1 Lecture 1

1.1 Administrative

1. Very useful course (Finance, Science, Gambling, Elections, Cryptography, Life Choices, etc.)
2. One of the top Probability departments in the world
3. Many unintuitive results in probability
4. Prerequisite: Calc 1-3
5. Textbook: Sheldon Ross
6. Syllabus
7. Will try to mix in statistics as motivating examples. Statistics is the inverse of probability.

1.2 Introduction

Mathematical methods for studying uncertainty arose hundreds of years ago. The simplest model involves phenomena where every outcome is equally likely (such as throwing dice, or flipping coins). As with any model, it is not an exact copy of the real world, but is designed to replicate some aspect. Modeling differences arise in the different schools of thought in probability/statistics (Frequentists, Bayesians). This is not an indication that the mathematics is incorrect, or illogical, but that there is freedom/disagreement in the modeling procedure. Frequentists propose that probabilities measure the long-term average occurrence rates when an event is repeated. Bayesians propose that probabilities represent a belief system about the chance of an event occurring, and repetition is not required. Later it will become more clear where they differ.

As probability matured, mathematicians started looking at experiments with infinitely many outcomes. For example, we could imagine an infinite sequence of coin tosses, or looking at all the points a dart could hit on a solid square. These lead to the situation where all outcomes have probability zero, but there are events with positive probability. It wasn’t until Kolomogrov (building on the work of Lebesgue) that we had a rigorous mathematical framework for handling such experiments, and this framework is still used today.
1.3 Math Prerequisites

1.3.1 Sets

Our model of probability will be based on set theory, and thus we begin with a review of the necessary concepts.

1.3.2 Set Review Exercises

1. Let $A = \{2, 3, 5\}$, $B = \{1, 3\}$, and suppose we are working in the universe $S = \{1, 2, \ldots, 8\}$.
   
   (a) What is $A \cup B$? What is $A \cap B$ (written $AB$ in our text)?
   
   (b) What is $B \setminus A$? What is $A \setminus B$?
   
   (c) What is $A^c$ (read “$A$ complement”)?
   
   (d) What is $A \cap B^c$? What is $(A \cap B^c) \cup (A \cap B)$?
   
   (e) What is $A \times B$?
   
   (f) Is $A \subset B$? Is $B \subset A$? Is $A \subset A$?

2. If $|A| = 5$ and $|B| = 10$ what can we say about $|A \cup B|$?

3. If $A_1, \ldots, A_n \subset S$, for some universe $S$, give an expression for $(\bigcup_{k=1}^{n} A_k)^c$ in terms of an intersection.

4. What is $\bigcup_{n=1}^{\infty} (3, n)$? Here $(3, n)$ denotes the open interval $\{x \in \mathbb{R} : 3 < x < n\}$.

5. What is $\bigcap_{n=1}^{\infty} (-1/n, 1/n)$?

6. What does it mean for $A_1, \ldots, A_n$ to form a partition of $S$?

1.3.3 Solutions

1. (a) $A \cup B = \{1, 2, 3, 5\}$ and $A \cap B = AB = \{3\}$.
   
   (b) $B \setminus A = \{1\}$ and $A \setminus B = \{2, 5\}$.
   
   (c) $A^c = S \setminus A = \{1, 4, 6, 7, 8\}$.
   
   (d) $A \cap B^c = A \setminus B = \{2, 5\}$ and $(A \cap B^c) \cup (A \cap B) = A$.
   
   (e) $\{(2, 1), (2, 3), (3, 1), (3, 3), (5, 1), (5, 3)\}$
   
   (f) No, No, Yes.

2. $10 \leq |A \cup B| \leq 15$ with all values possible. We get 10 if $A \subset B$ and 15 if $A \cap B = \emptyset$.

3. $(\bigcup_{k=1}^{n} A_k)^c = \bigcap_{k=1}^{n} A_k^c$. 

2
4. This union is the collection of all real numbers in at least one of the intervals. The result is thus \((3, \infty)\).

5. This intersection is the collection of all real numbers in all of the intervals. The result is thus \(\{0\}\).

6. It means two properties hold:
   
   (a) Pairwise disjointness: \(A_i \cap A_j = \emptyset\) for \(i \neq j\),
   
   (b) Cover: \(\bigcup_{i=1}^{n} A_i = S\).

We will often use partitions along with the distributive property:

\[C \cap \bigcup_{i=1}^{n} A_i = \bigcup_{i=1}^{n} C \cap A_i.\]

In general intersections distributed over unions and unions distribute over intersections:

\[(E \cup F) \cap G = (E \cap G) \cup (F \cap G)\quad \text{and}\quad (E \cap F) \cup G = (E \cup G) \cap (F \cup G).\]

1.3.4 Series

We will also need series/sums, as they frequently occur when computing probabilities. Here are some well known formulas:

1. \(\sum_{k=1}^{n} k = \frac{n(n+1)}{2}\) and \(\sum_{k=1}^{n} k^2 = \frac{n(n+1)(2n+1)}{6}\).

2. \(\sum_{k=0}^{n} r^k = \frac{1-r^{n+1}}{1-r}\) for \(r \neq 1\), and \(n + 1\) if \(r = 1\). Taking a limit gives

3. \(\sum_{k=0}^{\infty} r^k = \frac{1}{1-r}\) if \(|r| < 1\).

4. \(e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!}\) for all \(x \in \mathbb{R}\).

5. The harmonic series \(\sum_{n=1}^{\infty} \frac{1}{n}\) diverges. The \(p\)-series \(\sum_{n=1}^{\infty} \frac{1}{n^p}\) converges for \(p > 1\) and diverges for \(p \leq 1\) (can prove using integral test or Cauchy condensation test).

1.3.5 Series Exercises

Most of the series manipulations we have to make will involve modifying/pattern-matching with well known series.

1. Give a series equal to the expression \(\frac{1}{1-x^2}\). Where does it converge?

2. Give a (non-series) expression for \(\sum_{n=0}^{\infty} (-1)^n x^{3n}\) where \(|x| < 1\).

3. Give a (non-series) expression for \(\sum_{k=0}^{\infty} \frac{e^{t \lambda} - \lambda^k}{k!}\) where \(\lambda > 0\) and \(t \in \mathbb{R}\).

4. (*) Compute a non-series expression for \(\sum_{k=0}^{\infty} k x^k\) when \(|x| < 1\).
1.3.6 Solutions

1. \[ \sum_{n=0}^{\infty} (4x^2)^n = \sum_{n=0}^{\infty} 4^n x^{2n} \] and it converges for \(|4x^2| < 1\) or equivalently, when \(|x| < 1/2\).

2. \[ \frac{1}{1+x^2} \] since \((-1)^n = (-1)^{3n}.

3. As \(e^{tk\lambda} = (e^{t\lambda})^k\) we have
\[ e^{-\lambda} \sum_{k=0}^{\infty} \frac{(e^{t\lambda})^k}{k!} = e^{-\lambda} e^{e^{t\lambda}} = e^{\lambda(e^{t\lambda} - 1)}. \]

4. We can differentiate \(\sum_{k=0}^{\infty} x^k\) term-by-term (legal for convergent power series) to get
\[ \sum_{k=0}^{\infty} k x^{k-1} = \frac{d}{dx} \left( \frac{1}{1-x} \right) = \frac{1}{(1-x)^2}. \]
Thus the result is \(\frac{x}{(1-x)^2}\).

1.4 Counting

1.4.1 Multiplication, Permutations and Combinations

In situations where every outcome is equally likely computing a probability amounts to counting. As such we now go over various counting principles and formulas.

**Theorem 1** (Addition Principle). If \(A, B\) are sets with \(A \cap B = \emptyset\) then \(|A \cup B| = |A| + |B|\).

**Theorem 2** (Multiplication Principle). Each distinct element of a set is determined by first making one of \(n_1\) choices. For each of those choices, you then must make one of \(n_2\) choices, and so forth until you have made your \(k\)th choice. Then the set contains \(n_1 \cdot n_2 \cdots n_k\) elements.

**Example 3.** If there are 3 choices for appetizer, 4 choices for entree, and 2 choices for dessert, then there are 24 possible meals. Can draw a tree diagram.

This simple counting procedure can be used to count more complex objects.

Fix a set \(S\) of size \(n\).

1. The number of ordered lists of size \(k\) chosen from \(S\) with repeats allowed is \(n^k\).

2. The number of ordered lists of size \(k\) chosen from \(S\) with no repeats allowed is \(n(n-1)\cdots(n-k+1)\).

3. Number of subsets of \(S\) of size \(k\) is given by \(\binom{n}{k} = \frac{n^k}{k!(n-k)!} = \frac{n(n-1)\cdots(n-k+1)}{k!}\). Note that the second formula above allows us to evaluate \(\binom{n}{k}\) for any \(k \in \mathbb{Z}_{\geq 0}\) and \(n \in \mathbb{R}\). The special case of \(k = 0\) is always 1: \(\binom{n}{0} = 1\) for all \(n \in \mathbb{R}\) (this makes sense from a counting perspective when \(n \in \mathbb{Z}_{\geq 0}\) and we just define it to be 1 for general \(n \in \mathbb{R}\).)
1.4.2 Counting Exercises

1. How many sequences of 5 upper case English letters can be made if you are not allowed to reuse? If you are allowed to reuse?

2. How many ordered sequences with eight 0’s, and thirteen 1’s are possible? What about eight 0’s, thirteen 1’s and six 2’s?

3. Given a set \( S \) of \( n \) people, how many ways are there to choose a team of \( k \) people, and elect one of those \( k \) to be leader?

4. Given a group of 13 women and 10 men, how many ways are there to form a team of 5 women and 4 men?

5. (⋆) Suppose you have \( n \) identical pieces of candy. How many ways are there to distribute the candy to \( k \) (distinguishable) children? What if each piece of candy was distinguishable?

1.4.3 Solutions

1. No reuse: \( 26 \cdot 25 \cdot 24 \cdot 23 \cdot 22 = \binom{26}{5} \cdot 5! \) and with reuse is \( 26^5 \).

2. \( \frac{21!}{8!13!} = \binom{21}{8} = \binom{21}{13}, \frac{27!}{8!13!6!} = \binom{27}{8,13,6} \).

3. \( \binom{n}{k} \cdot k \)

4. \( \binom{13}{5} \binom{10}{4} \)

5. \( \binom{n+k-1}{k-1} \). The is the same as the number of solutions to the equation \( a_1 + \cdots + a_k = n \) where \( a_1, \ldots, a_k \) are non-negative integers.
2 Lecture 2

2.1 More Counting and Probability

2.1.1 Review Problems

1. How many different strings of length 12 can be made from the letters in “AABBBCCCDDE”?

2. A student must answer 7 out of 10 questions on an exam.
   (a) How many ways are there to take the exam?
   (b) If the student also must answer at least 3 out of the first 5 questions, how many ways are there?
   (c) Which is larger?

3. Ten people are split into 2 teams of 5 that will compete. How many possible match-ups are there?

4. A standard deck of 52 playing cards has 12 picture cards. How many ways are there to deal the cards to 4 players (each getting 13 cards) so that each player receives exactly 3 picture cards. Assume the order in which a given player receives the cards doesn’t matter.

2.1.2 Review Solutions

1. \( \frac{12!}{2!3!3!3!3!1!} = (\binom{12}{2,3,3,3,1}) \)

2. (a) \( \binom{10}{7} \)
   (b) \( \binom{5}{3} \binom{5}{4} + \binom{5}{3} \binom{5}{3} + \binom{5}{3} \binom{5}{2} \)
   (c) The first is bigger (less restrictive).

3. \( \binom{10}{5} / 2 \) (each pair of teams is counted twice)

4. \( \binom{12}{3,3,3,3} \binom{40}{10,10,10,10} \)

2.1.3 Multinomial Theorem

Consider the expressions \((x+y)^n\) and \((x+y+z)^n\). By looking at \((x+y)^2\) and \((x+y)^3\) we can see that they are sums of all possible strings of length \(n\) using \(x, y, \) and \(x, y, z\) respectively. Using the counting method above we can group the strings by their counts of \(x\)’s, \(y\)’s and \(z\)’s to obtain
**Theorem 4** (Multinomial Theorem). For any \(x, y, z \in \mathbb{R}\), and \(n \geq 0\) we have

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

and

\[
(x + y + z)^n = \sum_{0 \leq i, j, k \leq n \\ i+j+k=n} \binom{n}{i,j,k} x^i y^j z^k
\]

where

\[
\binom{n}{i,j,k} = \frac{n!}{i!j!k!}.
\]

This can be extended to any number of variables.

### 2.1.4 Counting Proofs

One type of proof technique the occurs frequently in combinatorics is counting the same quantity using two different methods. Here is an example.

**Theorem 5.**

\[
\binom{n}{k} = \frac{n(n-1) \cdots (n-k+1)}{k!}
\]

**Proof.** Consider all ordered sequences of length \(k\) taken from a set of size \(n\). By the multiplication principle there are \(n(n-1) \cdots (n-k+1)\) of these. Counted differently, first select one of the \(\binom{n}{k}\) subsets of size \(k\) to use. Then choose one of the \(k!\) orderings for these \(k\) elements. \(\square\)

**Theorem 6.**

\[
\frac{(2n)!}{2^n n!} = (2n - 1)(2n - 3) \cdots 1.
\]

**Proof.** Suppose we want to pair up \(2n\) people to form \(n\) teams of two. Give each person a number from \(1, \ldots, 2n\). If at each step we choose the lowest numbered unpaired player to pair up, we get the formula on the right. Now that we know the number of ways to pair up people, we can count the number of ways to order \(2n\) people. First choose the \(n\) pairs. Then choose one of the \(n!\) orderings of the pairs. Then choose one of the 2 orderings within each pair. This shows

\[
(2n)! = (2n - 1)(2n - 3) \cdots 1 \cdot n! \cdot 2^n.
\]

\(\square\)
2.1.5 Counting Proof Exercises

1. What is \((3x + 4y)^3\)?

2. Prove that \(\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k}\) using a counting proof.

3. Prove that \(\binom{n+m}{k} = \sum_{j=0}^{k} \binom{n}{j} \binom{m}{j}\) using a counting proof.

4. What is \(\binom{n}{0} - \binom{n}{1} + \binom{n}{2} - \binom{n}{3} + \cdots + (-1)^n \binom{n}{n}\)?

2.1.6 Solutions

1.
\[
\binom{3}{0}(3x)^3 + \binom{3}{1}(3x)^2(4y) + \binom{3}{2}(3x)(4y)^2 + \binom{3}{3}(4y)^3.
\]

2. Consider the set \(\{1, \ldots, n\}\). If we want to choose a \(k\) element subset we can either include \(n\) or not. If we include \(n\), we must then choose \(k-1\) elements from the remaining \(n-1\). If we don’t include \(n\) we must choose \(k\) elements from the remaining \(n-1\).

3. Consider \(n\) red numbers and \(m\) black numbers. Suppose we want to choose \(k\) total numbers. Then we can sum over the total quantity of red numbers we will choose.

4. This is \((1 + (-1))^n = 0\).

2.2 Axioms of Probability

Suppose we are modeling a random world or experiment. Let \(S\) (called the sample space) denote the set of all possible outcomes of our experiment, or possible states of our world. Subsets of \(S\) are called events. In a graduate course, we would get very precise on which subsets of \(S\) are allowed to be events. In this course, we won’t worry about this point. Define a function \(P\) (called a probability measure) whose domain is the set of events (remember events are subsets of \(S\)) and whose codomain is a real number such that

1. \(0 \leq P(E) \leq 1\),

2. \(P(S) = 1\),

3. (Countable additivity)
\[
P\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} P(E_i),
\]
where \(E_i\) are pairwise disjoint. That is, \(E_i \cap E_j = \emptyset\) for \(i \neq j\).
A few notes on the axioms:

1. This is it. All of the probability theory in this course will come from the 3 rules above.

2. The probability function (measure) $P$ takes subsets of $S$ as input, not elements of $S$.

3. Sometimes people include $P(\emptyset) = 0$ in the axioms, but we can just derive it. Let $E_1 = S$ and $E_k = \emptyset$ for $k \geq 2$. Then by countable additivity

$$1 = P(S) = P \left( \bigcup_{k=1}^{\infty} E_k \right) = P(E_1) + \sum_{k=2}^{\infty} P(E_k) = 1 + \sum_{k=2}^{\infty} P(\emptyset).$$

Thus we must have $P(\emptyset) = 0$.

We can derive various properties of probability from these axioms. For example:

**Theorem 7.** If $A, B$ are events with $A \subset B$ then $P(A) \leq P(B)$.

**Proof.** We can write $B = A \cup (B \cap A^c)$ so that

$$P(B) = P(A) + P(B \cap A^c) \geq P(A).$$

The above is valid, assuming we have finite additivity (which isn’t one of the axioms). To complete the proof we will establish finite additivity in the exercises below.

### 2.2.1 Axiom Exercises

1. For any $A \subset S$, derive an expression for $P(A^c)$ in terms of $P(A)$.

2. (Finite Additivity) For any $n$ events $A_1, \ldots, A_n \subset S$ with $A_i \cap A_j = \emptyset$ for $i \neq j$ prove that

$$P \left( \bigcup_{i=1}^{n} A_i \right) = \sum_{i=1}^{n} P(A_i).$$

3. For any events $A, B$ prove $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

4. (**) (Countable Subadditivity, or Boole’s Inequality) For any events $A_1, \ldots$ (not necessarily disjoint) we have

$$P \left( \bigcup_{i=1}^{\infty} A_i \right) \leq \sum_{i=1}^{\infty} P(A_i).$$
2.2.2 Solutions

1. Note that $P(A^c) + P(A) = P(S) = 1$ so $P(A^c) = 1 - P(A)$.

2. Define $A_{n+1} = \emptyset, A_{n+2} = \emptyset, \ldots$ and apply countable additivity.

3. Note that $P(A \cup B) = P(A) + P(B \cap A^c)$ (can help to look at a Venn diagram). Then note that $P(B \cap A^c) + P(B \cap A) = P(B)$.

4. Let $B_1 = A_1$, $B_2 = A_2 \cap A_1^c$, $B_3 = A_3 \cap A_2^c \cap A_1^c$, etc. In general $B_n = A_n \cap \bigcap_{k=1}^{n-1} A_k^c$. Then the $B_i$ are disjoint sets with the same union. Hence

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = P\left(\bigcup_{i=1}^{\infty} B_i\right) = \sum_{i=1}^{\infty} P(B_i) \leq \sum_{i=1}^{\infty} P(A_i).$$

2.2.3 Inclusion-Exclusion for Probabilities

We have already seen that

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

Using induction, you can extend this property $n$ sets.

$$P\left(\bigcup_{k=1}^{n} A_k\right) = \sum_{k=1}^{n} P(A_k) - \sum_{i<j} P(A_i \cap A_j) + \sum_{i<j<k} P(A_i \cap A_j \cap A_k) + \cdots + (-1)^{n-1} P(A_1 \cap A_2 \cap \cdots \cap A_n).$$

You can easily see this formula at work for $n = 3$ using a Venn diagram. The idea of the general proof is to write

$$\bigcup_{k=1}^{n+1} A_k = A_{n+1} \cup \bigcup_{k=1}^{n} A_k$$

and apply the case for 2 sets followed by the case for $n$ sets. It can also be shown that each time we subtract or add (before the last) we are overcorrecting. More precisely,

$$P\left(\bigcup_{k=1}^{n} A_k\right) \leq \sum_{k=1}^{n} P(A_k)$$

$$P\left(\bigcup_{k=1}^{n} A_k\right) \geq \sum_{k=1}^{n} P(A_k) - \sum_{i<j} P(A_i \cap A_j)$$

$$P\left(\bigcup_{k=1}^{n} A_k\right) \leq \sum_{k=1}^{n} P(A_k) - \sum_{i<j} P(A_i \cap A_j) + \sum_{i<j<k} P(A_i \cap A_j \cap A_k)$$

:
2.2.4 Examples of Probabilities

Example 8 (Equally Likely Outcomes). Let $S$ be finite, and define $P(A)$ for $A \subseteq S$ to be

$$P(A) = \frac{|A|}{|S|}.$$ 

We note that $0 \leq P(A) \leq 1$, $P(S) = 1$, and if $A_1, A_2, \ldots$ are disjoint then

$$P\left( \bigcup_{i=1}^{\infty} A_i \right) = P\left( \bigcup_{i=1}^{k} A_i \right) \quad (S \text{ is finite})$$

$$= \frac{|\bigcup_{i=1}^{k} A_i|}{|S|} \quad \text{(Defn of } P)$$

$$= \sum_{i=1}^{k} \frac{|A_i|}{|S|} \quad \text{(Addition Principle)}$$

$$= \sum_{i=1}^{k} P(A_i) \quad \text{(Defn of } P).$$

Above we use the fact that since $S$ is finite, there could not be infinitely many disjoint nonempty sets in our union. Thus there must be a $k$ so that $A_j = \emptyset$ for all $j > k$.

Example 9 (Arbitrary Coin Flip). Fix $p \in [0,1]$, let $S = \{H, T\}$, and define

$$P(\{H\}) = p, \quad P(\{T\}) = 1 - p, \quad P(\emptyset) = 0, \quad P(S) = 1.$$ 

Then all properties hold. We will often use the letter $q$ to denote $1 - p$.

Example 10 (General Finite Space). Let $S$ be a finite set with distinct elements $s_1, \ldots, s_n$. Define, for $i = 1, \ldots, n$,

$$P(\{s_i\}) = p_i$$

where each $p_i \in [0,1]$, and $\sum_{i=1}^{n} p_i = 1$. For more general events $A \subseteq S$, define

$$P(A) = \sum_{s_i \in A} P(\{s_i\}) = \sum_{s_i \in A} p_i.$$ 

This can be shown to have all the required properties.

This example includes both previous examples.

Example 11 (General Countably Infinite Space). Let $S$ be an infinite set with distinct elements $s_1, s_2, \ldots$. Define, for $i = 1, 2, \ldots$,

$$P(\{s_i\}) = p_i$$

where each $p_i \in [0,1]$, and $\sum_{i=1}^{\infty} p_i = 1$. For more general events $A \subseteq S$, define

$$P(A) = \sum_{s_i \in A} P(\{s_i\}) = \sum_{s_i \in A} p_i.$$ 

This can be shown to have all the required properties.
3 Lecture 3

3.1 Finishing Basic Probability

3.1.1 Review Exercises

1. Model flipping two fair coins using a sample space and a probability measure. Compute the probability of getting at least 1 head.

2. You are flipping a coin twice that has \( .75 \) chance of landing heads. What is the probability of getting 1 head?

3. Let \( S = \{1, 2, \ldots, 100\} \) with all outcomes equally likely. What is the probability of the event that a chosen integer isn’t divisible by 2, 3, or 5?

4. We want to distribute 10 identical pieces of candy to 4 (distinguishable) children. For each piece of candy we roll a 4-sided die and give it to the associated child. What is the probability that the children receive 2, 5, 1, 2 pieces of candy, respectively (i.e., first child receives 2, second receives 5, etc.)?

5. Suppose you are dealt an ordered sequence of \( k \) cards from a 52 card deck without replacement. Compute the probability of getting a particular \( k \) card (unordered) hand.

3.1.2 Solutions

1. Let \( S = \{(a_1, a_2) : a_i \text{ is } H \text{ or } T\} \) with all outcomes equally likely. This intuitive choice for the probability measure will later be seen to come from the fact that the coin flips are modeled as “independent” of each other. Then we have

\[
P(\text{at least one head}) = P(\{HT, TH, HH\}) = \frac{3}{4},
\]

where we wrote \((H,T)\) as \(HT\) for brevity.

2. Let \( S = \{(a_1, a_2) : a_i \in \{H, T\}\} \). We define \( P \) by

\[
P(\{(a_1, a_2)\}) = .75^{\text{# of heads}} \cdot .25^{\text{# of tails}}
\]

using the general finite space. As above, we will see later that this intuitive choice of measure comes from modeling the coin flips as independent. Note that

\[
.75^2 + .25^2 + 2(.75)(.25) = 1.
\]

The result is \(2(.75)(.25) = \frac{6}{16}.\)
3. Let $D_2, D_3, D_5$ denote the subsets of $S$ divisible by 2, 3, 5 respectively. Then

$$|D_2| = 50, |D_3| = 33, |D_5| = 20, |D_2D_3| = 16, |D_2D_5| = 10, |D_3D_5| = 6, |D_2D_3D_5| = 3.$$  

This gives 

$$|D_2 \cup D_3 \cup D_5| = 50 + 33 + 20 - 16 - 10 - 6 + 3 = 74.$$  

The answer is $1 - \frac{74}{100} = \frac{26}{100}.$

4. Here $S = \{(d_1, \ldots, d_{10}) : d_i \in \{1, 2, 3, 4\}\}$ and all outcomes are equally likely. The probability is thus 

$$\frac{10}{2,5,1,2}$$

since $|S| = 4^{10}$ and there are 

$$\binom{10}{2,5,1,2} = \frac{10!}{2!5!2!}$$

elements of $S$ that give the correct counts.

5. Each ordered hand occurs with an equal probability of $\frac{k!}{52 \cdot 51 \cdots (52 - k + 1)}$. Thus an unordered hand occurs with probability

$$\frac{k!}{52!} \cdot \frac{1}{52 \cdot 51 \cdots (52 - k + 1)} = \frac{1}{\binom{52}{k}}$$

as expected.

3.1.3 Aside on Measure Theory

Example 12 (Infinite Coin Flipping). Suppose we have a fair coin and wish to flip it infinitely many times. We could use the following sample space:

$$S = \{(a_1, a_2, \ldots) : a_i \in \{H, T\}\}.$$  

If the model accurately reflects coin flipping, we would want the following calculations to hold:

1. The probability the first flip is heads should be $1/2$.
2. The probability the first two flips are heads should be $1/4$.
3. In general, if we specify the first $k$ flips, we expect the probability to be $1/2^k$.  

13
This leads to the uncomfortable situation that the probability of any particular sequence is 0 since it must be less than \( \frac{1}{2^k} \) for all \( k \) by monotonicity. Thus any measure we choose must assign positive probability to some events, while assigning zero probability to every singleton event. Formulaically, we must have

\[
P(\{s\}) = 0,
\]

for any \( s \in S \). One key aspect to why this is possible is that \( S \) is an uncountable set, so it can have probability 1 even though all the singletons have probability zero. This is one of the nuances of countable additivity.

**Example 13** (Aside: Lebesgue Measure). Let \( S = [0, 1] \) and define \( P((a, b)) = b - a \). For more general events \( A \) we define

\[
P(A) = \inf \left\{ \sum_{i=1}^{\infty} (b_i - a_i) : \bigcup_{i=1}^{\infty} (a_i, b_i) \supset A \right\}.
\]

That is, we approximate general sets by looking at coverings by intervals. The infimum forces us to look at tighter and tighter coverings. In addition to allowing us to deal with sets of real numbers, we can use this model to study sequences of fair coin flips by treating each sequence as a binary expansion. One odd consequence of this theory is that not every subset of \([0, 1]\) can be called an event if we want countable additivity to hold. As such, we restrict to a special class of events called the measurable sets.

### 3.1.4 A Few More Interesting Examples

1. Assuming you are randomly dealt 5 cards from a standard deck of playing cards, what is the probability of getting two pair (a poker hand)? What about a full house?

2. How many students are needed in a class before it is likely that at least 2 have the same birthday (assume 365 days in a year, each day equally likely)?

3. Suppose \( n \) people put their ID cards down, the IDs are shuffled, and then randomly returned. What is the probability nobody gets their ID back? What if \( n \) is very large?

### 3.1.5 Solutions

1. Consider the sample space of all 5 card subsets of the 52 cards, where each is equally likely. To count the number of possible ways to get two pair, we first choose the two values for the pairs, and the value of the remaining card. Then we choose the suits for each pair and the extra card.

\[
\frac{\binom{13}{2} \cdot 11 \cdot \binom{4}{2} \cdot \binom{4}{1}}{\binom{52}{5}} \approx 0.0475.
\]
The probability of a full house is

\[
\frac{13 \cdot 12 \cdot \binom{4}{3} \binom{4}{2}}{\binom{52}{5}} \approx 0.0014.
\]

We could also count these using ordered hands. For two pair we have

\[
\frac{52 \cdot 3 \cdot 48 \cdot 3 \cdot 44 \cdot \frac{5!}{2!2!2!}}{52 \cdot 51 \cdot 50 \cdot 49 \cdot 48}.
\]

Here we count all possible ways of choosing hands in the order 11223, where the numbers denote values. Then we must consider all orderings of this string, but divide by an extra 2! since swapping the 1’s and 2’s doesn’t yield a distinct hand. To me this calculation seems trickier. For full house we have

\[
\frac{52 \cdot 3 \cdot 2 \cdot 48 \cdot 3 \cdot \frac{5!}{3!2!}}{52 \cdot 51 \cdot 50 \cdot 49 \cdot 48}.
\]

2. 23 students. The probability of \(k\) students and no matches is

\[
\frac{365 \cdot 364 \cdots (365 - k + 1)}{365^k}.
\]

Here the sample space is the collection of all sequences of length \(k\) where each value is between 1 and 365, and each sequence is equally likely. Even though 23 seems small compared to 365, the number of pairs of people is actually \(\binom{23}{2} = 253\).

3. The probability is roughly \(1/e\). To compute the precise probability we use inclusion-exclusion. Our sample space is the collection of all permutations of 1, \ldots, \(n\) with each equally likely (there are \(n!\) of them). Let \(A_i\) denote the event that the \(i\)th person picks up his ID (i.e., that the permutation has \(i\) in the \(i\)th spot). Then we can compute the probability of at least one person getting their correct ID by inclusion-exclusion:

\[
P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i) - \sum_{i<j} P(A_i A_j) + \cdots + (-1)^{n-1} P(A_1 \cdots A_n).
\]

Note that

\[
P(A_i) = 1/n, \quad P(A_i A_j) = 1/(n(n-1)), \quad P(A_i A_j A_k) = 1/(n(n-1)(n-2)), \ldots.
\]

In general, if there are \(p\) distinct \(A_i\) events we are intersecting the probability is \(\frac{(n-p)!}{n!}\).
This gives

\[
P\left(\bigcup_{i=1}^{n} A_i\right) = n \cdot \frac{1}{n} - \frac{1}{2} \frac{1}{n(n-1)} + \cdots + (-1)^{n-1} \frac{1}{n!} \\
= \sum_{k=1}^{n} (-1)^k \frac{n}{k} (n-k)! \frac{1}{n!} \\
= \sum_{k=1}^{n} (-1)^{k-1} k! \frac{1}{n!} \\
\approx 1 - e^{-1},
\]

where the approximation improves as \( n \) grows (error is less than \( 1/n! \)). The answer to our problem is \( 1/e \). It is interesting that the probability converges to a number strictly between 0 and 1. One intuition for the result is that the events are “weakly dependent” and well approximated by independent events.

3.2 Conditional Probability

3.2.1 Conditioning Formula

Conditioning is the process where we update our beliefs (probabilities) in light of newly found information. As you can imagine, this is a very practical skill to master. It turns out that conditioning can also be used to greatly simplify a calculation, by using the Law of Total Probability. Joseph Blitzstein says “Conditioning is the Soul of Statistics.”

Let \( A, B \) be events with \( P(B) > 0 \). We use the notation \( P(A|B) \) (read “probability of \( A \) given \( B \)” ) denote the probability that \( A \) assuming \( B \) will occur. In other words, assuming we have the information that \( B \) will occur, \( P(A|B) \) is the probability that \( A \) will occur. Formally, it is defined to be

\[
P(A|B) = \frac{P(A \cap B)}{P(B)}.
\]

We also write \( P(A|B, C) = P(A|B \cap C) = P(A|BC) \).

The motivation for this formula can be understood by looking at examples of sample spaces, and drawing Venn diagrams. It can also be instructive to look at the formula in the following form:

\[
P(A \cap B) = P(A|B)P(B).
\]

That is, the probability of \( A \) and \( B \) both occurring is the probability of \( B \) occurring, times the probability of \( A \) occurring given that \( B \) has occurred.

This can be repeated iteratively, to get the following formula:

**Theorem 14.** Assuming \( P(E_1) > 0, \ldots, P(E_n) > 0 \),

\[
P(E_1 \cap E_2 \cap \cdots \cap E_n) = P(E_1)P(E_2|E_1)P(E_3|E_1, E_2) \cdots P(E_n|E_1, \ldots, E_{n-1}).
\]
Proof. We simply write out all the formulas and cancel:

\[
P(E_1) \frac{P(E_2 \cap E_1)}{P(E_1)} \frac{P(E_3 \cap E_2 \cap E_1)}{P(E_2 \cap E_1)} \cdots \frac{P(E_n \cap \cdots \cap E_1)}{P(E_{n-1} \cap \cdots \cap E_1)} = P(E_n \cap \cdots \cap E_1).
\]

\[\square\]

Example 15 (Tree Diagrams). We are going to make 3 coin flips, but with the following rules. Each time you get a heads, the next coin should have half the chance of getting heads (i.e., if your first flip is heads, the second coin will have 1/4 probability of heads, and 3/4 probability of tails). What are the probabilities of getting 0, 1, 2, 3 heads? One way to solve this problem is to draw a tree whose leaves are all 8 possibilities. The probabilities on each edge are conditional, and to compute the probabilities of the leaves, we are using the previous theorem.

3.2.2 Conditioning Exercises

1. A jar contains a 4-sided die and a 6-sided die. We uniformly at random pick a die out of the jar and roll it. Assuming the roll is 3, what is the chance we picked the 6-sided die? Do you expect it to be bigger or smaller than 1/2?

2. Flip a fair coin 5 times in a row. Assuming we get a total of 3 heads, what is the probability the first flip was heads?

3. A standard deck is shuffled, and two cards are dealt face down. The dealer looks at both cards.

   (a) If the dealer tells you “At least one of the cards is an Ace”, what is the probability there are two Aces?

   (b) If the dealer tells you “One of the cards is the Ace of Spades”, what is the probability there are two Aces?

3.2.3 Solutions

1. Let \( A \) be the event of rolling a 3, and \( B \) be the event of drawing the 6-sided die. Then we have

\[
P(A \cap B) = 1/6 \cdot 1/2 \quad \text{and} \quad P(A) = 1/2 \cdot 1/6 + 1/2 \cdot 1/4.
\]

Then we have

\[
P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{1/12}{1/12 + 1/8} = \frac{1}{1 + 3/2} = \frac{2}{5}.
\]

Note that we would have gotten the same result for rolls 1, 2, 3, 4. For roll 5, 6 we know it is 6-sided with probability 1.
2. Let $A$ be the event of getting 3 heads, and let $B$ be the event that the first flip is heads. Then

$$P(A) = \left(\frac{5}{3}\right) \frac{1}{2^5} \quad \text{and} \quad P(A \cap B) = \left(\frac{4}{2}\right) \frac{1}{2^5}.$$

Thus we have

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{\left(\frac{4}{2}\right)}{\left(\frac{5}{3}\right)} = \frac{3}{5}.$$

3. (a) Let $A$ denote the event of getting two aces, and $B$ the event of getting at least one ace. Then we have

$$P(B) = \frac{4 \cdot 51 + 48 \cdot 4}{52 \cdot 51} \quad \text{and} \quad P(A \cap B) = \frac{4 \cdot 3}{52 \cdot 51}.$$

Thus the answer is

$$\frac{P(A \cap B)}{P(B)} = \frac{12}{4 \cdot 51 + 48 \cdot 4} = \frac{3}{99} = \frac{1}{33}.$$

(b) Let $A$ denote the event of getting two aces, and $B$ the event of getting the ace of spaces. Then we have

$$P(B) = \frac{2 \cdot 51}{52 \cdot 51} \quad \text{and} \quad P(A \cap B) = \frac{2 \cdot 3}{52 \cdot 51}.$$

Thus the answer is

$$\frac{P(A \cap B)}{P(B)} = \frac{6}{2 \cdot 51} = \frac{1}{17}.$$
4 Lecture 4

4.1 Conditional Probability

4.1.1 Review Exercises

1. Suppose a man has 2 children. Also assume girls and boys are equally likely, and each day of the week is equally likely.

   (a) If he says at least one of my children is a boy, what is the probability both are boys?

   (b) If he says at least one of my children is a boy born on a Monday, what is the probability both children are boys?

   (c) If he says his youngest child is a boy, what is the probability both are boys?

2. A sack contains 12 coins: 4 have heads on both sides, 4 have tails on both sides, and 4 are standard. A coin is randomly drawn from the sack, and flipped. You are shown the result of the flip, and it is a head. What is the probability the coin was double-headed?

4.1.2 Solutions

1. (a) \( P(\text{at least one boy}) = \frac{3}{4} \) and \( P(\text{both boys} \text{ and at least one boy}) = \frac{1}{4} \) giving

   \[
P(\text{both boys} | \text{at least one boy}) = \frac{1}{3}.
\]

   (b) \[
P(\text{at least one boy born on Monday}) = \frac{1 \cdot 14 + 13 \cdot 1}{14^2} = \frac{27}{14^2}
\]

   \[
P(\text{both boys} \text{ and at least one boy born on Monday}) = \frac{1 \cdot 7 + 6 \cdot 1}{14^2} = \frac{13}{14^2}
\]

   \[
P(\text{both boys} | \text{at least one boy born on Monday}) = \frac{13}{27}.
\]

   (c) \[
P(\text{youngest child is a boy}) = \frac{1}{2}
\]

   \[
P(\text{both boys and youngest is a boy}) = \frac{1}{4}
\]

   \[
P(\text{both boys} | \text{youngest is a boy}) = \frac{1}{2}.
\]
2. Let $A$ denote the event of getting a double-headed coin, and let $B$ denote the event of a head being shown. Then we have

\[ P(B) = \frac{1}{2} \quad \text{and} \quad P(A \cap B) = \frac{1}{3}. \]

Thus the answer is

\[ \frac{1}{3} \div \frac{2}{3} = \frac{2}{3}. \]

### 4.1.3 Conditional Beliefs

The function $P(\cdot | B)$ where $P(B) > 0$ represents our updated beliefs assuming that $B$ will occur. As such, we would hope that our new beliefs satisfy the axioms of probability. Furthermore, our beliefs shouldn’t care which order we receive information. That is, we want the following to hold: Let $P_B(A) = P(A|B)$ where $P(B) > 0$. Then we want

\[ P(A|B, C) = P_B(A|C). \]

When comparing $P(A)$ and $P(A|B)$ we call $P(A)$ the prior probability of $A$ occurring, and $P(A|B)$ the posterior probability of $A$ occurring given that we know $B$.

Since $P_B$ is a probability measure, all theorems we derive from the axioms apply to it. This means we can get conditional versions of any theorem we want. For example, we proved for any probability measure $P$ that Boole’s inequality holds:

\[ P \left( \bigcup_{i=1}^{\infty} A_i \right) \leq \sum_{i=1}^{\infty} P(A_n). \]

Since $P_B$ is a probability measure, we then immediately have

\[ P \left( \bigcup_{i=1}^{\infty} A_i | B \right) \leq \sum_{i=1}^{\infty} P(A_n | B), \]

where $B$ is any event with $P(B) > 0$.

### 4.1.4 Conditional Belief Exercises

1. **Exercise:** Fix a sample space $S$ a probability measure $P$ and an event $B$ with $P(B) > 0$. Prove that $P_B(\cdot) = P(\cdot | B)$ is a probability measure on the events of $S$ (i.e., satisfies the axioms of probability).

2. **Exercise:** Assume $A, B, C \subset S$ with $P(B) > 0$ and $P(C) > 0$. Prove that $P(A|B, C) = P_B(A|C)$ where $P_B$ is defined as in the previous exercise.

3. **Exercise:** State a formula for $P(A \cup B | C)$ assuming $P(C) > 0$.

4. **Exercise:** Assuming $P(E_1) > 0, \ldots, P(E_n) > 0$, give a simple proof by induction that

\[ P(E_1 \cap E_2 \cap \cdots \cap E_n) = P(E_1)P(E_2|E_1)P(E_3|E_2, E_1) \cdots P(E_n|E_1, \ldots, E_{n-1}) \]

for $n \geq 2$. 

20
4.1.5 Solutions

1. We must verify the 3 axioms of probability:

(a) \[
0 \leq \frac{P(A \cap B)}{P(B)} \leq \frac{P(B)}{P(B)} = 1.
\]

(b) \[
P_B(S) = \frac{P(B \cap S)}{P(B)} = \frac{P(B)}{P(B)} = 1.
\]

(c) Let \(A_1, \ldots\) be pairwise disjoint events. Then

\[
P_B\left(\bigcup_{i=1}^{\infty} A_i\right) = \frac{P(B \cap \bigcup_{i=1}^{\infty} A_i)}{P(B)} = \frac{P(\bigcup_{i=1}^{\infty} (B \cap A_i))}{P(B)} = \frac{\sum_{i=1}^{\infty} P(B \cap A_i)}{P(B)} = \sum_{i=1}^{\infty} P_B(A_i).
\]

2. \[
P_B(A|C) = \frac{P_B(A \cap C)}{P_B(C)} = \frac{P(A \cap C \cap B)/P(B)}{P(C \cap B)/P(B)} = P(A|B,C).
\]

3. \[
P(A \cup B|C) = P(A|C) + P(B|C) - P(A \cap B|C).
\]

4. The base case of \(n = 2\) is immediate by the definition of \(P(E_2|E_1)\). Assume that the result holds for any probability measure with \(n - 1\) or less terms. Then we have

\[
\frac{P(E_1 \cap E_2 \cap \cdots \cap E_n)}{P(E_1)} = P(E_2 \cap \cdots \cap E_n|E_1).
\]

But conditioning on \(E_1\) is a bonafide probability measure, and we can apply the induction hypothesis.

4.1.6 The Law of Total Probability, and Bayes’s Formula

The Law of Total Probability (abbreviated as LOTP by Joseph Blitzstein) allows you to decompose a probability
Theorem 16 (The Law of Total Probability). Let $A_1, A_2, \ldots, A_n \subset S$ be pairwise disjoint events with $P(A_k) > 0$ for all $k$ and $\bigcup_{i=1}^n A_n = S$. Stated differently, $A_1, \ldots, A_n$ are a partition of $S$ into events of positive probability. Then for any $B \subset S$ we have

$$P(B) = \sum_{i=1}^n P(B|A_i)P(A_i).$$

The real power of this theorem is that you can pick the $A_i$ however you want. A very common instance is where $A$ is some special event you have chosen, and the partition is $A, A^c$ giving:

$$P(B) = P(B|A)P(A) + P(B|A^c)P(A^c).$$

Next we come to a theorem that is as easy to prove as it is important. Assuming $P(A) > 0$ and $P(B) > 0$, it relates $P(A|B)$ and $P(B|A)$.

Theorem 17 (Bayes’s Formula). Let $B, A$ be events with $P(B) > 0$ and $P(A) > 0$. Then we have

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

You will prove this in a moment. One intuitive way to see why it is true is to draw a Venn diagram. Then it is easier to appreciate why $\frac{P(A)}{P(B)}$ is the scaling factor that turns $P(B|A)$ into $P(A|B)$. As we will learn later, Bayes’s formula is the basis for Bayesian statistical methods.

4.1.7 LOTP and Bayes Exercises

1. You meet a random individual that wants to play chess. Suppose that they are equally likely to be a master or a novice. If they are a master, you have a 5% chance to win. If they are a novice, you have a 70% chance to win. What is your chance to win?

2. You work at a company that has 3 machines $M_1, M_2, M_3$ that produce 30, 50, 20 percent of the products, respectively with defect rates of 1, 2, 3 percent, respectively. If you randomly sample a product produced and it is defective, what is the probability it was produced by $M_2$?

3. Suppose you roll a 100-sided die and then roll a $k$-sided die, where $k$ is the value you rolled on the 100-sided die. What is the probability that your second roll is a 51?

4. You were given a weighted coin with probability $p$ of getting heads, but you don’t know if $p = 1/3$ or $p = 3/4$. You currently believe these two options are equally likely. You flip the coin 4 times and get HTHH. What is the probability $p = 3/4$ given this information?
5. Prove the following variants of LOTP and Bayes’s Formula where $P(C) > 0$:

\[ P(B|C) = \sum_{i=1}^{n} P(B|A_i, C)P(A_i|C) \]

and

\[ P(A|B, C) = \frac{P(B|A, C)P(A|C)}{P(B|C)} \]

4.1.8 Solutions

1. Let $W$ be the event of winning, and $M$ be the event of playing a master. Then

\[ P(W) = P(W|M)P(M) + P(W|M^c)P(M^c) = .05 \cdot .5 + .7 \cdot .5 = .375. \]

2. We have the production probabilities

\[ P(M_1) = .3, \quad P(M_2) = .5, \quad P(M_3) = .2 \]

and the defect probabilities

\[ P(D|M_1) = .01, \quad P(D|M_2) = .02, \quad P(D|M_3) = .03. \]

Using these we compute

\[ P(M_2|D) = \frac{P(D|M_2)P(M_2)}{P(D|M_1)P(M_1) + P(D|M_2)P(M_2) + P(D|M_3)P(M_3)} \]

\[ = \frac{.02 \cdot .5}{.01 \cdot .3 + .02 \cdot .5 + .03 \cdot .2} \]

\[ \approx .526. \]

3. Let $A$ denote the event that the second roll is a 51, and let $F_k$ denote the event that the first roll is $k$. Then we have

\[ P(A) = \sum_{k=1}^{100} P(A|F_k)P(F_k) = \sum_{k=51}^{100} \frac{1}{k} \cdot \frac{1}{100} = \frac{1}{100} \sum_{k=51}^{100} \frac{1}{k} \approx \frac{\log 2}{100}. \]

4. We have that $P(A) = 1/2$ where $A$ is the event $p = 3/4$. Let $F$ be the event of flipping HTHH. Then we compute

\[ P(A|F) = \frac{P(F|A)P(A)}{P(F|A)P(A) + P(F|A^c)P(A^c)} \]

\[ = \frac{\frac{1}{2} \cdot \left(\frac{3}{4}\right)^3 \frac{1}{4}}{\frac{1}{2} \cdot \left(\frac{3}{4}\right)^3 \frac{1}{4} + \frac{1}{2} \cdot \left(\frac{1}{3}\right)^3 \frac{2}{3}} \]

\[ = \frac{27}{256} + \frac{2}{81} \]

\[ \approx .81. \]
5. They follow immediately from the standard variants since conditioning on $C$ gives a probability measure.
5 Lecture 5

5.1 Conditional Probability

5.1.1 Review Exercises

1. Suppose you roll a fair 6-sided die 12 times. What is the probability of getting each value exactly twice?

2. You keep rolling a fair 100-sided die until it is strictly larger than 60. Let $A_j$ denote the event that you roll a total of $j$ times.
   (a) What is $P(A_j)$?
   (b) What is $\sum_{j=1}^{\infty} P(A_j)$?
   (c) Let $N_j$ denote the event that you roll a total of $j$ times, and the last roll is a 90. Compute $P(N_j|A_j)$.
   (d) What is probability the last roll is 90? [Use LOTP and the previous parts.]
   (e) Suppose you roll a fair 100-sided die once. Let $C$ be the event of rolling strictly larger than 60, and let $D$ be the event of rolling a 90. What is $P(D|C)$?

