Introduction to Probability
Notes for Stat 110 (Harvard, Fall 2014)

Won I. Lee

This set of notes was written and compiled for the sake of students in the Fall 2014 edition of Stat 110 at Harvard, taught by Kevin Rader. As a result, it may differ in focus and material from previous and future iterations of the course, but the notes are provided in case future students may find them helpful in the quest to understand the subtleties and nuances of probability theory.[1]

1 Introduction, Combinatorics

In this section of the notes, I hope to lay out the "big picture" of each week's topics, focusing on why certain topics are important and how they relate to each other.

Why combinatorics/counting? Intuitively, probability is about the likelihood of some event happening, whether it be rain, stock prices rising, or marriage. And the way we often intuit these 'likelihoods' is by thinking about how many ways that event can occur, out of all the different possibilities. Of course, this is where the *combinatorics* part comes in - figuring out how to count things in a systematic and efficient way is key to figuring out the probability.

What is probability anyway? At its heart, probability is the "size of a set". Recall that $A \in S$ is a subset of the sample space, S, and we call such a subset an **event**. Thus, the **probability** $P(A)$ tells us the "size" of the set A, which equals the "likelihood" of it happening. But how do you typically calculate the "size" of a set? Well, for any *finite* things, we just tend to **count** the elements in the set (i.e. atoms in an object, molecules in a glass of water, pages in a book) - and this is why we need combinatorics!

So how about this conditional probability? Whenever I think conditional probability, I think of the **pebble world** example. Conditional probability asks, "*Given* that I am in set B, what is the probability of A occurring?" When we think about probability by itself, we count the elements of A over the entire sample space S; but when we know that we are *already in set* B, we only need to count A over B.

Useful Identifies

$$e = \frac{1}{0!} + \frac{1}{1!} + \frac{1}{2!} + \frac{1}{3!} + \frac{1}{4!} + \dots \qquad\qquad \binom{n}{k} = \binom{n}{n-k}$$

Set Theory

Sets and Subsets - A set is a collection of distinct objects. A is a subset of B if every element of A is also included in B.

Empty Set - The empty set, denoted $\emptyset$, is the set that contains nothing.

Set Notation - Note that $\mathbf{A} \cup \mathbf{B}$, $\mathbf{A} \cap \mathbf{B}$, and $\mathbf{A^c}$ are all sets too.

[1] While most of the notes and problems were written by me, some of the material and exercises were adapted from notes by W. Chen and S. Chiu, lecture slides/material by K. Rader, and notes/lectures given by J. Blitzstein. All errors are mine.

Union - $\mathbf{A} \cup \mathbf{B}$ (read $\mathbf{A}$ *union* $\mathbf{B}$) means $\mathbf{A}$ *or* $\mathbf{B}$

Intersection - $\mathbf{A} \cap \mathbf{B}$ (read $\mathbf{A}$ *intersect* $\mathbf{B}$) means $\mathbf{A}$ *and* $\mathbf{B}$

Complement - $\mathbf{A^c}$ (read $\mathbf{A}$ *complement*) occurs whenever $\mathbf{A}$ does not occur

Disjoint Sets - Two sets are disjoint if their intersection is the empty set (e.g. they don't overlap).

Partition - A set of subsets $\mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3, ...\mathbf{A}_n$ partition a space if they are disjoint and cover all possible outcomes (e.g. their union is the entire set). A simple case of a partitioning set of subsets is $\mathbf{A}, \mathbf{A^c}$

Principle of Inclusion-Exclusion - Helps you find the probabilities of unions of events.

$$P(\mathbf{A} \cup \mathbf{B}) = P(\mathbf{A}) + P(\mathbf{B}) - P(\mathbf{A} \cap \mathbf{B})$$

Counting (aka Combinatorics)

Naïve Definition of Probability - *If the likelihood of each outcome is equal*, the probability of any event happening is:

$$P(\text{Event}) = \frac{\text{number of favorable outcomes}}{\text{number of outcomes}}$$

Multiplication Rule - Let's say we have a compound experiment (an experiment with multiple components). If the 1st component has n_1 possible outcomes, the 2nd component has n_2 possible outcomes, and the rth component has n_r possible outcomes, then overall there are $n_1 n_2 \dots n_r$ possibilities for the whole experiment. **Example** There are 17 freshman dorms and 12 upperclassmen houses at Harvard, which means that there are a total of $17 * 12 = 204$ combinations of dorms that a Harvard student can be assigned to.

Factorial - $n! = 1 \cdot 2 \cdot 3 \cdots n$
i.e., $n!$ is the number of ways to order n people in line, by the multiplication rule: there are n choices for the first person, $n-1$ for the second, ..., and 1 for the last person. (**Note:** $0! = 1$)

The Binomial Coefficient is used often in combinatorics. $\binom{n}{k}$ (read n choose k) is the number of subsets of size k of a set of size n. (intuitively: choosing k people out of group of n *without replacement* where *order doesn't matter.*)

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

Sampling Table - The sampling tables describes the different ways to take a sample of size k out of a population of size n.

	Order Does Matter	**Order Doesn't Matter**
With Replacement	n^k	$\binom{n+k-1}{k}$
Without Replacement	$\frac{n!}{(n-k)!}$	$\binom{n}{k}$

Story Proofs

Definition - A proof by interpretation or application rather than by algebra or calculus. Examples of story proofs would be combinatorial proofs, which there are three examples of below, where two quantities are equated as it it shown that they are two ways of counting the same thing.

Example 1 - Symmetry Rule for Binomial Coefficients

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

If you want to choose k people out of n, you can either directly pick the k people to be included (LHS), or pick the $n-k$ to be not included (RHS).

Example 2 - Vandermonde Identity

$$\sum_{i=0}^{k} \binom{m}{i}\binom{n}{k-i} = \binom{m+n}{k}$$

A class contains m females and n males. We want to find all of the ways to form teams of k students. We can find the number of teams that include i females - there are $\binom{m}{i}$ ways to choose the females and $\binom{n}{k-i}$ ways to choose the males, for a total of $\binom{m}{i}\binom{n}{k-i}$ ways to choose a team with i females. We sum over i to get the total number of teams (LHS), which we know is also $\binom{m+n}{k}$ (RHS).

Conditional Probability and Bayes' Rule

Joint Probability - $P(\mathbf{A} \cap \mathbf{B})$ or $P(\mathbf{A}, \mathbf{B})$ - Probability of **A** *and/intersect* **B**.

Marginal (Unconditional) Probability - $P(\mathbf{A})$ - Probability of **A**

Conditional Probability - $P(\mathbf{A}|\mathbf{B})$ - Probability of **A** given **B** occurred.

Bayes' Rule - Bayes' Rule unites marginal, joint, and conditional probabilities. This is one of the central concepts in all of probability and statistics.

$$\boxed{P(\mathbf{A}|\mathbf{B}) = \frac{P(\mathbf{A} \cap \mathbf{B})}{P(\mathbf{B})} = \frac{P(\mathbf{B}|\mathbf{A})P(\mathbf{A})}{P(\mathbf{B})}}$$

Practice Problems

Example 1. Birth Order A certain family has 6 children, consisting of 3 boys and 3 girls. Assuming that all birth orders are equally likely, what is the probability that the 3 eldest children are the 3 girls?

Example 2. Making Teams

(a) How many ways are there to split a dozen people into 3 teams, where one team has 2 people, and the other two teams have 5 people each?

(b) How many ways are there to split a dozen people into 3 teams, where each team has 4 people?

Example 3. Robberies A city with 6 districts has 6 robberies in a particular week. Assume the robberies are located randomly, with all possibilities for which robbery occurred where equally likely. What is the probability that some district had more than 1 robbery?

Example 4. Big Die I roll a 120-sided die. What is the probability that my roll is divisible by 2, 3, or 5?

Example 5. Subsets and Combinatorics.

(a) How many subsets are there of an n-element set?

(b) How many subsets of size k are there of an n-element set?

(c) Simplify the following expression using a story proof:

$$\sum_{k=0}^{n} \binom{n}{k}$$

Example 6. Story Proofs.

(a) You want to pick a team of m members, among which there are k leaders. Use a story proof to prove the following result:

$$\binom{r}{m}\binom{m}{k} = \binom{r}{k}\binom{r-k}{m-k}$$

(b) You want to walk from the coordinate $(0,0)$ to (n,n), only taking steps in a positive direction each time. Use a story based on the number of ways to complete that walk to prove this special case of the Vandermonde Identity.

$$\sum_{k=0}^{n} \binom{n}{k}^2 = \binom{2n}{n}$$

Example 7. Birth Probabilities. A couple tells you that they plan on having two children.

a) What is the set of possible outcomes for the sequence of genders of their kids (the sample space)?

b) Let $\mathbf{A}$ be the event that at least one of their kids is a girl. Assuming that having a boy or a girl are equally likely. What is $P(\mathbf{A})$?

c) Let $\mathbf{B}$ be the event that at least one of their kids is a boy. What is $P(\mathbf{A} \cap \mathbf{B})$?

d) What is $P(\mathbf{B}|\mathbf{A})$?

e) Are $\mathbf{A}$ and $\mathbf{B}$ independent events? Are $\mathbf{A}$ and $\mathbf{B}$ disjoint events?

Example 8. Biased Coins Suppose that you have two coins. One of these coins is fair, meaning that it is heads with probability 1/2, and the other is heads with probability 3/4.

a) Let's say that you randomly select one of these two coins and you select each coin with probability 1/2. If you flip this randomly selected coin once, what is the probability that you flip a heads?

b) You randomly select a coin and flipped it two times, obtaining heads each time. What is the probability that you randomly selected the fair coin?

c) Given this information, what is now the probability that if you flip the coin one more time, it will come up as heads?

2 Conditional Probability, Random Variables

"Conditioning is the soul of statistics." As you proceed through the course, you will begin to see more and more just how central conditioning is to all of probability and statistics. In particular, one of the most important fields of research in today's statistics departments is in Bayesian methods, which relies completely on Bayes' Rule, which in turn relies completely on the idea of conditional probability.

"Random variables (a.k.a. distributions) are the bread and butter of statistics." You are just starting to see the tip of the iceberg of the world of random variables and distributions. These distributions are so important because they let us re-use all the results over and over again by applying them to different models (think Newton's second law, $F = ma$ applied to lots of physics problems; similarly, the Binomial distribution used to model many probability problems). Much of your future work (and challenges) will be in setting up the "right" distribution in tackling a problem.

First-step analysis. One concrete reason that conditioning is useful to you, *immediately*: it lets you do **first-step analysis**. (I bolded that again just to emphasize how amazing and important that is.) Essentially, doing problems in probability allows you to ask a genie/great lords of mathematics/God any particular aspect you want to know about the problem at hand, then **condition on that aspect.**

Disjointness Versus Independence

Disjoint Events - **A** and **B** are disjoint when they cannot happen simultaneously, or

$$\begin{aligned} P(\mathbf{A} \cap \mathbf{B}) &= 0 \\ \mathbf{A} \cap \mathbf{B} &= \emptyset \end{aligned}$$

Independent Events - **A** and **B** are independent if knowing one gives you no information about the other. **A** and **B** are independent if and only if one of the following equivalent statements hold:

$$\begin{aligned} P(\mathbf{A} \cap \mathbf{B}) &= P(\mathbf{A})P(\mathbf{B}) \\ P(\mathbf{A}|\mathbf{B}) &= P(\mathbf{A}) \end{aligned}$$

Conditional Independence - **A** and **B** are conditionally independent given **C** if: $P(\mathbf{A} \cap \mathbf{B}|\mathbf{C}) = P(\mathbf{A}|\mathbf{C})P(\mathbf{B}|\mathbf{C})$. Conditional independence does not imply independence, and independence does not imply conditional independence.

Unions, Intersections, and Complements

De Morgan's Laws - Gives a useful relation that can make calculating probabilities of unions easier by relating them to intersections, and vice versa. De Morgan's Law says that the complement is distributive as long as you flip the sign in the middle.

$$\begin{aligned} (\mathbf{A} \cup \mathbf{B})^c &\equiv \mathbf{A^c} \cap \mathbf{B^c} \\ (\mathbf{A} \cap \mathbf{B})^c &\equiv \mathbf{A^c} \cup \mathbf{B^c} \end{aligned}$$

Complements - The following are true.

$$\begin{aligned} \mathbf{A} \cup \mathbf{A}^c &= \Omega \\ \mathbf{A} \cap \mathbf{A}^c &= \emptyset \\ P(\mathbf{A}) &= 1 - P(\mathbf{A}^c) \end{aligned}$$

Principle of Inclusion-Exclusion

$$P(\mathbf{A} \cup \mathbf{B}) = P(\mathbf{A}) + P(\mathbf{B}) - P(\mathbf{A} \cap \mathbf{B})$$

$$P(\mathbf{A} \cup \mathbf{B} \cup \mathbf{C}) = P(\mathbf{A}) + P(\mathbf{B}) + P(\mathbf{C}) - P(\mathbf{A} \cap \mathbf{B}) - P(\mathbf{B} \cap \mathbf{C}) - P(\mathbf{A} \cap \mathbf{C}) + P(\mathbf{A} \cap \mathbf{B} \cap \mathbf{C})$$

Sometimes you can avoid Inclusion-Exclusion by using the complement. The probability that at least one of A_i happen is equal to 1 minus the probability that none of them happen.

$$\begin{aligned} P(A_1 \cup A_2 \cup \dots \cup A_n) &= 1 - P((A_1 \cup A_2 \cup \dots \cup A_n)^c) \\ &= 1 - P(A_1^c \cap \dots \cap A_n^c) \text{ (De Morgan's Laws)} \end{aligned}$$

For quintessential examples: think birthday problem, or check out de Montmort's matching problem on page 22 of the textbook.

Joint, Marginal, and Conditional Probabilities and Bayes' Rule

Joint Probability - $P(\mathbf{A} \cap \mathbf{B})$ or $P(\mathbf{A}, \mathbf{B})$ - Probability of **A** *and* **B**.

Marginal (Unconditional) Probability - $P(\mathbf{A})$ - Probability of **A**

Conditional Probability - $P(\mathbf{A}|\mathbf{B})$ - Probability of **A** given **B** occurred.

