STAT 410: Introduction to Probability Theory

Omkar Pathak*

January 21, 2023

These are my notes for UMD's STAT 410, which is an introduction to "the formulation and manipulation of probability models." These notes are taken live in class ("live-TEX"-ed). As a reference, I will also use the textbook based on Harvard's STAT 110, written by instructors Joe Blitzstein and Jessica Hwang, as well as some resources provided by MIT's 18.675 (Theory of Probability) professor Nike Sun. This course is taught by Prof. Mestiyage (Danul) Gunatilleka.

Contents

1 Tuesday, August 31, 2022 **5**
1.1 Probability spaces 5

2 Thursday, September 2, 2022 **8**
2.1 Consequences of the probability axioms 8

3 Tuesday, September 6, 2022 **11**
3.1 Random sampling 11

4 Thursday, September 8, 2022 **13**
4.1 Random variables 13
4.2 Conditional probability 15

5 Tuesday, September 13, 2022 **15**
5.1 Independence 15
5.2 Partitions 17
5.3 Law of Total Probability 17

6 Thursday, September 15, 2022 **17**
6.1 Applications of Bayes' theorem 17
6.2 Random variables (continued) 19

7 Tuesday, September 20, 2022 **19**
7.1 Probability distributions 19
7.2 Probability density functions 20

*Email: omkarp07@terpmail.umd.edu

7.3 Cumulative distribution function . . . 22
7.3.1 Cumulative distribution function of a discrete random variable 23
7.3.2 Cumulative distribution function of a continuous random variable 23
7.4 Finding the p.m.f. or p.d.f. from the c.d.f. . . . 23
7.5 General properties of cumulative distribution functions . . . 24
7.6 Named distributions . . . 24
7.6.1 Bernoulli distribution . . . 25
7.6.2 Binomial distribution . . . 25
7.6.3 Geometric distribution . . . 25
7.7 Further topics on sampling and independence . . . 26

8 Thursday, September 22, 2022 26
8.1 Expected value . . . 26
8.1.1 Infinite and noninfinite expectations . . . 27
8.2 Strong Law of Large Numbers . . . 27
8.2.1 Expectation of a function of a random variable . . . 29
8.2.2 Median and quantiles . . . 30

9 Tuesday, September 27, 2022 31
9.1 Variance . . . 31

10 Thursday, September 29, 2022 33
10.1 Named distributions (cont'd) . . . 33

11 Tuesday, October 4, 2022 34
11.1 Normal distribution . . . 34

12 Thursday, October 6, 2022 36
12.1 Finer points . . . 36
12.1.1 Density functions . . . 36
12.1.2 Integrals in probability . . . 36

13 Tuesday, October 11, 2022 37
13.1 Exam 1 Review . . . 37

14 Thursday, October 13, 2022 38

15 Tuesday, October 18, 2022 39
15.1 Exam #1 Review . . . 39
15.2 Poisson distribution . . . 39
15.3 Poisson processes . . . 42

16 Thursday, October 20, 2022 43
16.1 Moment generating functions . . . 44
16.2 Exponential distribution . . . 44
16.3 Gamma distribution . . . 45

17 Tuesday, October 25, 2022 **45**
17.1 Calculations of moments with the moment generating function 46
17.2 Joint distribution of discrete random variables 47

18 Thursday, October 27, 2022 **47**

19 Tuesday, November 1, 2022 **49**
19.1 Conditional distribution of a discrete random variable 49
19.2 Conditional distribution for jointly continuous random variables . . . 50
19.3 Independence 52
19.4 Expectation 53
19.4.1 Linearity of expectation 53
19.4.2 Variance 55

20 Thursday, November 3, 2022 **55**
20.1 Fubini and Tonelli's Theorems 55
20.1.1 Tonelli's Theorem 55
20.1.2 Fubini's Theorem 55
20.2 Effect of independence on expectation 57

21 Tuesday, November 8, 2022 **58**

22 Thursday, November 10, 2022 **58**
22.1 Further multivariate topics 58
22.1.1 Joint cumulative distribution function 58
22.2 Standard bivariate normal distribution 59
22.2.1 Infinitesimal method 60
22.2.2 Transformation of a joint density function 60
22.3 Sums of independent random variables 61
22.3.1 Sums and symmetry 61

23 Tuesday, November 15, 2022 **62**
23.1 Exam # 2 Review 62

24 Thursday, November 17, 2022 **65**

25 Tuesday, November 22, 2022 **65**

26 Thursday, November 29, 2022 **65**
26.1 Conditional expectation 65

27 Thursday, December 1, 2022 **69**
27.1 Markov's Inequality 69
27.2 Chebyshev's Inequality 70
27.3 Strong Law of Large Numbers (SLLN) 71

28 Tuesday, December 5, 2022 **72**
28.0.1 Machine Learning 73

28.1 Convergence of Sequences 73
28.2 Moving-Block Problem 74

29 Thursday, December 8, 2022 **74**
29.1 Weak Law of Large Numbers 74
29.2 Central Limit Theorem 75

30 Final Exam Practice Problem Solutions **76**

§1 Tuesday, August 31, 2022

§1.1 Probability spaces

Let's start by defining probability.

Definition 1.1. **Probability** is the mathematical study of uncertainty.

In probability, we are about to conduct an experiment and we would like to know how likely certain combinations of outcomes are.

Probability is usually used *before* an experiment or situation. In probability, we are *about* to conduct an experiment and have not conducted it yet. Consider the following scenario: you are about to roll a die and would like to know how likely it is (i.e. the probability) that you get an odd number.

If this were a fair die, the probability would be $\frac{1}{2}$. If it were a biased or unfair die, we cannot answer this without more information.

Let's consider another scenario: toss a coin until you see heads for the first time. What is the probability that heads appears on an even numbered toss? We'll encounter this and similar problems at a later point.

Definition 1.2. A **probability space** is an ordered triple $(\Omega, \mathcal{F}, P)$. Ω is the **sample space**, $\mathcal{F}$ is the **event space**, and P is the set of outcomes for which we would like to compute probabilities. In general, we have $\mathcal{F} \subseteq \mathbb{P}(\Omega)$, where $\mathbb{P}$ denotes the powerset. P is known as a **probability measure** and it's job is to assign (map) probabilities from the event space to the real numbers ($P : \mathcal{F} \to \mathbb{R}$). Note that on the same sample space and event spaces, we can have multiple probability measures.

Remark 1.3. Any function for P would not work. An intuitive explanation for this uses getting $P(\text{heads on the second toss}) < P(\text{heads on the second or third toss})$.

We place some restrictions on P (these are known as the **Axioms of Probability**):

- $P(A) \geq 0$ for any event A
- $P(\Omega) = 1$, $P(\emptyset) = 0$, where $\emptyset$ denotes the empty set.
- (Main computational tool for probability that we will use in class) For pairwise disjoint events (recall what this means) $A_1, A_2, \ldots, A_n$, $P(A_1 \cup A_2 \cup \cdots \cup A_i \cup \cdots \cup) = P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$.

There are certain set theory concepts explained above; recall what they are and what they mean.

Remark 1.4. While this may seem very intuitive, it is worth stating here: higher probability indicates that there is more chance of an event taking place, whereas a lower probability indicates that there is less chance of an event taking place.

Example 1.5

Recall the example from earlier: toss a coin till you see heads for the first time. What is the probability that heads appears on an even-numbered toss?

Solution. Here, $\Omega = \{(H), (T,H), (T,T,H), (T,T,T,H), \cdots\}$. Define $A_i = \{(T, \cdots, T, H)\}$, where there are $i-1$ Ts and an H at the ith position. Then, □

$$P(\bigcup_{i=1}^{\infty} A_{2i}) = \sum_{i=1}^{\infty} P(A_{2i}).$$

Remark 1.6. We can study "different coins" using different probability measures.

Let's suppose we have a fair coin. $P_1(A_i) = \frac{1}{2^i}$. If we have a coin with a bias, where $P(H) = \frac{1}{3}$. $P_2(A_i) = \frac{1}{3^i}$. If we go with P_1, $\sum_{i=1}^{\infty} P_i(A_{2i}) = \frac{1}{4} + \frac{1}{16} + \cdots$. This is left as an exercise.

The following is a question that is a bit beyond the scope of this class, but we will delve into it anyways: what is $\mathcal{F}$? Or rather, what properties should $\mathcal{F}$ have? We said that P had to have certain properties for it to make sense as a probability measure; what sort of properties does $\mathcal{F}$ have? So far, it seems intuitive that $\mathcal{F} = \mathbb{P}(\Omega)$.

Fact 1.7. $\mathcal{F} = \mathbb{P}(\Omega)$ does not always happen. Sometimes it can happen, whereas sometimes it doesn't. This is something we will study later on.

Consider the following experiment: we pick a number from the interval $[0,1]$ such that if i_1 and i_2 are subintervals of the same length of $[0,1]$, then $P(x \in i_1) = P(y \in i_2)$. If $\Omega = [0,1]$ and $F = \mathbb{P}(\Omega)$, then there is no probability measure on $\mathcal{F}$. Why this is the case isn't something we'll discuss. The important take is that even though we will be creating the event space as the powerset (or that any event that interests us will always have a probability assigned to it), this doesn't always have to be the case.

Remark 1.8. We often treat $\mathcal{F}$ as $\mathbb{P}(\Omega)$. The reason for this is that in practical situations, any $A \subseteq \Omega$ is such that $A \in F$. In practice, we usually give some $C \subseteq \mathcal{F}$ that "generates" $\mathcal{F}$. For example, if we have $I \subseteq [0,1]$ (e.g. I is a subinterval of $[0,1]$) and $P(I) = b - a$ for interval i with endpoints a, b, where $a < b$. Note that $\mathcal{F}$ is a **σ-algebra**.

We have some conditions about σ-algebras (these conditions also apply to $\mathcal{F}$, because $\mathcal{F}$ is a σ-algebra):

- $\emptyset \in \mathcal{F}$
- If $A \in \mathcal{F}$, then A^c (A-complement) is also in $\mathcal{F}$. Empty events will always be in $\mathcal{F}$. This is because $\Omega \in \mathcal{F}$, meaning that $\Omega^c \in \mathcal{F}$, further meaning that $\emptyset \in F$.
- If $A_1, \ldots A_n, \ldots \in \mathcal{F}$ (e.g.a sequence of events), then $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$.

The members of a σ-algebra are called **measurable sets**. The properties of a σ-algebra imply that countably many applications of the usual set operations is a safe way to produce new events.

To create the σ-algebra for the probability model we desire, we typically start with some nice sets that we want in $\mathcal{F}$, and let the axioms determine which other sets must also be members of $\mathcal{F}$. When Ω is $[0, 1]$, the real line, or some other interval of real numbers, the standard choice for $\mathcal{F}$ is to take the smallest σ-algebra that contains all subintervals of Ω. The members of this $\mathcal{F}$ are called **Borel sets**. Construction of probability spaces is beyond the scope of this course (as well as the textbook).

Another aspect of $\mathcal{F}$ as the event space is that $\mathcal{F}$ can represent *information*. The below example displays this idea.

Example 1.9

Suppose we are following two indistinguishable die. We can still model the sample space $\Omega = \{(i, j) : i, j \in \{1, 2, 3, 4, 5, 6\}\}$. By a judicious choice of $\mathcal{F}$, we can forbid the model from separating $(1, 3)$ and $(3, 1)$. This is how we define the event space $\mathcal{F}$:

- the singletons $\{(i, j)\}$ for integers $1 \leq i \leq 6$
- the sets of pairs $\{(i, j), (j, i)\}$ for $1 \leq i,\ j \leq 6$ such that $i \neq j$
- all unions of sets of the above two types, including the empty set

The probability measure P is the same as it would be with two distinguishable die, though the we have that events such as $(1, 3)$ and $(3, 1)$ are separate. This is because $\{(1, 3)\}$ is not a member of $\mathcal{F}$ and does not have a probability assigned to it, whereas the event $\{(1, 3), (3, 1)\}$ does have a probability assigned to it, namely $P(\{(1, 3), (3, 1)\} = \frac{2}{36}$. The point of the above example is that by restricting $\mathcal{F}$, we can model the information available to the observer of the experiment without changing Ω.

Example 1.10

Toss a coin twice. There are 4 possible outcomes. What are some σ-algebras of $\mathcal{F}$, the event space representing these possible outcomes?

The following are some σ-algebras: $\mathcal{F}_1 = \{\emptyset, \Omega\}, \mathcal{F}_2 = \{\Omega, \{(H, H), (H, T)\}, \{(T, H), (T, T)\}$.

Consequences of the probability axioms (i.e. properties of $\mathcal{F}$ and P):

- $\emptyset \in F$
- If $A_1, A_2, \ldots, A_n \in F$, then $\bigcup_{i=1}^{n} A_i \in \mathcal{F}$.

- If $A_1, \ldots, A_n, \ldots \in F$, $\bigcap_{i=1}^{\infty} A_i \in F$. This is essentially saying that if we have a collection of events, you can talk about all of these events that you're interested in.
- If $A_1, \ldots, A_n$ are pairwise disjoint, then $P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$. Recall that we have a similar idea with a finite sequence of A_i's.
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.
- If $A \subseteq B$, $P(A) \leq P(b)$.

Let's prove the second consequence:

Proof. For $i \geq n = 1$, take $A_i = \emptyset$ ($\emptyset \in \mathcal{F}$). Now, $A_1, A_2, \cdots, A_n, A_{n+1}, \ldots \in \mathcal{F}$. Therefore, $\bigcup_{i=1}^{n} A_i = \bigcup_{i=1}^{\infty} A_i \in F$ (note that the infinite union is a property of a σ-algebra, and therefore $\mathcal{F}$, because after n, $A_i = \emptyset$. This is a key concept of what we will cover in class, e.g. extending infinite series to finite ones. □

Now, let's prove the third consequence:

Proof. Note that we have $A_1^c, A_2^c, \ldots, A_n^c, \ldots \in F$, because complements of events are also in the event space. As a result, we have $\left(\bigcup_{i=1}^{\infty} A^c\right) \in \mathcal{F}$. Consequently, $\left(\bigcup_{i=1}^{\infty} A_i^c\right)^c = \bigcap_{i=1}^{\infty} A \in \mathcal{F}$. Note that we used DeMorgan's Laws here. This is a very key idea in set theory, and allows us to work with intersections and set theory. □

Let's also prove the fourth consequence:

Proof. This is similar to proving the second consequence. Take $A_i = \emptyset$ for $i > n$. Now, we have $\bigcup_{i=1}^{\infty} = \bigcup_{i=1}^{n} A_i$. □

§2 Thursday, September 2, 2022

§2.1 Consequences of the probability axioms

We started by reviewing certain properties of the probability space and the probability measure. These are all covered in Monday's notes.

The fourth property of $\mathcal{F}$ tells us that $P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$ if for every i, j with $i \neq j$, A_i and A_j are pairwise disjoint. Note that $P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$. How do we get from $\sum_{i=1}^{\infty} P(A_i)$ to $\sum_{i=1}^{n} P(A_i)$?

Before we do this, we have to prove that every A_i is pairwise disjoint, e.g. $A_i \cap A_j = \emptyset$ for any $i \neq j$. Let's do this in cases:

- Case 1: $i, j \leq n$

  Proof. We simply have $A_i \cap A_j = \emptyset$. This follows from conditions required for $P\left(\bigcup_{i=1}^{\infty} A_i\right) = P\left(\bigcup_{i=1}^{n} A_i\right)$. □

- Case 2: at least one of i or j is greater than n

Proof. Say $i > n$. Then, $A_i \cap A_j = \emptyset \cap A_j = \emptyset$. Say $j > n$. A similar result follows. We could have proved this with just one of i or j by using the term **without loss of generality**, e.g. **WLOG**. □

Let's prove the fifth consequence:

Proof. Notice that $A \cup B = A \cup (B - A) = A \cup (B \cap A^c)$. $P(A \cup B) = P(A \cup (B - A)) = P(A) + P(B - A)$. Also notice that $B = (A \cap B) \cup (B - A)$. Because these two sets are disjoint, we have that $P(B) = P(A \cap B) + P(B - A)$. From this, we can see that $P(A \cup B) = P(A) + P(B) - P(A \cup B)$. □

Let's prove the sixth consequence:

Proof. $B = A \cup (B - A)$ implies $P(B) = P(A) + P(B - A)$. Because $P(B - A) \geq 0$ by an axiom of the probability measure, we have $P(B) \geq P(A)$ proving $P(B) \geq P(A)$ if $A \subseteq B$. □

Remark 2.1. $P(A) = 1 - P(A^c)$.

Example 2.2

Suppose that Bob buys bananas at the grocery store with a probability of 0.7 and buys apples at the store with a probability of 0.4. Given that he buys both with a probability of of 0.3, compute the probability that:

1. he buys apples or bananas
2. he buys apples but not bananas

Solution. For the first part, we want to compute $P(A \cup B) = P(A) + P(B) - P(A \cap B)$, where A is the event where Bob buys apples and B is the event that Bob buys bananas. We have $P(A) = 0.4$, $P(B) = 0.7$, and $P(A \cup B) = 0.3$. Therefore, $P(A \cup B) = 0.4 + 0.7 - 0.3 = \boxed{0.8}$.

□

For the second part, we want to compute $P(A \cap B^c)$. This is left as an exercise.

Remark 2.3. The English words "and" and "but" refer to $\cap$, the word "or" refers to $\cup$, and the word "not" refers to the complement.

Remark 2.4. The **Principle of Inclusion-Exclusion** is discussed in the textbook; read this over.

Example 2.5

Pick a number from $[0, 1]$ so that for any interval i, $P(i)$ is the length of the interval i. $P\left(\left\{\frac{1}{2}\right\}\right) = 0$, because the length of the interval containing $\frac{1}{2}$ is 0.

The above example forces us to change our way of thinking about a probability of 0. Before, we considered events that had a probability of 0 as events that had no chance of happening. Now, we can say that an event with a probability of 0 only has this value because 0 is the only reasonable probability value that we can assign to the event.

Another example of this is below.

Example 2.6

Suppose we toss a fair coin until we get heads, at which point we stop. Let A_i be the event that we get heads at the ith toss. $P(A) = \frac{1}{2^i}$. $P(A_\infty) = \lim_{i\to\infty} P(A_i) = \lim_i \frac{1}{2^i} = 0$. Essentially, we have $0 \leq P(A_\infty) \leq P(A_i)$. Because we must have that $P(A_\infty)$ must be less than any positive number, we have $P(A_\infty) = 0$.

Remark 2.7. Examples like the one above usually only come up when we talk about infinite sample and event spaces; when these spaces are infinite, this usually does not occur. In the example above, we have that A_∞ almost surely does not take place.

Recall the meaning of **countable**: if A is countable, we have that $\exists f : \mathbb{N} \to A$, e.g. there is a function $\mathcal{F}$ in A such that we can map the natural numbers to it's co-domain. From the example with the number line, we have $P(x \in \{x_1, \ldots, x_n, \ldots\}) = P(x \in \{r_1\}) + \cdots + P(x \in \{r_n\}) = 0 + \ldots + 0 = 0$.

Fact 2.8. $\mathbb{Q}$ are countable.

Now, we'll talk about "games of chance," e.g. games that we will analyze with probability and solve problems with. In a "game of chance," each outcome is equally likely, e.g. for any $a, b \in \Omega$, $P(\{a\}) = P(\{b\}$. Below are two examples of games of chance and certain facts about their probability measures.

Example 2.9

Roll a fair die a fixed number of times. Then, $P(1,1,1) = P(6,5,3) = \cdots$.

Example 2.10

Draw 5 cards out of a well-shuffled deck.

$P(\text{one of diamonds}) = P(\text{three of diamonds}) = \cdots$.

Exercise 2.11. Justify why $\mathcal{F} = \mathbb{P}(\Omega)$ if $\Omega = [0, 1]$.

Proof. We know that every individual outcome must be in the powerset, because $\Omega = \{a_1, a_2, \cdots, a_n\}$ and Ω is finite. Note that $\{a_i\} \in F$ for each i. Let $A \subseteq \Omega$. Then, A can be written down as $\{a_{i1}, a_{i2}, \cdots, a_{ik}\}$. This is because we don't know which elements are in A, e.g. it could be the 5th, the 6th, etc. Now, we have that $A = \bigcup_{j=1}^{k} \{a_{ik}\}$. Note that $A \in \mathcal{F}$, because the unions of events are also events. As a result, because A is any subset of $\mathcal{F}$, $\mathbb{P}(\Omega)$. □

With the above result, we can take any subset A of Ω and find $P(A)$. $P(\{a_1, \ldots, a_n\}) = P(\{a_1\}) + \cdots + P(\{a_n\}) = \frac{1}{|\Omega|} + \frac{1}{|\Omega|} + \cdots + \frac{1}{|\Omega|}$

§3 Tuesday, September 6, 2022

§3.1 Random sampling

Here's a hint for one of the problems about random string generation: let q be the probability that the string we want **isn't** generated. Then, take $\lim_{n\to\infty} 1 - q^n$ to get 1, which means that the string we want will eventually be generated. This is question 51 in the text. Question 50 is the easier one of these two. You just have to adjust 51 to match 50, and you get your answer.