5.1.2 Solutions

1. \[
\frac{\binom{12}{2,2,2,2,2}}{6^{12}} = \frac{12!/(2!)^6}{6^{12}}.
\] This can also be written as \[
\frac{\binom{12}{2} \binom{10}{2} \binom{8}{2} \binom{6}{2} \binom{4}{2} \binom{2}{2}}{6^{12}}.
\]

2. Here we can define $S$ as follows:

\[S = \{(d_1, d_2, \ldots, d_n) : 1 \leq n, 1 \leq d_i \leq 60 \text{ for } i < n, 60 < d_n \leq 100\}\]

with \[P(\{(d_1, \ldots, d_n)\}) = \frac{1}{100^n}.
\]

This is a slight generalization of the general finite sample space called the general countable sample space (this has been added to the end of the Lecture 2 notes).

(a) \[\left(\frac{60}{100}\right)^{j-1} \frac{40}{100}\] for $j \geq 1$.

(b) \[
\sum_{j=1}^{\infty} P(A_j) = \frac{40}{100} \sum_{j=1}^{\infty} \left(\frac{60}{100}\right)^{j-1} = \frac{\frac{40}{100}}{1 - \frac{60}{100}} = 1.
\]
\[(c)\quad P(N_j|A_j) = \frac{P(N_j \cap A_j)}{P(A_j)} = \frac{P(N_j)}{P(A_j)} = \frac{\left(\frac{60}{100}\right)^{j-1} \frac{1}{100} \cdot \frac{40}{100}}{1} = \frac{1}{40}.\]

(d) Let \(M\) denote the event the last roll is a 90. Then we have
\[
P(M) = \sum_{j=1}^{\infty} P(N_j|A_j)P(A_j) = \frac{1}{40} \sum_{j=1}^{\infty} P(A_j) = \frac{1}{40}.
\]
Here we have applied the following countable version of the LOTP:

**Theorem 18** (Countable Law of Total Probability). If \(A_1, A_2, \ldots\) are pairwise disjoint with \(P(A_i) > 0\) for all \(i \geq 1\) then we have
\[
P(B) = \sum_{i=1}^{\infty} P(B|A_i)P(A_i).
\]

(e) \[
P(D|C) = \frac{P(D \cap C)}{P(C)} = \frac{1/100}{40/100} = \frac{1}{40}.
\]

### 5.1.3 Independence

Given two events \(A, B\), we say that they are independent if \(P(A \cap B) = P(A)P(B)\). Note that if \(P(B) > 0\), then this is equivalent to saying \(P(A|B) = P(A)\). That is, learning \(B\) doesn’t change our beliefs on \(A\).

**Example 19.** Suppose we flip 2 coins. Then the event that the first coin is heads is independent of the event that the second coin is heads.

We can extend this to multiple events, but there is a catch.

**Definition 20** (Independence). We say that the events \(A_1, \ldots, A_n\) are independent if
\[
P(A_{i_1} \cap \cdots \cap A_{i_k}) = P(A_{i_1}) \cdots P(A_{i_k})
\]
for any \(2 \leq k \leq n\) and \(1 \leq i_1 < i_2 < \cdots < i_k \leq n\).

It isn’t enough to have each pair be independent. Let’s do a concrete case. Suppose \(n = 3\). Then \(A_1, A_2, A_3\) are independent if
\[
\begin{align*}
P(A_1 \cap A_2) &= P(A_1)P(A_2), \\
P(A_1 \cap A_3) &= P(A_1)P(A_3), \\
P(A_2 \cap A_3) &= P(A_2)P(A_3), \\
P(A_1 \cap A_2 \cap A_3) &= P(A_1)P(A_2)P(A_3).
\end{align*}
\]
If we have an infinite collection of events, we say they are independent if every finite subcollection is independent.
5.1.4 Independence Exercises

1. Suppose we flip a fair coin 3 times. Let $H_i$ denote the event that the $i$th flip is heads. Show that $H_1, H_2, H_3$ are independent.

2. Show that if $E$ and $F$ are independent, then $E$ and $F^c$ are independent.

3. A circuit consists of $n$ components. Let $A_i$ denote the event that the $i$th component fails, and let $P(A_i) = p_i$. Assuming $A_1, \ldots, A_n$ are independent compute the following:
   
   (a) The probability at least one of the components fails.
   
   (b) The probability all of the components fail.

4. Suppose a can has 3 coins with head-probabilities $1/3, 1/2, 2/3$, respectively. We randomly pick out one coin, and flip it three times. Let $H_i$ be the event the $i$th flip is heads.
   
   (a) Are the $H_i$ independent?
   
   (b) Are the $H_i$ independent if we condition on which coin we chose?
   
   (c) What is the probability of getting 3 heads?

5.1.5 Solutions

1. We have $P(H_i) = 1/2$, $P(H_i H_j) = 1/4$ for $i \neq j$ and $P(H_1 H_2 H_3) = 1/8$.
   
   There is some sense in which independence holds here since we assumed it so. Even though we didn’t call it independence at the time, when we modeled coin flips or throwing dice we assumed the flips and dice didn’t interfere with each other, and chose probabilities accordingly. This implicitly amounted to assuming the dice/flips were independent.

2. We must prove that $P(EF^c) = P(E)P(F^c)$. Note that
   
   $$P(E) = P(EF) + P(EF^c) = P(E)(1 - P(F^c)) + P(EF^c).$$

   Thus $P(E)P(F^c) = P(EF^c)$.
   
   A more general result holds true:

   **Theorem 21.** Let $A_1, \ldots, A_n$ be independent events. Then $B_1, \ldots, B_n$ are independent events, where each $B_i$ is either $A_i$ or $A_i^c$.

   The above is really $2^n$ theorems in one, where for each version we choose to put complements on some of the $A_i$’s.

3. (a) $1 - P(A_1^c A_2^c \cdots A_n^c) = 1 - \prod_{i=1}^{n}(1 - p_i)$.

   (b) $P(A_1 A_2 \cdots A_n) = \prod_{i=1}^{n} p_i$. 

27
4. (a) No. Using LOTP:
\[
P(H_i) = 1 + 3 \cdot 2 + 3 \cdot 3 = \frac{1}{2}
\]
\[
P(H_1H_2) = 3 \cdot 1 + 1 \cdot 1 + 2 \cdot 2 = \frac{29}{108}
\]
Thus \(P(H_1)P(H_2) \neq P(H_1H_2)\).

(b) Yes, this is called conditional independence. More generally:

**Definition 22** (Conditional Independence). Let \(A_1, \ldots, A_n, B\) be events with \(P(B) > 0\). We say that \(A_1, \ldots, A_n\) are conditionally independent given \(B\) if they are independent with respect to the measure \(P_B(E) = P(E|B)\).

By our standard method for modeling coin flips we have
\[
P(H_i|C_1) = \left(\frac{1}{3}\right)^2 \quad \text{and} \quad P(H_1H_2|C_1) = \left(\frac{1}{3}\right)^3,
\]
where \(C_1\) is the event of picking the first coin. This shows that once we condition on the first coin the flips are independent. The same argument will show that conditioning on \(C_2\) or \(C_3\) also yields independence.

(c) By LOTP (letting \(C_i\) denote the event of choosing the \(i\)th coin):
\[
P(H_1H_2H_3) = P(H_1H_2H_3|C_1)P(C_1) + P(H_1H_2H_3|C_2)P(C_2) + P(H_1H_2H_3|C_3)P(C_3)
\]
\[
= \frac{1}{3} \left(\frac{1}{3^3} + \frac{1}{2^3} + \frac{2^3}{3^3}\right).
\]

5.1.6 Interesting Problems

1. (Monty Hall) You are on a game show with 3 doors. Behind one is a car, and the other two have goats. You choose a door. Afterward the host opens a door you didn’t pick that has a goat behind it (the host knows where the car is; if you picked the car he chooses randomly with equal chance). He then offers you the option to switch choices to the remaining closed door. Do you switch?

2. A test has been developed to diagnose a disease. It has a 2% chance of a false negative, and a 2% chance of a false positive. The disease infects 1 out of every 1000 people in the population. Assuming you take the test and it unfortunately comes back positive. What is the probability you have the disease?

3. You go to a casino and have decided to play a game. Each round you have a 51% chance of losing your bet, and a 49% chance of winning an amount equal to your bet. Letting \(E_i\) denote the event of winning the \(i\)th round, you may assume that the \(E_i\) are independent. You have $500 and have decided to bet $1 each time. What is the chance you can get to $550 before you lose the money you brought?
5.1.7 Solutions

1. You always switch, and win the car with probability 2/3. To see this, suppose you initially choose door 1. Let $C_i$ denote the event the car is behind door $i$, and let $W$ denote the event of winning the car assuming you always switch. Then we have

$$P(W) = P(W|C_1)P(C_1) + P(W|C_2)P(C_2) + P(W|C_3)P(C_3) = 0 \cdot \frac{1}{3} + 1 \cdot \frac{1}{3} + 1 \cdot \frac{1}{3} = \frac{2}{3}. $$

This shows that blindly switching gives $2/3$ chance of winning. Suppose we also take Monty’s (the host’s) chosen door into account. To that end, let $M_i$ denote the event that Monty opens door $i$. Note that

$$P(C_1|M_2) = \frac{P(M_2|C_1)P(C_1)}{P(M_2|C_1)P(C_1) + P(M_2|C_2)P(C_2) + P(M_2|C_3)P(C_3)}$$

$$= \frac{1/2 \cdot 1/3}{1/2 \cdot 1/3 + 0 \cdot 1/3 + 1 \cdot 1/3} $$

$$= \frac{1}{3}. $$

This shows the strategy of staying when Monty opens door 2 wins with probability 1/3 (so switching wins with probability 2/3). Nearly the same calculation gives the same result for when Monty opens door 3. Thus conditioning on Monty’s door choice gives the same winning probability when you switch. You can also do this with a tree diagram, and it could help illuminate what is going on. It can also help your intuition to think about the case with 1000 doors, when you choose a door, and Monty closes 998 goat-doors you haven’t picked.

One technical point. Above we said $P(C_1) = 1/3$. Strictly speaking, this isn’t known since the show doesn’t have to allocate the cars randomly. What we can say is that we randomly chose a door, and labeled the door we picked #1. Then $P(C_1) = 1/3$ makes sense.

2. Let $D$ denote the event of having the disease, $T$ denote testing positive. Then we have

$$P(D|T) = \frac{P(T|D)P(D)}{P(T|D)P(D) + P(T|D^c)P(D^c)} = \frac{.98 \cdot .001}{.98 \cdot .001 + .02 \cdot .999} \approx .047. $$

We can look at this as saying you have a prior probability of .001 of having the disease, and a positive test increased that to .047, but this number is nothing like .98.

In general, a person often has other symptoms that must be conditioned on, but this result still has real implications on the accuracy of medical tests.

Out of every 1000 people we expect 1 to be infected (.98 infected to also be diagnosed positive), and

$$999 \cdot .02 = 19.98 \approx 20 $$

people to be uninfected but get a positive test result. Thus given a positive test result, there is only about a 1 in 21 chance of being infected (actually .98 in 20.96).
3. This problem is called Gambler’s ruin. It will be helpful to solve a more general instance of this problem. Assume instead that you start with $0 \leq i \leq M$ dollars ($M = 550$ in our case). Let $p$ denote the chance of winning each round (.49 in our case) and $q = 1 - p$. Let $S_i$ denote the event of getting to $M$ dollars before going bankrupt when you being with $i$ dollars. Let $W$ denote the event of winning your first round. Then we have, for $0 < i < M$,

$$P(S_i) = P(S_i|W)P(W) + P(S_i|W^c)P(W^c) = P(S_{i+1})P(W) + P(S_{i-1})P(W^c) = pP(S_{i+1}) + qP(S_{i-1}).$$

If we let $s_i = P(S_i)$ we have

$$s_i = ps_{i+1} + qs_{i-1}$$

for $0 < i < M$. As our edge cases we have $s_0 = 0$ and $s_M = 1$. There are several ways to solve this problem, called a linear homogeneous recurrence equation. There is a general method for solving these equations:

**Theorem 23.** Let $c_2a_{n+2} + c_1a_{n+1} + c_0a_n = 0$ for $n \geq 0$ where the $c_i \in \mathbb{R}$ are fixed constants. Let

$$f(x) = c_2x^2 + c_1x + c_0 = c_2(x - r_1)(x - r_2)$$

be the associated polynomial with roots $r_1, r_2$. If $r_1 \neq r_2$ then solutions have the form $a_n = \alpha r_1^n + \beta r_2^n$ for some $\alpha, \beta \in \mathbb{R}$. Otherwise the solutions have the form $a_n = (\alpha + \beta n)r_1^n$.

Applying this to our problem we have

$$ps_{i+1} - s_i + qs_{i-1} = 0 \implies f(x) = px^2 - x + q = p(x - q/p)(x - 1).$$

If $q \neq p$ then solutions have the form $s_i = \alpha(q/p)^i + \beta$. To solve for $\alpha, \beta$ we plug in with $i = 0, M$:

$$s_0 = 0 = \alpha + \beta \implies \beta = -\alpha$$

$$s_M = 1 = \alpha(q/p)^M - \alpha = \alpha((q/p)^M - 1) \implies \alpha = \frac{1}{(q/p)^M - 1}.$$ 

Thus the solution for $p \neq q$

$$s_i = \frac{(q/p)^i - 1}{(q/p)^M - 1}.$$
If \( p = q = 1/2 \) then we have \( s_i = \alpha + \beta i \). Plugging in gives

\[
\begin{align*}
   s_0 &= 0 = \alpha \\
   s_M &= 1 = \beta M \implies \beta = \frac{1}{M}.
\end{align*}
\]

Thus if \( p = q = 1/2 \) then \( s_i = \frac{i}{M} \).

In summary, the probability of getting \( M \) dollars before going bankrupt when starting with \( i \) dollars is

\[
   s_i = \begin{cases} 
   \frac{\left(\frac{q}{p}\right)^i - 1}{\left(\frac{q}{p}\right)^M - 1} & \text{if } q \neq p, \\
   \frac{i}{M} & \text{if } q = p.
   \end{cases}
\]

When \( p = .49, q = .51, M = 550 \) and \( i = 500 \) we obtain \( .135 \). A much better strategy to use when you have unfavorable odds is bold play. In bold play you always bet the minimum of the total amount of money you have, and the amount you need to win your desired amount. In our example, the bold strategy would bet 50 dollars, and if it failed, would then bet 100 dollars, and if that failed 200 dollars, and then would have to bet the remaining 150 dollars. The odds given bold play are quite favorable (more than \( p + q(p + qp) = .87 \)).
6 Lecture 6

6.1 Finishing Conditional Probability

6.1.1 Review Exercises

1. Let $A, B, C$ be independent events with probabilities $p_A, p_B, p_C$ of occurring, respectively.

   (a) What is the probability all 3 events occurring?
   (b) What is the probability of none of them occurring?
   (c) What is the probability that exactly 1 occurs?

2. You have a 6-sided die that gives the value $i$ with probability $p_i$, where $i = 1, \ldots, 6$. If you roll the die repeatedly, what is the chance you will roll 3 before you get 4?

3. State the definition of events $A, B, C$ being conditionally independent given the event $D$.

4. Assuming the size of a set is roughly proportional to its probability in the following diagrams, which picture most closely depicts independent events $A, B$?

   ![Diagram](image)

6.1.2 Solutions

1. (a) $p_A p_B p_C$
   
   (b) $(1 - p_A)(1 - p_B)(1 - p_C)$
   
   (c) $p_A(1 - p_B)(1 - p_C) + (1 - p_A)p_B(1 - p_C) + (1 - p_A)(1 - p_B)p_C$

2. Using the idea from review exercise 2 in Lecture 5 we have

   $$P(\text{roll 3}|\text{roll 3 or 4}) = \frac{p_3}{p_3 + p_4}. $$

3. The events $A, B, C$ should be independent with respect to the probability measure $P_D(E) = P(E|D)$. More precisely, we need

   \[ P(AB|D) = P(A|D)P(B|D), \quad P(AC|D) = P(A|D)P(C|D), \quad P(BC|D) = P(B|D)P(C|D), \]
\[ P(ABC|D) = P(A|D)P(B|D)P(C|D). \]

4. The left picture.

### 6.1.3 Fallacies and Pitfalls

**Example 24** (Prosecutor’s Fallacy). In a famous trial, a woman was charged with murder when two of her infants died. An expert witness made the claim that the probability of a single infant death to occur randomly is 1 in 8500, so two deaths would be \((1/8500)^2\), or approximately 1 in 73 million. They said this probability is so unlikely, that she must be guilty. The first issue is that the events are definitely not independent. Even if they were, note that the above confuses the probability

\[ P(\text{evidence}|\text{guilty}) \] with \[ P(\text{guilty}|\text{evidence}), \]

and thus gets an entirely incorrect answer.

**Example 25** (Simpson’s Paradox). Suppose you hear that cigar smokers have a higher mortality rate from smoking related illnesses than cigarette smokers. What does this imply about the health risk of cigars vs. cigarettes? This is a complex issue, but let’s consider one small part of it. Would it surprise you to learn that amongst older people (let’s say above 60) cigars have a lower mortality rate, and the same is true for younger people? Must this data be inconsistent? The answer is no, and we will now see why.

More precisely, let our sample space be smokers in our study (of which there are equally many cigarette and cigar smokers), let \(D\) be the event of a person dying from a smoking related illness, let \(G\) be the event of being a cigar smoker, and let \(O\) be the event of being older. Then

\[
\begin{align*}
P(D|G) &> P(D|G^c), \\
P(D|G, O) &< P(D|G^c, O), \\
P(D|G, O^c) &< P(D|G^c, O^c).
\end{align*}
\]

The explanation is that cigar smokers tend to be older than cigarette smokers, and older people die more often than younger people in general. More precisely,

\[
\begin{align*}
P(D|G) &= P(D|G, O)P(O|G) + P(D|G, O^c)P(O^c|G), \\
P(D|G^c) &= P(D|G^c, O)P(O|G^c) + P(D|G^c, O^c)P(O^c|G^c).
\end{align*}
\]

As a concrete (fabricated) example, suppose there are 100 of each type of smoker with the following data:

<table>
<thead>
<tr>
<th>Cigar</th>
<th>Younger</th>
<th>Older</th>
<th>Cigarette</th>
<th>Younger</th>
<th>Older</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smoke Related Death</td>
<td>0</td>
<td>50</td>
<td>Smoke Related Death</td>
<td>30</td>
<td>9</td>
</tr>
<tr>
<td>Not Smoke Related</td>
<td>10</td>
<td>40</td>
<td>Not Smoke Related</td>
<td>60</td>
<td>1</td>
</tr>
</tbody>
</table>

Then cigars vs. cigarettes is 50/100 vs. 39/100, but for younger people it is 0/10 vs. 30/90 and for older people it is 50/90 vs. 9/10.
6.1.4 Fallacy Exercises

The table of mortality information conditioned on cigar smoking is:

<table>
<thead>
<tr>
<th>Cigar</th>
<th>Younger</th>
<th>Older</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smoke Related Death</td>
<td>0</td>
<td>50</td>
</tr>
<tr>
<td>Not Smoke Related</td>
<td>10</td>
<td>40</td>
</tr>
</tbody>
</table>

Compute the following quantities based on the table where $G$ denotes cigar smoking, $O$ denotes older, and $D$ denotes smoke related death.

1. \( P(D \cap O|G) \), \( P(D \cap O^c|G) \)
2. \( P(D^c \cap O|G) \), \( P(D^c \cap O^c|G) \)
3. \( P(O|G) \), \( P(O^c|G) \)
4. \( P(D|G,O) \), \( P(D|G,O^c) \)
5. \( P(D|G) \)

6.1.5 Solutions

1. 
\[
P(D \cap O|G) = \frac{50}{100}, \quad P(D \cap O^c|G) = \frac{0}{100}.
\]

2. 
\[
P(D^c \cap O|G) = \frac{40}{100}, \quad P(D^c \cap O^c|G) = \frac{10}{100}.
\]

3. 
\[
P(O|G) = \frac{90}{100}, \quad P(O^c|G) = \frac{10}{100}.
\]

4. 
\[
P(D|G,O) = \frac{50}{90}, \quad P(D|G,O^c) = \frac{0}{90}.
\]

5. 
\[
P(D|G) = \frac{50}{100}.
\]
6.2 Discrete Random Variables

6.2.1 Functions

A function \( f : X \rightarrow Y \) assigns to each element of \( X \) (the domain) exactly one element of \( Y \) (the codomain). The image (or range) of a function is the set \( f(X) \) defined by

\[
f(X) = \{ f(x) : x \in X \}.
\]

We say that a function is injective or one-to-one if \( f(x_1) = f(x_2) \) implies \( x_1 = x_2 \) for all \( x_1, x_2 \in X \). In the language of calculus/precalculus, injective functions pass the horizontal line test, and thus can be inverted (i.e., we get a function \( f^{-1} : f(X) \rightarrow X \)). We say a function is onto or surjective if \( f(X) = Y \), or in other words, every element of the codomain is assigned to some element of the domain by \( f \). Given any function (even non-invertible functions) we can form something called the preimage. More precisely, if \( A \subset Y \) then we have

\[
f^{-1}(A) = \{ x \in X : f(x) \in A \},
\]

called the preimage of \( A \) under \( f \). The functions we will deal will be real-valued (they may also take values in \( \mathbb{R}^n \), but this can just be looked at as a list of \( n \) real-valued functions). As such, we can perform algebra on functions in a natural way. For instance, if \( f, g : X \rightarrow \mathbb{R} \) then

\[
(f + g)(x) = f(x) + g(x) \quad \text{and} \quad (f^2)(x) = f(x)^2.
\]

6.2.2 Function Exercises

1. Let \( |X| = 5 \) and \( |Y| = 7 \).
   (a) How many functions \( f : X \rightarrow Y \) are there?
   (b) How many of them are injective?
   (c) How many are surjective?
   (d) \( \star \star \) How many have images of size 3?

2. Let \( f : \mathbb{R} \rightarrow \mathbb{R} \) be defined by \( f(x) = x^2 \). What is \( f^{-1}([1, 2]) \)? In other words, what is \( \{ x : f(x) \in [1, 2] \} \)?

3. Let \( X = \{(d_1, d_2) : 1 \leq d_i \leq 6 \} \) and \( Y = \mathbb{R} \). If \( f, g : X \rightarrow Y \) with \( f(x_1, x_2) = x_1 \) and \( g(x_1, x_2) = x_2 \), then what is \( (f + g)^{-1}(\{7\}) = \{ x \in X : (f + g)(x) = 7 \} \)?

6.2.3 Solutions

1. (a) There are \( 7^5 \) possible functions.
   (b) There are \( 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \) possible injective functions.
   (c) There are no possible surjective functions.
(d) We first solve a slightly simpler problem. Suppose $|Z| = 3$. How many surjective functions $f : X \to Z$ are there? The total number of functions from $X$ to $Z$ is $3^5$. The number of non-surjective functions from $X$ to $Z$ is $3^5 - \binom{3}{2}2^5 \cdot \binom{3}{1}1^5$. This formula is computed via inclusion-exclusion:

$$\sum_{A \subseteq Z} |F_A| - \sum_{A, B \subseteq Z, |A|=2, |B|=2} |F_A F_B|$$

where $F_A$ is the number of functions $g : X \to A$. Thus the number of surjective functions is $3^5 - \binom{3}{2}2^5 + \binom{3}{1}$. As there are $\binom{7}{3}$ choices for $Z$ the final answer is

$$\binom{7}{3} \left( 3^5 - \binom{3}{2}2^5 + \binom{3}{1} \right).$$

2. $[-\sqrt{2}, 1] \cup [1, \sqrt{2}]$.

3. $\{(1, 6), (2, 5), (3, 4), (4, 3), (5, 2), (6, 1)\}$.

6.2.4 Random Variables

Having built up the structure of probability theory, we need a way to represent random quantities (like the number of heads in a sequence of flips, etc.). We do this using random variables. Formally, a random variable is a function $X : S \to \mathbb{R}$ where $S$ is our sample space. We typically use capital letters near the end of the alphabet for random variables. [Technically a random variable is a measurable function on the sample space, but we are going to ignore this point, since we aren’t developing measure theory.] You can think of a random variable as a label on each point in the sample space.

**Example 26 (Coin Flips).** Suppose we flip a fair coin 3 times. Let $S$ be our standard associated sample space:

$$S = \{(f_1, f_2, f_3) : f_i \in \{H, T\}\}.$$ 

Define $X : S \to \mathbb{R}$ to be the number of heads that occurred. Then $X$ is a random variable.

6.2.5 Random Variable Exercises

1. Let $X$ be a random variable, and let $A \subseteq \mathbb{R}$. What does $P(X \in A)$ mean? If $x \in \mathbb{R}$, what does $P(X = x)$ mean?

2. Suppose we flip $n$ coins that are heads with probability $p$, and let $X$ denote the number of heads we get. What are $P(X = 0)$, $P(X = 1)$, $P(X = 2)$, and $P(X = 3)$?

3. Roll 2 fair 6-sided dice. Let $X$ represent the first die, and $Y$ the second. What is $P(X + Y = 7)$?
6.2.6 Solutions

1. We define \( P(X \in A) \) by

\[
P(X \in A) = P\{s \in S : X(s) \in A\} = P(X^{-1}(A)).
\]

Note that \( P(X = x) \) is the same as \( P(X \in \{x\}) \). Thus we can write

\[
P(X = x) = P\{s \in S : X(s) = x\} = P(X^{-1}(\{x\})).
\]

In both of these cases we have take preimages to turn statements about the value of \( X \) into events (subsets of \( S \)) whose probability can be determined.

2. The first step is to state the sample space

\[ S = \{(f_1,\ldots,f_n) : f_i \in \{H,T\}\}. \]

We use a general finite sample space with outcome probability given by

\[
P(\{(f_1,f_2,\ldots,f_n)\}) = p^{\# \text{ of heads}}(1 - p)^{\# \text{ of tails}}.
\]

Then \( X : S \to \mathbb{R} \) is defined by

\[ X((f_1,\ldots,f_n)) = \# \text{ of } f_i \text{ that are } H. \]

We then have

\[
P(X = 0) = (1 - p)^n,
\]

\[
P(X = 1) = \binom{n}{1} p(1 - p)^{n-1},
\]

\[
P(X = 2) = \binom{n}{2} p^2(1 - p)^{n-2},
\]

\[
P(X = 3) = \binom{n}{3} p^3(1 - p)^{n-3}.
\]

As an example, we explicitly calculate \( P(X = 2) \) using the definition:

\[
P(X = 2) = P(\{(f_1,\ldots,f_n) : \# \text{ of } f_i \text{ that are } H \text{ is } 2\}) = \binom{n}{2} p^2(1 - p)^{n-2}.
\]

Here there are \( \binom{n}{2} \) different sequences of \( n \) flips with 2 heads and each has a probability of \( p^2(1 - p)^{n-2} \) assigned to it.

3. Using question 3 from the previous set of exercises, we have \( P(X + Y = 7) = \frac{6}{36} = \frac{1}{6} \).
7 Lecture 7

7.1 Discrete Random Variables

7.1.1 Review Exercises

1. Let \( f : \mathbb{R} \rightarrow \mathbb{R} \) be defined by \( f(x) = e^{|x|} \). What is \( f^{-1}([2, 3]) \)?

2. Suppose we flip a fair coin 4 times, and let \( X \) denote the number of heads.
   
   (a) Give the sample space and probability measure.
   
   (b) Compute \( P(X = k) \) for all \( k \in \mathbb{R} \).

7.1.2 Solutions

1. \([−\log(3), −\log(2)] \cup [\log(2), \log(3)]\)

2. (a) \( S = \{(f_1, f_2, f_3, f_4) : f_i \in \{H, T\}\} \) with all outcomes equally likely.
   
   (b) \( P(X = k) = \binom{4}{k} \frac{1}{2^4} \) if \( 0 \leq k \leq 4 \) and 0 otherwise.

7.1.3 Definitions

We say a random variable \( X \) is discrete if its image is a countable set (finite or countably infinite). This implies there is a set \( \{s_1, s_2, \ldots\} \) such that

\[
\sum_{i=1}^{\infty} P(X = s_i) = 1.
\]

[Aside: There is a more general definition which says that says \( X \) takes on countably many values with probability 1, but we wont address this.]

At this point it may be difficult to understand what a non-discrete random variable would behave like, but we will get to those soon. The \textit{probability mass function} (PMF) \( p_X \) of a random variable \( X \) is defined as follows:

\[
p_X(x) = P(X = x).
\]

We say that two discrete random variables have the same distribution if they have the same probability mass function. In fact, for a given probability mass function there are infinitely many random variables that share it. Let’s consider some properties of a probability mass function:

1. Non-negativity: \( p(x) \geq 0 \) for all \( x \in \mathbb{R} \),

2. There is a countable set \( T = \{x_1, \ldots\} \) with \( \sum_{x_i \in T} p(x_i) = 1 \) and \( p(y) = 0 \) for \( y \in \mathbb{R} \setminus T \).
Conversely, any function satisfying these properties is the PMF for some random variable. Another function, the cumulative distribution function (CDF; also called the distribution function), can also be used to describe the distribution of a random variable. It is defined by \( F_X(x) = P(X \leq x) \). Unlike the PMF, the CDF will also be applicable later when we look at non-discrete distributions.

### 7.1.4 Discrete Random Variable Exercises

1. Let \( X \) denote the number of heads in 1 flip of a fair coin. Give the associated PMF and CDF.

2. Let \( Y \) denote the value when you roll a 6-sided die. Give the associated PMF and CDF.

3. Show that any function \( f : \mathbb{R} \rightarrow \mathbb{R} \) with the properties below is the PMF for some random variable.

   (a) Non-negativity: \( p(x) \geq 0 \) for all \( x \in \mathbb{R} \),

   (b) There is a countable set \( T = \{ x_1, \ldots \} \) with \( \sum_{x_i \in T} p(x_i) = 1 \) and \( p(y) = 0 \) for \( y \in \mathbb{R} \setminus T \).

4. \((⋆)\) Suppose you know the CDF \( F_X \) of the discrete random variable \( X \). How can you determine the associated PMF? How do you get the CDF from the PMF?

### 7.1.5 Solutions

1. The PMF is given by

   \[
   p_X(0) = 1/2, \quad p_X(1) = 1/2, \quad \text{and} \quad p_X(x) = 0 \text{ for } x \notin \{0, 1\}.
   \]

   The CDF is

   \[
   F_X(x) = \begin{cases} 
   0 & \text{if } x < 0, \\
   1/2 & \text{if } 0 \leq x < 1, \\
   1 & \text{if } x \geq 1.
   \end{cases}
   \]

2. The PMF is given by

   \[
   p_X(k) = 1/6 \text{ for } k = 1, \ldots, 6 \text{ and } p_X(x) = 0 \text{ otherwise.}
   \]

   The CDF is

   \[
   F_X(x) = \begin{cases} 
   0 & \text{if } x < 1, \\
   1/6 & \text{if } 1 \leq x < 2, \\
   2/6 & \text{if } 2 \leq x < 3, \\
   3/6 & \text{if } 3 \leq x < 4, \\
   4/6 & \text{if } 4 \leq x < 5, \\
   5/6 & \text{if } 5 \leq x < 6, \\
   1 & \text{if } x \geq 6.
   \end{cases}
   \]
3. Use $T$ as the sample space with $P(\{x_i\}) = p(x_i)$ (general countable space) and define $X : T \to \mathbb{R}$ by $X(x_i) = x_i$.

4. We get the PMF $p_X$ as follows:

$$p_X(x) = F_X(x) - F_X(x-)$$

where

$$F_X(x-) = \lim_{t \to x-} F_X(t),$$

the limit from the left side. In other words, the value of the PMF $x$ is the size of the jump in the CDF. Conversely, we obtain the CDF $F_X$ as follows:

$$F_X(x) = \sum_{x_i ; p_X(x_i) > 0} x_i \text{ if } x \leq x_i.$$

7.1.6 Expectation of Discrete Random Variables

The expected value of a discrete random variable is defined by

$$E[X] = \sum_{x \in \text{im}(X)} xp_X(x).$$

That is, it is a weighted average of the values taken by the random variable, where the weight is the probability. As the random variable is discrete, there can only be a countable number of values where the PMF is non-zero, so the sum makes sense. As a technical comment, we do require that the sum above converges absolutely. If not, we say the expectation doesn’t exist as there is no natural way to order the elements (conditionally convergent series can be rearranged to produce any value). We will learn later that if you repeat an experiment and average the results, the answer will converge to the expected value (the law of large numbers).

Also notice that the expectation formula only uses the PMF of a random variable, and not any information about its sample space. Thus two random variables with the same PMF will have the same expectation.

7.1.7 Expectation Exercises

1. Suppose you roll a fair $n$-sided die. What is the expected value?

2. Let $A \subseteq S$ be an event, and let $I_A$ denote the random variable defined by

$$I_A(x) = \begin{cases} 1 & \text{if } x \in A, \\ 0 & \text{otherwise}. \end{cases}$$

What is $E[I_A]$?
3. Suppose we flip $n$ coins with heads probability $p$, and let $X$ denote the number of heads.

(a) What is $p_X(k)$ for $k \in \mathbb{R}$?
(b) What is $\sum_{k=0}^{n} p_X(k)$?
(c) What is $E[X]$?

4. Let $X$ have PMF defined by

$$p_X(x) = \begin{cases} \frac{1}{2^x} & \text{if } x = 1, 2, 3, \ldots, \\ 0 & \text{otherwise}. \end{cases}$$

Compute $E[X]$.

7.1.8 Solutions

1. Let $X$ be a random variable denoting the value of the die. Then we have

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

2. We have

$$E[I_A] = 1 \cdot p_{I_A}(1) + 0 \cdot p_{I_A}(0) = P(A).$$

3. (a) $p_X(k) = \binom{n}{k} p^k (1 - p)^{n-k}$ for $k = 0, \ldots, n$ and 0 otherwise.
(b) 1.
(c) We have

$$E[X] = \sum_{k=1}^{n} k \binom{n}{k} p^k (1 - p)^{n-k}$$

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

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

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

$$= np.$$

The last equation either follows by the binomial theorem, or using the fact that we are summing $p_Y(k)$ where $Y$ is the number of heads in $n - 1$ flips.
4. We solve this in several ways. Our first method introduces an inner summation and then swaps the order of summation:

\[
E[X] = \sum_{k=1}^{\infty} \frac{k}{2^k} = \sum_{k=1}^{\infty} \sum_{j=1}^{\infty} \frac{1}{2^k} = \sum_{j=1}^{\infty} \sum_{k=j}^{\infty} \frac{1}{2^k} = \sum_{j=1}^{\infty} \frac{1}{2^{j-1}} = 2.
\]

As a second method we partially telescope our series:

\[
\left(1 - \frac{1}{2}\right) \sum_{k=1}^{\infty} \frac{k}{2^k} = \sum_{k=1}^{\infty} \frac{1}{2^k} = 1,
\]

so the sum is 2. As a final method, we can use the following theorem:

**Theorem 27.** Let \(X\) be a discrete random variable that only takes positive integer values (i.e., \(\text{im}(X) \subset \mathbb{Z}_{>0}\)). Then we have

\[
E[X] = \sum_{k=1}^{\infty} P(X \geq k).
\]

Then

\[
E[X] = \sum_{k=1}^{\infty} P(X \geq k) = \sum_{k=1}^{\infty} \frac{1}{2^{k-1}} = 2.
\]

This will be an extra-credit problem on Homework 7.

### 7.1.9 Function of a Random Variable

Given any random variable \(X : S \to \mathbb{R}\), and any function \(f : \mathbb{R} \to \mathbb{R}\) we can form a new random variable by composition: \(f(X) = f \circ X\). [Technically, we require a Borel measurable function \(f\), but we are ignoring issues of measurability.]

**Example 28** (Squaring Dice). Let \(X\) denote the value when you roll a 6-sided die. Let \(f : \mathbb{R} \to \mathbb{R}\) be defined by \(f(x) = x^2\). Then \(f(X) = X^2\) is the random variable taking on the values 1, 4, 9, 16, 25, 36, each with probability \(1/6\).
Note that if we let \( Y = f(X) \) then \( Y \) is a function from \( S \) to \( \mathbb{R} \), so it is also a random variable. Moreover, we have the following:

**Theorem 29.** Let \( X : S \to \mathbb{R} \) be a discrete random variable and let \( f : \mathbb{R} \to \mathbb{R} \). Then \( f(X) \) is a discrete random variable.

**Proof.** The image of \( X \) is countable, so the image of \( f(X) \) is countable as well. \( \square \)

We can look at a function of a random variable as “relabeling our labels.”

### 7.1.10 Function of an RV Exercises

1. Let \( X \) denote the value of rolling a fair 6-sided die, and let \( f(x) = (x - 4)^2 \).
   
   (a) What is \( P(f(X) = 1) \)?
   
   (b) What is \( P(f(X) > 4) \)?
   
   (c) What is \( E[X] \)?
   
   (d) What is \( E[f(X)] \)?

2. Let \( X \) denote the value of rolling a fair 6-sided die, and let \( Y \) denote the number of heads when you flip one fair coin.
   
   (a) What is \( E[Y] \)?
   
   (b) Let \( f : \mathbb{R} \to \mathbb{R} \) be defined by
   
   \[
   f(x) = \begin{cases} 
   1 & \text{if } x \geq 4, \\
   0 & \text{otherwise}.
   \end{cases}
   \]
   
   What is \( E[f(X)] \)?

### 7.1.11 Solutions

1. (a) \( P(f(X) = 1) = p_X(3) + p_X(5) = \frac{1}{3} \)
   
   (b) \( P(f(X) > 4) = p_X(1) = \frac{1}{6} \)
   
   (c) \[
   E[f(X)] = 9p_f(x)(9) + 4p_f(x)(4) + 1p_f(x)(1) = \frac{19}{6}.
   \]

2. (a) \( E[Y] = 1/2 \)
   
   (b) \( E[f(X)] = 1P(X \geq 4) = \frac{1}{2} \). Note that \( f(X) \) and \( Y \) have the same PMF, and thus have the same expectation.
8 Lecture 8

8.0.12 Review Exercises

1. Let $X$ denote the value of rolling a 20-sided die, and let $g : \mathbb{R} \to \mathbb{R}$ be defined as follows:

$$
g(x) = \begin{cases} 
1 & \text{if } x > 12, \\
0 & \text{if } x = 12, \\
-1 & \text{if } x < 12.
\end{cases}
$$

What is $E[g(X)]$?

2. You roll a 4-sided die and flip a fair coin. Let $X$ denote the value of the die, and let $Y$ denote the number of heads from the flip.

(a) What is $P(X + Y = 3)$?

(b) What is $P((X + Y)^2 = 4)$?

(c) What is $E[(X - 2Y)^2]$?

8.0.13 Solutions

1. Here we give a full solution.

$$
E[g(X)] = 1 \cdot p_{g(X)}(1) + (-1)p_{g(X)}(-1) = 1 \cdot P(g(X) = 1) + (-1)P(g(X) = -1) = 1 \cdot P(\{s \in S : g(X(s)) = 1\}) + (-1)P(\{s \in S : g(X(s)) = -1\}) = 1 \cdot \frac{8}{20} + (-1) \cdot \frac{11}{20} = -\frac{3}{20}.
$$

2. (a) Let $S = \{(d, f) : 1 \leq d \leq 4, f \in \{H, T\}\}$ with all outcomes equally likely. Then we have

$$
P(X + Y = 3) = P(\{s \in S : X(s) + Y(s) = 3\}) = P(\{(2, 1), (3, 0)\}) = \frac{2}{8}.
$$
(b) 
\[ P((X + Y)^2 = 4) = P(\{s \in S : (X(s) + Y(s))^2 = 4\}) \]
\[ = P(\{s \in S : X(s) + Y(s) = 2\}) \]
\[ = P(\{(2, 0), (1, 1)\}) \]
\[ = \frac{2}{8} \]

(c) Let \( Z = X - 2Y \). Then we have

\[ E[(X - 2Y)^2] = 0 \cdot p_{Z^2}(0) + 1 \cdot p_{Z^2}(1) + 4p_{Z^2}(4) + 9p_{Z^2}(9) + 16p_{Z^2}(16) \]
\[ = 1 \cdot P(Z \in \{-1, 1\}) + 4P(Z = 2) + 9P(Z = 3) + 16P(Z = 4) \]
\[ = 1 \cdot \frac{3}{8} + 4 \cdot \frac{2}{8} + 9 \cdot \frac{1}{8} + 16 \cdot \frac{1}{8} \]
\[ = \frac{36}{8} = \frac{9}{2} \]

8.0.14 Expectation of a Function of a Random Variable

Example 30.

Letting \( Y = f(X) \) we compute the following probabilities:

\[ P(Y = 7) = P(\{a, b\}) \text{ and } P(Y = 8) = P(\{c, d, e\}). \]

Alternatively, we can look at \( f \) as relabeling the labels assigned by \( X \). By unrelabeling we have

\[ P(Y = 7) = P(X = 1) = p_X(1) \text{ and } P(Y = 8) = P(X \in \{2, 3\}) = p_X(2) + p_X(3). \]

We can also compute \( E[Y] \) via unrelabeling:

\[ E[Y] = 7 \cdot p_Y(7) + 8 \cdot p_Y(8) = 7 \cdot p_X(1) + 8(p_X(2) + p_X(3)) = f(1)p_X(1) + f(2)p_X(2) + f(3)p_X(3). \]
Using the above example as motivation, we generalize the result below. To help us compute the expected value of functions of random variables we have the LOTUS theorem (which is essentially computing expectations via “unrelabeling”):

**Theorem 31** (Law of the Unconscious Statistician or LOTUS). *Let $X$ be a discrete random variable, and let $f : \mathbb{R} \to \mathbb{R}$ be a function. Then*

$$E[f(X)] = \sum_{x : p_X(x) > 0} f(x)p_X(x).$$

*To prove this we will use the following which says we can compute the probability $Y = y$ by “unrelabeling”.*

**Lemma 32.** *Let $X$ be a discrete random variable and let $Y = f(X)$ for some function $f$. Then we have*

$$p_Y(y) = \sum_{x : p_X(x) > 0 \atop f(x) = y} p_X(x).$$

**Proof.** We first partition $S$ by the $X$-labels:

$$S = \bigcup_{x : p_X(x) > 0} \{s : X(s) = x\},$$

a disjoint union that is countable by discreteness. Splitting by this partition we can write $A = \{s : Y(s) = y\}$ as

$$A = \bigcup_{x : p_X(x) > 0} \{s : X(s) = x\} \cap A$$

Thus

$$p_Y(y) = P(A) = \sum_{x : p_X(x) > 0 \atop f(x) = y} P(\{s : X(s) = x\} \cap A) = \sum_{x : p_X(x) > 0 \atop f(x) = y} p_X(x).$$

Now we prove the theorem.

**Proof.** Let $Y = f(X)$ so that

$$E[f(X)] = \sum_{y : p_Y(y) > 0} y p_Y(y)$$

$$= \sum_{y : p_Y(y) > 0} y \sum_{x : p_X(x) > 0 \atop f(x) = y} p_X(x)$$

$$= \sum_{y : p_Y(y) > 0} \sum_{x : p_X(x) > 0 \atop f(x) = y} f(x)p_X(x)$$

$$= \sum_{x : p_X(x) > 0} f(x)p_X(x).$$
The last equation holds since \( \sum p_X(x) = 1 \) and these \( x \)-values can be grouped by common \( y \)-value thus giving all possible \( y \)-values.

### 8.0.15 Expectation of a Function of an RV Exercises

1. Suppose \( X \) is a random variable that takes on integer values so that
\[
\sum_{k \in \mathbb{Z}} p_X(k) = 1.
\]

Give an expression for
\[
E \left[ \sin \left( \frac{\pi}{8} X \right) \right].
\]

2. Let \( f(x) = ax + b \) where \( a, b \in \mathbb{R} \) are fixed constants. Show that
\[
E[f(X)] = E[aX + b] = aE[X] + b,
\]
when \( X \) is a discrete random variable.