Bayes' Rule - Bayes' Rule unites marginal, joint, and conditional probabilities. This is *the most important concept of the week*, and one of the backbones of statistics.

$$\boxed{P(\mathbf{A}|\mathbf{B}) = \frac{P(\mathbf{A} \cap \mathbf{B})}{P(\mathbf{B})} = \frac{P(\mathbf{B}|\mathbf{A})P(\mathbf{A})}{P(\mathbf{B})}}$$

Law of Total Probability

This is a useful way to break up a harder problem into simpler pieces, conditioning on what we wish we knew. For any event B and set of events $\mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3, ...\mathbf{A}_n$ that partition a space, the following are true:

$$\begin{aligned} &P(\mathbf{B}) = P(\mathbf{B}|\mathbf{A}_1)P(\mathbf{A}_1) + P(\mathbf{B}|\mathbf{A}_2)P(\mathbf{A}_2) + ...P(\mathbf{B}|\mathbf{A}_n)P(\mathbf{A}_n) \\ &P(\mathbf{B}) = P(\mathbf{B} \cap \mathbf{A}_1) + P(\mathbf{B} \cap \mathbf{A}_2) + ...P(\mathbf{B} \cap \mathbf{A}_n) \end{aligned}$$

or in the simplest case where A is just any event

$$\begin{aligned} &P(\mathbf{B}) = P(\mathbf{B}|\mathbf{A})P(\mathbf{A}) + P(\mathbf{B}|\mathbf{A^c})P(\mathbf{A^c}) \\ &P(\mathbf{B}) = P(\mathbf{B} \cap \mathbf{A}) + P(\mathbf{B} \cap \mathbf{A^c}) \end{aligned}$$

Practice Problems

Example 9. A Trip to Oz Suppose that "Toto, I've a feeling I don't want to be in Kansas any more," and you'd like to be off in the wonderful land of Oz instead. The marginal probability of teenage girls being swept up by tornadoes and being flung to Oz is reported to be 0.042 (hence the Answer to Life, the Universe, etc.; source: Wizard). On the one hand, you may wait around for a tornado, in which case you have a probability of 0.0042 of getting to Oz; whereas on the other hand, you may become possibly the youngest tornado hunter ever, in which case you have a 0.42 probability of getting to Oz.

a) What is the probability that your childhood dreams are shattered and you spend your life meaninglessly without ever getting to Oz?

b) Is getting to Oz independent from chasing tornadoes?

c) What is the probability that Dorothy, having made it to Oz, was a tornado chaser?

Example 10. A Tale of Two Suspects A crime is committed by one of two suspects, A and B. Initially, there is equal evidence against both of them. In further investigation at the crime scene, it is found that the guilty party had a blood type found in 10% of the population. Suspect A does match this blood type, whereas the blood type of Suspect B is unknown.

a) Given this new information, what is the probability that A is the guilty party?

b) Given this new information, what is the probability that Bs blood type matches that found at the crime scene?

Example 11. Simple Stock Markets A model for the movement of the price of a stock supposes that on each day the stock's price either moves up 1 unit with probability p or moves down with probability $1-p$.

a) What is the probability that after two days the stock will be its original price?

b) What is the probability that after 3 days the stock will have increased by 1 unit?

Example 12. The Flippant Juror. (Frederick Mosteller, *Fifty Challenging Problems in Probability*) A three-man jury has two members each of whom independently has probability p of making the correct decision and a third member who flips a coin for each decision (majority rules). A one-man jury has probability p of making the correct decision. Which jury has the better probability of making the correct decision?

Example 13. A Simple Random Walk. A very drunk man goes for a walk. He sets off on a random walk along a line: at each step, he flips a fair coin (somehow) and either takes one step to the right of his current position or takes one step to the left of his current position, depending on whether the coin lands heads or tails.

a) What is the probability that after n steps, the man will end up exactly back in his original location?

b) It just so happens that at one end of the street, 2 steps to the man's left, there is a bar and at the other end of the street, 3 steps to the man's right, there is an AA clinic. Calculate the probability that the man will reach the bar before the AA clinic.

3 Expectation, Distributions, Discrete RVs

"Conditioning is the soul of statistics." *"Imma let you finish, but conditioning is one of the most important things you learn in this course of ALL TIME."* It lets you do **first-step analysis**! (Recall what that is.) Essentially, if you're thinking, "Gosh, I'd love to know if this thing happened before I can calculate my answer," then **condition on that event.**

Random variables, distributions, and what? RVs, PMFs, CDFs, PDFs, XYZs, ABCs, WTFs, ... There seems to be an alphabet soup of acronyms related to random variables. It's super unclear what exactly these random variables even *are*, much less what the rest of these things are. Think of the **distributions** as *blueprints* for a building. If you've ever been in an apartment complex, you'll know that these realtors often construct 20 apartments from the same blueprint. Each apartment is then a **random variable**: that is, a particular realization of the blueprint, or the distribution.

Now, these apartments, *due to the fact that they're from the same blueprint*, have a lot of features in common: for example, they all have 8 central common rooms, 4 corner suites, 2 elevators, and so forth. Similarly, if X and Y are random variables from the same distribution, they share a lot of common features: these *features* include the **probability mass function (PMF)** and the **cumulative distribution function (CDF)**.

Random Variables are the Bread and Butter of Statistics

Formal Definition - A random variable X is a *function* mapping the sample space S into the real line.

Descriptive Definition - Random variables are often denoted by capital letters, usually X and Y, and associate each *outcome*, such as $\{H, H, T\}$ with a *number*, such as 2 for the number of heads. This is because it's hard to do calculations with outcomes (how do I add $\{H, H, T\}$ and $\{T, H, T\}$?), whereas it's easy to do so with numbers. The "random" part comes from the fact that the outcomes are uncertain.

Examples - Let's say you now roll 2 die. Your sample space is:
$S = \{(1,1), (1,2), (1,3), (1,4), (1,5), (1,6), (2,1), \ldots, (6,5), (6,6)\}$ (for a total of 36 possible outcomes). Here are possible random variables you can define:

- X = Sum of two dice rolls (maps sample space to $\{2, 3, \ldots, 12\}$)
- X = Maximum of two dice rolls (maps sample space to $\{1, 2, \ldots, 6\}$)
- X = Absolute Difference of dice rolls (maps sample space to $\{0, 1, \ldots, 5\}$)
- X = Product of dice rolls (maps sample space to $\{1, 2, 3, 4, 5, 6, 8, 9, \ldots, 36\}$)
- X = Number of dice rolls that are prime (maps sample space to $\{0, 1, 2\}$)
- X = Number of dice you rolled (maps sample space to $\{2\}$) This r.v. is degenerate as there is no randomness.
- X = Twice the square of the maximum of your two dice rolls (maps sample space to $\{2, 8, \ldots, 72\}$). Functions of random variables are also random variables.

Random Variables Expressing Events - Let's say that X is the random variable expressing the sum of two dice rolls. $\{X = 7\}$ is the event that your dice summed to 7, and $\{X \leq 7\}$ is the event that your dice summed to at most 7. $\{X = 2, X = 3, X = 5, X = 7, X = 11\}$ is also an event (the event that X is prime), and so is $\{X = 2, X = 9\}$.

Conditioning is the Soul of Statistics

Law of Total Probability with $\mathbf{B}$ and $\mathbf{B^c}$ (special case of a partitioning set), and with Extra Conditioning (just add C!)

$$\begin{aligned}
P(\mathbf{A}) &= P(\mathbf{A}|\mathbf{B})P(\mathbf{B}) + P(\mathbf{A}|\mathbf{B^c})P(\mathbf{B^c}) \\
P(\mathbf{A}) &= P(\mathbf{A} \cap \mathbf{B}) + P(\mathbf{A} \cap \mathbf{B^c}) \\
P(\mathbf{A}|\mathbf{C}) &= P(\mathbf{A}|\mathbf{B}, \mathbf{C})P(\mathbf{B}|\mathbf{C}) + P(\mathbf{A}|\mathbf{B^c}, \mathbf{C})P(\mathbf{B^c}|\mathbf{C}) \\
P(\mathbf{A}|\mathbf{C}) &= P(\mathbf{A} \cap \mathbf{B}|\mathbf{C}) + P(\mathbf{A} \cap \mathbf{B^c}|\mathbf{C})
\end{aligned}$$

Law of Total Probability with a partitioning $\mathbf{B}_0, \mathbf{B}_1, \mathbf{B}_2, \mathbf{B}_3, \ldots, \mathbf{B}_n$, and applied to random variables $\mathbf{X}$, $\mathbf{Y}$.

$$\begin{aligned}
P(\mathbf{A}) &= \sum_{i=0}^{n} P(\mathbf{A}|\mathbf{B}_i)P(\mathbf{B}_i) \\
P(\mathbf{Y} = y) &= \sum_{k} P(\mathbf{Y} = y|\mathbf{X} = k)P(\mathbf{X} = k)
\end{aligned}$$

Bayes' Rule, and with Extra Conditioning (just add C!)

$$\begin{aligned}
P(\mathbf{A}|\mathbf{B}) &= \frac{P(\mathbf{A} \cap \mathbf{B})}{P(\mathbf{B})} = \frac{P(\mathbf{B}|\mathbf{A})P(\mathbf{A})}{P(\mathbf{B})} \\
P(\mathbf{A}|\mathbf{B}, \mathbf{C}) &= \frac{P(\mathbf{A} \cap \mathbf{B}|\mathbf{C})}{P(\mathbf{B}|\mathbf{C})} = \frac{P(\mathbf{B}|\mathbf{A}, \mathbf{C})P(\mathbf{A}|\mathbf{C})}{P(\mathbf{B}|\mathbf{C})}
\end{aligned}$$

PMF, CDF, and Independence

Probability Mass Function (PMF) (Discrete Only) gives the probability that a random variable takes on the value X.

$$P_X(x) = P(X = x)$$

Cumulative Distribution Function (CDF) gives the probability that a random variable takes on the value x or less (the second sum is only for discrete random variables):

$$F_X(x_0) = P(X \leq x_0) = P(X = 0) + P(X = 1) + \cdots + P(X = x_0)$$

Independence - Intuitively, two random variables are independent if knowing one gives you no information about the other. X and Y are independent if for ALL values of x and y:

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

Gambler's Ruin

Problem Statement - Gamblers A and B have a series of bets, and they bet \$1 with each other until one of them is bankrupt. A starts with $\$i$ and B starts with $\$(N - i)$. p is the probability that A wins a bet, and $q = 1 - p$ is the probability that B wins the bet. We wish to find P_i, which is the probability that A wins this game given that A starts with $\$i$.

We start off this problem by conditioning on the first step. If A wins the first bet, then the probability that A wins is P_{i+1}, and if A loses the first bet, then the probability that A wins is P_{i-1}. We know the boundary conditions are $P_0 = 0$ and $P_N = 1$. Thus, using first-step analysis:

$$P_i = P(\text{Win First Bet})P(\text{Win at } i+1) + P(\text{Lose First Bet})P(\text{Win at } i-1) = pP_{i+1} + qP_{i-1}$$

We start off the difference equation with the natural guess that $\boxed{P_i = x^i}$ (**default difference equations guess** that will usually work). We can now plug in $P_i = x^i$.

$$x^i = px^{i+1} + qx^{i-1}$$

$$\Rightarrow 0 = px^{i+1} - x^i + qx^{i-1}$$

The RHS of this equation is known as the **characteristic polynomial**. Assuming that $x \neq 0$, we can cancel out x^{i-1} and get:

$$0 = px^2 - x + q$$

The two roots of this polynomial are

$$\alpha_1 = \frac{1 + \sqrt{1 - 4p(1-p)}}{2p} \qquad \alpha_2 = \frac{1 - \sqrt{1 - 4p(1-p)}}{2p}$$

$$\Rightarrow \alpha_1 = 1 \qquad \alpha_2 = \frac{q}{p}$$

Thus we know that our characteristic polynomial is of the form:

$$P_i = c_1\alpha_1^i + c_2\alpha_2^i = c_1 1^i + c_2\left(\frac{q}{p}\right)^i = c_1 + c_2\left(\frac{q}{p}\right)^i$$

We plug in $P_0 = 0$ and $P_N = 1$ to solve for c_1 and c_2, and we get the general answer (where $p \neq q$). The solution where (where $p = q$) is more technical and will not be covered in this class.

Solution - Starting at point i, say that the state of ruin is at 0 and the state of success is at N. The gambler has a probability p of moving up one step, and a corresponding probability $q = 1 - p$ of moving down one step. The probability P_i that the gambler reaches N before 0 starting at state i is:

$$P_i = \begin{cases} \frac{1-\left(\frac{q}{p}\right)^i}{1-\left(\frac{q}{p}\right)^N}, & \text{if } p \neq q \\ \frac{i}{N}, & \text{if } p = q = \frac{1}{2} \end{cases}$$

Solving Difference Equations

1. Write out the difference equation, to express P_i in terms of the neighboring P_{i+1}, P_{i-1} by first-step analysis.

2. Guess $P_i = x^i$ to get the characteristic polynomial

3. Find the roots of the characteristic polynomial, $\alpha_1, \alpha_2, \ldots, \alpha_k$

4. If the roots are distinct, the solution is of the form $P_i = c_1\alpha_1^i + c_2\alpha_2^i + \cdots + c_k\alpha_k^i$. Plug in the boundary conditions to solve for $c_1, c_2, \ldots, c_k$

Distributions

A distribution describes the probability that a random variable takes on certain values. Some distributions are commonly used in statistics because they can help model real life phenomena. We can use Binomial distributions to model the number of homeruns a player hits per season, or a Bernoulli distribution to model the occurrence of any event.

Bernoulli and Binomial Distributions

Bernoulli The Bernoulli distribution is the simplest case of the Binomial distribution, where we only have one trial, or $n = 1$. Let us say that X is distributed Bern(p). We know the following:

Story. X "succeeds" (is 1) with probability p, and X "fails" (is 0) with probability $1 - p$.

Example. A fair coin flip is distributed Bern($\frac{1}{2}$).

PMF. The probability mass function of a Bernoulli is:

$$P(X = x) = p^x(1-p)^{1-x}$$

or simply

$$P(X = x) = \begin{cases} p, & x = 1 \\ 1-p, & x = 0 \end{cases}$$

Binomial Let us say that X is distributed Bin(n, p). We know the following:

Story X is the number of "successes" that we will achieve in n independent trials, where each trial can be either a success or a failure, each with the same probability p of success. We can also say that X is a sum of multiple independent $Bern(p)$ random variables. Let $X \sim \text{Bin}(n, p)$ and $X_j \sim \text{Bern}(p)$, where all of the Bernoullis are independent. We can express the following:

$$X = X_1 + X_2 + X_3 + \cdots + X_n$$

Example If Jeremy Lin makes 10 free throws and each one independently has a $\frac{3}{4}$ chance of getting in, then the number of free throws he makes is distributed Bin($10, \frac{3}{4}$), or, letting X be the number of free throws that he makes, X is a Binomial Random Variable distributed Bin($10, \frac{3}{4}$).

PMF The probability mass function of a Binomial is:

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

Binomial Coefficient $\binom{n}{k}$ is a function of n and k and is read n *choose* k, and means out of n possible distinguishable objects, how many ways can I possibly choose k of them? The formula for the binomial coefficient is:

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

Geometric, Negative Binomial, and Hypergeometric Distributions

Geometric Let us say that X is distributed Geom(p). We know the following:

Story X is the number of "failures" that we will achieve before we achieve our first success. Our successes have probability p.

Example If each pokeball we throw has a $\frac{1}{10}$ probability to catch Mew, the number of failed pokeballs will be distributed Geom($\frac{1}{10}$).