Let's now prove that $P\left(\bigcup_{i=1}^n A_i\right) \leq \sum_{i=1}^n P(A_i)$.

Proof. Let $B_1 = A_1$, $B_2 = A_2 - A_1$, $B_3 = A_3 - \bigcup_{i=1}^n A_i, \ldots B_n = A_n - \bigcup_{i=1}^{n-1} A_i$. Note that every B_i is pairwise disjoint with every other B_i, because we are removing all of the elements that appeared in any B_k before B_i. So, we have $B_i \cap B_j = \emptyset$ for $i \neq j$. As a result, we have that $P\left(\bigcup_{i=1}^n B_i\right) = \sum_{i=1}^n P(B_i) \leq P\left(\bigcup_{i=1}^n A_i\right)$. In other words, because $B_i \subseteq A_i$, we have that $P(B_i) \leq P(A_i)$ for all $1 \leq i \leq n$, by construction. □

There are a couple of takeaways from this proof. The first is that if we want to compute the probability of the union, we want to have pairwise disjoint sets. The second takeaway is that if we have sets that are not pairwise disjoint, we can construct a sequence of pairwise disjoint events for us to use (like we did with B_i in this example).

Recall when we discussed that $\mathcal{F}$ is not usually equal to $\mathbb{P}(\Omega)$, e.g. there is usually a C that **generates** $\mathcal{F}$. Let $C = \{\{a_1\}, \{a_2\}, \cdots, \{a_n\}\}$. If Ω is not countable, $C = \{\text{all singletons}\}$ **does not** generate $\mathbb{P}(\Omega)$. The key idea here is that we...

Recall from Thursday that we want to count the number of elements in the event of interest. Let $A \subseteq \Omega$ with $A \in F$. Then, we have $P(A) = \frac{|A|}{|\Omega|}$. How do we count the elements in $|A|$ or $|\Omega|$? We can count in the following ways:

- **Case 1:** with replacement, order matters
- **Case 2:** without replacement, order matters
- **Case 3:** with replacement, order does not matter
- **Case 4:** without replacement, order does not matter

Let's look at examples for all four cases:

Example 3.1 (Case 1)
You roll a fair die twice. What is the probability that the sum of numbers is even?

Here, our sample space/all possible outcomes can be represented with ordered pairs, e.g. $\Omega = \{(1,1),(1,2),\cdots,(6,6)\}$. We use the following counting principle here: if a task can be performed as a sequence of smaller tasks, $1,2,\cdots,n$, where the ith task can be performed in m_i ways, the total number of ways in which the task can be performed is $\Pi_{i=1}^{n} m_i$. Let $A = A_1 \cup A_2$ be the event that the sum is even. We want to express A as a collection of pairwise disjoint events for which we can compute the probability. Suppose A_1 represents the event where both numbers are odd, and A_2 represents the event where both numbers are even, e.g. $A_1 \cup A_2 = \emptyset$. Here, we have $P(A) = P(A_1 \cup A_2) = P(A_1) + P(A_2) = \frac{|A_1|}{\Omega} + \frac{|A_2|}{\Omega} = \frac{9}{36} + \frac{9}{36} = \boxed{\frac{1}{2}}$.

Let's say we toss this coin 10 times. Finding Ω is easy, e.g. it is simply 6^{10}. The concept here is same, but it takes more time. We're not going to focus on solving hard counting problems in this class, but we'll look at a few.

Before we look at Case 2, let's examine the following scenario: we have a collection of n objects in a box. When we pick the first element, we have n choices between what elements to pick from. When we pick the second, we have $n-1$ elements to choose from. When we pick the kth choice, we have $n-(k-1) = n-k+1$ choices. Because the order matters, we want to keep track of this with tuples. The number of possible tuples of objects that we can choose is $\frac{n \times n-1 \times \cdots \times (n-k+1) \times (n-k) \times \cdots \times 1}{(n-k) \times (n-k-1) \times \cdots \times 1} = \frac{n!}{(n-k)!}$. This is known as $P_{(n,k)}$ or nPk.

Example 3.2 (Case 2)

You create an alpha-numeric password that is **not** case-sensitive. This password features 8 non-repeated characters. You **uniformly at random** pick a password from an urn that contains all possible passwords. What is the probability that you picked a password that contains both letters and numbers?

Whenever you see this type of problem, you should try to compute Ω first. In this case, this is $P_{(36,8)}$. Now, we want to compute our event space. However, this seems like it will be quite tedious. Instead, we can compute the probability that a password contains only letters or only numbers. Let A be the event where we get a password with **both** numbers and letters. Then, we know that $P(A) = 1 - P(A^c)$, where A^c is the collection of passwords with only letters or only numbers, e.g. A^c = passwords with all numbers $\cup$ passwords with all letters. Both of the events in A^c are pairwise disjoint, so we can add the probabilities of each individual case.

Now, we want to compute $1 - P(N \cup L) = 1 - (P(N) + P(L)) = 1 - \left(\frac{10P8}{36P8} + \frac{26P8}{36P8}\right)$. This doesn't have to be simplified, so we can leave this answer as is.

We won't look at Case 3 in this class.

Before we get to Case 4, let's discuss what we can do in cases without replacement and where order doesn't matter. Essentially, we can modify the case of the problem where order does matter. If we have n objects and we want to pick k of them without

replacement, we can do this in $\frac{nPk}{k!} = \frac{n!}{(n-k!)(k!)}$. Essentially, we divide by $k!$ to ensure that we don't count multiple permutations of the same object again.

Example 3.3 (Case 4)

A poker hand consists of 5 cards. If the cards have distinct consecutive values, but not all of them are the same suit, we say that the hand is straight. What is the probability that you are dealt a straight hand from a well shuffled deck of cards? Note that we will use aces as a high or low card but **not** both.

Here, we don't care about the order, and the cards are dealt without replacement. Like in the above examples, the easiest thing to compute is the total number of possible outcomes, e.g. $|\Omega|$. In this case, it is $\binom{52}{5}$. Our strategy here will be to **count by type**. Instead of counting all of the different types of straight hands, we will count only straight hands of a certain type and then multiply by the different types of straight hands. Let's count the different types of straight hands with the numbers $2, 3, 4, 5, 6$. Because there are 4 suits, we have 4 options for each card. We have 4^5 total possibilities in this scenario, and must subtract 4 for the cases in which all of the cards are of the same suit. We have 10 different straights, so the total number of outcomes in our event space is $10 \times (4^5 - 4)$, so our probability is $\frac{10(4^5-4)}{\binom{52}{5}}$.

The quiz on Thursday will have 2 questions, and will cover the material that we covered up until and including today.

§4 Thursday, September 8, 2022

§4.1 Random variables

Today, we'll cover random variables (1.5) and conditional probability (Chapter 2).

Definition 4.1. A **random variable** is a (measurable) function $X : \Omega \to \mathbb{R}$. Measurable is in parentheses; we need this term for technical reasons.

The idea here is that we often care about values associated to what, who, etc. we pick, rather than what was actually picked. For example, in Monopoly, we roll a pair of dice, and the interesting outcome is the *sum* of the values of the dice. This idea of attaching a number to each outcome is captured by the notion of a random variable.

Random variables are usually denoted with capital letters such as X, Y, and Z. The value of a random variable X at sample point ω is $X(\omega)$. Notation-wise, we have $P(X \in A) = P(\{\omega : X(\omega) \in A\})$, where $\omega \in \Omega$.

Remark 4.2. Probability spaces sometime have extra information that is not relevant; random variables give us a way to focus on the salient characteristics via **distributions**. We will cover distributions at a later point.

Example 4.3

Suppose we roll a pair of die. We can introduce three random variables in this scenario: X_1 is the outcome of the first die, X_2 is the outcome of the second die, and S is the sum of the two die. For each sample point, $(i, j) \in \Omega$. $X_1(i, j) = i$, $X_2(i, j) = j$, $S(i, j) = X_1(i, j) + X_2(i, j) = i + j$. Our probability measure can be written as $P(\{S = 8\}) = \sum_{(i,j);i+j=8} P(\{X_1 = i, X_2 = j\}) = \frac{5}{36}$.

Example 4.4

Suppose we have $X(T) = 0$ and $X(H) = 1$. Then, we have $P(x \in \{0\}) = \frac{1}{2}$, because this is the same as asking $P(\omega \in \Omega : X(\omega) \in \{0\})$. This is because $P(\omega \in \Omega)$ is the same as asking for $P(T)$. On the other hand, we have $P(X \in [2, 5]) = 0$, because neither $X(T)$ or $X(H)$ will be in $[2, 5]$. Similarly, we have $P(X \in \{1\}) = \frac{1}{2}$, $P(X \in [0, 1]) = 1$, and $P(X \in (0, 1)) = 0$.

Example 4.5

Suppose $\Omega = [0, 1]$, and $P(I)$ for any interval is equal to $b - a$, where a and b are the endpoints of this interval. What is the probability that you get numbers from $\left[0, \frac{1}{2}\right]$ or $\left(\frac{1}{2}, 1\right]$? We have that $P(I_1) = \frac{1}{2}$ and $P(I_2) = \frac{1}{2}$. What is $\mathcal{F}$ here? It is **not** equal to $\mathbb{P}(\Omega)$. This is because there are not finitely many outcomes that are all equally likely (the set here is **uncountably infinite**, preventing us from creating the union of countably many sets).

Remark 4.6. If A is a singleton (e.g. $A = \{a\}$), we have that $P(X = a) = P(X \in \{a\})$. If A is an interval, then you express this with inequalities.

Example 4.7

Suppose you roll a fair die twice. What is the probability that the sum of the numbers is 8?

Note that $\Omega = \{(1, 1), (1, 2), \ldots, (6, 6)\}$. If we have $X : \Omega \to \mathbb{R}$, we can say $X(a, b) = a + b$. The purpose of this random variable is then to compute $P(X = 8)$. This is left as an exercise.

Definition 4.8. Let $X_1, X_2, \cdots, X_n$ be random variables defined on the same probability space. Then, $X_1, X_2, \cdots, X_n$ are independent if

$$P(X_1 \in B_1, X_2 \in B_2, \ldots, X_n \in B_n) = \prod_{k=1}^{n} P(X_k \in B_k)$$

for all choices of subsets $B_1, B_2, \ldots, B_n$ on the real number line.

The above definition can feel a bit impractical because it involves thinking of all possible subsets B_k. Fortunately, it simplifies in cases of interest. For example, if the random variables are discrete, then it is enough to check the condition for particular values.

Theorem 4.9

Discrete random variables $X_1, X_2, \ldots, X_n$ are independent if and only if

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = \prod_{k=1}^{n} P(X_k = x_k)$$

for all choices $x_1, x_2, \ldots, x_n$ of possible values of the random variables.

Proof. Try to see if you remember the proof from the textbook and mimic it here. □

§4.2 Conditional probability

We'll start with an example.

Example 4.10

Suppose you have a fair coin. You toss the coin and based on the outcome H/T, you roll either a fair six-sided die or a fair four-sided die.

How do we analyze this? We can use **conditional probability**.

Definition 4.11. $P(A|B) = \frac{P(A \cap B)}{P(B)}$ (where $P(A)$ and $P(B)$ are both greater than 0, for the sake of manipulation purposes). Analogously, $P(B|A) = \frac{P(B \cap A)}{P(A)}$.

Remark 4.12. Manipulating the above formula gives us $P(A \cap B) = P(A|B) \cdot P(B)$.

Theorem 4.13 (Bayes' Theorem, weaker version)

$P(B|A) = \frac{P(A|B) \cdot P(B)}{P(A)}$

Proof. This proof is left as an exercise (it was proven in class). □

Remark 4.14. We can use **tree diagrams** to compute probabilities visually.

§5 Tuesday, September 13, 2022

§5.1 Independence

Definition 5.1. Events A and B are **independent** if and only if $P(A \cap B) = P(A)P(B)$. Furthermore, events $A_1, A_2, \ldots, A_n$ are independent if for every collection $A_{i_1}, A_{i_2}, \ldots, A_{i_n}$, where $2 \leq k \leq n$ and $1 \leq i_1 \leq i_2 \leq \cdots \leq i_k \leq n$ and

$$P(A_{i_1} A_{i_2} \cdots A_{i_k}) = P(A_{i_1})P(A_{i_2}) \cdots P(A_{i_k}).$$

Example 5.2

Suppose we have independent events A, B, and C. Then, we have $P(A \cap B) = P(A)P(B)$, $P(B \cap C) = P(B)P(C)$, $P(A \cap C) = P(A)P(C)$, and $P(A \cap B \cap C) = P(A)P(B)P(C)$.

Theorem 5.3

Suppose events $A_1, A_2, \ldots, A_n$ are mutually independent. Then, for every collection $A_{i_1}, A_{i_2}, \ldots, A_{i_k}$, where $2 \leq k \leq n$ and $1 \leq i_1 \leq i_2 \leq \cdots \leq i_k \leq n$, we have

$$P(A_{i_1}^* A_{i_2}^* \cdots A_{i_k}^*) = P(A_{i_1}^*)P(A_{i_2}^*) \cdots P(A_{i_k}^*)$$

where each A_i^* represents either A_i or A_i^c.

Remark 5.4. The intuition behind the definition of independence is as follows: if A and B are independent, we have $P(A|B) = P(A)$, e.g. B has no effect on whether or not A takes place. Using the definition of conditional probability, we have $\frac{P(A \cap B)}{P(B)} = P(A)$, which implies $P(A \cap B) = P(A)P(B)$. However, we want to be able to apply the notion of independence to events that have a probability of 0. Yet, Bayes' theorem is only defined for events with positive probabilities. In other words, we have that independence depends on a broader class of events, not just the scope of events that work in Bayes' theorem.

Theorem 5.5 (Multiplication rule for n independent events)

If $A_1, A_2, \cdots, A_n$ are (mutually) independent events and all the conditional probabilities below make sense, then

$$P(A_1 A_2 \cdots A_n) = P(A_1)P(A_2|A_1)P(A_3|A_2 A_1) \cdots P(A_n|A_{n-1} \cdots A_2 A_1)$$

Example 5.6

Suppose that a fair coin comes up heads with probability $\frac{1}{2}$. Let B_n be the event such that you see heads on the nth toss. What is $P(B_n)$?

Solution. We can answer this using Bayes' theorem. However, we can also use independence to prove that a coin flip is not affected by the result of the previous flip, meaning that $P(B_n)$ is simply $\frac{1}{2}$. □

Remark 5.7. In real-world scenarios, independence is an assumption that needs to be justified carefully. For the "math-y" part (e.g. when we write down theorems and solve problems with these theorems), we usually assume independence. Computationally, when we assume independence, we have $P\left(\bigcap_{j=1}^k A_{ik}\right) = \prod_{j=1}^k P(A_{ij})$. Read the "Finer Points" section of Chapter 2 for more information about this.

§5.2 Partitions

A family of events $\{A_\lambda\}_{\lambda\in\Lambda}$ is said to be a **partition** of Ω if

i) The A_λ are pairwise disjoint

ii) $\bigcup_{\lambda\in\Lambda} A_\lambda = \Omega$

Remark 5.8. If we have a partition, we already have a collection of pairwise disjoint sets, meaning that we can manipulate these sets and use the additivity axiom of probability. Also, as a note, in the textbook, partitions are limited to finite or countably infinite partitions.

Example 5.9

Let B be an event. If we take $B = \bigcup_{\lambda\in\Lambda} B\cap A_i$ such that A_i is a partition of B. Then, we have that the $B\cap A_i$ are also pairwise disjoint. Let's say I is countable. Then, $P(B) = P\left(\bigcup_{i\in I} B\cap A_i\right) = \sum_{i\in I} P(B\cap A_i)$.

§5.3 Law of Total Probability

Theorem 5.10 (Law of Total Probability)

Suppose the events $A_1, A_2, \cdots, A_n$ are partitions of Ω, and that $P(A_i) > 0$. Then, for any event B, $P(B) = \sum_{i=1}^n P(B|A_i)\cdot P(A_i)$.

Proof. We proceed with Example 5.9 and Bayes' theorem (the "weaker" version). $P(B) = \sum_{i=1}^n P(B\cap A_i) = \sum_{i=1}^n \frac{P(B\cap A_i)}{P(A_i)}\cdot P(A_i) = \sum_{i=1}^n P(B|A_i)\cdot P(A_i)$. □

Remark 5.11. I called the use of Bayes' theorem the "weaker" version above because a version of Bayes' theorem was introduced earlier in these notes.

Theorem 5.12 (Bayes' theorem, general form)

Let $A_1, A_2, \cdots, A_n$ be a partition of Ω with $P(A_i) > 0$ and let B be an event with $P(B) > 0$. Then,

$$P(A_i|B) = \frac{P(B|A_i)\cdot P(A_i)}{\sum_{j=1}^n P(B|A_j)\cdot P(A_j)}$$

Proof. $P(A_i|B) = \frac{P(A_i\cap B)}{P(B)} = \frac{P(B|A_i)P(A_i)}{P(B)}$ (the denominator is reached using the Law of Total probability). Is this complete? If not, the proof is left as an exercise. □

§6 Thursday, September 15, 2022

§6.1 Applications of Bayes' theorem

Example 6.1

A doctor is called to see a sick child. The doctor has prior information that 90% of sick children in the neighborhood have the flu, while the other 10% are sick with measles. There are only two types of sicknesses in this world. A well-known symptom of measles is a rash. Assume that the probability of having a rash if one has measles is 0.95. However, children with flu will also occasionally develop a rash, and the probability of having a rash if one has the flu is 0.08. If the doctor finds a rash on the child, what is the probability that this child has measles?

Solution. To apply Bayes' theorem, we first need a partition of the sample space. Here, our sample space Ω consists of measles (M) and the flu (F). $M \cup F = \Omega$ and $M \cap F = \emptyset$, meaning that we have a partition. Now, by Bayes' theorem, have that

$$P(M|R) = \frac{P(R|M) \cdot P(M)}{P(R|M) \cdot P(M) + P(R|F) \cdot P(F)}.$$

Now, we can simply plug things in, and see that our probability is roughly 0.57. □

Remark 6.2. Many medical tests, such as those for COVID-19, advertise that they will be 95% accurate. However, this only means $P(+|H) = 0.95$; we also have to analyze $P(H|+)$ to see whether or not the test is actually reliable. Note that H represents the event that one has COVID-19.

Example 6.3

Suppose we have 3 cards, all 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 is colored black. The three cards are mixed in a hat, and 1 card is picked at random and put down on the ground. If the upper side of the chosen card is colored red, what is the probability that the other side of the card is black?

Solution. The intuitive argument here is that there are two cards with red sides. Only one of them has a side that is black. Hence, if we know that one side is red, it follows that the probability that the other side is black is $\frac{1}{2}$. Unfortunately, this is incorrect.

Here is an intuitive explanation for why the answer is $\frac{1}{3}$. Our Ω in this scenario is $\{(R_1, R_2), (R_2, R_1), (R, B), (B, R), (B_1, B_2), (B_2, B_2)\}$, where the cards are R_1, R_2 (both sides red), R, B (one side red, one side black), and B_1, B_2 (both sides black).

Let's now solve this using Bayes' theorem. Let RR, BB and RB denote respectively, the red card is chosen, the black card is chosen, and the card with both red and black sides is chosen. We want to find

$$P(RB|R) = \frac{P(RB \cap R)P(R)}{P(R)}$$
$$= \frac{P(RB \cap R)P(R)}{P(R|RR)P(RR) + P(R|RB)P(RB) + P(R|BB)P(BB)} = \frac{\frac{1}{2} \cdot \frac{1}{3}}{1 \cdot \frac{1}{3} + \frac{1}{2}\frac{1}{3} + 0} = \frac{1}{3}.$$

□

Refer to Ross' textbook (eighth edition) examples 3n/3l for another Bayes' theorem example.

§6.2 Random variables (continued)

Recall that a random variable is a function $X : \Omega \to \mathbb{R}$, where X is a Borel function. In other words, for "nice" subsets $A \subseteq \mathbb{R}$, $P(X \in A)$ can be calculated. Note that two random variables have the same distribution if and only if $P(X \in A) = P(Y \in A)$ for every Borel subset $A \subseteq R$.

Example 6.4

Toss a biased coin that comes up heads probability $\frac{1}{6}$. Let X be the number of tosses to heads. Then, we have $X((T, H)) = 2$, $X(H) = 1$, $X(T, T, H) = 3$, etc.

Example 6.5

Roll a fair die till you see 1, then stop. Let Y be the number of die rolls till you see the first 1. Then, we have $Y(1) = 1$, $Y((3, 5, 1)) = 3$, $Y((5, 3, 1)) = 3$, etc.

From the above two examples, we have that X and Y are clearly not the same function, in particular because their domains are completely different (recall the properties of function equality). But, X and Y have the same distribution, in particular because the coin is biased.

Theorem 6.6

If $P(X \in (-\infty, a]) = P(Y \in (-\infty, a])$ for all $a \in \mathbb{R}$, then X and Y have the same distribution.

