aligned}$$

With the above computation in mind, it is a small leap to understand the moment generating function of a Binomial random variable to be the product of n moment generating functions of Bernoulli random variables. Let us write this rigorously in the following

$$\begin{aligned} M_X(t) &= \mathbb{E}(\exp(tX)) \\ &= \mathbb{E}(\exp(tY_1 + \cdots + tY_n)) \\ &= \mathbb{E}(\exp(tY_1) \cdot \cdots \cdot \exp(tY_n)) \\ &= \mathbb{E}(\exp(tY_1)) \cdot \cdots \cdot \mathbb{E}(\exp(tY_n)) \\ &= M_{Y_1}(t) \dots M_{Y_n}(t) \\ &= (1 - p + p\exp(t))^n \end{aligned}$$

4.7 Geometric Random Variable

Consider a Bernoulli experiment, that is, a random experiment having two possible outcomes: either success or failure. We repeat the experiment until we get the first success, and then we count the number X of failures that we faced prior to recording the success. Since the experiments are random, X is a random variable. If the repetitions of the experiment are independent of each other, then the distribution of X, which we are going to study below, is called geometric distribution. For example, if we toss a coin until we obtain head, the number of tails before the first head has a geometric distribution.

Geometric random variable. In the following, let us formally define geometric random variable and let us find its expectation and variance. First, let us use the

definition to find its expectation. Suppose X is a Geometric random variable with the support $\mathbb{Z}_+$. The probability mass function of X is defined as

$$\mathbb{P}_X(x) = \begin{cases} (1-p)^x p & \text{if } x \in R_X \\ 0 & \text{if } x \notin R_X \end{cases}$$

Then the expected value of X is

$$\begin{aligned} \mathbb{E}X &= \sum_{x=0}^{\infty} x(1-p)^x p \\ &= p(1-p)\sum_{x=0}^{\infty} x(1-p)^{x-1} \text{ since } \tfrac{\partial}{\partial p}(1-p)^x = -x(1-p)^{x-1} \\ &= -p(1-p)\sum_{x=0}^{\infty} \tfrac{\partial}{\partial p}(1-p)^x \\ &= -p(1-p)\tfrac{\partial}{\partial p}\tfrac{1}{1-(1-p)} \text{ by geometric series} \\ &= -p(1-p)(-p^{-2}) \\ &= 1/p - 1 \\ &= \tfrac{1-p}{p} \end{aligned}$$

and this relies on the fact that $\sum_{x=0}^{\infty}(1-p)^x = \frac{1}{1-(1-p)}$. In a similar fashion (which we omit here), we can find $\mathbb{E}X^2 = \frac{2-3p+p^2}{p^2}$ and in this computation requires us to take partial derivative in the middle twice (just like the above). In the end, we have variance to be $\text{var}(X) = \frac{1-p}{p^2}$. With a similar procedure, we can find the moment generating function $M_X(t) = \frac{p}{1-(1-p)\exp(t)}$. For more, please see `www.statlect.com/probability-distributions/geometric-distribution/`

4.8 Poisson Random Variable

IMPORTANT A random variable X that takes on one of the values 0, 1, 2, ...is said to be Poisson random variable with parameter λ if, for some $\lambda > 0$,

$$\mathbb{P}(i) = \mathbb{P}(X = i) = e^{-\lambda}\frac{\lambda^i}{i!} \text{ for } i = 0, 1, 2, ...$$

and it has property $\lambda = np$.

We can check the following:

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

Some general applications that obey Poisson probability are

1. The number of misprints on a page (or a group of pages) of a book.
2. The number of people in a community who survive to age 100.
3. The number of wrong telephone numbers that are dialed in a day.
4. The number of packages of a dog biscuits sold in a particular store each day
5. The number of vacancies occurring during a year in the federal judicial system.
6. The number of α-particles discharged in a fixed period of time from some radioactive material.

Example 4.8.1. Suppose number of typographical errors on a single page of this book has a Poisson distribution with parameter $\lambda = \frac{1}{2}$. Calculate the probability that there is at least one error on your page.

Answer. Letting X denote the number of error on a page, we have

$$\mathbb{P}(X \geq 1) = 1 - \mathbb{P}(X = 0) = 1 - e^{-1/2} = 0.393$$

□

Example 4.8.2. Suppose that the probability that an item produced by a certain machine will be defective is 0.1. Find the probability that a sample of 10 items will contain at most 1 defective item.

Answer. The desired probability is

$$\binom{10}{0}(0.1)^0(0.9)^{10} + \binom{10}{1}(0.1)^1(0.9)^9 = 0.74$$

while Poisson approximation would yield similar results. □

An interesting proposition relates Poisson distribution to the exponential distribution.

Proposition 4.8.3. *The number of occurrences of an event within a unit of time has a Poisson distribution with parameter λ if the time elapsed between two successive occurrences of the event has an exponential distribution with parameter λ and it is independent of previous occurrences.*

Poisson Random Variable. Next, let us write out step-by-step procedure of finding the expectation and variance of a Poisson random variable.

$$\begin{aligned}
\mathbb{E}X &= \textstyle\sum_{x\in R_X} x\mathbb{P}(X) \\
&= \textstyle\sum_{x=0}^{\infty} x\exp(-\lambda)\lambda x\frac{1}{x!} \\
&= 0 + \textstyle\sum_{x=1}^{\infty} x\lambda^x\exp(-\lambda)/x! \\
&= \textstyle\sum_{y=0}^{\infty}(y+1)\lambda^{y+1}\exp(-\lambda)/(y+1)! \text{ by using change of variables: } y = x-1 \\
&= \lambda\textstyle\sum_{y=0}^{\infty}\lambda^y\exp(-\lambda)/y! \\
&= \lambda\cdot 1 \text{ because the summation results in one} \\
&= \lambda
\end{aligned}$$

and next to find variance we first find

$$\begin{aligned}
\mathbb{E}X^2 &= \textstyle\sum_{x\in R_X} x^2\mathbb{P}(x) \\
&= \textstyle\sum_{x=0}^{\infty} x^2\lambda^x\exp(-\lambda)/x! \\
&= 0 + \textstyle\sum_{x=1}^{\infty} x^2\lambda^x\exp(-\lambda)/x! \\
&= \textstyle\sum_{y=0}^{\infty}(y+1)^2\lambda^{y+1}\exp(-\lambda)/(y+1)! \text{ again, change of var.: } y = x-1 \\
&= \lambda\textstyle\sum_{y=0}^{\infty}(y+1)\lambda^y\exp(-\lambda)/y! \\
&= \lambda\big(\textstyle\sum_{y=0}^{\infty} y\lambda^y\exp(-\lambda)/y! + \sum_{y=0}^{\infty}(1)\lambda^y\exp(-\lambda)/y!\big) \\
&= \lambda(\mathbb{E}Y + 1) \\
&= \lambda(\lambda+1)
\end{aligned}$$

with simple computation variance can be found: $\text{var}(X) = \mathbb{E}X^2 - (\mathbb{E}X)^2 = \lambda^2+\lambda-\lambda^2 = \lambda$.

Last, we can find the moment generating function of a Poisson random variable. For any $t \in \mathbb{R}$, we can use the definition of a moment generating function to obtain

$$\begin{aligned}
M_X(t) &= \mathbb{E}(\exp(tX)) \\
&= \textstyle\sum_{x\in R_X}\exp(tx)\mathbb{P}(x) \\
&= \textstyle\sum_{x\in R_X}\exp(t)^x\lambda^x\exp(-\lambda)/x! \\
&= \exp(-\lambda)\textstyle\sum_{x=0}^{\infty}(\lambda\exp(t))^x/x! \text{ the summation term allows us to use Talor expansion} \\
&= \exp(-\lambda)\exp(\lambda\exp(t)) \\
&= \exp(\lambda(\exp(t)-1))
\end{aligned}$$

and we are done by using Taylor expansion: $\exp(a) = \sum_{x=0}^{\infty}(a)^x/x!$ while $a = \lambda\exp(t)$.

5 Continuous Random Variables

Go back to Table of Contents. Please click TOC

The previous chapter discussed discrete random variables, e.g. the random variable whose set of possible values is either finite or countably infinite. There also exist random variables whose set of possible values to be uncountable. Consider X be such a random variable. We say that X is continuous random variable if there exists a nonnegative function f, defined for all real $x \in (-\infty, \infty)$, having the property that for any set B of real numbers,

$$\mathbb{P}(X \in B) = \int_B f(x)dx$$

The function f is called the probability density function of the random variable X.

In words, the above equation states that the probability that X will be in B may be obtained by integrating the probability density function over the set B. Since X must assume some value, f must satisfy

$$1 = \mathbb{P}(X \in (-\infty, \infty)) = \int_{-\infty}^{\infty} f(x)dx$$

All probability statements about X can be answered in terms of f.

Example 5.0.1. Letting $B = [a, b]$, we obtain

$$\mathbb{P}(a \leq X \leq b) = \int_a^b f(x)dx$$

Example 5.0.2. IMPORTANT Suppose X is a continuous random variable whose probability density function is

$$f(x) = \begin{cases} C(4x - 2x^2) & 0 < x < 2 \\ 0 & \text{else} \end{cases}$$

1. What is the value of C?
2. Find $\mathbb{P}(X > 1)$.

Answer. We have the following

1. Since f is a probability density function, we must have
$$\int_{-\infty}^{\infty} f(x)dx = 1$$
and we can solve $C\int_0^2 (4x - 2x^2)dx = 1$. After integration, we have $C(2x^2 - \frac{2x^3}{3})|_{x=0}^2 = 1$ and we have result $C = \frac{3}{8}$.
2. $\mathbb{P}(X > 1) = \int_1^{\infty} f(x)dx = \frac{3}{8}\int_1^2 (4x - 2x^2)dx = \frac{1}{2}$.

□

Example 5.0.3. The amount of time in hours that a computer functions before breaking down is a continuous random variable with probability density function given by

$$f(x) = \begin{cases} \lambda e^{-x/100} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

What is the probability that

1. a computer will function between 50 and 150 hours before breaking down?
2. it will function for fewer than 100 hours?

Answer. We solve the parts accordingly

1. Since $1 = \int_{-\infty}^{\infty} f(x)dx = \lambda \int_0^{\infty} e^{-x/100}dx$, we can take integral and obtain $1 = -\lambda(100)e^{-x/100}|_0^{\infty} = 100\lambda$. We can solve for $\lambda = \frac{1}{\lambda}$. Then we can proceed to find the probability

$$\begin{aligned} \mathbb{P}(50 < X < 1500) &= \int_{50}^{150} \frac{1}{100} e^{-x/100}dx \\ &= -e^{-x/100}|_{50}^{150} \\ &= e^{-1/2} - e^{-3/2} \\ &= 0.383 \end{aligned}$$

2. I will leave this to you as an exercise.

□

5.1 Expectation and Variance of Continuous Random Variables

In discrete senses, we defined the expected value of a discrete random variable X by

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

If X is a continuous random variable having probability density function $f(x)$, then because

$$f(x) \approx \mathbb{P}(x \leq X \leq x + dx) \text{ for } dx \text{ small}$$

it is easy to see that the analogous definition is to define the expected value of X by

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

Example 5.1.1. Find $\mathbb{E}(X)$ when the density function of X is

$$f(x) = \begin{cases} 2x & \text{if } 0 \leq x \leq 1 \\ 0 & \text{else} \end{cases}$$

Answer. Solve the following

$$\begin{aligned} \mathbb{E}(X) &= \int xf(x)dx \\ &= \int_0^1 2x^2 \\ &= \frac{2}{3} \end{aligned}$$

□

Proposition 5.1.2. *If X is a continuous random variable with probability density function $f(x)$, then, for any real-valued function,*

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

Example 5.1.3. The density function of X is given by

$$f(x) = \begin{cases} 1 & \text{if } 0 \leq x \leq 1 \\ 0 & \text{else} \end{cases}$$

Find $E[e^X]$.

Answer. Let $Y = e^X$. We start by determining F_Y, the cumulative distribution function of Y. Now, for $1 \leq x \leq e$,

$$\begin{aligned} F_Y(x) &= \mathbb{P}(Y \leq x) \\ &= \mathbb{P}(e^X \leq x) \\ &= \mathbb{P}(X \leq \log(x)) \\ &= \int_0^{\log(x)} f(y)dy \\ &= \log(x) \end{aligned}$$

By differentiating $F_Y(x)$, we can conclude that the probability density function of Y is given by

$$f_Y(x) = \frac{1}{x} \text{ for } 1 \leq x \leq e$$

Hence,

$$\begin{aligned} E[e^X] = E[Y] &= \int_{-\infty}^{\infty} x f_Y(x)dx \\ &= \int_1^e dx \\ &= e - 1 \end{aligned}$$

□

Lemma 5.1.4. *For a nonnegative random variable Y,*

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

Lemma 5.1.5. *If a and b are constants, then*

$$E[aX + b] = aE[X] + b$$

The variance of a continuous random variable is defined exactly as it is for a discrete random variable, namely, if X is a random variable with expected value μ, then the variance of X is defined (for any type of random variable) by

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

The alternative formula,

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

Example 5.1.6. Recall the example above,

$$f(x) = \begin{cases} 2x & \text{if } 0 \leq x \leq 1 \\ 0 & \text{else} \end{cases}$$

Find $\text{Var}(X)$ for this random variable X.

Answer. First, we compute the second moment $E[X^2]$,

$$\begin{aligned} E[X^2] &= \int_{-\infty}^{\infty} x^2 f(x)dx \\ &= \int_0^1 2x^3 dx = \tfrac{1}{2} \end{aligned}$$

Hence, we obtain

$$\text{Var}(X) = \frac{1}{2} - \left(\frac{2}{3}\right)^2 = \frac{1}{18}$$

□

Note that we have the property $\text{Var}(aX+b) = a^2\text{Var}(X)$.

5.2 Uniform Random Variable

A random variable is said to be uniformly distributed over the interval $(0,1)$ if its probability density function is given by

$$f(x) = \begin{cases} 1 & 0 < x < 1 \\ 0 & \text{else} \end{cases}$$

Since this is a density function, the following properties hold (1) $f(x) \geq 0$ and (2) $\int_x f(x)dx = 1$.

In general, we say that X is a uniform random variable on the interval (α, β) if the probability density function of X is given by

$$f(x) = \begin{cases} \frac{1}{\beta-\alpha} & \text{if } \alpha < x < \beta \\ 0 & \text{else} \end{cases}$$

Since $F(a) = \int_{-\infty}^{a} f(x)dx$, it follows that

$$f(x) = \begin{cases} 0 & a \leq \alpha \\ \frac{1}{\beta-\alpha} & \text{if } \alpha < x < \beta \\ 1 & a \geq \beta \end{cases}$$

Example 5.2.1. Let X be uniformly distributed over (α, β). Find (a) $E[X]$ and (b) $\text{Var}(X)$.

Answer. We proceed accordingly

1. Compute
$$\begin{aligned} E[X] &= \int_{-\infty}^{\infty} xf(x)dx \\ &= \int_{\alpha}^{\beta} \frac{x}{\beta-\alpha}dx \\ &= \frac{\beta^2-\alpha^2}{2(\beta-\alpha)} \\ &= \frac{\beta+\alpha}{2} \end{aligned}$$

2. To find $\text{Var}(X)$, first calculate $E[X^2]$.
$$\begin{aligned} E[X^2] &= \int_{\alpha}^{\beta} \frac{1}{\beta-\alpha}x^2dx \\ &= \frac{\beta^3-\alpha^3}{3(\beta-\alpha)} \\ &= \frac{\beta^2+\alpha\beta+\alpha^2}{3} \end{aligned}$$

   Hence,
$$\text{Var}(X) = \frac{\beta^2+\alpha\beta+\alpha^2}{3} - \frac{(\alpha+\beta)^2}{4} = \frac{(\beta-\alpha)^2}{12}$$

□

Example 5.2.2. If X is uniformly distributed over $(0, 10)$, calculate the probability that $X < 3$.