PMF With $q = 1 - p$, the probability mass function of a Geometric is:

$$P(X = k) = q^k p$$

Negative Binomial Let us say that X is distributed NBin(r, p). We know the following:

Story X is the number of "failures" that we will achieve before we achieve our rth success. Our successes have probability p.

Example Thundershock has 60% accuracy and can faint a wild Raticate in 3 hits. The number of misses before Pikachu faints Raticate with Thundershock is distributed NBin$(3, .6)$.

PMF With $q = 1 - p$, the probability mass function of a Negative Binomial is:

$$P(X = n) = \binom{n + r - 1}{r - 1} p^r q^n$$

Hypergeometric Let us say that X is distributed Hypergeometric(w, b, n). We know the following:

Story In a population of b undesired objects and w desired objects, X is the number of "successes" we will have in a draw of n objects, without replacement.

Example 1) Let's say that we have only b Weedles (failure) and w Pikachus (success) in Viridian Forest. We encounter n of the Pokemon in the forest, and X is the number of Pikachus in our encounters. 2) The number of aces that you draw in 5 cards (without replacement). 3) You have w white balls and b black balls, and you draw b balls. X is the number of white balls you will draw in your sample. 4) Elk Problem - You have N elk, you capture n of them, tag them, and release them. Then you recollect a new sample of size m. How many tagged elk are now in the new sample?

PMF The probability mass function of a Hypergeometric is:

$$P(X = k) = \frac{\binom{w}{k}\binom{b}{n-k}}{\binom{w+b}{n}}$$

Practice Problems

Example 14. Properties of the Binomial.

a) Suppose $X \sim \text{Bin}(n_1, p)$ and independently $Y \sim \text{Bin}(n_2, p)$. What is the distribution of $X + Y$? Explain.

b) What is $P(X \geq 2)$?

c) What is $P(X + Y = 10)$? Assume that $n_1 + n_2 \geq 10$.

d) What is one reason why $X - Y$ cannot be Binomial?

e) Can you construct two random variables X and Y both distributed Bin$(3, \frac{1}{2})$ such that $P(X = Y) = 0$?

Example 15. The Triwizard Tournament. "Eternal glory" awaits those who win the Triwizard Tournament, held between Hogwarts, Durmstrang, and Beauxbatons. Due to the incredibly dangerous nature of the tournament, it was discontinued after 1792, brought back in 1994 because You-Know-Who wanted to be revived and such, and is brought back for a 20-year commemoration in 2014. This time, the schools decided to let the top 4 students compete in the tournament, with, for example, the top student in each of the four Houses of Hogwarts competing for the prize. Let H be the highest ranking in the tournament achieved by a Hogwarts student (because who cares about Durmstrang and Beauxbatons, am I right?). All 12! orderings of students are equally likely.

a) Determine the PMF of H.

b) Calculate $E(H)$, that is the expected highest ranking of any Hogwarts student.

Example 16. Gotta Catch ALL the Mews. Your pokeball has a constant p chance of success against Mew.

a) You plan to keep throwing Pokeballs until you catch it. What is the expected number of Pokeballs that will fail before you catch Mew?

b) You're actually cheating (hey, we all used Gameshark at some point), so you've set up the game such that every wild Pokemon encounter is with Mew. Since you love Stat 110, you've decided to try running into Mew n times, and seeing how many runs of consecutive successes and failures you have in catching Mew. Compute the expected number of runs.

Example 17. "I'm Gonna Make Him an Offer He Must Refuse." Suppose that you're at a carnival, and you've walked into a slightly shady-looking tent. Before you know it, you've been swindled out of \$5 to play a game with the dealer (it sounded like a good idea at the time!). In turns out that the \$5 was just to enter the tent; you have to shell out another \$50 just to play the game. This game takes the following stages:

1. In the first stage, the dealer rolls a 10-sided die, with equal probability of landing on any of $1, \dots, 10$. You must guess the number he rolled. If you are correct, your payoff increases by *\$110*, and you get to move on to round 2. If you guess wrong, then you lose \$15.

2. In the second stage, the dealer rolls a 100-sided die, with equal probability of landing on any of $1, \dots, 100$. You again guess the number he rolled. If you are correct, your payoff is whatever you got in round 1 plus *\$1000*! If you guess wrong, you get to keep the payoff from round 1.

There is one twist: you're a particularly devout and pious individual, so you can ask God (of whatever religious denomination you happen to be part of) what number the dealer has rolled. Unfortunately, your God is slightly cynical, so He/She/It only bothers to tell you whether the number is in the first half or second half (i.e. from 1 to 5 or 6 to 10, or 1 to 50 or 51 to 100). You can choose to walk away rather than playing the game, but you won't get a refund on your \$5 (remember, the dealer's actually a swindler).

a) What is the expected value of each of these options? (i.e. walk away now, play and pray to God on your first round, or play and pray to God on your second round)

b) How would these expected values change if your piousness *really* pays off, and you get divine inspiration on the first round (without asking God), allowing you to tell that the roll is between 4 and 10?

Challenge Problems

Example 18. A Betting Game. We play a game where you give me a number and then I roll two standard dice. If the sum of the two die equals the number you give me, you win the amount of your number *squared* in dollars. What number should you choose (to maximize expected return)?

Example 19. College Application Mishap. A lazy high school senior types up applications and envelopes to n different colleges. The letters are randomly put into the envelopes.

a) On average, how many applications went to the right college?

b) A match occurs if college A receives college B's envelope AND if college B receives college A's envelope. On average, how many matches occur?

4 Poisson, LotUS, Fundamental Bridge

Linearity of Expectation. As you saw in problem 4 and 5 on last week's problem set, **linearity of expectation** is one of the most magical, beautiful things to come out of this course. It works for independent, dependent, correlated, uncorrelated, any type of random variables you can realistically imagine. Note that linearity of expectation *directly implies* that a top-choice problem-solving strategy should be to break up a random variable into the sum of other random variables (since then you can just add up their expectations!).

Fundamental Bridge. You've split up a random variable into a bunch of indicator random variables – now what? Beautiful trick #2: **the fundamental bridge**. That is, $E[I_A] = P(A)$ for any event A! Thus, whenever we have a sum of expectations of indicators, we can simply sum their probabilities of occurrence, which by **symmetry**, can often be reduced to a simple sum (problem 4, or the Mew problem).

Law of the Unconscious Statistician (LotUS). Given its nickname and its semi-common-sense nature, you might be inclined to completely overlook **LotUS** when you first learn it. Yet LotUS is actually quite an intriguing and immensely useful result. As you'll learn later in the course, a function $g(X)$ of a random variable X may often follow completely different distributions (pmf/pdf) than the variable X itself; yet *LotUS* tells us that in terms of *expectations*, we can just treat $g(X)$ as if it followed the **same distribution** as X!

Poisson. Because we add **the Poisson distribution** as almost an afterthought in most introductions to probability, the true utility of the distribution remains unclear. Yet to be completely precise, the Poisson is probably one of the most useful distributions (if not the most!) you will learn in Stat 110. It is applied both in itself and as an approximation in an endless variety of scenarios; whenever you have a situation that involves *a count* of some sort, the Poisson is bound to play a role.

Law of the Unconscious Statistician (LotUS)

How do I find the expected value of a function of a random variable? Normally, you would find the expected value of X this way:

$$E(X) = \Sigma_x x P(X = x)$$

LotUS states that you can find the expected value of a *function of a random variable* g(X) this way:

$$E(g(X)) = \Sigma_x g(x) P(X = x)$$

What's a function of a random variable? A function of a random variable is also a random variable. For example, if X is the number of bikes you see in an hour, then $g(X) = 2X$ could be the number of bike wheels you see in an hour. Both are random variables.

What's the point? You don't need to know the PDF/PMF of $g(X)$ to find its expected value. All you need is the PDF/PMF of X.

Poisson Distribution

If X is distributed Pois(λ), we know the following:

Story There are rare events (low probability events) that occur many different ways (high possibilities of occurences) at an average rate of λ occurrences per unit space or time. The number of events that occur in that unit of space or time is X.

Example A certain busy intersection has an average of 2 accidents per month. Since an accident is a low probability event that can happen many different ways, the number of accidents in a month at that intersection is distributed Pois(2). The number of accidents that happen in two months at that intersection is distributed Pois(4)

PMF The PMF of a Poisson is:

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

Poisson Approximation

Binomial Random Variable Imagine that you model a particular situation using $X \sim \text{Bin}(n, p)$. Since we know the PMF of $\text{Bin}(n, p)$, we can, of course, compute the probability of every possible outcome $X = x$. Unfortunately, suppose that n is extremely large (i.e., $n \to \infty$) while p remains very small. This is a highly common case; as examples, consider traffic accidents at intersections, lottery wins, watching the sky for shooting stars, etc. In this case, it is difficult to compute the probabilities for the Binomial because of the factorial terms!

Poisson Approximation When we have $X \sim \text{Bin}(n, p)$ with $n \to \infty$ and $p \to 0$, with $n \cdot p = \lambda$ constant, then we can approximate X with a Poisson; i.e. $\boxed{X \approx \text{Pois}(np)}$. The Poisson can be envisioned as an infinite number of trials, each of which has a very small probability p of occurring, with a mean of np occurring per unit time, so this makes intuitive sense, and also works asymptotically.

Generalization Since a Binomial is just the sum of n independent Bernoulli trials, whenever we can break up a random variable:

$$Y = I_1 + I_2 + \cdots + I_n$$

with n very large, and each of the indicators having some small but constant probability p of success, we can approximate $Y \approx \text{Pois}(np)$.

Practice Problems

Example 20. Memoryless $\neq$ Amnesia. Suppose that you are waiting in line to buy some delicious Greenhouse Cafe pizza. After some careful modeling, you determine that the number of minutes from the moment you enter the line until you finally swipe your Boardplus and can enjoy your pizza follows $T \sim \text{Geom}(p)$ (this is a really strange line).

a) What is $P(T > t)$, the probability that you have to wait more than t minutes?

b) Suppose that you've already waited s minutes in line. What is the probability that you have to wait an additional t minutes, that is, $P(T > t + s | T > s)$?

c) Prove that the Geometric distribution is **memoryless**; that is: $\boxed{P(T > t + s | T > s) = P(T > t)}$.

Example 21. Fun with Unconscious Statisticians. In order to truly appreciate how wonderful LotUS is, you've decided to test out its magical powers for yourself by computing expectations for a bizarre array of random variables. For this problem, suppose that $X \sim \text{Bin}(n, p_1)$, $Y \sim \text{Geom}(p_2)$, and $Z \sim \text{Pois}(\lambda)$.

a) Suppose that $A = \frac{X}{n-X+1}$. Find $E[A]$.

b) Suppose that $B = \frac{1}{1+Z}$. (This is a famous example.) What is $E[B]$?

c) Find $E[e^{tY}]$.

Example 22. Random Variables and Transformations. Derive, through logic or algebra, the distributions of the following random variables.

a) Suppose that $X_1 \sim \text{Pois}(\lambda_1)$ and $X_2 \sim \text{Pois}(\lambda_2)$ independently. What is the distribution of $X_1 + X_2$?

b) Let $Y_k \sim \text{Geom}(p)$ independently. What distribution does $Z = \min_{k=1,\ldots,n} Y_k$ follow?

Example 23. Mixture Distributions & Expectations. Suppose that, as in the case of tornadoes, financial analysts and government officials have finally put together a "rating scale" for the current financial climate, ranging from 0 (extremely boring) to 7 (financial crisis of 2007-8). Denote this financial climate as F. As a starting model of the stock market, you've fit a model of one stock (say, the Dow Jones index); in a very crude approximation, you fit a discrete model for levels of the price, also ranging from 0 to 7, where 7 is the "highest" level for the stock price. Denote this stock as S. You note that the financial climate seems to follow $F \sim \text{Pois}(1)$; that is, it's usually extremely boring, but can sometimes hit crises.

a) Suppose that given $F = f$, $S \sim \text{Bin}(7, 1/f)$. What is the probability $P(S > 2)$?

b) What is the expected stock price, that is, $E[S]$?

c) What is the probability that the level of the stock price is greater than financial crisis rating?

Challenge Problems

Example 24. More LotUS and Linearity. More tricky problems involving the distributions of various random variables, and how to compute expectations of functions of random variables!

a) Suppose that $X \sim \text{Bin}(n, p)$ and $W|X \sim \text{Bin}(X, q)$. Find $E[W]$.

b) Suppose that $Y \sim \text{Pois}(\lambda)$. Find $E[Y!]$.

c) Suppose that $Z \sim \text{Geom}(p)$. Find $E[1/Z!]$.

d) Suppose that $Y \sim \text{Pois}(\lambda)$, and $V = \left(1 - \frac{1}{\lambda}\right)^Y$. Find $E[V]$.

Example 25. Splitting Up the Poisson. Let $X \sim \text{Pois}(\lambda_1)$, and $N \sim \text{Pois}(\lambda)$.

a) Find the PMF of $X|N$; that is, find $P(X = x|N = n)$. What distribution does it follow?

b) Let $Z = N - X$; find the PMF of Z, and name the distribution that it follows.

5 Continuous RVs, Transformations, Universality

This week covers a *ton* of new material, starting with continuous random variables, their PDF/CDFs, transformations between continuous random variables, universality, and so forth. As a result, these notes cover an extended overview section that maps out the scope of what has been covered, highlighting the key points and how these topics are linked. Please review this section if you feel lost, at a topical or conceptual level, while tackling problems or reviewing for the midterm! It's easy to get lost in this maze, so feel free to reach out with any questions, especially conceptual ones!

Continuous Random Variables. A continuous RV is one whose CDF can be represented as:

$$P(X < x) = \int_{-\infty}^{x} f_X(t)dt$$

Note that $P(X = x) = 0$ for every single x! In particular, $f_X(x) \neq P(X = x)$, unlike in the discrete case (consider what would happen if any of the $P(X = x) > 0$). On the other hand, we can think of:

$$P(X \in [x, x + \delta]) \approx f_X(x) \cdot \delta$$

With this little caveat, however, you can essentially do any calculation you could have done on discrete RVs (i.e. expected values, LotUS, variance, linearity) on continuous RVs, replacing *sums* by *integrals* whenever applicable.

What will be most useful in dealing with continuous RVs is to gain familiarity and comfort with moving from the *CDF* to the *PDF* and vice-versa, through differentiation and integration. Consequently, it will be very important for not only this topic, but also for the rest of the course, to gain a lot of practice doing this transition for all different sorts of continuous RVs, as the problem set this week asks you to do!

Transformation of Variables. This topic is rife with potential errors, and without practice it can be difficult to apply this idea to problems. The most important thing to remember about *transformation of variables* is the following formula, given that $Y = g(X)$ for some differentiable function g:

$$f_Y(y) = f_X(g^{-1}(y)) \cdot \left|\frac{dx}{dy}\right|$$

Note that $y = g(x) \Rightarrow x \in g^{-1}(y)$, in the sense that the function may not be one-to-one; for example, consider $y = x^2$; then $y = 1$ corresponds to both $x = \pm 1$. Thus, much care is necessary in transforming variables and finding their distributions, but with care the above formula is sufficient in most (one-variable) cases.