From this, we have that $P(X \subseteq C)$ is important, as it fully determines $P(X \in A)$ for all nice A. $F_X(C) = P(X \subseteq C)$ is called the **cumulative distribution function** (c.d.f.) of X. Recall that C is the set that will **generate** the event space $\mathcal{F}$. We will discuss the c.d.f. more in-depth at a later point.

Exercise 6.7. Try to use to Theorem 7.6 to show that $F_X(x) = F_Y(x)$, for X and Y in Examples 7.4 and 7.5.

§7 Tuesday, September 20, 2022

§7.1 Probability distributions

Definition 7.1. Let X be a random variable. The **probability distribution** of the random variable X is the collection of probabilities $P(X \in B)$ for sets B of real numbers.

Definition 7.2. A random variable X is a **discrete random variable** if there exists a finite or countably infinite set $\{k_1, k_2, \ldots, k_n\}$ of real numbers such that $\sum_i P(X = k_i) = 1$, where the sum ranges over the entire set of points $\{k_1, k_2, \ldots, k_i\}$.

Definition 7.3. The **probability mass function** (p.m.f.) of a discrete random variable X is the function p (or p_X) defined by $p(k) = P(X = k)$ for possible values k of X.

Definition 7.4. A random variable X is said to be **continuous** if there is some function $f_X(k) \geq 0$ such that $P(a < X < b) = \int_a^b f_X(k)\ dx$.

The key here is that for important classes of experiments, we can write down the probability mass function with a reasonable amount of work.

§7.2 Probability density functions

Recall from last class the cumulative distribution function (c.d.f.). Also recall the probability mass function (p.m.f.) $p(k) = P(X = k)$, which is a function from the set of possible values of X into $[0, 1]$. We can label the probability mass function with the random variable X as $p_X(k)$. Note that the values of the probability mass function must sum to 1 on account of the axioms of probability:

$$\sum_k p_X(k) = \sum_k P(X = k) = 1$$

where the sum is over all possible values k of X. Finally, recall that we compute the probability $P(X = B)$ for a subset $B \subset \mathbb{R}$ by

$$P(X \in B) = \sum_{k \in B} p_X(k)$$

where the sum runs over all possible values of k that lie in B. Note that the probability mass function can be represented as a bar chart. If $p_X(k) > 0$, the bar chart has a column centered at k with height $p_X(k)$.

Definition 7.5. Let X be a random variable. If a function f satisfies

$$P(X \leq b) = \int_{-\infty}^b f_k\ dx$$

for all real values b, then f is the **probability density function** (p.d.f.) of X.

In plain English, we have that this definition requires that the probability that the value of X lies in the interval $[-\infty, b]$ equals the area under the graph of f from $-\infty$ to b. An important (and somewhat surprising) technical fact is that if f satisfies the above definition, then

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

for *any* subset B of the real line for which integration makes sense. We prefer Definition 8.1 over the above definition because it ties in with the cumulative distribution function, and it is easier to check than a more general condition, such as the one above.

Example 7.6

The set B in the above definition can be an interval, bounded or unbounded, or any collection of intervals. Examples include:

$$P(a \leq X \leq b) = \int_a^b f(x)\ dx \quad \text{and} \quad P(X > a) = \int_a^\infty f(x)\ dx$$

for any real $a \leq b$.

Fact 7.7. If a random variable X has density function f, then point values have probability 0:

$$P(X = c) = \int_c^c f(x)\ dx = 0$$

for any real c.

It follows that a random variable with a density function is not discrete. It also follows that probabilities of intervals are not changed by including or excluding endpoints. For example, in Example 8.6, we have $P(a < X \leq b) = P(a < X < b) = \int_a^b f(x)\ dx$.

Which functions quality as probability density functions? Since probabilities are always nonnegative and $P(-\infty < X < \infty) = 1$, a density function f must satisfy:

$$f(x) \geq 0 \text{ for all } x \in \mathbb{R} \quad \text{and} \quad \int_\infty^\infty f(x)\ dx = 1.$$

Any function that satisfies the above condition will be called a probability density function of f.

Remark 7.8. No random variable X can have two different density functions. Section 3.6 in the textbook contains further comments on this point. Furthermore, random variables that have a density function are called **continuous** random variables. While this terminology, is not entirely accurate (as the variables do not necessarily have to be continuous on the sample space Ω), we will use it regardless because of its widely accepted use.

Remark 7.9. If X is continuous, we have

$$P(X \leq x) = \int_{-\infty}^x f_x(t)\ dx$$

If X is discrete, we have

$$P(X \leq x) = \sum_{t \leq x} p_x(t).$$

We will now cover the uniform distribution.

Definition 7.10. Let $[a, b]$ be a bounded interval on the real line. A random variable X has the **uniform distribution** on the interval $[a, b]$ if X has the density function

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{if } x \in [a, b] \\ 0 & \text{if } x \notin [a, b] \end{cases} \tag{1}$$

Abbreviate this by $X \sim \text{Unif}[a, b]$.

If $X \sim \text{Unif}[a, b]$ and $[c, d] \subset [a, b]$, then

$$P(c \leq X \leq d) = \int_c^d \frac{1}{b-a}\, dx = \frac{d-c}{b-a}$$

which is the ratio of the lengths of the intervals $[c, d]$ and $[a, b]$. Recall from Fact 8.7 that individual points make no difference to any probability calculation with a density function. Hence, in the above definition, we can drop one or both endpoints a and b if we so prefer, to define a uniform random variable on the half-open interval $(a, b]$ (or $[a, b)$) or on the open interval (a, b). It makes no difference to the probability calculation, because either way, $P(X = a \text{ or } X = b) = 0$.

It is important to note that the value at any given point of a density function f is not defined. A density function f satisfies probabilities of sets by integration, as in the examples provided by the textbook. When this function is multiplied by the length of a tiny interval, it approximates the probability of the interbal.

Fact 7.11. Suppose that a random variable X has density function f that is continuous at the point a. Then, for any small $\epsilon > 0$,

$$P(a < X < a + \epsilon) \approx f(a) \cdot \epsilon.$$

The fact above comes from the limit

$$\lim_{\epsilon \to 0} \frac{P(a<X<a+\epsilon)}{\epsilon} = \lim_{\epsilon \to 0} \frac{1}{\epsilon} \int_a^{a+\epsilon} f(x)\, dx = f(a).$$

The last limit is true because in a small interval around a point a, the continuous function f is close to the constant $f(a)$. Similarly to above, we have

$$P(a - \epsilon < X < a + \epsilon) = f(a) \cdot 2\epsilon \text{ and } P(a - \epsilon < X < a) = f(a) \cdot \epsilon$$

§7.3 Cumulative distribution function

Note that probability mass functions are defined only for discrete random variables and that density functions are defined only for continuous random variables. By contrast, the cumulative distribution function (c.d.f.) gives a way to describe the probability distribution of *any* random variable, including those that do not fall into discrete or continuous categories.

Definition 7.12. The **cumulative distribution function** (c.d.f) of a random variable X is defined by

$$F(s) = P(X \leq s) \text{ for all } s \in \mathbb{R}$$

It is important to note that the inequality is a $\leq$ in the equation; this means that the c.d.f. gives probabilities of left-open right-closed intervals in the form $(a, b]$:

$$P(a < X \leq b) = P(X \leq b) - P(X \leq a) = F(b) - F(a)$$

Knowing these probabilities is enough to determine the distribution of X completely; however, proving this fact is beyond the scope of this course.

§7.3.1 Cumulative distribution function of a discrete random variable

The connection between the p.m.f. and the c.d.f. of a discrete random variable is:

$$F(s) = P(X \leq s) = \sum_{k;k\leq s} P(X = k)$$

where the sum extends over those possible values k of X that are less than or equal to s. This is an application of the original definition of the p.m.f. with $B = (-\infty, s]$. Note the following features that distinguish the c.d.f. F of a discrete random variable X:

(i) The graph of F jumps exactly at the possible values of X and is constant between jumps; a function of this type is called a **step function** or **piecewise defined function**.

(ii) The probability of a value X equals the size of the jump of F.

§7.3.2 Cumulative distribution function of a continuous random variable

The connection between the c.d.f. of a continuous random variable F and its density function is

$$F(s) = P(X \leq s) = \int_{-\infty}^{s} f(x)\, dx$$

The second equality above comes from the definition of the p.d.f..

§7.4 Finding the p.m.f. or p.d.f. from the c.d.f.

Certain examples in the textbook illustrated how to utilize the p.m.f. or the p.d.f. of a random variable to find its c.d.f.. We will now analyze how to find the p.m.f. or the p.d.f. from the c.d.f.. First, however, we will address how to tell from the c.d.f. whether the random variable is discrete or continuous.

Fact 7.13. Let the random variable X have the c.d.f. F. Then,

(a) Suppose F is piecewise constant. Then, X is a discrete random variable. The possible values of X are the locations where F has jumps, and if x is such a point, then $P(X = x)$ equals the magnitude of the jump of F at x.

(b) Suppose F is continuous and the derivative $F'(x)$ exists everywhere on the real line, except possible at finitely many points. Then, X is a continuous random variable and $f(x) = F'(x)$ is the density function of X. If F is not differentiable at a certain point x, then the value $f(x)$ can be set arbitrarily.

We will not prove this fact; part (a) can be checked with the ideas of the first example in the textbook (finding the c.d.f. from a p.m.f.), and part (b) is discussed in Section 3.6 of the textbook. Note that this statement does not cover all possible cases, but is general enough for most practical needs.

§7.5 General properties of cumulative distribution functions

The examples in the textbook suggest that all c.d.f's share certain properties. These properties can also be used to characterize all possible c.d.f.'s:

Fact 7.14. Every cumulative distribution function F has the following properties:

(i) Monotonicity: if $s < t$, then $F(s) \leq F(t)$

(ii) Right continuity: for each $t \in \mathbb{R}$, $F(t) = \lim_{s \to t^+}$, where $s \to t^+$ means that s approaches t from the right

(iii) $\lim_{t \to -\infty} F(t) = 0$ and $\lim_{t \to \infty} F(t) = 1$

Part (i) is a consequence of the monotonicity of probability, according to which a larger event has a higher probability: for $s < t$, we have

$$F(s) = P(X \leq s) \leq P(X \leq t) = F(t)$$

Properties (ii) and (iii) in the c.d.f.'s can proved using properties from Chapter 1 of the textbook.

Remark 7.15. By definition, the c.d.f. F of a random variable X gives the probabilities $F(a) = P(X \leq a)$. A probability of the type $P(X < a)$ can also be expressed in terms of F, but this requires a limit.

Fact 7.16. Let X be a random variable with c.d.f F. Then, for any $a \in \mathbb{R}$,

$$P(X < a) = \lim_{s \to a^-} F(s)$$

We can use Fact 1.39 (from the textbook; recall what this is) to prove this identity. Note that when taking the limit of F from a, we can denote this as $F(a-)$; when taking the limit of F from the right, we use $F(a)$. A useful consequence of this fact is the following formula for point probabilities:

$$P(X = a) = P(X \leq a) - P(X < a) = F(a) - F(a-)$$

Note that the above expression is equal to the size of the jump in F at point a. This expression holds for *all* random variables.

§7.6 Named distributions

We will now discussed **named distributions**, which show up so often in probability that they have names. Named distributions can be either discrete or continuous.

§7.6.1 Bernoulli distribution

The Bernoulli random variable records the result of a single trial with two possible outcomes.

Definition 7.17. Let $0 \leq p \leq 1$. A random variable X has the **Bernoulli distribution** with success probability p if X is $\{0,1\}$-valued and satisfies $P(X = 1) = p$ and $P(X = 0) = 1 - p$. Abbreviate this by $X \sim \text{Ber}(p)$.

In general, you repeat a Bernoulli trial independently or previous trials until you succeed, and then you stop. Note that success usually occurs with an outcome of 1, and failure occurs with an outcome of 0.

§7.6.2 Binomial distribution

Suppose you repeat a Bernoulli trial a fixed number of times (n) in an independent way, and you count the number of successes. Let X be the number of successes. We are interested in finding $P(X = x)$. Start off by focusing on one specific outcome. Suppose that the first outcome was a success, the second and third were failures, the fourth was a success, etc.

In total, let there be k successes and $n-k$ failures. The probability of this happening is $p \cdot (1-p) \cdot (1-p) \cdot p \cdots$. If we compute this, there will be k p's, and $n-k$ $(1-p)$'s. Also note that we can have different combinations of successes and failures with the same amount, e.g. we could have had all successes at the beginning, and all failures at the end, etc.

Therefore, our probability is $\binom{n}{k} p^k (1-p)^{n-k}$.

Definition 7.18. Let n be a positive integer and $0 \leq p \leq 1$. A random variable X has the **binomial distribution** with parameters n and p if the possible values of X are $\{0, 1, \ldots, n\}$ and the probabilities are

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

for $k = 0, 1, \ldots, n$. Abbreviate this by $X \sim \text{Bin}(n, p)$.

The fact that binomial probabilities add to 1 is a particular case of the Binomial Theorem. We have that

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

§7.6.3 Geometric distribution

With a little leap of faith, we can consider *infinite sequences* of independent trials. This conceptual step is in fact necessary if we are to make sense of statements such as "when a fair die is rolled repeatedly, the long term frequency of fives is $\frac{1}{6}$. Imagine an infinite sequence of independent trials with success probability p. We won't worry about constructing the sample space; this won't be needed for us to do calculations. As before, let X_j be the outcome of the jth trial, with $X_j = 1$ if trial j is a success and $X_j = 0$ is the trial j is a failure. As in the binomial distribution, if $S_n = X_1 + X_2 + \cdots + X_n$ denotes the number of successes in the first n trials, then

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

We can extend this to an infinite sequence of independent trials to properly define the geometric distribution. Let N be the number of trials needed to see the first success in a number of independent trials with success probability p. Then, for any positive integer k,

$$P(N = k) = P(X_1 = 0, X_2 = 0, \ldots, X_{k-1} = 0, X_k = 1) = (1-p)^{k-1} p$$

This random variable has the geometric distribution that we encountered before.

Definition 7.19. Let $0 < p \leq 1$. A random variable X has the **geometric distribution** with success parameter p if the possible values of X are $\{1, 2, 3, \ldots\}$ and X satisfies $P(X = k) = p(1-p)^{k-1}$ for positive integers $k = 1, 2, 3, \ldots$ Abbreviate this using $X \sim \text{Geom}(p)$.

§7.7 Further topics on sampling and independence

We will now discuss **conditional independence**, as well as the **hypergeometric** distribution.

§8 Thursday, September 22, 2022

We've spend the past few classes looking at ways to describe the probability distribution of a random variable. Next, we turn to key quantities that represent useful partial information about a random variable, such as its expectation (also called the mean), variance, and moments.

§8.1 Expected value

Definition 8.1. The **expectation** or **mean** of a discrete random variable X is defined by

$$\mathbb{E}(X) = \sum_k k P(X = k)$$

where the sum ranges over all possible values k of X.

The expectation is also called the **first moment**, and another conventional symbol for it is $\mu = \mathbb{E}(X)$.

Definition 8.2. The expectation of a continuous random variable X is given by

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

where $f_X(x)$ represents the density function of X (as always).

Note the difference between how the expected value for a discrete random variable differs from that of a continuous random variable.

§8.1.1 Infinite and noninfinite expectations

Before proceeding further, we must take care of one technical point. Once we go beyond discrete random variables with finitely many values, infinite expectations can arise, and it can even happen that no value for the expectation can be given at all. Examples in the textbook illustrate this phenomenon (e.g. the famous St. Petersburg paradox). An expectation $\mathbb{E}(X)$ is **well-defined** if it has a defined value, which is either a finite number or $\pm\infty$.

§8.2 Strong Law of Large Numbers

We will now briefly introduce the Strong Law of Large Numbers. We will go over this in more detail further through the semester.

Theorem 8.3 (Strong Law of Large Numbers)
For almost every large sample $x_1, x_2, \ldots, x_n$,

$$\frac{\sum_{i=1}^{n} x_i}{n} \approx \mathbb{E}(X)$$

Now, let's go over some examples of calculating expected value.

Example 8.4
Roll a fair die. Let X be the number that comes up when you roll the die. Find $\mathbb{E}(X)$.

Solution. We want to find $\mathbb{E}(X) = \sum_{x=1}^{g} x p_X(x) = \frac{1}{6}\sum_{i=1}^{6} x = \frac{1}{6} \cdot \frac{6 \cdot 7}{2} = \frac{7}{2} = 3.5$. Note that this is because we have

$$p_X = \begin{cases} \frac{1}{6} & x = 1, 2, \ldots, 6 \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

□

Example 8.5
Flip a fair coin, where heads is 1 and tails is 0. What is the expected value of the coin flip.

Solution. The answer is $\frac{1}{2}$. This is because $\mathbb{E}(X) = \frac{1}{2} \cdot 1 + \frac{1}{2} \cdot 0 = \frac{1}{2}$. Note that we have $p_X(x) =$

$$\begin{cases} \frac{1}{2} & x = 0 \text{ or } x = 1 \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

□

Example 8.6

X is a uniform distribution on $[a, b]$, e.g. $X \sim [a, b]$. What is the expected value of X?

Solution.

$$p_X(x) = \begin{cases} 0 & \text{for any individual point in } [a, b] \\ \int_{-\infty}^{\infty} x f_X(x)\ dx & \text{in general} \end{cases} \tag{4}$$

Hence, we have $\mathbb{E}(X) = \int_{-\infty}^{\infty} x f_X(x)\ dx$, where $f_X(x)$ is the probability density function of X. This is equal to $\int_{-\infty}^{a} - \cdot x\ dx + \int_a^b x \frac{1}{b-1}\ dx + \int_b^{\infty} 0 \cdot dx = \frac{1}{b-a}\frac{x^2}{2}|_a^b = \frac{b^2-a^2}{2(b-a)} = \frac{a+b}{2}$.

□

From here on out, computing of expected values is going to involve quite a bit of integration, so review this.

Example 8.7

Suppose we have

$$\begin{cases} f_X(x) = 5x^4 & x \in [0, 1] \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

Find $\mathbb{E}(X)$.

Solution. We have $\mathbb{E}(X) = x \cdot f_X(x) = x \cdot \int_0^1 5x^4 = \frac{5}{6}$. □

Example 8.8

Suppose we have $X \sim U[0, 1]$. Compute $P(X < \frac{1}{4} | X \leq \frac{1}{2})$.

This is left as an exercise.

Remark 8.9. Every random variable will have an expected value.

Theorem 8.10

We say that the expected value of X exists if

$$\int_{-\infty}^{\infty} |x| f_X(x)\ dx < \infty$$

if X is continuous and

$$\sum_{n=1}^{\infty} x(n) p_x(n) < \infty$$

if X is discrete.

§8.2.1 Expectation of a function of a random variable

Taking a function of an existing random variable creates a new random variable, e.g. if X is a random variable on Ω, and g is a real-valued function defined on the range of X, then the composition $g(X)$ is a new random variable. Its value at a sample point $\omega \in \Omega$ is $g(X(\omega))$. Note that the notation $g(X)$ is used more commonly than $g \circ X$.

The **Law of the Unconscious Statistician (LOTUS)** gives us a way to compute the expected value of a function $g(X)$ of a random variable when one knows the probability distribution of X, but does not know the distribution of $g(X)$. Note: some books give this name, while others don't.

Theorem 8.11 (Law of the Unconscious Statistician, LOTUS)
If the random variable X has a discrete distribution, the expected value of $g(X)$ is equal to

$$E[g(X)] = \sum_x g(x) f_X(x)$$

where the sum is over all possible values x of X. If the random variable X has a continuous distribution and one knows its probability density function f_X (but not $f_{g(X)}$), the expected value of $g(X)$ is equal to

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

If one knows the cumulative probability distribution function F_X (but not $F_{g(X)}$), the expected value of X is given by the **Riemann-Stieltjes** integral:

$$E[g(X)] = \int_{-\infty}^{\infty} g(x) \mathrm{d}F_X(x)$$

Proof. Some notation may be off here, as this proof is (mostly) taken from the textbook. The formula for a discrete random variable is (mostly) proven in Example 3.24. The key is that the event $g(X) = y$ is the disjoint union of the events $\{X = k\}$ over those values of k that satisfy $g(k) = y$. This means that we have

$$\begin{aligned}
\mathbb{E}[g(X)] &= y \sum_y P(g(X) = y) \\
&= \sum_y \sum_{k:g(k)=y} P(X = k) \\
&= \sum_y \sum_{k:g(k)=y} g(k) P(X = k) \\
&= \sum_k g(k) P(X = k)
\end{aligned}$$

In the second equality above, we split the event $\{g(X) = y\}$ into the disjoint union of events $\{X = k\}$ over k with $g(k) = y$. In the third equality, we move y inside the inner sum and then notice that in the inner sum $y = g(k)$. In the final step, we merge all the terms to get a sum over all possible values k of X.