3. Let \( X \) be a discrete random variable. Give a formula for \( E[X^n] \) where \( n \) is a fixed non-negative integer. This is called the \( n \)th moment of \( X \).

4. The variance of a random variable \( X \) is defined to be \( \text{Var}[X] = E[(X - E[X])^2] \) (also called the second central moment).
   
   (a) Compute the variance of rolling a 6-sided die.
   (b) \( \star \) Assuming \( X \) is discrete, show that the variance is given by \( \text{Var}[X] = E[X^2] - E[X]^2 \).
   (c) Assuming \( X \) is discrete, derive a formula for \( \text{Var}[aX + b] \) where \( a, b \in \mathbb{R} \) are fixed constants.

### 8.0.16 Solutions

1. Using LOTUS we have
\[
E \left[ \sin \left( \frac{\pi}{8} X \right) \right] = \sum_{k \in \mathbb{Z}} \sin(k\pi/8)p_X(k).
\]
2. Using LOTUS we have

\[
E[f(X)] = \sum_{x:p_X(x)>0} f(x)p_X(x)
\]

\[
= \sum_{x:p_X(x)>0} (ax+b)p_X(x)
\]

\[
= a \sum_{x:p_X(x)>0} xp_X(x) + b \sum_{x:p_X(x)>0} p_X(x)
\]

\[
= aE[X] + b.
\]

3. Using LOTUS we have

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

4. (a) \(E[X] = 3.5\) so using LOTUS

\[
E[(X - 3.5)^2] = \frac{1}{6}((-2.5)^2 + (-1.5)^2 + (-.5)^2 + (.5)^2 + (1.5)^2 + (2.5)^2) = \frac{35}{12}
\]

(b) Using throughout that \(E[X]\) can be treated as a constant we have

\[
E[(X - E[X])^2] = E[X^2 - 2E[X]X + E[X]^2]
\]

\[
= E[X^2] - 2E[X]E[X] + E[E[X]^2] \quad \text{(Linearity needed)}
\]

\[
\]

The above argument works if we have a property of expectation called linearity. We will prove this in the next section.

(c) Note that

\[
\text{Var}[aX + b] = E[(aX + b - E[aX + b])^2]
\]

\[
= E[(aX + b - aE[X] - b)^2]
\]

\[
= a^2E[(X - E[X])^2]
\]

\[
= a^2\text{Var}[X].
\]
8.0.17 Introduction to Multiple Discrete Random Variables

Let $X, Y$ be (arbitrary) random variables defined on the same sample space. We say they are independent if

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

for all events $A, B \subset \mathbb{R}$. In words, if $X, Y$ are independent then knowing $X$ is in some $A$ doesn’t change the probability that $Y$ is in some other set $B$. An equivalent statement is that $X, Y$ are independent iff

$$P(X \leq x, Y \leq y) = P(X \leq x)P(Y \leq y) = F_X(x)F_Y(y).$$

The proof is done in a measure theory class.

For discrete random variables, there is a simpler formula. If $X, Y$ are discrete, they are independent iff

$$P(X = x, Y = y) = p_X(x)p_Y(y)$$

for all $x, y \in \mathbb{R}$.

This can be extended to $n$ random variables $X_1, \ldots, X_n$ on the same sample space by requiring

$$P(X_1 \in A_1, \ldots, X_n \in A_n) = P(X_1 \in A_1) \cdots P(X_n \in A_n)$$

to hold for all sets $A_1, \ldots, A_n \subset \mathbb{R}$. The result for PMFs can be extended to $n$ variables as well:

$$P(X_1 = x_1, \ldots, X_n = x_n) = p_{X_1}(x_1) \cdots p_{X_n}(x_n).$$

We now illustrate a general theme that will be true for the rest of the course (with the strong law of large numbers being one exception), and in the process prove a very useful theorem as a preview to later material. The theme is that knowing the PMF of a discrete random variable tells you everything you need to know about it. More generally, if we know

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$$

for a list of $n$ random variables (called the joint PMF), we know everything we need to know about their distributions, and how they interact.

Amazingly, the following result doesn’t require independence.

**Theorem 33 (Linearity of Expectation).** Let $X, Y$ be discrete random variables (defined on the same sample space) whose expectations exist. Then

$$E[X + Y] = E[X] + E[Y].$$

Furthermore, for any $a \in \mathbb{R}$ we have

$$E[aX] = aE[X].$$
This holds for general random variables, as we will see when we learn expectations in a more general context.

We first prove an intuitive and helpful lemma which is similar to the lemma we needed for LOTUS.

**Lemma 34.**

\[ P(X + Y = z) = \sum_{x:p_X(x) > 0, p_Y(y) > 0 \atop x + y = z} P(X = x, Y = y). \]

**Proof.** Let \( S \) denote our sample space and consider the event

\[ E = \{ s : X(s) + Y(s) = z \}. \]

As \( X \) is discrete, we can express this as a countable disjoint union:

\[ E = \bigcup_{x:p_X(x) > 0} \{ s : X(s) = x, Y(s) = z - x \}. \]

Using countable additivity we have

\[
P(X + Y = z) = P(E) = \sum_{x:p_X(x) > 0} P\{ s : X(s) = x, Y(s) = z - x \} = \sum_{p_X(x)>0, p_Y(y)>0 \atop x+y=z} P\{ s : X(s) = x, Y(s) = y \} = \sum_{p_X(x)>0, p_Y(y)>0 \atop x+y=z} P\{ X = x, Y = y \}.
\]

\[ \square \]

This extends to multiple random variables:

**Corollary 35.** If \( X_1, \ldots, X_n \) are discrete then

\[ P(X_1 + \cdots + X_n = z) = \sum_{x_1 + \cdots + x_n = z \atop p_{X_1}(x_1) > 0} P(X_1 = x_1, \cdots, X_n = x_n). \]

Now we prove the theorem.
Proof. Let \( Z = X + Y \). Then

\[
E[Z] = \sum_{p_Z(z) > 0} z P(Z = z)
\]

\[
= \sum_{p_Z(z) > 0} z \sum_{p_X(x) > 0, p_Y(y) > 0 \atop x + y = z} P\{X = x, Y = y\}
\]

\[
= \sum_{p_Z(z) > 0} \sum_{p_X(x) > 0, p_Y(y) > 0 \atop x + y = z} (x + y) P\{X = x, Y = y\}
\]

\[
= \sum_{p_X(x) > 0, p_Y(y) > 0} (x + y) P\{X = x, Y = y\}
\]

\[
= \sum_{p_X(x) > 0, p_Y(y) > 0} x P\{X = x\} + \sum_{p_Y(y) > 0} y P\{Y = y\}.
\]

The result about \( aX \) was proven earlier due to LOTUS.

Applying induction we obtain:

**Corollary 36.** Let \( X_1, \ldots, X_n \) be random variables whose expectations exist. Then we have

\[
E \left[ \sum_{k=1}^{n} X_k \right] = \sum_{k=1}^{n} E[X_k].
\]

### 8.0.18 Introduction to Multiple Discrete RVs Exercises

1. Suppose we roll \( n \) fair 6-sided dice, and let the random variable \( Y_k \) denote the value of the \( k \)th die. Show that the \( Y_k \), for \( k = 1, \ldots, n \) are independent.

2. Suppose \( X, Y \) are discrete random variables that take values in the integers (i.e., the PMFs only assign non-zero probability to integers).

   (a) What is the probability that \( X + Y = 3 \)?

   (b) Repeat the above assuming \( X, Y \) are independent.

3. Suppose we have an urn with 100 red balls, and 200 black balls. Suppose we draw \( k \leq 300 \) balls from the urn in sequence, and let \( X_i \) be the indicator of whether the \( i \)th ball drawn is black (i.e., 1 if black, 0 if not).

   (a) If the balls are drawn with replacement, what is the expected number of black balls?
(b) If the balls are drawn without replacement, what is the expected number of black balls?

(c) If the balls are drawn without replacement, are $X_1$ and $X_2$ independent?

4. In class of $n$ students, how many pairs of students are expected to have the same birthday (assume the year has 365 days each equally likely to be a birthday)?

8.0.19 Solutions

1. Consider $S = \{(d_1, \ldots, d_n) : 1 \leq d_k \leq 6\}$.

$$P(Y_1 = d_1, \ldots, Y_n = d_n) = P((d_1, \ldots, d_n)) = \frac{1}{6^n} = \prod_{k=1}^{n} P(Y_k = d_k).$$

2. (a) Here are two equivalent ways of writing this:

$$P(X + Y = 3) = \sum_{k \in \mathbb{Z}} P(X = k, Y = 3 - k) = \sum_{k \in \mathbb{Z}} P(X = 3 - k, Y = k).$$

This is the discrete version of an operation called a convolution.

(b)

$$P(X + Y = 3) = \sum_{k \in \mathbb{Z}} P(X = k, Y = 3 - k) = \sum_{k \in \mathbb{Z}} P(X = k)P(Y = 3 - k).$$

3. (a) The number of black balls drawn is given by $X_1 + X_2 + \cdots + X_k$. Since $E[X_i] = 2/3$ for all $i$ we obtain $2k/3$.

(b) We first show that $E[X_i] = 2/3$ for all $i$. One way to see this is that when counting possible sequences of draws, the problem is symmetric with respect to the positions, so every $X_i$ has the same probability of being 1 as the first. More explicitly, note that

$$P(X_i = 1) = \frac{200}{300!} \frac{299!}{(299 - (k-1))!} = \frac{200}{300}.$$

Thus $E[X_i] = 2/3$ so the full answer is again $2k/3$.

(c) No. Note that

$$P(X_1 = 1, X_2 = 1) = \frac{200 \cdot 199}{300!} \frac{298!}{(300 - k)!} = \frac{200 \cdot 199}{300 \cdot 299}.$$

This is not $P(X_1 = 1)P(X_2 = 1) = 4/9$.

4. Each pair of students has the same birthday with probability $1/365$ so the answer is

$$\frac{n^2}{365}.$$
9 Lecture 9

9.0.20 Review Exercises

1. (⋆)
   (a) Give a formula for $P(XY = k)$ as a sum over values $x, y$. [Hint: Use a formula similar to the formula for $P(X + Y = k)$.]
   (b) How does the previous formula change if $X, Y$ are independent?
   (c) Give a formula for $E[XY]$ using part (a). [You may assume it exists.]
   (d) How does the formula for $E[XY]$ simplify if $X, Y$ are independent?

2. (a) Compute a formula for $\text{Var}(X+Y)$ that involves $\text{Var}(X)$, $\text{Var}(Y)$ and other terms. [You may assume all expectations exist.]
   (b) Does the formula simplify if $X, Y$ are independent?

9.0.21 Solutions

1. (a)

   $$P(XY = k) = \sum_{x,y:xy=k} P(X = x, Y = y)$$

   (b)

   $$P(XY = k) = \sum_{x,y:xy=k} p_X(x)p_Y(y)$$

   (c)

   $$E[XY] = \sum_{k:p_{XY}(k)>0} kp_{XY}(k)$$

   $$= \sum_{k:p_{XY}(k)>0} \sum_{x,y:xy=k} xyP(X = x, Y = y)$$

   $$= \sum_{x,y:p_X(x)>0} \sum_{p_Y(y)>0} xyP(X = x, Y = y).$$
\[ E[XY] = \sum_{x,y: p_X(x) > 0, p_Y(y) > 0} xyP(X = x, Y = y) \]
\[ = \sum_{x,y: p_X(x) > 0, p_Y(y) > 0} xyp_X(x)p_Y(y) \]
\[ = \sum_{x:p_X(x) > 0} x \sum_{y:p_Y(y) > 0} yyp_X(x)p_Y(y) \]
\[ = \sum_{x:p_X(x) > 0} xp_X(x) \sum_{y:p_Y(y) > 0} yp_Y(y) \]
\[ = E[X]E[Y]. \]

We can actually get more out of this result using the following theorem:

**Theorem 37.** If \( X, Y \) are independent random variables and \( f, g : \mathbb{R} \to \mathbb{R} \) then \( f(X) \) and \( g(Y) \) are also independent.

The proof essentially proceeds by unrelabeling \( f(X) \) and \( g(Y) \).

**Proof.** For any \( A, B \subset \mathbb{R} \) we have

\[
P(f(X) \in A, g(Y) \in B) = P(X \in f^{-1}(A), Y \in g^{-1}(B))
\]
\[ = P(X \in f^{-1}(A))P(Y \in g^{-1}(B)) \quad \text{(Independence of \( X, Y \))} \]
\[ = P(f(X) \in A)P(g(Y) \in B). \]

\[ \square \]

Here is an example illustrating this concept.

**Example 38.** If \( X, Y \) are independent then \( X^2 \) and \( e^Y \) are independent. Also

\[ E[X^3 \sin(Y)] = E[X^3]E[\sin(Y)]. \]

This extends to multiple random variables:

**Theorem 39.** If \( X_1, \ldots, X_n \) are independent and \( f_i : \mathbb{R} \to \mathbb{R} \) for \( i = 1, \ldots, n \) then \( f_1(X_1), \ldots, f_n(X_n) \) are independent.
2. (a)

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

\[ = E[X^2] + E[Y^2] + 2E[XY] - (E[X]^2 + E[Y]^2 + 2E[X]E[Y]) \quad \text{(Linearity of Expectation)} \]

\[ = \text{Var}[X] + \text{Var}[Y] + 2(E[XY] - E[X]E[Y]). \]

The last term has a special definition, called the covariance of \( X \) and \( Y \):


where the last equality is obtained by FOIL. If we apply induction we get the following general result:

**Theorem 40.** Let \( X_1, \ldots, X_n \) be random variables whose variances exist. Then we have

\[ \text{Var}(X_1 + \cdots + X_n) = \sum_{i=1}^{n} \text{Var}(X_i) + 2 \sum_{i<j} \text{Cov}(X_i, X_j). \]

(b) If \( X, Y \) are independent then \( \text{Cov}(X, Y) = 0 \). Thus for independent random variables (we proved it for discrete, but it holds more generally) we have \( \text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) \). More generally (using the above theorem), if \( X_1, \ldots, X_n \) are independent then

\[ \text{Var}(X_1 + \cdots + X_n) = \sum_{i=1}^{n} \text{Var}(X_i). \]

### 9.1 Examples of Discrete Random Variables

We now discuss distributions of random variables. As we have seen, many interesting properties of a random variable, such as its expectation, can be computed knowing only the associated PMF. We say that two discrete random variables \( X, Y \) have the same distribution if they have the same PMF. This ignores any details about the sample spaces on which they are defined. For the next few sections, we discuss families of discrete distributions. We will use the notation

\[ X \sim \text{something} \]

to say that \( X \) has a distribution that is described by something, where something is a PMF, CDF, or some other description of a distribution. We will also say that \( X_1, \ldots, X_n \) are i.i.d. (independent and identically distributed) if they are independent and all have the same CDF (or PMF).
9.1.1 Binomial Random Variables

A random variable $X$ has a Bernoulli distribution, if its associated PMF assigns zero probability to all values but 0, 1. That is, there is some parameter $p \in [0, 1]$ with

$$p_X(0) = P(X = 0) = 1 - p \quad \text{and} \quad p_X(1) = P(X = 1) = 1.$$ 

We write $X \sim \text{Ber}(p)$ to signify that $X$ is distributed according to a Bernoulli distribution with parameter $p$. Notice that knowing the distribution has type Bernoulli, and knowing $p$ uniquely describes an associated PMF.

A random variable $X$ has a Binomial distribution with parameters $n, p$ if we have

$$p_X(k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, \ldots, n.$$ 

We write $X \sim \text{Bin}(n, p)$ to denote this. As we have seen earlier, the “story” of the Binomial distribution is that we we flip $n$ coins with probability $p$ of heads, and count the number of heads. More generally,

**Theorem 41.** Fix $p \in [0, 1]$ and let $X_1, \ldots, X_n \sim \text{Ber}(p)$ be independent random variables. Then $X_1 + \cdots + X_n \sim \text{Bin}(n, p)$.

**Proof.** We give two proofs.

1. Using our formula for the probability of a sum we have

$$P(X_1 + \cdots + X_n = k) = \sum_{x_1 + \cdots + x_n = k} P(X_1 = x_1, \ldots, X_n = x_n)$$

$$= \sum_{x_1 + \cdots + x_n = k} P(X_1 = x_1) \cdots P(X_n = x_n)$$

$$= \sum_{x_1 + \cdots + x_n = k} p^k (1-p)^{n-k}$$

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

2. We have already seen that the number of heads in $n$ flips of a coin with head probability $p$ is distributed as $\text{Bin}(n, p)$, and that the flips $Y_i$ are independent. This gives an example of $n$ i.i.d. $\text{Ber}(p)$ RVs whose sum gives a $\text{Bin}(n, p)$ random variable. In particular,

$$\binom{n}{k} p^k (1-p)^{k} = P(Y_1 + \cdots + Y_n = k) = \sum_{y_1 + \cdots + y_n = k} P(Y_1 = y_1, \ldots, Y_n = y_n).$$
But the value of the sum on the right is uniquely determined by the fact that $Y_i \sim \text{Ber}(p)$ and the $Y_i$ are independent. In other words, it gives the same calculation for any sequence of $n$ random variables $Y_i$ that are i.i.d. Ber($p$).

Just as a reminder, whenever we state a theorem with multiple random variables, we assume they all are defined on the same sample space. For example, in the theorem above it makes no sense to take the sum of random variables defined on different sample spaces.

The second proof above can be applied more generally. In particular, suppose we have a particular example of a sample space $S$ and random variables $Y_1, \ldots, Y_n : S \rightarrow \mathbb{R}$ that are i.i.d. with some CDF $F_1$. Furthermore, suppose we know that $Y_1 + \cdots + Y_n$ has CDF $F_2$. Then any $n$ i.i.d. random variables with CDF $F_1$ will have a sum with CDF $F_2$. This implies the “story” of a distribution has general applicability.

### 9.1.2 Binomial Exercises

1. Let $S = \{s_1, \ldots, s_n\}$ be a general finite space where $P(\{s_i\}) = \binom{n}{i} p^i (1-p)^{n-i}$ for some fixed $p \in [0,1]$. If $X(s_i) = i$ what is the distribution of $X$?

2. Suppose an urn has $N$ red balls and $M$ black balls. What is the distribution of the number of red balls you will obtain when choosing $k$ balls with replacement?

3. A company sells packages of 10 screws, and guarantees that if more than 1 is faulty, you get your money back. If each screw is defective with probability .01 independently, what is the probability that money will be returned to the customer?

4. A jury trial requires at least 8 of the 12 jurors to convict the defendant. Let $\alpha$ be the probability the defendant is guilty. Conditional on the innocence or guilt of the defendant, each juror makes the correct decision independently with probability $p$. What is the probability the jury renders the correct decision? The formula will have summations in it.

5. Let $Y \sim \text{Bin}(n, p)$. Compute the following:

   - (a) $\text{Var}[Y]$
   - (b) $\text{(*)}$ The mode of $Y$ (the value(s) where the PMF is largest).

### 9.1.3 Solutions

1. Bin$(n, p)$. Note the sample space has nothing to do with coin flips.

2. Bin$(k, N/(N + M))$
3. The number of defective screws $X$ in a pack has $X \sim \text{Bin}(10, .01)$. Thus we want

$$P(X > 1) = 1 - P(X \leq 1) = 1 - p_X(1) - p_X(0) = 1 - \binom{10}{1}(.01)^1(.99)^9 - \binom{10}{0}(.99)^{10} \approx .004.$$  

4. Let $C$ denote the event the jury is correct, and $G$ the event the defendant is guilty.

Then we have (by LOTP)

$$P(C) = P(C|G)P(G) + P(C|G^c)P(G^c) = P(X \geq 8) \cdot \alpha + P(X \geq 5) \cdot (1 - \alpha),$$

where $X \sim \text{Bin}(12, p)$.

5. (a) Let’s first compute $\text{Var}(X)$ where $X \sim \text{Ber}(p)$. This gives


Thus by our earlier result on summing variances for independent variables we have

$$\text{Var}[Y] = np(1-p)$$

if $Y \sim \text{Bin}(n, p)$.

(b) Here are some pictures to illustrate the PMFs of the Binomial distribution:
The mode is the value such that \( p_X(x) \) is largest, i.e., the highest point in the graph. Note that in the case of \( \text{Bin}(21, .5) \) there are two modes. If \( p = 0 \) the answer is 0 and if \( p = 1 \) the answer is \( n \). Otherwise, we ask the question “when is \( p_X(k + 1) \geq p_X(k) \)?” To answer this question, we determine when the following quotient is bigger than 1:

\[
\frac{p_X(k + 1)}{p_X(k)} = \frac{\binom{n}{k+1}p^{k+1}(1-p)^{n-k-1}}{\binom{n}{k}p^k(1-p)^{n-k}} = \frac{(n-k)p}{(k+1)(1-p)}.
\]

This is bigger than 1 when

\[
(n-k)p \geq (k+1)(1-p) \iff np \geq k+1-p \iff k \leq (n+1)p-1.
\]

Thus we have show that \( p_X(k + 1) \geq p_X(k) \) if and only if \( k \leq (n+1)p-1 \). Thus the mode is \( \lceil (n+1)p-1 \rceil \). Actually, when \((n+1)p\) is an integer this shows there are two modes at \((n+1)p-1\) and \((n+1)p\).

### 9.1.4 Hypergeometric Random Variable

Suppose we have an urn with \( N \) balls of which \( B \) are black. If we draw \( n \) balls with replacement, we saw above the number of black balls drawn is distributed as \( \text{Bin}(n, B/N) \). The hypergeometric distribution corresponds to the situation without replacement, which we write \( \text{HG}(n, B, N) \). Suppose \( X \sim \text{HG}(n, B, N) \). Then the PMF is given by

\[
p_X(k) = \binom{B}{k} \binom{N-B}{n-k} \binom{N}{n}, \quad k=0, \ldots, B,
\]

by a counting problem we have already done.

If \( N \to \infty \) while \( B/N \to p \), the difference between sampling with replacement and without replacement becomes negligible.

**Theorem 42.** Fix \( n \geq 0 \) and let \( X_N \sim \text{HG}(n, B_N, N) \) for all \( N \geq n \) with \( 0 \leq B_N \leq N \). Suppose \( B_N/N \to p \) as \( N \to \infty \). Then

\[
P(X_N = i) \to P(Y = i)
\]

for any fixed \( i \), where \( Y \sim \text{Bin}(n, p) \).

**Proof.** Note that

\[
P(X_N = i) = \binom{B_N}{i} \binom{N-B_N}{n-i} \binom{N}{n} \frac{B_N(B_N-1) \ldots (B_N-i+1)}{i!} \frac{(N-B_N) \ldots (N-B_N-(n-i)+1)}{(n-i)!} \frac{(N-1) \ldots (N-n+1)}{n!}
\]

\[
= \binom{n}{i} \binom{B_N}{N} \binom{B_N-i+1}{n-i+1} \binom{N-B_N}{N-i} \binom{N-1}{N-n+1}
\]

\[
\to \binom{n}{i} p^i (1-p)^{n-i}.
\]
9.1.5 Hypergeometric Exercises

1. Suppose we have $N$ balls in an urn of which $B$ are black, and we draw $n$ without replacement. Let $X_i$ be the indicator that the $i$th ball drawn is black. Let $Y \sim \text{HGeom}(n, B, N)$.

(a) What is $E[Y]$?
(b) What is $\text{Cov}(X_1, X_2)$?
(c) What is $\text{Cov}(X_1, X_1)$?
(d) What is $\text{Var}(Y)$?

9.1.6 Solutions

1. (a) By linearity of expectation $E[Y] = nB/N$.

(b) 

$$E[X_1X_2] - E[X_1]E[X_2] = \frac{B(B-1)}{N(N-1)} - \frac{B^2}{N^2} \cdot \frac{B}{N} \left( \frac{B-1}{N-1} - \frac{B}{N} \right)$$

$$= \frac{B}{N} \left( \frac{B-1}{N-1} - \frac{B}{N} \right)$$

$$= \frac{B}{N} \left( \frac{(B-1)N - B(N-1)}{N(N-1)} \right)$$

$$= -\frac{B(N-B)}{N^2(N-1)}.$$ 

(c) Since $X_i \sim \text{Ber}(B/N)$ we have

$$\text{Cov}(X_1, X_1) = \text{Var}(X_1) = \frac{B(N-B)}{N^2}.$$ 

(d) The variance of $Y$ is given by (since every pair is the same)

$$\text{Var}(Y) = \sum_{i=1}^{n} \text{Var}(X_i) + 2 \sum_{i < j} \text{Cov}(X_i, X_j)$$

$$= n \frac{B(N-B)}{N^2} - 2 \binom{n}{2} \frac{B(N-B)}{N^2(N-1)}$$

$$= \frac{nB(N-B)}{N^2} - \frac{n(n-1)B(N-B)}{N^2(N-1)}.$$ 

If we let $B/N = p$, the proportion of black balls, we get

$$np(1-p) - np(1-p) = np \frac{N-n}{N-1}.$$
This is the same as the Bin($n, p$) variances with the exception of the factor

$$\frac{N - n}{N - 1}$$

called the finite sample correction. Note that as $N \to \infty$ the correction limits to 1.


10 Lecture 10

10.0.7 Review Exercises

1. Earlier in the course we modeled the event of rolling $n$ fair 6-sided dice using the sample space

$$S = \{(d_1, \ldots, d_n) : 1 \leq d_i \leq 6\}$$

with measure determined by the fact that all outcomes are equally likely. We modeled the event of flipping $n$ coins with heads probability $p$ using

$$S = \{(f_1, \ldots, f_n) : f_i \in \{H, T\}\}$$

with measure determined using the general finite space with

$$P(\{(f_1, \ldots, f_n)\}) = p^{\text{#heads}}(1 - p)^{\text{#tails}}.$$

Looking back on what we now know, why did we choose these measures?

2. Suppose there is an urn with 10 balls of which 3 are black. You draw 5 balls without replacement, with $X_i$ indicating whether the $i$th draw is black. Let $\bar{X}$ denote the proportion of the 5 drawn balls that is black. That is

$$\bar{X} = \frac{1}{5} \sum_{i=1}^{5} X_i.$$

(a) What is $E[\bar{X}]$?

(b) What is $\text{Var}[\bar{X}]$?

10.0.8 Solutions

1. We wanted each die roll and each coin flip to be independent random variables.

2. Each $X_i$ is a $\text{Ber}(.3)$ random variable, and $\bar{X} = Y/5$ where $Y \sim \text{HGeom}(5, 3, 10)$. Thus

(a) $E[\bar{X}] = .3$ by linearity of expectation.

(b) $\text{Var}[\bar{X}] = \text{Var}[Y/5] = \frac{\text{Var}(Y)}{25} = \frac{1}{25} \cdot 5(.3)(.7)^{10-5} \approx .0233$.

10.0.9 Uniform Random Variables

The discrete uniform distribution is probably the second simplest after Bernoulli. Given a finite set of values $V = \{v_1, \ldots, v_n\} \subset \mathbb{R}$, the distribution $\text{Unif}(V)$ assigns equal probability $1/n$ to each. The most typical choice are the values $1, 2, \ldots, n$ modeling an $n$-sided die.
10.0.10 Geometric and Negative Binomial Random Variables

The Geometric distribution (Geom(\(p\))) models the number of flips you need to get your first head where \(0 < p \leq 1\) is the probability of getting heads. The Negative Binomial distribution (NBin(\(r, p\))) models the number of trials required to obtain \(r\) heads. Clearly Geom(\(p\)) = NBin(1, \(p\)).

Let \(X \sim \text{NBin}(r, p)\) then we have

\[
p_X(k) = \binom{k - 1}{r - 1} p^r (1 - p)^{k-r},
\]

by counting the number of possibilities where the last flip is heads. In the specific case of \(Y \sim \text{Geom}(p)\) we thus have

\[
p_Y(k) = p(1 - p)^{k-1}.
\]

There are several ways of computing the expectation:

1. If we only want the expectation of a Geom(\(p\)) random variable, it is given by

\[
E[X] = \sum_{k=1}^{\infty} kp(1 - p)^{k-1} = p \sum_{k=1}^{\infty} k(1 - p)^{k-1} = \frac{p}{(1 - (1 - p))^2}
\]

by taking derivatives. Thus \(E[X] = 1/p\). We will see another method for computing this later using conditional expectation.

2. Similarly we can compute the variance of a Geom(\(p\)) RV. Note that

\[
E[X^2] = \sum_{k=1}^{\infty} k^2 p(1 - p)^{k-1} = p \sum_{k=1}^{\infty} k^2 (1 - p)^{k-1}.
\]

Note that

\[
\sum_{k=1}^{\infty} k^2 x^{k-1} = \frac{d}{dx} \sum_{k=1}^{\infty} kx^k = \frac{d}{dx} \sum_{k=1}^{\infty} \frac{x^k}{(1-x)^2} = \frac{(1-x)^2 + 2x(1-x)}{(1-x)^4} = \frac{1+x}{(1-x)^3}.
\]

This shows \(E[X^2] = \frac{2-p}{p^2}\) and

\[
\text{Var}[X] = E[X^2] - E[X]^2 = \frac{2-p}{p^2} - \frac{1}{p^2} = \frac{1-p}{p^2}.
\]
Theorem 43. If \( X_1, \ldots, X_r \sim \text{Geom}(p) \) are independent then \( Y = X_1 + \cdots + X_r \sim \text{NBin}(r, p) \).

Proof. We can just compute this directly:

\[
p_Y(k) = \sum_{i_1 + \cdots + i_r = k} P(X_1 = i_1, \ldots, X_r = i_r)
= \sum_{i_1 + \cdots + i_r = k} P(X_1 = i_1) \cdots P(X_r = i_r)
= \sum_{i_1 + \cdots + i_r = k} p^r (1-p)^{k-r}
= \binom{k-1}{r-1} p^r (1-p)^{k-r},
\]

where we counted the number ways to partition \( k \) indistinguishable items into \( r \) bins assuming \( i_1, \ldots, i_r \geq 1 \) are the bin counts.

We could have also done a “story” type proof. In such a proof \( Y \sim \text{NBin}(r, p) \) is defined on a sample space of finite sequences of (possibly unfair) coin flips that end with their \( r \)th head, and \( X_k \) is the number of flips after the \((k - 1)\)th head, up until the \( k \)th head. Each \( X_k \) can then be shown to be independent and \( \text{Geom}(p) \) distributed.

This theorem shows \( \mathbb{E}[Y] = r/p \) by linearity. This works even if your negative binomial random variable isn’t a sum of geometrics, since the expectation only depends on the distribution.

We also could have just summed the series directly to obtain the expectation.

4. As we saw above, variances add for independent random variables. Thus the variance is

\[
\text{Var}[Y] = \frac{r(1-p)}{p^2}.
\]

10.0.11 Uniform, Geometric, and Negative Binomial Exercises

1. Let \( V = \{1, \ldots, n\} \) and let \( X \sim \text{Unif}(V) \).

   (a) What is \( \mathbb{E}[X] \)?
   (b) What is \( \text{Var}[X] \)?
   (c) How do the previous answers change if \( V = \{a, \ldots, b\} \) for some \( a, b \in \mathbb{Z} \) with \( a < b \)?
2. You are given a coin that is heads with probability 1/30.
   (a) What is the expected number of tosses till you get 5 heads?
   (b) What is the expected number of tails you get before seeing 5 heads?
   (c) What is the variance of the number of tosses till you get 5 heads?
   (d) What is the variance of the number of tails you get before seeing 5 heads?

3. Let $X \sim \text{Geom}(p)$. What is the probability that $X$ takes on an even value?

4. Let $X \sim \text{Geom}(p)$. Assuming $k, t \geq 1$, what is $P(X = k + t | X > k)$?

### 10.0.12 Solutions

1. (a) Same as rolling a fair $n$-sided die:

   $$E[X] = \frac{n + 1}{2}.$$ 

(b) Note that

   $$E[X^2] = \sum_{k=1}^{n} \frac{k^2}{n} = \frac{n(n + 1)(2n + 1)}{6n} = \frac{(n + 1)(2n + 1)}{6}$$

   giving

   $$\text{Var}[X] = E[X^2] - E[X]^2 = \frac{(n + 1)(2n + 1)}{6} - \frac{(n + 1)^2}{4} = \frac{4(2n^2 + 3n + 1) - 6(n^2 + 2n + 1)}{24} = \frac{2n^2 - 2}{24} = \frac{n^2 - 1}{12}.$$ 

(c) Letting $n = b - a + 1$ it is exactly the previous problem using $Y = X + a - 1$. Note that

   $$E[Y] = a - 1 + \frac{n + 1}{2} \quad \text{and} \quad \text{Var}[Y] = \frac{n^2 - 1}{12}.$$ 

2. (a) This is the expectation of a NBin(5, 1/30) RV giving $5/p = 150$.

(b) If $Y$ is the number of tails then $E[5 + Y] = 150$ so $E[Y] = 145$.

(c) $r(1 - p)/p^2 = 5(29/30)^2 = 5(29)(30) = 4350$

(d) Also 4350.
3. This is
\[
\sum_{k=0}^{\infty} p(1-p)^{2k+1} = p(1-p) \sum_{k=0}^{\infty} (1-p)^{2k} = \frac{p(1-p)}{1-(1-p)^2} = \frac{p(1-p)}{2p-p^2} = \frac{1-p}{2-p}.
\]

4. We have
\[
P(X = k + t | X > k) = \frac{P(X = k + t, X > k)}{P(X > k)} = \frac{p(1-p)^{k+t-1}}{\sum_{s=1}^{\infty} p(1-p)^{k+s-1}} = \frac{p(1-p)^{k+t-1}}{p(1-p)^k} = p(1-p)^{t-1} = P(X = t).
\]

This shows the Geometric distribution is “memoryless”.

10.0.13 Poisson Random Variable

The Poisson distribution is possibly the most important discrete distribution. If \(X \sim \text{Pois}(\lambda)\) with \(\lambda > 0\) then \(X\) has PMF given by

\[
p_X(k) = \frac{e^{-\lambda} \lambda^k}{k!}
\]

for \(k \geq 0\). One way to understand its value is as an approximation to the binomial distribution:

**Theorem 44** (Poisson Approximation to the Binomial). Let \(p_n \geq 0\) such that \(np_n \to \lambda\) as \(n \to \infty\). If \(X_n \sim \text{Bin}(n, p_n)\) then, for fixed \(k \geq 0\),

\[
p_{X_n}(k) = P(X_n = k) \to \frac{e^{-\lambda} \lambda^k}{k!}.
\]

**Proof.** Recall that

\[
p_{X_n}(k) = \binom{n}{k} p_n^k (1 - p_n)^{n-k} = \frac{n!}{(n-k)!k!} p_n^k (1 - p_n)^{n-k} = \frac{(np_n)^k}{k!} \cdot \frac{n!}{(n-k)!n^k} (1 - p_n)^{n-k}.
\]
By assumption we see \( np_n \to \lambda \) and since \( k \) is fixed we have
\[
\frac{n!}{(n-k)!n^k} = \frac{n(n-1) \cdots (n-k+1)}{n^k} \to 1.
\]

We are done if we can show that \((1 - p_n)^{n-k} \to e^{-\lambda}\). Note that
\[
(1 - p_n)^{n-k} = \left(1 - \frac{np_n}{n}\right)^{n-k}.
\]
Since \( np_n/n \to 0 \) we see the \(-k\) in the exponent can be removed. Taking logs gives
\[
\log \left(1 - \frac{np_n}{n}\right) = n \left(-\frac{np_n}{n} - O(1/n^2)\right) = -np_n - O(1/n) \to -\lambda.
\]

The error in the approximation of Bin\((n,p)\) by Pois\((np)\) is bounded by \(p\), but a proof of this is beyond the scope of the class.

The above theorem gives a good point of view when understanding the Poisson distribution. If we are modeling a phenomenon that has many rare events independent \((p_n \text{ small})\), a Poisson distribution may be accurate. Using more advanced methods than we shall study in this course (Stein-Chen method) some of these conditions can be relaxed. For example, if \(X_1, \ldots, X_k\) each have low chance of occurring (with possibly distinct probabilities \(p_1, \ldots, p_k\)) of occurring and are independent or “weakly dependent” (too technical to define here) then the distribution is well approximated by a Pois\((p_1 + \cdots + p_k)\) random variable. Here are some examples of events that have been successfully modeled with a Poisson random variable:

1. Number of misprints in a book, or on a page.
2. Number of blemishes per page of white bond paper, or defects on a computer screen.
3. Number of people in a community that live until age 100.
4. Number of customers arriving at a bank, or calling a call center.
5. Number of hits on a webpage.
6. Number of car accidents that will occur in a day.
7. Number of bid/offers being placed in a stock in a given period of time.
8. Number of particles discharged from radioactive material.

When modeling with the Poisson distribution, it is important to know what \(\lambda\) corresponds to. For example, with misprints in a book \(\lambda\) might denote the average number of errors per page. If we look at a 20 page chapter, we may want to know the distribution of errors in the chapter. Fortunately we have the following nice theorem:
Theorem 45. If $X \sim \text{Pois}(\lambda_1)$ and $Y \sim \text{Pois}(\lambda_2)$ are independent then $X + Y \sim \text{Pois}(\lambda_1 + \lambda_2)$.

Proof. Note that

$$P(X + Y = k) = \sum_{i=0}^{\infty} P(X = i, Y = k - i)$$

$$= \sum_{i=0}^{k} P(X = i) P(Y = k - i)$$

$$= \sum_{i=0}^{k} \frac{e^{-\lambda_1} \lambda_1^i}{i!} \frac{e^{-\lambda_2} \lambda_2^{k-i}}{(k-i)!}$$

$$= \frac{e^{-(\lambda_1 + \lambda_2)} k!}{k!} \sum_{i=0}^{k} \frac{1}{i!(k-i)!} \lambda_1^i \lambda_2^{k-i}$$

$$= \frac{e^{-(\lambda_1 + \lambda_2)} (\lambda_1 + \lambda_2)^k}{k!}.$$

Thus if each page is independent, the misprints in a 20 page chapter will have a \text{Pois}(20\lambda) distribution. Generalizing we have

Theorem 46. If $X_1 \sim \text{Pois}(\lambda_1), \ldots, X_n \sim \text{Pois}(\lambda_n)$ are independent then we have

$$X_1 + \cdots + X_n \sim \text{Pois}\left(\sum_{i=1}^{n} \lambda_i\right).$$

10.0.14 Poisson Exercises

1. Let $X \sim \text{Pois}(\lambda)$.

   (a) What is $E[X]$?
   (b) $(\star)$ What is $\text{Var}[X]$?
   (c) $(\star)$ What is the mode of $X$?

2. In a class of $n$ people let $X_{ij}$ be the indicator random variable that people $i, j$ have the same birthday. Assuming these indicators are weakly dependent, what is the Poisson approximation to the probability at least 1 pair of people have the same birthday? How about at least 2 pair?
10.0.15 Solutions

1. (a) We have

\[
E[X] = \sum_{k=1}^{\infty} kp_X(k) \\
= \sum_{k=1}^{\infty} ke^{-\lambda} \lambda^k \frac{1}{k!} \\
= \sum_{k=1}^{\infty} e^{-\lambda} \lambda^k \frac{(k-1)!}{(k-1)!} \\
= \lambda \sum_{k=1}^{\infty} e^{-\lambda} \lambda^{k-1} \frac{(k-1)!}{(k-1)!} \\
= \lambda .
\]

(b) We have

\[
E[X^2] = \sum_{k=1}^{\infty} k^2 p_X(k) \\
= \sum_{k=1}^{\infty} ke^{-\lambda} \lambda^k \frac{1}{(k-1)!} \\
= \lambda \sum_{k=1}^{\infty} ((k-1) + 1) e^{-\lambda} \lambda^{k-1} \frac{(k-1)!}{(k-1)!} \\
= \lambda(\lambda + 1) .
\]

Alternatively, we could compute

\[
E[X(X - 1)] = \lambda^2
\]

by an easier LOTUS computation. Thus

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

(c) We compute

\[
\frac{p_X(k+1)}{p_X(k)} = \frac{e^{-\lambda} \lambda^{k+1}}{(k+1)!} \frac{(k+1)!}{e^{-\lambda} \lambda^k k!} = \frac{\lambda}{k+1} .
\]

Thus the mode is \([\lambda - 1]\). If \(\lambda\) is an integer, then \(\lambda\) and \(\lambda - 1\) are both modes.

2. There are \(\binom{n}{2}\) indicators, and each has probability \(\frac{1}{365}\) of occurring. If we assume they are weakly dependent we can model this situation with

\[
Y \sim \text{Pois} \left( \frac{n(n-1)}{2 \cdot 365} \right)
\]

69
random variable. Thus we have, with $\lambda = \binom{n}{2}/365$,

$$P(Y \geq 1) \approx 1 - e^{-\lambda} \quad \text{and} \quad P(Y \geq 2) \approx 1 - e^{-\lambda} - \lambda e^{-\lambda}.$$ 

Note that with $n = 23$ we have

$$\lambda = 253/365, \quad P(Y \geq 1) \approx 0.500, \quad P(Y \geq 2) \approx 0.153.$$
11 Lecture 11

11.0.16 Review Exercises

1. If \( X \sim \text{Pois}(\lambda) \) for some \( \lambda > 0 \) is \( aX \) also Poisson for \( a > 0 \) with \( a \neq 1 \)?

2. Suppose the average number of car accidents at a given intersection is .7 per week. What is the chance there is more than 2 in a 4 week time span?

3. Coupon Collector: Each box of candy gives you 1 of 6 different toys, each being equally likely. On average how many boxes will you need to open to collect all 6 toys?

11.0.17 Solutions

1. No. If yes, then \( aX \) has expectation \( a\lambda \) so it would have to be a \( \text{Pois}(a\lambda) \) random variable with variance \( a\lambda \). But \( aX \) has variance \( a^2\lambda \), a contradiction.

2. Letting \( X \sim \text{Pois}(2.8) \) the answer is

\[
1 - p_X(0) - p_X(1) - p_X(2) = 1 - e^{-2.8} \left( 1 + 2.8 + \frac{2.8^2}{2} \right) \approx 0.5305.
\]

3. Let \( X_i \) be the number of boxes between the \((i-1)\)th distinct toy and the \(i\)th. Then

\[ X_k \sim \text{Geom} \left( \frac{6 - k + 1}{6} \right) \]

so the total expectation is, by linearity,

\[
\sum_{k=1}^{6} E[X_k] = \sum_{k=1}^{6} \frac{6}{7-k} = 14.7.
\]

11.0.18 Interesting Problems With Discrete Random Variables

1. Children Until Girl: Suppose a country has a policy where you may keep having children until you have a girl, but no more after your first girl. Assuming girls and boys are equally likely as children, what do you expect the ratio of girls to total children to be like after 1000000 children are born?

2. St. Petersburg Paradox: Suppose you keep flipping a fair coin until you get a heads. Your payoff is \( 2^{\text{number of flips}} \). How much would you pay to play this game?

3. Tagging Animals: You are studying dear in a particular region and want to know roughly how many there. You decide to uniformly draw 100 dear from the population, and tag them. You release them back into the population, and then uniformly draw 100 more dear. If 5 of the new dear are tagged, what is the number of total dear that yields the highest probability of seeing 5?
11.0.19 Solution

1. We expect the number of girls to be 500000 by linearity of expectation, so the ratio is expected to by 1/2.

2. If we compute the expectation, we get

\[ \sum_{k=1}^{\infty} \frac{2^k}{2^k} = \infty. \]

To come up with a reasonable estimate of what the game is worth, we assume that there is some amount of money that if won, the game runner cannot pay. For instance, if the game runner will flee the country if you win over 1000000 dollars then the expectation is

\[ \sum_{k=1}^{19} \frac{2^k}{2^k} = 19 = \lfloor \log_2(1000000 - 1) \rfloor. \]

3. If there are \( N \) total dear, \( T \) tagged dear, and we draw \( D \) dear, the number \( X \) of tagged dear has a HGeom\((D, T, N)\) distribution. Thus

\[ P(X = k) = \binom{T}{k} \frac{(N-T)}{D-k} \binom{N}{D-k} \binom{N-1-T}{D-k}. \]

Denote this value as \( f(N) \). To maximize \( N \) we look at the quantity \( f(N)/f(N - 1) \) giving

\[ \frac{f(N)}{f(N - 1)} = \frac{(N-T)}{D-k} \frac{(N-1)}{D} \binom{N}{D} \binom{N-1-T}{D-k} = \frac{N-D}{N} \cdot \frac{N-T}{N-T-D+k}. \]

This is at least 1 iff

\[ (N - D)(N - T) \geq N(N - T - D + k) \iff DT \geq Nk \iff \frac{DT}{k} \geq N. \]

Thus the answer is

\[ N = \left\lfloor \frac{DT}{k} \right\rfloor. \]

You can think of this as \( D/k \) being the fraction of dear we drew that were tagged, so if this fraction is representative of the full population, there should be \( D/k \cdot T \) total dear. In our example, this would be \( N = \lceil 10000/5 \rceil = 2000 \). Note, this isn’t saying what the size of the population is. This is saying that a choice of \( \lfloor DT/k \rfloor \) for the size of the population maximizes the probability that we get the given number of tagged dear \( k \). This is called a maximum likelihood estimator.
11.0.20 Meaning of Standard Deviation and Variance

For each of the discrete distributions we analyzed we computed the expectation and variance. Intuitively speaking, the expectation is the average value of the random variable and variance is how dispersed the values are about the average. We also saw that the standard deviation may be a more intuitive quantity than the variance, since it has the right units. Now we want to deepen our understanding of the standard deviation. We start with a simple theorem that is incredibly powerful.