Universality of the Uniform. The most important application of transformation of variables is *universality of the uniform.* Basically, universality states that if you have some random variable X with CDF F_X, then we have $F_X^{-1}(U) \sim X$ where $U \sim \text{Unif}[0, 1]$.

To understand this, consider what F_X^{-1} is actually doing. Recall that F_X takes values of x and maps them onto $[0, 1]$, which can be seen as the percentile of values of X that x is greater than. Thus, F_X^{-1} takes that percentile and converts it to the x value that is at that percentile, yielding a distribution that is identical to the distribution of X itself.

(*Exercise.* Prove that the PDF of $F_X^{-1}(U)$ is identical to the PDF of X.)

Poisson Process. This is the third time you hear about the Poisson (i.e. the distribution and the approximation being the first two). The *Poisson Process* links two distributions we have learned so far: the *Poisson* and the *Exponential.* In essence, it states that if you have a process that occurs at some given rate λ, but for which the actual occurrences are distributed $\text{Expo}(\lambda)$ in time, then the *number* of occurrences by a *fixed time* is distributed $\text{Pois}(\lambda)$.

(*Exercise.* Prove that if the number of occurrences of an event in time t is distributed Pois(λ), then if T is the time of the first occurrence, $P(T \leq t)$ is the CDF of the Expo(λ) distribution.)

Adam + Eve. Adam's Law is another name for *law of iterated expectation*, which is also known as *law of total expectation*, or otherwise known as the *tower property* - it has a lot of names. But simply, it just states that:

$$E[X] = E[E[X|Y]] = E_Y[E_X[X|Y]]$$

which is a fact that we have used previously. This law is supremely useful in problem-solving, as it allows us to condition on a random variable as necessary in order to derive the expectation.

Because of its name, the *law of total variance*, as a complement to Adam's Law, has been called Eve's Law, and it states:

$$\text{Var}(Y) = E[\text{Var}(Y|X)] + \text{Var}(E[Y|X])$$

Discrete vs. Continuous. The following table may help in the transition from discrete RVs to continuous RVs, and compare the "equivalencies" between these two sets of objects.

	Discrete	**Continuous**
$P(X \leq x) =$	$F(x)$ (CDF)	$F(x)$ (CDF)
To find probabilities,	Add over PMF $P(X = x)$	Integrate over PDF $f(x) = F'(x)$
$E(X) =$	$\sum_x xP(X = x)$	$\int_{-\infty}^{\infty} xf(x)dx$
$E(g(X)) =$	$\sum_x g(x)P(X = x)$ (LOTUS Discrete)	$\int_{-\infty}^{\infty} g(x)f(x)dx$ (LOTUS Continuous)

Continuous Random Variables

How to I find the probability that a CRV takes on a value within an interval? Use the CDF (or the PDF, see below). To find the probability that a CRV takes on a value in the interval $[a, b]$, subtract the respective CDFs.

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

What is the Cumulative Density Function (CDF)? It is the probability that $X \leq x$, that is:

$$F(x) = P(X \leq x)$$

With the following properties. 1) F is increasing. 2) F is right-continuous. 3) $F(x) \to 1$ as $x \to \infty$, $F(x) \to 0$ as $x \to -\infty$

What is the Probability Density Function (PDF)? The PDF, $f(x)$, is the derivative of the CDF.

$$F'(x) = f(x)$$

Or alternatively,

$$F(x) = \int_{-\infty}^{x} f(t)dt$$

Note that by the fundamental theorem of calculus,

$$F(b) - F(a) = \int_{a}^{b} f(x)dx$$

Thus to find the probability that a CRV takes on a value in an interval, you can integrate the PDF, thus finding the area under the density curve.

How do I find the expected value of a CRV? The sum becomes an integral!

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

Recall: Expected value is *linear*. This means that for *any* random variables X and Y and any constants a, b, c, the following is true:

$$E(aX + bY + c) = aE(X) + bE(Y) + c$$

Universality of Uniform

1) $F(X) \sim$ Unif$[0, 1]$. When you plug any random variable into its own CDF, you get a Unif$[0, 1]$ random variable. For example, let's say that a random variable X has a CDF $F(x) = 1 - e^{-x}$ (i.e. Expo(1)). By the Universality of the the Uniform, if we plug in X into this function then we get a uniformly distributed random variable.

$$F(X) = 1 - e^{-X} = U \sim \text{Unif}[0, 1]$$

2) $F^{-1}(U) \sim X$. When you put a Unif$[0, 1]$ into an inverse CDF, you get the corresponding random variable. If $U \sim$ Unif$[0, 1]$, then for the above $F(x)$, $F^{-1}(u) = \ln\left(\frac{1}{1-u}\right)$. Thus:

$$F^{-1}(U) = \ln\left(\frac{1}{1-U}\right) \sim X$$

The key point is that *for any continuous random variable X, we can transform it into a uniform random variable and back by using its CDF.*

Uniform Distribution (Continuous)

Let us say that U is distributed Unif(a, b). We know the following:

Properties of the Uniform For a uniform distribution, the probability of an draw from any interval on the uniform is proportion to the length of the uniform. The PDF of a Uniform is just a constant, so when you integrate over the PDF, you will get an area proportional to the length of the interval.

Example Ryan throws darts really badly, so his darts are uniform over the entire wall because they're equally likely to appear anywhere. Ryan's darts have a uniform distribution on the surface of the wall. The uniform is the only distribution where the probably of hitting in any specific region is proportion to the area of that region, and where the density of occurrence in any one specific spot is constant throughout the whole support.

PDF and CDF

$$\text{Unif}(0,1) \qquad f(x) = \begin{cases} 1 & x \in [0,1] \\ 0 & x \notin [0,1] \end{cases} \qquad F(x) = \begin{cases} 0 & x < 0 \\ x & x \in [0,1] \\ 1 & x > 1 \end{cases}$$

$$\text{Unif}(a,b) \qquad f(x) = \begin{cases} \frac{1}{b-a} & x \in [a,b] \\ 0 & x \notin [a,b] \end{cases} \qquad F(x) = \begin{cases} 0 & x < a \\ \frac{x-a}{b-a} & x \in [a,b] \\ 1 & x > b \end{cases}$$

Normal Distribution (Continuous) (a.k.a. Gaussian)

Let us say that X is distributed $\mathcal{N}(\mu, \sigma^2)$. We know the following:

Central Limit Theorem The Normal distribution is ubiquitous because of the central limit theorem, which states that averages of independent identically-distributed variables will approach a normal distribution regardless of the initial distribution.

Transformable Every time we stretch or scale the normal distribution, we change it to another normal distribution. If we add c to a normally distributed random variable, then its mean increases additively by c. If we multiply a normally distributed random variable by c, then its variance increases multiplicatively by c^2. Note that for every normally distributed random variable $X \sim \mathcal{N}(\mu, \sigma^2)$, we can transform it to the standard $\mathcal{N}(0,1)$ by the following transformation:

$$\frac{X-\mu}{\sigma} \sim \mathcal{N}(0,1)$$

Example Heights are normal. Measurement error is normal. By the central limit theorem, the sampling average from a population is also normal.

Standard Normal - The Standard Normal, denoted Z, is $Z \sim \mathcal{N}(0,1)$

PDF

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

CDF - It's too difficult to write this one out, so we express it as the function $\Phi(x)$

Exponential Distribution (Continuous)

Let us say that X is distributed Expo(λ). We know the following:

Story You're sitting on an open meadow right before the break of dawn, wishing that airplanes in the night sky were shooting stars, because you could really use a wish right now. You know that shooting stars come on average every 15 minutes, but it's never true that a shooting star is ever"due" to come because you've waited so long. Your waiting time is memoryless, which means that the time until the next shooting star comes does not depend on how long you've waited already.

Example The waiting time until the next shooting star is distributed Expo(4). The 4 here is λ, or the rate parameter, or how many shooting stars we expect to see in a unit of time. The expected time until the next shooting star is $\frac{1}{\lambda}$, or $\frac{1}{4}$ of an hour. You can expect to wait 15 minutes until the next shooting star.

All Exponentials are Scaled Versions of Each Other

$$Y \sim \text{Expo}(\lambda) \rightarrow X = \lambda Y \sim \text{Expo}(1)$$

PDF and CDF The PDF and CDF of a Exponential is:

$$f(x) = \lambda e^{-\lambda x}, \quad x \in [0, \infty) \qquad\qquad F(x) = P(X \leq x) = 1 - e^{-\lambda x}, \quad x \in [0, \infty)$$

Memorylessness The Exponential Distribution is the sole continuous memoryless distribution. This means that it's always "as good as new", which means that the probability of it failing in the next infinitesimal time period is the same as any infinitesimal time period. This means that for an exponentially distributed X and any real numbers t and s,

$$P(X > s + t | X > s) = P(X > t)$$

Given that you've waited already at least s minutes, the probability of having to wait an additional t minutes is the same as the probability that you have to wait more than t minutes to begin with. Here's another formulation.

$$X - a | X > a \sim \text{Expo}(\lambda)$$

Here are two consequences of the memoryless property.

1. If waiting for the bus is distributed exponentially with $\lambda = 6$, no matter how long you've waited so far, the expected additional waiting time until the bus arrives is always $\frac{1}{6}$, or 10 minutes. The distribution of time from now to the arrival is always the same, no matter how long you've waited.

2. The instantaneous failure rate $\left(\frac{f(x)}{1-F(x)}\right)$, or hazard function, is constant (try to think of the hazard function in terms of a conditional probability). That is, if the lifetime of a light bulb is distributed exponentially, the instantaneous rate of failure is always the same - the lightbulb will always have the same probability of failing in the next infinitesimal time unit for the duration of its lifetime. This is unlike real life light bulbs, whose failure rates grow as they grow older or my printer, which has a failure rate directly proportional to my current need for it.

Practice Problems

Example 26. Let's Prove It! In this problem, we will use the techniques of transformation of variables and establishing the PMF in an exercise to prove two of the central results: the *universality of the uniform* and the *Poisson process*. (Hopefully this will serve to give you intuition about the results, convince you that they're true, and give you practice in these techniques at the same time!)

a) Suppose that $U \sim \text{Unif}[0,1]$ and X has CDF F_X. Consider the random variable defined by $F_X^{-1}(U)$, and prove that its PDF is identical to the PDF of X.

b) Consider the Poisson process where the number of occurrences of an event in a fixed time t is given by $N(t) \sim \text{Pois}(\lambda t)$. Let T_i denote the time at which the i^{th} occurrence of the event happens. Find the probability $P(T_1 \leq t)$, and show that it is identically equal to the CDF of the *Exponential distribution*.

c) Similarly, consider the probability $P(T_{k+1} - T_k \leq t)$, and show that it also follows a Expo(λ) distribution.

Example 27. I Heard You Like PDFs ... So we are going to gain much practice converting to and from CDFs!

a) Consider $X \sim \text{Expo}(\lambda)$, i.e. its PDF is $f_X(x) = \lambda e^{-\lambda x}$. Derive the CDF of X, and show that it is a valid CDF.

b) Consider the Kumaraswamy distribution, defined by its CDF:

$$F_X(x) = 1 - (1 - x^a)^b$$

for $x \in [0,1]$. It is also closely related to the Beta distribution, of which the *arcsine distribution* is a special case. Show that the Kumaraswamy CDF is a valid CDF, and derive the PDF for the distribution.

c) Find the inflection point(s) of the Kumaraswamy distribution with parameters $a = 2, b = 5$.

Example 28. ... So I Put a Variable Transformation in Your Variable Transformation ...

a) Suppose that M is the number of minutes in a basketball game, and $M \sim \text{Expo}(\lambda)$. Find the PDF of the number of seconds in a basketball game, and in particular show that it is not $f_M(60 \cdot m)$.

b) Let T be the area of a triangle whose base equals its height, and whose height H is a random variable distributed as $H \sim \text{Unif}[0,1]$. Find the PDF of T.

c) Suppose that $X \sim \mathcal{N}(0,1)$, i.e. X is distributed standard normal. Find the PDF of $Y = \sigma X + \mu$, proving that $Y \sim \mathcal{N}(\mu, \sigma^2)$.

Example 29. ... So You Can Adam's Law While You Eve's Law! This example introduces the applications of Adam's Law and Eve's Law, and in particular considers their use for indicator random variables.

Suppose that you and a friend could really use a wish right now, so you're sitting outside pretending that airplanes in the night sky are shooting stars. In particular, suppose that JetBlue Airlines will fly a plane over the Boston skyline in J minutes, where $J \sim \mathcal{N}(5, 10)$ and Delta Airlines will fly a plane over in D minutes, where $D \sim \text{Unif}[2, 10]$. The airlines schedule their flights independently of each other. You and your friend decide that each of you gets to choose one of the airlines and can make a wish when a plane of that airline flies over.

a) What is the probability that JetBlue flies over before Delta?

b) Find the expectation of T, denoting the minutes that you have to wait before a plane of either airline flies over.

c) Find the variance of T.

6 Moments + Moment Generating Functions

This week is about *moments*, and how to generate them via *moment generating functions (MGFs).* As we move further into constructions based on continuous RVs and the idea of maximization, we must flex our calculus techniques more and more, so becoming (re-)comfortable with the idea of Taylor series and methods of integration would be very helpful.

Moments. The k^{th} *moment* of a distribution is defined as:

$$E[X^k]$$

which exists if $E|X^k| < \infty$ (or is said to be *integrable*).

What you are used to is probably the *central moment* of a distribution, such as the variance. We define the k^{th} *central moment* of a distribution as:

$$E[(X-\mu)^k]$$

Finally, we also deal with *standardized moments* of a distribution, such as the skewness and kurtosis. We define the k^{th} *standardized moment* of a distribution as:

$$E\left[\left(\frac{X-\mu}{\sigma}\right)^k\right]$$

Important moments and central moments include:

- **Mean**. $E[X]$ is the first moment.
- **Variance**. $E[(X-\mu)^2]$ is the second central moment.
- **Skewness**. $E\left[\left(\frac{X-\mu}{\sigma}\right)^3\right]$ is the third standardized moment.
- **Kurtosis**. $E\left[\left(\frac{X-\mu}{\sigma}\right)^4\right] - 3$ is essentially the fourth standardized moment.

(*Exercise.* Prove that any Normal distribution has a kurtosis of 0.)

(*Exercise.* Pictorially depict a typical right-skewed distribution, and show how the mean, mode, and median are related in such a distribution.)

Taylor Series. While a seemingly useless exercise in calculus, the Taylor series turns out to be one of the most useful tools for any mathematical discipline. It ties together the theory of functions with differential calculus and infinite summation, and is in general extremely useful for getting approximations of functions when the arguments are special in some way (i.e. near 0 or near the maximum).