The proof of the formula for a continuous random variable goes beyond the scope of this course and the textbook. □

The special case $g(x) = x^n$ is common enough to have its own name:

Definition 8.12. The nth **moment** of the random variable X is the expectation $\mathbb{E}(X^n)$. In the discrete case, the nth moment is calculated by

$$\mathbb{E}(X^n) = \sum_k k^n P(X = k)$$

If X has density function f, its nth moment is given by

$$\mathbb{E}(X^n) = \int_{-\infty}^{\infty} x^n f(x)\ dx$$

The second moment, $\mathbb{E}(X^2)$, is also called the **mean square**.

§8.2.2 Median and quantiles

When a random variable has rare, abnormally bad variables, its expectation may be a bad indicator of where the distribution lies. The median provides an alternative measure:

Definition 8.13. The **median** of a random variable X is any real value m that satisfies

$$P(X \geq m) \geq \tfrac{1}{2} \text{ and } P(X \leq m) \geq \tfrac{1}{2}$$

Note that the median does not have to be unique. However, when the cumulative distribution function is continuous and strictly increasing on the range of the random variable, the median is unique.

Definition 8.14. For $0 < p < 1$, the pth **quantile** of a random variable X is any real value x satisfying

$$P(X \geq x) \geq 1 - p \text{ and } P(X \leq x) \geq p.$$

The median is the 0.5th quantile. The quantiles corresponding the $p = 0.25$ and $p = 0.75$ are called the first and third **quartiles**.

§9 Tuesday, September 27, 2022

§9.1 Variance

The variance measures how much a random variable fluctuates around its mean.

Definition 9.1. Let X be a random variable with mean μ. The **variance** of X is defined by

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

An alternative symbol is $\sigma^2 - \text{Var}(X)$.

Definition 9.2. The square root of the variance is called the **standard deviation**. We have $\text{SD}(X) = \sigma = \sqrt{\text{Var}(X)}$.

While it may seem more intuitive to use $\mathbb{E}[X - \mu]$ to calculate the variance, it will become more evident later that the variance as defined with the square has many useful properties that would be lost if we attempted to work with $\mathbb{E}[X - \mu]$. As with the mean, there are two ways to calculate the variance for a random variable:

Theorem 9.3

Let X be a random variable with expected value (mean) μ. Then,

$$\text{Var}(X) = \textstyle\sum_k (k - \mu)^2 P(X = k)$$

if X is discrete; and

$$Var(X) = \int_{-\infty}^{\infty} (x - \mu)^2 f(x)\, dx$$

if X is continuous and has density function f.

According to the above formula, the variance of X is the expectation of the random variable $(X - \mu)^2$. Note that we can let $g(x) = (x - \mu)^2$ to to use LOTUS.

Observe that expanding the squares in Theorem 9.3 gives us another useful formula to calculate variance:

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

This formula is valid for *all* random variables:

Theorem 9.4

An alternative formula to calculate variance is as follows:

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

The mean and variance of a linear function of a random variable appear often enough to merit a separate function. A linear function g is of the form $g(x) = ax + b$. Such functions are also called **affine**.

Theorem 9.5

Let X be a random variable and a and b real numbers. Then,

$$E(aX + b) = aE(X) + b$$

and

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

provided the mean and variance are well-defined.

Proof. The first part of the proof is a consequence of the properties of sums, integrals, probabilities, and Theorem 9.3 Below is a derivation for X with density function f:

$$E(aX + b) = \int_{-\infty}^{\infty}(ax + b)fx\ dx = a\int_{-\infty}^{\infty} xf(x)\ dx + b\int_{-\infty}^{\infty} f(x)\ dx = aE(X) + b$$

To prove the second formula, we use the following:

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

□

The formula for variance can be interpreted as follows: since the variation measures fluctuation around the mean, shifting by b has no influence because the mean is shifted too, and since the variance is the expectation of squared deviations, the constant a must come out as a square.

The properties of linearity presented in Theorem 9.5 can be extended to sums of higher moments:

Theorem 9.6 (Linearity of Expectation)

For any choice of constants $a_0, a_1, \ldots, a_n$, we have

$$\mathbb{E}\left[\sum_{k=0}^{n} a_k X^k\right] = \sum_{k=0}^{n} a_k \mathbb{E}(X^k)$$

assuming each expected value (moment/mean) is finite.

In Chapter 8, we will see that linearity of expectation is even more general than the above theorem.

We end our coverage of expected value with the following theorem:

Theorem 9.7

For a random variable X, $\text{Var}(X) = 0$ if and only if $P(X = a) = 1$ for some real value a.

Proof. We will start with the backwards direction. If $P(X = a) = 1$, then X is discrete, $\mathbb{E}(X) = a$, and

$$\text{Var}(X) = \mathbb{E}[(X - a)^2] = (a - a)^2 P(X = a) = 0 \cdot 1 = 0$$

To prove the forward direction, we have to continue with the assumption that X is discrete. Suppose $\mu = \mathbb{E}(X)$ and

$$0 = \text{Var}(X) = \textstyle\sum_k (k - \mu)^2 P(X = k)$$

This sum is of nonnegative terms, meaning that it vanishes only if each individual term vanishes (e.g. becomes 0). For each term, note that we have

$$(k - \mu)^2 P(X = k) = 0 \text{ if and only if } k = \mu \text{ or } P(X = k) = 0$$

Because we need the above statement to be true for every random variable X, we need that $k = \mu$ for all k; thus, because we need $P(X = k) > 0$, and the only k-value with $P(X = k) > 0$ is $k = \mu$, we have $P(X = \mu) = 1$. □

§10 Thursday, September 29, 2022

To recap: so far, we've been discussing expected value, as well as how to interpret what it is.

§10.1 Named distributions (cont'd)

Definition 10.1. A random variable Z has the **standard normal distribution** (also called the **standard Gaussian distribution** if Z has density function

$$\varphi(x) = \tfrac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}$$

Abbreviate this by $Z \sim \mathcal{N}(0, 1)$.

The standard normal distribution is so important that it has its own notation. Instead of the generic f and F for the density and cumulative distribution function, we write φ for the standard normal density and

$$\Phi(x) = \tfrac{1}{\sqrt{2\pi}} \textstyle\int_{-\infty}^{x} e^{-\frac{s^2}{2}}\, ds$$

for the normal cumulative distribution function. where $x \in \mathbb{R}$. Note that there is no explicit antiderivative for φ, making it a bit difficult to work with the normal distribution (φ does integrate to 1, however, by property of the density function).

Theorem 10.2

To help us in computing the cumulative distribution function of $\Phi(x)$, we have

$$\int_{-\infty}^{\infty} e^{-\frac{s^2}{2}}\, ds = \sqrt{2\pi}$$

Proof. This is proved by computing the square of the integral as a double integral and switching to polar coordinates; it is in the textbook. □

To find explicit numerical values for $\Phi(x)$, we have to use a table, which gives the values of $\Phi(x)$ for $0 \le x \le 3.49$, accurate to four decimal digits. For larger x, $\Phi(x)$ will be closer to 0.0002 to 1. For negative values, we use symmetry, e.g. the fact that $\varphi(x) = \varphi(-x)$:

$$\Phi(-x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-x} 6 - xe^{-\frac{s^2}{2}}\, ds = \frac{1}{\sqrt{2\pi}} \int_{x}^{\infty} e^{-\frac{s^2}{2}}\, ds = 1 - \Phi(x)$$

Example 10.3

Let $Z \sim \mathcal{N}(0, 1)$. Find the numerical value of $P(-1 \le Z \le 1.5)$.

Solution. By properties of the density function and c.d.f., we have

$$\begin{aligned} P(-1 \le Z \le 1.5) &= \int_{-1}^{1.5} \varphi(s)\, ds = \int_{-\infty}^{1.5} \varphi(s) - \int_{-\infty}^{1} \varphi(s)\, ds \\ &= \Phi(1.5) - \Phi(-1) = \Phi(1.5) - (1 - \Phi(1)) \\ &\approx 0.9332 - (1 - 0.8413) = 0.7745 \end{aligned}$$

□

The second-to-last step above used the table of values for Φ.

§11 Tuesday, October 4, 2022

§11.1 Normal distribution

Today, we'll continue talking about the normal distribution.

Example 11.1

Find $z > 0$ so that a standard normal random variable Z has approximately $\frac{2}{3}$ probability of being in the interval $(-z, z)$.

Solution. We want $z > 0$ such that $P(-z < Z < z) = \frac{2}{3}$. Note that

$$P(-z < Z < z) = \Phi(z) - \Phi(-z) = \Phi(z) - (1 - \Phi(z)) = 2\Phi(z) - 1$$

Thus, we want z for which $\Phi(z) = \frac{1}{2}(1 + \frac{2}{3}) = \frac{5}{6} \approx 0.833$. This is because $2\Phi(z) = 1 + P(-z < Z < z)$, implying $\Phi(z) = \frac{1+P(-z<Z<z)}{2}$. We can use the table of values for Φ to see that $\Phi(0.96) = 0.8315$ and $\Phi(0.97) = 0.8340$. While we could use linear approximation to get a better approximation, this is not important. □

Theorem 11.2

Let $Z \sim \mathcal{N}(0, 1)$. Then, $\mathbb{E}(Z) = 0$ and $\text{Var}(Z) = \mathbb{E}(Z^2) = 1$.

Proof. The proof for this is in the textbook. It relies on using the fact that if f is an odd function (e.g. $f(-x) = -f(x)$), $\int_{-a}^{a} f(x)\, dx = 0$. □

The above theorem gives meaning to the parameters 0 and 1 in $Z \sim \mathcal{N}(0, 1)$: a standard normal variable has mean 0 and variance 1. The general family of normal distributions is obtained by linear (or affine) transformations of Z. Let μ be real, $\sigma > 0$, and $X = \sigma Z + \mu$. By Theorem 9.5, we have $\mathbb{E}(X) = \mu$ and $\text{Var}(X) = \sigma^2$. We can express the cumulative distribution function of X in terms of Φ:

$$F(X) = P(X \leq x) = P(\sigma Z + \mu \leq x) = P(Z \leq \tfrac{x-\mu}{\sigma}) = \Phi(\tfrac{x-\mu}{\sigma}).$$

The p.d.f. of X can now be obtained by differentiating F:

$$f(x) = F'(x) = \tfrac{d}{dx}[\Phi(\tfrac{x-\mu}{\sigma})] = \tfrac{1}{\sigma}\varphi(\tfrac{x-\mu}{\sigma}) = \tfrac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Above, we used the chain rule, as well as the fact that $\Phi'(x) = \varphi(x)$. This is the definition of the general normal random variable. We call the parameter μ the mean and σ^2 the variance, since we checked above that these are the mean and variance of X.

Definition 11.3. Let μ be real and $\sigma > 0$. A random variable X has the **normal distribution** with mean μ and variance σ^2 if X has density function

$$f(x) = \tfrac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

on the real line. Abbreviate this by $X \sim \mathcal{N}(\mu, \sigma^2)$.

The argument above can be used to prove the following theorem:

Theorem 11.4

Let μ be real, $\sigma > 0$, and suppose $X \sim \mathcal{N}(\mu, \sigma^2)$. Let $a \neq 0$, b be real, and $Y = aX + b$. Then, $Y \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$. That is, Y is normally distributed with mean $a\mu + b$ and variance $a^2\sigma^2$. In particular, $Z = \frac{X-\mu}{\sigma}$ is a standard normal random variable.

Example 11.5
Suppose $X \sim \mathcal{N}(-3, 4)$. Find the probability $P(X \leq -1.7)$.

Solution. Since $Z = \frac{X-(-3)}{2} \sim \mathcal{N}(0, 1)$, we have

$$P(X \leq -1.7) = P\left(\frac{X-(-3)}{2} \leq \frac{-1.7-(-3)}{2}\right) = P(Z \leq 0.65)$$

which is equivalent to $\Phi(0.65) \approx 0.7422$. □

Example 11.6
Suppose $X \sim \mathcal{N}(\mu, \sigma^2)$. What is the probability that that observed value of X deviates from μ by more than 2σ?

Solution. With $Z = \frac{X-\mu}{\sigma}$ we have

$$\begin{aligned} P(|X - \mu| > 2\sigma) &= P(X < \mu - 2\sigma) + P(X > \mu + 2\sigma) \\ &= P\left(\frac{X-\mu}{\sigma} < -2\right) + P\left(\frac{X-\mu}{\sigma} > 2\right) \\ &= P(Z < -2) + P(Z > 2) = 2(1 - P(Z \leq 2)) \\ &= 2(1 - \Phi(2)) \approx 2(1 - 0.9772) = 0.0456 \end{aligned}$$

□

The above solution gives us a rule of thumb: a normal random variable is within two standard deviations of its mean with probability over 95%.

§12 Thursday, October 6, 2022

§12.1 Finer points

Today, we will discuss certain finer points that were not covered throughout the chapter.

§12.1.1 Density functions

With density functions, we continue to sweep measure-theoretic details under the rug. Which functions can be density functions? Basically any function that can be legitimately integrated so that it can represent probabilities and that satisfies $f(x) \geq 0$ and $\int_{-\infty}^{\infty} f(x)\, dx = 1$.

§12.1.2 Integrals in probability

Today, we also discussed the use of integration in probability. Why does integration show up in probability? This is due to the concept of **relative frequency**. Look this up a bit more to understand what it means.

§13 Tuesday, October 11, 2022

§13.1 Exam 1 Review

Today, we will review the practice problems (1-7) provided on ELMS. Thursday's exam will have 5 questions and will be 75 minutes long (e.g. the entire duration of the class). Poisson processes (e.g. the topic of #8 on the study guide) will not be on the exam, as they have not been covered yet.

1. (a) *Proof.* Let $A_i = \emptyset$. Note that $\{A_i\}_{i\in\mathbb{N}}$ are pairwise disjoint, e.g. $A_i \cup A_j = \emptyset$ for all i, j. Furthermore, note that $P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i) = \sum_{i=1}^{\infty} P(\emptyset)$. Note that $0 \leq P(\emptyset) < \infty$. Now, we must have $P(\emptyset) < \infty$, because the probability measure is a function, and a function cannot assign a value to ∞. Assume $P(\emptyset) > 0$. Then $\sum_{i=1}^{\infty} P(\emptyset)$ would be infinite, which, as we discussed before, is not possible. Therefore, we must have $P(\emptyset) = 0$ in order to satisfy the probability axioms. Note that we didn't use the fact that $P(\omega) = 1$ at all in the above proof.

   □

   (b) This part was proved in class; it is left as an exercise for the reader.

   (c) This is false; suppose we have $A_1 = \omega$, $A_2 = \omega$, and $A_3 = \emptyset$. Then, $P(A_1 \cup A_2 \cup A_3) = 1 + 1 = 0$, if the formula holds. However, as probability cannot be greater than 1, this is false. Another example is $\omega = \{1, 2, 3\}$ and $P(\{1\}) = P(\{2\}) = P(\{3\}) = \frac{1}{3}$. Then, suppose we have $A_1 = \{1, 2\}$, $A_2 = \{2, 3\}$, and $A_3 = \{1, 3\}$. Then, $P(A_1 \cup A_2 \cup A_3) = 1$, but $P(A_1) + P(A_2) + P(A_3)3 \cdot \frac{2}{3} = 2$, which is not possible.

   (d) For part a), note that $P\left((\cup_{i=1}^{n} A_i^c)^c\right) = P\left(\cap_{i=1}^{n} A_i\right)$. Thus, by the definition of the probability of the complement, we are done. Part b) is done similarly.

2. Note that there are a total of $\binom{52}{5}$ hands we can be dealt. WLOG suppose we are choosing diamonds and clubs. Then, the amount of combinations are have are $\binom{13}{1}\binom{13}{4} + \binom{13}{2}\binom{13}{3} + \binom{13}{3}\binom{13}{2} + \binom{13}{4}\binom{13}{1}$. Also note that there are a total of $\binom{4}{2}$ ways to pick two hands. Let L be the amount of combinations that we calculated with diamonds and clubs. Then, our answer is $\boxed{\frac{\binom{4}{2} \cdot L}{\binom{52}{5}}}$.

   On the exam, questions such as this won't be as prevalent, as they are more combinatorics-oriented, as opposed to probability-oriented.

3. (a) Note that we need to compute $\sum_{n=1}^{\infty} \frac{k}{n(n+1)}$. Also note that this sum is equivalent to 1, by property of the probability mass function. Then, the nth term of the series can be written as $\frac{1}{n} - \frac{1}{n+1}$. We have $a_1 = 1 - \frac{1}{2}$, $a_2 = \frac{1}{2} - \frac{1}{3}, \ldots, a_n = \frac{1}{n} - \frac{1}{n+1}$. If you look at the nth partial sum, this is just $1 - \frac{1}{n+1}$, because the series telescopes. Also note that what we call the infinite sum is just the limit of the partial sums, e.g. $\sum_{n=1}^{\infty} \frac{1}{n(n+1)} = \lim_n S_n = 1$, meaning $k = 1$. On the exam, questions such as this will only rely on geometric and/or telescoping sums; there will be no advanced techniques.

(b) To compute the cumulative distribution function, if $x < 1$, $F_X(x) = 0$. Otherwise, for $x \geq 1$, we can analyze the jump between each point to see that $F_X(x) = \frac{1}{\lfloor x \rfloor - 1}$.

(c) $\mathbb{E}(X) = \sum_{n=1}^{\infty} n\frac{1}{n+1}$. The n terms will cancel out, and we will be left with $\sum_{n=1}^{\infty} \frac{1}{n+1}$. Because this sum diverges, we have $\mathbb{E}(X)$ does not exist.

4. (a) Note that our probability density function is $\int_0^1 \frac{k}{1+x^2} = k\int_0^1 \frac{1}{1+x^2} = k\arctan(x)|_0^1 = 1$. This implies that $k(\arctan(1) - \arctan(0)) = 1$, meaning that $k = \frac{4}{\pi}$. It is worth remembering the common trigonometric integrals, as they may appear on the exam.

(b) To compute the cumulative distribution function, if $x < 0$, the cdf is 0. If $x \geq 1$, the cdf is 1. Thus, we only have to compute $\int_0^x \frac{4}{\pi}\frac{1}{1+t^2}\,dt$. This integral is equivalent to $\frac{4}{\pi}\arctan(t)|_0^x = \frac{4}{\pi}(\arctan(x) - 0) = \boxed{\frac{4}{\pi}\arctan(x)}$.

(c) $\mathbb{E}(X) = \int_0^1 \frac{4}{\pi}\frac{x}{1+x^2}\,dx = \frac{2}{\pi}\int_0^1 \frac{2x}{1+x^2}\,dx = \frac{2}{\pi}\ln 1 + x^2|_0^1 = \frac{2\ln 2}{\pi}$. To compute $V(X)$, first note that $V(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2$. We have $\mathbb{E}(X^2) = \frac{4}{\pi}\int_0^1 \frac{x^2}{1+x^2}\,dx = \frac{4}{\pi}\int_0^1 \frac{x^2+1-1}{1+x^2} = \frac{4}{\pi}\int_0^1 1 - \frac{1}{1+x^2}\,dx$. Usually, questions this tricky won't be on the exam, but there will be an exception.

5. (a) This question has been done in class multiple times before; we can start by finding the cdf of Z, which is $P(Z \leq t) = P(e^X \leq t) = P(X \leq \ln t)$. For $t < 1$, $\ln t$ either does not exist or is negative; therefore, for $t < 1$, $P(Z \leq t) = 0$. For $1 < t \leq e$, $0 \leq \ln(t) = F_X(\ln t) \leq 1$. Thus, $P(X \leq \ln t) = \ln t$, using the cdf from the uniform distribution. For $t > e$, $P(X \leq \ln t) = 1$. Thus, $f_Z(z) = 0$ when $z < 1$, $\frac{1}{z}$ when $1 \leq z \leq e$, and 0 when $z > e$.

(b) $\mathbb{E}(Z)$ does exist, and can be found using the standard way to find expected value of a continuous random variable.

6. (a) This problem is a direct application of Bayes' theorem. Let J be the event where Alice orders juice. $P(J|A) = 0.56$, $P(J|B) = 0.27$, and $P(J|C) = 0.91$. We can simply use Bayes' theorem to find $P(C|J)$.

(b) Here, we want to find $P(A \cup B|J)$. We can use the fact that $P(A \cup B|J) = P(A|J) + P(B|J)$. We can once again use Bayes' theorem here.

7. (a) Computation.

(b) $P(9 < X < 12) = P\left(\frac{9-10}{4} < X < \frac{12-10}{4}\right) = P(-\frac{1}{4} < X < \frac{1}{2}) = P(X < \frac{1}{2}) - P(X < -\frac{1}{4})$. We can use the chart to find values of Φ for $X \leq \frac{1}{2}$ and $X \leq -\frac{1}{4}$. Note that $P(X \leq \frac{1}{2}) = P(X < \frac{1}{2})$, as $P(X = \frac{1}{2}) = 0$ (e.g. the probability of any individual point is 0).

§14 Thursday, October 13, 2022

Exam #1 is today.