**Theorem 47** (Markov’s Inequality). Let $X$ be a non-negative random variable whose expectation exists. Then for any $a \geq 0$

$$E[X] \geq a \cdot P(X \geq a).$$

**Proof.** We prove this here for discrete random variables (and later in the absolutely continuous setting). Note that

$$E[X] = \sum_{x:p_X(x) > 0, x < a} xp_X(x) + \sum_{x:p_X(x) > 0, x \geq a} xp_X(x) \geq 0 + \sum_{x:p_X(x) > 0, x \geq a} ap_X(x) = aP(X \geq a).$$

I think part of the mystery of Markov’s inequality is that it is hard to see how useful such a crude bound could be. For any random variable $X$ note that $Y = (X - E[X])^2$ is a non-negative random variable. Applying Markov’s inequality we obtain:

**Corollary 48** (Chebyshev’s Inequality). Let $X$ be a random variable whose variance exists. Then for any $a \geq 0$

$$\text{Var}(X) \geq a^2 \cdot P(|X - E[X]| \geq a).$$

Specifically, if we let $a = b\text{SD}(X) = k\sqrt{\text{Var}(X)}$ for some $k \geq 0$ we have

$$\frac{1}{k^2} \geq P(|X - E[X]| \geq k\text{SD}(X)).$$

Often people use $\mu_X$ to denote $E[X]$ and $\sigma_X$ to denote SD($X$) giving

$$P(|X - \mu_X| \geq k\sigma_X) \leq \frac{1}{k^2}.$$ 

Here $k$ is a non-negative real number (not necessarily an integer).

**Proof.** Let $Y = (X - \mu_X)^2$ so that $Y$ is non-negative and $E[Y] = \text{Var}(X)$. Applying Markov’s inequality gives

$$\text{Var}(X) \geq a^2 P((X - \mu_X)^2 \geq a^2)$$

for any $a \geq 0$. Note that $(X - \mu_X)^2 \geq a^2$ is equivalent to saying $|X - \mu_X| \geq a$ giving

$$\text{Var}(X) \geq a^2 P(|X - \mu_X| \geq a).$$

$\square$
Recall expressions like $|x - 9| \geq 4$ or $|x - 3| \leq 2$ mean that $x$ is within 4 units of 9, and $x$ is at least 3 units from 2, respectively. Now we can see what the standard deviation means. The probability that a random variable is more than $k$ standard deviations away from the mean is at most $\frac{1}{k^2}$. We will see that for many important random variables, the decay is much faster (i.e., much smaller than $\frac{1}{k^2}$). We will return to this topic later. Standard deviation also provides a scale invariant quantity that can be used to understand the RV better. Note that

$$SD(aX) = |a| \cdot SD(X)$$

for all $a \in \mathbb{R}$. Thus the standard deviation grows and shrinks accordingly as we scale a random variable. As such, the standard deviation is a good unit for measuring deviations from the mean.

11.0.21 Meaning of Standard Deviation Exercises

1. Suppose $X \sim \text{Bin}(100, .5)$.
   
   (a) Compute $P(|X - \mu_X| \geq 2\sigma_X)$ directly.
   
   (b) What does the Chebyshev bound give?

2. Let $X_1, \ldots, X_n$ be i.i.d. RVs (called a random sample) with $E[X_i] = \mu$ and $\text{Var}[X_i] = \sigma^2$. Define $\overline{X}_n$ (the sample mean) by

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

   (a) What is $E[\overline{X}_n]$?
   
   (b) What is $\text{Var}[\overline{X}_n]$?
   
   (c) What is $\text{SD}[\overline{X}_n]$?
   
   (d) How large must $n$ be to guarantee that $P(|\overline{X}_n - \mu| < .01) \geq .98$?

3. Suppose $X_1, \ldots, X_n$ form a random sample from a $\text{Ber}(p)$ distribution. Using Chebyshev, how large must $n$ be to guarantee that

$$P(|\overline{X}_n - p| < .01) \geq .98?$$

74
11.0.22 Solutions

1. (a) We have $\sigma_X = \sqrt{100(0.5)(0.5)} = 5$ so we must compute

$$P(|X - 50| \geq 10) = 1 - \sum_{k=41}^{59} P(X = k) = 1 - \sum_{k=41}^{59} \binom{100}{k} (0.5)^{100} \approx 0.0568879.$$

(b) $1/4$

2. (a) By linearity of expectation,

$$E[\bar{X}_n] = E \left[ \frac{1}{n} \sum_{i=1}^{n} X_i \right] = \frac{1}{n} \sum_{i=1}^{n} E[X_i] = \mu.$$

(b) By independence we get

$$\text{Var}[\bar{X}_n] = \text{Var} \left[ \frac{1}{n} \sum_{i=1}^{n} X_i \right] = \frac{\sigma^2}{n}.$$

(c)

$$\text{SD}[\bar{X}_n] = \frac{\sigma}{\sqrt{n}}.$$

It is important to note that $\sqrt{n}$ doesn’t grow that fast. Thus you need large samples to get a reasonable reduction in the standard deviation.

(d) By Chebyshev’s inequality,

$$P(|\bar{X}_n - \mu| \geq .01) \leq \frac{\text{Var}(\bar{X}_n)}{.01^2}$$

so we need

$$\frac{10000\sigma^2}{n} \leq .02 \iff n \geq 500000\sigma^2.$$

Part of the reason this bound is so large is that Chebyshev’s inequality assumes nothing about the distribution other than its mean and variance. As we saw in the first question, it can often being improved dramatically with more distributional information.

3. Here $\sigma^2 = p(1 - p)$ so using the previous problem we need $n \geq 500000p(1 - p)$. This is actually a huge overestimate. We will get a much lower approximate answer later using the central limit theorem.
11.1 Continuous Random Variables

In this section we will focus on absolutely continuous random variables. When studying discrete random variables we had probability mass functions. Using the physics analogy, we had point masses at different potential values of our random variable, with masses that sum to 1. The expectation was the center of mass. Now we move away from point masses to continuous slabs of material, where we must talk about densities (mass per unit length) instead of point mass. We turn densities into masses by integrating. The expectation will still be the center of mass, but we will need an integral instead of a sum to compute it.

11.1.1 Single Variable Calculus Review

We recall a few results from Calculus 1. Let \( f : [a, b] \to \mathbb{R} \) be integrable. If \( f \geq 0 \) then \( \int_a^b f \) represents the area of the region beneath \( f \). For \( g : \mathbb{R} \to \mathbb{R} \) we define

\[
\int_{-\infty}^b g(x) \, dx = \lim_{a \to -\infty} \int_a^b g(x) \, dx
\]

if the limit exists. Note that if \( g \geq 0 \) then

\[
G(b) = \int_{-\infty}^b g(x) \, dx
\]

is an increasing function. The FTC connects the definite integral which computes area to the antiderivative.

**Theorem 49** (Version of FTC).

1. Assume \( f : \mathbb{R} \to \mathbb{R} \) is integrable, and define \( F \) by

\[
F(x) = \int_{-\infty}^x f(t) \, dt.
\]

Then \( F'(x) = f(x) \) where \( f \) is continuous.

2. Suppose \( f : (a, b) \to \mathbb{R} \) is continuous and there is a function \( F : [a, b] \to \mathbb{R} \) that is continuous with \( F'(x) = f(x) \) for all \( x \in (a, b) \). Then

\[
\int_a^b f(x) \, dx = F(b) - F(a).
\]

11.1.2 Calculus Review Exercises

1. Which is bigger, or are they equal:

\[
\int_1^{100} e^u \, du \quad \text{or} \quad \int_1^{10} e^{2x} \, dx?
\]
2. Let $F(x) = \int_{-\infty}^{x} f(t) \, dt = \sin(e^{x^2})$ for all $x \in \mathbb{R}$. What is $\int_{-5}^{47} f(t) \, dt$?

3. Let $f(x) = e^{-|x|}$.
   
   (a) What is $F(x) = \int_{-\infty}^{x} f(t) \, dt$?
   
   (b) What is $F'(x)$?

4. If $F(x) = \int_{0}^{3x^2-2x} e^t \, dt$ then what is $F''(x)$?

5. Define $f(x)$ by

$$f(x) = \begin{cases} 
0 & \text{if } x < 0, \\
1 & \text{if } 0 \leq x < 1, \\
2 + x & \text{if } 1 \leq x < 2, \\
0 & \text{if } x \geq 2.
\end{cases}$$

   (a) What is $F(x) = \int_{-\infty}^{x} f(t) \, dt$?

   (b) What is $F'(x)$?

11.1.3 Solutions

1. Using $u$-substitution we know that

$$\int_{1}^{10} e^{x^2} 2x \, dx = \int_{1}^{10} e^u \, du.$$ 

   Since $2x > 1$ we see that

$$\int_{1}^{10} e^u > \int_{1}^{10} e^{x^2} \, dx.$$ 

2. 

$$\int_{-5}^{47} f(t) \, dt = F(47) - F(-5) = \sin(e^{47^2}) - \sin(e^{(-5)^2}).$$

3. (a) If $x \leq 0$ it is

$$F(x) = \int_{-\infty}^{x} e^t \, dt = e^x.$$ 

For $x > 0$ it is

$$F(x) = 1 + \int_{0}^{x} e^{-t} \, dt = 1 + [-e^{-t}]_0^x = 2 - e^{-x}.$$ 

(b) $F'(x) = f(x) = e^{-|x|}$
4. Let \( G(x) \) be defined by

\[
G(x) = \int_0^x e^{t^2} \, dt.
\]

If \( h(x) = 3x^2 - 2x \) then \( F = G \circ h \). Thus by the chain rule we obtain

\[
F'(x) = G'(h(x))h'(x) = e^{(3x^2-2x)^2}(6x - 2).
\]

5. (a) We have

\[
F(x) = \begin{cases} 
0 & \text{if } x < 0, \\
x & \text{if } 0 \leq x < 1, \\
1 + 2x + x^2/2 - 5/2 & \text{if } 1 \leq x < 2, \\
11/2 & \text{if } x \geq 2.
\end{cases}
\]

(b) \( F'(x) = f(x) \) except for \( x = 0, 1, 2 \).
12 Lecture 12

12.1 Continuous Random Variables

12.1.1 Review Questions

1. Define what it means for functions $f : \mathbb{R} \to \mathbb{R}$ to be increasing. Do the same for decreasing.

2. Suppose $f : \mathbb{R} \to \mathbb{R}$ is one-to-one and onto (bijective). Furthermore, suppose $f$ is differentiable with $f'(x) \neq 0$ for all $x$. What is $\frac{df^{-1}}{dx}$?

3. (⋆) For simplicity, assume $f : [a, b] \to \mathbb{R}$ is non-negative, bounded, and integrable. Show that $F(x) = \int_a^x f(t) \, dt$ is continuous for $x \in [a, b]$.

12.1.2 Solutions

1. We say that $f$ is increasing if $f(a) \leq f(b)$ whenever $a \leq b$. We say that $f$ is decreasing if $f(a) \geq f(b)$ whenever $a \leq b$.

2. We have

$$\frac{d}{dx} f^{-1}(x) = \frac{1}{f'(f^{-1}(x))}.$$

3. Let $B$ bound $f$, i.e., $0 \leq f(x) \leq B$ for all $x \in [a, b]$. Note that

$$|F(x + h) - F(x)| = \left| \int_x^{x+h} f(t) \, dt \right| \leq \left| \int_x^{x+h} B \, dt \right| = Bh \to 0.$$

This result is still true as long as $f$ has a finite integral.

12.1.3 General Random Variables

Up till this point we have studied discrete random variables. Recall that $X$ is discrete if there is some countable sequence of values $x_i$ so that

$$\sum_i P(X = x_i) = 1.$$

We also discussed their distributions, which we described using PMFs. Briefly, we also mentioned the associated CDF $F_X$ defined by

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

Discrete random variables always had discontinuous CDFs (discontinuous but right continuous). We now study random variables with continuous CDFs called continuous random variables (we use these to model continuous quantities like time, temperature, parameters
of distributions, etc.). For these random variables if you fix any value \( x \in \mathbb{R} \) we always have \( P(X = x) = 0 \). This removes any jumps in the CDF. We will actually study an even smaller subclass of random variables called absolutely continuous random variables. Absolutely continuous means the random variable has an associated density \( f_X \) which we call a PDF (probability density function). The formal definition goes as follows: The random variable \( X \) has PDF \( f_X \) if the CDF can be written as

\[
F_X(x) = \int_{-\infty}^{x} f_X(x) \, dx.
\]

More generally we can write

\[
P(X \in A) = \int_A f_X(x) \, dx.
\]

In contrast to a PMF, the values of a PDF aren’t probabilities, but “probability per unit length”. If \( f_X(3) = 5 \) and \( f_X \) is continuous at 3 then

\[
P(X \in (3 - \epsilon/2, 3 + \epsilon/2)) \approx 5\epsilon,
\]

for small \( \epsilon > 0 \).

### 12.1.4 Continuous Random Variable Exercises

1. Let \( X \) be random variable with CDF \( F_X(x) \) defined by

\[
F_X(x) = \begin{cases} 
0 & \text{if } x \leq 0, \\
x & \text{if } 0 < x \leq 1, \\
1 & \text{if } x \geq 1.
\end{cases}
\]

(a) Is \( X \) a continuous random variable?

(b) Is \( X \) an absolutely continuous random variable? If so, give a density function \( f_X \).

(c) What is \( P(X = 1) \)?

(d) What is \( P(.25 < X \leq .75) \)?

2. Let \( X \) be a random variable with PDF \( f_X(x) \) defined by

\[
f_X(x) = \begin{cases} 
x & \text{if } 0 \leq x < 1 \\
x - 1 & \text{if } 1 \leq x < 2 \\
0 & \text{otherwise}.
\end{cases}
\]

(a) Is \( X \) a continuous random variable?

(b) What is \( P(0 < X \leq 1) \)?

(c) What is \( P(.5 < X \leq 1.5) \)?

(d) Give the CDF \( F_X \) and plot it.
3. Let $X$ be a random variable with CDF $F_X(x)$ defined by

$$F_X(x) = \begin{cases} 
0 & \text{if } x < -1, \\
1/3 & \text{if } -1 \leq x < 0, \\
1/3 + x/3 & \text{if } 0 \leq x < 1, \\
5/6 + (x - 1)^2/6 & \text{if } 1 \leq x < 2, \\
1 & \text{if } x \geq 2.
\end{cases}$$

(a) Is $X$ a continuous random variable?
(b) Is $X$ a discrete random variable?
(c) What is $P(X = -1)$? What is $P(X = 0)$? What is $P(X = 1)$?
(d) What is $P(-1 < X \leq 3/2)$? What is $P(-1 \leq X \leq 3/2)$?

4. Let $X$ be a random variable with PDF $f_X(x)$.

(a) What is $\int_{-\infty}^{\infty} f_X(x) \, dx$?
(b) What is $P(X = a)$ where $a \in \mathbb{R}$ is fixed?
(c) Is $X$ continuous?
(d) How do you compute $P(a < X \leq b)$ and $P(a \leq X \leq b)$ for some $a, b \in \mathbb{R}$ with $a < b$?
(e) $(\star)$ What is $P(X \in \mathbb{Q})$ where $\mathbb{Q}$ is the set of rational numbers?

12.1.5 Solutions

1. (a) Yes.
   (b) Yes, $f_X(x) = 1$ if $x \in [0, 1]$ and 0 otherwise. It turns out there are infinitely many $f_X$ that work, since we can modify $f_X$ on any finite set (actually any measure zero set) and it gives the same integral.
   (c) $P(X = 1) = 0$ since $X$ is continuous.
   (d) $P(.25 < X \leq .75) = F_X(.75) - F_X(.25) = .5$

2. (a) Yes. All absolutely continuous random variables are continuous (since integrals are continuous functions of their upper limit).
   (b) $$\int_{0}^{1} f_X(x) \, dx = \frac{1}{2}.$$
   (c) $$\int_{.5}^{1.5} f_X(x) \, dx = \int_{.5}^{1} x \, dx + \int_{1}^{1.5} x - 1 \, dx = \frac{1}{2}. $$
(d) The CDF is given by

\[ F_X(x) = \begin{cases} 
0 & \text{if } x \leq 0, \\
x^2/2 & \text{if } 0 \leq x < 1, \\
1/2 + (x - 1)^2/2 & \text{if } 1 \leq x < 2, \\
1 & \text{if } x \geq 2.
\end{cases} \]

3. (a) No, there are jumps at \( x = -1, 1 \).

(b) No. The jumps don’t add up to a probability of 1. Another way to see this is that the function isn’t always horizontal away from the jumps.

(c) \( P(X = -1) = 1/3, P(X = 0) = 0, P(X = 1) = 1/3 \).

(d)

\[ P(-1 < X \leq 3/2) = F_X(3/2) - F_X(-1) = \frac{5}{6} + \frac{1}{24} - \frac{1}{3} = \frac{13}{24} \]

and

\[ P(-1 \leq X \leq 3/2) = P(-1 < X \leq 3/2) + P(X = -1) = \frac{21}{24}. \]

4. (a)

\[ \int_{-\infty}^{\infty} f_X(x) \, dx = P(X \in \mathbb{R}) = 1. \]

(b)

\[ P(X = a) = \int_{a}^{a} f_X(x) \, dx = 0. \]

(c) Yes, by review exercise 2.

(d) Both are \( \int_{a}^{b} f_X(x) \, dx \), since \( P(X = a) = 0 \).

(e) Let \( q_1, q_2, \ldots \) be a listing of all rational numbers. We know \( P(X = q_i) = 0 \), so by countable additivity,

\[ P(X \in \mathbb{Q}) = \sum_{i=1}^{\infty} P(X = q_i) = 0. \]
13 Lecture 13

13.1 Continuous Random Variables

13.1.1 Review Exercises

1. Compute
\[
\int_0^1 x \sin(x) \, dx.
\]

2. (Extra Credit from Midterm) Suppose you have an urn with \(B\) black balls and \(R\) red balls. You keep drawing balls without replacement until you get \(b\) black balls (where \(1 \leq b \leq B\)). Compute the expected number of draws required. Partial credit for a finite summation, full credit for a concise expression. [Hint: For the concise expression, use linearity of expectation to compute the number of red balls drawn.]

13.1.2 Solutions

1. Integrating by parts,
\[
\int_0^1 x \sin(x) \, dx = [-x \cos(x)]_0^1 - \int_0^1 -\cos(x) \, dx = -\cos(1) + [\sin(x)]_0^1 = \sin(1) - \cos(1).
\]

2. Let \(X\) denote the number of draws. For the partial credit solution, we can compute the probability that drawing \(b\) balls will require \(k\) turns. This is given by
\[
P(X = k) = \frac{\binom{B}{b} \binom{R}{k-b} b(k-1)!}{(R+B)(R+B-1) \cdots (R+B-k+1)} = \frac{b \binom{B}{b} \binom{R}{k-b}}{k \binom{R+B}{k}}.
\]
This gives
\[
E[X] = \sum_{k=b}^{R+b} b \frac{\binom{B}{b} \binom{R}{k-b}}{\binom{R+B}{k}}.
\]

For the full credit solution, let \(Y_i\) denote the indicator of whether the \(i\)th red ball is used. Then \(Z = Y_1 + \cdots + Y_R\) is the total number of red balls used. We now compute \(E[Y_i]\). Assume after you draw the \(b\)th black ball, you continue drawing just to see what you would have gotten. The probability that the \(i\)th red ball is drawn before \(b\) black balls is \(\frac{b}{B+1}\) by looking at their relative orderings. Thus the answer is
\[
E[X] = E[Z + b] = \frac{Rb}{B+1} + b = \frac{b(R+B+1)}{B+1}.
\]

This is also \(b\) times the expected number of draws to get 1 black ball (which it must be by linearity).
13.1.3 Functions of an Absolutely Continuous Random Variable

Just as with discrete random variables, we want to compose functions with our absolutely continuous random variables. As an example, suppose \( X \) has PDF \( f_X(x) = 1 \) for \( x \in [0, 1] \) and 0 otherwise (we will call this a uniform random variable on \([0, 1]\)). Then we can form random variables like \( X^2 \), \( 2X \), and \( e^X \) by composing functions. Remember as always that a random variable is a function from a sample space to \( \mathbb{R} \) (for continuous random variables, the sample spaces need to be more complex though). It may make sense for you to review these concepts since they can get jumbled up:

<table>
<thead>
<tr>
<th>Discrete</th>
<th>Absolutely Continuous</th>
</tr>
</thead>
<tbody>
<tr>
<td>( X : S \to \mathbb{R} )</td>
<td>( X : S \to \mathbb{R} )</td>
</tr>
<tr>
<td>( F_X(x) = P(X \leq x) )</td>
<td>( F_X(x) = P(X \leq x) )</td>
</tr>
<tr>
<td>PMF: ( p_X : \mathbb{R} \to \mathbb{R}_{\geq 0} )</td>
<td>PDF: ( f_X : \mathbb{R} \to \mathbb{R}_{\geq 0} )</td>
</tr>
<tr>
<td>( P(X \in A) = \sum_{x \in A} p_X(x) )</td>
<td>( E[X] = \int_{-\infty}^{\infty} x f_X(x) , dx )</td>
</tr>
</tbody>
</table>

We will learn about expectations in the next section. Going back to functions, if \( X : S \to \mathbb{R} \) is a random variable then we can take a function \( g : \mathbb{R} \to \mathbb{R} \) and compose the two to obtain a new random variable \( g(X) = g \circ X \).

13.1.4 Functions of an RV Exercises

1. Suppose \( X \) has PDF \( f_X(x) = 1 \) for \( x \in [0, 1] \) and 0 otherwise.
   (a) What is the CDF for \( X? \)
   (b) What is the CDF for \( X^2? \)
   (c) What is the PDF for \( X^2? \)
   (d) What is the PDF for \( e^X? \)

2. Suppose \( X \) has PDF \( f_X(x) = e^{-x} \) for \( x \geq 0 \) and 0 otherwise.
   (a) What is the PDF for \( X^2? \)
   (b) What is the PDF for \(-X^2? \)
   (c) What is the PDF for \( f(X) \) defined by \( f(x) = \begin{cases} \log(x) & \text{if } x > 0, \\ 0 & \text{otherwise.} \end{cases} \)

3. Suppose \( X \) has PDF \( f_X(x) = 1/4 \) for \( x \in [-2, 2] \) and 0 otherwise. What is the PDF of \( X^2? \)

4. (*) Let \( X \) have PDF \( f_X \) with \( P(a < X < b) = 1 \) for some (possibly infinite) interval \((a, b)\). Let \( g \) be a strictly monotonic function with \( g' \neq 0 \) (increasing or decreasing) defined on \((a, b)\). Give the PDF for \( g(X) \).
13.1.5 Solutions

1. (a) Integrating gives

\[ F_X(x) = \begin{cases} 
0 & \text{if } x < 0, \\
 x & \text{if } 0 \leq x < 1, \\
1 & \text{if } x \geq 1.
\end{cases} \]

(b) Let \( Y = X^2 \). For \( y < 0 \) we see \( F_Y(y) = 0 \). Otherwise we have

\[ F_Y(y) = P(Y \leq y) = P(X^2 \leq y) = P(|X| \leq \sqrt{y}) = P(X \leq \sqrt{y}) = F_X(\sqrt{y}). \]

(c) Differentiating we get, for \( y \geq 0 \),

\[ f_Y(y) = \frac{d}{dy} F_X(\sqrt{y}) = \frac{f_X(\sqrt{y})}{2\sqrt{y}} = \frac{1}{2\sqrt{y}} 1_{[0,1]}(y). \]

Here \( 1_A(y) \) is the function that is 1 for \( y \in A \) and 0 otherwise. We can thus write

\[ f_Y(y) = \begin{cases} 
0 & \text{if } y < 0, \\
 \frac{1}{2\sqrt{y}} & \text{if } 0 \leq y < 1, \\
0 & \text{if } y \geq 1.
\end{cases} \]

(d) Letting \( Z = e^X \) we have, for \( z > 0 \),

\[ F_Z(z) = P(Z \leq z) = P(e^X \leq z) = P(X \leq \log(z)) = F_X(\log(z)). \]

Differentiating gives

\[ f_Z(z) = \frac{1}{z} \frac{1}{2\sqrt{y}} 1_{[0,1]}(\log(z)) \]

for \( z > 0 \), and 0 otherwise. Written differently,

\[ f_Z(z) = \begin{cases} 
0 & \text{if } z \leq 1, \\
 \frac{1}{z} & \text{if } 1 < z \leq e, \\
0 & \text{if } z > e.
\end{cases} \]

2. (a) Letting \( Y = X^2 \) we have, for \( y \geq 0 \),

\[ F_Y(y) = P(X^2 \leq y) = P(X \leq \sqrt{y}) = F_X(\sqrt{y}). \]

Differentiating we obtain, for \( y \geq 0 \),

\[ f_Y(y) = \frac{f_X(\sqrt{y})}{2\sqrt{y}} = \frac{e^{-\sqrt{y}}}{2\sqrt{y}} \]

and 0 otherwise.

(b) We solve this in two ways. Let \( Z = X^2 \).
i. Proceeding as before gives, for $z \leq 0$,

$$F_Z(z) = P(-X^2 \leq z) = P(X^2 \geq -z) = P(X \geq \sqrt{-z}) = 1 - F_X(\sqrt{-z}).$$

Differentiating we obtain, for $z \leq 0$,

$$f_Z(z) = -f_X(\sqrt{-z}) \left( -\frac{1}{2\sqrt{-z}} \right) = \frac{e^{-\sqrt{-z}}}{2\sqrt{-z}},$$

and 0 otherwise. To see the connection with problem 4, note that we obtained:

$$f_Z(z) = f_X(\sqrt{-z}) \left| -\frac{1}{2\sqrt{-z}} \right|.$$

ii. We use the solution to (b) to solve (c) which simply reflects the PDF around the vertical axis getting the same answer as above.

(c) Let $W = f(X)$ giving

$$F_W(w) = P(f(X) \leq w) = P(X \leq e^w) = F_X(e^w).$$

Differentiating gives

$$f_W(w) = f_X(e^w)e^w = e^{-w}e^w.$$

Unlike before, this PDF is non-zero on all of $\mathbb{R}$.

3. Let $Y = X^2$ so that, for $y \geq 0$,

$$F_Y(y) = P(X^2 \leq y) = P(|X| \leq \sqrt{y}) = P(-\sqrt{y} \leq X \leq \sqrt{y}) = F_X(\sqrt{y}) - F_X(-\sqrt{y}).$$

Differentiating gives, for $y \geq 0$,

$$f_Y(y) = \frac{f_X(\sqrt{y})}{2\sqrt{y}} - \frac{f_X(-\sqrt{y})}{-2\sqrt{y}} = \frac{1}{4\sqrt{y}} \mathbb{1}_{[0,4]}(y).$$

4. Suppose the image of $(a,b)$ under $g$ is $(\alpha, \beta)$. We must have $g' > 0$ or $g' < 0$ everywhere. If $g' > 0$ then $g$ is increasing. Then, for $y \in (\alpha, \beta)$ we have

$$F_{g(X)}(y) = P(g(X) \leq y) = P(X \leq g^{-1}(y)) = F_X(g^{-1}(y))$$

so differentiating gives

$$f_{g(X)}(y) = f_X(g^{-1}(y)) \cdot \frac{1}{g'(g^{-1}(y))}.$$ 

If $g' < 0$ then $g$ is decreasing and we have, for $y \in (\alpha, \beta),$

$$F_{g(X)}(y) = P(g(X) \leq y) = P(g^{-1}(y) \leq X) = 1 - F_X(g^{-1}(y)).$$
Differentiating gives
\[ f_g(x)(y) = -f_X(g^{-1}(y)) \cdot \frac{1}{g'(g^{-1}(y))}. \]
Since \( g' \) is negative, this can be written as
\[ f_g(x)(y) = f_X(g^{-1}(y)) \cdot \left| \frac{1}{g'(g^{-1}(y))} \right|, \]
which works for both \( g' > 0 \) and \( g' < 0 \). Thus we have proven:

**Theorem 50.** Let \( X \) have PDF \( f_X \) with \( P(a < X < b) = 1 \) for some (possibly infinite) interval \((a,b)\). Let \( g \) be a strictly monotonic function with \( g' \neq 0 \) (increasing or decreasing) defined on \((a,b)\). Suppose that \( g \) takes values in \((\alpha,\beta)\). Then \( Y = g(X) \) has PDF given by
\[ f_Y(y) = f_X(g^{-1}(y)) \left| \frac{1}{g'(g^{-1}(y))} \right|, \]
for \( y \in (\alpha,\beta) \), and 0 otherwise.

If we specialize this to linear functions we obtain:

**Corollary 51.** Let \( X \sim f_X \) and let \( g(x) = ax + b \) with \( a \neq 0 \). Then if \( Y = g(X) \) we have
\[ f_Y(y) = f_X \left( \frac{y-b}{a} \right) \left| \frac{1}{a} \right|. \]

### 13.1.6 Expectation and Variance

Let \( X \) be a random variable with PDF \( f_X \). Then we define the expectation by
\[ E[X] = \int_{-\infty}^{\infty} x f_X(x) \, dx. \]
We say the expectation exists if
\[ \int_{-\infty}^{\infty} |x| f_X(x) \, dx \]
is finite (analogous to absolute convergence from the discrete case). Here we have replaced the “discrete” summation with the “continuous” integral. The definition of variance actually goes unchanged:
\[ \text{Var}[X] = E[(X - E[X])^2] = E[X^2] - E[X]^2. \]
The only difference is that when \( X \) has a PDF, we must compute the expectation using an integral. Just as in the discrete case we have that expectation is linear, but we wont be able to prove this until later when we define joint PDFs. We also still have LOTUS, which we prove now:
**Theorem 52** (LOTUS). Let $X \sim f_X(x)$ and let $g : \mathbb{R} \to \mathbb{R}$. Then

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

*Proof.* We only give the proof in the case that $g$ satisfies the constraints of Exercise 4 from the previous set. The more general result is true, and is proven when studying measure theoretic probability. As such, suppose $g' > 0$ or $g' < 0$, that $X$ takes values in some interval $(a, b)$, and that $g(X)$ takes values in some interval $(\alpha, \beta)$. Recall that

$$f_Y(y) = f_X(g^{-1}(y)) \cdot \left| \frac{1}{g'(g^{-1}(y))} \right|.$$  

Using this we have

$$\int_{-\infty}^{\infty} g(x)f_X(x) \, dx = \int_{a}^{b} g(x)f_X(x) \, dx = \int_{\alpha}^{\beta} u f_X(g^{-1}(u)) \left| \frac{1}{g'(g^{-1}(u))} \right| \, du$$

by doing the substitution $u = g(x)$. We needed the absolutely values, since if $g^{-1}$ is decreasing, the limits of integration would be swapped. But this is equal to

$$E[Y] = \int_{\alpha}^{\beta} y f_Y(y) \, dy = \int_{-\infty}^{\infty} y f_Y(y) \, dy$$

completing the proof.  

\[\square\]

### 13.1.7 Expectation Exercises

1. Let $X \sim f_X$ where $f_X(x) = 1/(b - a)$ for $x \in [a, b]$ and 0 elsewhere. Here $b > a$ are fixed constants.
   
   (a) What is $E[X]$?
   
   (b) What is $\text{Var}[X]$?
   
   (c) Let $t \in \mathbb{R}$ be fixed. Compute $E[e^{tX}]$.

2. Let $X \sim f_X$ where $f_X(x) = \alpha e^{-\alpha x}$ for $x \geq 0$ and zero elsewhere. Here $\alpha > 0$ is a fixed constant.
   
   (a) What is $E[X]$?
   
   (b) What is $\text{Var}[X]$?
   
   (c) Determine for which $t \in \mathbb{R}$ the value $E[e^{tX}]$ exists, and compute this value.

3. Let $X \sim f_X$ where $f_X(x) = \frac{\theta}{x^2}$ for $x \geq 1$ and 0 elsewhere.
   
   (a) What is $C$?
   
   (b) What is $E[X]$?
13.1.8 Solutions

1. (a) 
   \[ E[X] = \int_{-\infty}^{\infty} x f_X(x) \, dx = \int_{a}^{b} \frac{x}{b-a} \, dx = \frac{b^2 - a^2}{2(b-a)} = \frac{b + a}{2}. \]

(b) 
   \[ E[X^2] = \int_{-\infty}^{\infty} x^2 f_X(x) \, dx = \int_{a}^{b} \frac{x^2}{b-a} \, dx = \frac{b^3 - a^3}{3(b-a)} = \frac{a^2 + ab + b^2}{3}. \]
   Thus we have 
   \[ \text{Var}[X] = E[X^2] - E[X]^2 = \frac{a^2 + ab + b^2}{3} - \frac{a^2 + 2ab + b^2}{4} - \frac{a^2 + 2ab + b^2}{12} = \frac{(a - b)^2}{12}. \]

(c) 
   \[ E[e^{tx}] = \int_{-\infty}^{\infty} e^{tx} f_X(x) \, dx = \int_{a}^{b} \frac{e^{tx}}{b-a} \, dx = \frac{e^{tb} - e^{ta}}{t(b-a)}. \]
   If we look at this as a function of \( t \), it is called the Moment Generating Function of \( X \). We will learn more about them later in the course. This is also called the Laplace transform of \( f_X \).

2. (a) Integrating by parts, 
   \[ E[X] = \int_{0}^{\infty} x e^{-\alpha x} \, dx = \left[ \frac{\alpha e^{-\alpha x} x}{-\alpha} \right]_{0}^{\infty} - \frac{1}{-\alpha} \int_{0}^{\infty} \alpha e^{-\alpha x} \, dx = 0 + \frac{1}{\alpha} = \frac{1}{\alpha}. \]

(b) Integrating by parts, 
   \[ E[X^2] = \int_{0}^{\infty} x^2 e^{-\alpha x} \, dx = \left[ \frac{\alpha x^2 e^{-\alpha x}}{-\alpha} \right]_{0}^{\infty} - \frac{1}{-\alpha} \int_{0}^{\infty} 2\alpha xe^{-\alpha x} \, dx = 0 + \frac{2}{\alpha^2} = \frac{2}{\alpha^2}. \]
   Thus we have 
   \[ \text{Var}[X] = E[X^2] - E[X]^2 = \frac{2}{\alpha^2} - \frac{1}{\alpha^2} = \frac{1}{\alpha^2}. \]

(c) We assume \( t < \alpha \) so the integral exists: 
   \[ E[e^{tx}] = \int_{0}^{\infty} e^{tx} e^{-\alpha x} \, dx = \left[ \frac{\alpha e^{(t-\alpha)x}}{t-\alpha} \right]_{0}^{\infty} = \frac{\alpha}{\alpha - t}. \]

3. (a) We have 
   \[ 1 = \int_{1}^{\infty} Cx^{-2} \, dx = \left[ \frac{C x^{-1}}{-1} \right]_{1}^{\infty} = C \]
   so \( C = 1 \).

(b) Note that 
   \[ \int_{1}^{\infty} x f_X(x) \, dx = \int_{1}^{\infty} \frac{1}{x} \, dx = \left[ \log(x) \right]_{1}^{\infty} = \infty \]
   so the expectation doesn’t exist.
13.2 Examples of Absolutely Continuous Random Variables

13.2.1 Uniform and Exponential Random Variables

We say that $X$ has a (continuous) uniform distribution on $[a, b]$ and write $X \sim \text{Unif}(a, b)$ if it has the PDF

$$f_X(x) = \begin{cases} \frac{1}{b-a} & \text{if } a \leq x \leq b, \\
0 & \text{otherwise.} \end{cases}$$

We saw earlier that this distribution has

$$E[X] = \frac{a + b}{2} \quad \text{and} \quad \text{Var}[X] = \frac{(b - a)^2}{12}.$$ 

We say that $X$ has an exponential distribution and write $X \sim \text{Exp}(\lambda)$ (with $\lambda > 0$) if $X$ has PDF

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x > 0, \\
0 & \text{otherwise.} \end{cases}$$

This has

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

Like the geometric random variable, the exponential distribution has a memoryless property:

**Theorem 53 (Memoryless).** If $X \sim \text{Exp}(\lambda)$ then

$$P(X > s + t | X > s) = P(X > t)$$

for any $s, t \geq 0$.

If we think of $X$ as being the lifetime of a lightbulb, it says knowing that $X$ has lived at least $t$ hours doesn’t give us information about how much longer it will last.
14 Lecture 14

14.1 Examples of Absolutely Continuous Random Variables

14.1.1 Uniform and Exponential Review Exercises

1. Let $X$ be a continuous random variable with CDF $F_X$ and assume that $F_X$ is strictly increasing. Let $g(x) = F_X(x)$. What is the distribution of $g(X)$?

2. The median of a random variable is a value $x$ for which
   
   $$P(X \geq x) \geq 1/2 \quad \text{and} \quad P(X \leq x) \geq 1/2.$$

   For absolutely continuous random variables, this simplifies to a value $x$ for which $P(X \leq x) = 1/2$. Compute the median of $X$ when
   
   (a) $X \sim \text{Unif}(a,b),$
   (b) $X \sim \text{Exp}(\lambda)$.

3. Prove that the exponential distribution has the memoryless property:

   **Theorem 54 (Memoryless).** If $X \sim \text{Exp}(\lambda)$ then
   
   $$P(X > s + t | X > s) = P(X > t)$$

   for any $s, t \geq 0$.

4. A random variable $X$ has a Weibull distribution with parameters $a > 0$ and $b > 0$ if $X$ has PDF $f_X$ with

   $$f_X(x) = \begin{cases} 
   \frac{b}{a^b}x^{b-1}e^{-(x/a)^b} & \text{for } x > 0, \\
   0 & \text{for } x \leq 0. 
   \end{cases}$$

   Show that if $X^b$ has an exponential distribution, and determine the parameter.

14.1.2 Solutions

1. Let $Y = g(X)$ so that, for $y \in (0,1),$

   $$F_Y(y) = P(Y \leq y) = P(F_X(X) \leq y) = P(X \leq F_X^{-1}(y)) = F_X(F_X^{-1}(y)) = y.$$ 

   Thus $Y \sim \text{Unif}(0,1)$. This is an important result in simulation, since it shows that if $Y \sim \text{Unif}(0,1)$ and $F$ is some CDF then $F^{-1}(Y)$ is distributed according to $F$. We emphasize this in a theorem:

   **Theorem 55.** Let $F$ be a strictly increasing CDF of some distribution, and let $X \sim \text{Unif}(0,1)$. Then $Y = F^{-1}(X)$ has CDF $F$. 

91
This is true more generally than strictly increasing CDFs, but we won’t cover that.

2. (a) We must find \( m \) such that
\[
\frac{1}{2} = \int_{-\infty}^{m} f_X(x) \, dx.
\]
Looking for \( m \) in \((a, b)\) we have
\[
\frac{1}{2} = \int_{a}^{m} f_X(x) \, dx = \frac{m - a}{b - a} \iff m = \frac{b + a}{2}.
\]
(b) Looking for \( m \geq 0 \) we compute
\[
\frac{1}{2} = \int_{0}^{m} \lambda e^{-\lambda x} \, dx = 1 - e^{-\lambda m} \iff e^{-\lambda m} = \frac{1}{2} \iff \lambda m = \log(2) \iff m = \frac{\log(2)}{\lambda}.
\]

3. We compute
\[
P(X > s + t | X > s) = \frac{P(X > s + t)}{P(X > s)} = \frac{e^{-\lambda(s+t)}}{e^{-\lambda s}} = e^{-\lambda t} = P(X > t),
\]
where we used that \( F_X(x) = 1 - e^{-\lambda x} \) for \( x \geq 0 \) above several times.

4. Let \( Y = X^b \) and \( g(y) = y^b \) giving, for \( y > 0 \),
\[
f_Y(y) = f_X(g^{-1}(y)) \left| \frac{1}{g'(g^{-1}(y))} \right| = f_X(y^{1/b}) \left| \frac{1}{by^{(b-1)/b}} \right| = \frac{by^{(b-1)/b}e^{-(y^{1/b}/a)^b}a^{b}by^{(b-1)/b}}{a^{b}} = e^{-y/a^b},
\]
the PDF for a \( \text{Exp}(a^{-b}) \) random variable.

14.1.3 Normal Distribution

A variable \( X \) has a normal (or Gaussian) distribution with parameters \( \mu \in \mathbb{R} \) and \( \sigma^2 > 0 \) if it has PDF \( f_X \) given by
\[
f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.
\]
We write \( X \sim \mathcal{N}(\mu, \sigma^2) \). We will see that
\[
E[X] = \mu \quad \text{and} \quad \text{Var}[X] = \sigma^2.
\]
When $\mu = 0$ and $\sigma^2 = 1$ we call it the standard normal distribution, with PDF

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2}}.$$

The normal distribution is the most important distribution for reasons we will see later. For now I want to highlight some of the interesting features of the PDF. Firstly, where does the $\sqrt{2\pi}$ come from? Kind of amazing that $\pi$ would play a role here. To see this we employ a famous trick that oddly involves squaring the quantity we wish to compute:

$$\left( \int_{-\infty}^{\infty} e^{-x^2/2} \, dx \right)^2 = \int_{-\infty}^{\infty} e^{-x^2/2} \, dx \int_{-\infty}^{\infty} e^{-y^2/2} \, dy$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-(x^2+y^2)/2} \, dx \, dy$$

$$= \int_{0}^{2\pi} \int_{0}^{\infty} e^{-r^2/2} \, dr \, d\theta$$

$$= 2\pi \int_{0}^{\infty} e^{-r^2/2} \, dr$$

$$= 2\pi \left[ -e^{-u} \right]_{0}^{\infty}$$

$$= 2\pi.$$

Taking square-roots we get our result for the standard normal. The second point is to understand how the standard normal relates to the general normal. As you will prove, we have the following result.

**Theorem 56.** Let $X \sim \mathcal{N}(0, 1)$ and let $Y = aX + b$ where $a \neq 0$. Then $Y \sim \mathcal{N}(b, a^2)$.

The standard normal PDF has a bell shape centered at 0 with rapidly decaying tails. To get the general normal PDF we simply shift and scale it appropriately. We can also use this backwards:

**Corollary 57 (Standardization).** If $Y \sim \mathcal{N}(\mu, \sigma^2)$ then $\frac{Y-\mu}{\sigma} \sim \mathcal{N}(0, 1)$.

Due to this corollary, if you only know the standard normal CDF (often denoted $\Phi$), you can figure out the general normal CDF.

**Example 58.** Suppose $X \sim \mathcal{N}(\mu, \sigma^2)$. What is $P(X \leq x)$? Note that

$$P(X \leq x) = P\left( \frac{X - \mu}{\sigma} \leq \frac{x - \mu}{\sigma} \right) = \Phi \left( \frac{x - \mu}{\sigma} \right).$$

The reason the normal distribution is so important is the central limit theorem, which we will prove later. It basically says that sums of i.i.d. random variables look normal, regardless of the distribution (as long as it has a mean and variance).

Here are some other properties of the normal distribution. Assume $X \sim \mathcal{N}(\mu, \sigma^2)$. 93
1. $P(|X - \mu| < \sigma) \approx .68$, $P(|X - \mu| < 2\sigma) \approx .95$, and $P(|X - \mu| < 3\sigma) \approx .997$.

2. The normal distribution is symmetric about its mean (i.e., the PDF has $f_X(\mu - x) = f_X(\mu + x)$).

### 14.1.4 Normal Exercises

1. An expert witness in a paternity suit testifies that gestation time (in days) is approximately normally distributed with $\mu = 270$ and $\sigma^2 = 100$.
   
   (a) If the defendant was outside of the country starting 290 days and ending 240 days before the child’s birth, what is the probability he is the father?
   
   (b) What is the probability a child will gestate for more than 300 days?

2. Show that if $X \sim \mathcal{N}(0, 1)$ and $Y = aX + b$ with $a \neq 0$ then $Y \sim \mathcal{N}(b, a^2)$.

3. Compute $E[X]$ and $\text{Var}[X]$ for a standard normal random variable. What about for a general normal random variable (use previous part)?

4. Let $f(x) = ax + b$ where $a \neq 0$.
   
   (a) If $X \sim \text{Unif}(c, d)$ is $f(X)$ uniformly distributed? If yes, what are its parameters?
   
   (b) If $X \sim \text{Exp}(\lambda)$ what are $E[f(X)]$ and $\text{Var}[f(X)]$? Is $f(X)$ exponentially distributed? If yes, what is $\lambda$?

### 14.1.5 Solutions

1. (a) The probability is approximately $1 - .95 - .0235 = 1 - .9735 = .0265$.
   
   (b) The probability is approximately .0015.

2. We have

$$f_Y(y) = f_X \left( \frac{y - b}{a} \right) \left| \frac{1}{a} \right| = \frac{1}{\sqrt{2\pi a^2}} e^{-\frac{1}{2} \left( \frac{y - b}{a} \right)^2}.$$  

3. Let $X \sim \mathcal{N}(0, 1)$. Then we have

$$E[X] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} xe^{-x^2/2} \, dx = 0$$

since the integrand is odd. Continuing, we have

$$E[X^2] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^2 e^{-x^2/2} \, dx$$

$$= \left[ -\frac{1}{\sqrt{2\pi}} xe^{-x^2/2} \right]_{-\infty}^{\infty} - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} -e^{-x^2/2} \, dx$$

$$= 1.$$
Since $E[X] = 0$ we see $\text{Var}[X] = E[X^2] = 1$ and we are done. By the previous part, we see that if $Y \sim \mathcal{N}(\mu, \sigma^2)$ then $E[Y] = \mu$ and $\text{Var}[Y] = \sigma^2$.