Important Taylor series include the following:

- **Exponential**. The quintessential Taylor series expansion is done on the exponential function, e^x:

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots = \sum_{n=0}^{\infty} \frac{x^n}{n!}$$

- **Sine/Cosine**. We have:

$$\sin(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \cdots = \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n+1)!} x^{2n+1}$$

$$\cos(x) = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \cdots = \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n)!} x^{2n}$$

- **Geometric series**. We can derive a number of useful formulae for odd-looking polynomials through the geometric series:

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

$$\frac{1}{(1-x)^2} = \sum_{n=0}^{\infty} nx^{n-1}$$

(*Exercise.* Prove that the Taylor series expansion of the exponential function is $e^x = 1 + x + x^2/2 + x^3/3! + \cdots$.)

Moment Generating Functions (MGFs). The clever idea behind MGFs is the fact that the moments of a distribution completely characterize it; if we knew all the moments, we would "know" the distribution. Thus, one idea was to "string together the moments" in one function, called the MGF, that would contain in it all the "information" regarding the moments. If that sounds incredibly vague, that's because it is! While that may serve as a useful conceptual background for people who already know what an MGF is, for starters it's more useful to think of the MGF as like a CDF or PDF: a function that is another "property" of a distribution like the Normal or Exponential. It is *unique* to the distribution at hand, and thus comparing MGFs can tell us when two distributions are the same. Moreover, differentiating the MGF can yield any moment we desire.

We *define* the moment generating function as:

$$M_X(t) = E[e^{tX}]$$

Note that the right-hand side is just a function of X; here again LotUS comes to the rescue for computation!

There are two important things to note about MGFs:

1. **Generating moments**. We compute the *moments* of the distribution of X as follows:

$$E[X^k] = M_X^{(k)}(0)$$

What that means is to take the k^{th} derivative of $M_X(t)$, and evaluate it at $t = 0$.

2. **Summing independent RVs**. Since if X_1, X_2 are independent, we can evaluate $E[f(X_1)g(X_2)] = E[f(X_1] \cdot E[g(X_2)]$, the MGF of the sum of independent random variables is just the product of their MGFs:

$$M_{X_1+\cdots+X_n}(t) = M_{X_1}(t) \cdots M_{X_n}(t)$$

This makes the computation extremely easy, especially when the X_i are i.i.d. from some distribution. In this case:

$$M_{X_1+\cdots+X_n}(t) = M_{X_i}(t)^n$$

(*Exercise.* Prove that the MGF actually generates moments; that is, prove that $M_X^{(k)}(0) = E[X^k]$.)

Practice Problems

Example 30. Let's Prove It! Round Two. In this problem, we will prove some interesting properties about the Normal distribution, and make you comfortable with doing computations with moments.

a) Prove that an arbitrary $\mathcal{N}(\mu, \sigma^2)$ distribution has a kurtosis of 0.

b) Prove that any such Normal distribution has skewness of 0.

c) Prove that any such Normal distribution has every odd standardized moment equal to 0. Discuss the implications for the shape of the Normal distribution.

Example 31. Mean, Median, and Mode. Any introductory statistics courses teaches the idea of the mean, median, and mode as describing the shape of a distribution. We will explore how to compute these moments and what they actually imply about the distribution.

The *Gamma distribution* is one of the most important continuous distributions, as the generalization of the Exponential, including such staples of statistical inference as the Chi-squared distribution. (Take a look at your textbook cover and note the central place that the Gamma occupies on the cover!)

The PDF of the Gamma distribution for $X \sim \text{Gamma}(\alpha, \beta)$ is of the form:

$$f_X(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$$

a) Find the mode of the distribution in terms of its parameters, α, β.

b) Find the mean of the distribution in terms of its parameters, and compare it to the value for the mode. Discuss what it implies about the shape of the distribution.

c) Given your answer to part (b), do you expect the skewness of the distribution to be positive or negative? Check your answer by computing the skewness.

Example 32. Moment Generating Functions and Convolutions. Summing two independent random variables results in an operation referred to as a *convolution.* MGFs make this computation tractable. A well-known derivative distribution of the Gamma is the *Chi-squared distribution*, or χ^2_k, which has the PDF:

$$f(x) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} e^{-x/2}$$

Chi-squared distributions commonly arise in statistics because the square of a standard normal distribution is χ^2_1.

a) Show that the sum of k independent χ^2_1 random variables is distributed as χ^2_k.

b) Show that if $X \sim \mathcal{N}(0, 1)$, then $Y = X^2 \sim \chi^2_1$ by finding the PDF of Y and showing that it is the correct form as the above.

7 Joint Distributions, Multivariate RVs

One random variable is often not enough to solve a problem or model a situation. For example, we may deal with two coins rather than one; less trivially, we may be interested in differences between two Normal or Exponential distributions. Consequently, we add another random variable to the mix. So far in the course, we have dealt with situations where these random variables were independent, uncorrelated, or otherwise simplifiable (recall the minimum of n independent Expo RVs, or the sum of two independent Normals).

However, these are often "toy" situations that help for simplifying calculations in certain cases, but do not work more generally. In section, I hinted at a more general method of dealing with multiple random variables and their expectations, which we now embark upon: namely, *joint distributions.*

When we have two or more random variables, we can no longer consider them separately in the general case; that is, probabilities such as $P(X < x, Y > y)$ cannot be broken down into terms dealing only with X and Y separately. Instead, we have to consider sums/integrals over the distribution of *both* X and Y at the *same time*, which makes computations potentially much trickier. However, the conceptual framework we've developed thus far will mostly stay the same!

Joint Distributions. The following three figures pretty much sum up everything we need to understand conceptually about joint distributions: what they are, and how they relate to each marginal and conditional distribution. We also have the following ways to derive marginal and conditional PDFs from joint PDFs:

$$f_X(x) = \int f_{X,Y}(x,y)dy$$

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

where the integral is understood to be taken on the entire support of Y. Moreover, we have the following important fact about joint PDF/PMFs:

$$\boxed{\text{If the joint PDF or PMF can be factored as the product of two PDFs or PMFs, each only}}$$
$$\boxed{\text{dependent on one of the RVs, then the RVs are } \textit{independent.}}$$

Splitting Joints by Conditionals. Once we know the joint distribution (PMF or PDF), we theoretically know everything: we can, in theory, compute any expectation, probability, marginal probability, etc. that we wish. Of course, the computation may not be analytically feasible, but that's why WolframAlpha exists (as well as numerical integration tools).

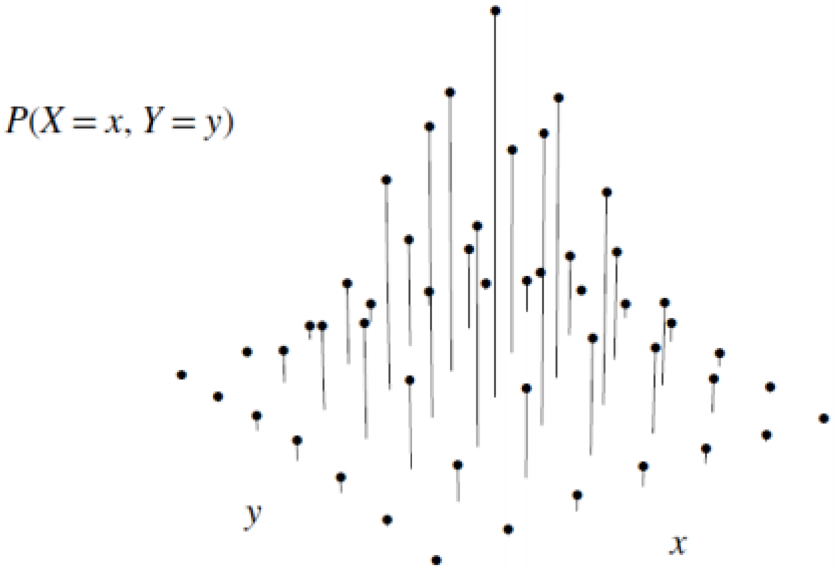
The hard part, then, is deriving the joint distribution. Except in certain special cases, such as the Multinomial and Multivariate Normal distributions to be covered later, it's actually difficult if not impossible to simply "write down" the joint distribution. For example, if $X \sim \text{Pois}(\lambda)$ and $Y \sim \text{Geom}(p)$ without independence (i.e. some correlation ρ), what would be the joint PMF of (X, Y)?

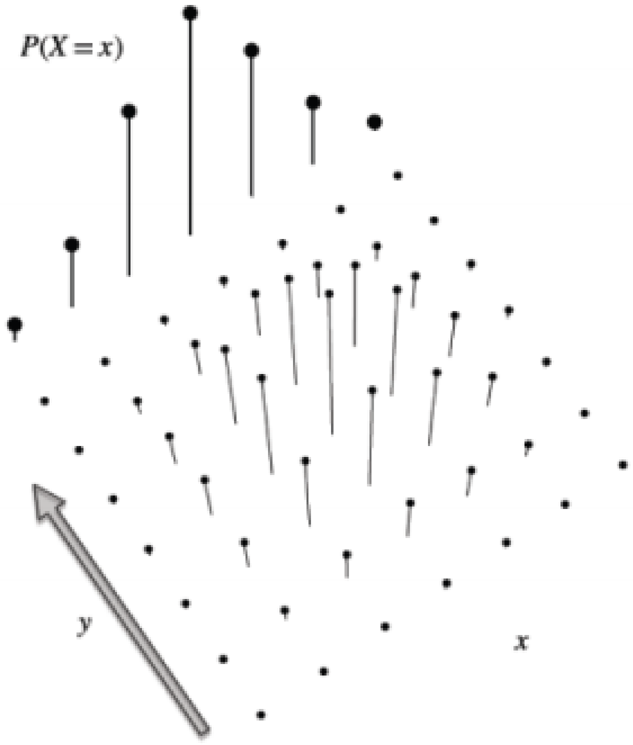
Thus, we try to simplify the situation by splitting the joint PMF/PDF into a sequence of *conditional PMFs.* The quintessential example is the *chicken-and-egg problem*, covered in lecture, which went as follows: a chicken lays a random number of eggs, $N \sim \text{Pois}(\lambda)$, and they hatch with probability p independently. If X is the number that hatch and Y is the number that do not, such that $X + Y = N$, then:

$$P(X = x, Y = y) = P(X = x, N = x + y) = P(X = x|N = x + y) \cdot P(N = x + y)$$

In a similar way, suppose we have X and Y such that we know, conditional on $Y = y$, what the distribution of X is; that is, we know the distribution of $X|Y = y$. In this case, we can write:

$$P(X = x, Y = y) = P(X = x|Y = y)P(Y = y)$$

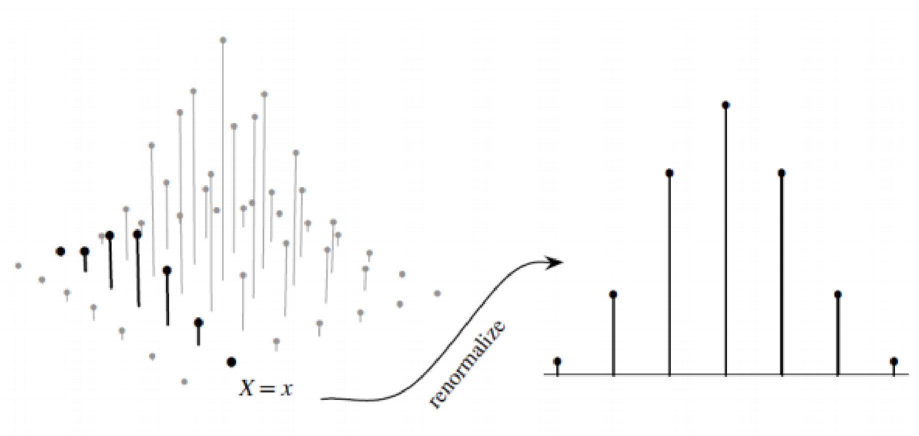




both terms that we know. For continuous distributions, this would translate to:

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

Multidimensional LotUS. Everything we know about unconscious statisticians follows through in the multivariate case. In particular, it is correct that:



$$E[g(X,Y)] = \sum_x \sum_y g(x,y) p_{X,Y}(X=x, Y=y) \text{ or } \int\int g(x,y) f_{X,Y}(x,y) dxdy$$

However, be careful that you do *both* sums/integrals! When we have a problem with n random variables and we would like to find the expectation, we need to do n sums or integrals. The most often encountered cases that 2-D LotUS comes into play are functions $g(X,Y)$ such as:

- **Differences.** $g(X,Y) = X - Y$
- **Sums.** $g(X,Y) = X + Y$
- **Products.** $g(X,Y) = XY$
- **Exponents.** $g(X,Y) = e^{X+Y} = e^X e^Y$, which shows up in MGF calculations for multiple RVs.

(*Exercise.* Prove in the general case that linearity of expectation holds: that is, when $g(X,Y) = X + Y$, show that $E[g(X,Y)] = E[X] + E[Y]$ by using the 2-D LotUS formula.)

Covariance and Correlation. As noted above, generally we cannot consider any special relationship like independence or uncorrelatedness between the RVs under consideration. Thus, we'd like to see how, in fact, the RVs are related: are they highly correlated, or roughly independent, or very negatively correlated? To do so, we define the *covariance*:

$$\text{Cov}(X,Y) = E[(X - E[X])(Y - E[Y])] = E[XY] - E[X]E[Y]$$

Note that the first expectation is the "definition" of covariance, but as with the variance, we almost always use the second formula for calculations. In addition, note that this is very similar to the variance formula, in fact:

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

but for arbitrary X, Y, since we no longer have squared terms, there is no guarantee that $\text{Cov}(X,Y) \geq 0$. Moreover, while the covariance can give us the *sign* of the relationship (that is, if X is high, is Y high or low?), but says very little about the *magnitude* of the relationship. This is because we don't know what *units* the X, Y are being measured in; if we scale up X by 100, then covariance goes up by a factor of 100 as well! (This is a deja vu with the problem we had with variance, and why we use standard deviation instead.)

Thus, we introduce the *correlation*:

$$\mathrm{Corr}(X,Y) = \rho_{X,Y} = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y}$$

where $\sigma_X = \sqrt{\mathrm{Var}(X)}$ and $\sigma_Y = \sqrt{\mathrm{Var}(Y)}$. This standardizes the correlation such that $|\rho_{X,Y}| \leq 1$. Values of $\rho_{X,Y}$ near ± 1 imply that the variables X, Y are very highly correlated (either positive or negatively). In fact, if $|\rho_{X,Y}| = 1$, then we can write $Y = aX + b$ in a linear relationship!

Finally, we recall an important property for calculating variances:

$$\mathrm{Var}(X_1 + \cdots + X_n) = \sum_{i=1}^{n} \mathrm{Var}(X_i) + 2 \sum_{i<j} \mathrm{Cov}(X_i, X_j)$$

(*Exercise.* Prove that $|\rho_{X,Y}| \leq 1$ for any arbitrary choice of X and Y.)