Answer. Compute $\mathbb{P}(X < 3) = \int_0^3 \frac{1}{10}dx = \frac{3}{10}$. □

Example 5.2.3. IMPORTANT Buses arrive at a specified stop at 15-minute intervals starting at 7AM. That is, they arrive at 7, 7:15, 7:30, 7:45, and so on. If a passenger arrives at the stop at a time that is uniformly distributed between 7 and 7:30, find the probability that he waits

1. less than 5 minutes for a bus;
2. more than 10 minutes for a bus.

Answer. We proceed accordingly

1. Compute
$$\begin{aligned}\mathbb{P}(10<X<15)+\mathbb{P}(25<X<30) &= \int_{10}^{15}\tfrac{1}{30}dx+\int_{25}^{30}\tfrac{1}{30}dx\\ &= \tfrac{1}{30}\end{aligned}$$
2. Compute
$$\mathbb{P}(0<X<5)+\mathbb{P}(15<X<20)=\frac{1}{3}$$

□

5.3 Normal Random Variables

We say that X is a normal random variable, or simply that X is normally distributed, with parameters μ and σ^2 if the density of X is given by
$$f(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-(x-\mu)^2/2\sigma^2}$$
for $-\infty<x<\infty$. The density function is a bell-shaped curve that is symmetric about μ.

Example 5.3.1. Find $E[X]$ and $\text{Var}(X)$ when X is a normal random variable with parameters μ and σ^2.

Answer. Let us start by finding the mean and variance of the standard normal random variable $Z=(X-\mu)/\sigma$. We have
$$\begin{aligned}E[Z] &= \int_{-\infty}^{\infty}xf_Z(x)dx\\ &= \tfrac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}xe^{-x^2/2}dx\\ &= -\tfrac{1}{\sqrt{2\pi}}e^{-x^2/2}|_{-\infty}^{\infty}\\ &= 0\end{aligned}$$

Thus,
$$\begin{aligned}\text{Var}(Z) &= E[Z^2]\\ &= \tfrac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}x^2e^{-x^2/2}dx,\ \text{IBP: let } \mu=x \text{ and } d\nu=xe^{-x^2/2}\\ &= \tfrac{1}{\sqrt{2\pi}}\big(-xe^{-x^2/2}|_{-\infty}^{\infty}+\underbrace{\int_{-\infty}^{\infty}e^{-x^2/2}dx}_{=1}\big)\\ &= 1\end{aligned}$$

Because $X=\mu+\sigma Z$, the preceding yields the results
$$E[X]=\mu+\sigma E[Z]=\mu$$
and
$$\text{Var}(X)=\sigma^2\text{Var}(Z)=\sigma^2$$

□

Conventionally, we denote the cumulative distribution function of a standard normal random variable by $\Phi(x)$. That is,

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

and we have table

Table 5.1 Area $\Phi(x)$ Under the Standard Normal Curve to the Left of X.

X	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

Figure 2: Area $\Phi(x)$ from page 190 in [1]

Example 5.3.2. IMPORTANT If X is a normal random variable with parameters $\mu = 3$ and $\sigma^2 = 9$, find (a) $\mathbb{P}(2 < X < 5)$, (b) $\mathbb{P}(X > 0)$, and (c) $\mathbb{P}(|X - 3| > 6)$.

Answer. We proceed accordingly

1. Compute

$$\begin{aligned} \mathbb{P}(2 < X < 5) &= P\left(\tfrac{2-3}{3} < \tfrac{X-3}{3} < \tfrac{5-3}{3} \right) \\ &= \Phi(\tfrac{2}{3}) - \Phi(-\tfrac{1}{3}) \\ &= 0.3779 \end{aligned}$$

2. Compute
$$\begin{aligned}\mathbb{P}(X>0) &= P\left(\tfrac{X-3}{3} > \tfrac{0-3}{3}\right)\\ &= \mathbb{P}(Z>-1)\\ &= \Phi(1) = 0.8413\end{aligned}$$

3. Compute
$$\begin{aligned}\mathbb{P}(|X-3|>6) &= \mathbb{P}(X>9)+\mathbb{P}(X<-3)\\ &= P\left(\tfrac{X-3}{3} > \tfrac{9-3}{3}\right) + P\left(\tfrac{X-3}{3} < \tfrac{-3-3}{3}\right)\\ &= \mathbb{P}(X>2)+\mathbb{P}(Z<-2)\\ &= 0.0456\end{aligned}$$

□

Example 5.3.3. An expert witness in a paternity suit testifies that the length (in days) of human gestation is approximately normally distributed with parameters $\mu = 270$ and $\sigma^2 = 100$. The defendant in the suit is able to prove that he was out of the country during a period that begun 290 days before the birth of the child and ended 240 days before the birth. If the defendant was, in fact, the father of the child, what is the probability that the mother could have had the very long or very short gestation indicated by the testimony?

Answer. Let X denote the length of the gestation, and assume that the defendant is the father. Then the probability that the birth could occur within the indicated period is
$$\begin{aligned}\mathbb{P}(X>290 \text{ or } X<240) &= \mathbb{P}(X>290)+\mathbb{P}(X<240)\\ &= \mathbb{P}(\tfrac{X-270}{10}>2)+\mathbb{P}(\tfrac{X-270}{10}<-3)\\ &= 1-\Phi(2)+1-\Phi(3)\\ &= 0.0241\end{aligned}$$

□

Remark 5.3.4. Please be aware that the problem can ask you "or" instead of "and". In that case, the properties we learned from set theory follow. You should check the intersection between the two events accordingly.

Example 5.3.5. If X, the gain from an investment, is a normal random variable with mean μ and variance σ^2, then because the loss is equal to the negative of the gain, the VAR of such an investment is that value ν such that
$$0.01 = \mathbb{P}(-X > \nu)$$
We compute the following
$$\begin{aligned}0.01 &= P\left(\tfrac{-X+\mu}{\sigma} > \tfrac{\nu+\mu}{\sigma}\right)\\ &= 1-\Phi(\tfrac{\nu+\mu}{\sigma})\end{aligned}$$
and from table we know $\Phi(2.33) = 0.99$ so we know $\frac{\nu+\mu}{\sigma} = 2.33$. That is, $\nu = \text{VAR} = 2.33\sigma - \mu$. Consequently, among set of investments all of whose gains are normally distributed, the investment having the smallest VAR is the one having the largest value of $\mu - 2.33\sigma$.

Theorem 5.3.6. *The DeMoivre-Laplace Theorem. If S_n denotes the number of successes that occur when n independent trials, each resulting in a success of probability p, are performed, then, for any $a < b$,*

$$P\left(a \leq \frac{S_n - np}{\sqrt{n\mathbb{P}(1-p)}} \leq b \right) \to \Phi(b) - \Phi(a)$$

Example 5.3.7. Let X be the number of times that a fair coin that is flipped 40 times lands on heads. Find the probability that $X = 20$. Use the normal approximation and then compare it with the exact solution.

Answer. To employ normal approximation, note that because the binomial is a discrete integer-valued random variable, whereas the normal is a continuous random variable, it is best to write $\mathbb{P}(X = i)$ as $\mathbb{P}(i - 1/2 < X < i + 1/2)$ before applying the normal approximation (this is called the continuity correction). Hence, we compute

$$\begin{aligned}
\mathbb{P}(X = 20) &= \mathbb{P}(19.5 < X < 20.5) \\
&= P\left(\tfrac{19.5-20}{\sqrt{10}} < \tfrac{X-20}{\sqrt{10}} < \tfrac{20.5-20}{\sqrt{10}} \right) \\
&= P\left(-1.6 < \tfrac{X-20}{\sqrt{10}} < 1.6 \right) \\
&= \Phi(0.16) - \Phi(-0.16) \\
&= 0.1272
\end{aligned}$$

□

5.4 Exponential Random Variable

A continuous random variable whose probability density function is given, for some $\lambda > 0$, by

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}$$

is said to be an exponential random variable with parameter λ. The cumulative distribution $F(a)$ of an exponential random variable is given by

$$\begin{aligned}
F(a) &= \mathbb{P}(X \leq a) \\
&= \int_0^a \lambda e^{-\lambda x} dx \\
&= -e^{-\lambda x}|_0^a \\
&= 1 - e^{-\lambda a} \text{ for } a \geq 0
\end{aligned}$$

Note that $F(\infty) = \int_0^\infty \lambda e^{-\lambda x} = 1$.

Example 5.4.1. IMPORTANT Let X be an exponential random variable with parameter λ. Calculate (a) $E[X]$ and (b) $\text{Var}(X)$.

Answer. We solve the following accordingly

1. We use $E[X^n] = \int_0^\infty x^n \lambda e^{-\lambda} dx$. Integrating by parts (with $\lambda e^{-\lambda x} = d\nu$ and $\mu = x^n$) yields

$$\begin{aligned}
E[X^n] &= -x^n e^{-\lambda x}|_0^\infty + \int_0^\infty e^{-\lambda x} n x^{n-1} dx \\
&= 0 + \tfrac{n}{\lambda} \int_0^\infty \lambda e^{-\lambda x} x^{n-1} dx \\
&= \tfrac{n}{\lambda} E[X^{n-1}]
\end{aligned}$$

   Letting $n = 1$ and $n = 2$ gives us

$$E[X] = \frac{1}{2} \text{ and } E[X^2] = \frac{2}{\lambda} E[X] = \frac{2}{\lambda^2}$$

2. We have variance
$$\text{Var}(X) = \frac{2}{\lambda^2} - \left(\frac{1}{\lambda}\right)^2 = \frac{1}{\lambda^2}$$

□

Example 5.4.2. Suppose that the number of miles that a car can run before its battery wears out is exponentially distributed with an average value of 10,000 miles. If a person desires to take a 5000-mile trip, what is the probability that he or she will be able to complete the trip without having to replace the car battery? What can be said when the distribution is not exponential? (Assume the parameter $\lambda = 1/10$).

Answer. The desired probability is
$$\mathbb{P}(\text{remaining lifetime} > 5) = 1 - F(5) = e^{-5\lambda} = 0.607$$
However, if the lifetime distribution F is not exponential, then the relevant probability is
$$\mathbb{P}(\text{lifetime} > t + 5 | \text{lifetime} > t) = \frac{1 - F(t+5)}{1 - F(t)}$$
where t is the number of miles that the battery had been in use prior to the start of the trip. Therefore, if the distribution is not exponential, additional information is needed (namely, the value of t) before the desired probability can be calculated. □

6 Jointly Distributed Random Variables

Go back to Table of Contents. Please click TOC

6.1 Joint Distribution Functions

We understand from above sections with probability distributions for single random variable. However, we are often interested in probability statements concerning two or more random variables. In order to deal with such probabilities, we define for any two random variables X and Y, the joint cumulative probability distribution function of X and Y by

$$F(a,b) = \mathbb{P}(X \leq a, Y \leq b) \text{ for } -\infty < a, b < \infty$$

The distribution of X can be obtained from the joint distribution of X and Y as follows

$$\begin{aligned} F_X(a) &= \mathbb{P}(X \leq a) \\ &= \mathbb{P}(X \leq a, Y < \infty) \\ &= \mathbb{P}(\lim_{b\to\infty} \{X \leq a, Y \leq b\}) \\ &= \lim_{b\to\infty} \mathbb{P}(\{X \leq a, Y \leq b\}) \\ &= \lim_{b\to\infty} F(a,b) \\ &= F(a,\infty) \end{aligned}$$

Note that the preceding set of equalities, we have once again made use of the fact that probability is a continuous set function. Similarly, the cumulative distribution function of Y is given by

$$\begin{aligned} F_Y(b) &= \mathbb{P}(Y \leq b) \\ &= \lim_{a\to\infty} F(a,b) \\ &= F(\infty, b) \end{aligned}$$

In the case when X and Y are both discrete random variables, it is convenient to define the joint probability mass function of X and Y by

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

the probability mass function of X can be obtained from $\mathbb{P}(x,y)$ by

$$\begin{aligned} p_X(x) &= \mathbb{P}(X = x) \\ &= \textstyle\sum_{x:\mathbb{P}(x,y)>0} \mathbb{P}(x,y) \end{aligned}$$

Similarly, we have

$$p_Y(y) = \sum_{x:\mathbb{P}(x,y)>0} \mathbb{P}(x,y)$$

We say that X and Y are jointly continuous if there exists a function $f(x,y)$, defined for all real x and y, having the property that for every set C of pairs of real numbers (that is, C is a set in the two-dimensional plane),

$$\mathbb{P}((X,Y) \in C) = \iint_{(x,y)\in C} f(x,y)dxdy$$

Example 6.1.1. IMPORTANT The joint density function of X and Y is given by

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

Compute (a) $\mathbb{P}(X > 1, Y < 1)$, (b) $\mathbb{P}(X < Y)$, and (c) $\mathbb{P}(X < a)$.

Answer. Please refer to the following

- Compute
$$\begin{aligned}\mathbb{P}(X>1, Y<1) &= \int_0^1 \int_1^\infty 2e^{-x}e^{-2y}dxdy \\ &= \int_0^1 2e^{-2y}\big(-e^{-x}|_1^\infty\big)dy \\ &= e^{-1}\int_0^1 2e^{-2y}dy \\ &= e^{-1}(1-e^{-2})\end{aligned}$$
- Compute
$$\begin{aligned}\mathbb{P}(X<Y) &= \iint_{(x,y):x<y} 2e^{-x}e^{-2y}dxdy \\ &= \int_0^\infty 2e^{-2y}(1-e^{-y})dy \\ &= \int_0^\infty 2e^{-2y}dy - \int_0^\infty 2e^{-3y}dy \\ &= 1-\tfrac{2}{3} \\ &= \tfrac{1}{3}\end{aligned}$$
- Compute
$$\begin{aligned}\mathbb{P}(X<a) &= \int_0^a \int_0^\infty 2e^{-2y}e^{-x}dydx \\ &= \int_0^a e^{-x}dx\end{aligned}$$

□

Example 6.1.2. The joint density of X and Y is given by

$$f(x,y) = \begin{cases} e^{-(x+y)} & \text{if } 0<x<\infty, 0<y<\infty \\ 0 & \text{otherwise}\end{cases}$$

Find the density function of the random variable X/Y.

Answer. Start by computing the distribution function of X/Y. For $a>0$,

$$\begin{aligned}F_{X/Y}(a) &= \mathbb{P}\left\{\tfrac{X}{Y}\le a\right\} \\ &= \iint_{x/y\le a} e^{-(x+y)}dxdy \\ &= \int_0^\infty \int_0^{ay} e^{-(x+y)}dxdy \\ &= \int_0^\infty (1-e^{-ay})e^{-y}dy \\ &= \left\{-e^{-y}+\tfrac{e^{-(a+1)y}}{a+1}\right\}\Big|_0^\infty \\ &= 1-\tfrac{1}{a+1}\end{aligned}$$

Differentiation shows the density function of X/Y is given by $f_{X/Y}(a) = 1/(a+1)^2$ for $0<a<\infty$. □

6.2 Independent Random Variables

The random variables X and Y are said to be independent if, for any two sets of real numbers A and B,

$$\mathbb{P}(X\in A, Y\in B) = \mathbb{P}(X\in A)\mathbb{P}(Y\in B)$$

In other words, X and Y are independent if, for all A and B, the events $E_A = \{X\in A\}$ and $F_B = \{Y\in B\}$ are independent.