§15 Tuesday, October 18, 2022

(Read finer points for Chapter 3, Chapter 5, and start Chapter 6)

§15.1 Exam #1 Review

A common mistake on Exam #1 was on Problem 1b), where most people answered no. The correct answer is yes. Consider mutually exclusive events A_1 and A_2. This means $P(A_1 \cup A_2) = 0$. Now, we have $P(A_1) + P(A_2) - P(A_1 \cup A_2) = P(A_1)P(A_2)$. Try to complete the rest of the solution here.

For problem 5b), part ii), another mistake was that many people tried computing the integral for X. However, one can use $V(X) = E(X^2) - [E(X)]^2$ to get $E(X^2) = V(X) + [E(X)]^2$. We know $V(X) = 1$, meaning $E(X^2) = 1 + 100 = 101$. The reason why we know $V(X) = 1$ is because σ^2 is a parameter in $N(\mu, \sigma^2)$, meaning $\sigma = \sqrt{1} = 1$, in this case.

§15.2 Poisson distribution

Definition 15.1. Let $\lambda > 0$. A random variable X has the **Poisson Distribution** with parameter λ if X is a nonnegative integer and has the probability mass function

$$P(X = k) = e^{-\lambda}\frac{\lambda^k}{k!}$$

for $k \in \{0, 1, 2, \ldots\}$. Abbreviate this by $X \sim \text{Poisson}(\lambda)$.

The Poisson distribution is different from most other discrete probability distributions because we've described most discrete distributions (Bernoulli, binomial, geometric) with experiments that are easily understandable. The Poisson distribution, on the contrary, isn't as easily described with such an experiment.

Note that the formula for a Poisson distribution is indeed a probability mass function because it can be rewritten as $e^{\lambda}\sum_{k=0}^{\infty}\frac{\lambda^k}{k!}$, which yields $1 = \sum_{k=0}^{\infty} e^{-\lambda}\frac{\lambda^k}{k!}$. In general, the Poisson distribution tends to model "rare" events, which is a bit of a hand-wavy statement. However, we'll take this for granted and assume it models such "rare" events.

Let's compute the mean and variance of a Poisson random variable.

Example 15.2

Let $X \sim \text{Poisson}(\lambda)$. Then, $\mathbb{E}[X] = \lambda$ and $\text{Var}(X) = \lambda$.

Proof. $\mathbb{E}[X] = \sum_{k=0}^{\infty} ke^{-\lambda}\frac{\lambda^k}{k!} = \sum_{k=1}^{\infty} e^{-\lambda}\frac{\lambda^k}{(k-1)!} = \lambda\sum_{k=1}^{\infty} e^{-\lambda}\frac{\lambda^{k-1}}{(k-1)!} = \lambda\sum_{j=0}^{\infty} e^{-\lambda}\frac{\lambda^j}{j!}$, which is equivalent to λ, $\sum_{j=0}^{\infty} e^{-\lambda}\frac{\lambda^j}{j!}$ is the Taylor Series of e^{λ}.

In the second equality, we dropped the $k = 0$ term and canceled a factor of k. In the next to last equality, we changed the summation index to $j = k - 1$. Finally, we

summed the Poisson probability mass function to 1.

To calculate the variance, we require $\mathbb{E}[X^2]$. Due to the factorial in the denominator, it is more convenient instead to compute $\mathbb{E}[X(X-1)]$, and then use the equality $\mathbb{E}[X^2] = \mathbb{E}[X(X-1)] + \mathbb{E}[X]$, which was justified earlier in the textbook. We have:

$$\begin{aligned}\mathbb{E}[X(X-1)] &= \sum_{k=0}^{\infty} k(k-1)e^{-\lambda}\frac{\lambda^k}{k!} \\ &= \sum_{k=2}^{\infty} e^{-\lambda}\frac{\lambda^k}{(k-2)!} \\ &= \lambda^2 \sum_{j=0}^{\infty} e^{-\lambda}\frac{\lambda^j}{j!} = \lambda^2\end{aligned}$$

Then, we have

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

□

We will now look at the Poisson limit of the binomial. This limit is sometimes called the **Law of Rare Events**. Informally speaking, this law says that if successes are very rare in a sequence of independent trials, then the number of successes is well-approximated by a Poisson random variable. We didn't go over Theorem 15.3 in class, but it's here for completeness.

Theorem 15.3 (Law of Rare Events)

Let $\lambda > 0$ and consider positive integers n for which $\frac{\lambda}{n} < 1$. Let $S_n \sim \text{Bin}(n, \lambda/n)$. Then,

$$\lim_{n\to\infty} P(S_n = k) = e^{-\lambda}\frac{\lambda^k}{k!}$$

for all $k \in \{0, 1, 2, \ldots\}$

In plain English, Theorem 15.4 says that if S_n counts the number of successes in n independent trials, and the mean $\mathbb{E}(S_n) = \lambda$ does not change with n, then as $n \to \infty$, the distribution of S_n approaches the Poisson(λ) distribution. This is an important example of a **limit in distribution**. Below is a proof:

Proof. We rearrange the binomial probability cleverly and then simply observe where the different pieces converge as $n \to \infty$:

$$\begin{aligned} P(S_n = k) &= \binom{n}{k}\left(\frac{\lambda}{n}\right)^k\left(1-\frac{\lambda}{n}\right)^{n-k} \\ &= \frac{n(n-1)\cdots(n-k+1)}{k!}\frac{\lambda^k}{n^k}\left(1-\frac{\lambda}{n}\right)^n\frac{1}{\left(1-\frac{\lambda}{n}\right)^k} \\ &= \frac{\lambda^k}{k!}\left(1-\frac{\lambda}{n}\right)^n\left[1\cdot\left(1-\frac{1}{n}\right)\cdot\left(1\frac{2}{n}\right)\cdots\left(1-\frac{k-1}{n}\right)\right] \end{aligned}$$

As $n \to \infty$, the above expression becomes $\frac{\lambda^k}{k!}\cdot e^{-\lambda}\cdot 1\cdot 1 = e^{-\lambda}\frac{\lambda^k}{k!} = \text{Poisson}(\lambda)$. Note that the proof relied on the fact that $\lim_{n\to\infty}\left(1+\frac{x}{n}\right)^n = e^x$. □

Theorem 15.3 states that the distribution $\text{Bin}\left(n, \frac{\lambda}{n}\right)$ gets closer and closer to $\text{Poisson}(\lambda)$ as $n \to \infty$. However, what if we want a statement for a fixed n? We would want to quantify the error in the approximation of the binomial distribution with the Poisson distribution. The following theorem does exactly this:

Theorem 15.4

Let $X \sim \text{Bin}(n, p)$ and $Y \sim \text{Poisson}(np)$. Then for any subset $A \subseteq \{0, 1, 2, \ldots, \}$, we have

$$|P(X \in A) - P(Y \in A)| \leq np^2$$

The proof of this theorem is not in the textbook. However, for a bit of explanation, note that $\mathbb{E}(X) = \mathbb{E}(Y)$. Furthermore, suppose we have $n = 5$ and $p = 0.9$. $np^2 = 4.05$. This piece of information is useless, as we already know $-1 \leq |P(X \in A) - P(X \in B)| \leq 1$, by the definition of probability, and the bound 4.05 does not give us a better bound.

Example 15.5

Suppose we have $n = 300$ and $p = \frac{1}{300}$ and $A = \{99\}$. Bound $P(X \in A)$, where $X \sim \text{Bin}\left(300, \frac{1}{300}\right)$.

Solution. By Theorem 15.4, we have $|P(X \in A) - P(X \in B)| < 300 \cdot \left(\frac{1}{300}\right)^2$. With some algebraic manipulation, we have $P(X \in B) - \frac{1}{300} \leq P(X \in A) \leq P(X \in B) + \frac{1}{300}$. This is useful. Suppose $A = \{99\}$. Now, our goal is to compute $P(X = 99)$, where $X \sim \text{Bin}\left(300, \frac{1}{300}\right)$. We now have $P(X \in A) = \binom{300}{99}\left(\frac{1}{300}\right)^{99}\frac{299}{300}^{201}$. This is quite difficult to compute. On the other hand, suppose $Y \sim \text{Poisson}(1)$. $P(Y \in A) = e^{-1}\frac{1^{99}}{99!} = e^{-1}\frac{1}{99!}$. While 99! isn't easy to compute, this expression is much simpler than that generated by $P(X \in A)$. Thus, by Theorem 15.4, we have $\frac{e^{-1}}{99!} - \frac{1}{300} \leq P(X = 99) \leq \frac{e^{-1}}{99!} + \frac{1}{300}$.

We can see, from this example, that the Poisson distribution serves as a good approximation for the Binomial distribution. However, note that we need p to be comparatively small when compared with n. □

Fact 15.6 (Poisson approximation for counting rare events)**.** Assume that the random variable X counts the occurrences of rare events that are not strongly dependent on each other. Then, the distribution of X can be approximated with a Poisson(λ) distribution for $\lambda = \mathbb{E}[X]$. That is,

$$P(X = k) \text{ is close to } e^{-\lambda}\frac{\lambda^k}{k!} \text{ for } k \in \{0, 1, 2, \dots, \}$$

Note that the above isn't a mathematically rigorous statement: we did not define what rare events or strong dependence mean. However, the Poisson distribution can still be used to model many real-world phenomena.

§15.3 Poisson processes

Suppose we want to model the times of shark attacks on a particular beach. In a given time interval, the number of shark attacks can be modeled with a Poisson random variable. If we assume that the average rate of attacks is constant in time, the mean of the Poisson random variable should be proportional to the length of the time interval. A useful simplifying assumption is that the numbers of occurrences in nonoverlapping time intervals are independent (two intervals are nonoverlapping if they share at most an endpoint). The independence assumption would be precisely true if the Poisson random variables were approximate counts coming from underlying independent Bernoulli random variables. The below assumptions define Poisson processes. We write $|I|$ for the length of an interval I. For example, $|I| = b - a$ for $I = [a, b]$.

Definition 15.7. The **Poisson process** with intensity (or rate) $\lambda > 0$ is a collection of random points in the half-line $[0, \infty)$. Poisson processes have the following properties:

- The points are distinct (that is, there cannot be more than one point at any given position on $[0, \infty)$
- The number of points in a bounded interval $I \subset [0, \infty)$, which is denoted by $N(I)$, has Poisson distribution with parameter $\lambda \cdot |I|$. For example, if $I = [a, b]$, $N(I) \sim \text{Poisson}(\lambda(b - a))$.
- If $I_1, I_2, \dots, I_n$ are nonoverlapping intervals in $[0, \infty)$, then the random variables $N(I_1), N(I_2), \dots, N(I_n)$ are mutually independent (recall what this means)

The object introduced in the above definition can also be called the **Poisson point process**. The name Poisson process is reserved for the random function of time, $N_t = N([0, t])$, that counts the number of occurrences by time t. For simplicity, we ignore this distinction. Now, we'll compute some interesting probabilities related to Poisson probabilities:

Example 15.8

Example 4.35 in the textbook.

Solution. Each part of this exercise relies on using the Poisson distribution.

(a) Note that the number of customers between 9AM and 10AM is a Poisson process with parameter 5. Thus, we have

$$P(\text{no customers between 9AM and 10AM}) = P(N[9,10] = 0) = e^{-5} \approx 0.00674$$

(b) We look at the customers in the intervals $[9, 10]$, $[10, 10:30]$, and $[2, 3:30]$. These are nonoverlapping intervals, so $N([9,10])$, $N([10,10.5])$, and $N([14,15.5])$ are independent Poisson random variables with parameters 5, $5 \cdot \frac{1}{2}$, and $5 \cdot \frac{3}{2}$. This gives us

$$\begin{aligned} P(N([9,10]) = 2, N([10,10.5]) = 3, N([2,3.5] = 5) \\ &= P(N([9,10]) = 2) \cdot P(N([10,10.5]) = 3) \cdot P(N[(2,3.5]) = 5) \\ &= \frac{5^2}{2!}e^{-5} \cdot \frac{(5/2)^3}{3!}e^{-5/2} \cdot \frac{(15/2)^2}{5!}e^{-15/2} \\ &= \frac{5^{10} \cdot 3^5}{2!3!5!2^8}e^{-15} \approx 0.00197 \end{aligned}$$

(c) Here, we want to find $P(N([10,10:30] = 3 | N([10,12] = 12))$. We can now apply the formula for conditional probability.

$$\begin{aligned} P(N([10,10.5] = 3 | N([10,12]) = 12) &= \frac{P(N([10,10.5]) = 3, N([10,12] = 12))}{P(N([10,12]) = 12} \\ &= \frac{P(N([10,10.5]) = 3, N([10.5,12]) = 9}{P(N([10,12]) = 12} \\ &= \frac{P(N([10,10.5]) = 3 \cdot P(N(10.5,12))}{P(N([10,12]) = 12)} \\ &= \frac{(5/2)^3}{3!}e^{-\frac{5}{2}} \cdot \frac{(15/2)^9}{9!}e^{-\frac{15}{2}} \cdot \left(e^{-10}\frac{10^{12}}{12!}\right)^{-1} \\ &= \binom{12}{3}\left(\frac{1}{4}\right)^3\left(\frac{3}{4}\right)^9 \approx 0.258 \end{aligned}$$

□

Note: there will be a quiz on Thursday, with the problem similar to Examples 4.35 and/or 4.36 in the textbook.

§16 Thursday, October 20, 2022

Today, we'll start Chapter 5 in the textbook.

§16.1 Moment generating functions

Up until now, we have described distributions of random variables with probability mass functions, probability density functions, and cumulative distribution functions. The moment generating function (m.g.f.) offers an alternative way to characterize probability distributions. Furthermore, as then name suggests, it can also be used to compute moments of a random variable.

Before, we start discussing the m.g.f, however, let's review moments.

Theorem 16.1

If $\mathbb{E}[X^n] < \infty, \mathbb{E}[X^j] < \infty$ for any $j \leq \infty$.

Proof. $\mathbb{E}[X^n] = \int_{-\infty}^{\infty} x^n f_X(x)\ dx$. As we are assuming $X > 0$, we can rewrite the above integral as $\int_0^{-\infty} x^n f_X(x)\ dx$. Now, note that $\mathbb{E}[X^j] = \int_0^{\infty} x^j f_X(x)\ dx = \int_0^1 x^J f_X(x)\ dx + \int_1^{\infty} x^j f_X(x)\ dx$. □

Definition 16.2. The **moment generating function** of a random variable X is defined by $M(t) = \mathbb{E}(e^{tX})$. It is a function of the real variable t. We write $M(t)$ as $M_X(t)$ if we wish to distinguish the random variable X.

Remark 16.3. $M_X(t)$ written in its power series for $M_X(t) = \sum_{n=0}^{\infty} \frac{\mu_j}{n!} t^n$, where $\mu_n = \mathbb{E}[X^n]$.

Example 16.4 (Moment generating function of a discrete random variable)

Let X be a discrete random variable with p.m.f.

$$P(X = -1) = \tfrac{1}{3},\ P(X = 4) = \tfrac{1}{6}, \text{ and } P(X = 9) + \tfrac{1}{2}$$

Find the moment generating function of X.

Solution. We can use LOTUS to calculate the m.g.f., with $g(x) = e^{tx}$ and $g(X) = e^{tX}$. We have $M_X(t) = \mathbb{E}[e^{tx}] = \sum_k e^{tk} P(X = k) = \frac{1}{3}e^{-t} + \frac{1}{6}e^{4t} + \frac{1}{2}e^{9t}$. □

§16.2 Exponential distribution

Recall that the geometric distribution is a discrete probability distribution that models waiting times, such as the first time in a sequence of coin lips that tails appears. This section develops a continuous counterpart, for modeling waiting times such as the first arrival of a customer at a post office. This distribution is defined below.

Definition 16.5. Let $0 < \lambda < \infty$. A random variable X has the **exponential distribution** with parameter λ if X has density function

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

on the real line. Abbreviate this by $X \sim \text{Exp}(\lambda)$. The $\text{Exp}(\lambda)$ distribution is called the exponential distribution with **rate** λ.

By integrating the density function, we find the cdf of the $\text{Exp}(\lambda)$ distribution:

$$F(t) = \int_{-\infty}^{t} f(x)\ dx = \int_0^t \lambda e^{-\lambda x}\ dx = 1 - e^{-\lambda t}$$

for $t \geq 0$. Note that $F(t) = 0$ for $t < 0$. By letting $t \to \infty$ in the above integrals, we see that $\int_{-\infty}^{\infty} f(x)\ dx = 1$, meaning f is indeed a probability density function for the exponential distribution.

Example 16.6

Let X be a random variable. Find the mean and variance of $X \sim \text{Exp}(\lambda)$.

Solution. $\mathbb{E}[X] = \int_0^\infty x\lambda e^{-\lambda x}\ dx = \frac{1}{\lambda}$ and $\mathbb{E}[X^2] = \int_0^\infty x^2\lambda e^{-\lambda x}\ dx = \frac{2}{\lambda^2}$. Then, $\text{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}$. □

§16.3 Gamma distribution

Definition 16.7. Let $r, \lambda > 0$. A random variable X has the **gamma distribution** with parameters (r, λ) if X is nonnegative and has probability density function

$$f_X(x) = \frac{\lambda^r x^{r-1}}{\Gamma(r)} e^{-\lambda x} \text{ for } x \geq 0$$

with $f_X(x) = 0$ for $x < 0$. We abbreviate this $X \sim \text{Gamma}(X)$. Note that $\Gamma(r) = (r-1)!$.

§17 Tuesday, October 25, 2022

Let's first look at a few examples of calculating moments.

Example 17.1

Let X be a continuous random variable with probability density function

$$f(x) = \begin{cases} \frac{e^x}{e-1} & \text{if } 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

Find the moment generating function of X.

Solution. To find the expectation of $g(x) = e^{tx}$, we have

$$M_X(t) = \mathbb{E}[e^{tX}] = \int_{-\infty}^{\infty} e^{tx} f(x)\ dx = \int_0^1 e^{tx} \frac{e^x}{e-1}\ dx = \frac{1}{e-1}\int_0^1 e^{(t+1)x}\ dx$$

The last integral above splits into two cases. If $t = -1$, we have $M_X(t) = \frac{1}{e-1}\int_0^1 = \frac{1}{e-1}$. If $t \neq -1$, $M_X(t) = \frac{1}{e-1}\frac{e^{(t+1)}}{t+1}\Big|_{x=0}^{x=1} = \frac{e^{(t+1)}-1}{(e-1)(t+1)}$. □

Remark 17.2. Notice from its definition and the examples above that the moment generating function is *not random*, but rather a function of the variable t. Since $e^0 = 1$, we have $M(0) = \mathbb{E}[e^{0 \cdot X}] = \mathbb{E}[1] = 1$ for all random variables.

Example 17.3 (Moment generating function of the Poisson distribution)
Let $X \sim \text{Poisson}(\lambda)$. Calculate the moment generating function of X.

Solution. We have

$$\mathbb{E}[e^{tX}] = \sum_{k=0}^{\infty} e^{tk} \frac{e^{-\lambda}\lambda^k}{k!} = e^{-\lambda} \sum_{k=0}^{\infty} \frac{(e^t\lambda)^k}{k!} = e^{-\lambda} \cdot e^{e^t\lambda} = e^{\lambda(e^t-1)}.$$

where we used the series expansion of the exponential function. Thus, for $X \sim \text{Poisson}(\lambda)$, we have $M_X(t) = e^{\lambda(e^t-1)}$. □

Example 17.4 (Moment generating function of the normal distribution)
Let $Z \sim \mathcal{N}(0,1)$. Calculate the moment generating function of Z.

Solution. Fill this in ASAP. □

§17.1 Calculations of moments with the moment generating function

Because the moment generating function (in this course and the textbook) is finite around the origin and can be differentiated there, we can use them to find moments. Let $M(t) = \mathbb{E}[e^{tX}]$ and consider the following calculation:

$$M'(t) = \frac{d}{dt}\mathbb{E}[e^{tX}] = \mathbb{E}\left[\frac{d}{dt}e^{tX}\right] = \mathbb{E}[Xte^{tX}].$$

Substituting $t = 0$ gives the formula $M'(0) = \mathbb{E}[X]$. However, the fact that we can move the differentiation inside the expectation is not self-evident (MATH410 discusses how obvious this fact is). However, in the case that X takes only finitely many values, this step is straightforward, because the derivative of a sum is equal to the sum of the derivatives. Thus, we have

$$M'(t) = \frac{d}{dt}\sum_k e^{kt}P(X=k) = \sum_k ke^{kt}P(X=k) = \mathbb{E}[Xe^{tX}]$$

Returning to the general case, we can continue to differentiate as many times as we please by taking the derivative inside the expectation. Write $M^{(n)}$ for the nth derivative of the function M:

$$M^{(n)}(t) = \frac{d^n}{dt^n}\mathbb{E}[e^{tX}] = \mathbb{E}\left[\frac{d^n}{dt^n}e^{tX}\right] = \mathbb{E}[X^n e^{tX}].$$

Taking $t = 0$ gives the following formula:

Theorem 17.5

When the moment generating function $M(t)$ of a random variable X is finite in an interval around the origin, the moments of X are given by

$$\mathbb{E}(X^n) = M^{(n)}(0)$$

§17.2 Joint distribution of discrete random variables

We have come to a major juncture in the course where our point of view shifts from a single random variable to larger collections of random variables. We have seen some instances of this; for example, in our discussions in sequences of trials. The remainder of this course (as well as probability theory in general) is concerned with studying collections of random variables in various situations.