Practice Problems

Example 33. Let's Prove It! Round Three. In this problem, we will prove some interesting facts using the general joint distribution framework, in hopes of becoming more comfortable with joint PDFs and PMFs.

a) Prove if $g(X,Y) = X + Y$, then $E[g(X,Y)] = E[X] + E[Y]$; that is, prove linearity of expectation in the general case.

b) Prove that the correlation coefficient, $\rho_{X,Y}$, is always between ± 1.

c) Prove that if $\rho_{X,Y} = \pm 1$, then we can express Y as a linear function of X; that is, $Y = aX + b$. Find a, b.

Example 34. Thinking in 3-D. Consider choosing a random point in the unit sphere:

$$\{(x,y,z) : x^2 + y^2 + z^2 \leq 1\}$$

. We can represent this "random choice" as three random variables, each representing a coordinate: (X, Y, Z).

a) Find the joint PDF of X, Y, Z.

b) Find the marginal distribution of X.

c) Find the probability that (X, Y, Z) lies in a smaller sphere of radius 0.5.

Example 35. Unconscious in 2-D. Suppose that $X \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$, and that the two RVs are independent.

a) Find the MGF of $X + Y$ using only 1-D LotUS by using a property of two independent Normal distributions.

b) Find the MGF of $X + Y$ using the joint distribution of X, Y and 2-D LotUS. Confirm that your answer matches what was derived in part (a).

Example 36. Whims of Fate ... and Chinese Food. This is it: today is the day you will meet *the one.* Fortunately, some divine being has tipped you off to this cosmic event, so you know that he/she/other will arrive at some point between 5 - 6 pm. As per your usual tradition, you sit alone at a table at Yenching, waiting for the one to arrive. Said divine being has also alerted you to the fact that because your soulmate has opted to take a bus that arrive at 5 pm, his/her/its arrival time follows an *Kumaraswamy distribution* (that is, $\mathrm{Beta}(1,3)$) between 5 and 6 pm.

There is one twist, though. Your soulmate will promptly leave and you will never ever see him/her/it again unless you are at your most charming. Your charm levels also happen to spike between 5 - 6 pm, but its arrival in that hour follows a $\mathrm{Beta}(2,2)$ distribution. Assume that for 10 minutes after your charm levels spike, you are charming enough to hold your soulmate's attention.

a) What is the probability that you meet your soulmate and are able to hold his/her/its attention, living happily ever after?

b) Suppose now that rather than taking public transportation, your soulmate has decided to drive to Cambridge by car. Thus, his/her/its arrival time now follows Unif$[0, 1]$ between 5 and 6 pm. What is the probability of living happily ever after?

8 Poisson Splitting, MVN, Multinomial

Poisson Splitting. The idea of *Poisson splitting* is often linked to the *Poisson process* that we learned earlier, regarding the fact that if the occurrence of events are distributed in a Pois manner, then the time between events is distributed Expo. It's extremely confusing unless you know exactly what you're looking for, because these two terms are used interchangeably, together, incorrectly, and generally in all possible permutations, although there is (at least in my mind) a very clear distinction. This is how it conceptually originated:

Suppose that we have two independent "counting" processes, i.e. processes where events happen over time, and we are counting the number that occur within a given time t. Suppose these are $N_1(t) \sim \text{Pois}(\lambda_1 t)$ and $N_2(t) \sim \text{Pois}(\lambda_2 t)$ Then, $N(t) = N_1(t) + N_2(t)$ is also Poisson with $\lambda = \lambda_1 + \lambda_2$.

But the point is that we knew all this; it's not specific to the *Poisson process*, but rather just a generic property of the Poisson distribution: if you have two random variables, independently distributed Poisson, then their sum is Poisson with the sum of the rate parameters.

Similarly, the idea of *Poisson splitting* is for all Poisson distributed random variables, not specific to the idea of the Poisson process (meaning we *don't* have to consider events occurring across time only). It goes as follows:

Suppose $N \sim \text{Pois}(\lambda)$. Imagine radioactive particles being emitted, calls coming into a call center, or chickens laying eggs. Then, N is *split* into two types, 1 and 2, represented by X and Y respectively. This can represent radioactive particles being detected or not; calls being received or not; or eggs hatching or not. Suppose that for each event that occurs, whether it is of type 1 or type 2 is distributed Bernoulli; that is, $I_i \sim \text{Bern}(p)$ where $I_i = 1$ if type 1 and $I_i = 0$ if type 2. Then, we have the following result:

1) $X \perp Y$, that is, X and Y are independent.

2) $X \sim \text{Pois}(\lambda p)$ and $Y \sim \text{Pois}(\lambda(1-p))$

Now that we've learned about the *Multinomial distribution* (next section), we can consider the natural generalization of Poisson splitting into n categories, and it works! Suppose that we have n types, and $X_1, \ldots, X_k$ are the counts of events in each type, with each type having probability p_i of occurrence. That is, $X_1, \ldots, X_k \sim \text{Mult}_k(p_1, \ldots, p_k)$. Then, we have:

1) $X_1, \ldots, X_k$ are independent.

2) $X_1 \sim \text{Pois}(\lambda p_1), \ldots, X_k \sim \text{Pois}(\lambda p_k)$.

Multinomial, Binomial, and Hypergeometric. The *Multinomial distribution* is one of the most useful, often used, yet confusing distributions. There seems to be a clear intuition behind the distribution: we have some k categories, each with probability p_i of occurrence, and at each of the n trials, we throw a ball into one of the k categories.

There are important conditions that must hold for the Multinomial distribution to work:

a) All of the n trials are independent.

b) The probabilities p_i stay constant throughout all n trials.

c) The number of categories k stay constant throughout all n trials.

Some examples of when such a distribution might be useful are when you want to model people making one of k choices, especially in survey data when responses fall into one of the k categories, or when we sample with replacement from a population falling into k types (i.e. ethnicities).

Important properties of the Multinomial include (supposing that $X_1, \ldots, X_k \sim \text{Mult}_k(p_1, \ldots, p_k)$:

a) Every $X_i \sim \text{Bin}(p_i)$. (You can consider this as a “one-versus-all” format.)

b) *Lumping*: Any $X_i+X_j \sim \text{Bin}(p_i+p_j)$. Moreover, the $k-1$ random vector given by $X_1, ;X_i+X_j, \dots, X_k \sim \text{Mult}_{k-1}(p_1, \dots, p_i+p_j, \dots, p_k)$.

That seems well enough - but one can go seriously awry when one of the assumptions listed above do not hold. For example, if we sample without replacement from a population, we can never use the Multinomial since the probabilities are changing. This is akin to the situation with the Binomial and Hypergeometric (the “sampling distributions”): when we sample with replacement, we use Binomial, whereas if it is without replacement, then we use Hypergeometric. While there does exist a “Multivariate Hypergeometric” distribution, it is not commonly used in practice. The way to go would be to model each X_i separately as a Hypergeometric in such a case.

(*Exercise.* What is the sample space (“pebble world” representation of the Multinomial distribution? What are the equally-likely pebbles? How does the random variable map from this pebble to the probability?)

Multivariate Normal. Now we finally arrive at the mother of all multivariate distributions, the *Multivariate Normal distribution.* As you may have noticed, unlike in the univariate case, we don't really have many multivariate distributions; the Multivariate Normal and Multinomial just about sum it up. Of the two, the Multivariate Normal is far and away the more significant and widely used - in fact, pretty much every sophisticated application you see involves the Multivariate Normal.

Without further ado, here are equivalent definitions of the Multivariate Normal:

a) *Axiomatic.* $X_1, \dots, X_k \sim \mathcal{N}(\mu, \Sigma)$ if every linear combination $a_1X_1 + \cdots + a_kX_k \sim \mathcal{N}$.

b) *Constructive.* $X = X_1, \dots, X_k \sim \mathcal{N}(\mu, \Sigma)$ if $X = WZ + \mu$, where $Z = (Z_1, \dots, Z_k)$ is a random k-vector of independent standard Normals, W is a matrix, and mu is the mean of X.

There are two particularly powerful and magical properties of the Multivariate Normal.

a) *Linear combinations.* Any linear combination of the X_i is itself Normal (equivalent to the first definition); in particular, all the $X_i \sim \mathcal{N}(\mu_i, \Sigma_{ii})$ marginally.

b) *Uncorrelated implies independence.* In general, uncorrelated does not imply independence (though vice-versa works). However, for a Multivariate Normal, if $\text{Cov}(X_i, X_j) = 0$, that is X_i, X_j are uncorrelated, then they are automatically independent.

The PDF of a Multivariate Normal is given as:

$$f_X(x) = \frac{1}{\sqrt{(2\pi)^k|\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)$$

Practice Problems

Example 37. Poisson Splitting Your Way through Shopping. Suppose that you shop $N \sim \text{Pois}(10)$ courses during shopping period. You only care about Statistics and Folklore & Mythology courses; of the total number of courses in both departments combined, 0.7 are from F&M while 0.3 are from Stat. Moreover, of these courses, 0.9 are bad and 0.1 are good. You've shopped 4 bad courses, but you forgot the number of good courses you shopped. What is the PMF, conditional on having shopped 4 bad courses, of the number of good courses you shopped?

Example 38. Morbid Chicken-and-Egg. Suppose two clownfish, Marlin and Coral, lay a $\text{Pois}(\lambda)$ number of eggs, denoted N. Unfortunately, the Great Barrier Reef has been known for its barracuda attacks, and the probability of any given egg surviving a barracuda attack is p.

a) What is the distribution of the number of eggs that survive and potentially star in a Disney movie (call it X), given N?

b) What is the correlation between X and N?

Example 39. Lumping in with the Higgs. On July 4, 2012, the state of humanity's understanding of the universe took a large step forward with the discovery of the Higgs boson. Put simply, the way to discover the Higgs boson is to collide particles at high enough energies so that the boson is produced. Unfortunately, it decays extremely quickly into a bunch of elementary particles. For simplicity, let's suppose that a Higgs boson decays into: 1) fermion pair; 2) W boson pair; 3) Z boson pair. The events occur with probabilities $p, q, 1-p-q$.

a) What is the PMF of F, W, Z, signifying the number of times each of the decay events occur, given that we produce n Higgs bosons?

b) What is the distribution of the number of fermion pair decays?

c) Find the maximum likelihood estimate of p, q given observations $F = f, W = w, Z = z$.

d) Unfortunately, our detectors are not sensitive enough to distinguish between fermion pair and W boson pair decays; that is, we know $F + W$ but not each separately. Find the MLE of the relevant parameter given $F + W = B = b, Z = z$.

9 Variable Transformations, Gamma, and Sampling

Variable Transformations, Redux. Let's recall, from a few sections past, our formula for determining the distribution of a transformed variable. Suppose X has PDF f_X, and $Y = g(X)$. Then, we saw that:

$$\boxed{f_Y(y) = f_X(g^{-1}(y)) \cdot \left|\frac{dx}{dy}\right|}$$

That's all there is to it! (Though, also recall that there could be some tricky problems when the function $g(\cdot)$ was not 1-1; for example, we had problems with $g(x) = x^2$.) It turns out that there is a natural generalization to the multivariate case, although things become more complicated because we don't know what a "derivative" means in multiple dimensions. We define such a derivative as the *Jacobian*:

$$\frac{d\mathbf{x}}{d\mathbf{y}} = \begin{pmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} & \cdots & \frac{\partial x_1}{\partial y_n} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} & \cdots & \frac{\partial x_2}{\partial y_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial x_n}{\partial y_1} & \frac{\partial x_n}{\partial y_2} & \cdots & \frac{\partial x_n}{\partial y_n} \end{pmatrix}$$

This is the natural generalization of the idea of a *gradient*, which is simply the Jacobian for a function from $\mathbb{R}^n \to \mathbb{R}$. Thus, rather than having n rows, it only has 1 - that is, it's a vector. Unfortunately, we can't use the Jacobian by itself because multiplying a PDF by an $n \times n$ matrix doesn't quite make sense. So we have to use the determinant, $\left|\frac{d\mathbf{x}}{d\mathbf{y}}\right|$ instead, and we finally have the equivalent formula for multivariate distributions:

$$\boxed{f_{\mathbf{Y}}(\mathbf{y}) = f_{\mathbf{X}}(\mathbf{x}) \cdot \left|\frac{d\mathbf{x}}{d\mathbf{y}}\right|}$$

One common application of this is 2-by-2 transformations (note that it's generally very complicated for anything ≥ 3 dimensions). In particular, consider:

$$Y_1 = g(X_1, X_2) = X_1 \cdot X_2$$

In this case, how do we deal with the variable transformation since we have a transformation from $\mathbb{R}^2 \to \mathbb{R}$? Our Jacobian matrix is no longer square, and we can't quite take a determinant of a non-square matrix. As a result, we have a very common method that will most likely come in handy: *just let $Y_2 = X_2$!* That is:

$$Y_1 = X_1 \cdot X_2, Y_2 = X_2 \Rightarrow X_1 = \frac{Y_1}{Y_2}, X_2 = Y_2$$

Then, our Jacobian is simply:

$$\frac{d\mathbf{x}}{d\mathbf{y}} = \begin{pmatrix} \frac{1}{y_2} & -\frac{y_1}{y_2^2} \\ 0 & 1 \end{pmatrix}$$

and so our Jacobian determinant is just $\left|\frac{d\mathbf{x}}{d\mathbf{y}}\right| = \frac{1}{y_2}$, and our transformed density is given by:

$$f_{Y_1,Y_2}(y_1, y_2) = f_{X_1,X_2}(y_1/y_2, y_2) \cdot \frac{1}{y_2}$$

Now, our originally desired density was for $Y_1 = g(X_1, X_2)$, which we can find by our usual method of "integrating out" Y_2. That is:

$$f_{Y_1}(y_1) = \int f_{X_1,X_2}(y_1/y_2, y_2)\frac{1}{y_2}dy_2$$

(*Exercise.* Using the method outlined above, find the PDF of $g(X_1, X_2) = \frac{X_1}{X_2}$ where $X_1 \sim \text{Expo}(\lambda)$ and $X_2 \sim \Gamma(n, 1/\lambda)$.)

Gamma Distribution. We've hinted at the *Gamma distribution* repeatedly in lecture and section thus far, so it's finally time to define exactly what this definition is!

We say that: $X \sim \text{Gamma}(\alpha, \beta)$ if it has the following PDF:

$$\boxed{f_X(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}}$$

(Note that there are other *parameterizations*, or possible ways to write the PDF using, for example, $\beta = 1/\theta$.)