It can be shown by using the three axioms of probability that the above equation will follow if and only if, for all a, b,

$$\mathbb{P}(X\le a, Y\le b) = \mathbb{P}(X\le a)\mathbb{P}(Y\le b)$$

Hence, in terms of the joint distribution function F of X and Y, X and Y are independent if

$$F(a,b) = F_X(a)F_Y(b) \text{ for all } a, b$$

Proposition 6.2.1. *The continuous (discrete) random variables X and Y are independent if and only if their joint probability density (mass) function can be expressed as*

$$f_{X,Y}(x,y) = h(x)g(y), -\infty < x < \infty, -\infty < y < \infty$$

Answer. Let us give the proof in the continuous case. First, note that independence implies that the joint density is the product of the marginal densities of X and Y, so the preceding factorization will hold when the random variables are independent. Now, suppose that

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

Then

$$\begin{aligned} 1 &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)dxdy \\ &= \int_{-\infty}^{\infty} h(x)dx \int_{-\infty}^{\infty} dy \\ &= C_1C_2 \end{aligned}$$

where $C_1 = \int_{-\infty}^{\infty} h(x)dx$ and $C_2 = \int_{-\infty}^{\infty} g(y)dy$. Also,

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy = C_2h(x) \\ f_Y(y) &= \int_{-\infty}^{\infty} f_{X,Y}(x,y)dx = C_1g(y) \end{aligned}$$

Since $C_1C_2 = 1$, it follows that

$$f_{X,Y}(x,y) = f_X(x)f_Y(y)$$

□

Example 6.2.2. IMPORTANT Let X, Y, Z be independent and uniformly distributed over $(0,1)$. Compute $\mathbb{P}(X \geq YZ)$.

Answer. Since

$$f_{X,Y,Z}(x,y,z) = f_X(x)f_Y(y)f_Z(z) = 1, 0 \leq x \leq 1, 0 \leq y \leq 1, 0 \leq z \leq 1$$

we have

$$\begin{aligned} \mathbb{P}(X \geq YZ) &= \iiint_{x \geq yz} f_{X,Y,Z}(x,y,z)dxdydz \\ &= \int_0^1 \int_0^1 \int_{yz}^1 dxdydz \\ &= \int_0^1 \int_0^1 (1-yz)dydz \\ &= \int_0^1 (1-\tfrac{z}{2})dz \\ &= \tfrac{3}{4} \end{aligned}$$

□

6.3 Sums of Independent Random Variables

It is often important to be able to calculate the distribution of $X + Y$ from the distributions of X and Y when X and Y are independent. Suppose that X and Y are

independent, continuous random variables having probability density functions f_X and f_Y. The cumulative distribution function of $X+Y$ is obtained as follows:

$$\begin{aligned} F_{X+Y}(a) &= \mathbb{P}(X+Y \leq a) \\ &= \iint_{x+y\leq a} f_X(x) f_Y(y) dxdy \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{a-y} f_X(x) f_Y(y) dxdy \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{a-y} f_X(x) dx f_Y(y) dy \\ &= \int_{-\infty}^{\infty} f_X(a-y) f_Y(y) dy \end{aligned}$$

The cumulative distribution function F_{X+Y} is called convolution of the distributions F_X and F_Y (the cumulative distribution functions of X and Y, respectively.

By differentiating the above equation, we find that the probability density function f_{X+Y} of $X+Y$ is given by

$$\begin{aligned} f_{X+Y}(a) &= \frac{d}{da} \int_{-\infty}^{\infty} F_X(a-y) f_Y(y) dy \\ &= \int_{-\infty}^{\infty} \frac{d}{da} F_X(a-y) f_Y(y) dy \\ &= \int_{-\infty}^{\infty} f_X(a-y) f_Y(y) dy \end{aligned}$$

Let us explore the relationship of two random variables. Recall gamma random variable has a density of the form

$$f(y) = \frac{\lambda e^{-\lambda y} (\lambda y)^{t-1}}{\Gamma(t)}, 0 < y < \infty$$

An important property of this family of distributions is that for a fixed value of λ, it is closed under convolutions.

Proposition 6.3.1. *If X and Y are independent gamma random variables with respective parameters (s,λ) and (t,λ), then $X+Y$ is a gamma random variable with parameters $(s+t,\lambda)$.*

$$\begin{aligned} f_{X+Y}(a) &= \frac{1}{\Gamma(s)\Gamma(t)} \int_0^a \lambda e^{-\lambda(a-y)} [\lambda(a-y)]^{t-1} \lambda e^{-\lambda y} (\lambda y)^{t-1} dy \\ &= K e^{-\lambda a} \int_0^a (a-y)^{s-1} y^{t-1} dy \\ &= K e^{-\lambda a} a^{s+t-1} \int_0^1 (1-x)^{s-1} x^{t-1} dx, \text{ by letting } x = \frac{y}{a} \\ &= C e^{-\lambda a} a^{s+t-1} \end{aligned}$$

where C is a constant that does not depend on a. But, as the preceding is a density function and thus must integrate to 1, the value of C is determined, and we have

$$f_{X+Y}(a) = \frac{\lambda e^{-\lambda a} (\lambda a)^{s+t-1}}{\Gamma(s+t)}$$

Hence, the result is proved.

Proposition 6.3.2. *If X_i, $i=1,...,n$, are independent random variables that are normally distributed with respective parameters μ_i, σ_i^2, $i=1,...,n$, then $\sum_{i=1}^n X_i$ is normally distributed with parameters $\sum_{i=1}^n \mu_i$ and $\sum_{i=1}^n \sigma_i^2$.*

Proposition 6.3.3. *If X and Y are independent Poisson random variables with respective parameters λ_1 and λ_2, compute the distribution of $X+Y$.*

6.4 Bivariate Transformations

IMPORTANT This part we discuss bivariate transformations $(X,Y) \to (U,V)$ where $U = g_1(X,Y)$ and $V = g_2(X,Y)$ are univariate transformations of the random vector (X,Y). Let us suppose that the transformation $(x,y) \to (g_1(x,y), g_2(x,y))$ to be

bijective and differentiable. We have that if $B \subset \mathbb{R}^2$, then $(U,V) \in B$ if and only $(X,Y) \in A$ where $A := \{(x,y) : (g_1(x,y), g_2(x,y)) \in B\}$. Thus,

$$\mathbb{P}((U,V) \in B) = \mathbb{P}((X,Y) \in A) \Rightarrow \int_B f_{U,V}(u,v)d(u,v) = \int_A f_{X,Y}(x,y)d(x,y)$$

Thus, the joint pdfs $f_{U,V}(u,v)$ and $fX,Y(x,y)$ must be related through the change of variables procedure for $(x,y) \to (u,v)$. In one dimension, recall from the previous review session that this involved the derivative of the transformation g. However, now there are two transformations g_1 and g_2. Here we introduce a concept called Jacobian matrix, and the determinant is

$$J := \det \begin{pmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{pmatrix} = \frac{\partial x}{\partial u}\frac{\partial y}{\partial v} - \frac{\partial y}{\partial u}\frac{\partial x}{\partial v}$$

Then the rectangle of differentials in X,Y space with area $dx \times dy$ goes to a parallelogram in U,V space with area roughly the determinant of the Jacobian J. Thus, we should have $d(x,y) = |U|d(u,v)$ and

$$f_{U,V}(u,v) = f_{X,Y}(h_1(u,v), h_2(u,v))|J|,$$

where $(u,v) \to (h_1(u,v), h_2(u,v))$ is the inverse transformation of $(x,y) \to (g_1(x,y), g_2(x,y))$.

Example 6.4.1. Let X,Y be independent $\mathcal{N}(0,1)$ random variables. Consider the transformation $U = X/Y$ and $V = |Y|$ (for $Y = 0$, we can let $(U,V) = (1,1)$ or any value since $\mathbb{P}(Y = 0) = 0$ and so this case is negligible). This transformation is not one-to-one, but is one-to-one when restricted to either positive or negative values of y. Let

$$A_1 = \{(x,y) : y > 0\}, A_2 = \{(x,y) : y < 0\}, A_0 = \{(x,y) : y = 0\}$$

which partition $\mathcal{A} = \mathbb{R}^2$. For either A_1 or A_2, if $(x,y) \in A_i$, $v = |y| > 0$ and for a fixed $v = |y|$, $u = x/y$ can be any real number. Thus,

$$\mathcal{B} = \{(u,v) : v > 0\}$$

The inverse transformations from $\mathcal{B}$ to A_1 and $\mathcal{B}$ to A_2 are given by $(u,v) \to (uv,v)$ and $(u,v) \to (-uv,-v)$. The determinants of both Jacobians are v. Then, using the fact that

$$f_{X,Y}(x,y) = \frac{1}{2\pi}e^{-x^2/2}e^{-y^2/2}$$

we have using our transformation law to obtain

$$f_{U,V}(u,v) = \frac{v}{\pi}e^{-(u^2+1)v^2/2}, u \in \mathbb{R}, v \in (0,\infty)$$

From this the marginal pdf of U can be computed to be

$$f_U(u) = \int_0^\infty \frac{v}{\pi}e^{-(u^2+1)v^2/2}dv = \frac{1}{\pi(u^2+1)}, u \in \mathbb{R}$$

This is the pdf for Cauchy distribution. In other words, the ratio of two independent standard normal random variables is a Cauchy random variable.

6.5 Hierarchical Models and Mixture Distributions

The topics of hierarchical models and mixture distribution are direct application from some of the most basic theorems discussed in jointly distributed random variables. In other words, this part we discuss some nice properties related with multiple random variables.

Theorem 6.5.1. ***Law of Iterated Expectation*** *If X, Y are any two random variables, then $\mathbb{E}(X) = \mathbb{E}[\mathbb{E}(X|Y)]$.*

Proof. Letting $f(x, y)$ denote the joint pdf of X, Y, we have

$$\mathbb{E}(X) = \iint xf(x,y)dxdy = \int \Big[\int xf(x|y)dx \Big] f_Y(y)dy = \mathbb{E}[\mathbb{E}[X|Y]]$$

□

Theorem 6.5.2. ***Law of total Variance*** *For any two random variables X, Y,*

$$var(X) = \mathbb{E}[var(X|Y)] + var(\mathbb{E}[X|Y])$$

provided that the expectations exist.

Proof. This is another one of those proofs where we decompose a square error as a sum of two squares:

$$\begin{aligned} \text{var}(X) &= \mathbb{E}[(X - \mathbb{E}[X])^2] \\ &= \mathbb{E}[(X - \mathbb{E}[X|Y] + \mathbb{E}[X|Y] - \mathbb{E}[X])^2] \\ &= \mathbb{E}[(X - \mathbb{E}(X|Y))^2] + \mathbb{E}[(\mathbb{E}(X|Y) - \mathbb{E}[X])^2] \end{aligned}$$

We obtained the last term by expanding the square and applying the linearity of expectation. We did the cross term disappear? We can finish by verifying

$$\begin{aligned} \mathbb{E}[(X - \mathbb{E}(X|Y))^2] &= \mathbb{E}[\mathbb{E}[(X - \mathbb{E}[X|Y])^2|Y]] = \mathbb{E}[\text{var}(X|Y)] \\ \mathbb{E}[(\mathbb{E}(X|Y) - \mathbb{E}[X])^2] &= \text{var}(\mathbb{E}[X|Y]) \end{aligned}$$

□

Example 6.5.3. IMPORTANT **Binomial-Poisson Hierarchy** An insect lays a large number of eggs, each surviving with probability p. On the average, how many eggs will survive? Let the "large number" of eggs laid be a random variable $Y \sim \text{Poisson}(\lambda)$ and let the number of survivors be a random variable X. Assuming each egg's survival is independent, we have $X|Y \sim \text{Binomial}(Y, p)$. This gives a hierarchical model. We have

$$\begin{aligned} \mathbb{P}(X = x) &= \textstyle\sum_{y=0}^{\infty} \mathbb{P}(X = x, Y = y) \\ &= \textstyle\sum_{y=0}^{\infty} \mathbb{P}(X = x|Y = y)\mathbb{P}(Y = y) \\ &= \textstyle\sum_{y=x}^{\infty} \Big(\binom{y}{x} p^x (1-p)^{y-x} \Big) \Big(\frac{e^{-\lambda}\lambda^y}{y!} \Big) \\ &= \frac{(\lambda p)^x e^{-\lambda}}{x!} \textstyle\sum_{y=x}^{\infty} \frac{((1-p)\lambda)^{y-x}}{(y-x)!} \\ &= \frac{(\lambda p)^x e^{-\lambda}}{x!} \textstyle\sum_{y=x}^{\infty} \frac{((1-p)\lambda)^{t}}{t!}, \text{ let } t = y - x \\ &= \frac{(\lambda p)^x}{x!} e^{-\lambda p} \\ &= \frac{(\lambda p)^x}{x!} e^{-\lambda p} \end{aligned}$$

so that $X \sim \text{Poisson}(\lambda p)$. Thus, $\mathbb{E}X = \lambda p$ to answer the original question.

Example 6.5.4. IMPORTANT Let X and Y be iid $\mathcal{N}(0, 1)$ random variables. Consider $Z := \text{sign}(Y)X$ where $\text{sign}(y) := 1$ if $y > 0$ and $\text{sign}(y) := -1$ if $y \leq 0$.

(a). Find the distribution of Z

(b). Compute the covariance of X and Z

(c). Determine $\mathbb{P}(X+Z=0)$

(d). Are X and Z independent? (Give a precise mathematical argument)

Answer. Let us address them one by one.

(a). We have

$$\begin{aligned}\mathbb{P}(Z\leq z_0) &= \mathbb{P}(X\leq z_0|\text{sign}(Y)=1)\mathbb{P}(\text{sign}(Y)=1)\\ &\quad +\mathbb{P}(-X\leq z_0|\text{sign}(Y)=-1)\mathbb{P}(\text{sign}(Y)=-1)\\ &= \tfrac{1}{2}\mathbb{P}(X\leq z_0)+\tfrac{1}{2}\mathbb{P}(X\geq -z_0)\\ &= \mathbb{P}(X\leq z_0),\ \text{because } X \text{ is symmetric}\end{aligned}$$

which implies that $Z\sim\mathcal{N}(0,1)$.

(b). The covariance is

$$\begin{aligned}\text{cov}(X,Z) &= \mathbb{E}XZ-\mathbb{E}X\mathbb{E}Z\\ &= \mathbb{E}[\mathbb{E}[XZ|\text{sign}(Y)]]\\ &= \tfrac{1}{2}\mathbb{E}[XZ|\text{sign}(Y)=1]+\tfrac{1}{2}\mathbb{E}[XZ|\text{sign}(Y)=-1]\\ &= \tfrac{1}{2}\mathbb{E}X^2-\tfrac{1}{2}\mathbb{E}X^2=0\end{aligned}$$

(c). We have

$$\mathbb{P}(X+Z=0)=\frac{1}{2}\mathbb{P}(X+X=0)+\frac{1}{2}\mathbb{P}(X-X=0)=1/2$$

(d). No, X and Z are not independent. If they are independent, then $X+Z$ would be Gaussian as well, but $\mathbb{P}(X+Z=0)>0$, which is a contradiction.

□

Example 6.5.5. IMPORTANT Let $X\in\mathbb{R}^d$ be a centered normal random vector and $A\in\mathbb{R}^{d\times d}$ a fixed symmetric matrix. Denote by Y an independent copy of X. Show that

$$X^TAX-Y^TAY\overset{d}{=}2X^TAY$$

Hint: $(X\pm Y)/\sqrt{2}$ are iid random vectors following the same distribution as X.