4. (a) As usual, we compute the PDF. Letting $Y = f(X)$ we have

$$f_Y(y) = f_X \left( \frac{y - b}{a} \right) \frac{1}{\left| \frac{1}{a} \right|} \frac{1}{|c|d} \left( \frac{y-b}{a} \right).$$

If $a > 0$ we have

$$c \leq \frac{y - b}{a} \leq d \iff ac + b \leq y \leq ad + b.$$ 

If $a < 0$ we have

$$c \leq \frac{y - b}{a} \leq d \iff ad + b \leq y \leq ac + b.$$ 

In either case, it is a uniform distribution.

(b) Firstly, we have

$$E[aX + b] = aE[X] + b = \frac{a}{\lambda} + b \quad \text{and} \quad \text{Var}(aX + b) = a^2 \text{Var}(X) = \frac{a^2}{\lambda^2}.$$ 

Note that if $b \neq 0$ or $a \leq 0$ the distribution cannot be exponential (since an exponential must be positive, and has positive probability of being close to zero). If $b = 0$ and $a > 0$ we have, for $y > 0$,

$$f_Y(y) = f_X \left( \frac{y - b}{a} \right) \frac{1}{\frac{1}{a}} \frac{1}{\lambda} e^{-\frac{\lambda y}{a}},$$

which is the PDF of an $\text{Exp}(\lambda/a)$ distribution.

### 14.1.6 Cauchy and Lognormal Distributions

We say that $X$ has a Cauchy distribution if its PDF has the form

$$f_X(x) = \frac{C}{1 + x^2}.$$ 

We will learn more about it in the exercises. We say that $Y$ has a Lognormal distribution when $Y = e^X$ and $X$ is normal. This is an important distribution in finance, as it is used to model stock fluctuations. The idea is to model percentage moves in a stock with a normal distribution, since the size of a move should depend on the price of the stock.
15 Lecture 15

15.1 Finishing Continuous Distributions

15.1.1 Review Exercises

1. Suppose you are able to generate draws from a Unif(0,1) distribution. In other words, you have some procedure in a programming language that will give values that are uniformly distributed on (0,1). How can you use this procedure to generate draws from an Exp(\(\lambda\)) distribution?

2. We saw that a Cauchy distribution has

\[
f_X(x) = \frac{C}{1 + x^2}
\]

for some \(C\). What is \(C\)?

3. What is the expectation of a Cauchy random variable?

4. Let \(X \sim \mathcal{N}(\mu, \sigma^2)\).

   (a) What is the PDF of \(Y = e^X\)?

   (b) Compute \(E[Y]\).

15.1.2 Solutions

1. Recall the CDF of an Exp(\(\lambda\)) distribution is

\[
F(x) = 1 - e^{-\lambda x}
\]

for \(x \geq 0\). Inverting we have

\[
F^{-1}(y) = -\frac{\log(1 - y)}{\lambda}.
\]

Thus if \(X \sim \text{Unif}(0,1)\) then

\[
-\frac{\log(1 - X)}{\lambda} \sim \text{Exp}(\lambda).
\]

We take each value generated by our procedure, and plug it into \(F^{-1}(y)\) to get our draws.

2. Note that

\[
1 = \int_{-\infty}^{\infty} \frac{C}{1 + x^2} \, dx = C [\arctan(x)]_{-\infty}^{\infty} = C\pi,
\]

so \(C = 1/\pi\).
3. We will show the expectation doesn’t exist. To see this, note that
\[ \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{|x|}{1 + x^2} \, dx = \frac{1}{\pi} \int_{0}^{\infty} \frac{2x}{1 + x^2} \, dx = \frac{1}{\pi} \int_{1}^{\infty} \frac{du}{u} = \frac{[\log(u)]_{1}^{\infty}}{\pi} = \infty. \]

4. (a) Let \( g(x) = e^x \) so that \( Y = g(X) \) and, for \( y > 0 \),
\[ f_Y(y) = f_X(\log(y)) \frac{1}{y} = \frac{1}{y \sqrt{2\pi \sigma^2}} e^{-\frac{1}{2} \left( \frac{\log(y) - \mu}{\sigma} \right)^2}. \]

(b) By LOTUS we have
\[
E[Y] = \frac{1}{\sqrt{2\pi \sigma^2}} \int_{-\infty}^{\infty} e^{x} e^{-\frac{1}{2} \left( \frac{x - \mu}{\sigma} \right)^2} \, dx
= \frac{1}{\sqrt{2\pi \sigma^2}} \int_{-\infty}^{\infty} e^{-\frac{1}{2\sigma^2} (x^2 + \mu^2 - 2x\mu - 2x\sigma^2)} \, dx
= \frac{1}{\sqrt{2\pi \sigma^2}} \int_{-\infty}^{\infty} e^{-\frac{1}{2\sigma^2} (x-(\mu+\sigma^2))^2 + \frac{\sigma^2}{2} + \mu} \, dx
= e^{\frac{\sigma^2}{2} + \mu}.
\]

### 15.2 Multivariate Distributions

Up to this point we have spent our time studying a variety of different distributions, both discrete and continuous. In this section we look at joint distributions which show not only how random variables behave, but how they are related. We had a preview of this topic when looking at independent random variables.

### 15.3 Multivariable Calculus Review

#### 15.3.1 Integrating over Regions Exercises

1. Let \( f : \mathbb{R}^2 \rightarrow \mathbb{R} \) be defined as follows:
\[ f(x, y) = \begin{cases} 
1 & \text{if } 0 \leq x \leq y \leq 1, \\
0 & \text{otherwise}.
\end{cases} \]

What is \( \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) \, dx \, dy \)?
2. Consider the function \( f(x, y) = 4 - (x^2 + y^2) \). The graph is a surface of revolution of a parabola.

(a) Let \( R = \{ (x, y): 0 \leq x, y \leq 1 \} \). What is \( \iint_R f(x, y) \, dx \, dy \)?

(b) Let \( S = \{ (x, y): 0 \leq x^2 + y^2 \leq 4 \} \). What is \( \iint_S f(x, y) \, dx \, dy \)?
(c) Let $T$ be the triangular region in the $xy$-plane enclosed by the lines $y = 1 - x$, $x = 0$, and $y = 0$. What is $\iint_T f(x, y) \, dx \, dy$?
3. Let \( f(x, y) = e^{-(x+y)} \).

(a) Let \( g \) be defined by

\[
g(x, y) = \begin{cases} 
1 & \text{if } 0 \leq x, y \leq 2, \\
0 & \text{otherwise}.
\end{cases}
\]

What is \( \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)g(x, y) \, dx \, dy \)?

(b) Let \( R = \{(x, y) : 0 \leq x, y \leq 1\} \) and let \( T = 2R = \{(x, y) : 0 \leq x, y \leq 2\} \). Which of the following two integrals is bigger:

\[
\int_{R} f(x, y) \, dx \, dy \quad \text{or} \quad \int_{T} f(x/2, y/2) \, dx \, dy?
\]
(c) Let $R = \{(x, y) : 1 \leq x, y \leq 4\}$ and $T = \{(x, y) : 1 \leq x, y \leq 2\}$. Which of the following two integrals is bigger:

$$\int_{R} f(x/10, y/10) \, dx \, dy \quad \text{or} \quad \int_{T} f(x^2/10, y^2/10) \, dx \, dy?$$

4. Let $g(x, y) = x^2 + 2xy + y^3$ and let $f$ be defined by

$$f(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} e^{-(2t^2+u^2)} \, dt \, du.$$

(a) What are $\frac{\partial g}{\partial x}(x, y)$ and $\frac{\partial g}{\partial y}(x, y)$?
(b) What are $\frac{\partial f}{\partial x}(x, y)$ and $\frac{\partial f}{\partial y}(x, y)$?
(c) What is $\frac{\partial^2 f}{\partial x \partial y}(x, y)$?

15.3.2 Solutions

1. By inspection, we can see the volume under the surface is $1/2$. More precisely,

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) \, dx \, dy = \int_{0}^{1} \int_{x}^{1} 1 \, dy \, dx = \int_{0}^{1} [y]_{x}^{1} \, dx = \int_{0}^{1} 1 - x \, dx = [x - x^2/2]_{0}^{1} = 1/2.$$
2. (a) Integrating we have

\[
\int_{0}^{1} \int_{0}^{1} 4 - (x^2 + y^2) \, dx \, dy = \int_{0}^{1} \left[ 4x - x^3/3 - xy^2 \right]_{0}^{1} \, dy \\
= \int_{0}^{1} 4 - 1/3 - y^2 \, dy \\
= \left[ \frac{11y}{3} - \frac{y^3}{3} \right]_{0}^{1} \\
= \frac{10}{3}.
\]

(b) If we change to polar coordinates, we have

\[
\int\int_{S} 4 - (x^2 + y^2) \, dx \, dy = \int_{0}^{2\pi} \int_{0}^{2} (4 - r^2) r \, dr \, d\theta \\
= \int_{0}^{2\pi} \left[ \frac{4r^2}{2} - \frac{r^4}{4} \right]_{0}^{2} \, d\theta \\
= \int_{0}^{2\pi} 8 - 4 \, d\theta \\
= 8\pi.
\]

Note above that we converted \((x, y)\) into \((r, \theta)\). This is the multivariate form of a \(u\)-substitution. To understand how this works, let \(T : (0, \infty) \times (0, 2\pi) \rightarrow \mathbb{R}^2\) be defined by

\[T(r, \theta) = (r \cos \theta, r \sin \theta) = (x, y)\] .

Then we write the Jacobian:

\[
\frac{d(x, y)}{d(r, \theta)} = \det \begin{pmatrix} \cos \theta & -r \sin \theta \\ \sin \theta & r \cos \theta \end{pmatrix} = r \cos^2 \theta + r \sin^2 \theta = r.
\]

This leads to

\[
\int\int_{S} f(x, y) \, dx \, dy = \int_{0}^{2\pi} \int_{0}^{2} f(T(r, \theta)) \left| \frac{d(x, y)}{d(r, \theta)} \right| \, dr \, d\theta
\]
as we did above.

Alternatively, we can try to integrate over cartesian coordinates:

\[
\int\int_{S} 4 - (x^2 + y^2) \, dx \, dy = \int_{-2}^{2} \int_{-\sqrt{4-x^2}}^{\sqrt{4-x^2}} 4 - x^2 - y^2 \, dy \, dx,
\]

but the resulting integral is much harder to evaluate.
(c) Integrating we have
\[
\int_0^1 \int_0^{1-x} 4 - x^2 - y^2 \, dy \, dx = \int_0^1 \left[ 4y - x^2 y - y^3 \right]_0^{1-x} \, dx
\]
\[
= \int_0^1 4(1-x) - x^2(1-x) - \frac{(1-x)^3}{3} \, dx
\]
\[
= \int_0^1 4 - 4x - x^2 - x^3 - \frac{1}{3} + x - x^2 + \frac{x^3}{3} \, dx
\]
\[
= \left[ \frac{11x}{3} - \frac{3x^2}{2} - \frac{2x^3}{3} + \frac{x^4}{3} \right]_0^1
\]
\[
= 22 - 9 - 4 + 2
\]
\[
= \frac{11}{6}.
\]

3. (a)
\[
\int_0^2 \int_0^2 e^{-(x+y)} \, dx \, dy = \int_0^2 e^{-x} \, dx \int_0^2 e^{-y} \, dy = (1 - e^{-2})^2.
\]

(b) The second is larger. If we let \((u,v) = G(x,y) = (x/2, y/2)\) then we have
\[
\int_R f(u,v) \, du \, dv = \int_T f(G(x,y)) \left| \frac{d(u,v)}{d(x,y)} \right| \, dx \, dy = \int_T f(x/2, y/2) \cdot \frac{1}{4} \, dx \, dy.
\]

(c) The first is larger. If we let \((u,v) = G(x,y) = (x^2, y^2)\) then we have
\[
\int_R f(u/10, v/10) \, du \, dv = \int_T f(G(x,y)/10) \left| \frac{d(u,v)}{d(x,y)} \right| \, dx \, dy = \int_T f(x^2/10, y^2/10)4xy \, dx \, dy.
\]

4. (a) We have
\[
\frac{\partial g}{\partial x}(x,y) = 2x + 2y \quad \text{and} \quad \frac{\partial g}{\partial y}(x,y) = 2x + 3y^2.
\]

(b) We have
\[
\frac{\partial f}{\partial x}(x,y) = \int_{-\infty}^{\infty} e^{-(2t^2+x^2)} \, dt \quad \text{and} \quad \frac{\partial f}{\partial y}(x,y) = \int_{-\infty}^{\infty} e^{-(2y^2+u^2)} \, du.
\]

(c) We have
\[
\frac{\partial^2 f}{\partial x \partial y}(x,y) = e^{-(2y^2+x^2)}.
\]
16 Lecture 16

16.1 Multivariate Distributions

We will spend most of our time focusing on bivariate distributions, but all of our work will extend to general multivariate distributions. They are just harder to visualize.

16.1.1 Discrete Bivariate Distributions

Let $X, Y : S \to \mathbb{R}$ be discrete random variables. The joint PMF $p_{X,Y} : \mathbb{R}^2 \to \mathbb{R}$ is defined as follows:

$$p_{X,Y}(x, y) = P(X = x, Y = y) = P(X = x \text{ and } Y = y).$$

If $X, Y$ take on finitely many values, it is convenient to represent the joint PMF as a table.

We can also define a joint CDF $F_{X,Y} : \mathbb{R}^2 \to \mathbb{R}$ by

$$F_{X,Y}(x, y) = P(X \leq x, Y \leq y).$$

16.1.2 Discrete Bivariate Exercises

1. Suppose $X, Y$ have a joint PMF given by the following table:

<table>
<thead>
<tr>
<th>$X$</th>
<th>0</th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1/10</td>
<td>2/10</td>
<td>3/10</td>
</tr>
<tr>
<td>1</td>
<td>1/15</td>
<td>2/15</td>
<td>3/15</td>
</tr>
</tbody>
</table>

(a) Compute $P(X = 1, Y = 1)$.

(b) Compute $P(Y = 2)$.

(c) Compute $P(X = Y)$.

(d) Compute $P(Y = 2 | X = 1)$.

(e) Are $X, Y$ independent?

2. Suppose there are two coins in a can with head probabilities $1/2, 2/3$, respectively. We randomly draw a coin from the can and flip it twice. Let $X$ denote head probability of the coin we choose, and let $Y$ denote the number of heads we obtain from the 2 flips.

(a) Give the joint PMF of $X, Y$ as a table.

(b) Use the table to compute the PMF of $X$.

(c) Use the table to compute the PMF of $Y$.

(d) Use the table to determine if $X, Y$ independent.

3. Let $A \subset \mathbb{R}^2$ and let $X, Y$ be discrete random variables with joint PMF $p_{X,Y}$. Give a formula for $P((X, Y) \in A)$.
16.1.3 Solutions

1. (a) \( P(X = 1, Y = 1) = \frac{2}{15}. \)
(b) \( P(Y = 2) = \frac{3}{10} + \frac{3}{15}. \)
(c) \( P(X = Y) = \frac{1}{10} + \frac{2}{15}. \)
(d) \( P(Y = 2|X = 1) = \frac{\frac{3}{15}}{\frac{1}{15} + \frac{2}{15} + \frac{3}{15}}. \)
(e) Summing rows and columns we have
\[
\begin{align*}
P(X = 0) &= \frac{3}{5}, \quad P(X = 1) = \frac{2}{5}, \quad P(Y = 0) = \frac{1}{6}, \\
P(Y = 1) &= \frac{2}{6}, \quad P(Y = 2) = \frac{3}{6}.
\end{align*}
\]
Checking each entry, we see that \( P(Y = i)P(X = j) = P(Y = i, X = j) \) showing that \( X, Y \) are independent.

2. (a)

\[
\begin{array}{c|ccc|c}
 & 0 & 1 & 2 & \text{Total} \\
\hline
X \ 1/2 & \frac{1}{8} & \frac{2}{8} & \frac{1}{8} & \frac{4}{8} \\
2/3 & \frac{1}{18} & \frac{2}{9} & \frac{4}{18} & \frac{9}{18} \\
\hline
\text{Total} & \frac{13}{52} & \frac{34}{52} & \frac{25}{52} & \\
\end{array}
\]

(b) \( p_X(1/3) = p_X(2/3) = \frac{1}{2}. \)
(c) \( p_Y(0) = \frac{13}{72}, \quad p_Y(1) = \frac{34}{72}, \quad p_Y(2) = \frac{25}{72}. \)
(d) No. Note that \( p_X(1/3)p_Y(0) = \frac{52}{576} \neq \frac{1}{8}. \)

3.
\[
P((X, Y) \in A) = \sum_{(x,y) \in A} p_{X,Y}(x,y).
\]

16.1.4 Absolutely Continuous Joint Distributions

Let \( X, Y : S \to \mathbb{R} \) be random variables. We say that \( X, Y \) have an absolutely continuous joint distribution if the joint CDF \( F_{X,Y} \) can be expressed as
\[
F_{X,Y}(x, y) = P(X \leq x, Y \leq y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{X,Y}(s, t) \, dt \, ds,
\]
for some function $f : \mathbb{R}^2 \to [0, \infty)$. Here $f_{X,Y}$ is called the joint PDF of $X, Y$. For more general regions $A \subset \mathbb{R}^2$ we have

$$P((X, Y) \in A) = \int \int_A f_{X,Y}(s, t) \, ds \, dt.$$ 

Even if $X, Y$ have absolutely continuous distributions, it doesn’t mean the pair has a joint PDF (a technical point we won’t investigate). That said, if $X, Y$ have a joint PDF then $X, Y$ each have PDF’s separately. To see this note that

$$P(X \leq x) = P(X \leq x, Y < \infty) = \int_{-\infty}^{\infty} \int_{-\infty}^{x} f_{X,Y}(s, t) \, dt \, ds.$$ 

This implies

$$f_X(s) = \int_{-\infty}^{\infty} f_{X,Y}(s, t) \, dt$$

is the PDF of $X$ called the marginal PDF of $X$. Analogously, we have

$$f_Y(t) = \int_{-\infty}^{\infty} f_{X,Y}(s, t) \, ds$$

called the marginal PDF of $Y$. This operation is akin to summing over columns/rows in the joint PMF table. The plot below shows a joint PDF, and then gives the two marginal PDFs superimposed. You can think of the marginal PDFs shown as acting like the total columns and rows in the joint PMF table, where instead of summing we integrated.

In general, we know that $X, Y$ are independent iff

$$P(X \leq x, Y \leq y) = F_{X,Y}(x, y) = F_X(x)F_Y(y).$$
For $X, Y$ with a joint PDF $f_{X,Y}$ we have another test for independence: $X, Y$ are independent iff

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

for some functions $g, h : \mathbb{R} \to \mathbb{R}$. [Aside: This is not quite true, as PDFs are only defined almost everywhere.] Up to constant factors we have $g = f_X$ and $h = f_Y$.

As a reminder, we already saw that if $X, Y$ are independent random variables, then so are $g(X)$ and $h(Y)$ for any $g, h : \mathbb{R} \to \mathbb{R}$. This is true for all random variables.

### 16.1.5 Continuous Joint Distribution Exercises

1. If $f_{X,Y}(x, y)$ is the joint PDF of $X, Y$ what is

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

2. If $(X, Y)$ is an absolutely continuous joint distribution with joint CDF $F_{X,Y}(x, y)$, how do you obtain the joint PDF?

3. Suppose $X, Y$ have joint PDF

$$f_{X,Y}(x, y) = \begin{cases} 
  c(4 - (x^2 + y^2)) & \text{if } x^2 + y^2 \leq 4, \\
  0 & \text{otherwise.}
\end{cases}$$

   (a) Find $c$.
   (b) Compute the probability $P(0 \leq X \leq 1, 0 \leq Y \leq 1)$.

4. $(\star)$ Suppose that $f_{X,Y}$ is the joint PDF of $X, Y$ with $f_{X,Y}(x, y) = g(x)h(y)$ for some $g, h : \mathbb{R} \to \mathbb{R}$. Show that $g(x) = C_1 f_X(x)$ and $h(y) = C_2 f_Y(y)$ for some $C_1, C_2$ with $C_1 C_2 = 1$. Using this, show that $X, Y$ are independent.

5. If $f_{X,Y}(x, y) = 1$ for $0 \leq x, y \leq 1$ are $X, Y$ independent?

6. For each of the following PDFs:

   (i) Compute $C$.
   (ii) Determine if $X, Y$ are independent.
   (iii) Compute $P(Y \geq X)$.

   (a)

   $$f_{X,Y}(x, y) = \begin{cases} 
   C e^{-2x-3y} & \text{if } x, y \geq 0, \\
   0 & \text{otherwise.}
\end{cases}$$

   (b)

   $$f_{X,Y}(x, y) = \begin{cases} 
   C(x + y) & \text{if } 0 \leq x \leq 1, 0 \leq y \leq 1, \\
   0 & \text{otherwise.}
\end{cases}$$
(c) 

\[ f_{X,Y}(x, y) = \begin{cases} 
  C & \text{if } 0 \leq x^2 + y^2 \leq 1, \\
  0 & \text{otherwise.} 
\end{cases} \]

(d) 

\[ f_{X,Y}(x, y) = \begin{cases} 
  Cx^2y & \text{if } 0 \leq x^2 \leq y \leq 1, \\
  0 & \text{otherwise.} 
\end{cases} \]

16.1.6 Solutions

1. \[ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x, y) \, dx \, dy = 1. \]

2. \[ f_{X,Y}(x, y) = \frac{\partial^2 F_{X,Y}(x, y)}{\partial x \partial y}. \]

3. (a) In the calculus review we computed the integral. Thus \( c = \frac{1}{8\pi} \).

(b) In the calculus review we computed the integral. Thus

\[ P(0 \leq X \leq 1, 0 \leq Y \leq 1) = \frac{10}{3(8\pi)}. \]

4. Note that

\[ 1 = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x, y) \, dx \, dy = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x)h(y) \, dx \, dy = \left( \int_{-\infty}^{\infty} g(x) \, dx \right) \left( \int_{-\infty}^{\infty} h(y) \, dy \right). \]

Next note that the marginal \( f_X \) is given by

\[ f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y) \, dy = \int_{-\infty}^{\infty} g(x)h(y) \, dy = g(x) \int_{-\infty}^{\infty} h(y) \, dy. \]

Similarly, we have

\[ f_Y(y) = h(y) \int_{-\infty}^{\infty} g(x) \, dx. \]

Thus

\[ f_X(x)f_Y(y) = \left( g(x) \int_{-\infty}^{\infty} h(y) \, dy \right) \left( h(y) \int_{-\infty}^{\infty} g(x) \, dx \right) = g(x)h(y). \]
Computing the CDF gives

\[ F_{X,Y}(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{X,Y}(s, t) \, dt \, ds \]
\[ = \int_{-\infty}^{x} \int_{-\infty}^{y} g(s)h(t) \, dt \, ds \]
\[ = \int_{-\infty}^{x} \int_{-\infty}^{y} f_X(s)f_Y(t) \, dt \, ds \]
\[ = \left( \int_{-\infty}^{x} f_X(s) \, ds \right) \left( \int_{-\infty}^{y} f_Y(t) \, dt \right) \]
\[ = F_X(x)F_Y(y). \]

5. Yes, we can write \( f_{X,Y}(x, y) = 1_{[0,1]}(x)1_{[0,1]}(y). \)

6. (a) i. Integrating gives

\[ 1 = \int_{0}^{\infty} \int_{0}^{\infty} C e^{-2x-3y} \, dx \, dy = C \int_{0}^{\infty} e^{-2x} \, dx \int_{0}^{\infty} e^{-3y} \, dy = \frac{C}{6}, \]

so \( C = 6. \)

ii. Yes. Let \( g(x) = 6e^{-2x} \) for \( x \geq 0 \) and \( h(y) = e^{-3y} \) for \( y \geq 0. \) Here \( X \sim \text{Exp}(2), Y \sim \text{Exp}(3) \) are independent.

iii. We have

\[ P(Y \geq X) = \int_{0}^{\infty} \int_{x}^{\infty} 6e^{-2x-3y} \, dy \, dx \]
\[ = 6 \int_{0}^{\infty} \int_{x}^{\infty} e^{-2x} e^{-3y} \, dy \, dx \]
\[ = 6 \int_{0}^{\infty} e^{-2x} \left[ \frac{e^{-3y}}{-3} \right]_{x}^{\infty} \, dx \]
\[ = 2 \int_{0}^{\infty} e^{-5x} \, dx \]
\[ = \frac{2}{5}. \]

(b) i. Integrating gives

\[ 1 = \int_{0}^{1} \int_{0}^{1} C(x + y) \, dx \, dy = C \int_{0}^{1} \left[ \frac{x^2}{2} + xy \right]_{0}^{1} \, dy = C \int_{0}^{1} \frac{1}{2} + y \, dy = C, \]

so \( C = 1. \)
ii. No. Computing the marginals we have

\[ f_X(x) = \int_0^1 x + y \, dy = \left[ xy + \frac{y^2}{2} \right]_0^1 = x + \frac{1}{2}. \]

Symmetrically we have

\[ f_Y(y) = y + \frac{1}{2}. \]

Note that \( f_{X,Y}(x, y) \neq f_X(x)f_Y(y). \)

iii. We have

\[ P(Y \geq X) = \int_0^1 \int_y^1 x + y \, dx \, dy \]
\[ = \int_0^1 \left[ \frac{y^2}{2} + y^2 \right]_0^1 \, dx \]
\[ = \left[ \frac{y^3}{2} \right]_0^1 \]
\[ = \frac{1}{2}. \]

(c) i. The area is \( \pi \) so \( C = 1/\pi. \)

ii. No. Suppose \( f_{X,Y}(x, y) = g(x)h(y). \) Note that \( f_{X,Y}(5/6, 5/6) = 0, \) so either \( g(5/6) = 0 \) or \( h(5/6) = 0. \) But this implies either \( f_{X,Y}(5/6, 0) = 0 \) or \( f_{X,Y}(0, 5/6) = 0, \) which both aren’t true. In general, if the region where the PDF is positive isn’t rectangular, the random variables cannot be independent.

iii. The probability is \( 1/2, \) since the disk is symmetric.

(d) i. Integrating gives

\[ 1 = \int_{-1}^1 \int_{-x^2}^1 Cx^2y \, dy \, dx \]
\[ = C \int_{-1}^1 \left[ \frac{x^2y^2}{2} \right]_{-x^2}^1 \, dx \]
\[ = C \int_{-1}^1 \frac{x^2}{2} - \frac{x^6}{2} \, dx \]
\[ = C \left[ \frac{x^3}{6} - \frac{x^7}{14} \right]_{-1}^1 \]
\[ = C \left[ \frac{1}{3} - \frac{1}{7} \right], \]

implying \( C = 21/4. \)

ii. No, the region isn’t rectangular.
iii. We have

$$P(Y \geq X) = P(X \leq 0) + P(X \geq 0, Y \geq X)$$

$$= \frac{1}{2} + \frac{21}{4} \int_{0}^{1} \int_{x}^{1} x^2 y \, dy \, dx$$

$$= \frac{1}{2} + \frac{21}{4} \int_{0}^{1} \left[ \frac{x^2 y^2}{2} \right]_x^1 \, dx$$

$$= \frac{1}{2} + \frac{21}{4} \int_{0}^{1} \frac{x^2}{2} - \frac{x^4}{2} \, dx$$

$$= \frac{1}{2} + \frac{21}{4} \int_{0}^{1} \left[ \frac{x^3}{6} - \frac{x^5}{10} \right]_0^1$$

$$= \frac{1}{2} + \frac{21}{60}.$$  

### 16.1.7 Probability Picture Problems

When dealing with a uniform density over some set, probabilities are proportional to areas. In these cases we can solve probability problems using geometric methods.

1. Two people will arrive at a coffee shop between 6:00 and 7:00 (uniformly and independently). What is the chance they arrive within 15 minutes of each other?

2. You have a stick that you break randomly (uniformly and independently) at two points. What is the chance the 3 segments are the sides of a triangle? [Three segments form a triangle if the sum of the two smaller lengths is greater than the third.]

### 16.1.8 Solutions

1. We can model this as 2 uniform i.i.d. Unif(0, 1) random variables describing where each person arrives between 6 and 7. Their joint distribution is just uniform on the unit square, so the probability is the area of the required region. This is drawn below:
The area of the region is $1 - \left( \frac{3}{4} \right)^2 = \frac{7}{16}$.

2. The answer is $1/4$. Since we have a uniform distribution on the unit square, this is simply a question about the area of the event in question. Thus, to solve this, we mark the regions on the square where we obtain a triangle:

To better understand this, note that we don’t obtain a triangle when any of the sides has length more than .5. By symmetry, let’s focus on the half where $y < x$ (i.e., below the line $y = x$). In this region $y$ represents our first cut, so we must have $y < .5$. Our
last segment must also be less than .5, so we must have $x > .5$. Finally, we must have $x - y < .5$, so we take values above the line $y = x - .5$. This gives the lower triangle below. By symmetry, we also get the upper triangle giving a total of $1/4$ for the area.
17 Lecture 17

17.1 More Bivariate Distributions

17.1.1 Review Exercises

1. Let $X, Y \sim \text{Unif}(0, 1)$ be independent. What is the probability that $\max(X, Y) > 1/2$?

2. Suppose $X \sim \text{Exp}(\lambda)$ and $Y \sim \text{Bin}(n, p)$ are independent random variables. How would you describe their joint distribution?

3. (a) (*) Define $\Gamma : \mathbb{R}_{>0} \to \mathbb{R}$ by

   \[
   \Gamma(a) = \int_0^\infty x^{a-1}e^{-x} \, dx.
   \]

   Show that $\Gamma(a) = (a - 1)\Gamma(a - 1)$ for $a > 1$, and $\Gamma(n) = (n - 1)!$ if $n$ is a positive integer.

   (b) (**) Show, for $a, b > 0$,

   \[
   B(a, b) = \int_0^1 x^{a-1}(1 - x)^{b-1} \, dx = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a + b)},
   \]

   called the Beta function.

17.1.2 Solutions

1. Here are two ways to do this:

   (a) The complement is $P(X \leq 1/2, Y \leq 1/2) = P(X \leq 1/2)P(Y \leq 1/2)$ by independence. This is $1/4$ so the answer is $3/4$.

   (b) Draw the square and mark the region where the max is at least $1/2$. To do this divide the region into 4 equal sized squares. The only region you will not mark is the lower left square, again giving $3/4$.

2. We will discuss this more in a moment, but we will used a joint PDF/PMF given by

   \[
   f_{X,Y}(x, y) = \lambda e^{-\lambda x} \binom{n}{y} p^y (1 - p)^{n-y},
   \]

   for $x \geq 0$ and $y = 0, 1, \ldots, n$. This satisfies,

   \[
   P(X \leq x_0, Y \leq y_0) = \int_{-\infty}^{x_0} \sum_{y\leq y_0} f_{X,Y}(x, y).
   \]
3. (a) Using integration by parts we have, for \( a > 1 \),

\[
\Gamma(a) = \int_0^\infty x^{a-1}e^{-x} \, dx = \left[-x^{a-1}e^{-x}\right]_0^\infty - \int_0^\infty -(a-1)x^{a-2}e^{-x} \, dx = (a-1)\Gamma(a-1).
\]

The factorial result follows by noting

\[
\Gamma(1) = \int_0^\infty x^0 e^{-x} \, dx = 1.
\]

(b) We compute \( \Gamma(a)\Gamma(b) \) and then get sneaky:

\[
\Gamma(a)\Gamma(b) = \left( \int_0^\infty x^{a-1}e^{-x} \, dx \right) \left( \int_0^\infty y^{b-1}e^{-y} \, dy \right)
\]

\[
= \int_0^\infty \int_0^\infty x^{a-1}y^{b-1}e^{-x-y} \, dx \, dy.
\]

Now we make the substitution \( u = x + y, \ v = \frac{x}{x+y} \) giving

\[
x = uv \quad \text{and} \quad y = u - uv
\]

and

\[
\frac{d(x,y)}{d(u,v)} = \det \begin{pmatrix} v & u \\ 1 - v & -u \end{pmatrix} = -vu - u(1-v) = -u.
\]

Thus we get

\[
\int_0^\infty \int_0^\infty x^{a-1}y^{b-1}e^{-x-y} \, dx \, dy = \int_0^\infty \int_0^1 (uv)^{a-1}(u-uv)^{b-1}e^{-uv}e^{-(u-uv)}u \, dv \, du
\]

\[
= \int_0^\infty \int_0^1 u^{a+b-1}v^{a-1}(1-v)^{b-1}e^{-u} \, dv \, du
\]

\[
= \left( \int_0^\infty u^{a+b-1}e^{-u} \, du \right) \left( \int_0^\infty v^{a-1}(1-v)^{b-1} \, dv \right)
\]

\[
= \Gamma(a+b) \, B(a,b).
\]

Note that a special case of this is when \( a, b > 0 \) are integers. Then we have

\[
B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} = \frac{(a-1)!(b-1)!}{(a+b-1)!}.
\]

Based on the last exercise above, we define the following distribution.

**Definition 59** (Beta Distribution). We say that \( X \sim \text{Beta}(a,b) \) for \( a, b > 0 \) if \( X \) has PDF \( f_X \) given by

\[
f_X(x) = \frac{1}{B(a,b)}x^{a-1}(1-x)^{b-1},
\]

for \( x \in [0,1] \).
The importance of the Beta distribution will become clear shortly. It is often used to model parameters to other distributions, like the $p$ in the Binomial. To better understand the Beta PDF, note that $a$ controls the behavior near 0, and $b$ controls the behavior near 1. If $a < 1$ then the PDF tends to infinity as $x \to 0$. If $a$ is large, then the PDF decays to zero quickly as $x \to 0$. Finally, note that if $a = b = 1$ then we get a Unif(0, 1) distribution.

17.1.3 Mixed Bivariate Distributions

It is also possible to have joint distributions where one variable is discrete and the other is continuous. These distributions will have a function $f_{X,Y}$ like a joint PDF/PMF satisfying

$$ P(X \in A, Y \in B) = \int_A \sum_{y \in B} f_{X,Y}(x, y) \, dx. $$

17.1.4 Marginal and Conditional Distributions

We have already seen that if you have a joint PMF $p_{X,Y}$ given as a table, then we can extract probabilities such as $P(X = x)$ and $P(Y = y)$ by summing rows or columns of the table. Formulaically we have

$$ P(X = x) = \sum_{y : p_{X,Y}(x,y) > 0} p_{X,Y}(x,y) = \sum_{y : p_{X,Y}(x,y) > 0} P(X = x, Y = y). $$

When we have a joint distribution of $X, Y$, we call the individual distributions of $X, Y$ their marginal distribution. Using the above, if $p_{X,Y}(x,y)$ is the joint PMF, then

$$ p_X(x) = \sum_{y : p_{X,Y}(x,y) > 0} p_{X,Y}(x,y) $$

is the marginal PMF of $X$. Analogously, if $f_{X,Y}$ is the joint PDF then

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

is the marginal PDF of $X$.

The plot below shows a joint PDF, and then gives the two marginal PDFs superimposed. You can think of the marginal PDFs shown as acting like the total columns and rows in the joint PMF table, where instead of summing we integrated.
The marginal distribution of $X$ is computed by aggregating over all possible values of $Y$. We can also ask for the distribution of $X$ when we are given a value of $Y$. For example, suppose $X, Y$ have joint PMF $p_{X,Y}$ and note that

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

where the equation above defines the notation on the left. We call $p_{X|Y}(x|y)$ the conditional PMF of $X$ given $Y = y$. To obtain it, we simply look at the appropriate row (or column, depending on how the table is laid out), and then divide all quantities by their sum.

By analogy we define the conditional PDF to be

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = \frac{f_{X,Y}(x,y)}{\int_{-\infty}^{\infty} f_{X,Y}(t,y) \, dt}.$$ 

You can also think of the conditional PDF by the equivalent equation

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

or by thinking of it as the function $g(x,y)$ such that

$$\int_B f_Y(y) \left( \int_A g(x,y) \, dx \right) \, dy = \int_B \int_A f_{X,Y}(x,y) \, dx \, dy = P(X \in A, Y \in B).$$

Up until this point, if $X, Y$ have a joint PDF $f_{X|Y}$ we would have no way of assigning meaning to the expression $P(X \in A| Y = y)$ since the event $Y = y$ has probability 0 (as $Y$ is continuous). We now define this to be

$$P(X \in A| Y = y) = \int_A f_{X|Y}(x|y) \, dx.$$
In terms of our old definition, this can be achieved by considering
\[ P(X \in A | Y \in [y - \epsilon, y + \epsilon]) \]
and letting \( \epsilon \to 0^+ \).

Recall that conditional probability measures are actual probability measures that give our updated beliefs given information. Analogously, conditional PMFs and PDFs are PMFs and PDFs that incorporate information, such as the value of the other RV in our bivariate distribution. To see this, let \( f_{X|Y} \) be a conditional PDF. Then
\[
\int_{-\infty}^{\infty} f_{X|Y}(x|y) \, dx = \int_{-\infty}^{\infty} \frac{f_{X,Y}(x,y)}{f_Y(y)} \, dx = \frac{1}{f_Y(y)} \int_{-\infty}^{\infty} f_{X,Y}(x,y) \, dx = \frac{f_Y(y)}{f_Y(y)} = 1.
\]
The same can be see for a conditional PMF where we sum instead of integrating.

If \( X, Y \) have a mixed joint PDF/PMF \( f_{X,Y} \) with \( X \) continuous and \( Y \) discrete then the marginals are given by
\[
f_X(x) = \sum_y f_{X,Y}(x,y) \quad \text{and} \quad p_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y) \, dx
\]
and the conditionals are given by
\[
f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{p_Y(y)} \quad \text{and} \quad f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)}.
\]

### 17.1.5 Marginal and Conditional Exercises

1. The following table describes the joint PMFs of \( X, Y \):

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.06</td>
<td>0.02</td>
<td>0.04</td>
<td>0.08</td>
<td>0.20</td>
</tr>
<tr>
<td>2</td>
<td>0.15</td>
<td>0.05</td>
<td>0.10</td>
<td>0.20</td>
<td>0.50</td>
</tr>
<tr>
<td>3</td>
<td>0.09</td>
<td>0.03</td>
<td>0.06</td>
<td>0.12</td>
<td>0.30</td>
</tr>
<tr>
<td>Total</td>
<td>0.30</td>
<td>0.10</td>
<td>0.20</td>
<td>0.40</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Below the PMF is depicted as a plot.
(a) What is the marginal PMF of $X$?
(b) What is the marginal PMF of $Y$?
(c) What is the conditional PMF $p_{X|Y}(x|3)$?
(d) What is the conditional PMF $p_{Y|X}(y|2)$?

2. Suppose $X, Y$ have joint PDF

\[
f_{X,Y}(x, y) = \begin{cases} \frac{21}{4} x^2 y & \text{if } 0 \leq x^2 \leq y \leq 1, \\ 0 & \text{otherwise}. \end{cases}
\]

Below the picture on the left shows where the joint PDF is positive. The picture on the right shows the joint PDF, and the marginals along side of it.

Below the picture on the left shows the PDF with a plane slicing it through $y = 2/3$. The picture on the right shows the intersection of the joint PDF and the slicing plane. Up to a constant multiple, this is the conditional PDF $f_{X|Y}(x|2/3)$. 

119
(a) Compute the marginal distribution of $X$.
(b) Compute the marginal distribution of $Y$.
(c) Compute the conditional PDF $f_{X|Y}(x|y)$ for $y \in (0,1)$.
(d) Compute the conditional PDF $f_{Y|X}(y|x)$ for $x \in (0,1)$.

3. Suppose that $X \sim f_X$ represents the probability a given coin is heads where

$$f_X(x) = 6x(1-x) \quad \text{for } x \in [0,1]$$

and 0 otherwise (Beta(2, 2) distributed). We flip the coin 5 times. Conditional on the value of $X = x$, we assume that $Y$, the number of heads, has a Bin(5, $x$) distribution. Depicted below is the joint PDF/PMF where we have filled in the plots on the right to show the areas that sum to 1.

(a) Give the joint PDF/PMF of the mixed distribution of $X, Y$.
(b) What is the marginal distribution of $Y$?
(c) Given we get 5 heads, what is the conditional distribution of \( X \)?

(d) Compute \( P(X \geq 1/2|Y = 5) \).

### 17.1.6 Solutions

1. (a) The marginal PMF of \( X \) is given by

\[
p_X(1) = 0.2, \quad p_X(2) = 0.5, \quad p_X(3) = 0.3,
\]

with all other values zero. This is taken directly from the totals column.

(b) The marginal PMF of \( Y \) is given by

\[
p_Y(1) = 0.3, \quad p_Y(2) = 0.1, \quad p_Y(3) = 0.2, \quad p_Y(4) = 0.4,
\]

with all other values 0.

(c) To compute \( p_{X|Y}(x|3) \) we simply copy the column where \( Y = 3 \) and scale by the total. This gives

\[
p_{X|Y}(1|3) = \frac{0.04}{2}, \quad p_{X|Y}(2|3) = \frac{0.1}{2}, \quad p_{X|Y}(3|3) = \frac{0.06}{2}.
\]

(d) To compute \( p_{Y|X}(y|2) \) we use the row where \( X = 2 \) scaled by the total. This gives

\[
p_{Y|X}(1|2) = \frac{0.15}{2}, \quad p_{Y|X}(2|2) = \frac{0.5}{2}, \quad p_{Y|X}(3|2) = \frac{0.1}{2}, \quad p_{Y|X}(4|2) = \frac{0.2}{2}.
\]

2. (a) To obtain the marginal we integrate giving, for \( -1 \leq x \leq 1 \),

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

\[
= \frac{21}{4} \int_{x^2}^{1} x^2 y \, dy
\]

\[
= \frac{21}{4} \left[ \frac{x^2 y}{2} \right]_{x^2}^{1}
\]

\[
= \frac{21}{4} \left( \frac{x^2}{2} - \frac{x^6}{2} \right).
\]
(b) Again we integrate giving, for $0 \le y \le 1$

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

$$= \frac{21}{4} \int_{-\sqrt{y}}^{\sqrt{y}} x^2 y \, dx$$

$$= \frac{21}{4} \left[ \frac{x^3 y}{3} \right]_{-\sqrt{y}}^{\sqrt{y}}$$

$$= \frac{21}{4} \frac{2 y^{5/2}}{3}$$

$$= \frac{7}{2} y^{5/2}.$$

(c) The conditional PDF is obtained by scaling by the marginal. We get, for $0 \le y \le 1$ and $x^2 \le y$,

$$f_{X|Y}(x|y) = \frac{\frac{21}{4} x^2 y}{\frac{7}{2} y^{5/2}} = \frac{3}{2} x^2 y^{-3/2}.$$

(d) Repeating we get, for $-1 \le x \le 1$ and $x^2 \le y \le 1$,

$$f_{Y|X}(y|x) = \frac{\frac{21}{4} x^2 y}{\frac{4}{3} (\frac{x^2}{2} - \frac{x^4}{2})} = \frac{2y}{1 - x^4}.$$

3. (a) The joint PDF/PMF is given by, for $x \in [0, 1]$ and $y \in \{0, \ldots, 5\}$,

$$f_{X,Y}(x, y) = 6x(1-x)\binom{5}{y} x^y (1-x)^{5-y}.$$

(b) Integrating we obtain, for $y = 0, 1, \ldots, 5$,

$$p_Y(y) = \int_{0}^{1} 6x(1-x)\binom{5}{y} x^y (1-x)^{5-y} \, dx$$

$$= 6 \binom{5}{y} \int_{0}^{1} x^{y+1}(1-x)^{6-y} \, dx$$

$$= 6 \binom{5}{y} B(y+2, 7-y)$$

$$= 6 \binom{5}{y} \frac{(y+1)!(6-y)!}{8!}$$

$$= \frac{(y+1)(6-y)}{56}.$$
(c) We have
\[ f_{X|Y}(x|5) = \frac{f_{X,Y}(x,5)}{p_Y(5)} = \frac{6x(1-x)x^5}{\frac{6}{56}} = 56x^6(1-x). \]

This has a Beta distribution with parameters B(7, 2). We will learn more about this later.

(d) We integrate to get
\[
\int_{1/2}^{1} 56x^6(1-x) \, dx = 56 \left[ \frac{x^7}{7} - \frac{x^8}{8} \right]_{1/2}^{1} = \frac{247}{256} \approx 0.965.
\]
18 Lecture 18

18.1 More Bivariate Distributions

18.1.1 Review Problems

1. Suppose $X, Y$ are independent with joint PDF $f_{X,Y}$. What is the conditional PDF $f_{X|Y}$?

2. Suppose $X$ has an absolutely continuous distribution with PDF $f_X$. Suppose $Y$, conditional on $X = x$ has a PMF $p_{Y|X}(y|x)$.

   (a) What is the mixed joint PDF/PMF $f_{X,Y}$?
   (b) Give formulas for the PMF or PDFs of the marginal distributions of $X$ and $Y$?
   (c) If instead $U$ was discrete with PMF $p_U$ and $V|U = u$ had a PDF $f_{V|U}$ what is the joint PDF/PMF and the marginal of $V$?

3. Suppose $X, Y$ have the joint PDF/PMF

   $$f_{X,Y}(x, y) = \frac{xy^{x-1}}{3}, \quad \text{for } x = 1, 2, 3 \text{ and } 0 < y < 1.$$ 

   (a) Check that $f_{X,Y}$ is a valid joint PDF/PMF.
   (b) Compute $P(X \geq 2, Y \geq 1/2)$.

18.1.2 Solutions

1. Note that

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

   so as expected, knowing $Y = y$ doesn’t change our beliefs on $X$.