Recall from previous sections that the Gamma is intimately related to a number of other important continuous distribution:

a) *Exponential*: $\boxed{\text{Expo}(\lambda) = \text{Gamma}(1, \lambda)}$

Moreover, if we have $X_1, \ldots, X_n \sim \text{Expo}(\lambda)$ i.i.d., then: $X_1 + \cdots + X_n \sim \text{Gamma}(n, \lambda)$.

b) *Chi-square*: $\boxed{\chi^2_n = \text{Gamma}(n/2, 2)}$ (That is, the Chi-square distribution with n degrees of freedom is simply the Gamma with parameters $n/2, 2$.)

c) *Beta*: By the Bank-Post Office relation, we have that if $X \sim \text{Gamma}(\alpha, \theta)$ and $Y \sim \text{Gamma}(\beta, \theta)$, then: $\boxed{\frac{X}{X+Y} \sim \text{Beta}(\alpha, \beta)}$

More generally, the Gamma happens to be one of the most frequently used distributions in not just probability theory, but statistics as a whole (especially in Bayesian statistics), so it will keep coming up again and again!

The story of the Gamma distribution is the same as the *Negative Binomial* was for the *Geometric*: that is, the Exponential distribution is the waiting time until the first success (or occurrence), whereas the Gamma counts until α occurrences (in which α does not need to be an integer!).

Sampling, Confidence Intervals, and Stat 110 + 1. This week's problem set takes us deeper into the world of sampling distributions and confidence intervals, part of the domain we call *statistical inference*, and which will be the main topic of Stat 111. Here, we give a brief overview of what these concepts mean, and how to do computations such as deriving sampling distributions and forming confidence intervals.

Sampling distributions are the distributions of *statistics*, or functions of the data. We suppose that the data are *random samples*, or i.i.d. random variables $X_1, \ldots, X_n$ drawn independently from some population density f_X. A statistic is a function $g(X_1, \ldots, X_n)$. Some common examples are:

- *Sample mean*: This is probably the most common and most well-known statistic:

$$g(X_1, \ldots, X_n) = \frac{X_1 + \cdots + X_n}{n}$$

It is often used in conjunction with Normal distributions or others in which the average is an important quantity, such as the Binomial.

- *Sample count*: This is usually used for counting distributions where the sums of the random variables are well-known, such as the Poisson or Exponential:

$$g(X_1, \ldots, X_n) = X_1 + \cdots + X_n$$

- *Sample max/min*: One final example is the maximum or minimum or a set of data points, given by:

$$g(X_1, \ldots, X_n) = \max(X_1, \ldots, X_n)$$
$$g(X_1, \ldots, X_n) = \min(X_1, \ldots, X_n)$$

Now, although we know the distribution of the individual X_i, it is sometimes not trivial to determine the distribution of these statistics, which are called *sampling distributions*.

However, there are special cases in which computing the sampling distribution, or the distribution of the statistic, is relatively easy. For example, we know that the sums of independent Poissons are Poisson themselves, and the sums of independent Normals are also Normal.

Confidence intervals are a relatively difficult concept to understand, and I'm personally not certain that I fully understand it myself. However, the basic idea and calculations, though slightly non-intuitive, are fairly approachable, and we illustrate it using the general framework and tying in sampling distributions.

Suppose that we observe some data $X_1, \ldots, X_n$. We compute the *statistic* (such as a sample mean), and find the distribution of the statistic, or *sampling distribution*. Call the statistic S, and suppose it follows some CDF F_S. Meanwhile, the distribution of the statistic will depend on the distribution of the data (such as Expo or Normal), which in turn depend on some parameter θ (i.e. λ for the Expo and μ, σ^2 for the Normal).

Unfortunately, we don't know this parameter θ; that's why we're collecting the data - to try to conduct *inference* on what the most likely value of this parameter is, given the data we have observed. In other words, what values of θ make it plausible for us to observe the data $X_1, \ldots, X_n$? We can determine this by looking at the "middle 95%" of the sampling distribution. If the observed statistic, which is a function of the observed data, falls within this middle 95%, then that value of the parameter θ is "fine" in some sense. However, if the observed statistic falls outside the middle 95% of the distribution given some θ, then we have *statistically significant* evidence against that θ at the 5% significance level.

Thus, in order to determine the 95% *confidence interval* for θ, we consider all the values of θ that yield sampling distributions whose middle 95% contain our observed statistic. Typically, that means making sure that the observed statistic is not in the lowest 2.5% or the highest 2.5% of the sampling distribution given θ.

This is best illustrated through an example, so please refer to the **Practice Problems** section for this week to get a better sense of what's going on!

Practice Problems

Example 40. Transforming Variables Left and Right. To try our hand at multivariate variable transformations, let's do a few examples!

a) Suppose that $X_1 \sim \text{Expo}(\lambda)$ and $X_2 \sim \text{Gamma}(n, 1/\lambda)$. Derive the PDF of $\frac{X_1}{X_2}$ and show that it is the PDF of the *Pareto distribution*, which has PDF:

$$\frac{\alpha x_m^\alpha}{x^{\alpha+1}}$$

where x_m is the minimum value of the support of the Pareto.

b) Suppose that X and Y are independently standard Normal, and we are considering their polar coordinates R and θ. Find the distribution of θ, and explain what this means about the bivariate Normal.

Example 41. Gamma Properties and the Bank-Post Office. For this problem, suppose that $X \sim \text{Gamma}(a, \lambda)$ and $Y \sim \text{Gamma}(b, \lambda)$, with X, Y independent.

a) Show that $X + Y \sim \text{Gamma}(a + b, \lambda)$ in three ways: 1) using a convolution; 2) using MGFs; and 3) using a story proof.
(*Hint*: Consider the Poisson process and how the Gamma distribution relates.)

b) Suppose that X represents the amount of time that Ryan takes to prepare his section materials, and Y represents the amount of time that Ryan takes to teach section and office hours every week (note: $a \gg \lambda$ for this case!). To see how unbalanced a life he is leading, Ryan is interested in looking at the ratio X/Y; to report his hours for his paycheck, he is also interested in looking at the total $X + Y$. Are these two quantities independent of each other? Give a proof in either case.

Example 42. Sampling Distributions and Confidence Intervals. Let's work with the sampling distribution of Exponential random variables and derive some confidence intervals. For this problem, suppose that $X_1, \ldots, X_n \sim \text{Expo}(\lambda)$ i.i.d., where each represents an observation.

a) Find the sampling distribution of the sample total, $T = \sum_{i=1}^{n} X_i$.

b) Find the sampling distribution of the sample mean, $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$. Show how to do a simple transformation to convert it to a distribution that is free of the scale parameter.

c) Suppose that we observe a sample mean value of $\bar{x} = 13$ for 10 observations. Using the results derived above, find the 95% confidence interval for the value of the scale parameter λ.

10 Conditional Moments, Inequalities

Conditional Expectation. The idea of conditional expectation is actually quite simple, in formal terms. When we do a normal expectation, we basically take an integral:

$$E[X] = \int x f_X(x) dx$$

When we a conditional expectation, all we do is *update the probability* so that we consider probabilities given that some other random variable/event is some value:

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

Thus, you see that the only thing that has changed is the probability density, which went from a marginal density to a conditional density given the value of y.

As we saw earlier in the course and in section, computing conditional expectations is particularly useful when we want to break up a more complicated expectation (think back to why we used conditional probabilities - first-step analysis - and indicator random variables!). To do so, we employ *Adam's Law*, which states that:

$$E[X] = E[E[X|Y]]$$

This law goes by a number of different names, such as *law of iterated expectation* or *tower property.* What exactly does an expectation inside an expectation mean? Basically:

$$E[X] = \int E[X|Y = y] f_Y(y) dy$$

In other words, we compute the expectation of X when $Y = y$, and then do a weighted average (integral) over all possible values of Y.

Conditional Variance. Conditional variance is very similar to conditional expectation, for obvious reasons. Again, we do the same calculation as in a normal (marginal) variance, except use a PDF/PMF that is updated given the random variable takes a certain value or an event occurs. Recall that:

$$\text{Var}(X) = \int (x - \mu)^2 f_X(x) dx$$

Thus, taking our cue from the conditional expectation formula, we can simply write:

$$\text{Var}(X|Y = y) = \int (x - \mu)^2 f_{X|Y}(x|y) dx$$

with the integral appropriately changed to a sum in the case of a discrete RV. In particular, note that our simplified formula for the variance also holds in the case of the conditional variance:

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

And, just as in the case of the conditional expectation, we really care about the conditional variance to compute the variance of some random variable X that depends on the value of another random variable Y (i.e. consider the number of hatched eggs X, which is easy to compute once we know the total number of eggs N). To calculate this, we use the analogue of Adam's Law, which is called *Eve's Law*:

$$\text{Var}(X) = \text{Var}(E[X|Y]) + E[\text{Var}(X|Y)]$$

Hierarchical Models. This is probably one of the most useful concepts to come out of Bayesian analysis, and so I wanted to take a moment to discuss this in light of our ideas on conditional expectation/variance. *Hierarchical modeling* is a really enlightening idea, and it goes like this. Suppose that we observe some random variable Y_i: we can imagine it's something like the height, weight, IQ, ... of an individual in a given population. We believe that the individual has some *true* height, weight, IQ, ... that we will call μ_i. Unfortunately, our measuring devices (and life's circumstances, i.e., you curled up to sleep last night so your spinal cord is not as stretched out as usual) render our measurements inaccurate, so there is some distribution of our measurement Y_i that is "centered at" the true value μ_i.

Note that *if we knew* μ_i, then life would be easy. Let us denote $F_{Y|\mu}$ as the CDF representing the distribution of Y_i, given that we know the true value μ_i. In other words, this distribution represents our *measurement uncertainty.* Now, we don't know μ_i, so we need some distribution over the population regarding these μ as well. Let's call this distribution CDF F_μ. So in summary we have:

$$Y|\mu \sim F_{Y|\mu}$$
$$\mu \sim F_\mu$$

A common and widely-used example is $F_{Y|\mu} = \mathcal{N}(\mu, \sigma_Y^2)$ and $F_\mu = \mathcal{N}(\theta, \sigma_\mu^2)$ where both the σ_Y, σ_μ and θ are known quantities.

In any case, we can actually derive the expectation of Y, despite the fact that we don't know the marginal distribution! In other words, we simply use Adam's Law:

$$E[Y] = E[E[Y|\mu]] = \int \left[\int y f_{Y|\mu}(y|\mu) dy \right] f_\mu(\mu) d\mu$$

(This looks complicated, but take a moment to decipher the integral and you'll figure it out!) We can do the exact same thing using Eve's Law to find the conditional variance!

Inequalities. Though they may seem boring, inequalities are actually the bread-and-butter of more mathematical probability theory. Important ones include:

- *Cauchy-Schwarz*: For any random variables X, Y with finite variance:

  $$|E[XY]| \leq \sqrt{E[X^2]E[Y^2]}$$

  (Recall that we used Cauchy-Schwarz to prove that the correlation must be between -1 and 1! *Exercise.* Go back to the definition of correlation, and use Cauchy-Schwarz to prove that it has magnitude ≤ 1.)

- *Jensen*: If X is a random variable and:

  a) g is a *convex* function: $E[g(X)] \geq g(E[X])$

  b) g is a *concave* function: $E[g(X)] \leq g(E[X])$

- *Markov*: $P(|X| \geq a) \leq \frac{E|X|}{a}$

- *Chebyshev*: $P(|X - \mu| \geq a) \leq \frac{\sigma^2}{a^2}$

Practice Problems

Example 43. Let's Prove It! Round 3. In this problem, we will prove various properties of the conditional expectation, variance, and Adam's/Eve's Laws so that you get a better feel for how they can be used.

a) Show that the variance decomposition formula holds for the conditional variance; that is:

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

b) Prove Eve's Law.

c) Let X, Y be random variables, and $W = Y - E[Y|X]$ be the residual of Y given a predicted value $E[Y|X]$. Show that both $E[W]$ and $E[W|X]$ are exactly 0. (That is, $E[Y|X]$ is an unbiased predictor for Y.)

Example 44. Normal-Normal Hierarchical Model. Let's suppose that we are measuring heights in a population. We believe that our measurement device gives the correct measurement, but with some measurement error, so that our measured heights are distributed:

$$H_i \sim \mathcal{N}(\theta_i, 2^2)$$

where H_i is the measured height and θ_i is the true height. We don't know the true height of the person we are measuring (otherwise we wouldn't be measuring them!), and can represent this uncertainty by putting a distribution on the heights over the population as:

$$\theta_i \sim \mathcal{N}(65, 10^2)$$

where height is measured in inches.

a) You are waiting in the measuring room, and the next person has yet to enter. What is your expectation of the measured height of that person, and what is the variance?

b) What is the distribution of H_i, given that you don't know θ_i?

c) What is the correlation between the measured heights and the true heights?

Example 45. Beta-Binomial Hierarchical Model. Suppose that you are flipping a coin. Unfortunately, you don't know how biased the coin is (all real coins are biased, after all!); in other words, you don't know p, the probability of heads, for the coin. You model this uncertainty regarding p by setting:

$$p \sim \text{Beta}(2, 2)$$

which is centered at 1/2 but with some uncertainty. Suppose you flip the coin 100 times and count the number of heads, that is:

$$Y|p \sim \text{Bin}(100, p)$$

a) What is the expected number of heads in the 100 tosses? What is the variance in the number of heads?

b) Suppose that you observe 70 heads among the 100 tosses. What is your updated distribution for p given the data? (Hint: use Bayes' rule.)

Example 46. When Are Inequalities Useful? Let us go over some reasons why inequalities can be useful, illustrate their properties, and get some practice applying them.

a) Visually illustrate why Jensen's inequality holds for a generic convex function and a discrete RV taking two values.

b) Practice makes perfect!

   i) $E[\log(X)]$ vs. $\log(E[X])$
   ii) $E[|X|]$ vs. $\sqrt{E[X^2]}$
   iii) $P(X > c)$ vs. $\frac{E[X^3]}{c^3}$
   iv) $E[\min(X, Y)]$ vs. $\min(E[X], E[Y])$

c) Toss a fair coin 100 times. Use Chebyshev's inequality to find the probability that the number of heads is between 40 and 60. Then, use a Poisson approximation to derive another estimate.

11 Central Limit Theorem, LLN, and Markov Chains

Central Limit Theorem (CLT). This is the godfather of all limit theorems, and is invoked in almost every topic that involves probability. It serves as the basis for the infamous "bell curve" - we'll come back to this shortly. What does the CLT state? Namely, that:

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} \mathcal{N}(0,1)$$

as long as the two moments, μ, σ^2, are finite. In practical terms, what we care about is the following:

$$\bar{X}_n \approx \mathcal{N}(\mu, \sigma^2/n)$$

That is, the sample mean is distributed approximately Normal, around the true mean μ with variance σ^2/n, as long as n is large.

Note that $\bar{X}_n$ is essentially a large sum of i.i.d. random variables. Suppose that we have some sum, $Y = X_1 + \cdots + X_n$, where all the X_i are independent RVs. In some nonrigorous sense, we can claim that Y then follows a Normal distribution (which we can make more rigorous with some finesse and using the CLT). This is why we observe the "bell curve" in the real world: suppose that Y is the height of an individual. The height is affected by a large number of X_i, such as genetics, diet, socioeconomic status, activities, and so forth; thus, since Y is the sum of a large number of RVs, it will be approximately Normal!