Answer. Consider

$$\begin{aligned}\tfrac{1}{\sqrt{2}}(X\pm Y) &\overset{\text{d}}{=} X \text{ while } Y=X\\ \text{thus } X=_d \tfrac{1}{\sqrt{2}}(X+Y) &\quad\text{and}\quad Y=_d\tfrac{1}{\sqrt{2}}(X-Y)\\ \text{together } (X,Y) &=_d (\tfrac{X+Y}{\sqrt{2}},\tfrac{X-Y}{\sqrt{2}})\\ X^TY &=_d \tfrac{1}{2}(X^TX-X^TY+Y^TX-Y^TY)\\ X^TAY &=_d \tfrac{1}{2}(X^TAX-Y^TAY)\\ 2X^TAY &=_d X^TAX-Y^TAY\end{aligned}$$

which is desired and here $=_d$ means left-hand-side has the same distribution as the right-hand-side. □

Example 6.5.6. IMPORTANT Suppose X_1,X_2 are iid $\mathcal{N}(0,1)$.

(a). Find the joint distribution of X_1+X_2 and X_1-X_2.

(b). Show that $2X_1X_2$ has the same distribution as $X_1^2-X_2^2$.

Answer. We address the following.

1. Let $Y_1 = X_1 + X_2$ and $Y_2 = X_1 - X_2$, and then the Jacobian would be

$$J = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$$

and thus we have determinant to be -2. In addition, notice that we have the following identity

$$x_1^2 + x_2^2 = \frac{1}{2}[(x_1 + x_2)^2 + (x_1 - x_2)^2 = \frac{1}{2}(y_1^2 + y_2^2)]$$

Thus, by law of transformation, we have joint density of (Y_1, Y_2) to be

$$f_{Y_1,Y_2}(y_1, y_2) = \frac{1}{4\pi} e^{-\frac{1}{4}(y_1^2 + y_2^2)}$$

and thus Y_1, Y_2 are iid $\mathcal{N}(0, 2)$.

□

Example 6.5.7. Let X and Y be iid $\mathcal{N}(0, 1)$ random variables, and define $Z = \min(X, Y)$. Prove that $Z^2 \sim \chi_1^2$.

Answer. We claim the cdf of Z^2 to be $F_{Z^2}(z) = 1 - 2F_X(-\sqrt{z})$. We have

$$\begin{aligned} F_{Z^2}(z) &= \mathbb{P}(\min(X, Y)^2 \leq z) \\ &= \mathbb{P}(-\sqrt{z} \leq \min(X, Y) \leq \sqrt{z}) \\ &= \mathbb{P}(\min(X, Y) \leq \sqrt{z}) - \mathbb{P}(\min(X, Y) \leq -\sqrt{z}) \\ &= (1 - \mathbb{P}(\min(X, Y) > \sqrt{z}) - (1 - \mathbb{P}(\min(X, Y) > -\sqrt{z}) \\ &= \mathbb{P}(\min(X, Y) > -\sqrt{z}) - \mathbb{P}(\min(X, Y) > \sqrt{z}) \\ &= \mathbb{P}(X > -\sqrt{z})\mathbb{P}(Y > -\sqrt{z}) - \mathbb{P}(X > \sqrt{z})(\mathbb{P}(Y > \sqrt{z}) \end{aligned}$$

where we use the independence of X and Y to establish the last equality. Since X and Y are identically distributed, $\mathbb{P}(X > a) = \mathbb{P}(Y > a) = 1 - F_X(a)$. So,

$$F_{Z^2}(z) = (1 - F_X(-\sqrt{z}))^2 - (1 - F_X(\sqrt{z}))^2 = 1 - 2F_X(-\sqrt{z})$$

and thus differentiating gives pdf

$$f_{Z^2}(z) = \frac{1}{\sqrt{2\pi}} e^{-z/2} z^{-1/2}$$

This is the pdf of a $\chi^2(1)$ random variable. □

7 Properties of Expectation

Go back to Table of Contents. Please click TOC

7.1 Introduction

This section we develop and exploit additional properties of expected values. Recall expected value of the random variable X

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

where X is a discrete random variable with probability mass function $\mathbb{P}(x)$, and by

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

when X is a continuous random variable with probability density function $f(x)$.

7.2 Expectation of Sums of Random Variables

Let us begin by introducing one of the most important properties in expectation of random variables.

Proposition 7.2.1. *If X and Y have a joint probability mass function $\mathbb{P}(x,y)$, then*

$$\mathbb{E}[g(X,Y)] = \sum_y \sum_x g(x,y)\mathbb{P}(x,y)$$

If X and Y have a joint probability density function $f(x,y)$, then

$$E[g(X,Y)] = \int_{-\infty}^{\infty} g(x,y)f(x,y)dxdy$$

Let us prove the above property.

Proof. Suppose we have random variables X and Y that are jointly continuous with joint density function $f(x,y)$ and when $g(X,Y)$ is a nonnegative random variable. Because $g(X,Y) \geq 0$, we have

$$\mathbb{E}[g(X,Y)] = \int_0^{\infty} \mathbb{P}(g(X,Y) > t)dt$$

We can write

$$\mathbb{P}(g(X,Y) > t) = \iint_{(x,y):g(x,y)>t} f(x,y)dydx$$

shows that

$$\mathbb{E}[g(X,Y)] = \int_0^{\infty} \iint_{(x,y):g(x,y)>t} f(x,y)dydxdt$$

Interchanging the order of integration gives

$$\begin{aligned} E[g(X,Y)] &= \int_x \int_y \int_{t=0}^{g(x,y)} f(x,y)dtdydx \\ &= \int_x \int_y g(x,y)f(x,y)dydx \end{aligned}$$

Thus, the result is proven when $g(X,Y)$ is a nonnegative random variable. □

An application of such property can be used in the following application.

Example 7.2.2. IMPORTANT An accident occurs at a point X that is uniformly distributed on a road of length L. At the time of the accident, an ambulance is at a location Y that is also uniformly distributed on the road. Assuming that X and Y are independent, find the expected distance between the ambulance and the point of the accident.

Answer. We want to compute $\mathbb{E}[|X-y|]$. We have the joint density function of X and Y to be

$$f(x,y)=\frac{1}{L^2}, 0<x<L, 0<y<L$$

and it follows from the property above

$$\mathbb{E}[|X-Y|]=\frac{1}{L^2}\int_0^L\int_0^L |x-y|dydx$$

Now we can do the math

$$\begin{aligned}\int_0^L |x-y|dy &= \int_0^x (x-y)dy+\int_x^L (y-x)dy\\ &= \frac{x^2}{2}+\frac{L^2}{2}-\frac{x^2}{2}-x(L-x)\\ &= \frac{L^2}{2}+x^2-xL\end{aligned}$$

Therefore,

$$\begin{aligned}\mathbb{E}[|X-Y|] &= \frac{1}{L^2}\int_0^L \left(\frac{L^2}{2}+x^2-xL\right)dx\\ &= \frac{L}{3}\end{aligned}$$

□

An important application of the above property is the following. Suppose $\mathbb{E}[X]$ and $\mathbb{E}[Y]$ are both finite and let $g(X,Y)=X+Y$. Then, in the continuous case,

$$\begin{aligned}\mathbb{E}[X+Y] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x+y)f(x,y)dxdy\\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}xf(x,y)dydx+\int_{-\infty}^{\infty}yf(x,y)dxdy\\ &= \mathbb{E}[X]+\mathbb{E}[Y]\end{aligned}$$

The same result holds in general; thus, whenever $\mathbb{E}[X]$ and $\mathbb{E}[Y]$ are finite,

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

Example 7.2.3. Let $X_1,...,X_n$ be independent and identically distributed random variables having distribution function F and expected value μ. Such as sequence of random variables is said to constitute a sample from the distribution F. The quantity

$$\bar{X}=\sum_{i=1}^n \frac{X_i}{n}$$

is called the sample mean. Compute $\mathbb{E}[\bar{X}]$.

Answer. Compute

$$\begin{aligned}\mathbb{E}[\bar{X}] &= \mathbb{E}\left[\sum_{i=1}^n \frac{X_i}{n}\right]\\ &= \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^n X_i\right]\\ &= \frac{1}{n}\sum_{i=1}^n \mathbb{E}[X_i]\\ &= \mu \text{ since } \mathbb{E}X_i\equiv\mu\end{aligned}$$

We conclude that the expected value of the sample mean is μ, the mean of the distribution. When the distribution mean μ is unknown, the sample mean if often used in statistics to estimate it. □

7.3 Moments of the Number of Events that Occur

Let us look at an example.

Example 7.3.1. Suppose that there are N distinct types of coupons and that, independently of past types collected, each new one obtained is type j with probability p_j, $\sum_{j=1}^{N} p_j = 1$. Find the expected value and variance of the number of different types of coupons that appear among the first n collected.

Answer. We will find it more convenient to work with the number of uncollected types. Let Y equal the number of types of coupons collected, and let $X = N - Y$ denote the number of uncollected types. With A_i defined as the event that there are no type i coupons in the collection, X is equal to the number of the events $A_1, ..., A_N$ that occur. Because the types of the successive coupons collected are independent, and, with probability $1 - p_i$ each new coupon is not type i, we have

$$\mathbb{P}(A_i) = (1 - p_i)^n$$

Hence, $\mathbb{E}[X] = \sum_{i=1}^{N}(1 - p_i)^n$, from which it follows that

$$\mathbb{E}[Y] = N - \mathbb{E}[X] = N - \sum_{i=1}^{N}(1 - p_i)^n$$

Similarly, because each of the n coupons collected is neither a type i nor a type j coupon, with probability $1 - p_i - p_j$, we have

$$\mathbb{P}(A_i A_j) = (1 - p_i - p_j)^n, i \neq j$$

Thus,

$$\mathbb{E}[X(X-1)] = 2\sum_{i<j} \mathbb{P}(A_i A_j) = 2\sum_{i<j}(1 - p_i - p_j)^n$$

or

$$\mathbb{E}[X^2] = 2\sum_{i<j}(1 - p_i - p_j)^n + \mathbb{E}[X]$$

Hence, we obtain

$$\begin{aligned} \text{var}(Y) &= \text{var}(X) \\ &= \mathbb{E}[X^2] - (\mathbb{E}X)^2 \\ &= 2\textstyle\sum_{i<j}(1 - p_i - p_j)^n + \sum_{i=1}^{N}(1 - p_i)^n - \left(\sum_{i=1}^{N}(1 - p_i)^n\right)^2 \end{aligned}$$

□

7.4 Covariance, Variance of Sums, and Correlations

The following proposition shows that the expectation of a product of independent random variables is equal to the product of their expectations.

Proposition 7.4.1. *If X and Y are independent, then, for any functions h and g,*

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

Answer. Suppose that X and Y are jointly continuous with joint density $f(x,y)$. Then we have

$$\begin{aligned}\mathbb{E}[g(X)h(Y)] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x)h(y)f(x,y)dxdy \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x)h(y)f_X(x)f_Y(y)dxdy \\ &= \int_{-\infty}^{\infty} h(y)f_Y(y)dy \int_{-\infty}^{\infty} g(x)f_X(x)dx \\ &= \mathbb{E}[h(Y)]\mathbb{E}[g(X)]\end{aligned}$$

□

Definition 7.4.2. IMPORTANT The covariance between X and Y, denoted by $\text{Cov}(X,Y)$, is defined by

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

Upon expanding the right side of the preceding definition, we see that

$$\begin{aligned}\text{Cov}(X,Y) &= \mathbb{E}[XY - \mathbb{E}[X]Y - X\mathbb{E}[Y] + \mathbb{E}[Y]\mathbb{E}[X]] \\ &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] - \mathbb{E}[X]\mathbb{E}[Y] + \mathbb{E}[X]\mathbb{E}[Y] \\ &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]\end{aligned}$$

Proposition 7.4.3. *There are the following properties:*

- $cov(X,Y) = cov(Y,X)$
- $cov(X,X) = var(X)$
- $cov(aX,Y) = aCov(X,Y)$
- $cov(\sum_{i=1}^n X_i, \sum_{j=1}^m Y_j) = \sum_{i=1}^n \sum_{j=1}^m cov(X_i,Y_j)$

7.5 Conditional Expectation

If X and Y are jointly discrete random variables, then the conditional probability mass function of X, given that $Y=y$, is defined for all y such that $\mathbb{P}(Y=y)>0$, by

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

It is therefore natural to define, in this case, the conditional expectation of X given that $Y=y$, for all values of y such that $p_Y(y)>0$, by

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

7.6 Moment Generating Functions

IMPORTANT The moment generating function $M(t)$ of the random variable X is defined for all real values of t by

$$M(t) = \mathbb{E}[e^{tX}] = \begin{cases} \sum_x e^{tx}\mathbb{P}(x) & \text{if } X \text{ is discrete with mass function } \mathbb{P}(x) \\ \int_{-\infty}^{\infty} e^{tx}f(x)dx & \text{if } X \text{ is continuous with density } f(x) \end{cases}$$

We call $M(t)$ the moment generating function because all of the moments of X can be obtained by successively differentiating $M(t)$ and then evaluating the result at $t=0$. For example,

$$\begin{aligned}M'(t) &= \tfrac{d}{dt}\mathbb{E}[e^{tX}] \\ &= \mathbb{E}[\tfrac{d}{dt}(e^{tX})] \\ &= \mathbb{E}[Xe^{tX}]\end{aligned}$$

where we have assumed that the interchange of the differentiation and expectation operators is legitimate. That is, we have assumed that

$$\frac{d}{dt}\left[\sum_x e^{tx}\mathbb{P}(x)\right] = \sum_x \frac{d}{dt}[e^{tx}\mathbb{P}(x)]$$

in the discrete case and

$$\frac{d}{dt}\left[\int e^{tx}f(x)dx\right] = \int \frac{d}{dt}[e^{tx}f(x)]dx$$

in the continuous case. This assumption can almost always be justified and, indeed, is valid for all of the distributions considered in this book. Hence, from above the first derivative of moment generating function, evaluated at $t = 0$, we obtain

$$M'(0) = \mathbb{E}[X]$$

Similarly,

$$\begin{aligned} M''(t) &= \tfrac{d}{dt}M'(t) \\ &= \tfrac{d}{dt}\mathbb{E}[Xe^{tX}] \\ &= \mathbb{E}\left[\tfrac{d}{dt}(Xe^{tX}\right] \\ &= \mathbb{E}[X^2e^{tX}] \end{aligned}$$

Thus, we have

$$M''(0) = \mathbb{E}[X^2]$$

In general, the nth derivative of $M(t)$ is given by

$$M^n(t) = \mathbb{E}[X^n e^{tX}], n \geq 1$$

implying that

$$M^n(0) = \mathbb{E}[X^n], n \geq 1$$

Example 7.6.1. IMPORTANT If X is a binomial random variable with parameters n and p, then

$$\begin{aligned} M(t) &= \mathbb{E}[e^{tX}] \\ &= \textstyle\sum_{k=0}^n e^{tk}\binom{n}{k}p^k(1-p)^{n-k} \\ &= \textstyle\sum_{k=0}^n \binom{n}{k}(pe^t)^k(1-p)^{n-k} \\ &= (pe^t + 1 - p)^n \end{aligned}$$

where the last equality follows from the binomial theorem. Differentiation yields

$$M'(t) = n(pe^t + 1 - p)^{n-1}pe^t$$

Thus, we have

$$\mathbb{E}[X] = M'(0) = np$$

Differentiating a second time yeilds

$$M''(t) = n(n-1)(pe^t + 1 - p)^{n-2}(pe^t)^2 + n(pe^t + 1 - p)^{n-1}pe^t$$

so

$$\mathbb{E}[X^2] = M''(0) = n(n-1)p^2 + np$$

The variance of X is given by

$$\begin{aligned} \text{Var}(X) &= \mathbb{E}[X^2] - (\mathbb{E}(X))^2 \\ &= n(n-1)p^2 + np - n^2p^2 \\ &= n\mathbb{P}(1-p) \end{aligned}$$

Example 7.6.2. IMPORTANT If X is a Poisson random variable with parameter λ, then

$$\begin{aligned} M(t) &= \mathbb{E}[e^{tX}] \\ &= \sum_{n=0}^{\infty} \frac{e^{tn} e^{-\lambda} e^{-\lambda} \lambda^n}{n!} \\ &= e^{-\lambda} \sum_{n=0}^{\infty} \frac{(\lambda e^t)^n}{n!} \\ &= e^{-\lambda} e^{\lambda} e^{t} \\ &= \exp(\lambda(e^t - 1)) \end{aligned}$$

Differentiating yields

$$\begin{aligned} M'(t) &= \lambda e^t \exp(\lambda(e^t - 1)) \\ M''(t) &= (\lambda e^t)^2 \exp(\lambda(e^t - 1)) + \lambda e^t \exp(\lambda(e^t - 1)) \end{aligned}$$

Thus,

$$\begin{aligned} \mathbb{E}[X] &= M'(0) = \lambda \\ \mathbb{E}[X^2] &= M''(0) = \lambda^2 + \lambda \\ \text{var}(X) &= \mathbb{E}[X^2] - (\mathbb{E}(X))^2 \\ &= \lambda \end{aligned}$$

Hence, both the mean and the variance of the Poisson random variable equal λ.