If $X_1, X_2, \ldots, X_n$ are random variables defined on a sample space ω, we can regard them as coordinates of the **random vector** $(X_1, X_2, \ldots, X_n)$. This vector is a random variable in the sense of being a function that maps ω to $\mathbb{R}^n$. The probability distribution of $(X_1, X_2, \ldots, X_n)$ can be adapted to the assignment of probabilities $P((X_1, X_2, \ldots, X_N) \in B$ (where B is a subset of $\mathbb{R}^n$).

Example 17.6

Suppose you have \$50. You play the following game: toss a fair coin; if it comes up heads you win \$2, and if it comes up tails, you lose \$1. How much do you expect to have at the end of 100 tosses?

The above is an example of a scenario that would be better analyzed with a random vector.

Definition 17.7. The probability distribution of a random vector is called a **joint distribution**, and the probability distributions of the individual coordinates X_j are called **marginal distributions**.

A joint distribution can again be described by a probability mass function in the discrete case, and a probability density function in the jointly continuous case. In general, one can define a multivariate cumulative distribution function.

Definition 17.8. Let $X_1, X_2, \ldots, X_n$ be discrete random variables, all defined on the same sample space. Their **joint probability mass function** is defined by

$$p(k_1, k_2, \ldots, k_n) = P(X_1 = k_1, X_2 = k_2, \ldots, X_n = k_n)$$

for all possible values $k_1, k_2, \ldots, k_n$ of $X_1, X_2, \ldots, X_n$.

As with other notations, we can write the joint probability mass function as $p_{X_1, X_2, \ldots, X_n}(k_1, k_2, \ldots, k_n)$.

§18 Thursday, October 27, 2022

Today, we'll continue discussing joint distributions of random variables. Recall from last lecture the joint probability mass function. This function has the same properties as the probability mass function for a single random variables, e.g.

$$\sum_{k_1, k_2, \ldots, k_n} p_{X_1, X_2, \ldots, X_n}(k_1, k_2, \ldots, k_n) = 1 \text{ and } p_{X_1, X_2, \ldots, X_n} \geq 0$$

Definition 18.1. A collection of random variables is said to be **jointly continuous** if there exists a function, called the **joint probability density function** such that

$$P((X_1, X_2, \dots X_n) \in B) = \int \int \int \cdots \int f_{X_1, X_2, \dots, X_N}(x_1, x_2, \dots, x_n)\ dx_1 dx_2 \dots dx_n$$

where there are $|B|$ (the cardinality of B) integrals.

Definition 18.2. The probability mass function and probability density function of one random variable (that is part of a random vector) are known as the **marginal probability mass function** and **marginal probability density function**, respectively.

Theorem 18.3

Let f be the joint density function of $X_1, \dots, X_n$. Then, each random variable X_j has a density function f_{X_j} that can be obtained by integrating away the other variables away from f:

$$f_{X_j}(x) = \underbrace{\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty}}_{n-1 \text{ integrals}} f(x_1, \dots, x_{j-1}, x, x_{j+1}, \dots, x_n)\ dx_1, \dots, dx_{j-1}, dx_{j+1}, \dots, dx_n$$

In words, this formula states that to compute $f_{X_j}(x)$, place x in the jth coordinate inside f and integrate away the other $n-1$ variables. For two random variables X and Y, the formula is

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\ dy$$

A proof of Theorem 18.3 is provided in the textbook.

Definition 18.4 (Conditional pmf/pdf)**.** If (X, Y) is a random vector consisting of discrete random variables, we have

$$p_{X,Y}(x, y) = p_X(x) \cdot p_{Y|X}(y|x)$$

where p is the joint probability mass function. If (X, Y) is a random vector consisting of continuous random variables, we have

$$f_{X,Y}(x, y) = f_X(x) \cdot f_{Y|X}(y|x)$$

where f is the joint probability density function.

§19 Tuesday, November 1, 2022

Today, we'll combine two of the central themes in this course: conditional probabilities and joint distributions of random variables in order to define and study conditional distributions of random variables. Conditional distributions are fundamental to the study of stochastic processes.

§19.1 Conditional distribution of a discrete random variable

Definition 19.1. Let X be a discrete random variable and B an event with $P(B) > 0$. The **conditional probability mass function** of X, given B, is the function $p_{X|B}$ defined as follows for all possible values k of X:

$$p_{X|B}(k) = P(X = k|B) = \frac{P(\{X = k\} \cap B)}{P(B)}$$

The conditional pmf $p_{X|B}(k)$ behaves like a regular pmf: each of its values are non-negative, and the sum of these values is 1. A justification of this is in the textbook. The key idea in the proof is that each of the events $\{X = k\} \cap B$ are disjoint for different values of k, and their union over k is B.

Applying the Law of Total Probability to the event $A = \{X = k\}$, we obtain the following:

Theorem 19.2

Let Ω be a sample space, X a discrete random variable on Ω, and $B_1, \ldots, B_n$ a partition of ω such that each $P(B_i) > 0$. Then, the unconditional pmf of X can be calculated by averaging the conditional pmfs:

$$p_X(k) = \sum_{i=1}^{n} p_{X|B_i}(k) P(B_i)$$

The averaging idea extends to expectations:

Theorem 19.3

Let Ω be a sample space, X a discrete random variable on Ω, and $B_1, \ldots, B_n$ a partition of ω such that each $P(B_i) > 0$. Then,

$$\mathbb{E}[X] = \sum_{i=1}^{n} \mathbb{E}[X|B_i] P(B_i)$$

To be fully precise, we need to assume that all expectations and conditional expectations are well-defined in Fact 10.4.

A proof of the above is presented in the textbook. The next step is to let the partition in Theorems 19.4 and 19.4 come from another discrete random variable Y. Here are the key definitions:

Definition 19.4. The X and Y be discrete random variables. Then, the conditional probability mass function of X given $Y = y$ is the following two-variable function:

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

§19.2 Conditional distribution for jointly continuous random variables

In the previous section, we defined a conditional probability mass function $p_{X|Y}(x|y)$ for X by conditioning on the event $\{Y = y\}$ for a discrete random variable Y. If Y is a continuous random variable, we run into a problem: $P(Y = y) = 0$, so we cannot condition on this event. However, ratios of density functions turn out to be meaningful, and enable us t o define conditional density functions:

Definition 19.5. Let X and Y be jointly continuous random variables with joint density function $f_{X,Y}(x, y)$. The conditional density function of X, given $Y = y$, is denoted by $f_{X|Y}(x|y)$ and defined as

$$f_{X|Y}(x|y) \frac{f_{X,Y}(x, y)}{f_Y(y)}$$

for y such that $f_Y(y) > 0$.

It is verified in the textbook that $\int_{-\infty}^{\infty} f_{X|Y}(x|y)dx = 1$. With the conditional pdf defined, we can establish the analogues of the definitions and facts we defined with the conditional pmf. Just as an ordinary density function is used to calculate probabilities and expectations, a conditional density function is used to calculate conditional probabilities and conditional expectations. The definition below gives the continuous counterparts of the discrete formulas:

Definition 19.6. The conditional probability that $X \in A$, given $Y = y$, is

$$P(X \in A \mid Y = y) = \int_A f_{X|Y}(x|y)\ dx$$

Example 19.7

Suppose X and Y are jointly continuous random variables with a joint probability density function given by

$$f_{X,Y}(x,y) = \begin{cases} k & (x,y) \in R \\ 0 & \text{otherwise} \end{cases}$$

where R is given by $-1 \leq x \leq 1$ and $0 \leq y \leq \sqrt{1-x^2}$. Find the value of k, as well as $f_X(x)$ and $f_Y(y)$. Also find

- $P\left(0 < X < \frac{1}{2}\right)$
- $P(Y \geq X)$
- $f_{Y|X}(y|x)$ and $f_{X|Y}(x|y)$
- $P(0 < Y < \frac{1}{2} | X = 0)$.

Solution. To find k, note that

$$\begin{aligned} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{x,y}(x,y)\, dxdy &= 1 \\ \int_{-1}^{1} \int_{0}^{\sqrt{1-x^2}} k\, dydx &= 1 \\ \frac{\pi k}{2} = 1 \to k &= \frac{2}{\pi} \end{aligned}$$

To find the marginal pdf of X, we will fix the value of x and sum (or integrate, if the random variables are continuous) over all possible values of y, e.g.

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} f_{x,y}(x,y)\, dy = \int_{0}^{\sqrt{1-x^2}} \frac{2}{\pi}\, dy \\ &= \frac{2}{\pi}\sqrt{1-x^2} \end{aligned}$$

To compute $P(0 < X < \frac{1}{2})$, we want

$$\int_0^{\frac{1}{2}} \frac{2}{\pi}\sqrt{1-x^2}\, dx = \frac{2}{\pi} \int_0^{\frac{1}{2}} \sqrt{1-x^2}$$

$P(0 < X < \frac{1}{2})$ can be computed from the above integral.

To find $P(Y \geq X)$, we have

$$P(Y \geq X) = \iint_R f_{X,Y}(x,y)\ dxdy$$

As before, $P(Y \geq X)$ can be computed using the above formula. It may be easier to split the region R into two regions R_1 and R_2, and sum the areas obtained over both areas.

To compute $f_{X|Y}(x|xy)$ and $f_{Y|X}(y|x)$, we can use the formulas $f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)}$. We have

$$\begin{aligned} f_{Y|X}(y|x) &= \frac{f_{X,Y}(x,y)}{f_X(x)} \\ &= \frac{\frac{2}{\pi}}{\frac{2}{\pi}\sqrt{1-x^2}} \\ &= \frac{1}{\sqrt{1-x^2}} \end{aligned}$$

$f_{X|Y}(x|y)$ can be computed similarly. To compute $P\left(0 < Y < \frac{1}{2} | X = 0\right)$, we can compute $\int_0^{\frac{1}{2}} f_{Y|X}(y|0)\ dy = \int_0^{\frac{1}{2}} \frac{1}{1}\ dy = \frac{1}{2}$.

□

§19.3 Independence

Recall our formula for conditional probability of joint random variables. We have

$$\begin{aligned} f_{Y|X}(y|x) &= \frac{f_{X,Y}(x,y)}{f_X(x)} \\ \rightarrow f_{X,Y}(x,y) &= f_X(x) f_{Y|X}(y|x) \end{aligned}$$

If X and Y are independent, we have $f_{Y|X}(y|x) = f_Y(y)$, and $f_{X,Y}(x,y) = f_X(x)f_Y(y)$. The next theorem presents this fact more generally:

Theorem 19.8

$X_1, \ldots, X_n$ are independent if and only if the joint density (or mass) function is the product of the marginal distributions, e.g.

$$f_{X_1,\ldots,X_N}(x_1,\ldots,x_n) = f_{X_1}(x_1) f_{X_2}(x_2) \cdots f_{X_n}(x_n)$$

Note that the textbook presents Theorem 19.2 separately for the discrete and continuous case; both cases are presented simultaneously above.

Example 19.9

Recall the experiment from last class, where a coin toss was followed by a die roll. Let X be a random variable representing the outcome of the coin toss, and Y represent the outcome of the die roll. Are X and Y independent?

Theorem 19.10

Suppose $X_1, \ldots, X_{m+n}$ are independent random variables. Let $f : \mathbb{R}^m \to \mathbb{R}$ and let $g : \mathbb{R}^n \to \mathbb{R}$ be real-valued functions of multiple variables. Define random variables $Y = f(X_1, \ldots, X_m)$ and $Z = g(X_{m+1}, \ldots, X_{m+n})$. Then, Y and Z are random variables.

Remark 19.11. Note carefully that Y and Z are functions of *distinct* independent random variables.

Solution. Considering the experiment, X and Y are clearly not independent. If $X = 0$ (a toss of heads), $Y = 5$ is possible, whereas $X = 1$ and $Y = 5$ are impossible. This implies $P_{Y|X}(5|0) \neq P_{Y|X}(5|1)$. If these events were independent, then these two probabilities should be independent. However, as they are not, X and Y are not independent. Computing the probabilities is left as an exercise to the reader. □

§19.4 Expectation

In the case with one random variable, we have $\mathbb{E}(X) = \int_{-\infty}^{\infty} x f_X(x)\ dx$. In the joint random variable case, what is $\mathbb{E}(X_1, \ldots, X_n) = \int_{-\infty}^{\infty} > \cdot f_X(x)/dx$? Consider $g(X_1, \ldots, X_n)$. We have $\mathbb{E}(g(X_1, \ldots, X_N)) = \int_{-\infty}^{\infty} y\ f_Y(y)\ dy$ where $y = g(X)$. To compute $\mathbb{E}(g(X_1, \ldots, X_n))$, we can use LOTUS:

$$\mathbb{E}(g(X_1, \ldots, X_n)) = \int_{-\infty}^{\infty} g(X_1, \ldots, X_n) f_{X_1, \ldots, X_N}(X_1, \ldots, X_n)\ dx_1 \cdots dx_n$$

In the presence of independence, the above formula becomes

$$\mathbb{E}(g(X_1, \ldots, X_n)) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(X_1, \ldots, X_n) f_{X_1, \ldots, X_N}(X_1, \ldots, X_n)\ dx_1 \cdots dx_n$$

Having multiple integrals with independence is useful, as it allows one to change the order of integration.

§19.4.1 Linearity of expectation

Linearity is arguably one of the most important properties of expected value.

Theorem 19.12 (Linearity of Expectation)
If $X_1, \ldots, X_N$ are discrete random variables, we have

$$\mathbb{E}\left(\sum_{i=1}^{n}(a_i X_i)\right) = \sum_{i=1}^{n} a_i \mathbb{E}(X_i)$$

A similar result holds for the continuous case.

Proof. We will first prove this Theorem for the continuous random variable. We have

$$\begin{aligned}
\mathbb{E}(aX + bY) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(ax + by)(f_{X,Y}(x,y)\ dxdy \\
&= a\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x f_{X,Y}(x,y)\ dxdy + b\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y f_{X,Y}(x,y)\ dxdy \\
&= a\mathbb{E}(X) + b\mathbb{E}(Y)
\end{aligned}$$

The proof for the case in which X and Y are discrete random variables is similar, and done by induction. It is left as an exercise. □

Before we move on, we must note that there are two ways to compute $\mathbb{E}(X)$. Previously, we used the formula $\mathbb{E}(X) = \int_{-\infty}^{\infty} x f_X(x)\ dx$. Now, we can use $\mathbb{E}(X) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x f_{X,Y}(x,y)\ dxdy$. How do we verify that both formulas give us the same result? We have

$$\begin{aligned}
\mathbb{E}(X) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x f_{X,Y}(x,y)\ dxdy &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x f_{X,Y}(x,y)\ dydx \\
&= \int_{-\infty}^{\infty} x\int_{-\infty}^{\infty} f_{X,Y}(x,y)\ dydx \\
&= \int_{-\infty}^{\infty} x f_X(x) dx
\end{aligned}$$

The last step above follows from the definition of $f_X(x)$, as $f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\ dx$, by Theorem 18.3.

Theorem 19.13
Let $X_1, \ldots, X_n$ be independent random variables. Then, for all functions $g_1, \ldots, g_n$ for which the expectations below are well defined, we have

$$\mathbb{E}\left[\prod_{k=1}^{n} g_k(X_k)\right] = \prod_{k=1}^{n}\mathbb{E}[g_k(X_k)]$$

§19.4.2 Variance

Theorem 19.14

Assume the random variables $X_1, \ldots, X_n$ are independent and have finite variances. Then,

$$\text{Var}(X_1 + X_2 + \cdots + X_n) = \text{Var}(X_1) + \text{Var}(X_2) + \cdots + \text{Var}(X_n)$$

A proof of the above theorem is presented in the textbook.

§20 Thursday, November 3, 2022

To recap, so far, we've covered sections 6.1-6.3, 10.1-10.2, and 8.1 in the textbook. Before the second exam on 11/17, our goal is to cover 6.4, 8.2-8.3, and 10.3 in the book. Another important point to note is that we have covered several *named distributions* with single random variables. We will not cover all of these for joint distributions. On an exam, the pdf/pmf of a named distribution will be given; these are not required to memorize.

§20.1 Fubini and Tonelli's Theorems

We will now discuss certain important theorems related to integrals. So far, we have assumed that the order of integration doesn't matter when computing iterated integrals. These two theorems will help explain how

§20.1.1 Tonelli's Theorem

Theorem 20.1 (Tonelli's Theorem)

Given certain technical conditions are met (such conditions are beyond the scope of this class), and $f_{X,Y}(x, y) > 0$ for random variables X and Y,

$$\iint_{A \times B} f(x, y)\, dxdy = \int_B \left(\int_A f(x, y)\, dx \right) dy = \int_A \left(\int_B f(x, y)\, dx \right) dy$$

Tonelli's Theorem implies that for pdfs, switching the order of integration does not matter. While there are certain technicalities that are needed for Tonelli's Theorem to be invoked, they are always met in our context.

§20.1.2 Fubini's Theorem

Theorem 20.2 (Fubini's Theorem)

Given certain technical conditions are met (such conditions are beyond the scope of this class), if either

$$\int_A \int_B |f(x,y)| \, dxdy < \infty$$

or

$$\int_B \int_A |f(x,y)| \, dydx < \infty$$

we have

$$\int_A \int_B |f(x,y)| \, dxdy = \int_B \int_A |f(x,y)| \, dydx = \iint_{A \times B} |f(x,y)| \, dxdy$$

Fubini's Theorem tells us that if we can evaluate one of these integrals and obtain a finite answer, we can evaluate all of these integrals and will obtain the same answer.

Example 20.3

Let X, Y be jointly continuous random variables with joint pdf

$$f_{X,Y}(x,y) = \begin{cases} cx+1 & x,y \geq 0, x+y<1 \\ 0 & \text{otherwise} \end{cases}$$

1. Find c
2. Compute the marginal probability density functions of X and Y
3. Find $P(Y < 2X^2)$
4. Find $\mathbb{E}[X^2Y^2]$

Solution. For 1), as any pdf must integrate to 1, we have $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dxdy = 1$. It helps to sketch the region that we want to integrate over using the fact that $x+y<1$ and graphing $x+y=1$. Now, we can set up an iterated integral. Outside of the triangular region, the integral would evaluate to 0, meaning we only have to look inside the region. We can set up the integral as follows: $\int_0^1 \int_0^{1-x} (cx+1)\,dydx = 1 \to \int_0^1 (cx+1)y|_0^{1-x}\,dx = \int_0^1 cx - cx^2 + 1 - x \to \frac{cx^2}{2} - \frac{cx^3}{3} + x - \frac{x^2}{x}|0^1 = \frac{c}{2} - \frac{c}{3} + 1 - \frac{1}{2} - 0 = 0 \to \frac{c}{6} + \frac{1}{2} = 1 \to c = 3$.

To find $f_X(x)$, we can fix x and look at the change in y and compute $\int_{-\infty}^{\infty} f_{X,Y} f(x,y)\ dy = \int_0^{1-x}(3x+1)\ dy = (3x+1)(1-x)$ for $x \in [0,1]$. To find $f_Y(y)$, we can fix y and look at the change in x and compute $\int_{-\infty}^{\infty} f_{X,Y}(x,y)\ dx \to \int_0^{1-y}(3x+1)\ dx = \frac{3x^2}{2} + x|_0^{1-y} = \frac{3(1-y)^2}{2} + (1-y) = (1-y)\frac{3(1-y)+2}{2} = \frac{(1-y)(5-3y)}{2}$. for $y \in [0,1]$

To find $P(Y < 2X^2)$, we can draw a diagram where the curves $x+y=1$ and $y=2x^2$ intersect. After sketching this region, it would be better to split the region up into two places using the point at where $2x^2$ and $x+y=1$ intersect. Setting up the integral, we see as x changes from 0 to $\frac{1}{2}$, the change in y is bounded above by the curve $y=2x^2$; as x changes from $\frac{1}{2}$ to 1, the change in y is bounded above by the curve $x+y=1$. We can sum the results from integrals over both regions to obtain the area of the entire region. Our integrals become $\int_0^{\frac{1}{2}} \int_0^{2x^2}(3x+1)\ dydx + \int_{\frac{1}{2}}^1 \int_0^{1-x}(3x+1)\ dydx$. These integrals now become $\int_0^{\frac{1}{2}} 2x^2(3x+1)\ dx + \int_{\frac{1}{2}}^1 (3x+1)(1-x)\ dx$.

