36-217 Notes

Michael You

Spring 2017

Contents

1 1/17/2017: Lecture nulla

1.1 Probabilistic Model

- Start from an **idealized** description of the problem
- Identify **all** the things that could happen, called the **sample space**
- Assign probabilities to outcomes
- Random Variable → the computation you care about
- Expected value, an average value for the thing you care about

The book we will be using in this class is Bersetkas and Tsitsiklis: Intro to Probability, 2$^{\text{nd}}$ Edition

1.2 Sample Space

The set of all outcomes

Example 1.1. A coin flip has $\{\text{Heads}, \text{Tails}\}$

Example 1.2. You're running a factory and wondering how many widgets you can produce before production stops $\implies \{0, 1, 2, \dots\}$

Example 1.3. You wonder how long it takes a ^{235}U nucleus to decay, $[0, \infty)$

1.3 Set Notation

We introduce the following notation in this class,

- $\in$: **set inclusion symbol**. Tells you $x \in S$ that "x is *in* S"
- $\emptyset$: **empty set**, the set of no elements
- Ω: The **universal set**, or the set of all the elements we care about

1.4 Kinds of Sets

- **Finite**: $\{x_i\}_{i=1}^n$
- **Countably infinite**: e.g. $\mathbb{Z}$ or any set with a bijection to integers
- **Continuous**: e.g. $[0, \infty)$. Uncountable.

Example 1.4. If we flip a coin 3 times, we have the outcomes,

$$\{HHH\}, \{HHT\}, \{HTH\}, \{HTT\}, \{THH\}, \{THT\}, \{TTH\}, \{TTT\}$$

the probability of each outcome is $\frac{1}{8}$ assuming a fair coin.

Example 1.5. If we don't care about order, then

$$\Omega : \{HHH, HHT, HTT, TTT\}$$
$$P(X_i) : \{\frac{1}{8}, \frac{3}{8}, \frac{3}{8}, \frac{1}{8}\}$$

It is generally easier to include **ordering** when we are counting.

Definition 1.1. We define $S \subset T$ if $\forall s \in S$, $s \in T$. That is, S is a **subset** of T.

Definition 1.2. We define the **union** of S and T as

$$S \cup T = \{x | x \in S \text{ or } x \in T\} \tag{1}$$

Definition 1.3. We define the **intersection of** S and T as

$$S \cap T = \{x | x \in S \text{ and } x \in T\} \tag{2}$$

Definition 1.4. We define the **complement** of S as

$$S^c = \{x | x \in \Omega \text{ and } x \notin S\} \tag{3}$$

2 1/19/2017: Lecture I, Set Algebra, Probability Laws

2.1 Set Algebra

Set **union** and **intersection** satisfy,

1. **Commutativity**:

$$S \cup T = T \cup S$$
$$S \cap T = T \cap S$$

2. **Associativity**:

$$S \cup (T \cup V) = (S \cup T) \cup V$$
$$S \cap (T \cap V) = (S \cap T) \cap V$$

3. **Distributivity**:

$$S \cup (T \cap V) = (S \cup T) \cap (S \cup V)$$
$$S \cap (T \cup V) = (S \cap T) \cup (S \cap V)$$

4. $(S^c)^c = S$
5. $S \cap S^c = \emptyset$
6. $S \cup \Omega = \Omega$
7. $S \cap \Omega = S$
8. $(S \cup T)^c = S^c \cap T^c$
9. $(S \cap T)^c = S^c \cup T^c$

2.2 DeMorgan's Laws

Definition 2.1. **DeMorgan's Laws** state that

a) $\left(\bigcup_{i=1}^{n} S_i\right)^c = \bigcap_{i=1}^{n} S_i^c$

b) $\left(\bigcap_{i=1}^{n} S_i\right)^c = \bigcup_{i=1}^{n} S_i^c$

Definition 2.2. Two sets A and B are **disjoint** if $A \cap B = \emptyset$

Corollary 1. A collection of sets $\{S_i\}_{i=1}^n$ is disjoint if

$$\bigcap_{i=1}^n S_i = \emptyset$$

Definition 2.3. A **partition** of a set S is a collection of sets that are disjoint and whose union gives S. These sets can be thought of as mutually exclusive events that make up a sample space.