Example 7.6.3. Let us find the first and second moment of exponential distribution withy parameter λ.

$$\begin{aligned} M(t) &= \mathbb{E}[e^{tX}] \\ &= \int_0^{\infty} e^{tx} \lambda e^{-\lambda x} dx \\ &= \frac{\lambda}{\lambda - t} \text{ for } t < \lambda \end{aligned}$$

We note from this derivation that for the exponential distribution, $M(t)$ is defined only for values of t less than λ. Differentiation of $M(t)$ yields

$$M'(t) = \frac{\lambda}{(\lambda - t)^2}, M''(t) = \frac{2\lambda}{(\lambda - t)^3}$$

Hence,

$$\mathbb{E}[X] = M'(0) = \frac{1}{\lambda}, \mathbb{E}[X^2] = M''(0) = \frac{2}{\lambda^2}$$

The variance of X is given by

$$\begin{aligned} \text{var}(X) &= \mathbb{E}[X^2] - (\mathbb{E}(X))^2 \\ &= \frac{1}{\lambda^2} \end{aligned}$$

Let us summarize the moment generating function in the following table.

Table 7.1 Discrete Probability Distribution.

	Probability mass function, $p(x)$	Moment generating function, $M(t)$	Mean	Variance
Binomial with parameters n, p; $0 \le p \le 1$	$\binom{n}{x} p^x (1-p)^{n-x}$ $x = 0, 1, \ldots, n$	$(pe^t + 1 - p)^n$	np	$np(1-p)$
Poisson with parameter $\lambda > 0$	$e^{-\lambda} \frac{\lambda^x}{x!}$ $x = 0, 1, 2, \ldots$	$\exp\{\lambda(e^t - 1)\}$	λ	λ
Geometric with parameter $0 \le p \le 1$	$p(1-p)^{x-1}$ $x = 1, 2, \ldots$	$\frac{pe^t}{1 - (1-p)e^t}$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
Negative binomial with parameters r, p; $0 \le p \le 1$	$\binom{n-1}{r-1} p^r (1-p)^{n-r}$ $n = r, r+1, \ldots$	$\left[\frac{pe^t}{1 - (1-p)e^t}\right]^r$	$\frac{r}{p}$	$\frac{r(1-p)}{p^2}$

Table 7.2 Continuous Probability Distribution.

	Probability density function, $f(x)$	Moment generating function, $M(t)$	Mean	Variance
Uniform over (a, b)	$f(x) = \begin{cases} \frac{1}{b-a} & a < x < b \\ 0 & \text{otherwise} \end{cases}$	$\frac{e^{tb} - e^{ta}}{t(b-a)}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Exponential with parameter $\lambda > 0$	$f(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}$	$\frac{\lambda}{\lambda - t}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Gamma with parameters $(s, \lambda), \lambda > 0$	$f(x) = \begin{cases} \frac{\lambda e^{-\lambda x}(\lambda x)^{s-1}}{\Gamma(s)} & x \geq 0 \\ 0 & x < 0 \end{cases}$	$\left(\frac{\lambda}{\lambda - t}\right)^s$	$\frac{s}{\lambda}$	$\frac{s}{\lambda^2}$
Normal with parameters (μ, σ^2)	$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2} \quad -\infty < x < \infty$	$\exp\left\{\mu t + \frac{\sigma^2 t^2}{2}\right\}$	μ	σ^2

Example 7.6.4. IMPORTANT Calculate the distribution of $X + Y$ when X and Y are independent Poisson random variables with means respectively λ_1 and λ_2.

Answer. We compute the following

$$\begin{aligned} M_{X+Y}(t) &= M_X(t) M_Y(t) \\ &= \exp(\lambda_1(e^t - 1)) \exp(\lambda_2(e^t - 1)) \\ &= \exp((\lambda_1 + \lambda_2)(e^t - 1)) \end{aligned}$$

Hence, $X + Y$ is Poisson distributed with mean $\lambda_1 + \lambda_2$. □

It is also possible to define the joint moment generating function of two or more random variables. This is done as follows: for any n random variables $X_1, ..., X_n$, the joint moment generating function, $M(t_1, .l.., t_n)$, is defined, for all real values of $t_1, ..., t_n$, by

$$M(t_1, ..., t_n) = \mathbb{E}[e^{t_1 X_1 + \cdots + t_n X_n}]$$

The individual moment generating functions can be obtained from $M(t_1, ..., t_n)$ by letting all but one of the t_j's be 0. That is,

$$M_{X_i}(t) = \mathbb{E}[e^{tX_i}] = M(0, ..., 0, t, 0, ..., 0)$$

where the t is in the ith place.

It can be proven that the joint moment generating function $M(t_1, ..., t_n)$ uniquely determines the joint distribution of $X_1, ..., X_n$. This result can then be used to prove that the n random variables $X_1, ..., X_n$ are independent if and only if

$$M(t_1, ..., t_n) = M_{X_1}(t_1) \dots M_{X_n}(t_n)$$

For the proof in one direction, if the n random variables are independent, then

$$\begin{aligned} M(t_1, ..., t_n) &= \mathbb{E}[e^{(t_1 X_1 + \cdots + t_n X_n)}] \\ &= \mathbb{E}[e^{t_1 X_1} \dots e^{t_n X_n}] \\ &= \mathbb{E}[e^{t_1 X_1}] \dots \mathbb{E}[e^{t_n} X_n] \text{ by independence} \\ &= M_{X_1}(t_1) \dots M_{X_n}(t_n) \end{aligned}$$

8 Limit Theorems

Go back to Table of Contents. Please click TOC

8.1 Introduction

The most important theoretical results in probability theory are limit theorems. Of these, the most important are those classified either under the heading laws of large numbers of under the heading central limit theorems. Usually, theorems are considered to be laws of large numbers if they are concerned with stating conditions under which the average of a sequence of random variables converges (in some sense) to the expected average.

8.2 Convergence Theory

This subsection we discuss the most important component in regarding to describing the limiting behaviors of a random variable or a sequence of random variables. Specifically, we address some important convergence theories.

Point-wise Convergence Let $\{X_n\}$ be a sequence of random variables defined on a sample space Ω. Let us consider a single sample point $\omega \in \Omega$ and a generic random variable X_n belonging to the sequence. Here X_n is a function such that $X_n : \Omega \to \mathbb{R}$. However, once we fix ω, the realization $X_n(\omega)$ associated to the sample point ω is just a real number. By the same token, once we fix ω, the sequence $\{X_n(\omega)\}$ is just a sequence of real numbers. Therefore, for a fixed ω, it is very easy to assess whether the sequence $\{X_n(\omega)\}$ is convergent; this is done by using the usual definition of convergence of sequences of real numbers. If, for a fixed ω, the sequence $\{X_n(\omega)\}$ is convergent, we denote its limit by $X(\omega)$, to underline that the limit depends on the specific ω we have fixed.

Definition 8.2.1. Let $\{X_n\}$ be a sequence of random variables defined on a sample space Ω. We say that $\{X_n\}$ is point-wise convergent to a random variable X defined on Ω if and only if $\{X_n(\omega)\}$ converges to $X(\omega)$ for all $\omega \in \Omega$. We use the following notation to indicate convergence point-wise:

$$X_n \to X \text{p.w.}$$

8.3 Chebyshev's Inequality and the Weak Law of Large Numbers

Let us start with Markov's Inequality. IMPORTANT

Proposition 8.3.1. *If X is a random variable that takes only nonnegative values, then for any value $a > 0$,*

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}[X]}{a}$$

Proof. For $a > 0$, let

$$I = \begin{cases} 1 & \text{if } X \geq a \\ 0 & \text{otherwise} \end{cases}$$

and note that, since $X \geq 0$, then we have $I \leq \frac{X}{a}$. Taking expectations of the preceding inequality yields

$$\mathbb{E}[I] \leq \frac{\mathbb{E}[X]}{a}$$

which, because $\mathbb{E}[I] = \mathbb{P}(X \geq a)$, proves the result. □

Please see the following example. The code is written in programming language R.

```
# Package
library(quantmod)

# Get Data
getSymbols('FB')
data <- FB
head(data); tail(data)
plot(data[,4], main = "Chart: Stock Price ($)")

# Define Return
head(data[,4]); head(lag(data[,4]))
return <- data[,4]/lag(data[,4]) - 1
summary(return);
plot(return, main = "Chart: Return of Stock Price")
hist(return, breaks = 100, main = "Histogram of Returns")

# Markov's Inequality
a <- 0.02
p <- mean(na.omit(ifelse(return > a, 1, 0))); p
expected.value <- mean(na.omit(return)); expected.value
expected.value/a

# Summarize in table
Summary <- cbind(
Probability.of.Event = p,
Expectation.over.Arbitrary.Value = expected.value/a
); Summary


# Define Function
Markov.Inequality <- function(a = 0.1) {
# Get Data
getSymbols('FB')
data <- FB
head(data); tail(data)
plot(data[,4], main = "Chart: Stock Price ($)")

# Define Return
head(data[,4]); head(lag(data[,4]))
return <- data[,4]/lag(data[,4]) - 1
summary(return);
plot(return, main = "Chart: Return of Stock Price")
hist(return, breaks = 100, main = "Histogram of Returns")

# Markov's Inequality
a <- a
p <- mean(na.omit(ifelse(return > a, 1, 0))); p
expected.value <- mean(na.omit(return)); expected.value
expected.value/a

# Summarize in table
```

```
Summary <- cbind(
Probability.of.Event = p,
Expectation.over.Arbitrary.Value = expected.value/a
); Summary

# Output
return(Summary)
}

# Run
lapply(c(0.01, 0.05, 0.1, 0.15, 0.2), Markov.Inequality)
```

However, this will not give us correct answer. Who can spot the problem? Please review the following.

Example 8.3.2.
```
# The first line does not satisfy the inequality
# Can anybody spot the mistake?
# Ans:

# Define Function
Markov.Inequality <- function(a = 0.1) {
# Get Data
getSymbols('AAPL')
data <- AAPL
head(data); tail(data)
plot(data[,4], main = "Chart: Stock Price ($)")

# Define Return
head(data[,4]); head(lag(data[,4]))
return <- data[,4]/lag(data[,4]) - 1
summary(return); plot(return, main = "Chart: Return of Stock Price")
hist(return, breaks = 100, main = "Histogram of Returns")

# Markov's Inequality
a <- a
number.of.pos.obs <- sum(na.omit(ifelse(return > 0, 1, 0)));
number.of.pos.obs
number.of.event <- sum(na.omit(ifelse(return > a, 1, 0)));
number.of.event
p <- number.of.event/number.of.pos.obs; p
expected.value <- mean(na.omit(return[ifelse(return > 0, 1, 0) == 1,]));
expected.value
expected.value/a

# Summarize in table
Summary <- cbind(
Value.of.Interest = a,
Probability.of.Event = p,
Expectation.over.Arbitrary.Value = expected.value/a
); Summary

# Output
return(Summary)
```

```
}

# Run
lapply(c(0.001, 0.005, 0.01, 0.05, 0.1, 0.15, 0.2), Markov.Inequality)
Report <- matrix(unlist(lapply(c(0.001, 0.005, 0.01, 0.05, 0.1, 0.15, 0.2),
Markov.Inequality)), nrow = 3);
Report <- t(Report); colnames(Report) <- c("Value.of.Interest",
"Prob.of.Event", "Exp.Over.Arbi.Value"); Report
```

Proposition 8.3.3. *IMPORTANT Chebyshev's Inequality. If X is a random variable with finite mean μ and variance σ^2, then for any value $k > 0$,*

$$\mathbb{P}((X-\mu)^2 \geq k) \leq \frac{\sigma^2}{k^2}$$

Proof. Since $(X-\mu)^2$ is a nonnegative random variable, we can apply Markov's Inequality (with $a = k^2$) to obtain

$$\mathbb{P}((X-\mu)^2 \geq k^2) \leq \frac{\mathbb{E}[(X-\mu)^2]}{k^2}$$

But since $(X-\mu)^2 \geq k^2$ if and only if $|X-\mu| \geq k$, than what is written above is equivalent to

$$\mathbb{P}(|X-\mu| \geq k) \leq \frac{\mathbb{E}[(X-\mu)^2]}{k^2} = \frac{\sigma^2}{k^2}$$

and we are done. □

```
# Get Data
getSymbols('FB')

# Define Function
Chebyshev.Inequality <- function(k = 0.1) {
# Check
if (k < 0) {
return(
print(paste(
"Error␣Message:␣Chebyshev␣Inequality␣requires␣k␣to
be␣nonnegative.␣Please␣check␣the␣value␣of␣k."
)
)
)
} else {
# Get Data
data <- AAPL
head(data); tail(data)
plot(data[,4], main = "Chart:␣Stock␣Price␣($)")

# Define Return
head(data[,4]); head(lag(data[,4]))
return <- data[,4]/lag(data[,4]) - 1
summary(return); plot(return, main = "Chart:␣Return␣of␣Stock␣Price")
hist(return, breaks = 100, main = "Histogram␣of␣Returns")

# Markov's Inequality
```

```
mu <- mean(na.omit(return))
p <- mean(na.omit(as.numeric(abs(return - mu) > k)))
sigma <- var(na.omit(return))

# Summarize in table
Summary <- cbind(
Value.of.Interest = k,
Probability.of.Event = p,
Variance.over.Arbitrary.Value.Square = sigma/k^2
)

# Output
return(Summary)
}
}

# Run
#lapply(c(0.001, 0.005, 0.01, 0.05, 0.1, 0.15, 0.2), Chebyshev.Inequality)
Report <- matrix(unlist(lapply(c(0.001, 0.005, 0.01, 0.05, 0.1, 0.15, 0.2),
Markov.Inequality)), nrow = 3);
Report <- t(Report); colnames(Report) <- c("Value.of.Interest",
"Prob.of.Event", "Var.over.Arbi.Value.Sq"); Report

# What happen if k is negative?
Chebyshev.Inequality(-0.1)
```

Example 8.3.4. If X is uniformly distributed over the interval $(0, 10)$, then what is the probability that X and value 5 has distance greater than 4.