To find $\mathbb{E}[X^2Y^2]$, we want to compute $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x^2y^2 f_{X,Y}(x,y)\ dydx$. Note that we can use either $dxdy$ or $dydx$ in the integral, as Fubini's Theorem tells us that the order of integration doesn't matter. Our integral becomes $\int_0^1 \int_0^{1-x} x^2y^2(3x+1)\ dydx$. □

§20.2 Effect of independence on expectation

Independence is truly where probability becomes something that doesn't involve just setting up an integral and computing it. As explained earlier, independence (in the 2-variable case) is when $f_{X,Y}(x,y) = f_X(x)f_Y(y)$. Joint random variables are independence if and only if their joint distribution is equal to the product of the marginal distributions.

In the scope of expected value, independence doesn't affect the linearity of expectation. It does, however, affect

Theorem 20.4

If random variables X and Y are independent we have $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$.

Proof. In the continuous case, we have $\mathbb{E}[XY] = \int_{-\infty}^{\infty} xyf_{X,Y}(x,y)\ dxdy$. This, in turn, is equal to

$$\begin{aligned}\int_{-\infty}^{\infty} xyf_X(x)f_Y(y)\ dxdy &= \int_{-\infty}^{\infty} yf_Y(y)\left[\int_{-\infty}^{\infty} xf_X(x)\ dx\right]dy \\ &= \int_{-\infty}^{\infty} f_X(x)\ dx \cdot \int_{-\infty}^{\infty} yf_Y(y)\ dy \\ &= \mathbb{E}[X]\cdot\mathbb{E}[Y]\end{aligned}$$

The reason why we were able to take out the $\int_{-\infty}^{\infty} xf_X(x)\ dx$ from the integral is because it is constant with respect to y.

□

§21 Tuesday, November 8, 2022

Class is canceled today.

§22 Thursday, November 10, 2022

Class is on Zoom today. Today, we will cover Section 6.4 of the textbook, which is about computing the pdf of a function of a random vector. Recall that in the past, we covered questions such as "If $X \sim \mathcal{U}[0, 10]$, what is the pdf of $Z = e^X$?" We will attempt to cover similar questions with joint distributions in Section 7.1.

We will also try to cover Section 7.1, which discusses questions such as "What is the pdf of $X + Y$, if X and Y are random vectors and both X and Y are normally distributed wiht mean μ_X and μ_Y and variance σ_X^2 and σ_Y^2. We have that $X + Y \sim \mathcal{N}(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$. Many classical results in statistics depend on the above insight. This result is used extensively in statistical inference and hypothesis testing, which is covered in STAT401.

We will also cover Section 10.3, which is about conditional expectation. We've looked at conditional probability, In the past, we've looked at questions such as $p_{X|Y}(X > 13, Y = 10)$. With conditional expectation, we'll attempt to look at questions such as $\mathbb{E}[X|Y = 10]$.

Now, we'll start Section 6.4.

§22.1 Further multivariate topics

We will now cover the joint cumulative distribution function and then the bivariate normal distribution.

§22.1.1 Joint cumulative distribution function

We have discussed two ways of describing the joint distribution of multiple random variables $X_1, X_2, \ldots, X_n$. Discrete random variables have a joint probability mass function, whereas jointly continuous random variables have a joint probability density function. These are natural extensions of their one-dimensional counterparts. We can also extend the cumulative distribution function to several variables:

Definition 22.1. The **joint cumulative distribution function** of random variables $X_1, X_2, \ldots, X_n$ is a function of n variables defined as

$$F(s_1, \ldots, s_n) = P(X_1 \leq s_1, \ldots, X_n \leq s_n)$$

for all $s_1, \ldots, s_n \in \mathbb{R}$.

The joint cdf completely identifies the joint distribution; while it is a bit cumbersome for computations, it can help find the joint density function. We can discuss this connection for two jointly continuous random variables. The general case is similar.

If X and Y have joint density function f, their joint cumulative distribution function is

$$F(x,y) = P(X \leq x, Y \leq y) = \int_{-\infty}^{x} \int_{-\infty}^{y} F(s,t)\, dtds$$

In the opposite direction, if we assume that the joint density function is continuous at a point (x,y), then the density function can be obtained from the joint cdf by taking the mixed partial derivative and using the Fundamental Theorem of Calculus:

$$\frac{\partial^2}{\partial x \partial y} F(x,y) = f(x,y)$$

The joint cdf can also be used to identify if random variables are independent or not. The condition is the factorization of the joint cdf as the product of the marginal cdfs.

Theorem 22.2

Random variables $X_1, X_2, \ldots, X_n$ are independent if and only if

$$f_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n) = \prod_{k=1}^{n} f_{X_k}(x_k)$$

where $f_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n)$ is the joint pdf of $X_1, X_2, \ldots, X_n$ and $f_{X_k}(x_K)$ is the marginal pdf of X_i.

§22.2 Standard bivariate normal distribution

We will now discuss the joint pdf for two *dependent* normal random variables. Let $-1 < \rho < 1$ and suppose that the joint density function of the random variables X and Y is given by

$$f(x,y) = \frac{1}{2\pi\sqrt{1-\rho^2}} e^{-\frac{x^2+y^2-2\rho xy}{2(1-\rho)^2}}$$

The multivariate distribution defined by the joint density function is called the **standard bivariate normal** with parameter ρ. Note that we must check that this is a probability density function, and that X and Y both have the standard normal distribution. This is in the textbook. The probabilistic meaning o

§22.2.1 Infinitesimal method

As with single-variable density functions, it should be kept in mind that the value $f(x_1, x_2, \ldots, x_n)$ is *not* the probability of an event. Rather, the value of $f(x_1, \ldots, x_n)$ can be used to approximate the probability that $(X_1, \ldots, X_n)$ lies in a small n-dimensional cube around the point $(x_1, \ldots, x_n)$.

Theorem 22.3

Suppose $X_1, \ldots, X_n$ have joint density function f and f is continuous at the point $(a_1, \ldots, a_n)$. Then, for small $\epsilon > 0$, we have

$$P(X_1 \in (a_1, a_1 + \epsilon), \ldots, X_n \in (a_n, a_n + \epsilon)) \approx f(a_1, \ldots, a_n) \cdot \epsilon^n$$

As in the 1-dimensional case, the precise meaning of this statement is that

$$\lim_{\epsilon \to 0} \epsilon^{-n} P(X_1, \in (a_1 + a_1 + \epsilon), \ldots, X_n \in (a_n, a_n + \epsilon)) = f(a_1, \ldots, a_N)$$

§22.2.2 Transformation of a joint density function

Consider two jointly continuous random variables (X, Y) with joint density function $f_{X,Y}$. Let another pair of random variables (U, V) be defined as functions of $(X, Y) : U = g(X, Y), V = h(X, Y)$. The goal is to find the joint density function $f_{U,V}$ of (U, V) through a multivariate change of variables. We need to be slightly technical about the assumptions we make when we try to find this joint pdf.

Let K be a region of the xy-plane so that $f_{X,Y}(x, y) = 0$ outside K. This implies $P\{(X, Y) \in K\} = 1$. Let $G(x, y) = (g(x, y), h(x, y))$ be a one-to-one function that maps K onto a region L on the plane. Denote the inverse of G with $G^{-1}(u, v) = (q(u, v), r(u, v))$. This means that q and r are functions from the region L to K that satisfy

$$u = g(q(u, v), r(u, v)) \text{ and } v = h(q(u, v), r(u, v)).$$

Assume that the functions q and f satisfy the following conditions:

(i) q and r both have continuous partial derivatives with respect to u and v on L

(ii) The Jacobian of these partial derivatives does not vanish anywhere on L

Theorem 22.4

If functions q and r satisfy both of the conditions above, the joint density function of (U, V) is given by

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

for $(u,v) \in L$ and $f_{U,V}(u,v) = 0$ outside L.

The proof of this theorem entails checking that

$$\mathbb{E}[w(U,V)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} w(u,v) f_{U,V}(u,v)\, dudv$$

when $f_{U,V} = f_{X,Y}(q(u,v), r(u,v))|J(u,v)|$. Theorem 22.4 generalizes to higher dimensions; when there are n variables, the Jacobian becomes an $n \times n$ matrix of partial derivatives of the inverse of G.

Next Thursday, we will have the second midterm exam, which will focus on sections 4.4, 4.5, 5.1, Chapter 6, 7.1, 10.1, and 10.2 (excluding computations of conditional expectations). Tuesday will be a review day.

§22.3 Sums of independent random variables

We will now look at deriving distributions of sums of independent random variables and demonstrate how to use symmetry to simplify complicated calculations.

§22.3.1 Sums and symmetry

Suppose X and Y are random variables. Let's describe the distribution of $X + Y$ in both the discrete and jointly continuous cases, as well as when X and Y are independent.

Suppose first that X and Y are discrete with the pmf $p_{X,Y}$. Then, $X + Y$ must also be discrete, and its pmf can be computed by breaking up the event $\{X + Y = n\}$ into the disjoint union of the events $\{X = k, Y = n - k\}$ over all possible values of k:

$$p_{X+Y}(n) = P(X + Y = n) = \sum_k p_{X,K}(k, n-k)$$

As X and Y are independent, we can write the joint probability mass function as the product of the marginal probability mass functions and get $p_{X+Y}(n) = \sum_k p_X(k)p_Y(n-k)$. The same argument with $X = n - l$ and $Y = l$ gives $\sum_k p_X(n - l)p_Y(n)$. This operation that produces the new probability mass function p_{X+Y} from p_X and p_Y is a type of product of probability mass functions.

Definition 22.5. The above operation is called the **convolution** and is denoted by $p_X * p_Y$:

$$p_X * p_Y = \sum_k p_X(k)p_Y(n-k) = \sum_l p_X(n-l)p_Y(l)$$

Now, let's look at jointly continuous random variables X and y with joint pdf $f_{X,Y}$. First, we can identify the cdf of $X+Y$:

$$\begin{aligned} F_{X+Y}(z) = P(X+Y \leq z) &= \iint_{x+y\leq z} f_{X,Y}(x,y)dxdy \\ &= \int_{-\infty}^{\infty}\left(\int_{-\infty}^{\infty} f_{X,Y}(x,y)dy\right)dx = \int_{-\infty}^{\infty}\left(\int_{-\infty}^{z} f_{X,Y}(x,w-x)dw\right)dx \\ &= \int_{-\infty}^{z}\left(\int_{-\infty}^{\infty} f_{X,Y}(x,w-x)dx\right)dw \end{aligned}$$

By definition, the cdf F_{X+Y} and the pdf f_{X+Y} are related by

$$F_{X+Y}(z) = \int_{-\infty}^{z} f_{X+Y}(w)\ dw$$

We can compare the final integral above and obtain $f_{X+Y} =$

$$\int_{-\infty}^{\infty} f_{X,Y}(x, z-x)dx$$

If X and Y are independent, we can simplify the integrand to $f_X(x)f_Y(z-x)\ dx$. The integral $\int_{-\infty}^{\infty} f_X(x)f_Y(z-x)dx$ is again called the **convolution** of f_X and f_Y, and denoted by $f_X * f_Y$. A similar argument yields $f_{X+Y}(z) = \int_{-\infty}^{\infty} f_X(z-x)f_Y(x)\ dx$ Fact 7.1 in the textbook summarizes these findings.

§23 Tuesday, November 15, 2022

§23.1 Exam # 2 Review

Today is a review day for the second exam. We will review the practice problems, emphasizing the difficult ones:

1. (a) We want to compute $P(N_1 = 10)$. As the cars arriving is a Poisson process, our answer becomes $e^{-10} \cdot \frac{10^{10}}{10!} \cdot \left(\frac{1}{2}\right)^{10}$.
   (b) Our answer is $e^{-y} \cdot \frac{10^y}{y!} \cdot \binom{y}{10} \cdot \left(\frac{1}{2}\right)^{10} \cdot \left(\frac{1}{2}\right)^{y-10}$.

(c) This part is a bit more intricate. The first thing to note is that the number of arrivals should be exactly 10, not more or less. Our answer becomes

$$\sum_{i=10}^{\infty} e^{-10} \cdot \frac{10^i}{i!} \cdot \binom{i}{10} \cdot \left(\frac{1}{2}\right)^j \cdot \left(\frac{1}{2}\right)^{j-10}$$

2.

3. We have

$$p_X(x) = \begin{cases} p_1 & x = 1 \\ 1 - p_1 & x = 0 \\ 0 & \text{otherwise} \end{cases}$$

and

$$p_Y(y) = \begin{cases} \frac{1}{2} & x = 1 \\ \frac{1}{2} & x = 0 \\ 0 & \text{otherwise} \end{cases}$$

When

4. The first thing you want to do for questions like this is determine the region you will be integrating over. To find the value of k, we can solve for k with the integral

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x, y)\ dxdy = 1$$

As y is bounded between 0 and $\sqrt{x}$, we will instead use

$$\begin{aligned} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x, y)\ dydx &= 1 \\ &= \int_0^1 \int_0^{\sqrt{x}} kxy\ dydx = 1 \\ &= k \int_0^1 \frac{x^2}{2}\ dx \\ &= \frac{k}{2} \int_0^1 x^2 = \frac{k}{6} = 1 \end{aligned}$$

Thus, $k = 4$, meaning $f_{X,Y}(x, y) = 6xy$. To find $f_Y(y)$, we must determine the bounds of our integral.

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\ dx = \int_{y^2}^{1} 6xy\ dx$$

Computing $f_X(x)$ is similar:

$$\int_{-\infty}^{\infty} 6xydy = \int_{0}^{\sqrt{x}} 6xy\ dy$$

To compute $P(Y > X)$, we can draw a picture of the region that we would like to integrate over. This region is the region above the two points of intersection of $y = \sqrt{x}$ and $y = x$. After analyzing this, our integral becomes

$$\int_0^1 \int_x^{\sqrt{x}} 6xy\ dydx$$

Note that our bounds for x and $\sqrt{x}$ are as such because on the interval $[0, 1]$, $\sqrt{x} > x$. We can look at the region that we sketched to determine the bounds. A key strategy in these problems is to draw a picture and sketch the region that you are working with. To check if X and Y are independent, we can check if $f_{X,Y}(x,y) = f_X(x)f_Y(y)$. After doing this, we have that X and Y are not independent. Finding $f_{X|Y}(x|y)$ is a relatively straightforward application of

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$$

where $f_Y(y) \neq 0$ and $(x,y) \in \mathbb{R}^2$. I won't include the calculations for $\mathbb{E}[X|Y = y]$ and $\text{Var}X|Y = y$; they won't be on the exam, and are left as an exercise.

5. Let $Z = X + Y$ and $W = X - Y$. We would like to obtain $f_Z(z)$ and $f_W(w)$. We obtain $X = \frac{Z+W}{2}$ and $Y = \frac{Z-W}{2}$. We must now compute the Jacobian determinant with $h_1 = \frac{Z+W}{2}$ and $h_2 = \frac{Z-W}{2}$. Our Jacobian is

$$\begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} \end{pmatrix} \tag{6}$$

Thus, our Jacobian determinant is $\frac{1}{4} + \frac{1}{4} = \frac{1}{2}$. We have $f_{W,Z}(w,z) = f_{X,Y}(h_2,h_1) \cdot \frac{1}{2} = f_{X,Y}\left(\frac{Z-W}{2}, \frac{Z+W}{2}\right) \cdot \frac{1}{2}$. Now, we have

$$f_Z(z) = \int_{-\infty}^{\infty} f_{W,Z}(w,z)\ dw$$

and

$$f_W(w) = \int_{-\infty}^{\infty} f_{W,Z}(w,z)\ dz$$

6. Let X_i be a random variable such that $X_i(k)$ represents the digit in the ith position for a particular number. We will consider a simpler version of this problem, looking at the set $\{00, 01, \ldots, 99\}$. The distribution of each X_i is equal to $\sum_{i=0}^{9} \frac{1}{0} \cdot i = \frac{1}{0} \cdot \frac{9(10)}{2}$. We want the expected value $\mathbb{E}(X_1 + \cdots + X_n)$ (we go until X_n because the highest value in the original problem was $10^n - 1$. Using linearity of expectation, we have $\mathbb{E}(X_1 + \cdots + X_n) = \sum_{i=0}^{n} \mathbb{E}(X_i) \approx 4.5$. Questions requiring certain clever insights will likely not appear on the exam without hints; if one does appear, it will likely not be worth too many points.

§24 Thursday, November 17, 2022

Exam #2 is today.

§25 Tuesday, November 22, 2022

As Thanksgiving break is close, there will be a recorded lecture in place of class today. We will start discussing **covariance**. This will be done a bit differently than in the textbook, which starts by giving the definition of covariance and jumping into computations.

Let X and Y be jointly continuous random variables with $\mu_X = \mathbb{E}(X)$ and $\mu_Y = \mathbb{E}(Y)$, as usual. Let's now compute $\text{Var}(X + Y)$. We have

$$\text{Var}(X + Y) = \mathbb{E}\left[(X + Y - (\mu_X + \mu_Y))^2\right]$$

Using linearity of expectation, we obtain

$$\mathbb{E}\left[(X - \mu_X)^2 + (Y - \mu_Y)^2\right]$$

§26 Thursday, November 29, 2022

§26.1 Conditional expectation

Today, we'll discuss **conditional expectation**. This has many applications to **stochastic processes**, which in turn has applications to mathematical finance and **Markov decision processes**. At this point, after taking this class, we are probably good to take a class on stochastic processes that does not involve measure theory. At UMD, STAT650 (Stochastic Processes) does this.

Let's start with an example.

Example 26.1

Tos a fair coin followed by a

- tetrahedral die roll (if the result of the first toss is heads)
- six-sided die roll (if the result of the first toss is tails)

Let X represent the outcome of the coin toss toss ($X = 0$ represents tails, and $X = 1$ represents heads) and Y represent the outcome of the die roll.

We have $\mathbb{E}[Y|X = 1] = 2.5$. This is because if the first toss is heads, the tetrahedral die will be rolled, which has an expected value of 2.5. Similarly, $\mathbb{E}[Y|X = 0] = 3.5$.

Definition 26.2. Let X and Y be discrete random variables. The **conditional expectation** of X given $Y = y$ is

$$\mathbb{E}[X|Y = y] = \sum_x x f_{X|Y}(x|y)$$

where $f_{X|Y}(x|y)$ is the conditional probability mass function of X and Y.

The conditional expectation defined for continuous random variables is similar:

Definition 26.3. Let X and Y be continuous random variables. The **conditional expectation** of X given $Y = y$ is

$$\mathbb{E}[X|Y = y] = \int_{-\infty}^{\infty} x p_{X|Y}(x|y)\ dx$$

where $p_{X|Y}(x|y)$ is the conditional probability density function of X and Y.

The above conditional pmf and conditional expectations satisfy the properties of a pmf and expectation.

Theorem 26.4

Let X and Y be discrete random variables. Then,

$$p_X(x) = \sum_y p_{X|Y}(x|y) p_Y(y)$$

and

$$\mathbb{E}(X) = \sum_y \mathbb{E}[X|Y = y] p_Y(y)$$

The sums extend over those values y such that $p_Y(y) > 0$.

Remark 26.5. Rearranging Definition 19.5 gives $p_{X,Y}(x,y) = p_{X|Y}(x|y)p_Y(y)$; we can express the joint pmf in terms of the conditional and marginal pmfs. Note that if $p_Y(y) = 0$, we have that $p_{X,Y}(x,y)$ will also be 0. With this identity, we can define a joint pmf when a marginal and conditional pmf are given.

Definition 26.6. The conditional expectation of $g(X)$, given $Y = y$, is

$$\mathbb{E}[g(X)|Y = y] = \int_{-\infty}^{\infty} g(x) f_{X|Y}(x|y) dx$$

The quantities above are defined for y such that $f_Y(y) > 0$.

Here are the identities for conditional probability and conditional expectation after the Law of Total Probability is applied:

Theorem 26.7

Let X and Y be random variables. Then,

$$\mathbb{E}(X) = \sum \mathbb{E}(X|Y_i)P(Y_i)$$

where each $\{Y_i\}$ is a finite partition of the sample space.

Theorem 26.8

Let X and Y be jointly continuous. Then,

$$f_X(x) = \int_{-\infty}^{\infty} f_{X|Y}(x|y) f_Y(y)\ dy$$

for any function g for which the expectations below make sense, we have

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} \mathbb{E}[g(X)]|Y = y] f_Y(y)\ dy$$

The integrals above seem to violate the definitions of conditional probability and conditional expectation because they integrate over all $-\infty < y < \infty$, even though $f_{X|Y}(x|y)$ and $\mathbb{E}[g(X)|Y = y]$ are defined only for those y that satisfy $f_Y(y) > 0$. However, this is not a problem; if $f_Y(y) = 0$, we can regard the entire integrand as zero, and so these values of y do not actually contribute to the integral. A proof of the above theorem is provided in the textbook. Proving the first part is trivial after substituting $f_{X,Y}(x,y) = f_{X|Y}(x|y)f_Y(y)$. Proving the second part is done using a string of equalities that follow the definitions provided earlier.

Remark 26.9. We can use the identity $f_{X,Y}(x,y) = f_{X|Y}f_Y(y)$ to define the joint density function $f_{X,Y}$ from a marginal and a conditional density function.