$$S = \bigcup_{i=1}^n S_i \quad \text{AND} \quad \bigcap_{i=1}^n S_i = \emptyset$$

Example 2.1.

$$\begin{aligned} A &= (A \cap B) \cup (A \cap B^c) \\ &= A \cap (B \cup B^c) && \text{(Distributivity)} \\ &= A \cap \Omega \\ &\boxed{A}. \end{aligned}$$

Example 2.2.

$$\begin{aligned} (A \cap B) \cap (A \cap B^c) &= A \cap A \cap B \cap B^c && \text{(Associativity)} \\ &= A \cap (B \cap B^c) \\ &= A \cap \emptyset \\ &= \boxed{\emptyset} \end{aligned}$$

As a bonus, we also have that the intersection of a set with the partition sets is also a partition.

2.3 Probability Axioms

$P(A)$ probability law assigns a number to each event A,

1. **Non-negativity**: $P(A) \geq 0$
2. **Additivity**: If A and B are disjoint events, then

$$P(A \cup B) = P(A) + P(B)$$

3. **Normalization**: $P(\Omega = 1$. Something *has to happen*

Example 2.3. Coin flip of H or T.

$$\begin{aligned} &\text{Possible Events: } \{H\}, \{T\}, \emptyset, \Omega \\ &P(\{H, T\}) = 0 \\ &P(\{H\}) = \frac{1}{2} \\ &P(\{T\}) = \frac{1}{2} \\ &P(\{\emptyset\}) = 0 \end{aligned}$$

Example 2.4. Prove $P(\emptyset) = 0$

$$\begin{aligned} P(\Omega \cup \emptyset) &= P(\Omega) + P(\emptyset) \\ P(\Omega) &= P(\Omega) + P(\emptyset) \implies \boxed{P(\emptyset) = 0}. \end{aligned}$$

For our coin flip, we can use example 2.4 to show that

$$P(\{H\}) = 1 - P(\{T\})$$

if our coin is binary.

Definition 2.4. Finite and countably infinite sets are **discrete**. Discrete sample spaces can be partitioned into single events.

Example 2.5. Given $A \subset B$, show $P(A) \leq P(B)$ Let $B = A \cup C$, where A and C are disjoint, then

$$\begin{aligned} P(B) &= P(A \cup C) \\ &= P(A) + P(C) && \text{(Additivity)} \\ \implies \boxed{P(B) \geq P(A)} & && \text{(Nonnegaitivity of } P(C)\text{)} \end{aligned}$$

2.4 Making Probability Laws

1. **Measure**: assume frequencies of previous outcomes imply probabilities of future outcomes
2. **Exploit Independence**
3. **Symmetry**: Choose a sample space where each outcome is equally likely
4. **Make a conanical choice**: everything is already done for you!

3 1/24/2017: Lecture II, Combinatorics

3.1 Counting

Definition 3.1. Basic Counting Principle. If experiment 1 has m outcomes and experiment 2 has n outcomes, experiment 1 and 2 together have mn outcomes.

Example 3.1. We want to know the probability that $n, n \leq 365$ people all have distinct birthdays.

$$\prod_{i=0}^{n-1} \frac{365 - i}{365}$$

because we adjust the total possible birthdays after a person claims a birthday.

We can extend this example to get an approximation using the fact that

$$e^{-x} \approx 1 - x, \text{ for } |x| << 1 \tag{4}$$

then

$$\begin{aligned} \prod_{i=0}^{n-1} \frac{365 - i}{365} &= \prod_{i=0}^{n-1} \left(1 - \frac{i}{365}\right) \\ &= \prod_{i=0}^{n-1} e^{-i/365} \\ &= \boxed{\exp\left[-\sum_{i=1}^{n-1} \frac{i}{365}\right]} \end{aligned}$$

notice that as n increases, the probability that everyone has different birthdays decreases.