In general, the following distributions are often approximated by the Normal because of their reasonably bell-shaped nature and computational cost:

- *Binomial*: The Normal approximation to the Binomial is probably the most common of all the CLT applications in approximating distributions. If $X \sim \text{Bin}(n,p)$, then we have:

$$X \approx \mathcal{N}(np, np(1-p))$$

- *Poisson*: The curious thing about the Poisson is that both its mean and variance are equal to λ, its rate parameter. Thus, if $X \sim \text{Pois}(\lambda)$, then:

$$X \approx \mathcal{N}(\lambda, \lambda)$$

- *Negative Binomial*: Less common, but useful, is the Normal approximation to the Negative Binomial; if $X \sim \text{NBin}(r,p)$, then:

$$X \approx \mathcal{N}(r/p, r(1-p)/p^2)$$

- *Gamma*: Finally, the Gamma distribution is sometimes approximated by the Normal (and this therefore includes the Exponential and Chi-squared distributions); if $X \sim \text{Gamma}(\alpha, \beta)$, then:

$$X \approx \mathcal{N}(\alpha/\beta, \alpha/\beta^2)$$

Note that many of the distributions named are *discrete*, whereas the Normal is *continuous*. Consequently, it is often advisable to make the *continuity correction* to improve our approximation by correcting for the fact that we are approximating a discrete distribution by a continuous one. That is, if $X \approx Z \sim \mathcal{N}(\mu, \sigma^2/n)$, then:

$$P(X = x) \approx P(x - 1/2 \leq Z \leq x + 1/2)$$

Supposing that $Z \sim \mathcal{N}(0,1)$, then we simply have:

$$P(X = x) \approx \Phi(x + 1/2) - \Phi(x - 1/2)$$

which has rendered a potentially difficult calculation into one that is extremely straightforward!

Law of Large Numbers (LLN). When people talk about "the" Law of Large Numbers, there are actually two: the Strong and the Weak. (In mathematics, the idea of "strong" implies that the conclusions of the theorem are tighter and more significant than the "weak" version.) In the case of the LLN, the strong version tells us that the sample mean $\bar{X}_n$ will, in fact, converge to the expectation μ (with probability 1, but this is a technical point we won't worry about), whereas the weak version says that the sample mean converges in probability. (The distinction is that if you converge almost surely, as in the strong version, then after some N, *every single* sample mean will be within some tiny interval around the true mean μ, whereas the weak version says that "most of" the sample means will be in the tiny interval around μ).

What we care about, though, is not the distinction, but rather the application of the LLN to real-world problems. The main reason that the LLN is useful is because it *justifies taking averages.* In fact, we've grown so accustomed to the LLN that we mindlessly (or unconsciously?) take sample means and expect it to "work" in any experiment, measurement, and so forth, in giving us a reasonably accurate estimate. The only reason that such an assumption would hold, however, is because of the LLN! (In fact, without the strong version, it would be flawed to assume that taking a sample mean will always yield such results.)

Markov Chains. *"The future depends only on the past through the present."* This seemingly philosophical statement (which can probably be found in some fortune cookie somewhere) aptly describes the basic *Markov property* that distinguishes Markov chains for other kinds of stochastic processes, namely that:

$$P(X_{n+1} = x | X_n = x_n, \ldots, X_1 = x_1) = P(X_{n+1} = x | X_n = x_n)$$

Thus, we can completely characterize a homogeneous Markov chain (i.e. one where the probability above stays constant regardless of n) through its *transition matrix*:

$$q_{ij} \equiv P(X_{n+1} = j | X_n = i) \text{ and } Q = (q_{ij})$$

Note carefully that the *row index*, that is i, denotes the current state, whereas the *column index*, j, denotes the future state! As a result, summing up over all the columns in any given row yields 1: $\sum_j q_{ij} = 1$ for every i.

The reason that transition matrices are so useful is because they characterize not only the one-step change, but also the n-step change for any n; that is, given that you know the starting state k, you can figure out the probabilities of any state j at any time n by knowing the transition matrix. This is because:

$$q_{ij}^{(n)} \equiv P(X_n = j | X_0 = i) = (Q^n)_{ij}$$

that is, $q_{ij}^{(n)}$ is simply the i, j element of the matrix Q^n, which is the transition matrix taken to the n^{th} power.

What we are most concerned about in any Markov chain is its *stationary distribution.* The stationary distribution of a Markov chain is a probability distribution over its states $i = 1, \ldots, M$ such that:

$$\mathbf{s}Q = \mathbf{s}$$

In other words, if we start with some probability distribution over the states given by $\mathbf{s} = (s_1, \ldots, s_M)$, then no matter how long the Markov chain runs, our probability distribution over the states will *still be* $\mathbf{s}$! This is a profound statement, and the most important characterization of a Markov chain. Thus, we are interested in whether a Markov chain possesses a stationary distribution, whether said distribution is unique, and so forth. It turns out that while existence is quite general, other properties require us to specify Markov chains further.

This leads to some important definitions about Markov chains:

a) *Recurrent / transient*: State i of a Markov chain is *recurrent* if there is always a positive probability of returning to i; that is, for every state j that is reachable from i, it must be the case that i is reachable from j. Otherwise, the state is *transient.* (Intuition: you have probability 1 of returning to a recurrent state.)

b) *Reducible / irreducible*: A Markov chain is *irreducible* if every state can be reached by every other state; otherwise, it is *reducible.* (Intuition: the graphical model works best; if the entire graph is connected, it is irreducible, whereas if there are separate components, it is reducible.)

c) *Periodic / aperiodic*: The *period* of a state i is the greatest common divisor of all the path lengths leading back to itself; if the period is 1, then i is *aperiodic*, whereas if it is > 1, it is *periodic.* (Intuition: If there are cycles leading back to the state i, then it is generally periodic.)

Again, why are these definitions important, and why do we care? Namely, because they allow us to prove increasingly definite properties of Markov chains and their *stationary distributions.* We have that:

a) Any Markov chain with a *finite state space* has a stationary distribution (existence).

b) Any *irreducible* Markov chain has a *unique* stationary distribution.

c) Any *irreducible* and *aperiodic* Markov chain has the property that $P(X_n = i) \to s_i$ (that is, the probability of being in state i converges to the stationary distribution probability s_i) as $n \to \infty$.

One additional property that a Markov chain can possess that is of interest is called *reversibility.* A Markov chain is *reversible* with respect to some distribution $\pi = (\pi_1, \ldots, \pi_M)$ if:

$$\pi_i q_{ij} = \pi_j q_{ji}$$

Intuitively, this expression says that the probability of starting at state i multiplied by the probability of going from i to j is the same as the probability of starting at state j and going from j to i; which is exactly what the name "reversibility" implies. The key point about reversible Markov chains is that the distribution π is automatically a *stationary distribution*!

One example that frequently pops up is the undirected graph, in which we have a random walk over the vertices and the transition probabilities depend on the number of edges that are linked to the vertex. In this case, reversibility allows an easy proof of the fact that the stationary distribution is given by:

$$s = \left(\frac{d_1}{\sum d_i}, \ldots, \frac{d_M}{\sum d_i} \right)$$

Practice Problems

Example 47. Inequalities and Limit Theorems. Let $X_1, X_2, \ldots$ be i.i.d. positive r.v. with mean a, and $Y_1, Y_2, \ldots$ be i.i.d. positive r.v. with mean b. Assume the X_i's are also independent of the Y_j's.

a) Prove that

$$E\left(\frac{X_1 + \ldots + X_n}{Y_1 + \ldots + Y_n} \right) \geq \frac{a}{b}$$

b) Show that

$$\frac{X_1 + \ldots + X_n}{Y_1 + \ldots + Y_n} \longrightarrow \frac{a}{b}$$

with probability 1 as $n \to \infty$.

Example 48. I Can Has PageRank? This problem will develop both the theory and solution method of the basic PageRank algorithm used by Google. PageRank defines the importance of a page, r_i, as follows:

$$r_i \equiv \sum_{j \in I_i} \frac{r_j}{|O_j|}$$

where I_i is the set of pages j that have links going into page i, and O_j is the set of pages that j links to. Because of the recursiveness, we need to use an iterative algorithm to determine the values r_i.

Consider your typical Web surfer (i.e. a Stat 110 student who is procrastinating on his/her problem set). Suppose the person randomly clicks links on a page to other pages; this defines a transition matrix across all pages. The one problem arises when a page has no outlinks – that is, other pages link to it, but it doesn't link to any other pages. Now our surfer is stuck; but fear not, for we have the almighty address bar, which the surfer will use to transition at random to any other page. Thus, all our entries in the row with no outlinks is simply $1/n$, where n is the total number of pages on the Interwebs.

a) In what sense does this transition matrix make sense? Is it in fact a valid transition matrix? What quantity would be of interest in determining the importance of a page?

b) Suppose we had a pretty boring Internet with three pages: `icanhas.cheezburger.com`, `lolcats.com`, and `stat110.net`. (On second thought, that'd be just about all we need anyway!) For whatever reason, suppose that Stat 110 links to both ICanHasCheezburger and LolCats; LolCats links to ICanHasCheezburger; and ICanHasCheezburger does not link to any page. Find the transition matrix and determine the stationary distribution.

c) Suppose now that we had a separate cluster of pages, namely `facebook.com` and `youtube.com`, that linked to each other but not to any of the first three pages. Find the new transition matrix. What problems does this pose for our stationary distribution? How would you fix this without biasing the ranking?

d) Suppose that we are back to the three-page world, and Stat 110 wants to boost its ranking. Unfortunately, it can't really affect what's going on at ICanHasCheezburger or LolCats. Determine the best way to improve Stat 110's PageRank.

Example 49. Let's Play Chess. In chess, the king can move one square at a time in any direction (horizontally, vertically, or diagonally). A king is wandering around on an otherwise empty 8 by 8 chessboard, where for each move all possibilities are equally likely. Find the stationary distribution of this chain.

Example 50. Financial Markov Chains. Suppose we are thinking about a model where you are rich, poor, or in-between, and a stock of interest is at a price that is high or low. The stock transitions between these prices with probability p.

Ideally, you'd like to buy a stock when it is low and you are rich, so if these two conditions coincide, you buy the stock with probability 1. Unfortunately, the stock is expensive, so buying it makes you in-between with probability 1. However, buying the stock does not influence its price (you're a small investor) so it can be high or low with the usual probabilities.

When you are poor, you never buy stocks and transition with probability 1 into the in-between stage in the next period (you earn some money from income). When you are in-between, sometimes you submit to an impulse buy and purchase the stock if it is low, making you poor, with probability α. If you are rich, you also sometimes buy on impulse even if the stock price is high, with probability α.

a) Draw the state-space diagram for this Markov chain, with the corresponding probabilities.

b) Determine the transition matrix for this Markov chain. State whether it is aperiodic and/or irreducible, and which states are recurrent/transient.

c) Find the stationary distribution for this Markov chain, if one exists, and discuss what this says about the model.

Distribution	**PDF and Support**	**EV**	**Variance**	**MGF**
Bernoulli Bern(p)	$P(X=1)=p$ $P(X=0)=q$	p	pq	$q+pe^t$
Binomial Bin(n,p)	$P(X=k)=\binom{n}{k}p^k(1-p)^{n-k}$ $k\in\{0,1,2,\dots n\}$	np	npq	$(q+pe^t)^n$
Geometric Geom(p)	$P(X=k)=q^kp$ $k\in\{0,\ 1,\ 2,\ \dots\}$	q/p	q/p^2	$\frac{p}{1-qe^t}, qe^t<1$
Negative Binom. NBin(r,p)	$P(X=n)=\binom{r+n-1}{r-1}p^rq^n$ $n\in\{0,\ 1,\ 2,\ \dots\}$	rq/p	rq/p^2	$(\frac{p}{1-qe^t})^r, qe^t<1$
Hypergeometric Hypergeometric(w,b,n)	$P(X=k)=\binom{w}{k}\binom{b}{n-k}/\binom{w+b}{n}$ $k\in\{0,1,2,\dots,n\}$	$\mu=\frac{nw}{b+w}$	$\frac{w+b-n}{w+b-1}n\frac{\mu}{n}(1-\frac{\mu}{n})$	–
Poisson Pois(λ)	$P(X=k)=\frac{e^{-\lambda}\lambda^k}{k!}$ $k\in\{0,\ 1,\ 2,\ \dots\}$	λ	λ	$e^{\lambda(e^t-1)}$
Uniform Unif(a,b)	$f(x)=\frac{1}{b-a}$ $x\in(a,b)$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	$\frac{e^{tb}-e^{ta}}{t(b-a)}$
Normal $\mathcal{N}(\mu,\sigma^2)$	$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-(x-\mu)^2/(2\sigma^2)}$ $x\in(-\infty,\infty)$	μ	σ^2	$e^{t\mu+\frac{\sigma^2t^2}{2}}$
Exponential Expo(λ)	$f(x)=\lambda e^{-\lambda x}$ $x\in(0,\infty)$	$1/\lambda$	$1/\lambda^2$	$\frac{\lambda}{\lambda-t}, t<\lambda$
Gamma Gamma(a,λ)	$f(x)=\frac{1}{\Gamma(a)}(\lambda x)^a e^{-\lambda x}\frac{1}{x}$ $x\in(0,\infty)$	a/λ	a/λ^2	$\left(\frac{\lambda}{\lambda-t}\right)^a, t<\lambda$
Beta Beta(a, b)	$f(x)=\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}x^{a-1}(1-x)^{b-1}$ $x\in(0,1)$	$\mu=\frac{a}{a+b}$	$\frac{\mu(1-\mu)}{(a+b+1)}$	–
Chi-Squared χ_n^2	$\frac{1}{2^{n/2}\Gamma(n/2)}x^{n/2-1}e^{-x/2}$ $x\in(0,1)$	n	$2n$	$(1-2t)^{-n/2}, t<1/2$
Multivar Uniform A is support	$f(x)=\frac{1}{\|A\|}$ $x\in A$	–	–	–
Multinomial Mult$_k(n,\vec{p})$	$P(\vec{X}=\vec{n})=\binom{n}{n_1\dots n_k}p_1^{n_1}\dots p_k^{n_k}$ $n=n_1+n_2+\dots+n_k$	$n\vec{p}$	$\text{Var}(X_i)=np_i(1-p_i)$ $\text{Cov}(X_i,X_j)=-np_ip_j$	$\left(\sum_{i=1}^k p_ie^{t_i}\right)^n$

Cauchy-Schwarz	**Markov**	**Chebychev**	**Jensen**
$\|E(XY)\|\leq\sqrt{E(X^2)E(Y^2)}$	$P(X\geq a)\leq\frac{E\|X\|}{a}$	$P(\|X-\mu_X\|\geq a)\leq\frac{\sigma_X^2}{a^2}$	g convex: $E(g(X))\geq g(E(X))$ g concave: $E(g(X))\leq g(E(X))$