Answer. Let us work this out a step at a time.

- First, we compute expectation: $\mathbb{E}(X) = \frac{a+b}{2} = \frac{10}{2} = 5$;
- Second, we compute variance: $\mathbb{E}(X) = \frac{(b-a)^2}{12} = \frac{100}{12} = \frac{25}{3}$
- Compute the result using Chebyshev's Inequality:

$$\mathbb{P}(|X - 5| > 4) = \frac{25/3}{16} = 0.52$$

□

Theorem 8.3.5. *Weak Law of Large Numbers. Let $X_1, X_2, ...$ be a sequence of independent and identically distributed random variables, each having finite mean $\mathbb{E}[X_i] = \mu$. Then, for any $\epsilon > 0$,*

$$\mathbb{P}\left\{\left|\frac{X_1 + \cdots + X_n}{n} - \mu\right| \geq \epsilon\right\} \to 0 \text{ as } n \to \infty$$

Proof. We shall prove this theorem only under the additional assumption that the random variables have a finite variance σ^2. Now, since

$$\mathbb{E}\left[\frac{X_1 + \cdots + X_n}{n}\right] = \mu \text{ and } \text{Var}\left(\frac{X_1 + \cdots + X_n}{n}\right) = \frac{\sigma^2}{n}$$

it follows from Chebyshev's Inequality that

$$\mathbb{P}\left\{\left|\frac{X_1+\cdots+X_n}{n}-\mu\right|\geq\epsilon\right\}\to 0 \text{ as } n\to\infty$$

□

Definition 8.3.6. Convergence in probability A sequence of random variables $X_1,...,X_n$ converges in probability to a random variable X if, for every $\epsilon>0$,

$$\lim_{n\to\infty}\mathbb{P}(|X_n-X|\geq\epsilon)=0,$$

or equivalently $\lim_n \mathbb{P}(|X_n-X|<\epsilon)=1$. We will abbreviate this as $X_n\xrightarrow{\text{P}}X$.

8.4 The Weak Law of Large Numbers

Theorem 8.4.1. ***Weak Law of Large Numbers*** *Let $X_1,...,X_n$ be iid random variables with $\mathbb{E}X_i=\mu$ and $var(X_i)=\sigma^2$ and assume the variance is finite. Define $\bar{X}_n:=\frac{1}{n}\sum_{i=1}^n X_i$. Then we have $X_n\xrightarrow{P}\mu$.*

Proof. We have for every $\epsilon>0$:

$$\mathbb{P}(|\bar{X}_n-\mu|\geq\epsilon)=\mathbb{P}((\bar{X}_n-\mu)^2\geq\epsilon)\leq\frac{\text{var}(\bar{X}_n}{\epsilon^2}=\frac{\sigma^2}{n\epsilon^2}\to 0$$

as $n\to\infty$ which completes the proof. □

Remark 8.4.2. The sample variance of a sequence of iid random variables can be shown consistent using the same procedure. One can go to index page to look for "consistency of sample variance".

8.5 The Central Limit Theorem

The Central Limit Theorem is one of the most remarkable results in probability theory. Loosely put, it states that the sum of a large number of independent random variables has a distribution that is approximately normal. Hence, it not only provides a simple method for computing approximate probabilities for sums of independent random variables, but also helps explain the remarkable fact that the empirical frequencies of so many natural populations exhibit bell-shaped (that is, normal) curves.

Theorem 8.5.1. *The Central Limit Theorem. Let $X_1,X_2,...$ be a sequence of independent and identically distributed random variables, each having mean μ and variance σ^2. Then the distribution of*

$$\frac{X_1+\ldots X_n-n\mu}{\sigma\sqrt{n}}$$

tends to the standard normal as $n\to\infty$. That is, for $-\infty<a<\infty$,

$$\mathbb{P}\left\{\frac{X_1+\cdots+X_n-n\mu}{\sigma\sqrt{n}}\right\}\to\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{a}e^{-x^2/2}dx \text{ as } n\to\infty$$

Lemma 8.5.2. *Let $Z_1,Z_2,...$ be a sequence of random variables having distribution functions F_{Z_n} and moment generating functions M_{Z_n}, $n\geq 1$, and let Z be a random variable having distribution function F_Z and moment generating function M_Z. If $M_{Z_n}(t)\to M_Z(t)$ for all t, then $F_{Z_n}(t)\to F_Z(t)$ for all t at which $F_Z(t)$ is continuous.*

If we let Z be a standard normal random variable, then, since $M_Z(t) = e^{t^2/2}$, it follows from above lemma that if $M_{Z_n}(t) \to e^{t^2/2}$ as $n \to \infty$, then $F_{Z_n}(t) \to \Phi(t)$ as $n \to \infty$.

Now let us produce the following proof.

Proof. Suppose $\mu = 0$ and $\sigma^2 = 1$. We prove the theorem under the assumption that the moment generating function of the X_i, $M(t)$, exists and is finite. Now the moment generating function of $X_i/\sqrt{n}$ is given by

$$\mathbb{E}\left[\exp\left\{\frac{tX_i}{\sqrt{n}}\right\}\right] = M\big(\frac{t}{\sqrt{n}}\big)$$

Thus, the moment generating function of $\sum_{i=1}^n X_i/\sqrt{n}$ is given by $\big[M\big(\frac{t}{\sqrt{n}}\big)\big]^n$. Let

$$L(t) = \log M(t)$$

and note that

$$\begin{aligned} L(0) &= 0 \\ L'(0) &= \tfrac{M'(0)}{M(0)} \\ &= \mu \\ &= 0 \\ L''(0) &= \tfrac{M(0)M''(0) - [M'(0)]^2}{[M(0)]^2} \\ &= \mathbb{E}[X^2] \\ &= 1 \end{aligned}$$

Now, to prove the theorem, we must show that $[M(t/\sqrt{n})]^n \to e^{t^2/2}$ as $n \to \infty$, or, equivalently, that $nL(t/\sqrt{n}) \to t^2/2$ as $n \to \infty$. To show this, note that

$$\begin{aligned} \lim_{n\to\infty} \tfrac{L(t/\sqrt{n})}{n^{-1}} &= \lim_{n\to\infty} \tfrac{-L'(t/\sqrt{n})n^{-3/2}t}{-2n^{-2}}, \text{ by L'Hopital's Rule} \\ &= \lim_{n\to\infty}\left[\tfrac{L'(t/\sqrt{n})t}{2n^{-1/2}}\right] \\ &= \lim_{n\to\infty}\left[\tfrac{-L''(t/\sqrt{n})n^{-3/2}t^2}{-2n^{-3/2}}\right], \text{ again by L'Hopital's Rule} \\ &= \lim_{n\to\infty}\left[L''\big(\tfrac{t}{\sqrt{n}}\tfrac{t^2}{2}\right] \\ &= \tfrac{t^2}{2} \end{aligned}$$

Thus, the central limit theorem is proven when $\mu = 0$ and $\sigma^2 = 1$. The result now follows in the general case by considering the standardized random variables $X_i^* = (X_i - \mu)/\sigma$ and applying the preceding result, since $\mathbb{E}[X_i^*] = 0$, $\text{var}(X_i^*) = 1$. □

Theorem 8.5.3. *Central limit Theorem for independent random variables. Let $X_1, X_2, ...$ be a sequence of independent random variables having respective means and variances $\mu_i = \mathbb{E}[X_i]$, $\sigma_i^2 = Var(X_i)$. If (a) the X_i are uniformly bounded – that is, if for some M, $\mathbb{P}(|X_i| < M) = 1$ for all i, and (b) $\sum_{i=1}^\infty \sigma_i^2 = \infty$ then we have*

$$\mathbb{P}\left\{\frac{\sum_{i=1}^n (X_i - \mu_i)}{\sqrt{\sum_{i=1}^n \sigma_i^2}} \le a\right\} \to \Phi(a) \to \text{ as } n \to \infty$$

Theorem 8.5.4. ***Delta Method*** *Let Y_n be a sequence of random variables that satisfies $\sqrt{n}(Y_n - \theta) \to \mathcal{N}(0, \sigma^2)$ in distribution. For a given function g and specific value of θ, suppose that $g'(\theta)$ exists and is not 0. Then*

$$\sqrt{n}[g(Y_n) - g(\theta)] \to_d \mathcal{N}(0, \sigma^2[g'(\theta)]^2)$$

while the symbol $\to_d$ means convergence in distribution

Proof. The Taylor expansion of $g(Y_n)$ around $Y_n = \theta$ is

$$g(Y_n) = g(\theta) + g'(\theta)(Y_n - \theta) + \mathcal{O}(n)$$

while $\mathcal{O}(n)$, the remainder of the Taylor expansion, vanishes. Since by the Weak Law of Large Numbers we have $Y_n \to \theta$ in probability, we can use Slutsky's Theorem to obtain

$$\sqrt{n}[g(Y_n) - g(\theta)] =_{\text{d}} g'(\theta)\sqrt{n}(Y_n - \theta)$$

while here the symbol $=_{\text{d}}$ means the left-hand-side has the same asymptotic distribution of the right-hand-side. □

Theorem 8.5.5. ***Second-order Delta Method*** *Let Y_n be a sequence of random variables that satisfies $\sqrt{n}(Y_n - \theta) \to \mathcal{N}(0, \sigma^2)$ in distribution. For a given function g and a specific value of θ, suppose that $g'(\theta) = 0$ and $g''(\theta)$ exists and is not 0. Then we have*

$$n[g(Y_n) - g(\theta)] \to_d \sigma^2 \frac{g''(\theta)}{2}\chi_1^2$$

Proof. The idea here is to apply Taylor expansion just like the proof we show in Delta Method. Consider

$$g(Y_n) = g(\theta) + g'(\theta)(Y_n - \theta) + \frac{1}{2}g''(\theta)(Y_n - \theta)^2 + \mathcal{O}(n)$$

while the remainder $\mathcal{O}(n)$ vanishes. We know that by Central Limit Theorem we have $Y_n - \theta$ goes to a normal distribution so that means $(Y_n - \theta)^2$ follows χ_1^2. In other words, we may write

$$g(Y_n) - g(\theta) \approx \frac{1}{2}g''(\theta)(Y_n - \theta)^2$$

and by Slutsky's Theorem, we obtain

$$n(g(Y_n) - g(\theta)) \to_{\text{d}} \sigma^2 \frac{1}{2}g''(\theta)\chi_1^2$$

which is desired. □

8.6 The Strong Law of Large Numbers

The strong law of large numbers is probably the best-known result in probability theory. 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 8.6.1. *Let $X_1, X_2, ...$ be a sequence of independent and identically distributed random variables, each having a finite mean $\mu = \mathbb{E}[X_i]$. Then, with probability 1,*

$$\frac{X_1 + \cdots + X_n}{n} \to \mu \text{ as } n \to \infty$$

Remark 8.6.2. Here we say converges in probability and what we mean is the following

$$\mathbb{P}(\lim_{n\to\infty} (X_1 + \cdots + X_n)/n = \mu) = 1$$

8.7 Other Inequalities

We are sometimes confronted with situations in which we are interested in obtaining an upper bound for a probability of the form $\mathbb{P}(X - \mu \geq a)$, where a is some positive value and when only the mean $\mu = \mathbb{E}[X]$ and variance $\sigma^2 = \text{var}(X)$ of the distribution of X are known. Naturally, since $X - \mu \geq a > 0$ implies that $|X - \mu| \geq a$, it follows from Chebyshev's inequality that

$$\mathbb{P}(|X - \mu| \geq a) \leq \frac{\sigma^2}{a^2} \text{ when } a > 0$$

Proposition 8.7.1. *One-sided Chebyshev Inequality. If X is a random variable with mean 0 and finite variance σ^2, then, for any $a > 0$,*

$$\mathbb{P}(X \geq a) \leq \frac{\sigma^2}{\sigma^2 + a^2}$$

Answer. Let $b > 0$ and note that

$$X \geq a \text{ is equivalent to } X + b \geq a + b$$

Hence,

$$\begin{aligned} \mathbb{P}(X \geq a) &= \mathbb{P}(X + b \geq a + b) \\ &\leq \mathbb{P}((X + b)^2 \geq (a + b)^2) \end{aligned}$$

where the inequality is obtained by noting that since $a + b > 0$, $X + b \geq a + b$ implies $(X + b)^2 \geq (a + b)^2$. Upon applying Markov's inequality, the preceding yields that

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}[(X + b)^2]}{(a + b)^2} = \frac{\sigma^2 + b^2}{(a + b)^2}$$

Letting $b = \sigma^2/a$ [which is easily seen to be the value of b that minimizes $(\sigma^2 + b^2)/(a + b)^2$] gives the desired result. □

Proposition 8.7.2. *If $\mathbb{E}[X] = \mu$ and $Var(X) = \sigma^2$, then, for $a > 0$,*

$$\mathbb{P}(X \geq \mu + a) \leq \frac{\sigma^2}{\sigma^2 + a^2}$$

$$\mathbb{P}(X \leq \mu - a) \leq \frac{\sigma^2}{\sigma^2 + a^2}$$

Proposition 8.7.3. *Chernoff Bounds.*

$$\begin{aligned} \mathbb{P}(X \geq a) &\leq e^{-ta} M(t) \text{ for all } t > 0 \\ \mathbb{P}(X \leq a) &\leq e^{-ta} M(t) \text{ for all } t < 0 \end{aligned}$$

Since the Chernoff bounds hold for all t in either the positive or negative quadrant, we obtain the best bound on $\mathbb{P}(X \geq a)$ by using the t that minimizes $e^{-ta} M(t)$.

Proposition 8.7.4. *Jensen's Inequality. If $f(x)$ is a convex function, then*

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

provided that the expectations exist and are finite.

9 Homework

Go back to Table of Contents. Please click TOC

This section is private. Please inquire us for more information in class.

10 Exam Review

Go back to Table of Contents. Please click TOC

This section we review some of the most important concepts based on the coverage of Midterm 1, Midterm 2, and the Final exam. In general, we respect our students and we believe our students have the capabilities to study the materials without any issues unless told otherwise. If students face difficulty, we are here to help. However, we do not make the assumption that someone is lack the knowledge of a certain area in any topics as well as all the pre-requisite of statistics. In regard of this assumption, we encourage all students to speak up and let us know the questions they have in order to tackle them one by one.

10.1 1st Midterm

The following notions are important for this midterm.

1. Chapter 1. Counting Principle, Permutation, Combinations, Binomial Theorem (and its Propositions).
2. Chapter 2. Sample Space, Union, Intersection, Complement, Mutually Exclusive, Inclusion-Exclusion Identity
3. Chapter 3. Conditional Probability, Multiplication Rule of Probability, Bayes's Formula, Denominator of Bayes's Formula by Law of Total Probability.

Counting

Example 10.1.1. An image has size 6 by 6. In computer vision, scholars like to use a small filter to extract information from the image. This action is called convolution operation which we denote as $\odot$. The requirement for two matrices to apply this operation is that they need to be the same dimension. Suppose the filter has size 3 by 3 and we start to apply $\odot$ from the first row and the first column. We roll this filter towards the right. Once we hit the end we move to the next row and do the same.

- What is the dimension for the output matrix?
- If we apply $\odot$ on the odd rows and columns, what is the dimension for the output matrix?
- If the filter has size 2×3, what is the dimension for the output matrix?

Answer. We compute the following:

- First, compute the size of the output matrix to be $6-3+1=4$ and the dimension is 4×4.
- Next, $(6-3+1)/2 = 2$ and the dimension is 2×2.
- Third, we compute the new number of rows and columns respectively. For number of rows, we have $6-2+1=5$ and for number of columns, we have $6-3+1=4$. Thus, the dimension of the output matrix is 5×4.