Definition 3.2. We define a **factorial** as

$$n! = n \cdot (n-1) \cdots 1 = \boxed{\prod_{i=1}^{n} i} \tag{5}$$

Definition 3.3. We define a **permutation** as

$${}_nP_k = \frac{n!}{(n-k)!} \tag{6}$$

Definition 3.4. We define a **combination** as

$${}_nC_k = \binom{n}{k} = \frac{n!}{k!(n-k)!} \tag{7}$$

Example 3.2. Suppose we wanted to choose a committee of 7 people from 10, with one person being the head. How many ways can we do this?

$$\begin{gathered}\#\text{ ways group} \cdot \#\text{ choices leader} = \text{total} \\ \binom{10}{7} \cdot 7 = 840\end{gathered}$$

Theorem 1. *The* ***Binomial Theorem*** *states that*

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k} \tag{8}$$

Example 3.3.

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

Example 3.4. Consider

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

if we have $p = \frac{1}{2}$, then

$$\begin{aligned}(\frac{1}{2} + \frac{1}{2})^n &= \sum_{k=0}^{n} \binom{n}{k} \frac{1}{2}^k (\frac{1}{2})^{n-k} \\ 1 &= \frac{1}{2^n} \sum_{k=0}^{n} \binom{n}{k} \\ &\implies \boxed{\sum_{k=0}^{n} \binom{n}{k} = 2^n}.\end{aligned}$$

4 1/25/2017: Lecture III, Conditioning

4.1 Definition

Definition 4.1. **Conditional** probability of A given B is

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{9}$$

Example 4.1. Let A be the event that you get more heads than tails in 3 flips of a coin, let B be the event that the first flip is heads, then

$$\begin{aligned} P(A|B) &= \frac{P(A\cap B)}{P(B)} \\ &= \frac{\frac{3}{8}}{\frac{1}{2}} \\ &= \boxed{\frac{3}{4}} \end{aligned}$$

We check that conditional probability satisfies the probability axioms,

1. **Nonnegativity**: $\frac{P(A\cap B)}{P(B)} \geq 0$ because $P(A \cap B)$ is nonnegative, and $P(B) > 0$ from assumption
2. **Normalization**: Let $B = \Omega$, then
$$P(B|B) = \frac{P(B\cap B)}{P(B)} = \frac{P(B)}{P(B)} = 1.$$
3. **Additivity**:
$$\begin{aligned} P(A\cup B|C) &= \frac{P((A\cup B)\cap C)}{P(C)} \\ &= \frac{P((A\cap C)\cup(B\cap C))}{P(C)} \\ &= \frac{P(A\cap C)+P(B\cap C)}{P(C)} \\ &= P(A|C)+P(B|C). \end{aligned}$$

Example 4.2. Monty Hall Problem. Basically, when the host opens a door and shows you a goat, he is giving you a better chance if you switch doors. Nonintuitive, but makes sense if you do some extension of the problem to more doors.

4.2 Multiplication Rule

Definition 4.2. Multiplication Rule. We have $P(A \cap B) = P(A) \cdot P(B|A)$, directly from conditional probability. In general, we can have

$$P\left(\bigcap_{i=1}^{n} A_i\right) = P(A_1)P(A_2|A_1)P(A_3|A_2\cap A_1)\dots P\left(A_n\middle|\bigcap_{i=1}^{n-1} A_i\right) = \boxed{\prod_{i=1}^{b} P\left(A_i\middle|\bigcap_{i=1}^{n-1} A_i\right)} \tag{10}$$

You can prove this by writing out all of the conditional probabilities in fractional form, which results in a telescoping cancellation. An intuitive way to see this is to think about how you move from A_i to A_{i+1}, and how you got to A_i in the first place.