Remark 26.10. A function of a random variable X, e.g. $g(X)$, is also a random variable. Furthermore, the conditional expectation of two random variables, e.g. $\mathbb{E}[Y|X = x]$ is also a random variable.

Theorem 26.11

If X and Y are independent, $\mathbb{E}[Y|X = x] = \mathbb{E}[Y]$.

Proof. This follows directly from $f_{Y|X}(y|x) = f_Y(y)$ for two discrete random variables, and $p_{Y|X}(y|x) = p_Y(y)$ for two continuous random variables. □

Theorem 26.12

If X and Y are random variables, $\mathbb{E}(\mathbb{E}[X|Y]) = \mathbb{E}[X]$.

Proof. This is left as an exercise. Try this on your own. □

Example 26.13

Suppose you toss a fair coin. What is greater: the expected number of tosses to obtain two heads in a row (HH), or the expected number of tosses to obtain a head and a tail in a row (HT)?

Solution. We will first tackle the first part of the question, e.g. what is the expected number of tosses to obtain two heads in a row (HH)? The key observation here is that if we see a tail on the first coin flip, it will cause the entire process to "restart." In other words, whenever a tails is flipped, any streak will be ruined and will force us to restart. Let the expected number of coin flips be k, e.g. $\mathbb{E}[X] = k$. We now have three possible classes:

1. If a tail appears on the first coin flip, we have one wasted flip and have to do k more. The probability of this event is $\frac{1}{2}$ and the total number of flips will be $k + 1$
2. If a heads appears on the first flip and a tails appears on the second flip, we will have two wasted flips and will have to do k more flips to reach our goal. Therefore, the expected number of flips will be $k + 2$, and the probability of this event is $\frac{1}{4}$
3. If we get two consecutive heads on two consecutive flips of the coin, we have $k = 2$; the probability of this is $\frac{1}{4}$

We can now use the Law of Total Expectation: from our cases above, we have $\mathbb{E}(X|T) = 1 + \mathbb{E}(X)$, $\mathbb{E}(X|HT) = 2 + \mathbb{E}(X)$, and $\mathbb{E}(X|HH) = 2$.

$$\mathbb{E}(X) = \mathbb{E}(X|T) \cdot P(T) + \mathbb{E}(X|HT) \cdot P(HT) + \mathbb{E}(X|HH) \cdot P(HH)$$

Solving gives us $\mathbb{E}(X) = 6$. We can use a similar approach for $\mathbb{E}(Y)$, where Y represents the expected number of coin flips to obtain HT, to obtain 4. □

§27 Thursday, December 1, 2022

Today, we'll continue discussing expected value. Like other parts of probability, there are some surprising results here as well. We want to look at expected value beyond just computations. First, recall that expected value is a weighted average. Now, note that expected value is usually used as a summary metric in many places. When looking at such metrics, we should ask **why** expected value is a good summary. There are two reasons that explain this: **Markov**'s inequality and **Chebyshev**'s inequalities. Furthermore, the **Weak Law of Large Numbers**, the **Strong Law of Large Numbers**, and the **Central Limit Theorem** attempt to put expected value in perspective of the experiment that is being conducted. These three are known as **convergence theorems**; in order to understand what they are saying, we need to understand convergence. We will spend today and Tuesday discussing this.

§27.1 Markov's Inequality

Theorem 27.1

Let X be a nonnegative random variable. Then, for any $c > 0$,

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

Proof. First, let's introduce some notation. Let χ_A be an **indicator function** defined as follows:

$$\chi_A = \begin{cases} 1 & x \in A \\ 0 & x \notin A \end{cases} \tag{7}$$

We have $\mathbb{E}[\chi_A] = P(A)$: suppose $A = \left[0, \frac{1}{2}\right]$ and $\mathcal{U} \sim [0, 1]$. Then, we have $P(A) = \frac{1}{2}$. Note that $\mathbb{E}[\chi_A] = \int_{-\infty}^{\infty} \chi_A f_{\mathcal{U}}(x)\ dx = \int_0^1 \chi_A f_{\mathcal{U}}(x)\ dx = P(A)$.

Now, we will proceed with the proof. Let $x \geq 0$ and $c > 0$. Take $A = \{w : \chi(w) \geq c\}$. First, note that $X \geq X \cdot X_A$. If $\chi \geq c$, we have $X\chi_A = X \cdot 1 = X$. If $X < c$, then $X\chi_A = X \cdot 0 = 0$. Thus, if $X \geq c$, we have

$$X \geq X\chi_A \geq c\chi'_A$$

Let $Y = c\chi_A$ be a random variable. Then, $X \geq Y \geq 0$. Now, we have

$$\begin{aligned}\mathbb{E}[X]\mathbb{E}[Y] &= \mathbb{E}[c\chi_A] \\ &= c\mathbb{E}(\chi_A) \\ &= cP(w \in A) \\ &= cP(X \geq c) \to P(X \geq c) \leq \frac{\mathbb{E}[X]}{c}\end{aligned}$$

Above, we used the fact that the expected value of an indicator random variable $\mathbb{E}[\chi_A]$ is equal to the probability of $P(A)$. □

Remark 27.2. Markov's inequality is a very general result: we are simply presenting a random variable X that could be either discrete or continuous, or neither. Furthermore, the inequality is essentially saying that one cannot stray too far from the mean. For example, if $\mathbb{E}[X] = 10000$ and $c = 1$, we have $P(X \geq 1) \leq \frac{100000}{1}$, which is a rather useless statement, as we already know $P(X \geq 1) \leq 1$. On the other hand, suppose $c = 10^{10}\mathbb{E}[X]$. Then, we have

$$P(X \geq 10^{10}\mathbb{E}[X]) \leq \frac{\mathbb{E}[X]}{10^{10}\mathbb{E}[X]} = \frac{1}{10^{10}}$$

This tells us that the further away from the mean we go, the less likely we are to see $X \geq 10^{10}\mathbb{E}[X]$.

§27.2 Chebyshev's Inequality

Theorem 27.3 (Chebyshev's Inequality)

Let X be a random variable with finite mean and variance. Then,

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

Proof. Since $|X - \mu| \geq 0$ and $c > 0$, the inequality $|X - \mu| \geq c$ is equivalent to the inequality $(X - \mu)^2 \geq c^2$. Now, we an apply Markov's inequality to the random variable $(X - \mu)^2$:

$$\begin{aligned}P(|X - \mu| \geq c) &= P((X - \mu)^2 \geq c^2) \\ &\leq \frac{\mathbb{E}[(X - \mu)^2]}{c^2} = \frac{\sigma^2}{c^2}\end{aligned}$$

□

Remark 27.4. Note that we did not require the mean and variance to be finite in Markov's inequality; this is important in Chebyshev's inequality.

Now, we'll start discussing the convergence theorems. They will tell us more about the role of expected value in the experiment we are conducting.

§27.3 Strong Law of Large Numbers (SLLN)

In words, the Strong Law of Large Numbers tells us that if we repeat the experiment in an identical and independent manner, the average of the observations is almost always approximately equal to $\mathbb{E}[X]$. We will discuss this more in detail with a singular experiment.

Suppose we flip a fair coin. Let $\Omega_1 = \{H, T\}$ and $X(H) = 1$ and $X(T) = 0$. Note that X is an indicator variable; $\chi_{\{X\}} = P(\{H\})$. Then, we have $\mathbb{E}[X] = \frac{1}{2}$. Consider $\Omega_2 = \Omega_1 \times \Omega_1 = \{(H,H),(H,T),(H,T),(T,T)\}$. Let $X_1(x,y) = X(x)$; in other words, X_1 will only tell us about the outcome of the first coin toss in a sequence of two tosses. Similarly, let $X_2(x,y) = X(y)$; in other words, X_2 will only tell us about the outcome of the second coin toss in a sequence of two tosses. If we want to look at three tosses, we can consider $\Omega_3 = \Omega_1 \times \Omega_1 \times \Omega_1$. Similarly, if we want to consider 4 tosses, we can consider $\Omega_3 = \Omega_1 \times \Omega_1 \times \Omega_1$. If we want to have an infinite sample space, we can multiply Ω_1 by itself infinitely many times to obtain the space Ω_∞. This space will contain all possible sequences of heads and tails.

Define

$$X_i(\omega) = \begin{cases} 1 & \text{if the } i\text{th toss is heads} \\ 0 & \text{if the } i\text{th toss is tails} \end{cases} \tag{8}$$

Note that the X_i's are independent identically distributed random variables, and $F_{X_i} = F_{X_j}$ (e.g. the cdf's of any X_i, X_j are the same). Now, we have that the average of the first n possible outcomes is

$$\overline{X_n} = \frac{\sum_{i=1}^n X_i}{n}$$

Theorem 27.5 (Informal Strong Law of Large Numbers)

For the *independent and identically distributed* (iid) random variables $X_1, X_2, \ldots X_n$, we have

$$\lim_{n\to\infty} \frac{\sum_{i=1}^n X_i(\omega)}{n} = \mathbb{E}(X) = \overline{X}_n$$

where $\overline{X}_n$ is called the **sample mean** of the sequence.

Remark 27.6. Clearly, not every possible outcome gives you $\lim_{n\to\infty} \overline{X_n}(\omega) = \mathbb{E}[X]$. If $\omega = (H, H, \ldots, H, \ldots)$, we have that $P(\omega)$ is almost 0, and that it is very likely that ω will never occur. Furthermore, the random variables $X_1, X_2, \ldots, X_n$ capture the notion of repeating a process indefinitely under identical circumstances in an independent manner.

§28 Tuesday, December 5, 2022

We have finished all of the quizzes in this class; the final homework is due on Monday, December 12.

Recall how we formed the sample space Ω_∞ when covering the Strong Law of Large Numbers. A space formed by a Cartesian product such as $\Omega_1 \times \Omega_1 \times \cdot \times \Omega_1 \times \cdots$ is called a **product space**. While we won't discuss product spaces in this class, they are used extensively in probability.

Theorem 28.1 (Strong Law of Large Numbers)

Let $X_1, \ldots, X_n, \ldots$ be independent and identically distributed random variables with $\mathbb{E}[X_i] = \mu < \infty$. The sequence

$$\overline{X_n} = \frac{\sum_{i=1}^n X_i}{n}$$

converges *almost surely* to μ.

Informally, as we discussed last lecture, the Strong Law of Large Numbers can be explained as follows: A sequence of random variables $Y_1, Y_2, \ldots, Y_n, \ldots$ of the same sample spaces converges *almost surely* to Z if removing a set $\mathcal{E}$ of exceptions the $Y_\pi(\omega) \to Z(\omega)$ for all $\omega \in \Omega - \mathcal{F}$.

The Strong Law of Large Numbers has numerous powerful applications, such as in **Monte Carlo Methods**. Suppose we want to approximate π. Let's look at the function $g(x) = \sqrt{1 - x^2}$. We know that $\int_0^1 \sqrt{1 - x^2}\, dx$ will give us $\frac{\pi}{4}$. Thus, we have

$$\pi = 4 \int_0^1 \sqrt{1 - x^2}\, dx$$

Thus, approximating π boils down to approximating $\int_0^1 \sqrt{1 - x^2}\, dx$. We can do this using the Strong Law of Large Numbers. Suppose $X \sim \mathcal{U}[0, 1]$. Using LOTUS to compute $\mathbb{E}[g(X)]$, we obtain

$$\int_{-\infty}^{\infty} g(x) f_X(x)\, dx$$

This integral is precisely $\int_0^1 g(x) f_X(x)\, dx$. We can approximate this integral b generating a collection of i numbers drawn from $\mathcal{U}[0, 1]$, computing a value of g for each of these numbers, and then calculating $\sum_{i=1}^n g(x_i)$. By the Strong Law of Large Numbers, we have that

§28.0.1 Machine Learning

The Strong Law of Large Numbers is used extensively in **machine learning** (ML); since ML is talked about quite often these days, let's talk about it a bit here. Recall the example talked about earlier this year with a probability distribution P_θ. having to decide whether an image contains a cat or not. Unfortunately, we cannot find the parameter θ too easily; this brings us to define the **loss function**. Define

$$L(x, \theta) = \begin{cases} 0 & \text{correct label} \\ 1 & \text{incorrect label} \end{cases}$$

While our model will not be perfect, we want to ensure that it'll make the least amount of mistakes as possible. Therefore, we want to find the function P_θ that minimizes $\mathbb{E}[L(x, \theta)]$. Using the Strong Law of Large Numbers, we have that

$$\mathbb{E}[L(x, \theta)] \approx \sum_{i=1}^{n} L(x_i, \theta)$$

While we will only have an approximation of $\mathbb{E}[L(x, \theta)]$, it will be close enough and will allow us to obtain the best P_θ. The main takeaway from this brief interlude is that the Strong Law of Large Numbers has many useful and practical applications.

§28.1 Convergence of Sequences

Now, we will discuss the pure math aspect of the Strong Law of Large Numbers. Recall the notion of **convergence** from calculus. Suppose we have a sequence a_n. If the sequence convergences to l, it can have any type of behavior until it gets close to the nth term in the sequence, where the terms will start to cluster in or around l. This is a very intuitive definition of convergence, but covers what we want, which is **pointwise convergence**.

Suppose we have a sequence of functions $f_1, f_2, \ldots, f_n, \ldots$ that are all defined on $[0, 1]$.

Definition 28.2. A function f_n converges to f **pointwise** if for every x in the domain of f_n, $f_n(x) \to f(x)$.

Let's cover an example. Let $f_n(x) : [0, 1] \to \mathbb{R} = x^n$. At 0, we have $f_1(0) = 0$, $f_2(0) = 2$, and so on. Each $f_n(0)$ will be equal to 0. At $x = \frac{1}{2}$, we have $f_1\left(\frac{1}{2}\right) = \frac{1}{2}$, $f_2\left(\frac{1}{2}\right) = \frac{1}{4}$ and so on. We have $\lim_{n\to\infty} f_n\left(\frac{1}{2}\right) = 0$. Finally, let's look at $x = 1$. As each $f_n(1) = 1$, we have $\lim_{n\to\infty} f_n(1) = 1$. Thus, we can say f converges pointwise to $f : [0, 1] \to \mathbb{R}$ if

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

Definition 28.3. **Almost sure convergence** is convergence outside a set of measure 0 (e.g. a set that has probability 0).

In our above example, we have f_n converges almost surely to $g : [0, 1] \to \mathbb{R}$ where $g(x) = 0$.

Definition 28.4. A sequence of random variables $X_1, \ldots, X_n$ **converges in probability** to a random variable X, written as $X_n \xrightarrow{p} X$, if

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

for all $\epsilon > 0$.

Remark 28.5. Convergence in probability is an important concept. In statistics, it is used to capture the notion that larger samples give better approximations. Convergence in probability is, especially at a first glance, a bit harder to understand than almost sure convergence. However, it is a useful concept that shows up in many real-life scenarios. Furthermore, almost all of the interesting theorems under almost sure convergence hold in convergence probability.

§28.2 Moving-Block Problem

This is the second problem on the homework due next Monday. It is also known as the Typewrite problem. We covered this in class; try to tackle it with convergence.

§29 Thursday, December 8, 2022

Today is the last lecture for this class. We will finish discussing convergence, as well as the Strong/Weak Law of Large Numbers and the Central Limit Theorem.

Definition 29.1. Let $X_1, \ldots, X_n$ be a sequence of random variables on the same sample space. We say $\{X_n\}_{n \in N}$ converges almost surely to Y if $P(\lim_{n \to \infty} X_n = Y) = 1$, or equivalently $P(X_n \neq Y) = 0$.

Definition 29.2. We say $\{X_n\}_{n \in N}$ converges **in probability** to Y (e.g. converges in measure) if for every epsilon greater than 0,

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

Fact 29.3. If X_n converges almost surely to Y, then X_n converges in probability to Y.

§29.1 Weak Law of Large Numbers

Theorem 29.4 (Weak Law of Large Numbers)
Let $X_1, \ldots, X_n, X_{n+1}$ be independent identically distributed random variables with $\mathbb{E}[X_1] = \mu < \infty$. Then,

$$\lim_{n \to \infty} P(|\overline{X} - \mu| \leq \epsilon) = 0$$

Proof. Assume $\text{Var}(X_1) = \sigma^2 < \infty$ (e.g. the variance of X_1 is bounded and finite). Now, we can use Chebyshev's Inequality as follows:

$$P(|\overline{X} - \mu| \geq \epsilon) \leq \frac{\text{Var}(\overline{X})}{\epsilon^2} = \frac{\text{Var}(X)}{n\epsilon^2}$$

which goes to 0 as $n \to \infty$. □

§29.2 Central Limit Theorem

Now, we'll discuss the **Central Limit Theorem**, which is the cornerstone of classical statistics. First, we'll discuss **convergence in distribution**.

Definition 29.5. Let $X_1, \ldots, X_n$ be a sequence of random variables, possibly not defined on the same sample space. X_n converges to Y **in distribution**, often denoted as $X_n \xrightarrow{D} Y$, if for any $a \in \mathbb{R}$,

$$\lim_{n \to \infty} P(X_n \leq a) = P(Y \leq a)$$

For "large n," we have $P(X_n \leq a) \approx P(Y \leq a)$.

Remark 29.6. While the above definition doesn't explicitly require each random variable to be on the same sample space, this is usually not what we will be working with.

Now, we can introduce the Central Limit Theorem:

Theorem 29.7 (Central Limit Theorem)
Let $X_1, \ldots, X_n, \ldots$ be independent identically distributed random variables with $\mathbb{E}[X_1] = \mu$ and $\text{Var}(X_1) = \sigma^2$. Then, for any $a, b \in \mathbb{R}$ with $a < b \leq \infty$,

$$\lim_{n \to \infty} P\left(a \leq \frac{S_n - n\mu}{\sqrt{n}\sigma} \leq b\right) = \Phi(b) - \Phi(a)$$

where Φ (as usual) refers to the cdf of the normal distribution. Note that we will have $P(X \leq \infty) = 1$.

Remark 29.8. Essentially, the Central Limit Theorem states that, certain conditions, the sum of a large number of random variables is *approximately normal*.

Let's now look at an example using the CLT:

Example 29.9

Suppose 1000 fair die are rolled and the values of each roll are added. Estimate the probability that the sum is at least 3600.

Solution. Let X_i denote a random variable that gives you the outcome of the ith roll. $X_1, \ldots, X_{1000}$ are independent identically distributed random variables. We have $\mathbb{E}[X_1] = 3.5$ and $\text{Var}(X_1) = \frac{35}{12}$. Now, we apply the Central Limit Theorem. Let $S_{1000} = \sum_{i=1}^{1000} X_i$; by the Central Limit Theorem, $P(S_{1000} \geq 3600) = P\left(\frac{S_{1000} - \mathbb{E}[S_{1000}]}{\sqrt{\text{Var}(S_{1000})}}\right) \geq \frac{3600 - \mathbb{E}[S_{1000}]}{\sqrt{\text{Var}(S_{1000})}}$. This probability, in turn, is equivalent to

$$1 - P\left(\frac{S_{1000} - \mathbb{E}[S_{1000}]}{\sqrt{\text{Var}(S_{1000})}}\right) \leq \frac{3600 - 1000 \cdot 3.5}{\sqrt{\text{Var}(S_{1000})}}$$

Note that we plugged in $1000 \cdot 3.5 = 3500$ for $\mathbb{E}[S_{1000}]$ and $\sqrt{\frac{35000}{12}}$ for $\sqrt{\text{Var}(S_{1000})}$. □

§30 Final Exam Practice Problem Solutions

The following are my solutions to the practice problems for topics covered after Exam # 2. They are for my reference.

1. This one was quite long (so I won't type it all out), but relied only on the definitions of the conditional pmf, conditional expectation, and some integration by parts techniques, which are worth reviewing. $\mathbb{E}(X|Y)$ and $\mathbb{E}(Y|X)$ did indeed exist.

2.

3. We can let X be a random variable mapping a sequence of tosses achieving HHH to the number of tosses needed to achieve HHH and let $\mathbb{E}[X] = k$. We now have the following cases:

   - If the first toss is a tails (T), the sequence resets and there are now an average of $k + 1$ more tosses needed; this occurs with probability $\frac{1}{2}$.
   - If the first 2 tosses are HT, the sequence resets and there are now an average of $k + 2$ more tosses needed; this occurs with probability $\frac{1}{4}$.
   - If the first three tosses are HHT, the sequence resets and there are now an average of $k + 3$ more tosses required; this occurs with probability $\frac{1}{8}$.
   - If the first three tosses are HHH, the sequence does not reset and 3 tosses are required; this occurs with probability $\frac{1}{8}$.

Now, we can create the equation

$$\begin{aligned}\mathbb{E}[X] = k &= \frac{1}{2}(k+1) + \frac{1}{4}(k+2) + \frac{1}{8}(k+3) + \frac{1}{8}(3) \\ \rightarrow \frac{k}{8} &= \frac{14}{8} \rightarrow \boxed{k = 14}\end{aligned}$$

Thus, there are an average of $\boxed{14}$ tosses required to see three consecutive heads.

4.