□

Example 10.1.2. Let A be a set with $|A| = n < \infty$. How many distinct subsets does A have?

Answer. Assume $A = \{a_1, a_2, ..., a_n\}$. We can look at this problem in the following way. To choose a subset B, we perform the following experiment. First we decide whether or not $a_1 \in B$ (so two choices), ..., and finally we decide whether or not $a_n \in B$ (so again two choices). By the multiplication principle, the total number of subsets is then given by $2 \times 2 \times \cdots \times 2 = 2^n$. To check out answer, let's assume $A = \{1, 2\}$. Then our formula states that there are 4 possible subsets. Indeed the subsets are

- $\{\} = \emptyset$
- $\{1\}$
- $\{2\}$
- $\{1, 2\}$

□

Example 10.1.3. In the literature of modern day Artificial Intelligence, a branch of AI research that studies language problem is called Natural Language Processing (short for NLP). In this field, there are the following sub-branches of studies: Recurrent Neural Network (denoted as R), Long Short Term Memory (denoted as L), and bi-directional RNN (denoted as B). There are 10 papers studying Recurrent Neural Network and 8 papers studying bi-directional RNN.

- There are 4 papers studying both Long Short Term Memory and Recurrent Neural Network.
- There are 3 papers studying both Recurrent Neural Network and bi-directional RNN.
- The total number of papers studying Recurrent Neural Network or Long Short Term Memory or bi-directional RNN is 21.
- There is no paper studying both Long Short Term Memory and bi-directional RNN.

How many papers studying only Recurrent Neural Network?

Proof. There are two approaches. First, let us draw a Venn Diagram.

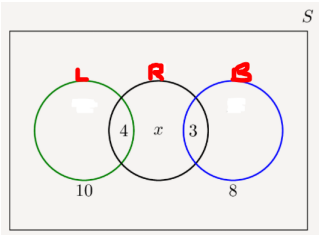


Figure 3: This is the Venn Diagram

and the goal is simply to find $x = 21 - 10 - 8 = 3$.

Alternatively, we can use the formula for inclusion-exclusion. Hence, we write the following

$$\begin{aligned} 21 &= |L| + |R| + |B| \\ &= |L| + |R| + |B| - |L \cap R| - |L \cap B| - |R \cap B| + |L \cap R \cap B| \\ &= 10 + 8 + |B| - 0 - 4 - 3 + 0 \end{aligned}$$

which solves for $|B| = 10$. Then in the field of B, we subtract the number of papers with overlaps in other field to obtain the final answer, $10 - 4 - 3 = 10 - 7 = 3$.

Remark. Both approaches (either using Venn Diagram or inclusion-exclusion formula) should get you the same answer. □

Permutation

Example 10.1.4. There are n people sitting in a row and this will give us $n!$ different arrangements. A classical example (Homework #1) is the following. Consider delegates from 10 countries with R, F, E, U and rest of the 6 more countries sitting in a row. We want to satisfy two premises (1) F and E are sitting together, i.e. FE or EF. (2) R and U are not sitting together.

Answer. The key is to work out (1) first and then subtract the compliment of (2) to obtain the answer.

(1) We want

1	2	3	4	5	6	7	8	9	10
F	E	$-$	$-$	$-$	$-$	$-$	$-$	$-$	$-$
$-$	F	E	$-$	$-$	$-$	$-$	$-$	$-$	$-$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$-$	$-$	$-$	$-$	$-$	$-$	$-$	$-$	F	E

so this is total of 9 possible outcomes and with FE and EF being different arrangement. The rest of the 8 countries may sit in different arrangement. That is, (2)(9)(8!) possibilities.

(2) We want the compliment of (2). (2) says R and U are not sitting together. The compliment of this event is referring to the situation while R and U are indeed sitting together. While satisfying (1) the same time, we have compliment of (2) to be FERU, FEUR, EFRU, EFUR. so this is $2 \cdot 2$ and the rest 8 countries may sit in different arrangements.

Thus, we have final answer (1) minus compliment of (2). This gives us

$$(2)(9)(8!) - (2)(2)(8!) = 564,440$$

□

Example 10.1.5. Another classical example is the grid problem from homework and also textbook. One can refer to the following. Consider a grid that has size 3 by 4. There are dots at each intersection and assume we can only move along the lines. We want to move from A to B that requires 3 ups and 4 rights. It does not matter which move you go first.

Answer. Thus, we have

$$\binom{3+4}{3} = \binom{7}{3} = \binom{7}{4} = 35$$

□

Positive Solution

Example 10.1.6. Given an equation in the form of

$$\sum_{i=1}^{r} x_i = n$$

you need to be familiar how to solve this type of problems. Be aware of both premises (1) assume positive solution, and (2) assume non-negative solution for X_i's.

You should recall formula

$$\binom{n-1}{r-1} \text{ for positive integer-valued vector } (x_1, ..., x_r)$$

and

$$\binom{n+r-1}{r-1} \text{ for non-negative integer-valued vector } (x_1, ..., x_r)$$

A classical example is from Homework 1. Given 8 identical blackboards to be identically distributed among 4 schools, we want to find out

(a) How many distributions are possible? This is

$$X_1 + X_2 + X_3 + X_4 = 8$$

while allowing "0" so we have

Answer.

$$\binom{8+4-1}{4-1} = \binom{11}{3} = 165$$

□

(b) How many if at least 1 is distributed to each school? This is not allowing "0". Hence, we have

Answer.

$$\binom{8-1}{4-1} = \binom{7}{3} = 35$$

□

Example 10.1.7. Common practice in data science suppose that a data frame has rows and columns. Each column is a variable. If a data set has p columns, then the data has p variables. Sometimes multiple variables together can form an impact to help us make predictions. We call this impact interactions (or interaction effect). An interaction effect formed by two variables would be called a two-way interaction.

- From p variables, what is the total number of possible choices to form 2-way interaction? Denote this number as n.
- If only one of these 2-way interactions is important, what is the chance of finding it?
- Suppose we allocate n choices to among r possible computers. We denote the evaluation of whether each choice of 2-way interaction is important a job. In other words, we have n jobs here. Each computer must certain positive number of jobs. How many different possible ways to distribute them?
- Why is it better to use more than one computers? Justify your answer.

Answer. Suppose we have a data with p columns (i.e. p variables).

- From p variables to choose two variables, we have a total of $\binom{p}{2}$ possible choices.
- If two variables form an interaction, out of p variables there are $\binom{p}{2}$ possible ways of choosing 2-way interactions. If there are only one of them that is important, the chance of picking out this interaction would be $1/\binom{p}{2}$
- Suppose we have n number of jobs (or choices) and we want to distribute n jobs to r computers. Another assumption is non-negative solutions for each computers. Hence, we can apply the formula $\binom{n-1}{r-1}$.
- Suppose $n = 3$ and $r = 2$. We have $c_1 + c_2 = 3$ where c_1 and c_2 refer to two computers. The problem assumes that each computer takes a second to finish a job. Here based on the additive formula, we have solutions $\binom{3-1}{2-1} = 2$ and alternatively we can list out all of them using brute force, i.e. $\{(1,2),(2,1)\}$. For each of the solution, we spend 2 seconds in total i.e. because $\max(1,2) = \max(2,1) = 2$. If we do not have two computers and use only one computer for all the jobs, we would have spend 3 seconds. This means using more computers is always faster then using one computers.

□

Bayes' Formula and Conditional Probability

Example 10.1.8. Bayes' Formula is definitely within the scope of this exam. You should be familiar with all sorts of formulas in this arena. Let us recall Bayes' Formula

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

Please also be aware of the identity (aka multiplication rule of probability).

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

You should also be aware of the related formulas as one can always derive Bayes' formula in fancy ways depending on the problem.

A classical example is from Homework 3. An ectopic pregnancy is twice as likely to develop when the pregnant woman is a smoker as it is a non-smoker. If 32 percent of women of childbearing age are smokers, what percentage of women having ectopic pregnancies are smokers?

Answer. Let E be the event that a pregnant women has an ectopic pregnancy, and S be the event that they are smokers. We know that $\mathbb{P}(E|S) = 2\mathbb{P}(E|S^c)$ and $\mathbb{P}(S) = 0.32$. Then by Bayes' Theorem, we find

$$\begin{aligned}
\mathbb{P}(S|E) &= \frac{\mathbb{P}(E|S)\mathbb{P}(S)}{\mathbb{P}(E|S)\mathbb{P}(S)+\mathbb{P}(E|S^c)\mathbb{P}(S^c)} \\
&= \frac{2\mathbb{P}(E|S^c)(0.32)}{2\mathbb{P}(E|S^c)(0.32)+\mathbb{P}(E|S^c)(1-0.32)} \\
&= \frac{0.64}{0.64+0.68} \\
&= \frac{0.64}{1.32} = \frac{32}{66}
\end{aligned}$$

□

One can also refer to another problem in Homework 3. A total of 48 percent of the women and 37 percent of the men who took a certain "quit smoking" class remained nonsmokers for at least one year after completing the class. These people then attended a success party at the end of a year. If 62 percent of the original class was male,

1. what percentage of those attending the party were women?
2. what percentage of the original class attended the party?

Answer. We answer the question in the following

1. Let A be the event that a person attends a party. Let W be the event that the person is a woman, and $M = W^c$ be the event that this person is a man. Then by Bayes' Theorem
$$\begin{aligned}\mathbb{P}(W|A) &= \frac{\mathbb{P}(A|W)\mathbb{P}(W)}{\mathbb{P}(A|W)\mathbb{P}(W)+\mathbb{P}(A|M)\mathbb{P}(M)} \\ &= \frac{0.48\times 0.38}{0.48\times 0.38+0.37\times 0.62} \\ &= 0.44\end{aligned}$$
2. By the law of total probability, we have that
$$\mathbb{P}(A) = \mathbb{P}(A|W)\mathbb{P}(W) + \mathbb{P}(A|M)\mathbb{P}(M) = 0.48\times 0.38 + 0.37\times 0.62 = 0.41$$

□

10.2 2nd Midterm

The following notions are important for this midterm.

1. Chapter 1. Counting Principle, Permutation, Combinations, Binomial Theorem (and its Propositions).
2. Chapter 2. Sample Space, Union, Intersection, Complement, Mutually Exclusive, Inclusion-Exclusion Identity
3. Chapter 3. Conditional Probability, Multiplication Rule of Probability, Bayes's Formula, Denominator of Bayes's Formula by Law of Total Probability.
4. Chapter 4. Density function (PDF), Distribution function (CDF), Expectation (Mean), Variance, Famous distributions (binomial, Poisson, geometric, negative binomial).
5. Chapter 5. Uniform. Normal. Exponential. Memoryless Property. Application on Memoryless Property.

Expected Value and Variance

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

The expected value of X is a weighted average of the possible values that X can take on, each value being weighted by the probability that X assumes it.

Example 10.2.1. Find $E[X]$, where X is the outcome when we roll a fair die.

Answer. Since $\mathbb{P}(1) = \mathbb{P}(2) = \mathbb{P}(3) = \mathbb{P}(4) = \mathbb{P}(5) = \mathbb{P}(6) = 1/6$, we obtain

$$E[X] = 1\big(\frac{1}{6}\big) + 2\big(\frac{1}{6}\big) + 3\big(\frac{1}{6}\big) + 4\big(\frac{1}{6}\big) + 5\big(\frac{1}{6}\big) + 6\big(\frac{1}{6}\big) = 7/2$$

□

Example 10.2.2. Calculate $\text{Var}(X)$ if X represents the outcome when a fair die is rolled.

Answer. You can easily find $E[X] = \frac{7}{2}$. Now, we find

$$\begin{aligned} E[X^2] &= 1^2(1/6) + 2^2(1/6) + 3^2(1/6) + 4^2(1/6) + 5^2(1/6) + 6^2(1/6) \\ &= (91)(1/6) \end{aligned}$$

and thus we have variance

$$\text{Var}(X) = \frac{91}{6} - \left(\frac{7}{2}\right)^2 = \frac{35}{12}$$

□

PDF/PMF and CDF/CMF

Example 10.2.3. Suppose X is a continuous random variable whose probability density function is

$$f(x) = \begin{cases} C(4x - 2x^2) & 0 < x < 2 \\ 0 & \text{else} \end{cases}$$

1. What is the value of C?
2. Find $\mathbb{P}(X > 1)$.

Answer. We have the following

1. Since f is a probability density function, we must have

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

and we can solve $C\int_0^2 (4x - 2x^2)dx = 1$. After integration, we have $C(2x^2 - \frac{2x^3}{3})|_{x=0}^2 = 1$ and we have result $C = \frac{3}{8}$.

2. $\mathbb{P}(X > 1) = \int_1^{\infty} f(x)dx = \frac{3}{8}\int_1^2 (4x - 2x^2)dx = \frac{1}{2}$.

□

Example 10.2.4. The amount of time in hours that a computer functions before breaking down is a continuous random variable with probability density function given by

$$f(x) = \begin{cases} \lambda e^{-x/100} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

What is the probability that

1. a computer will function between 50 and 150 hours before breaking down?
2. it will function for fewer than 100 hours?

Answer. We solve the parts accordingly

1. Since $1 = \int_{-\infty}^{\infty} f(x)dx = \lambda\int_0^{\infty} e^{-x/100}dx$, we can take integral and obtain $1 = -\lambda(100)e^{-x/100}|_0^{\infty} = 100\lambda$. We can solve for $\lambda = \frac{1}{100}$. Then we can proceed to find the probability

$$\begin{aligned} \mathbb{P}(50 < X < 1500) &= \int_{50}^{150} \tfrac{1}{100} e^{-x/100}dx \\ &= -e^{-x/100}|_{50}^{150} \\ &= e^{-1/2} - e^{-3/2} \\ &= 0.383 \end{aligned}$$

2. I will leave this to you as an exercise.

□

In general, we say that X is a uniform random variable on the interval (α, β) if the probability density function of X is given by

$$f(x) = \begin{cases} \frac{1}{\beta-\alpha} & \text{if } \alpha < x < \beta \\ 0 & \text{else} \end{cases}$$

Since $F(a) = \int_{-\infty}^{a} f(x)dx$, it follows that

$$f(x) = \begin{cases} 0 & a \leq \alpha \\ \frac{1}{\beta-\alpha} & \text{if } \alpha < x < \beta \\ 1 & a \geq \beta \end{cases}$$

Example 10.2.5. Let X be uniformly distributed over (α, β). Find (a) $E[X]$ and (b) $\text{Var}(X)$.

Answer. We proceed accordingly

1. Compute
$$\begin{aligned} E[X] &= \int_{-\infty}^{\infty} xf(x)dx \\ &= \int_{\alpha}^{\beta} \frac{x}{\beta-\alpha}dx \\ &= \frac{\beta^2-\alpha^2}{2(\beta-\alpha)} \\ &= \frac{\beta+\alpha}{2} \end{aligned}$$

2. To find $\text{Var}(X)$, first calculate $E[X^2]$.
$$\begin{aligned} E[X^2] &= \int_{\alpha}^{\beta} \frac{1}{\beta-\alpha}x^2dx \\ &= \frac{\beta^3-\alpha^3}{3(\beta-\alpha)} \\ &= \frac{\beta^2+\alpha\beta+\alpha^2}{3} \end{aligned}$$

   Hence,
$$\text{Var}(X) = \frac{\beta^2+\alpha\beta+\alpha^2}{3} - \frac{(\alpha+\beta)^2}{4} = \frac{(\beta-\alpha)^2}{12}$$

□

Example 10.2.6. If X is uniformly distributed over $(0, 10)$, calculate the probability that $X < 3$.