2. (a) $f_{X,Y}(x, y) = p_{Y|X}(y|x)f_X(x)$

   (b) The marginal of $X$ is just $f_X$. The marginal of $Y$ is

   $$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x, y) \, dx.$$ 

   (c) Here we have $f_{U,V}(u, v) = p_U(u)f_{V|U}(v|u)$ and

   $$f_V(v) = \sum_u p_U(u)f_{V|U}(v|u).$$

3. (a) To check it is valid note that

   $$\sum_{x=1}^{3} \int_0^1 \frac{xy^{x-1}}{3} \, dy = \sum_{x=1}^{3} \left[ \frac{y^x}{3} \right]_0^1 = \sum_{x=1}^{3} \frac{1}{3} = 1.$$ 

124
\[
P(X \geq 2, Y \geq 1/2) = \frac{1}{3} \sum_{x=2}^{3} \int_{1/2}^{1} x y^{x-1} \, dy
\]
\[
= \frac{1}{3} \sum_{x=2}^{3} \left[ y^{x} \right]_{1/2}^{1}
\]
\[
= \frac{1}{3} \sum_{x=2}^{3} \left( 1 - \left( \frac{1}{2} \right)^{x} \right)
\]
\[
= \frac{2 - 1/4 - 1/8}{3}
\]
\[
= \frac{13}{24}.
\]

18.1.3 LOTP and Bayes’s Theorem for Random Variables

Now that we have conditional PDFs and PMFs we can state versions of LOTP and Bayes’s Theorem.

**Theorem 60** (Law of Total Probability for RVs). Let \( X, Y \) have a bivariate joint distribution that is discrete, (absolutely) continuous, or mixed. Let \( f_{X,Y} \) denote the joint PDF, PMF, or PDF/PMF. If \( Y \) is discrete then we have

\[
f_X(x) = \sum_{y} f_{X|Y}(x|y) f_Y(y).
\]

*If \( Y \) is continuous then we have

\[
f_X(x) = \int_{-\infty}^{\infty} f_{X|Y}(x|y) f_Y(y) \, dy.
\]

*Here we used \( f \) even in the discrete case just for economy of notation (we really should have used \( p \) in that case).*

The proof is nothing more than seeing that our definitions guarantee

\[
f_{X|Y}(x|y) f_Y(y) = f_{X,Y}(x,y),
\]

and then we compute the marginal distribution of \( X \). Similarly we have:

**Theorem 61** (Bayes’s Theorem for RVs). Let \( X, Y \) have a bivariate joint distribution that is discrete, (absolutely) continuous, or mixed. Let \( f_{X,Y} \) denote the joint PDF, PMF, or PDF/PMF. Then we have

\[
f_{X|Y}(x|y) = \frac{f_{Y|X}(y|x) f_X(x)}{f_Y(y)}.
\]

*Here again we used \( f \) even in the discrete case just for economy of notation (we really should have used \( p \) in that case for the marginal and joint distributions).*
The denominator in Bayes’s theorem can be computed as a marginal, or using LOTP for RVs. Since \( f_{X|Y}(x|y) \) is treated as a function of \( x \), and the denominator in Bayes’s theorem only depends on \( y \) we have:

**Corollary 62.**

\[
 f_{X|Y}(x|y) \propto f_{Y|X}(y|x) f_X(x). 
\]

Thus if we want to know the distribution of \( X \) given \( Y = y \) it is enough to just compute the numerator, as the constant is uniquely determined (since PDFs must integrate to 1 and PMFs must sum to 1).

### 18.1.4 Bayes’s Theorem Exercises

1. Suppose \( X \sim \text{Unif}(0, 1) \) represents the probability of getting heads, and conditional on \( X \) we see that \( Y \) has a \( \text{Bin}(n, X) \) distribution (i.e., we flip the coin \( n \) times). Depicted below are plots of the joint distribution for \( n = 5 \).

   ![Joint Distribution Plot](image)

   (a) What is the conditional PDF of \( f_{X|Y}(x|y) \)?
   (b) Compute the marginal PMF \( p_Y(y) \).

2. More generally, suppose \( X \sim \text{Beta}(a, b) \) (for some \( a, b > 0 \)) and \( Y|X \sim \text{Bin}(n, X) \). What is the conditional PDF of \( f_{X|Y}(x|y) \)?

3. Suppose \( X \sim \text{Unif}(0, Y) \) (continuous uniform) conditional on the value of \( Y \). Furthermore, suppose \( Y \) has PDF given by

   \[
   f_Y(y) = \frac{2}{y^2} 
   \]

   for \( y \geq 2 \) and 0 otherwise (Pareto distributed). What is the conditional PDF \( f_{Y|X}(y|x) \)?

4. Suppose \( X \sim \mathcal{N}(Y, 1) \) conditional on \( Y \) where \( Y \sim \mathcal{N}(0, 1) \). What is \( f_{Y|X}(y|x) \)?
18.1.5 Solutions

1. (a) Applying Bayes’s theorem we have

\[ f_{X|Y}(x|y) = \frac{f_{Y|X}(y|x)f_X(x)}{f_Y(y)} = \frac{n^y x^y (1-x)^{n-y}}{f_Y(y)} = C x^y (1-x)^{n-y}, \]

for \( x \in (0,1) \) and \( y = 0, \ldots, n \). So \( X|Y = y \sim \text{Beta}(y+1, n-y+1) \) and \( C = 1/B(y+1, n-y+1) \).

(b) We can either use the previous part to see that, for \( y = 0, \ldots, n \),

\[ p_Y(y) = \binom{n}{y} B(y+1, n-y+1) = \frac{n!}{y!(n-y)!} \cdot \frac{y!(n-y)!}{(n+1)!} = \frac{1}{n+1}, \]

or we can use LOTP to obtain

\[ p_Y(y) = \int_0^1 p_Y(y|x) f_X(x) \, dx = \int_0^1 \binom{n}{y} x^y (1-x)^{n-y} \, dx = \binom{n}{y} B(y+1, n-y+1). \]

Thus \( Y \sim \text{Unif}\{0,1, \ldots, n\} \).

2. Applying Bayes’s again we have

\[ f_{X|Y}(x|y) = \frac{f_{Y|X}(y|x)f_X(x)}{f_Y(y)} = C x^{a+y} (1-x)^{b+y-n}, \]

for \( x \in (0,1) \) and \( y = 0, \ldots, n \), so \( X|Y = y \sim \text{Beta}(a+y, b+n-y) \) and

\[ C = \frac{1}{B(a+y, b+n-y)}. \]

This is an example of what is called a conjugate prior distribution for Bernoulli sampling. We will state this as a theorem

**Theorem 63.** Suppose \( X \sim \text{B}(a,b) \) (the prior distribution) for some \( a,b > 0 \) and conditional on \( X \) let \( Y_1, \ldots, Y_n \) be a random sample from \( \text{Ber}(X) \) (i.e., \( Y = Y_1 + \cdots + Y_n \sim \text{Bin}(n,X) \)). Then the posterior distribution \( X|Y = y \) is \( \text{Beta}(a+y, b+n-y) \).

That is, if the prior distribution of the probability is Beta, and we take random Bernoulli samples then the posterior distribution is still Beta. This makes updating your beliefs very convenient.

3. Applying Bayes’s theorem we have

\[ f_{Y|X}(y|x) = \frac{f_{X|Y}(x|y)f_Y(y)}{f_X(x)} = \frac{C}{y^3}, \]
for \( y \geq 2 \), \( 0 < x < y \), and 0 otherwise. If \( x < 2 \) we have

\[
1 = \int_{2}^{\infty} \frac{C}{y^3} \, dy = \left[ \frac{C}{-2y^2} \right]_{2}^{\infty} = \frac{C}{8}
\]
giving \( C = 8 \). Otherwise, if \( x \geq 2 \), we have

\[
1 = \int_{x}^{\infty} \frac{C}{y^3} \, dy = \left[ \frac{C}{-2y^2} \right]_{x}^{\infty} = \frac{C}{2x^2}
\]
giving \( C = 2x^2 \).

4. Applying Bayes’s theorem with \( x, y \in \mathbb{R} \) we have

\[
f_{Y|X}(y|x) = \frac{f_{X|Y}(x|y)f_Y(y)}{f_X(x)} = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2} (x-y)^2} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} \frac{1}{f_X(x)} e^{-y^2/2} = e^{-y^2+xy} \propto e^{-(y-x/2)^2} = e^{-\frac{1}{2}z^2}.
\]

Thus \( Y|X = x \sim \mathcal{N} \left( \frac{x}{2}, \frac{1}{2} \right) \) and \( C = \frac{1}{\sqrt{\pi}} \).

18.1.6 Convolutions and 2d LOTUS

Suppose \( X, Y \) have a discrete joint distribution. Then we have already seen that

\[
P(X + Y = k) = \sum_{x+y=k} p_{X,Y}(x,y) = \sum_{x} p_{X,Y}(x,k-x)
\]

and if \( X, Y \) are independent this simplifies to

\[
P(X + Y = k) = \sum_{x+y=k} p_X(x)p_Y(y) = \sum_{x} p_X(x)p_Y(k-x).
\]

We now extend this idea to the continuous case. As such, suppose \( X, Y \) have joint PDF \( f_{X,Y} \), and let \( Z = X + Y \). We want to know the PDF of \( Z \). This is given by either of the following two formulas:

\[
f_Z(z) = \int_{-\infty}^{\infty} f_{X,Y}(x, z-x) \, dx = \int_{-\infty}^{\infty} f_{X,Y}(z-y, y) \, dy.
\]
If $X,Y$ are independent, we obtain:

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(x)f_Y(z-x) \, dx = \int_{-\infty}^{\infty} f_X(z-y)f_Y(y) \, dy.$$ 

The integral above is called the convolution of $f_X$ and $f_Y$ and is written:

$$f_Z(z) = (f_X \ast f_Y)(z).$$

To see this is indeed the PDF note that (letting $u = y + x$),

$$P(X + Y \leq z) = \int_{-\infty}^{\infty} \int_{-\infty}^{z-x} f_{X,Y}(x,y) \, dy \, dx = \int_{-\infty}^{\infty} \int_{-\infty}^{z} f_{X,Y}(x,u-x) \, du \, dx = \int_{-\infty}^{z} \int_{-\infty}^{\infty} f_{X,Y}(x,u-x) \, dx \, du.$$

Thus the convolution is the PDF, as it integrates to the CDF.

For discrete random variables, we proved linearity of expectation, and that independent random variables have 0 covariance. We now establish these results for variables with joint PDFs. To do this, we first state 2d LOTUS, which is proven in a measure theoretic course.

**Theorem 64 (2d LOTUS).** Let $X,Y$ be random variables, and let $g : \mathbb{R}^2 \to \mathbb{R}$. If $X,Y$ have joint PMF $p_{X,Y}$ then

$$E[g(X,Y)] = \sum_{x,y} g(x,y)p_{X,Y}(x,y).$$

If $X,Y$ have joint PDF $f_{X,Y}$ then

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

If $X,Y$ have joint PDF/PMF $f_{X,Y}$ then

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

As with our other results, 2d LOTUS extends to $n$-dimensional distributions with $g : \mathbb{R}^n \to \mathbb{R}$. Using LOTUS we will prove a few important results.

**Theorem 65 (Linearity of Expectation).** If $X,Y$ are random variables then $E[X + Y] = E[X] + E[Y]$. 

129
Proof. We will only prove the case where \( X, Y \) have a joint PDF. Then, (using 2d LOTUS with \( g(x, y) = x + y \))

\[
E[X + Y] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x + y) f_{X,Y}(x, y) \, dx \, dy \\
= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xf_{X,Y}(x, y) \, dx \, dy + \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} yf_{X,Y}(x, y) \, dx \, dy \\
= \int_{-\infty}^{\infty} xf_X(x) \, dx + \int_{-\infty}^{\infty} yf_Y(y) \, dy.
\]

This result can be extended to multiple variables by using \( n \)-dimensional LOTUS to obtain:

**Theorem 66 (Linearity of Expectation).** If \( X_1, \ldots, X_n \) are random variables then

\[
E[X_1 + \cdots + X_n] = E[X_1] + \cdots + E[X_n].
\]

Next we use 2d LOTUS to prove that independent random variables are uncorrelated (a term we will learn later).

**Theorem 67.** Let \( X, Y \) be independent RVs. Then \( E[XY] = E[X]E[Y] \).

Proof. Again we prove the case where \( X, Y \) have joint PDF \( f_{X,Y} \). Then we have (using 2d LOTUS with \( g(x, y) = xy \))

\[
E[XY] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xyf_{X,Y}(x, y) \, dx \, dy \\
= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xyf_X(x) \, dx \, dy \\
= \left( \int_{-\infty}^{\infty} xf_X(x) \, dx \right) \left( \int_{-\infty}^{\infty} yf_Y(y) \, dy \right) \\
= E[X]E[Y].
\]

Again we have the \( n \)-dimensional extension:

**Theorem 68.** If \( X_1, \ldots, X_n \) are independent RVs then

\[
E[X_1 \cdots X_n] = E[X_1] \cdots E[X_n].
\]

From these results we have the important corollary:

**Corollary 69.** If \( X_1, \ldots, X_n \) are independent then

\[
\text{Var}(X_1 + \cdots + X_n) = \text{Var}(X_1) + \cdots + \text{Var}(X_n).
\]
18.1.7 Convolution and 2d LOTUS Exercises

1. Given that $X, Y$ are independent Unif(0,1) random variables, compute the PDF of $Z = X + Y$.

2. Let $X, Y$ be i.i.d. Unif(0,1) random variables.
   
   (a) Compute $E[X + Y]$.
   
   (b) Compute $E[|X - Y|]$.
   
   (c) Compute $E[(X - Y)^2]$.

18.1.8 Solutions

1. We know ahead of time that $X + Y$ should have a positive density on $[0, 2]$ but let’s see what we obtain. Using the convolution formula we have

   $$f_Z(z) = \int_{-\infty}^{\infty} f_X(z-t)f_Y(t)\, dt.$$

   The integrand is not zero when $0 \leq z - t \leq 1$ and $0 \leq t \leq 1$. If $0 \leq z \leq 1$ we have

   $$f_Z(z) = \int_{0}^{z} 1\, dt = z$$

   and if $1 < z \leq 2$ we have

   $$f_Z(z) = \int_{z-1}^{1} 1\, dt = 2 - z.$$

   Thus the PDF is a triangle.

2. (a) $E[X + Y] = 1$ by linearity.
   
   (b) By 2d LOTUS with $g(x, y) = |x - y|$ we have

   $$E[|X - Y|] = \int_{0}^{1} \int_{0}^{1} |x - y|\, dy\, dx$$

   $$= 2 \int_{0}^{1} \int_{0}^{x} x - y\, dy\, dx$$

   $$= 2 \int_{0}^{1} \left[ xy - \frac{y^2}{2} \right]_{0}^{x} dx$$

   $$= \int_{0}^{1} x^2\, dx$$

   $$= \frac{1}{3}.$$
(c) By 2d LOTUS with \( g(x, y) = (x - y)^2 \) we have

\[
E[(X - Y)^2] = \int_0^1 \int_0^1 (x - y)^2 \, dy \, dx
\]

\[
= \int_0^1 \left[ \frac{(y - x)^3}{3} \right]_0^1 \, dx
\]

\[
= \int_0^1 x^3 + (1 - x)^3 \, dx
\]

\[
= \frac{1}{6}.
\]

### 18.1.9 Bivariate Normal Distribution

One of the most important bivariate distributions is the bivariate normal distribution. The simplest example occurs when \( X, Y \) are i.i.d. \( \mathcal{N}(0, 1) \) random variables. Then the joint PDF is given by

\[
f_{X,Y}(x, y) = \frac{1}{2\pi} e^{-\frac{1}{2}(x^2+y^2)}.
\]

The plot on the left shows this PDF. If we look at the distribution of \( U, V \) where \( U = X - 1 \) and \( V = Y + 1 \) we get the plot on the right.

The plot on the right is given by the PDF

\[
f_{U,V}(u, v) = \frac{1}{2\pi} e^{-\frac{1}{2}(u+1)^2+(v-1)^2}.
\]

In terms of contour plots, we have
In general, a bivariate normal distribution $U = (U, V)$ takes the form

$$U = AX + b$$

where $A$ is an invertible $2 \times 2$ matrix, $b$ is a $2 \times 1$ vector and $X$ is the $2 \times 1$ vector $\begin{pmatrix} X \\ Y \end{pmatrix}$ with $X, Y$ i.i.d. $\mathcal{N}(0, I)$. This clearly can be generalized to $n$ dimensions. As we saw above, when $b \neq 0$, it simply moves the PDF around $\mathbb{R}^2$. Thus, we will focus on the case when $b = 0$ and focus on $A$.

The first interesting fact about bivariate normal distributions is that the marginals are always normal. This comes from the following fact:

**Theorem 70.** If $X \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$ are independent, then $X + Y \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.

Although we could prove this with a long and nasty calculation using convolutions, we will give a short proof later using moment generating functions. Moreover, it is not hard to show (we may prove this later) that if

$$\begin{pmatrix} U \\ V \end{pmatrix} = A \begin{pmatrix} X \\ Y \end{pmatrix}$$

for $X, Y$ i.i.d. $\mathcal{N}(0, I)$ then

$$AA^T = \begin{pmatrix} \text{Var}(U) & \text{Cov}(U, V) \\ \text{Cov}(U, V) & \text{Var}(V) \end{pmatrix}.$$  

This is called the covariance matrix of $U, V$, and often denoted by $\Sigma$. Interestingly, the conditional distributions are also normal, but we will not show this.

To figure out the joint PDF of $U, V$ we will need the following theorem and corollary which basically amount to multivariate $u$-substitution.

**Theorem 71.** Let $g$ be a continuously differentiable bijection between regions $R, S \subset \mathbb{R}^2$. Suppose $X, Y$ has joint PDF $f_{X,Y}$ positive on $R$, and let $(U, V) = g(X, Y)$. If $\frac{d(u,v)}{d(x,y)} \neq 0$ for all $(x, y) \in R$ then

$$f_{U,V}(u,v) = f_{X,Y}(g^{-1}(u,v)) \left| \frac{d(x,y)}{d(u,v)}(u,v) \right|,$$

for all $(u, v) \in S$.

We obtain the following corollary by letting $g$ be a linear map:

**Corollary 72.** Let $X = (X_1, \ldots, X_n)$ be a random vector (a list of $n$ random variables) with joint PDF $f_X$ and let $A$ be an invertible $n \times n$ matrix. Then $U = AX$ has joint PDF given by

$$f_U(u) = \frac{1}{|\det A|} f_X(A^{-1}u).$$
To use this, let’s write the joint PDF of $X$ in the following way:

$$f_X(x) = \frac{1}{2\pi} e^{-\frac{1}{2}x^T x}.$$  

Then we can apply the corollary above to find the joint PDF of $U = AX$:

$$f_U(u) = \frac{1}{\det A} f_X(A^{-1} u) = \frac{1}{2\pi \det A} e^{-\frac{1}{2}u^T A^{-T} A^{-1} u}.$$  

If we let $\Sigma = AA^T$ then $\Sigma^{-1} = A^{-T} A^{-1}$ and $\det \Sigma = (\det A)^2$ giving

$$f_U(u) = \frac{1}{2\pi \sqrt{\det \Sigma}} e^{-\frac{1}{2}u^T \Sigma^{-1} u}.$$  

If we consider $b \neq 0$ we get the following full formula for the general bivariate normal distribution:

$$f_U(u) = \frac{1}{2\pi \sqrt{\det \Sigma}} e^{-\frac{1}{2}(u-b)^T \Sigma^{-1} (u-b)}.$$  

We denote this distribution as $\mathcal{N}(b, \Sigma)$. It is typical to use $\mu$ instead of $b$ as the name of the first parameter.

18.1.10 Bivariate Normal Exercises

1. Let $A = \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix}$. The plot and contour plot for $U = AX$ are given below (where $X$ is a vector of i.i.d. $\mathcal{N}(0, 1)$ as above):

   (a) What are the marginal distributions? Are the two components of $U$ independent?
   (b) What is the joint PDF?
   (c) What is the conditional distribution $U|V = v$?
2. Let \( \mathbf{A} = \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix} \), \( \mathbf{b} = \begin{pmatrix} 3 \\ 4 \end{pmatrix} \), and let \( \mathbf{U} = \mathbf{A} \mathbf{X} \). What are the marginal distributions?

3. Let \( \mathbf{A} = \begin{pmatrix} 4 & -1 \\ 4 & 1 \end{pmatrix} \). The plot and contour plot for \( \mathbf{A} \mathbf{X} \) are given below (where \( \mathbf{X} \) is a vector of i.i.d. \( \mathcal{N}(0,1) \) as above):

   ![Plot and Contour Plot](image)

(a) What is the joint PDF?
(b) What are the marginal distributions?
(c) Are the components \( U, V \) of \( \mathbf{A} \mathbf{X} \) independent?

18.1.11 Solutions

1. (a) Here \( U = 3X \) and \( V = Y \) so \( U \sim \mathcal{N}(0,9), V \sim \mathcal{N}(0,1) \) and they are independent.
   (b) By independence we immediately have
   \[
   f_{U,V}(u,v) = \frac{1}{6\pi} e^{-\frac{1}{2}((\frac{u}{3})^2+v^2)}.
   \]
   (c) Same as \( U \) since \( U, V \) are independent.

2. Here \( U = 3X + 3 \) and \( V = Y + 4 \) so \( U \sim \mathcal{N}(3,9) \) and \( V \sim \mathcal{N}(4,1) \).

3. (a) Using our formula we have
   \[
   \Sigma = \mathbf{A} \mathbf{A}^T = \begin{pmatrix} 17 & 15 \\ 15 & 17 \end{pmatrix}, \quad \det(\Sigma) = 64, \quad \Sigma^{-1} = \frac{1}{64} \begin{pmatrix} 17 & -15 \\ -15 & 17 \end{pmatrix}.
   \]
   Thus the joint PDF is given by
   \[
   f_{\mathbf{u}}(\mathbf{u}) = \frac{1}{2\pi} \cdot 8 \exp \left( -\frac{1}{2} \mathbf{u}^T \Sigma^{-1} \mathbf{u} \right)
   \]
or as components

\[ f_{U,V}(u,v) = \frac{1}{16\pi} \exp \left( -\frac{1}{2} \cdot \frac{64}{64}(17u^2 - 30uv + 17v^2) \right). \]

(b) Using \( \Sigma \) above we see \( U, V \sim \mathcal{N}(0, 17) \).

(c) No, the marginals don’t multiply to give the joint PDF.

Another way to derive the results above is to notice that

\[
\begin{pmatrix}
4 & -1 \\
4 & 1
\end{pmatrix}
= \begin{pmatrix}
\sqrt{2}/2 & -\sqrt{2}/2 \\
\sqrt{2}/2 & \sqrt{2}/2
\end{pmatrix}
\begin{pmatrix}
4\sqrt{2} & 0 \\
0 & \sqrt{2}
\end{pmatrix}.
\]

A 45° counterclockwise rotation matrix multiplied by a diagonal matrix. This says that we obtain \( U, V \) by stretching \( X, Y \) and then rotating. Using something called the Singular Value Decomposition (or SVD) from linear algebra, you can prove that every invertible matrix applied to \( X, Y \) effectively just scales and then rotates.
19 Lecture 19

19.1 More on Expectation

19.1.1 Conditional Expectation

For both continuous and absolutely continuous random variables we have defined the expectation $E[X]$. We also saw that if $X, Y$ have a joint distribution, then we can compute the conditional distributions of $X|Y$ and $Y|X$. Using these distributions, we can compute expectations, and they are called the conditional expectation.

**Definition 73** (Conditional Expectation). Let $X|Y = y$ have the conditional PDF $f_{X|Y}(x|y)$. Then the conditional expectation of $X$ given $Y = y$ is defined by

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

If instead $X|Y = y$ has conditional PMF $p_{X|Y}(x|y)$ then we have

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

Analogously we define conditional variance.

**Definition 74** (Conditional Variance). We define

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

As before, the second equality above holds by foil realizing that $E[X|Y = y]$ is a constant. Although it looks a little weird, the conditional expectation and variance are no different from the regular expectation and variance, since conditional PDFs/PMFs are bonafide PDFs/PMFs. The conditional expectation and conditional variance of $X$ given $Y = y$ are just the expectation and variance of the distribution of $X|Y = y$. As such, results such as linearity and LOTUS still hold. Here we state a useful result whose proof we will not give.

**Theorem 75.** Let $X, Y$ be random variables and let $g : \mathbb{R}^2 \to \mathbb{R}$ be a function. If $Z = r(X, Y)$ then

$$E[Z|Y = y] = E[r(X, Y)|Y = y] = E[r(X, y)|Y = y].$$

The above says that if we are conditioning on $Y = y$, then we can always replace $Y$ by a constant in the expression we are taking the expectation of.

The next definition makes perfect sense mathematically, but takes a little while to grasp. Note that $E[X|Y = y]$ will give us a different real number for each possible value of $y$. That is, we can define a function $g$ by

$$g(y) = E[X|Y = y]$$

for each possible value $y$ of $Y$. As such, we can create a new random variable $g(Y)$ which we denote $E[X|Y]$. Remember that $E[X|Y = y]$ is a number, and $E[X|Y]$ is a random variable. This definition allows the following very useful (and actually intuitive) theorem:
**Theorem 76** (Law of Iterated Expectation). If $X,Y$ are random variables then

$$E[E[X|Y]] = E[X].$$

**Proof.** We prove this for joint PMFs and joint PDFs. Suppose $X,Y$ have joint PMF $p_{X,Y}$. Then we have

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

$$= \sum_y \sum_x xp_{X|Y}(x|y)p_Y(y)$$

$$= \sum_y \sum_x xp_{X,Y}(x,y)$$

$$= \sum_x xp_X(x)$$

$$= E[X].$$

If instead $X,Y$ have a joint PDF $f_{X,Y}$ then we have

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

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

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

$$= \int_{-\infty}^{\infty} xf_X(x) dx$$

$$= E[X].$$

\[\square\]

### 19.1.2 Conditional Expectation Exercises

1. The following table describes the joint PMFs of $X,Y$:

<table>
<thead>
<tr>
<th>$X$</th>
<th>3</th>
<th>4</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.1</td>
<td>0.3</td>
<td>.4</td>
</tr>
<tr>
<td>2</td>
<td>0.4</td>
<td>0.2</td>
<td>.6</td>
</tr>
<tr>
<td>Total</td>
<td>0.5</td>
<td>0.5</td>
<td>1</td>
</tr>
</tbody>
</table>

(a) What is $E[X|Y = y]$ for $y = 3,4$?
(b) What is $E[Y|X = x]$ for $x = 1, 2$?
(c) What is $E[X|Y]$?

2. Let $X \sim \text{Unif}(0, 1)$ and conditional on $X = x$, let $Y|X = x \sim \text{Unif}(x, 1)$.
   (a) What is $E[Y|X = x]$ for $x \in (0, 1)$?
   (b) What is $E[Y]$?

3. Let $X \sim \text{Beta}(a, b)$ and let $Y|X = x \sim \text{Bin}(n, x)$.
   (a) What is $E[Y]$?
   (b) What is $E[X|Y = y]$?
   (c) What is $E[E[X|Y]]$?

4. Suppose $X, Y$ are independent with PDFs $f_X$ and $f_Y$. What is $E[X|Y = y]$?

5. Let $X = c$. That is $X$ is the constant random variable only taking on the value $c \in \mathbb{R}$ ($p_X(c) = 1$). Let $Y$ be any other random variable with PMF $p_Y$. (All the results will hold for any $Y$ actually.)
   (a) Compute $E[X]$.
   (b) Show that $X, Y$ are independent.
   (c) Compute $E[X|Y = y]$ for some value $y$ with $p_Y(y) > 0$.

   (b) More generally, show that $E[h(X)Y|X] = h(X)E[Y|X]$.

19.1.3 Solutions

1. (a) 
   
   $E[X|Y = 3] = 1 \cdot \frac{0.1}{0.5} + 2 \cdot \frac{4}{0.5} = \frac{9}{5}$
   
   and 
   
   $E[X|Y = 4] = 1 \cdot \frac{0.3}{0.5} + 2 \cdot \frac{2}{0.5} = \frac{7}{5}$
   
   (b) 
   
   $E[Y|X = 1] = 3 \cdot \frac{0.1}{0.4} + 4 \cdot \frac{0.3}{0.4} = \frac{15}{4}$
   
   and 
   
   $E[Y|X = 2] = 3 \cdot \frac{4}{0.6} + 4 \cdot \frac{2}{0.6} = \frac{20}{6}$.

   (c) Let $Z = E[X|Y]$. Then $p_Z(9/5) = 1/2$ and $p_Z(7/5) = 1/2$.

2. (a) Note that $Y|X = x \sim \text{Unif}(x, 1)$ so $E[Y|X = x] = \frac{1 + x}{2}$.
(b) Using the law of iterated expectation we have

\[ E[Y] = E[E[Y|X]] = E \left[ \frac{1 + X}{2} \right] = \frac{1}{2} + \frac{1}{4} = \frac{3}{4}. \]

3. (a) Using the law of iterated expectation we have

\[ E[Y] = E[E[Y|X]] = E[nX] = \frac{na}{a+b}. \]

(b) As we saw last class, we know that \( X|Y = y \sim B(a + y, b + n - y) \). Thus

\[ E[X|Y = y] = \frac{a + y}{a + b + n}. \]

(c) From the law of iterated expectation this is just \( E[X] = \frac{a}{a+b} \). Alternatively we could directly compute this:

\[
E[E[X|Y]] = E \left[ \frac{a + Y}{a + b + n} \right] = \frac{a}{a + b + n} + \frac{an}{a(a + b + n)} = \frac{a(a + b + n)}{(a + b)(a + b + n)} = \frac{a}{a + b}.
\]

4. From last class we saw that \( f_{X|Y} = f_X \) since \( X, Y \) are independent. Thus the answer is \( E[X] \).

5. (a) \( E[X] = cp_X(c) = c \)

(b) \( p_{X,Y}(x, y) = p_X(x)p_Y(y) \).

(c) Since \( X, Y \) are independent \( E[X|Y = y] = E[X] = c \). This is the same as saying \( E[c|Y = y] = c \), which at first may look a little confusing, since you need to compute the joint distribution of a number and \( Y \), but we showed this is possible in this problem.

6. (a) Note that \( E[XY|X = x] = E[xY|X = x] = xE[Y|X = x] = xE[Y|X] \) so \( E[XY|X] = XE[Y|X] \).

(b) Similarly \( E[h(X)Y|X = x] = E[h(x)Y|X = x] = h(x)E[Y|X = x] = h(X)E[Y|X] \).
19.1.4 A Few Questions About Coins

1. You are going to flip a coin with heads probability $p$ until you get a heads. Let $X_1$ be the indicator random variable of whether the first flip is heads, and let $Y$ denote the number of flips up to and including the first heads.

   (a) Compute $E[Y|X_1 = 1]$ and $E[Y|X_1 = 0]$.
   (b) Compute $E[Y]$.

2. Let $X$ be the number of flips of a fair coin until you get two heads in a row, and let $Y$ be the number of flips of a fair coin until you get a heads followed by a tails. Which is larger: $E[X]$ or $E[Y]$?

19.1.5 Solutions

1. (a) We have $E[Y|X_1 = 1] = 1$ and $E[Y|X_1 = 0] = 1 + E[Y]$, since getting a tail means you flip once and restart.

2. Let $X_1, X_2$ denote the first two flips on our way to 2 heads. Then we have

$$E[X] = E[E[X|X_1]]$$

$$= \frac{1}{2} E[X|X_1 = 1] + \frac{1}{2} E[X|X_1 = 0]$$

$$= \frac{1}{2} E[X|X_1 = 1] + \frac{1}{2} (1 + E[X]).$$

We haven’t covered the following natural extension, but we could take this further:

$$E[X|X_1 = 1] = E[E[X|X_1 = 1, X_2]]$$

$$= \frac{1}{2} E[X|X_1 = 1, X_2 = 1] + \frac{1}{2} E[X|X_1 = 1, X_2 = 0]$$

$$= \frac{1}{2} \cdot 2 + \frac{1}{2} (2 + E[X]).$$

This gives

$$E[X] = \frac{1}{2} \left( \frac{1}{2} \cdot 2 + \frac{1}{2} (2 + E[X]) \right) + \frac{1}{2} (1 + E[X])$$

which simplifies to

$$\frac{1}{4} E[X] = \frac{3}{2} \implies E[X] = 6.$$
A similar style of calculation would show that $E[Y] = 4$. Alternatively, we could observe this is the sum of two $\text{Geom}(1/2)$ random variables, the first being number of flips till getting heads. The second being number of flips after first heads before getting tails. Thus $E[X] = 6 > 4 = E[Y]$.

19.1.6 Covariance and Correlation

When computing the variance of the sum of random variables we saw the covariance. We will now study it more carefully. Recall the definition:

**Definition 77 (Covariance).** If $X, Y$ are random variables then the covariance is defined by


Later we will prove many properties of the covariance. Note that covariance is positive if $X, Y$ tend to be larger or smaller than their mean simultaneously. The problem with covariance is that it scales with $X, Y$. For example, we will see that $\text{Cov}(aX, Y) = a \text{Cov}(X, Y)$. Thus, if $X$ is initially measured in kilometers, and then we change it to meters the covariance is scaled by a factor of 1000. Thus if we learn

$$\text{Cov}(X, Y) = 1000 \quad \text{and} \quad \text{Cov}(U, V) = 1000000$$

we don’t know which pair of variables has a stronger relationship. For this reason, we introduce the correlation:

**Definition 78 (Correlation).** If $X, Y$ are random variables the the correlation is defined by

$$\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{E[X - \mu_X, Y - \mu_Y]}{\sigma_X \sigma_Y} = E\left[\frac{X - \mu_X}{\sigma_X}, \frac{Y - \mu_Y}{\sigma_Y}\right].$$

If we scale $X$ then both the covariance and the standard deviation are scaled, so the correlation is unchanged (i.e., it is scale-invariant). In fact, we will prove the following fact:

**Theorem 79.** For any random variables $X, Y$,

$$-1 \leq \rho(X, Y) \leq 1.$$  

If $\rho(X, Y) = 1$ then $Y = aX + b$ with $a > 0$. If $\rho(X, Y) = -1$ then $Y = aX + b$ with $a < 0$.

The above is equivalent to the statement

$$E[UV]^2 \leq E[U^2]E[V^2]$$

which is a version of the Cauchy-Schwarz inequality.

We say $X, Y$ are uncorrelated if $\text{Cov}(X, Y) = 0$ or equivalently that $\rho(X, Y) = 0$. We proved earlier that independent random variables are uncorrelated. We will see in the exercises that the converse is false.
20 Lecture 20

20.1 More on Expectation

20.1.1 Review Exercises

1. What is the expected number of rolls to get 3 immediately followed by 4 when repeatedly rolling a fair 6-sided die?

2. If \(X, Y\) are random variables with joint PDF \(f_{X,Y}\), guess the formula for \(f_W\) where \(W = X - Y\).

3. Suppose \(X_1, \ldots, X_n\) form a random sample from a Unif(0, 1) distribution. Let \(U = \max_i(X_i)\) and let \(V = \min_i(X_i)\).

   (a) What is the probability that \(U \leq u\)?
   (b) What is the probability that \(V \geq v\)?
   (c) What is the probability that \(U \leq u\) and \(V \geq v\)?
   (d) What is the probability that \(U \leq u\) and \(V \leq v\)?
   (e) What is the joint PDF \(f_{U,V}\)?

20.1.2 Solutions

1. Let \(X\) denote the number of rolls to get 3 then 4, and \(Y\) the number of remaining rolls to get a 3 then four assuming your previous roll was 3. Then we have

   \[
   E[X] = 1 + \frac{5}{6}E[X] + \frac{1}{6}E[Y] \\
   E[Y] = 1 + \frac{4}{6}E[X] + \frac{1}{6}E[Y].
   \]

Solving this system gives

\[
E[X] = 1 + \frac{5}{6}E[X] + \frac{1}{6} \left( \frac{6}{5} + \frac{4}{5}E[X] \right) = \frac{6}{5} + \frac{29}{30}E[X] \implies E[X] = 36.
\]

A similar calculation shows the expected number of rolls to get two 3’s in a row is 42.

2. \(f_W(w) = \int_{-\infty}^{\infty} f_{X,Y}(t, t - w) \, dt\).

3. (a) \(P(X_1 \leq u, \ldots, X_n \leq u) = \prod_{i=1}^{n} P(X_i \leq u) = u^n\) for \(u \in (0, 1)\).
   (b) \(P(X_1 \geq v, \ldots, X_n \geq v) = \prod_{i=1}^{n} P(X_i \geq v) = (1 - v)^n\) for \(v \in (0, 1)\).
(c) $P(v \leq X_1 \leq u, \ldots, v \leq X_n \leq u) = \prod_{i=1}^{n} P(v \leq X_i \leq u) = (u - v)^n$ for $0 < v < u < 1$.

(d) $P(U \leq u, V \leq v) = P(U \leq u) - P(U \leq u, V \geq v) = u^n - (u - v)^n$ for $0 < v < u < 1$.

(e) Taking partial derivatives of the joint CDF above gives

$$\frac{\partial^2}{\partial u \partial v} F_{U,V}(u,v) = n(n-1)(u-v)^{n-2}.$$}

We will note something now that will come up again later. Fix any $\epsilon > 0$. Part a) shows that $P(|U - 1| > \epsilon) \to 0$ as $n \to \infty$. We say “$U$ converges to 1 in probability”.

20.1.3 Covariance and Correlation

**Definition 80** (Correlation). If $X, Y$ are random variables the the correlation is defined by

$$\rho(X,Y) = \frac{\text{Cov}(X - \mu_X, Y - \mu_Y)}{\sigma_X \sigma_Y} = \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y} = E \left[ \frac{X - \mu_X}{\sigma_X} \cdot \frac{Y - \mu_Y}{\sigma_Y} \right].$$

If we scale $X$ then both the covariance and the standard deviation are scaled, so the correlation is unchanged (i.e., it is scale-invariant). In fact, we will prove the following fact:

**Theorem 81.** For any random variables $X,Y$ with $\sigma_X, \sigma_Y > 0$,

$$-1 \leq \rho(X,Y) \leq 1.$$  

If $\rho(X,Y) = 1$ then $Y = aX + b$ with $a > 0$. If $\rho(X,Y) = -1$ then $Y = aX + b$ with $a < 0$.

**Proof.** Define $U,V$ by

$$U = \frac{X - \mu_X}{\sigma_X} \quad \text{and} \quad V = \frac{Y - \mu_Y}{\sigma_Y}.$$  

Then $\rho(U,V) = \rho(X,Y)$ (see exercises).

$$E[(U + V)^2] = E[U^2] + E[V^2] + 2E[UV] = \text{Var}(U) + \text{Var}(V) + 2\rho(U,V) = 2 + 2\rho(U,V)$$  

and


This shows

$$2 + 2\rho(U,V) \geq 0 \quad \text{and} \quad 2 - 2\rho(U,V) \geq 0.$$  

Together these imply $-1 \leq \rho(U,V) \leq 1$. The only way either side can be an equality is if $E[(U + V)^2] = 0$ or $E[(U - V)^2] = 0$. Note that (the following only holds with probability 1, but we will ignore this detail)

$$E[(U + V)^2] = 0 \implies \frac{X - \mu_X}{\sigma_X} = -\frac{Y - \mu_Y}{\sigma_Y}.$$
and
\[ E[(U - V)^2] = 0 \Rightarrow \frac{X - \mu_X}{\sigma_X} = \frac{Y - \mu_Y}{\sigma_Y} \]
giving the result.

The above is equivalent to the statement
\[ E[UV]^2 \leq E[U^2]E[V^2] \]
which is a version of the Cauchy-Schwarz inequality.

We say \( X,Y \) are uncorrelated if \( \text{Cov}(X,Y) = 0 \) or equivalently that \( \rho(X,Y) = 0 \). We proved earlier that independent random variables are uncorrelated. We will see in the exercises that the converse is false.

### 20.1.4 Covariance and Correlation Exercises

1. Show that \( \text{Cov} \) is a symmetric bilinear form. That is:
   (a) \( \text{Cov}(X,Y) = \text{Cov}(Y,X) \),
   (b) \( \text{Cov}(aX,Y) = a\text{Cov}(X,Y) \),
   (c) \( \text{Cov}(X + Z,Y) = \text{Cov}(X,Y) + \text{Cov}(Z,Y) \).

   Also show \( \text{Cov}(X + b,Y) = \text{Cov}(X,Y) \).

2. Prove that
\[
\text{Cov}\left(\frac{X - \mu_X}{\sigma_X}, \frac{Y - \mu_Y}{\sigma_Y}\right) = \frac{\text{Cov}(X,Y)}{\sigma_X\sigma_Y} = E\left[\frac{X - \mu_X}{\sigma_X} \cdot \frac{Y - \mu_Y}{\sigma_Y}\right].
\]

3. Show that if \( U,V \) are standardized versions of \( X,Y \), respectively, then \( \rho(X,Y) = \rho(U,V) \).

4. Show that \( Y = aX + b \) with \( a \neq 0 \) implies \( |\rho(X,Y)| = 1 \).

5. Let \( X \sim \mathcal{N}(0,1) \) and let \( Y = X^2 \).
   (a) Are \( X,Y \) independent?
   (b) Are \( X,Y \) uncorrelated?
20.1.5 Solutions

1. (a) Just swap $X, Y$ in definition.
(b) Note
(c) Using linearity of expectation we have
Finally, we have

2. Since standardized random variables have zero mean, we see the last term equals the first term. To see the last equals the middle, simply factor out $\frac{1}{\sigma_X \sigma_Y}$.

3. Standardized random variables have zero mean, and standard deviation 1. Thus standardizing standardized random variables does nothing. This shows
$$\rho(X, Y) = \text{Cov}(U, V) = \text{Cov} \left( \frac{U - \mu_U}{\sigma_U}, \frac{V - \mu_V}{\sigma_V} \right) = \rho(U, V).$$

4. Note that
$$\frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{\text{Cov}(X, aX + b)}{|a|\sigma_X^2} = \frac{a}{|a|},$$
which is 1 or $-1$ depending on the sign of $a$.

5. (a) No. Knowing $X$ determines $Y$ exactly.
(b) Yes. $E[X] = 0$ and $E[XY] = E[X^3] = 0$ since
$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^3 e^{-\frac{1}{2}x^2} \, dx$$
has an odd integrand. Thus
$$E[XY] - E[X]E[Y] = 0 - 0 = 0.$$

20.1.6 Moment Generating Functions

Earlier we defined the following function:

**Definition 82** (Moment Generating Function). For a random variable $X$, define the moment generating function (MGF) $M_X(t)$ by
$$M_X(t) = E[e^{tX}].$$
We say the moment generating function exists if the expectation $E[e^{tX}]$ exists for all $t$ in some open interval $(a, b)$ containing 0.
As we will see in the exercises, moment generating functions have many nice properties. The following is beyond the scope of this class:

**Theorem 83.** If random variables $X, Y$ have the same moment generating function in some interval $(a, b)$ containing 0, then they have the same distribution.

The above theorem essentially says that if the MGF exists, then we “un-MGF” to get the distribution back (i.e., the inverse Laplace transform exists). The reason it is called a moment generating function is the following theorem whose proof is also beyond the scope of the class:

**Theorem 84.** If $X$ has an MGF $M_X$ (i.e., the MGF of $X$ exists) then

$$M_X(t) = E[e^{tX}] = E \left[ \sum_{k=0}^{\infty} \frac{t^k X^k}{k!} \right] = \sum_{k=0}^{\infty} \frac{t^k}{k!} E[X^k].$$

Furthermore, differentiating $n$ times and evaluating at 0 gives

$$M_X^{(n)}(0) = E[X^n].$$

### 20.1.7 Moment Generating Function Exercises

1. Compute the MGF of $X$ for the following distributions:
   (a) $X \sim \text{Ber}(p)$,
   (b) $X \sim \text{Bin}(n, p)$,
   (c) $X \sim \mathcal{N}(0, 1)$.

2. Let $X$ have MGF $M_X(t)$ and $Y$ have MGF $M_Y(t)$.
   (a) Give the MGF of $aX + b$ where $a \neq 0$.
   (b) Give the MGF of $X + Y$ assuming $X, Y$ are independent.

3. Use MGFs to prove that if $X_1, \ldots, X_n \sim \text{Ber}(p)$ are a random sample then $X_1 + \cdots + X_n \sim \text{Bin}(n, p)$.

4. (a) Let $X \sim \mathcal{N}(\mu, \sigma^2)$. What is $M_X(t)$?
   (b) Let $X \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$ be independent. What is $M_{X+Y}(t)$?

### 20.1.8 Solutions

1. (a)
   $$E[e^{tX}] = pe^t + (1 - p).$$

147
\[ E[e^{tX}] = \sum_{k=0}^{n} e^{kt} \binom{n}{k} p^k (1 - p)^{n-k} = \sum_{k=0}^{n} \binom{n}{k} (pe^t)^k (1 - p)^{n-k} = (pe^t + (1 - p))^n. \]

(c)

\[
E[e^{tX}] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{tx} e^{-\frac{1}{2}x^2} \, dx
= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(x-t)^2 + \frac{t^2}{2}} \, dx
= e^{\frac{t^2}{2}}.
\]

2. (a)

\[ M_{aX+b}(t) = E[e^{(aX+b)t}] = e^{tb} E[e^{atX}] = e^{tb} M_X(at). \]

(b)

\[ M_{X+Y}(t) = E[e^{t(X+Y)}] = E[e^{tX}e^{tY}] = E[e^{tX}]E[e^{tY}] = M_X(t)M_Y(t). \]

More generally we have the following:

**Theorem 85.** If \( X_1, \ldots, X_n \) are independent random variables each with MGF \( M_{X_i}(t) \) and \( Y = X_1 + \cdots + X_n \) then

\[ M_Y(t) = \prod_{i=1}^{n} M_{X_i}(t). \]

3. Let \( X \sim \text{Ber}(p) \) and \( Y \sim \text{Bin}(n,p) \). We showed \( M_X(t)^n = M_Y(t) \). Since \( M_X(t)^n \) is the moment generating function of \( X_1 + \cdots + X_n \), the result follows.