Answer. Compute $\mathbb{P}(X < 3) = \int_0^3 \frac{1}{10}dx = \frac{3}{10}$. □

We say that X is a normal random variable, or simply that X is normally distributed, with parameters μ and σ^2 if the density of X is given by

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

for $-\infty < x < \infty$. The density function is a bell-shaped curve that is symmetric about μ.

Example 10.2.7. Find $E[X]$ and $\text{Var}(X)$ when X is a normal random variable with parameters μ and σ^2.

Answer. Let us start by finding the mean and variance of the standard normal random variable $Z = (X - \mu)/\sigma$. We have

$$\begin{aligned} E[Z] &= \int_{-\infty}^{\infty} x f_Z(x) dx \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x e^{-x^2/2} dx \\ &= -\frac{1}{\sqrt{2\pi}} e^{-x^2/2} |_{-\infty}^{\infty} \\ &= 0 \end{aligned}$$

Thus,

$$\begin{aligned} \text{Var}(Z) &= E[Z^2] \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^2 e^{-x^2/2} dx, \text{ IBP: let } \mu = x \text{ and } d\nu = x e^{-x^2/2} \\ &= \frac{1}{\sqrt{2\pi}} \Big(- x e^{-x^2/2} |_{-\infty}^{\infty} + \underbrace{\int_{-\infty}^{\infty} e^{-x^2/2} dx}_{=1} \Big) \\ &= 1 \end{aligned}$$

Because $X = \mu + \sigma Z$, the preceding yields the results

$$E[X] = \mu + \sigma E[Z] = \mu$$

and

$$\text{Var}(X) = \sigma^2 \text{Var}(Z) = \sigma^2$$

□

Example 10.2.8. If X, the gain from an investment, is a normal random variable with mean μ and variance σ^2, then because the loss is equal to the negative of the gain, the VAR of such an investment is that value ν such that

$$0.01 = \mathbb{P}(-X > \nu)$$

We compute the following

$$\begin{aligned} 0.01 &= P\left(\frac{-X+\mu}{\sigma} > \frac{\nu+\mu}{\sigma} \right) \\ &= 1 - \Phi(\frac{\nu+\mu}{\sigma}) \end{aligned}$$

and from table we know $\Phi(2.33) = 0.99$ so we know $\frac{\nu+\mu}{\sigma} = 2.33$. That is, $\nu = \text{VAR} = 2.33\sigma - \mu$. Consequently, among set of investments all of whose gains are normally distributed, the investment having the smallest VAR is the one having the largest value of $\mu - 2.33\sigma$.

Remark 10.2.9. Please repeat the above analysis for

1. Discrete: Binomial, Poisson, Geometric.
2. Continuous: Uniform, Normal, Exponential.

Remark 10.2.10. Resources:

1. StatLect Website: `https://www.statlect.com/probability-distributions/`
2. Univariate Distribution Relationships: `http://www.math.wm.edu/~leemis/chart/UDR/UDR.html`

10.3 Final Exam

The following notions are important for this final exam.

1. Chapter 1. Counting Principle, Permutation, Combinations, Binomial Theorem (and its Propositions).
2. Chapter 2. Sample Space, Union, Intersection, Complement, Mutually Exclusive, Inclusion-Exclusion Identity
3. Chapter 3. Conditional Probability, Multiplication Rule of Probability, Bayes's Formula, Denominator of Bayes's Formula by Law of Total Probability.
4. Chapter 4. Density function (PDF), Distribution function (CDF), Expectation (Mean), Variance, Famous distributions (binomial, Poisson, geometric, negative binomial).
5. Chapter 5. Uniform. Normal. Exponential. Memoryless Property. Application on Memoryless Property.
6. Chapter 6. Joint Cumulative Probability Distribution Function, Joint Probability Mass Function
7. Chapter 7. MGF, Expectation, Variance.
8. Chapter 8. Markov, Chebyshev.

Please be aware of the following:

- Exam is cumulative from Chapter 1 to Chapter 8.
- Midterm I & II are very good reference of the final. For problems asking topics discussed in Chapter 5 and before, Midterm I & II provide very good insight.
- For problems asking topics discussed in Chapter 6 and after, please refer to sample exam.
- Yellow highlight from "This is important." appear in the text can be valuable reference.

Convergence Theory

The concept of convergence is crucial in probability theory. However, since this is beginning level of probability theory course, we do not expect complicated and rigorous proof in order to obtain full credits. In this regard, we make clear of the following concepts.

We have (i) convergence in probability, i.e. recall the Weak Law of Large Numbers, (ii) almost sure convergence, i.e. this is the Strong Law of Large Numbers (and usually it is outside of the scope of the exam), (iii) convergence in distribution (this is usually required topic for this course). In general, we can describe their relationship in the following diagram

$$\text{Convergence a.s.} \Rightarrow \text{Convergence in probability} \Rightarrow \text{Convergence in distribution}$$

The following let us discuss them one by one in the order of the appearances in the above diagram.

Convergence in Distribution. A sequence of random variables $X_1, ..., X_n$ convergences in distribution to a random variable X if the cdf converge, or

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$$

at all points of x where $F_X(x)$ is continuous. We denote this by $X_n \xrightarrow{\text{d}} X$

Example 10.3.1. If $X_1, ..., X_n$ are iid inform(0,1), we have

$$\mathbb{P}(|X_{(n)} - 1| \geq \epsilon) = \mathbb{P}(X_{(n)} \geq 1 + \epsilon) + \mathbb{P}(X_{(n)} \leq 1 - \epsilon) = \mathbb{P}X_{(n)} \leq 1 - \epsilon$$

which is

$$\mathbb{P}(X_{(n)} \leq 1 - \epsilon) = \mathbb{P}(X_i \leq 1 - \epsilon, i = 1, ..., n) = (1 - \epsilon)^n$$

which gives us

$$\mathbb{P}(n(1 - X_{(n)} \leq t) \to 1 - \exp^{-t}$$

Thus, we have shown that the random variable $n(1 - X_{(n)})$ converges in distribution to an exp(1) random variable.

Convergence in Probability. A sequence of random variables $X_1, ..., X_n$ converges in probability to a random variable X if, for every $\epsilon > 0$,

$$\lim_{n \to \infty} \mathbb{P}(|X_n - X| \geq \epsilon) = 0$$

or equivalently, $\lim_n \mathbb{P}(|X_n - X| < \epsilon) = 1$. We will denote this as $X_n \xrightarrow{\text{P}} X$.

The important theorem here is the Weak Law of Large Numbers. Let $X_1, ..., X_n$ be iid random variables with $\mathbb{E}(X_i) = \mu$ and $\text{var}(X_i) = \sigma^2 < \infty$. Define $\bar{X}_n : - = (1/n) \sum_{i=1}^n X_i$. Then, $X_n \xrightarrow{\text{P}} \mu$. The proof is directly from the Markov's Inequality.

Proof. We have for every $\epsilon > 0$,

$$\mathbb{P}(|\bar{X}_n - \mu| \geq \epsilon) = \mathbb{P}((\bar{X}_n - \mu)^2 \geq \epsilon^2) \leq \frac{\text{var}(\bar{X}_n)}{\epsilon^2} = \sigma^2/(n\epsilon^2) \to 0 \text{ as } n \to \infty$$

□

Consistency of Sample Variance. Can you show that the sample variance is consistent? We say that an estimator of a parameter is a consistent estimator if that this estimator converges to the true estimator in probability.

Example 10.3.2. Suppose we have a sequence $X_1, ..., X_n$ of iid random variables with $\mathbb{E}(X_i) = \mu$ and $\text{var}(X_i) = \sigma^2 < \infty$. Let us define

$$S_n^2 := \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$$

and the goal is to show WLLN. Using Chebyshev's Inequality, we have

$$\mathbb{P}(|X_n^2 - \sigma^2| \geq \epsilon) \leq \frac{\mathbb{E}(S_n^2 - \sigma^2)^2}{\epsilon^2} = \frac{\text{var}S_n^2}{\epsilon^2}$$

and thus, a sufficient condition that S_n^2 converges in probability to σ^2 is that $\text{var}(S_n^2) \to 0$ as $n \to \infty$.

An interesting theorem here is the following. Suppose $X_n \xrightarrow{\text{P}} X$ and that h is a continuous function. Then $h(X_n) \xrightarrow{\text{P}} h(S_n^2)$ is a consistent estimator of σ.

Another important theorem is the Slutsky's Theorem. IF $X_n \xrightarrow{\text{d}} X$ and $Y_n \xrightarrow{\text{P}} c$, a constant, then we have

- $Y_n X_n \xrightarrow{\text{d}} cX$
- $X_n + Y_n \xrightarrow{\text{d}} X + c$

Example 10.3.3. Suppose $X_1, ..., X_n$ are iid with $\mathbb{P}(X_i = \pm 1) = 1/2$. Define

$$Y_i := \prod_{j=1}^{i} X_j, \text{ for } i = 1, ..., n$$

1. Find the joint distribution of (Y_1, Y_2).
2. Derive the limiting distribution of $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} Y_i$

Answer. Let us address the following.

1. From the problem statement, we can see that Y_1 and Y_2 can only take values $\{-1, 1\}$ with equal probability. One can simply write all of them out. This will be the answer for (i). In other words, we have
$$\begin{aligned}
&\mathbb{P}(Y_1 = 1, Y_2 = 1) = 1/4 \\
&\mathbb{P}(Y_1 = 1, Y_2 = -1) = 1/4 \\
&\mathbb{P}(Y_1 = -1, Y_2 = 1) = 1/4 \\
&\mathbb{P}(Y_1 = -1, Y_2 = -1) = 1/4
\end{aligned}$$
2. Here notice that these Y_i's are mutually independent. Thus, from CLT, we can conclude $\bar{Y}_n \xrightarrow{\text{d}} \mathcal{N}(0, 1/n)$. This can be rewritten as $\sqrt{n}\bar{Y}_n \xrightarrow{\text{d}} \mathcal{N}(0, 1)$. The reason to rewrite the above is because we want the right-hand-side to have non-degenerating variance. When we had $1/n$, this degenerates to 0 as $n \to \infty$. If we factor n out, then we can avoid this problem, aka the asymptotic variance does not go to zero.

Note here the terminology "limiting distribution" is equivalent as "asymptotic distribution", "limiting behavior", or "asymptotic behavior". These key words all refer to the same framework of problems. When n goes to ∞, what happens? What does the distribution look like under large sample size? The answer shares the same philosophy of Law of Large Number (both weak and strong forms) and Central Limit Theorem. Hence, to avoid any ambiguity, the idea to solve this type of problems would always be some adaptation of the Central Limit Theorem. □

Example 10.3.4. Suppose $X_1, ..., X_n$ are iid having an exponential distribution with mean 1. Show that

$$\max_{1 \leq k \leq n} \frac{X_k}{\log n} \xrightarrow{\text{P}} 1 \text{ as } n \to \infty$$

where the arrow with P stacked on the top denotes convergence in probability.

Answer. To show convergence in probability, it suffices to show, for all $\epsilon > 0$,

$$\begin{aligned}
\mathbb{P}(|X_k/\log n - 1| > \epsilon) &= \mathbb{P}(X_{(n)}/\log n > 1 + \epsilon) + \mathbb{P}(X_{(n)}/\log n < 1 - \epsilon) \\
&= \mathbb{P}(X_{(n)} > (1+\epsilon)\log n) + \mathbb{P}(X_{(n)} < (1-\epsilon)\log n) \\
&= 1 - \mathbb{P}(X_{(n)} \leq (1+\epsilon)\log n) + (1 - \exp(-(1-\epsilon)\log n))^n \\
&= 1 - (1 - e^{-((1+\epsilon)\log n)})^n + (1 - e^{-(1-\epsilon)\log n})^n \\
&= 1 - (1 - n^{-(1+\epsilon)})^n + (1 - n^{-(1-\epsilon)})^n
\end{aligned}$$

Now consider

$$\begin{aligned}
\lim_{n\to\infty} n \log(1 - n^{-(1+\tilde{\epsilon})}) &= \lim_{n\to\infty} \frac{\log(1 - n^{-(1+\tilde{\epsilon})})}{1/n}, \text{ for any } \tilde{\epsilon} \in \mathbb{R} \\
&= \lim_{n\to\infty} \frac{(-1-\tilde{\epsilon})n^{-2-\tilde{\epsilon}}}{1 - n^{-(1+\tilde{\epsilon})}} \frac{1}{1/n^2} \\
&= \lim_{n\to\infty} \frac{-(1+\tilde{\epsilon})}{1 - n^{-(1+\tilde{\epsilon})}} \frac{1}{n^2} \frac{1}{n^{\tilde{\epsilon}}} n^2 \\
&= \lim_{n\to\infty} \frac{-(1+\tilde{\epsilon})}{n^{\tilde{\epsilon}} - 1/n}, \text{ by some algebra} \\
&= \lim_{n\to\infty} \frac{1+\tilde{\epsilon}}{n^{\tilde{\epsilon}}}
\end{aligned}$$

and next we discuss the following:

- When $1+\tilde{\epsilon} > 1$, the limit is zero. Since $\log(0) = 1$, we obtain $\lim_n (1-e^{-(1+\epsilon)})^n = 1$
- When $1+\tilde{\epsilon} < 1$, the limit goes to negative infinity. Since $\log(-\infty) = 0$, we obtain $\lim_n (1-e^{-(1-\epsilon)})^n = 0$

Thus, we have concluded that

$$\lim_{n\to\infty} \mathbb{P}(|\max_k X_k / \log n - 1| > \epsilon) = 0$$

for all $\epsilon > 0$ which suffices to show convergence in probability. □

Example 10.3.5. Suppose that $X_1, ..., X_{2n}$ are iid U[0,1]. Let $Y_i = X_{2i-1} + X_{2i}$ for $1 \leq i \leq n$.

(i). Find the limiting distribution of Y_1.

(ii). Find the limiting distribution of $\sqrt{n}(2 - Y_{(n)})$ as $n \to \infty$.

Answer. Let us address (i) and (ii) respectively in the following.

(i). Consider the scenarios: (i) $t \in [0,1]$ and (ii) $t \in [1,2]$. We can also visualize this using geometry (draw graphs). Here let us write it out with formulas. When $t \in [0,1]$, the probability is $\mathbb{P}(X_1+X_2 \leq t) = t^2/2$. When $t \in [1,2]$, the probability is $\mathbb{P}(X_1 + X_2 \leq t) = 1 - (2-t^2)/2 = -t^2/2 + 2t + 1$.

(ii). To find the limiting distribution of $\sqrt{n}(2 - Y_{(n)})$ as $n \to \infty$, consider

$$\begin{aligned} \mathbb{P}\sqrt{n}(2 - Y_{(n)} > \alpha) &= \mathbb{P}(Y_{(n)} < 2 - \alpha/\sqrt{n}) \\ &= \mathbb{P}(Y_1 < 2 - \alpha/\sqrt{n})^n \end{aligned}$$

Note that as n goes to infinity, we have $\alpha/n \to 0$, so $2 - \alpha/n$ falls in [1,2]. Thus, we have

$$\begin{aligned} \mathbb{P}(Y_1 < 2 - \alpha/2)^n &= (-(2-\alpha/\sqrt{n})^2 + 2(2-\alpha/2) - 1)^n \\ &= (1 - (\alpha^2/2)/n)^n \\ &= e^{-\alpha^2/2} \end{aligned}$$

which is the desired limiting distribution.

□

Index

References

[1] Ross, S. “A First Course in Probability”, 9th Edition.