4. (a) If \( Z \sim \mathcal{N}(0,1) \) then \( X = \sigma Z + \mu \sim \mathcal{N}(\mu, \sigma^2) \) so

\[ M_X(t) = e^{\mu t} e^{-(\sigma t)^2/2} = e^{\mu t + \sigma^2 t^2/2}. \]

(b) We have

\[ M_{X+Y}(t) = e^{t\mu_1 + \sigma_1^2 t^2/2} e^{t\mu_2 + \sigma_2^2 t^2/2} = e^{t(\mu_1 + \mu_2) + (\sigma_1^2 + \sigma_2^2) t^2/2}. \]

This shows the sum has a \( \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2) \) distribution.
21 Lecture 21

21.1 More on Expectation

21.1.1 Review Exercises

1. (a) Compute $M_X(t)$ when $X \sim \text{Unif}(a,b)$.
   (b) Compute $M_X(t)$ when $X \sim \text{Geom}(p)$.
   (c) Compute $M_X(t)$ when $X \sim \text{NBin}(r,p)$.

2. Fix $a,b \in \mathbb{R}$, $\sigma > 0$, and suppose $Y_k \sim \mathcal{N}(ka + b, \sigma^2)$ for $k = 1, \ldots, 1000$ are independent.
   (a) Give a function $g : \mathbb{R}^3 \to \mathbb{R}$ so that $E[g(Y_1, Y_2, Y_3)] = a$.
   (b) Give $\epsilon$ to guarantee that
   $$P(|a - g(Y_1, Y_2, Y_3)| < \epsilon) \geq 0.997.$$
   (c) Suppose you have $Y_1$ and could have one of $Y_k$ for $k = 2, \ldots, 1000$ to make a formula to estimate $a$. Which would you pick to give the best estimate?
   (d) If you had $Y_1$, $Y_2$, and $Y_6$ what $g$ would you choose (as in part a)?

21.1.2 Solutions

1. (a) We have
   $$E[e^{tX}] = \frac{1}{b-a} \int_a^b e^{tx} \, dx = \frac{e^{tb} - e^{ta}}{t(b-a)}.$$
   (b) We have
   $$E[e^{tX}] = \sum_{k=1}^\infty e^{tk}p(1-p)^{k-1} = \frac{p}{1-p} \sum_{k=1}^\infty (e^t(1-p))^k = \frac{pe^t}{1 - e^t(1-p)},$$
   valid when $e^t(1-p) < 1$, i.e., $t < -\log(1-p)$.
   (c) We can guess the answer since $r$ independent $\text{Geom}(p)$ random variables summed together gives an $\text{NBin}(r,p)$ distribution. Alternatively, we can do the following
calculation:

\[ E[e^{tX}] = \sum_{k=r}^{\infty} \binom{k}{k-r} e^{tk} p^r (1 - p)^{k-r} \]

\[ = \sum_{k=r}^{\infty} \binom{k}{k-r} (pe^t)^r (1 - p)e^{t(k-r)} \]

\[ = (pe^t)^r \sum_{k=r}^{\infty} \binom{-r}{k-r} (1 - p)e^{t(k-r)} \]

\[ = (pe^t)^r (1 - (1 - p)e^t)^{-r} \quad \text{(Newton’s Binomial Formula)} \]

\[ = \left( \frac{pe^t}{1 - (1 - p)e^t} \right)^r, \]

where we need \( t < -\log(1 - p) \) to apply the Binomial formula.

2. (a) Some potential choices would be

\[ g_1(y_1, y_2, y_3) = y_2 - y_1 \quad \text{and} \quad g_2(y_1, y_2, y_3) = \frac{y_3 - y_1}{2}. \]

(b) Using the notation from the previous part, note that

\[ g_1(Y_1, Y_2, Y_3) = Y_2 - Y_1 \sim \mathcal{N}(a, 2\sigma^2), \]

so we need \( \epsilon = 3\sqrt{2\sigma^2} = 3\sqrt{2}\sigma \). Alternatively, note that

\[ g_2(Y_1, Y_2, Y_3) = \frac{Y_3 - Y_1}{2} \sim \mathcal{N}(a, \sigma^2/2), \]

so we need \( \epsilon = 3\sqrt{\sigma^2/2} = \frac{3}{\sqrt{2}}\sigma \). Thus \( g_2 \) is superior to \( g_1 \) in this respect.

(c) Choose \( Y_{1000} \) and use \( g(Y_1, Y_{1000}) = \frac{Y_{1000} - Y_i}{999} \). Then our estimator has the lowest variance of the possible choices given.

(d) It turns out the best choice is

\[ g(Y_1, Y_2, Y_6) = \frac{(-2)Y_1 + (-1)Y_2 + 3Y_6}{14}, \]

where the weight on \( Y_i \) is the distance from \( i \) to the mean \( \frac{1+2+6}{3} = 3. \) The denominator is the sum of the squares of the coefficients. This can be found by calculus methods or by linear algebra. The formula, when generalized to arbitrary \( x \) and \( y \)-values, is called the ordinary least squares estimator for the slope coefficient \( a \). Requiring \( E[g(Y_1, \ldots, Y_n)] = a \) means the estimator is unbiased. In this problem we are looking for a BLUE estimator. This means best linear unbiased estimator. Best means lowest variance, and linear means that \( g \) is a linear function of the \( Y_i \).
21.2 Limit Theorems

21.2.1 Markov and Chebyshev for Absolutely Continuous RVs

We have already proven these results for discrete random variables when investigating the meaning of standard deviation. Now we establish these results for random variables with PDFs.

**Theorem 86** (Markov’s Inequality). If $X$ is a non-negative random variable then for any $a \geq 0$ we have

$$aP(X \geq a) \leq E[X].$$

**Proof.** We give the proof for $X$ with PDF $f_X$. Note that

$$E[X] = \int_0^\infty xf_X(x) \, dx \geq \int_a^\infty xf_X(x) \, dx \geq a\int_a^\infty f_X(x) \, dx = aP(X \geq a).$$

**Corollary 87** (Chebyshev’s Inequality). If $X$ is a random variable such that $\mu_X$ and $\sigma_X$ exist then

$$P(|X - \mu_X| > k\sigma) \leq \frac{1}{k^2}$$

and for $a > 0$,

$$P(|X - \mu_X| > a) \leq \frac{\text{Var}(X)}{a^2}.$$

**Proof.** Let $Y = (X - \mu_X)^2$, note that for $a > 0$

$$P(|X - \mu_X| > a) = P(|X - \mu_X|^2 > a^2),$$

and then apply Markov’s inequality.

As before, we have seen that Chebyshev’s inequality is very powerful since it applies to any random variable with a mean and variance. That said, this is also its weakness, as it must be a weak enough bound to accommodate such generality. In our next set of topics, we will learn about normal approximations, and obtain much stronger approximate bounds.

21.2.2 The Law of Large Numbers

One way to think of a coin having heads probability $p$ is to imagine repeatedly flipping it and asking what proportion of the flips yield heads. We will now add rigor to this intuition.

**Theorem 88** (Strong Law of Large Numbers). Let $X_n$ for $n = 1, 2, \ldots$ be an i.i.d. sequence of random variables (defined on the same sample space), and suppose $E[X_1] = \mu$. Then

$$P\left(\lim_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} X_n = \mu\right) = 1.$$
Let’s take a minute to understand what the statement is saying. Let $S$ be the sample space and define the event $E \subset S$ by

$$E = \left\{ s \in S : \lim_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} X_n(s) = \mu \right\}.$$ 

Then the strong law of large numbers says that $P(E) = 1$. That is, aside from a negligible set of rare outcomes, the sequence of averages tends to $\mu$. The proof of this statement is beyond the scope of the course. Instead we will prove a weaker statement:

**Theorem 89** (Weak Law of Large Numbers). Let $X_n$ be an i.i.d. sequence of random variables with mean $E[X_1] = \mu$ and variance $\text{Var}(X_1) = \sigma^2$. For any $\epsilon > 0$,

$$P \left( \left| \frac{1}{N} \sum_{n=1}^{N} X_n - \mu \right| > \epsilon \right) \to 0$$

as $N \to \infty$.

*Proof.* Let $\overline{X}_N$ denote the average of $X_1, \ldots, X_N$. Applying Chebyshev’s inequality,

$$P(|\overline{X}_N - \mu| > \epsilon) \leq \frac{\text{Var}(\overline{X}_N)}{\epsilon^2} = \frac{\sigma^2}{N\epsilon^2} \to 0.$$

\[\square\]

The weak law also says that the successive averages approach the mean, but it does so in a different way. It says the probability that the averages differ from $\mu$ by more than $\epsilon$ goes to 0 as $N \to \infty$.

### 21.2.3 Central Limit Theorem

The law of large numbers tells you that successive averages of an i.i.d. sequence converge to the mean. In some ways, the next theorem tells us something even more profound. Before we state it, let’s recall some of our earlier results. We first showed that the Hypergeometric distribution converges to the Binomial distribution as the size of the urn grew large assuming the proportion of black balls converges to some limit. Later we showed that the Binomial distribution converges to the Poisson distribution if $n$ grows large, and $np$ converges to some limit $\lambda$. In each case we showed that the PMFs, or equivalently the CDFs, converged to the required distribution. Such a statement is called convergence in distribution. Here we prove an amazing result of this type.

**Theorem 90** (Central Limit Theorem). Let $X_n$ for $n = 1, 2, \ldots$ be a sequence of i.i.d. random variables with mean $E[X_1] = \mu$ and variance $\text{Var}(X_1) = \sigma^2$. Then

$$P \left( \frac{X_1 + X_2 + \cdots + X_n - n\mu}{\sigma \sqrt{n}} \leq a \right) \to \Phi(a),$$

where $\Phi$ is the CDF of an $\mathcal{N}(0,1)$ random variable.
The proof will need the following lemma, whose proof is beyond the scope of the course.

**Lemma 91.** If \( Y_n \) for \( n = 1, 2, \ldots \) is a sequence of random variables with MGFs \( M_{Y_n}(t) \), and suppose the MGFs converge pointwise to the MGF \( M_Z(t) \) of a random variable \( Z \). Then the CDF of \( Y_n \) converges to the CDF of \( Z \) pointwise (at points of continuity of the CDF of \( Z \)).

The last technical point of the lemma about continuity wont apply to us, as we are going to use a continuous random variable \( Z \).

**Proof.** We first assume that \( \mu = 0 \) and \( \sigma^2 = 1 \). If we can prove the result in this case, we then obtain the general result by un-standardization. We also assume that each \( X_i \) has common MGF \( M_X(t) \). Let \( Y \) be defined by

\[ Y_n = \frac{X_1 + X_2 + \cdots + X_n}{\sqrt{n}} \]

so that

\[ M_{Y_n}(t) = M_X \left( \frac{t}{\sqrt{n}} \right)^n. \]

Using our taylor expansion theorem from last class, we have

\[ M_X \left( \frac{t}{\sqrt{n}} \right) = \sum_{k=1}^{\infty} \frac{t^k}{k! n^{k/2}} E[X^k] = 1 + \frac{t^2}{2n} + O \left( \frac{1}{n^{3/2}} \right). \]

Thus we have (can be proven by taking logs)

\[ M_X \left( \frac{t}{\sqrt{n}} \right)^n \rightarrow e^{t^2/2} \]

which is the MGF of a standard normal random variable. The lemma proves the theorem for standardized random variables.

Now assume \( \mu, \sigma^2 \) are arbitrary. Then the above proof applies to \( \frac{X_i - \mu}{\sigma} \) giving the full result.

An alternative way to write the limit is

\[ P \left( \frac{\sqrt{n}(X_n - \mu)}{\sigma} \leq a \right) \rightarrow \Phi(a). \]

The theorem says that \( \sum_{i=1}^{n} X_i \) is approximately a \( N(n\mu, n\sigma^2) \) random variable for \( n \) large, and it doesn’t matter what distribution the \( X_i \) have. This is a pretty incredible result. We could also say that \( \overline{X}_n \) is approximately \( N(\mu, \sigma^2/n) \) distributed.
21.2.4 Correction for Continuity

When approximating integer valued distributions, the central limit theorem can sometimes give inaccurate results. For instance, suppose we want to approximate $X \sim \text{Bin}(n, p)$ (a sum of $n$ Ber($p$) random variables) distribution with $n$ large. If we approximate with $Y \sim \mathcal{N}(np, np(1-p))$ we get some weird answer such as

$$P(X = k) \approx P(Y = k) = 0.$$ 

To obtain better accuracy, we can use the following correction:

$$P(a \leq X \leq b) = P(a-1 < X < b+1) \approx P\left(a - \frac{1}{2} \leq Y \leq b + \frac{1}{2}\right) = F\left(b + \frac{1}{2}\right) - F\left(a - \frac{1}{2}\right),$$

where $F$ is the CDF of a $\mathcal{N}(np, np(1-p))$ random variable. Using this correction we have

$$P(X = k) = P(k \leq X \leq k) \approx F(k + 1/2) - F(k - 1/2),$$

which is more accurate.

21.2.5 Limit Theorem Exercises

1. Suppose a random sample of size $n = 12$ is taken from a Unif(0, 1). Approximate

$$P(|\bar{X}_n - 1/2| \leq 0.1).$$

2. (From DeGroot) Suppose that the distribution of the number of defects on any given bolt of cloth is the Poisson distribution with mean 5, and the number of defects on each bolt is counted for a random sample of 125 bolts. Determine the probability that the average number of defects per bolt in the sample will be less than 5.5.

21.2.6 Solutions

1. We approximate $\bar{X}_n$ using the central limit theorem with $\mathcal{N}(1/2, 1/144)$. If $F$ is the associated CDF, the probability is

$$F(.6) - F(.4) \approx 0.7698607,$$

which is computed via statistical software on a computer (I used the statistical package R).

2. We approximate $\bar{X}_n$ with $\mathcal{N}(5, 5/125)$ to get

$$P(\bar{X}_n \leq 5.5) \approx F(5.5) = 0.9937903.$$
22 Lecture 22

22.1 Limit Theorems

22.1.1 CLT Review Exercises

1. On a computer we generate a large random sample \( X_1, \ldots, X_n \sim \text{Ber}(p) \) where \( p \in (0, 1) \) is unknown.

   (a) How can we use the law of large numbers to estimate \( p \)?

   (b) Let \( Y \) be our estimate of \( p \) from the previous part. Using Chebyshev, how large must \( n \) be to guarantee that

   \[
   P(|Y - p| < .01) \geq .99
   \]

   Remember that \( p \) is unknown.

   (c) Using the central limit theorem, how large must \( n \) be so that

   \[
   P(|Y - p| < .01) \geq .99
   \]

   Remember that \( p \) is unknown.

2. Suppose we have a computer procedure to generate independent random pairs \((X_i, Y_i)\) for \( i = 1, 2, \ldots \). Each such pair is uniformly drawn from the square \([-1, 1] \times [-1, 1]\) (i.e., constant density).

   (a) Give a method for approximating \( \pi \) using the above scheme.

   (b) Let \( Z \) denote your approximation for \( \pi \). How many pairs must you generate to guarantee that

   \[
   P(|Z - \pi| < .001) \geq .95
   \]

22.1.2 Solutions

1. (a) We expect \( \overline{X}_n \) to approximate \( p \) well when \( n \) is sufficiently large.

   (b) Since we don’t know what \( p \) is, we use the largest possible variance of \( \overline{X}_n \). Since

   \[
   \text{Var}(\overline{X}_n) = \frac{p(1-p)}{n},
   \]

   the max occurs at \( \frac{1}{4n} \). By Chebyshev we have

   \[
   .01^2P(|Y - p| \geq .01) \leq \frac{1}{4n} \quad \Rightarrow \quad .01^2(1 - P(|Y - p| < .01)) \leq \frac{1}{4n}
   \]

   \[
   \Rightarrow \quad 1 - \frac{2500}{n} \leq P(|Y - p| < .01).
   \]

   155
We want this quantity to be at least .99 so we need
\[ .99 \leq 1 - \frac{2500}{n} \implies \frac{2500}{n} \leq .01 \implies 25000 \leq n. \]

(c) Here we approximate with a \( \mathcal{N}(p, 1/(4n)) \) random variable to see
\[ P(|Y - p| < .01) \approx F(p + .01) - F(p - .01). \]
At this point, we are sort of stuck since we don’t know \( n \). We can instead convert to standard normals so that
\[ \frac{\sqrt{n}(Y - p)}{1/2} \approx \mathcal{N}(0, 1) \]
and
\[ P(|Y - p| < .01) = P(2\sqrt{n} |Y - p| < .02\sqrt{n}) \approx 2\Phi(.02\sqrt{n}) - 1. \]
Here we used the symmetry of the normal PDF giving, for positive \( b \),
\[ \Phi(b) - \Phi(-b) = \Phi(b) - (1 - \Phi(b)) = 2\Phi(b) - 1. \]
This gives
\[ 2\Phi(.02\sqrt{n}) - 1 \geq .99 \implies \Phi(.02\sqrt{n}) \geq .995 \implies \sqrt{n} \geq 50\Phi^{-1}(.995) = 128.7915 \]
Thus \( n \geq 16588. \)

2. (a) Let \( Z_i \) be the indicator of whether \( X_i^2 + Y_i^2 < 1 \), i.e., whether it lies in the disk of radius 1. Then \( Z_i \sim \text{Ber}(p) \) with \( p = \pi/4 \). Applying the law of large numbers, we have \( 4Z_n \approx \pi \).

(b) Our approximation is \( 4Z_n \) and
\[ \text{Var}(4Z_n) = 16\text{Var}(Z_n) = \frac{16\pi}{4} \left( \frac{1 - \pi}{4} \right) = \frac{\pi(4 - \pi)}{n}. \]
This means
\[ \frac{\sqrt{n}(4Z_n - \pi)}{\sqrt{\pi(4 - \pi)}} \approx \mathcal{N}(0, 1). \]
Thus
\[ P(|4Z_n - \pi| < .001) = P \left( \frac{\sqrt{n}|4Z_n - \pi|}{\sqrt{\pi(4 - \pi)}} < \frac{.001\sqrt{n}}{\sqrt{\pi(4 - \pi)}} \right) \approx 2\Phi \left( \frac{.001\sqrt{n}}{\sqrt{\pi(4 - \pi)}} \right) - 1. \]
Proceeding we have
\[ 2\Phi \left( \frac{.001\sqrt{n}}{\sqrt{\pi(4 - \pi)}} \right) - 1 \geq .95 \implies \Phi \left( \frac{.001\sqrt{n}}{\sqrt{\pi(4 - \pi)}} \right) \geq .975 \implies \sqrt{n} \geq 1000\Phi^{-1}(.975)\sqrt{\pi(4 - \pi)} = 3218 \implies n \geq 10359515. \]
23 Homework 1 Solutions

Due Tuesday, July 7th at the beginning of class

1. Fix $a_0, d \in \mathbb{R}$ and let $a_k = a_{k-1} + d$ for all $k \geq 1$. Show that $\frac{1}{n} \sum_{k=0}^{n-1} a_k$ is given by $(a_{n-1} + a_0)/2$.

Solution. Since $a_k = a_0 + kd$ and $a_{n-1} = a_0 + (n-1)d$ we have

$$\sum_{k=0}^{n-1} a_k = \sum_{k=0}^{n-1} a_0 + kd = n a_0 + d \sum_{k=0}^{n-1} k = n a_0 + \frac{d n(n-1)}{2}.$$ 

Note that

$$\frac{a_0 + a_{n-1}}{2} = \frac{a_0 + a_0 + (n-1)d}{2} = a_0 + \frac{(n-1)d}{2},$$

so dividing our above result by $n$ gives the answer.

2. Let $S = \{1, 2, \ldots, n\}$.

(a) By counting the number of subsets of $S$ in two ways, prove that

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

(b) How many ways are there to choose 2 disjoint subsets $A, B$ from $S$? For example, if $n = 10$ then some distinct choices are:

$$A = \{1, 3\}, \quad B = \{4\}$$
$$A = \{4\}, \quad B = \{1, 3\}$$
$$A = \{1, 2, 3\}, \quad B = \emptyset$$
$$A = \emptyset, \quad B = \emptyset$$
$$A = \emptyset, \quad B = \{1, 2, \ldots, 10\}.$$ 

Solution.
(a) Each element of $S$ is either in our subset or not. Thus we have 2 options, and we must make this choice $n$ times giving $2^n$ possible subsets. Alternatively, we can sum over the possible sizes of the subset. For a given size $k$, there are $\binom{n}{k}$ subsets of that size.

Another way to solve this problem is to apply the binomial theorem to $(1 + 1)^n$.

(b) Here each element is in $A$, $B$, or neither. Since we have 3 options and must make this choice $n$ times, the solution is $3^n$.

3. You have a team of 15 (distinguishable) players and must choose which 11 will play. There are two kinds of roles ($A, B$) for each player chosen. Seven will play role $A$, and four will play role $B$.

(a) If 8 are able to play role $A$ and 7 are able to play role $B$, how many ways are there to choose who will play each role?

(b) If 8 are able to play role $A$, 5 are able to play role $B$ and 2 can play either, how many ways are there to choose who will play each role?

**Solution.**

(a) $\binom{8}{7} \binom{7}{4}$

(b) We can sum over the ways of assigning the 2 special people to roles. Each can play either role $A$, role $B$, or not play at all. Here we count the number of ways to choose roles in each case.

i. Both play role $A$: $\binom{8}{5} \binom{5}{4}$.

ii. Both play role $B$: $\binom{8}{7} \binom{7}{5}$.

iii. First plays role $A$, second plays role $B$ or vice-versa: $2 \binom{8}{5} \binom{5}{3}$.

iv. First plays role $A$ and second doesn’t play, or vice-versa: $2 \binom{8}{5} \binom{5}{4}$.

v. First plays role $B$ and second doesn’t play, or vice-versa: $2 \binom{8}{7} \binom{7}{3}$.

vi. Neither play: $\binom{8}{7} \binom{7}{4}$.

Thus the final answer is

\[
\binom{8}{5} \binom{5}{4} + \binom{8}{7} \binom{7}{5} + 2 \binom{8}{6} \binom{6}{3} + 2 \binom{8}{5} \binom{5}{4} + 2 \binom{8}{7} \binom{7}{3} + \binom{8}{7} \binom{7}{4} .
\]
24 Homework 2

Due Wednesday, July 8th at the beginning of class

1. Prove that if $A, B$ are events then

\[ P(A \cap B) \geq P(A) + P(B) - 1. \]

*Solution.* Note that

\[ 1 \geq P(A \cup B) = P(A) + P(B) - P(A \cap B). \]

By moving 1 to the right and $P(A \cap B)$ to the left we get the result.

2. You roll a fair 6-sided die and a fair 4-sided die.

   (a) Give a sample space for the outcomes, and define $P$.
   (b) Use your sample space to compute the probability of rolling a 3 on the first die.
   (c) Use your sample space to compute the probability that the sum is 7.
   (d) Suppose instead you rolled a fair 6-sided die, and then a fair $k$-sided die where $k$ is the value you got on your first roll. What is the sample space, what is $P$, and what is the chance of the sum being 7? (If you roll a 1-sided die, you always get 1.)

*Solution.*

(a) Let $S = \{(d_1, d_2) : 1 \leq d_1 \leq 6, 1 \leq d_2 \leq 4\}$ and

\[ P(A) = \frac{|A|}{|S|} = \frac{|A|}{24}. \]

Alternatively, you could have said that

\[ P(\{(d_1, d_2)\}) = \frac{1}{24} \]

for any $(d_1, d_2)$ and defined $P(A)$ as in our general finite space (end of lecture 2 notes).

(b) If $A$ denotes the event of rolling a 3 on the first die, we have

\[ P(A) = \frac{4}{24} = \frac{1}{6}. \]

(c) There are 4 ways to get a sum of 7, so the probability is $\frac{1}{6}$. 159
(d) Here the sample space is

\[ S = \{(d_1, d_2) : 1 \leq d_1 \leq 6, 1 \leq d_2 \leq d_1\}. \]

We can treat this as a general finite space (see end of notes from lecture 2) with

\[ P((d_1, d_2)) = \frac{1}{6} \cdot \frac{1}{d_1}. \]

If \( A \) denotes the event the sum is 7 then we have

\[ A = \{(4, 3), (5, 2), (6, 1)\} \]

so

\[ P(A) = \frac{1}{6} \cdot \frac{1}{4} + \frac{1}{6} \cdot \frac{1}{5} + \frac{1}{6} \cdot \frac{1}{6} = \frac{1}{6} \cdot \frac{74}{120} = \frac{37}{360}. \]
25  Homework 3 Solutions

Due Thursday, July 9th at the beginning of class

1. Suppose you are randomly dealt 7 cards from a standard 52 card deck without replacement.

(a) Give a sample space and probability measure.
(b) What is the probability of getting 4 of one value and 3 of another?
(c) What is the probability at least 5 of them have the same suit?
(d) (⋆) What is the probability at least 3 cards have the same value?
(e) What is the probability of getting 3 distinct pairs, while not getting 3-of-a-kinds (i.e., can’t get 3 of the same value)?

Solution. Let $D$ denote the set of all cards in the deck.

(a) $S = \{ B : B \subset D, |B| = 7 \}$ and let $P$ be determined by treating every outcome as equally likely. Alternatively, we could use

$$S = \{ (d_1, d_2, \ldots, d_7) : d_i \in D, d_i \text{ are distinct} \}.$$

(b) 

$$\frac{13 \cdot 12 \cdot \binom{4}{3} \binom{48}{4}}{\binom{52}{7}} = \frac{13 \cdot 12 \cdot 4}{\binom{52}{7}}.$$

If you used ordered hands, you would get the same probability but it would be written slightly differently (this goes for the rest of the solutions as well).

(c) The number of hands with at least 5 of the same suit is

$$4 \binom{13}{5} \binom{39}{2} + 4 \binom{13}{6} \binom{39}{1} + 4 \binom{13}{7}$$

thus the answer is

$$4\binom{13}{5} \binom{39}{2} + 4 \binom{13}{6} \binom{39}{1} + 4 \binom{13}{7}.$$

(d) Let $A_i$ denote the event that at least 3 cards have value $i$, and assume the values are $1, \ldots, 13$. Then we have, by inclusion-exclusion,

$$P \left( \bigcup_{i=1}^{13} A_i \right) = \sum_{i=1}^{13} P(A_i) - \sum_{i<j} P(A_iA_j) + \sum_{i<j<k} P(A_iA_jA_k) + \cdots.$$

Note that it is impossible to have more than 2 values with at least 3 cards, so only the first two sums are non-zero. Computing the parts we have

$$P(A_i) = \frac{\binom{4}{3} \binom{48}{4}}{\binom{52}{7}}.$$

161
and
\[ P(A_iA_j) = \frac{\binom{3}{4}^2 44 + 2\binom{4}{4} \binom{3}{3}}{\binom{52}{7}}. \]

Thus the answer is
\[ 13\frac{\binom{4}{4} \binom{48}{4} + \binom{4}{3} \binom{48}{3}}{\binom{52}{7}} - \left( \frac{13}{2} \right) \frac{\binom{4}{3}^2 44 + 2\binom{4}{4} \binom{3}{3}}{\binom{52}{7}}. \]

(e)
\[ \frac{10\binom{13}{3} \binom{4}{3} \binom{4}{1}}{\binom{52}{7}}. \]

2. Suppose you roll a 4-sided die once, and then flip a coin 4 times that is heads with probability \( \frac{k}{5} \) where \( k \) is the roll of the die. Given that you get 3 heads, what is the probability you rolled a 3?

**Solution.** Let \( A \) denote the event of rolling a 3, and let \( B \) denote the event of getting 3 heads. Then we have

\[
P(B) = \sum_{k=1}^{4} \frac{1}{4} \left( \binom{4}{3} \right) \left( \frac{k}{5} \right)^3 \left( 1 - \frac{k}{5} \right)
\]

\[
P(A \cap B) = P(B|A)P(A) = \left( \binom{4}{3} \right) \left( \frac{3}{5} \right)^3 \left( \frac{2}{5} \right) \frac{1}{4}
\]

\[
P(A|B) = \frac{P(A \cap B)}{P(B)}
\]

\[
= \frac{\left( \binom{4}{3} \right) \left( \frac{3}{5} \right)^3 \left( 1 - \frac{k}{5} \right) \frac{1}{4}}{\frac{1}{4} \sum_{k=1}^{4} \left( \binom{4}{3} \right) \left( \frac{k}{5} \right)^3 \left( 1 - \frac{k}{5} \right)}
\]

\[
= \frac{\left( \frac{3}{5} \right)^3 \left( 1 - \frac{k}{5} \right)}{\sum_{k=1}^{4} \left( \frac{k}{5} \right)^3 \left( 1 - \frac{k}{5} \right)}
\]

\[
= \frac{27}{73}.
\]
26 Homework 4

Due Monday, July 13th at the beginning of class

1. [Taken from DeGroot] In a certain city 30 percent of the people are Conservative, 50 percent are Liberals, and 20 percent are Independent. We also know that 65 percent of Conservatives voted, 82 percent of Liberals voted, and 50 percent of Independents voted. If a person in the city is selected at random and we learn they didn’t vote, what is the probability they are Liberal?

Solution. Let $C, L, I$ be the events denoting Conservative, Liberal, and Independent, respectively. Let $V$ be the event denoting voting. Then we have

$$P(L|V^c) = \frac{P(V^c|L)P(L)}{P(V^c|L)P(L) + P(V^c|I)P(I) + P(V^c|C)P(C)}$$

$$= \frac{.18 \cdot .5}{.18 \cdot .5 + .35 \cdot .2 + .5 \cdot .3}$$

$$\approx .305.$$  

2. You roll a fair $n$-sided die $k$ times.

(a) Give a sample space and an associated probability measure.

(b) What is the chance that your rolls are strictly increasing? That is, the second roll is larger than the first, and third is larger than the second, etc.

Solution.

(a) $S = \{(d_1, \ldots, d_k) : 1 \leq d_i \leq n\}$ where all outcomes are equally likely.

(b) $\binom{n}{k} \cdot \frac{k(n-k)}{n^k}$.

3. Give a counting argument to show that

$$\binom{n}{2} = \binom{k}{2} + k(n-k) + \binom{n-k}{2}$$

where $2 \leq k \leq n$.

Solution. The number of subsets of size 2 taken from a set of size $n$ is given by $\binom{n}{2}$. Suppose the elements are broken into two groups, the first of size $k$ and the remainder of size $n-k$. Then to choose 2 elements we could either pick two from the first group, two from the second group, or one from each group. These three options have $\binom{k}{2}$, $k(n-k)$ and $\binom{n-k}{2}$ choices each, respectively.
4. For any event $A$, let an $A$-certificate be a contract that requires the seller to give the buyer $1000$ if the event $A$ occurs, and 0 otherwise. For a particular (sports-based) experiment, your friend is willing to buy or sell any $A$-certificate for $P_{\text{friend}}(A) \cdot 1000$ dollars, where $P_{\text{friend}}$ reflects your friend’s belief system. Assume there are disjoint events $B, C$ such that

$$P_{\text{friend}}(B) + P_{\text{friend}}(C) \neq P_{\text{friend}}(B \cup C).$$

Show how making a few trades with your friend can earn you some money. This shows how a belief system that doesn’t adhere to the axioms of probability can be exploited in the context of betting.

**Solution.** Let $B, C$ be the disjoint events described in the statement. If

$$P_{\text{friend}}(B) + P_{\text{friend}}(C) > P_{\text{friend}}(B \cup C)$$

then we should sell him a $B$-certificate and a $C$-certificate and purchase a $(B \cup C)$-certificate from him. Due to the inequality above, this set of transactions nets us

$$1000(P_{\text{friend}}(B) + P_{\text{friend}}(C) - P_{\text{friend}}(B \cup C))$$

dollars. Analogously, if

$$P_{\text{friend}}(B) + P_{\text{friend}}(C) < P_{\text{friend}}(B \cup C)$$

then we sell a $(B \cup C)$-certificate and purchase a $B$-certificate and a $C$-certificate again netting us money.

If $(B \cup C)^c$ occurs, none of the contracts pay and we keep our money. If $(B \cup C)$ occurs then he must pay us 1000 and we must pay him 1000 so again we make money in total. Stated a different way, holding a $B$-certificate and a $C$-certificate is equivalent to holding a $(B \cup C)$-certificate, so they should have the same price.

5. Suppose you roll a 100-sided die $n$ times. The die isn’t fair, and gives the value $i$ with probability $p_i$ where $\sum_{i=1}^{100} p_i = 1$. What is the probability your $n$th roll is different from all previous rolls? [Hint: Condition on the value of the last roll and apply the LOTP.]

**Solution.** Let $A_i$ denote the event that the last roll is $i$, and let $B$ denote the event that the last roll is different from all previous. Then we have

$$P(B) = \sum_{i=1}^{100} P(B|A_i)P(A_i).$$

Note that

$$P(A_i) = p_i, \quad P(B|A_i) = \frac{P(B \cap A_i)}{P(A_i)} = \frac{(1 - p_i)^{n-1}p_i}{p_i} = (1 - p_i)^{n-1}$$
6. Suppose an urn has \( M \) black balls and \( N \) white balls. You draw \( k \) balls from it without replacement where \( 1 \leq k \leq M + N \).

(a) Give a sample space and a probability measure. [Treat every ball as distinct, and either use subsets or ordered sequences.]

(b) What is the probability that \( j \) of the \( k \) balls are black?

**Solution.**

(a) Suppose the balls are numbered \( 1, \ldots, M + N \) where the first \( M \) are black. If we use unordered sets as our sample space then we get

\[
S = \{ B : B \subset \{1, \ldots, M + N\}, |B| = k \}
\]

with all outcomes equally likely. If we use ordered sequences then we get

\[
S = \{(b_1, \ldots, b_k) : b_i \in \{1, \ldots, M + N\}, b_i \text{ are distinct}\}
\]

where again all outcomes are equally likely.

(b) For the unordered sample space we get

\[
\frac{\binom{M}{j} \binom{N}{k-j}}{\binom{M+N}{k}}.
\]

For the ordered sample space we get

\[
\frac{\binom{M}{j} \binom{N}{k-j} k!}{(M + N)(M + N - 1) \cdots (M + N - k + 1)}
\]

which is the same answer.
27 Homework 5

Due Tuesday, July 14th at the beginning of class

1. Suppose we flip two fair coins, with all 4 outcomes equally likely. Let $A$ be the event that the first coin is heads, let $B$ be the event the second coin is heads, and let $C$ be the event both coins match. Show that all pairs of the events are independent, but that the triple $A, B, C$ isn’t independent.

 SOLUTION. Note that

$$P(AB) = \frac{1}{4} = P(A)P(B)$$
$$P(BC) = \frac{1}{4} = P(B)P(C)$$
$$P(AC) = \frac{1}{4} = P(A)P(C)$$

$$P(ABC) = \frac{1}{4} \neq P(A)P(B)P(C).$$

2. A stock is currently priced at $77. Each day it either goes up $1 with probability .5 or down $1 with probability .5. What is the chance it will hit $100 before it hits $50? [Hint: Use result from Lecture 5.]

 SOLUTION. This is equivalent to the Gambler’s Ruin problem where bankruptcy corresponds to the stock hitting 50. Thus we have $p = .5$, $M = 50$ (since $100 - 50 = 50$) and we compute

$$s_{27} = \frac{27}{50}.$$

3. Suppose again we are playing a game with 100 doors behind which there are 2 cars and 98 goats. You randomly pick a door (for simplicity, assume you pick door #1), and then the host opens 96 goat doors (randomly chosen from the goat doors among doors 2-100). You are given the opportunity to keep your current door, or pick one of the 3 other doors.

(a) Assuming you randomly choose a remaining door that isn’t door 1, what is your chance of winning a car?

(b) What is your chance of winning a car if you stick with your original pick?

 SOLUTION.

(a) If $W$ is the event of winning and $C_1$ is the event the car is behind door 1 we have

$$P(W) = P(W|C_1)P(C_1) + P(W|C^c_1)P(C^c_1) = \frac{1}{3} \cdot \frac{2}{100} + \frac{2}{3} \cdot \frac{98}{100} = \frac{198}{300}. $$
(b) This is the same as your original chance of picking the correct door:

\[ \frac{2}{100}. \]
Homework 6

Due Wednesday, July 15th at the beginning of class

1. Suppose players $A$ and $B$ alternate rolling pairs of fair 6-sided dice. The game is won when one player rolls dice that sum to 7. If $A$ rolls first, what is the probability that $A$ eventually wins?

Solution. First note that there is a $\frac{1}{6}$ chance of rolling a sum of 7 with two dice. Following the technique described in Lecture 5 Review Exercise 2, we compute

$$P(A \text{ gets 7} | A \text{ or } B \text{ gets 7}) = \frac{\frac{1}{6}}{\frac{1}{6} + \frac{1}{6} - \frac{1}{36}} = \frac{6}{11}.$$ 

Alternatively, we could compute this using an infinite sum:

$$\sum_{k=0}^{\infty} \left( \frac{5}{6} \cdot \frac{5}{6} \right)^k \frac{1}{6} = \frac{\frac{1}{6}}{1 - \frac{25}{36}} = \frac{6}{11}.$$ 

2. A fair 6-sided die is rolled 10 times. Let the $X$ denote the sum of the rolls, and $Y$ denote the product.

(a) Give a sample space and probability measure modeling the rolls.

(b) Show how $X, Y$ are defined as random variables (directly use the definition of a random variable).

(c) $(\star)$ What is $P(X = 10)$?

(d) $(\star)$ What is $P(X \leq 12)$?

(e) What is $P(Y \text{ is even})$?

Solution.

(a) Define $S$ by

$$S = \{(f_1, f_2, \ldots, f_{10}) : 1 \leq f_i \leq 6\}$$

with all outcomes equally likely. Here our measure is determined by treating all outcomes as equally likely.

(b) Define $X, Y : S \to \mathbb{R}$ by

$$X((f_1, \ldots, f_{10})) = \sum_{i=1}^{10} f_i \quad \text{and} \quad Y((f_1, \ldots, f_{10})) = \prod_{i=1}^{10} f_i.$$

(c) $\frac{1}{6^{10}}$
(d) \[
\frac{1}{6^{10}} + \frac{10}{6^{10}} + \frac{10 + \binom{10}{2}}{6^{10}} = \frac{66}{6^{10}}.
\]

If done with partitions, this same formula would be written
\[
\frac{\binom{9}{9}}{6^{10}} + \frac{\binom{10}{1}}{6^{10}} + \frac{\binom{11}{2}}{6^{10}}.
\]

(e) We compute the probability the product isn’t odd:
\[
1 - \frac{1}{2^{10}} = \frac{2^{10} - 1}{2^{10}}.
\]
29 Homework 7

Due Thursday, July 16th at the beginning of class

1. Depicted below is the PMF $p_X$ for a random variable $X$:

(a) Draw the CDF $F_X$.
(b) What is $E[X]$?
(c) If $g(x) = e^{(x-2)^2}$, what is $E[g(X)]$?

Solution.

(a) CDF:

(b) We have

$E[X] = 1 \cdot p_X(1) + 3 \cdot p_X(3) + 6 \cdot p_X(6) = .5 + 3(.25) + 6(.25) = \frac{11}{4} = 2.75$.

(c) Letting $Y = g(X)$ we have

$E[X] = e \cdot p_Y(e) + e^{16} \cdot p_Y(e^{16}) = e(.75) + e^{16}(.25)$.

2. We have an urn containing 20 black balls and 10 white balls.

(a) Suppose we draw 7 balls from the urn with replacement. Compute $p_X(x)$ for all $x \in \mathbb{R}$ where $X$ is the number of black balls drawn.
(b) Suppose instead we draw 7 balls from the urn without replacement. Compute \( p_Y(y) \) for all \( y \in \mathbb{R} \) where \( Y \) is the number of black balls drawn.

Solution.

(a) If \( k \notin \{0, 1, \ldots, 7\} \) we have \( p_X(k) = 0 \). Otherwise,
\[
p_X(k) = \binom{7}{k} \left( \frac{2}{3} \right)^k \left( \frac{1}{3} \right)^{n-k}.
\]

(b) Again, for \( k \notin \{0, 1, \ldots, 7\} \) we have \( p_Y(k) = 0 \). Otherwise,
\[
p_Y(k) = \frac{\binom{20}{k} \binom{10}{7-k}}{\binom{30}{7}}.
\]

3. (⋆) Prove the following theorem.

**Theorem 92.** Let \( X \) be a discrete random variable that only takes positive integer values (i.e., \( \text{im}(X) \subset \mathbb{Z}_{>0} \)). Then we have
\[
E[X] = \sum_{k=1}^{\infty} P(X \geq k).
\]

[Hint: Use the technique of introducing an inner summation, and then swapping summations like our first computation of \( \sum k \frac{1}{2^k} \).

Solution.

Proof.
\[
\sum_{k=1}^{\infty} P(X \geq k) = \sum_{k=1}^{\infty} \sum_{j=k}^{\infty} p_X(j)
= \sum_{j=1}^{\infty} \sum_{k=1}^{\infty} p_X(j)
= \sum_{j=1}^{\infty} j p_X(j)
= E[X].
\]
30 Homework 8 Solutions

Due Monday, July 20th at the beginning of class

1. (DeGroot) A civil engineer is studying a left-turn lane that is long enough to hold 4 cars. Let $X$ be the number of cars in the lane at the end of a randomly chosen red light. The engineer believes that the probability that $X = x$ is $C(x + 1)(8 - x)$ for $x = 0, \ldots, 4$ (the possible values of $X$) where $C > 0$ is the same for all $x$.

   (a) Find the PMF of $X$.
   (b) Find the probability that $X$ will be at least 3.

Solution.

(a) Solving we have

$$ C \left[ (1)(8) + (2)(7) + (3)(6) + (4)(5) + (5)(4) \right] = 80C $$

so $C = 1/80$. Thus $p_X(x) = (x + 1)(8 - x)/80$ for $x = 0, \ldots, 4$, and 0 otherwise.

(b) The answer is

$$ P(X \geq 3) = p_X(3) + p_X(4) = \frac{4 \cdot 5 + 5 \cdot 4}{80} = \frac{1}{2}. $$

2. A group of $n$ people all have distinct heights. They are waiting in a straight line at the bank (one person in front of the other), with all orderings of the people equally likely. A person can see ahead if they are taller than everyone in front of them.

   (a) What is the probability that the $i$th person in line can see ahead (where the first person is at the front of the line, the second is behind the first, etc)?

   (b) What is the expected number of people in line that can see ahead? [Hint: Linearity of expectation.]

Solution.

(a) The $i$th person in line can see if he is the largest of the first $i$ people. There is a $1/i$ chance of this occurring. For a more explicit calculation, let $S$ denote the space of all $n!$ possible orderings with each equally likely. The number of orderings with the $i$th person able to see ahead is

$$ \binom{n}{i} (i - 1)! (n - i)! = \frac{n!}{i}, $$

as there are $\binom{n}{i}$ choices for the first $i$ people, there are $(i - 1)!$ orderings of them that have the $i$th person tallest, and there are $(n - i)!$ orderings of the remaining $n - i$ people.

172
(b) By linearity, the expected number is

$$\sum_{i=1}^{n} \frac{1}{i}.$$

3. (DeGroot) Suppose that when a machine is adjusted properly 50 percent of the items made are high quality, and 50 percent are medium quality. If the machine is adjusted improperly, then 25 percent are of high quality, and 75 percent are of medium quality. Assume the machine is adjusted improperly 10 percent of the time. [As usual, conditional on the adjustment of the machine, the items are independent.]

(a) The machine is adjusted, and then you look at 5 produced items. If 4 out of 5 are high quality, what is the chance it was adjusted properly?

(b) After the 5 above were produced, you look at another item, and find it is of medium quality (making a total of 6 observed items). What is your new posterior probability that the machine was adjusted properly?

Solution.

(a) Let $A$ be the event the machine is adjusted properly, and $E$ the event that 4 out of 5 are of high quality. Then

$$P(A|E) = \frac{P(E|A)P(A)}{P(E|A)P(A) + P(E|A^c)P(A^c)}$$

$$= \frac{\binom{5}{4}(0.5)^5(0.9)}{\binom{5}{4}(0.5)^5(0.9) + \binom{5}{1}(0.25)^4(0.75)(0.1)}$$

$$= \frac{96}{97}$$

$$\approx 0.99.$$  

As all of the $\binom{5}{4}$ coefficients cancel, we see that saying 4 out of 5 are high quality, or specifying a particular sequence like $HHHHM$ or $MHHHH$ give the same conditional probability. This reflects the idea that when collecting independent samples and updating beliefs, the order of the samples doesn’t have information.

(b) We solve this in two different ways.

i. For our first solution, let $A$ denote the event the machine is adjusted properly, and $F$ the event we have 4 out of 5 high quality, followed by a medium quality.
This gives

\[
P(A|F) = \frac{P(F|A)P(A)}{P(F|A)P(A) + P(F|A^c)P(A^c)}
\]

\[
= \frac{\binom{5}{1}(.5)^6(.9)}{\binom{5}{1}(.5)^6(.9) + \binom{5}{2}(.25)^4(.75)^2(.1)}
\]

\[
= \frac{64}{65}
\]

\[
\approx .984.
\]

As stated earlier, we could have also just used any sequence with 4 H’s (for high quality) and 2 M’s (for medium quality) as the binomial coefficients just cancel.

ii. For our second solution, let \( P_E \) be the probability measure defined by \( P_E(B) = P(B|E) \) where \( E \) is defined as in part (a). Let \( G \) denote the event of getting a medium quality item. Then we have

\[
P_E(A|G) = \frac{P_E(G|A)P_E(A)}{P_E(G|A)P_E(A) + P_E(G|A^c)P_E(A^c)}
\]

\[
= \frac{.596}{.596 + .751}
\]

\[
= \frac{64}{65}
\]

\[
\approx .984.
\]

Note that \( P_E(G|A) = P(G|A) \) because conditional on the machine’s adjustment the items are independent.

4. Suppose we have two coins, one with probability \( p_1 \) of getting heads, and the other probability \( p_2 \). We flip each coin once. Letting \( X \) denote the total number of heads, compute the PMF \( p_X(k) \) for all \( k \in \mathbb{R} \).

**Solution.** For \( k \notin \{0, 1, 2\} \) we have \( p_X(k) = 0 \). Otherwise,

\[
p_X(0) = (1 - p_1)(1 - p_2), \quad p_X(1) = p_1(1 - p_2) + (1 - p_1)p_2, \quad p_X(2) = p_1p_2.
\]

5. Suppose we flip \( n \) coins each obtaining heads with probability \( 0 < p < 1 \).

(a) Give the sample space and probability measure.

(b) Let \( X_k \) denote the indicator random variable for the \( k \)th flip being heads (i.e., 1 if heads, 0 if tails). Show the \( X_k, k = 1, \ldots, n \) are independent random variables.
Solution. 

(a) Let $S$ be defined by 

$$S = \{(f_1, \ldots, f_n) : f_i \in \{H, T\}\}.$$ 

We use the general finite sample space where 

$$P(\{(f_1, \ldots, f_n)\}) = p^H (1 - p)^T.$$ 

(b) Note that 

$$P(X_1 = x_1, \ldots, X_n = x_n) = p^H (1 - p)^T = \prod_{i=1}^{n} P(X_i = x_i)$$ 

for any $x_1, \ldots, x_n$. Thus they are independent. We could have also written the formula above as 

$$p^{\sum_{i=1}^{n} x_i}(1 - p)^{n - \sum_{i=1}^{n} x_i}.$$ 

6. Let $n$ be a fixed positive integer, and let $X$ be a random variable that takes on the values $1, \ldots, n$. Suppose $p_X(k) = Ck$ for $k = 1, \ldots, n$, where $C > 0$ is the same for all $k$. 

(a) Determine $C$. 

(b) What is $E[X]$? 

Solution. 

(a) Solving for $C$ we have 

$$1 = \sum_{k=1}^{n} Ck = Cn(n+1) \quad \Rightarrow \quad C = \frac{2}{n(n+1)}.$$ 

(b) We have 

$$E[X] = \sum_{k=1}^{n} \frac{2k^2}{n(n+1)} = \frac{2n(n+1)(2n+1)}{6n(n+1)} = \frac{2n+1}{3}.$$
31 Homework 9

Due Tuesday, July 21th at the beginning of class

1. Suppose that a large lot with 10000 manufactured items has 30 percent defective items and 70 percent nondefective. You choose a subset of 10 items to test.

(a) What is the probability that at most 1 of the 10 test items is defective?
(b) Approximate the previous answer using the binomial distribution.

Solution.
(a) If $X \sim \text{HGeom}(10, 3000, 10000)$ then we have

$$p_X(0) + p_X(1) = \frac{\binom{7000}{10}}{\binom{10000}{10}} + \frac{\binom{3000}{1}}{\binom{10000}{10}} \approx 0.149.$$ 

(b) If $Y \sim \text{Bin}(10, .3)$ we have

$$p_Y(0) + p_Y(1) = (.7)^{10} + \binom{10}{1}(.3)(.7)^9 \approx 0.149.$$ 

2. Fix $0 < p < 1$ and suppose there is a coin that obtains a head with probability $p$. We flip the coin 17 times and get a total of 5 heads. Given this information, what is the chance that 3 of those heads occurred in the first 10 flips?

Solution. Let $F$ denote the event 3 occurred in the first 10 flips, and let $E$ denote the event of getting 5 heads out of 17. Then we have

$$P(F|E) = \frac{P(FE)}{P(E)} = \frac{\binom{10}{3}\binom{7}{2}p^5(1-p)^{12}}{\binom{17}{5}p^5(1-p)^{12}} = \frac{\binom{10}{3}\binom{7}{2}}{\binom{17}{5}}.$$ 

In general, if $X \sim \text{Bin}(n_1, p)$ and $Y \sim \text{Bin}(n_2, p)$ are independent then

$$P(X = k|X + Y = j) = \frac{\binom{n_1}{k}\binom{n_2}{j-k}}{\binom{n_1+n_2}{j}}.$$ 

That is, conditional on the sum being $j$ we get a $\text{HGeom}(j, n_1, n_1 + n_2)$ distribution.
32 Homework 10
Due Wednesday, July 22nd at the beginning of class

1. You are rolling a fair 30-sided die.
   
   (a) What is the expected number of rolls required to get a 27?

   (b) What is the expected number of rolls required to get a 27, followed later by a 28?
       For example,
       
       \[9, 28, 1, 27, 27, 3, 28\]
       
       would be an example of requiring 7 rolls and
       
       \[9, 3, 27, 28\]
       
       would be 4 rolls.

   (c) What is the expected number of rolls required to get a 27, and a 28 in either order? For example,
       
       \[9, 27, 3, 28\]
       
       would be an example of 4 rolls and
       
       \[1, 28, 14, 28, 5, 27\]
       
       would be an example of 6 rolls. [Hint: Think of the number of rolls to get 27 or 28, and then the number of rolls to get the remaining value.]

   (d) What is the expected number of rolls to get all 30 possible values (in any order)?

   Solution.

   (a) Letting \(X \sim \text{Geom}(1/30)\), this is \(E[X] = 30\).

   (b) Letting \(X, Y \sim \text{Geom}(1/30)\) this is \(E[X + Y] = 60\).

   (c) Letting \(X \sim \text{Geom}(2/30)\) and \(Y \sim \text{Geom}(1/30)\) this is \(E[X + Y] = 15 + 30 = 45\).

   (d) This is \(E[\sum_{i=1}^{30} X_i]\) where \(X_i \sim \text{Geom}((31 - i)/30)\). Thus the answer is

   \[
   \sum_{i=1}^{30} \frac{30}{31 - i} \approx 119.85.
   \]

2. Suppose that a book contains an average of \(\lambda\) misprints per page.

   (a) What is the probability that 10 pages will contain at most 1 misprint?

   (b) What is the probability that \(n\) pages will contain at most 3 misprints?

   (c) \((\star)\) If the book has \(n\) pages, what is the probability that there will be at least \(m\) pages that each contain more than \(k\) misprints?
Solution.

(a) \( e^{-10\lambda}(1 + 10\lambda) \).

(b) \[ e^{-n\lambda} \left( 1 + n\lambda + \frac{(n\lambda)^2}{2} + \frac{(n\lambda)^3}{3!} \right) . \]

(c) Let \( p \) be defined by

\[ p = 1 - e^{-\lambda} \sum_{i=0}^{k} \frac{\lambda^i}{i!} . \]

Then the answer is

\[ \sum_{i=m}^{n} \binom{n}{i} p^i (1 - p)^{n-i} . \]
33 Homework 11
Due Thursday, July 23nd at the beginning of class

1. You keep flipping a coin until you get a head. You are paid $2^k$ dollars. Suppose that if you make more than $2^{10}$ dollars, you will only receive $2^{10}$ dollars. What is the expected payoff of this game?

Solution.

$$\sum_{k=1}^{10} \frac{2^k}{2^k} + \sum_{k=11}^{\infty} \frac{2^{10}}{2^k} = 11.$$

2. Suppose $X_1, \ldots, X_n$ form a random sample from a Pois($\lambda$) distribution. Use Chebyshev to determine how large $n$ must be to guarantee that

$$P(|\bar{X}_n - \mu| < .05) \geq .9,$$

where $\mu = E[X_i]$.

Solution. By Chebyshev

$$P(|\bar{X}_n - \lambda| \geq .05) \leq \frac{\lambda}{.05^2 n}$$

giving

$$1 - P(|\bar{X}_n - \lambda| < .05) \leq \frac{\lambda}{.05^2 n}.$$

Rearranging we have

$$P(|\bar{X}_n - \lambda| < .05) \geq 1 - \frac{\lambda}{.05^2 n}.$$

Thus we need

$$1 - \frac{\lambda}{.05^2 n} \geq .9 \iff .1 \geq \frac{\lambda}{.05^2 n} \iff n \geq 4000\lambda.$$

3. Suppose $F(x) = \int_{e^x}^{x^2} e^t \, dt$. Compute $F'(x)$. [Hint: Write as a difference of two integrals of the form $\int_0^{x^2}$ and $\int_0^{e^x}$ and apply the chain rule.]

Solution. As hinted we have

$$F(x) = \int_0^{x^2} e^t \, dt - \int_0^{e^x} e^t \, dt.$$

Differentiating using the Fundamental Theorem of Calculus and the chain rule gives

$$F'(x) = 2xe^{x^2} - e^x e^x.$$
1. Let $X$ have PDF $f_X$ and let $a, b \in \mathbb{R}$. Show that

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

**Solution.** By LOTUS we have

$$E[aX + b] = \int_{-\infty}^{\infty} (ax + b) f_X(x) \, dx = a \int_{-\infty}^{\infty} x f_X(x) \, dx + b \int_{-\infty}^{\infty} f_X(x) \, dx = aE[X] + b.$$ 

2. Let $U \sim \text{Unif}(0, 1)$, $X = U^2$ and $Y = e^X$. Compute $E[Y]$ (leave answer as an integral).

**Solution.** Note that $Y = e^{U^2}$ so by LOTUS we have

$$E[Y] = \int_{0}^{1} e^{u^2} \, du.$$ 

If instead we try to compute the PDFs first (a harder alternative) we have, for $x \in (0, 1)$,

$$f_X(x) = f_U(\sqrt{x}) \frac{1}{2\sqrt{x}} = \frac{1}{2\sqrt{x}}$$

and, for $y \in (1, e)$,

$$f_Y(y) = f_X(\log(y)) \cdot \frac{1}{y} = \frac{1}{2y \sqrt{\log(y)}}.$$ 

Then we obtain

$$E[Y] = \int_{1}^{e} \frac{1}{2\sqrt{\log(y)}} \, dy$$

which gives the same answer.

3. (From DeGroot) Suppose that a person has a lottery ticket from which she will win $X$ dollars, where $X \sim \text{Unif}(0, 4)$. Suppose her utility function is $U(x) = x^\alpha$ for $x \geq 0$ and 0 otherwise, where $\alpha > 0$ is some fixed constant. For how many dollars is she willing to sell the ticket (i.e., what is $E[U(X)]$)?

**Solution.** By LOTUS we have

$$E[U(X)] = \int_{-\infty}^{\infty} U(x) f_X(x) \, dx = \frac{1}{4} \int_{0}^{4} x^\alpha \, dx = \frac{1}{4} \left[ \frac{x^{\alpha+1}}{\alpha + 1} \right]_0^4 = \frac{4^{\alpha}}{\alpha + 1}.$$ 

Alternatively, we could proceed the hard way and compute the PDF of $Y = U(X)$: For $y \in (0, 4^\alpha)$ we have

$$f_Y(y) = f_X(y^{1/\alpha}) \frac{1}{\alpha y^{\frac{\alpha-1}{\alpha}}} = \frac{1}{4\alpha y^{\frac{\alpha-1}{\alpha}}}.$$
Then we have

\[ \int_0^{4^\alpha} \frac{y}{4\alpha y^{\frac{\alpha-1}{\alpha}}} \, dy = \int_0^{4^\alpha} \frac{y^{1/\alpha}}{4\alpha} \, dy = \frac{1}{4\alpha} \left[ \frac{y^{\frac{\alpha+1}{\alpha}}}{\frac{\alpha+1}{\alpha}} \right]_0^{4^\alpha} = \frac{4^{\alpha+1}}{4(\alpha + 1)} = \frac{4^\alpha}{\alpha + 1}. \]
35  Homework 13

Due Thursday, July 30th at the beginning of class

1. Compute $E[e^{tX}]$ where $X \sim \mathcal{N}(0, 1)$. [Hint: Complete the square in the exponent.]

Solution.

\[
E[e^{tX}] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{tx} e^{-x^2/2} \, dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-1/2(x^2-2tx)} \, dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-1/2(x-t)^2+t^2} \, dx = e^{t^2/2} \int_{-\infty}^{\infty} e^{-1/2(x-t)^2} \, dx = e^{t^2/2},
\]

by noting that we are integrating the PDF of a $\mathcal{N}(t, 1)$ RV.

2. Consider a 1 meter stick and suppose you break it into two pieces $X$ meters from the end, where $X \sim \text{Unif}(0, 1)$. What is the expected length of the longer piece (after it is broken)?

Solution. If we break it $x \geq 1/2$ meters from the end, then the longer piece has length $x$. Otherwise, it has length $1-x$. This gives:

\[
\int_{1/2}^{1} x \, dx + \int_{0}^{1/2} 1-x \, dx = \left[ \frac{x^2}{2} \right]_{1/2}^{1} + \left[ x - \frac{x^2}{2} \right]_{0}^{1/2} = \frac{1}{2} - \frac{1}{8} + \frac{1}{2} - \frac{1}{8} = \frac{3}{4}.
\]
Suppose $X$ has PDF $f_X$ given by

$$f_X(x) = \begin{cases} \frac{\alpha x_0^\alpha}{x^{\alpha+1}} & \text{if } x \geq x_0, \\ 0 & \text{if } x < x_0, \end{cases}$$

where $x_0 > 0$ and $\alpha > 0$ are given fixed parameters. What is the distribution of $\log(X/x_0)$? Give PDF and the name of the distribution.

**Solution.** Let $Y = g(X)$ where $g(x) = \log(x/x_0)$. Since $x \geq x_0$ we see that $g(x)$ takes values in $[0, \infty)$ on which $g^{-1}(y) = x_0 e^y$. Then we have, for $y \geq 0$,

$$f_Y(y) = \left| \frac{1}{g'(g^{-1}(y))} \right| f_X(g^{-1}(y))$$

$$= \frac{\alpha x_0^\alpha}{(x_0 e^y)^{\alpha+1}} \cdot e^y$$

$$= \alpha e^{-(\alpha+1)y} e^y$$

$$= \alpha e^{-\alpha y}.$$

Thus $Y \sim \text{Exp}(\alpha)$.

2. Suppose the temperature in a location (measured in Fahrenheit) is normally distributed with mean 68 degrees and a standard deviation of 4 degrees. What is the distribution of the temperature measured in Celsius?

**Solution.** Let $X$ denote the temperature in Fahrenheit so that $X \sim \mathcal{N}(68, 16)$. Then $Y$, the temperature in Celsius, is given by

$$Y = \frac{5(X - 32)}{9}.$$

Thus $Y \sim \mathcal{N}\left(20, \frac{400}{81}\right)$ (i.e., mean is 20, variance is $400/81$ and standard deviation is $20/9$).

3. Let $Y$ have a lognormal distribution with parameters 3 and 1.44 (i.e., $Y = e^X$ where $X \sim \mathcal{N}(3, 1.44)$). Compute $P(Y \leq 6.05)$ approximately. [Hint: Don’t use the lognormal PDF/CDF.]

**Solution.** Since $\sigma_X = \sqrt{1.44} = 1.2$ we have

$$P(Y \leq 6.05) \approx P(X \leq 1.8) = P(X \leq \mu_X - \sigma_x) \approx .16.$$
4. Given that $X \sim \text{Exp}(\lambda)$, compute $E[e^{-(X-\lambda/2)^2}]$. Your answer should not be left as an integral.

Solution. Using LOTUS we have

$$E[e^{-(X-\lambda/2)^2}] = \int_0^\infty \lambda e^{-(x-\lambda/2)^2} e^{-\lambda x} \, dx$$

$$= \int_0^\infty \lambda e^{-x^2 - \lambda^2/4} \, dx$$

$$= \lambda e^{-\lambda^2/4} \int_0^\infty e^{-x^2} \, dx$$

$$= \lambda e^{-\lambda^2/4} \sqrt{2\pi} \cdot \frac{1}{\sqrt{2\pi} \cdot 1/2} \int_0^\infty e^{-x^2/2} \, dx$$

$$= \lambda e^{-\lambda^2/4} \frac{1}{2} \sqrt{\pi},$$

since the rest is the integral of a $\mathcal{N}(0,1/2)$ PDF.

5. Let $f(x,y) = x^2 y$ for $0 \leq x^2 \leq y \leq 1$, and 0 otherwise.

(a) Compute $\int_{-\infty}^\infty \int_{-\infty}^\infty f(x,y) \, dx \, dy$. [Hint: Integrate over region below.]

(b) Compute $\int_R f(x,y) \, dx \, dy$ where $R = \{(x,y) : x \geq y\}$. [Hint: Integrate over region below.]
Solution.

(a) We have
\[
\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) \, dx \, dy = \int_{-1}^{1} \int_{x^2}^{1} x^2 y \, dy \, dx
\]
\[
= \int_{-1}^{1} x^2 \left( \frac{1}{2} - \frac{x^4}{2} \right) \, dx
\]
\[
= \left[ \frac{x^3}{6} - \frac{x^7}{14} \right]_{-1}^{1}
\]
\[
= \frac{8}{42}.
\]

(b) We have
\[
\int_{0}^{1} \int_{x^2}^{1} x^2 y \, dx \, dy = \int_{0}^{1} x^2 \left( \frac{x^2}{2} - \frac{x^4}{2} \right) \, dx
\]
\[
= \left[ \frac{x^5}{10} - \frac{x^7}{14} \right]_{0}^{1}
\]
\[
= \frac{1}{35}.
\]

6. Let \( f(x, y) = y^2 \) for \( (x, y) \in [0, 2] \times [0, 1] \) and 0 otherwise. Compute \( \iint_{R} f(x, y) \, dx \, dy \) where \( R \) is given by:

(a) \( R = \{(x, y) : x + y > 2\} \).

(b) \( R = \{(x, y) : y < 1/2\} \).

(c) \( R = \{(x, y) : x \leq 1\} \).

(d) \( R = \{(x, y) : x = 3y\} \). [Hint: What is the volume under the graph of \( f \) when you are integrating over a line?]

Solution.
(a) 
\[
\int_1^2 \int_{2-x}^1 y^2 \, dy \, dx = \int_1^2 \frac{1}{3} - \frac{(2-x)^3}{3} \, dx \\
= \frac{1}{3} - \int_0^1 \frac{u^3}{3} \, du \\
= \frac{1}{3} - \frac{1}{12} \\
= \frac{1}{4}.
\]

You could have also used the other order of integration to get the same answer:
\[
\int_0^1 \int_{2-y}^2 y^2 \, dx \, dy.
\]

(b) 
\[
\int_0^2 \int_0^{1/2} y^2 \, dy \, dx = \int_0^2 \frac{1}{24} \, dx = \frac{1}{12}.
\]

(c) 
\[
\int_0^1 \int_0^1 y^2 \, dx \, dy = \int_0^1 \frac{1}{3} \, dx = \frac{1}{3}.
\]

(d) The integral is zero.
37 Homework 15

Due Tuesday, August 4th at the beginning of class

1. (From DeGroot) The joint PMF of $X, Y$ are given by the following table:

<table>
<thead>
<tr>
<th>X</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.08</td>
<td>0.07</td>
<td>0.06</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>1</td>
<td>0.06</td>
<td>0.10</td>
<td>0.12</td>
<td>0.05</td>
<td>0.02</td>
</tr>
<tr>
<td>2</td>
<td>0.05</td>
<td>0.06</td>
<td>0.09</td>
<td>0.04</td>
<td>0.03</td>
</tr>
<tr>
<td>3</td>
<td>0.02</td>
<td>0.03</td>
<td>0.03</td>
<td>0.03</td>
<td>0.04</td>
</tr>
</tbody>
</table>

Determine the following:
(a) $P(Y > 2)$
(b) $P(X \leq 2$ and $Y \leq 2$)
(c) $P(X > Y)$.
(d) $P(X = 3|Y = 3)$.

Solution.

(a) $.01 + .05 + .04 + .03 + .01 + .02 + .03 + .04 = .23.$
(b) $.08 + .07 + .06 + .06 + .10 + .12 + .05 + .06 + .09 = .69.$
(c) $.06 + .06 + .03 + .05 + .03 + .02 = .25.$
(d) $.03/(.01 + .05 + .04 + .03) = .03/.13 \approx .2308.$

2. (From DeGroot) Suppose that the joint PDF of two random variables $X$ and $Y$ is as follows:

$$f(x, y) = \begin{cases} 
  c(x^2 + y) & \text{for } 0 \leq y \leq 1 - x^2, \\
  0 & \text{otherwise.}
\end{cases}$$

(a) What is $c$?
(b) Determine $P(0 \leq X \leq 1/2)$.
(c) Determine $P(Y \leq X + 1)$.
(d) Determine $P(Y = X^2)$.

Solution.
(a) We have

\[
1 = c \int_{-1}^{1} \int_{0}^{1-x^2} x^2 + y \, dy \, dx
= c \int_{-1}^{1} \left[ x^2 y + \frac{y^2}{2} \right]_{0}^{1-x^2} \, dx
= c \int_{-1}^{1} x^2(1-x^2) + \frac{(1-x^2)^2}{2} \, dx
= \frac{c}{2} \int_{-1}^{1} -1(1-x^4) \, dx
= \frac{c}{2} \left[ x - \frac{x^5}{5} \right]_{-1}^{1}
= c \left( 1 - \frac{1}{5} \right),
\]
so \( c = \frac{5}{4} \).

(b) We have

\[
P(0 \leq X \leq 1/2) = \frac{5}{4} \int_{0}^{1/2} \int_{0}^{1-x^2} x^2 + y \, dy \, dx
= \frac{5}{8} \int_{0}^{1/2} 1-x^4 \, dx
= \frac{5}{8} \left[ x - \frac{x^5}{5} \right]_{0}^{1/2}
= \frac{5}{8} \left( \frac{1}{2} - \frac{1}{160} \right)
= \frac{5 \cdot 79}{8 \cdot 160}
= \frac{79}{256}.
\]

(c) We compute

\[
P(X < 0, Y \leq X + 1) + P(X \geq 0, Y \leq X + 1) = P(X < 0, Y \leq X + 1) + 1/2.
\]
\[
\frac{5}{4} \int_{-1}^{0} \int_{0}^{x+1} x^2 + y \, dy \, dx + \frac{1}{2} = \frac{5}{4} \int_{-1}^{0} \left[ x^2 y + \frac{y^2}{2} \right]_{0}^{x+1} \, dx + \frac{1}{2}
\]
\[
= \frac{5}{4} \int_{-1}^{0} x^2(x + 1) + \frac{(x + 1)^2}{2} \, dx + \frac{1}{2}
\]
\[
= \frac{5}{8} \int_{-1}^{0} 2x^3 + 3x^2 + 2x + 1 \, dx + \frac{1}{2}
\]
\[
= \frac{5}{8} \left[ \frac{x^4}{2} + x^3 + x^2 + x \right]_{-1}^{0} + \frac{1}{2}
\]
\[
= \frac{5}{8} \left( -\frac{1}{2} + 1 - 1 + 1 \right) + \frac{1}{2}
\]
\[
= \frac{5}{16} + \frac{1}{2}
\]
\[
= \frac{13}{16}.
\]

(d) 0.
38  Homework 16

Due Wednesday, August 5th at the beginning of class

1. (From DeGroot) The following table shows joint probabilities for each year of student in a given high school, and the number of times they visited the museum:

<table>
<thead>
<tr>
<th></th>
<th>Never</th>
<th>Once</th>
<th>More than once</th>
</tr>
</thead>
<tbody>
<tr>
<td>Freshmen</td>
<td>0.08</td>
<td>0.10</td>
<td>0.04</td>
</tr>
<tr>
<td>Sophmores</td>
<td>0.04</td>
<td>0.10</td>
<td>0.04</td>
</tr>
<tr>
<td>Juniors</td>
<td>0.04</td>
<td>0.20</td>
<td>0.09</td>
</tr>
<tr>
<td>Seniors</td>
<td>0.02</td>
<td>0.15</td>
<td>0.10</td>
</tr>
</tbody>
</table>

(a) What is the probability a random student has never visited the museum?
(b) If a student has visited more than once, what is the probability she is a senior?

Solution.

(a) $0.08 + 0.04 + 0.04 + 0.02 = 0.18$.
(b) $\frac{0.10}{0.04 + 0.04 + 0.09 + 0.10} = \frac{10}{27} \approx 0.37$.

2. (From DeGroot) Suppose students take Math and Music aptitude tests, each giving test results in $[0, 1]$. Suppose a random student’s Math and Music scores are given by $X, Y$, respectively, which have the joint PDF

$$f_{X,Y}(x, y) = \frac{2}{5}(2x + 3y)$$

for $0 \leq x, y \leq 1$, and 0 otherwise.

(a) What is the probability a student will score higher than .8 on the Math test?
(b) What is the probability a student will score higher than .8 on the Math test given they score .3 on the Music test?
(c) What is the probability a student will score higher than .8 on the Music test given they score .3 on the Math test?

Solution. We first compute the marginal distributions:

$$f_X(x) = \frac{2}{5} \int_0^1 2x + 3y \, dy = \frac{2}{5} \left(2x + \frac{3}{2}\right)$$

and

$$f_Y(y) = \frac{2}{5} \int_0^1 2x + 3y \, dx = \frac{2}{5} (1 + 3y).$$
(a) We have

\[ P(X \geq 8) = \frac{2}{5} \int_{8}^{1} 2x + \frac{3}{2} \, dx \]

\[ = \frac{2}{5} \left[ x^2 + \frac{3}{2} x \right]_{8}^{1} \]

\[ = \frac{2}{5} \left( 1 - \frac{16}{25} + \frac{3}{10} \right) \]

\[ = \frac{2}{5} \cdot \frac{33}{50} \]

\[ = \frac{33}{125} \]

\[ = 0.264. \]

(b) We have

\[ P(X \geq 0.8|Y = .3) = \int_{.8}^{1} f_{X|Y}(x, .3) \, dx \]

\[ = \int_{.8}^{1} f_{X,Y}(x, .3) \, dx \frac{f_{Y}(.3)}{f_{Y}(.3)} \]

\[ = \int_{.8}^{1} \frac{2}{5} (2x + 3(.3)) \, dx \]

\[ = \int_{.8}^{1} \frac{2}{5} (1 + 3(.3)) \, dx \]

\[ = \int_{.8}^{1} \frac{2x + .9}{1.9} \, dx \]

\[ = \frac{1}{1.9} \left[ x^2 + .9x \right]_{.8}^{1} \]

\[ = \frac{1}{1.9} \left( 1 - \frac{16}{25} + \frac{9}{50} \right) \]

\[ = \frac{27}{95} \]

\[ \approx 0.28421. \]
(c) We have

\[
P(Y \geq 0.8|X = .3) &= \int_{.8}^{1} f_{Y|X}(y|3) \, dy \\
&= \int_{.8}^{1} \frac{f_{X,Y}(3, y)}{f_X(3)} \, dy \\
&= \int_{.8}^{1} \frac{\frac{2}{5} (2(3) + 3y)}{\frac{2}{5} (2(3) + \frac{3}{2})} \, dy \\
&= \frac{10}{21} \int_{.8}^{1} \frac{6}{10} + 3y \, dy \\
&= \frac{10}{21} \left[ \frac{6y}{10} + \frac{3y^2}{2} \right]_{.8}^{1} \\
&= \frac{10}{21} \left( \frac{6}{10} \frac{3}{2} - \frac{48}{50} \right) \\
&= \frac{33}{30} \\
&= \frac{1050}{330} \\
&\approx 0.31429.
\]
39 Homework 17

Due Thursday, August 6th at the beginning of class

1. Let \( X \sim \text{Beta}(a, b) \). Compute \( E[X] \) and \( \text{Var}(X) \).

   **Solution.** Firstly,
   \[
   E[X] = \frac{1}{B(a, b)} \int_0^1 x^a (1-x)^{b-1} \, dx \\
   = \frac{B(a+1, b)}{B(a, b)} \cdot \frac{1}{B(a+1, b)} \int_0^1 x^a (1-x)^{b-1} \, dx \\
   = \frac{B(a+1, b)}{B(a, b)}.
   \]

   Simplifying we have
   \[
   E[X] = \frac{\Gamma(a+1)\Gamma(b)\Gamma(a+b)}{\Gamma(a)\Gamma(b)\Gamma(a+b+1)} = \frac{a}{a+b},
   \]

   using the identity \( \Gamma(x) = (x-1) \Gamma(x-1) \) which holds for all real \( x > 1 \). Similarly, we have
   \[
   E[X^2] = \frac{1}{B(a, b)} \int_0^1 x^{a+1} (1-x)^{b-1} \, dx \\
   = \frac{B(a+2, b)}{B(a, b)} \\
   = \frac{(a+1)a}{(a+b)(a+b+1)}.
   \]

   Thus
   \[
   \text{Var}[X] = \frac{(a+1)a}{(a+b)(a+b+1)} - \frac{a^2}{(a+b)^2} = \frac{a[(a+1)(a+b) - a(a+b+1)]}{(a+b)^2(a+b+1)} = \frac{ab}{(a+b)^2(a+b+1)}.
   \]

2. Let \( X \sim \text{Exp}(\lambda_1) \) and \( Y \sim \text{Exp}(\lambda_2) \) be independent with \( \lambda > 0 \).

   (a) Compute the PDF of \( X + Y \).

   (b) Compute \( E[X^2Y^2] \). [Hint: Can use 2d LOTUS, or that functions of independent RVs are independent.]

   **Solution.**
(a) Letting $Z = X + Y$ we have, if $\lambda_1 \neq \lambda_2$, 

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(t) f_Y(z-t) \, dt$$

$$= \int_0^z \lambda_1 \lambda_2 \int_0^z e^{-\lambda_2 z} e^{(\lambda_2 - \lambda_1) t} \, dt$$

$$= \lambda_1 \lambda_2 \int_0^z e^{-\lambda_2 z} \left[ \frac{e^{(\lambda_2 - \lambda_1) t}}{\lambda_2 - \lambda_1} \right]_0^z$$

$$= \frac{\lambda_1 \lambda_2}{\lambda_2 - \lambda_1} e^{-\lambda_2 z}(e^{(\lambda_2 - \lambda_1) z} - 1)$$

$$= \frac{\lambda_1 \lambda_2}{\lambda_2 - \lambda_1} (e^{-\lambda_1 z} - e^{-\lambda_2 z}).$$

If $\lambda_1 = \lambda_2$ it is much cleaner:

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(t) f_Y(z-t) \, dt$$

$$= \lambda_1 \lambda_2 \int_0^z e^{-\lambda_2 z} \, dt$$

$$= \lambda_1 \lambda_2^{z} e^{-\lambda_2 z},$$

which is the PDF of a Gamma($2, \lambda_2$) random variable.

3. We say that $X \sim \text{Gamma}(a, b)$ (with $a, b > 0$) if the PDF $f_X$ is given by

$$f_X(x) = \frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx}$$

for $x > 0$ and 0 otherwise, with $\Gamma(a)$ defined in the lecture notes. Suppose $X \sim \text{Gamma}(a, b)$ and $Y \mid X \sim \text{Pois}(X)$.

(a) What is the conditional PDF $f_{X\mid Y}(x\mid y)$? [Hint: Very fast using Bayes’s Theorem.]

(b) What is the joint PDF/PMF $f_{X,Y}$?

Solution.

(a) Note that

$$f_{X\mid Y}(x\mid y) \propto f_{Y\mid X}(y\mid x) f_X(x) \propto e^{-x} \frac{x^y}{y!} x^{a-1} e^{-b x} \propto x^{y+a-1} e^{-(b+1)x}$$

for $y = 0, 1, \ldots, x > 0$. Thus $X\mid Y = y \sim \text{Gamma}(y + a, b + 1)$ and has PDF

$$f_{X\mid Y}(x\mid y) = \frac{(b + 1)^{y+a}}{\Gamma(y + a)} x^{y+a-1} e^{-(b+1)x}.$$
(b) The joint PDF/PMF is given by

\[ f_{X,Y}(x, y) = f_{Y|X}(y|x) f_X(x) = e^{-x^y} \frac{x^y}{y!} \frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx}, \]

for \( x > 0 \) and \( y = 0, 1, \ldots \).
40 Homework 18

Due Monday, August 10th at the beginning of class

1. Suppose a chicken has $N$ eggs where $N \sim \text{Pois}(\lambda)$ for some $\lambda > 0$. Conditional on $N = n$ let $X$, the number of eggs that hatch, be distributed as $X|N = n \sim \text{Bin}(n,p)$. That is, conditional on $N = n$ each egg hatches independently with probability $p$. Let $Y$ denote the number of eggs that do not hatch.

(a) What is the distribution of $Y|N = n$?

(b) What are $E[X]$ and $E[Y]$?

(c) What is $P(X = i, Y = j|N = n)$ for fixed $i, j \geq 0$? [Hint: What is the value if $i + j \neq n$? If $i + j = n$ this is the same as $P(X = i|N = n)$.]

(d) What is $P(X = i, Y = j)$ for fixed $i, j \geq 0$? [Hint: Use LOTP and note almost every term in series is zero.]

(e) In the previous part we computed the joint PMF of $X, Y$. Compute the marginal distributions of $X$ and $Y$. [Hint: Both have same type of distribution. Use solution to b) to help with the algebra.]

(f) Are $X, Y$ independent?

Solution.

(a) $Y|N = n \sim \text{Bin}(n, 1 - p)$

(b)

$$E[X] = E[E[X|N]] = E[Np] = \lambda p \quad \text{and} \quad E[Y] = E[E[Y|N]] = E[N(1-p)] = \lambda (1-p).$$

(c) For $i + j \neq n$, the probability is zero since $P(X = i, Y = j, N = n) = 0$. If $i + j = n$ then we have

$$P(X = i, Y = j|N = n) = P(X = i|N = n) = \binom{n}{i} p^i (1-p)^{n-i}.$$

(d)

$$P(X = i, Y = j) = \sum_{n=0}^{\infty} P(X = i, Y = j|N = n) P(N = n) = P(X = i, Y = j|N = i + j) P(N = i + j) = \binom{i + j}{i} p^i (1-p)^{n-i} e^{-\lambda} \frac{\lambda^{i+j}}{(i+j)!}. $$
(e) Factoring the above, note that

\[
\binom{i+j}{i} p^i (1-p)^{n-i} e^{-\lambda \frac{\lambda^i + j}{(i+j)!}} = \frac{e^{-\lambda p} (\lambda p)^i}{i!} \frac{e^{-\lambda(1-p)} (\lambda(1-p))^j}{j!}.
\]

Thus \(X \sim \text{Pois}(\lambda p)\) and \(Y \sim \text{Pois}(\lambda(1-p))\).

(f) Yes, since we factored the PMF.

2. Let \(X, Y\) have a constant density over the triangular region bounded by the lines \(y = x\), \(y = -x\), and \(y = 1\).

(a) What is the joint PDF \(f_{X,Y}\)?

(b) Are \(X, Y\) independent?

(c) What is the marginal PDF \(f_X\)?

(d) What is the marginal PDF \(f_Y\)?

(e) What is the distribution of \(Y|X = x\)?

(f) What is the distribution of \(X|Y = y\)?

(g) What is \(E[Y|X = x]\)?

(h) What is \(\text{Var}[Y|X = x]\)?

(i) What is \(E[Y]\)?

Solution.

(a) \(f_{X,Y}(x, y) = 1\) for \(x\) in the triangular region. That is \(-1 \leq x \leq 1\) and \(|x| \leq y \leq 1\).

(b) No, the region isn’t rectangular.

(c) We have

\[
f_X(x) = \int_{|x|}^{1} 1 \, dx = 1 - |x|,
\]

for \(|x| \leq 1\).

(d) We have

\[
f_Y(y) = \int_{-y}^{y} 1 \, dx = 2y,
\]

for \(0 \leq y \leq 1\).

(e) We have

\[
f_{Y|x}(y|x) = \frac{1}{1 - |x|}
\]

for \(-1 \leq x \leq 1\) and \(|x| \leq y \leq 1\). That is \(\text{Unif}(|x|, 1)\).
(f) We have
\[ f_{X|Y}(x|y) = \frac{1}{2|y|} \]
for 0 ≤ y ≤ 1 and −|y| ≤ x ≤ |y|. That is Unif(−|y|, |y|).

(g) \( E[Y|X = x] = \frac{1+|x|}{2} \).

(h) \( \text{Var}[Y|X = x] = \frac{(1-|x|)^2}{12} \).

(i) By the law of iterated expectation and then LOTUS
\[
E[Y] = E[E[Y|X]] = \int_{-1}^{1} E[Y|X = x]f_x(x) \, dx = \int_{-1}^{1} \frac{1+|x|}{2} (1-|x|) \, dx = \int_{0}^{1} 1-x^2 \, dx = \frac{2}{3}.
\]
You could have also just used the marginal of y, and either integrated it, or noticed \( Y \sim \text{Beta}(2,1) \).

3. In class we defined the conditional variance \( \text{Var}(X|Y = y) \). We define \( \text{Var}(X|Y) \) analogously to \( E(X|Y) \). Prove that
\[ \text{Var}(X) = E[\text{Var}(X|Y)] + \text{Var}(E[X|Y]). \]

Solution. Note that
and
\[ \text{Var}(E[X|Y]) = E[E[X|Y]^2] - E[E[X|Y]]^2 = E[E[X|Y]^2] - E[X]^2. \]
Thus \( E[\text{Var}(X|Y)] + \text{Var}(E[X|Y]) = E[X^2] - E[X]^2 = \text{Var}(X) \).

4. The Gamma distribution was defined in the previous homework. Let \( X \sim \text{Gamma}(a,b) \) for some \( a > 0 \) and \( b > 0 \).

(a) Let \( Y = 1/X \). Compute the PDF \( f_Y \) of \( Y \). We say \( Y \sim \text{InvGamma}(a,b) \).

(b) Compute \( E[Y] \) where \( a > 1 \). [Hint: Let \( u = 1/y \), compare to Gamma PDF, and remember \( \Gamma(a) = (a-1)\Gamma(a-1) \). Alternatively, use LOTUS.]

(c) Suppose \( W \sim \text{InvGamma}(a,b) \) with \( a > 1 \) and \( b > 0 \) and that \( Z|W = w \sim \mathcal{N}(0, w) \). What is the distribution of \( W|Z = z \)? [Hint: InvGamma is a conjugate prior here.]

Solution.

(a) Let \( g(x) = 1/x \) so that \( g'(x) = -x^{-2} \) and \( g^{-1}(y) = 1/y \). Then
\[
f_Y(y) = f_X(g^{-1}(y)) \left| \frac{1}{g'(g^{-1}(y))} \right| = \frac{b^{a}}{\Gamma(a)} y^{1-a} e^{-b/y} y^{-2} = \frac{b^{a}}{\Gamma(a)} y^{-1-a} e^{-b/y}.
\]
(b) By LOTUS we have (spotting the Gamma($a - 1, b$) PDF)

$$
\int_{0}^{\infty} \frac{1}{x} f_X(x) \, dx = \frac{b^a}{\Gamma(a)} \int_{0}^{\infty} x^{a-2} e^{-bx} \, dx = \frac{b^a}{b^{a-1}} \cdot \frac{\Gamma(a-1)}{\Gamma(a)} = \frac{b}{a-1}.
$$

(c) By Bayes’s theorem,

$$
\begin{align*}
    f_{W|Z}(w|z) & \propto f_{Z|W}(z|w) f_W(w) \\
    & \propto \frac{1}{w^{3/2}} e^{-\frac{z^2}{2w}} w^{-1-a} e^{-b/w} \\
    & = w^{-3/2-a} e^{-(z^2/2+b)/w},
\end{align*}
$$

so $Z|W = w \sim \text{InvGamma}(a+1/2, z^2/2 + b)$.

5. Let $X \sim \text{Gamma}(a, b)$ for some $a > 1$ and $b > 0$ and let $Y|X = x \sim \text{Exp}(x)$.

(a) What is $E[Y]$? [Hint: InvGamma.]

(b) What is the distribution of $X|Y = y$?

**Solution.**

(a) By the law of iterated expectation (and problem 4b),

$$
E[Y] = E[E[Y|X]] = E \left[ \frac{1}{X} \right] = \frac{b}{a-1}.
$$

(b) By Bayes’s theorem,

$$
\begin{align*}
    f_{X|Y}(x|y) & \propto f_{Y|X}(y|x) f_X(x) \\
    & \propto xe^{-y/x} x^{a-1} e^{-bx} = x^a e^{-(y+b)x},
\end{align*}
$$

so $X|Y = y \sim \text{Gamma}(a+1, y+b)$. 

199
41 Homework 19

Due Tuesday, August 11th at the beginning of class

1. Let $X, Y \sim \mathcal{N}(0, 1)$ be i.i.d. and let $U, V$ given by

$$U = aX + bY + c \quad \text{and} \quad V = dX + eY + f$$

have a bivariate normal distribution (here $a, b, c, d, e, f \in \mathbb{R}$ with $ae - bd \neq 0$).

(a) What is $\text{Cov}(X, Y)$?
(b) What is $\text{Cov}(U, V)$?
(c) What are the marginal distributions of $U, V$?

Solution.

(a) $\text{Cov}(X, Y) = 0$ since they are independent.
(b) $\text{Cov}(U, V) = \text{Cov}(aX + bY + c, dX + eY + f) = ad \text{Var}(X) + be \text{Var}(Y) + (ae + bd) \text{Cov}(X, Y) = ad + be$.
(c) From our result in class on sums of normal random variables we have

$$U \sim \mathcal{N}(c, a^2 + b^2) \quad \text{and} \quad V \sim \mathcal{N}(f, d^2 + e^2).$$

2. (a) Compute the MGF of $X$ with PDF $f_X(x) = 2x$ for $x \in (0, 1)$.
(b) Compute the MGF of $Y \sim \text{Geom}(p)$ with $p \in (0, 1)$.
(c) Compute the MGF of $aX + bY$ for $a, b \neq 0$ assuming $X, Y$ are independent and defined in the previous two parts.

Solution.
(a) Integrating by parts,

\[ M_X(t) = E[e^{tX}] = \int_0^1 2xe^{tx} \, dx \]
\[ = 2 \left[ \frac{xe^{tx}}{t} \right]_0^1 - 2 \int_0^1 \frac{e^{tx}}{t} \, dx \]
\[ = \frac{2e^t}{t} - 2 \left[ \frac{e^{tx}}{t^2} \right]_0^1 \]
\[ = \frac{2e^t}{t} - \frac{2(e^t - 1)}{t^2} \]
\[ = \frac{2((t-1)e^t + 1)}{t^2}. \]

The above works when \( t \neq 0 \). As usual, we always have \( M_X(0) = 1 \).

(b) We have

\[ M_Y(t) = E[e^{tY}] = \sum_{k=1}^{\infty} e^{tk}p(1-p)^{k-1} \]
\[ = \frac{p}{1-p} \sum_{k=1}^{\infty} (e^t(1-p))^k \]
\[ = \frac{pe^t(1-p)}{(1-p)(1-e^t(1-p))} \]
\[ = \frac{pe^t}{1-e^t(1-p)}. \]

(c) We have

\[ M_{aX+bY}(t) = M_X(at)M_Y(bt) = \frac{2((at-1)e^{at} + 1)}{(at)^2} \cdot \frac{pe^{bt}}{1-e^{bt}(1-p)}. \]

3. This problem from DeGroot gives a little practice with higher dimensional distributions. Suppose \( X \) has PDF \( f_X \) given by

\[ f_X(x) = \frac{1}{n!} x^n e^{-x} \]

for \( x > 0 \) where \( n \) is a non-negative integer. Conditional on \( X = x \) let \( Y_1, \ldots, Y_n \) (same \( n \)) be i.i.d. with PDF

\[ g_Y(y|x) = \frac{1}{x} \]

for \( 0 < y < x \).
(a) Give the joint PDF $f_{X,Y_1,...,Y_n}(x,y_1,...,y_n)$.

(b) Give the marginal joint PDF $f_{Y_1,...,Y_n}(y_1,...,y_n)$. [Hint: The lower bound on the integral will be $\max_i y_i$.]

(c) Give the conditional PDF $f_{X|Y_1,...,Y_n}(x|y_1,...,y_n)$.

**Solution.**

(a) We have

$$f_{X,Y_1,...,Y_n}(x,y_1,...,y_n) = \frac{e^{-x}}{n!},$$

where $0 < y_i < x$ for all $i = 1, \ldots, n$.

(b) As hinted we compute the following integral:

$$f_{Y_1,...,Y_n}(y_1,...,y_n) = \int_{\max_i y_i}^{\infty} \frac{e^{-x}}{n!} \, dx = \frac{e^{-\max_i y_i}}{n!},$$

for $y_1, \ldots, y_n > 0$.

(c) Dividing we have

$$f_{X|Y_1,...,Y_n}(x|y_1,...,y_n) = \frac{f_{X,Y_1,...,Y_n}(x,y_1,...,y_n)}{f_{Y_1,...,Y_n}(y_1,...,y_n)} = e^{\max_i y_i - x},$$

for $x > \max_i y_i$. 

202