Example 4.3. Aircraft radar example from Bersetkas. Given $P(R|A) = 0.99$ and $P(R|A^c) = 0.1$, where A is the event that an aircraft is present and R is the event that the alarm sounds.

We have, assuming $P(A) = 0.05$

$$\begin{aligned} P(R\cap A) &= 0.0495 \\ P(R^c\cap A) &= 0.0005 \\ P(R\cap A^c) &= 0.095 \\ P(R^c\cap A^c) &= 0.855 \end{aligned}$$

thus,

$$P(A|R) = \frac{P(A\cap R)}{P(R)} = \frac{0.0495}{0.1455} = \boxed{34\%}.$$

4.3 Total Probability

Theorem 2. ***Total Probability Theorem:*** *Let $\{A_i\}_{i=1}^n$ be a partition of Ω with $P(A_i) > 0 \forall i$, then for any event B,*

$$P(B) = \sum_{i=1}^n P(B|A_i)P(A_i). \tag{11}$$

Proof. We have

$$B = \bigcup_{i=1}^n (B \cap A_i)$$

because A_i form a partition. Then

$$\begin{aligned} P(B) &= \sum_{i=1}^n P(B \cap A_i) && \text{(Additivity)} \\ &= \sum_{i=1}^n P(B|A_i)P(A_i) && \text{(Def. of conditional probability)} \end{aligned}$$

□

5 1/31/2017: Lecture IV, Bayes' Theorem

5.1 Definition

Some examples about conditional probability and how probabilities change when you are given information.

Theorem 3. *(Bayes' Theorem) Given A, B, and $P(A), P(B) \neq 0$,*

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)} \tag{12}$$

Proof. Start with

$$\begin{aligned} P(A \cap B) &= P(A)P(B|A) && \text{(def. cond. prob.)} \\ &= P(B)P(A|B) && \text{(def. cond. prob.)} \\ \implies P(A)P(B|A) &= P(B)P(A|B) \\ \implies P(A|B) &= \frac{P(A)P(B|A)}{P(B)} \end{aligned}$$

□

Corollary 2. If we have a partition of A, $\{A_i\}_{i=1}^n$, then we have

$$\begin{aligned} P(B) &= \sum_{i=1}^n P(B|A_i)P(A_i) \\ \implies P(A_i|B) &= \frac{P(B|A_i)P(A_i)}{\sum_{i=1}^n P(B|A_i)P(A_i)} \end{aligned}$$

Example 5.1. Suppose we have a fair coin that comes up H with probability $\frac{1}{2}$ and a biased coin that comes up with heads with probability $\frac{3}{4}$. What is the probability that we have a biased coin given that a single flip gives us heads.

Let the probability that we select the biased coin be f.

$$\begin{aligned}
P(B|H) &= \frac{P(H|B)P(B)}{P(H)} \\
&= \frac{\frac{3}{4}f}{f \cdot \frac{3}{4} + (1-f) \cdot \frac{1}{2}} \\
&= \frac{3f}{f+2}
\end{aligned}$$

What if we flip heads n times?

$$\begin{aligned}
P(B|H_n) &= \frac{P(H_n|B)P(B)}{P(H_n)} \\
&= \frac{\left(\frac{3}{4}\right)^n f}{f \cdot \left(\frac{3}{4}\right)^n + (1-f) \cdot \left(\frac{1}{2}\right)^n} \\
&= \frac{3^n \cdot f}{3^n \cdot f + 2^n - 2^n \cdot f}
\end{aligned}$$

Although we can't compute this limit exactly (with the skills we have), we can simulate the experiment electronically. We see that as the distribution matches the biased coin better, the more likely it is the biased coin.

6 2/2/2017: Lecture V, Independence

6.1 Definition

Definition 6.1. Independent events A and B satisfy

$$P(A \cap B) = P(A)P(B) \tag{13}$$

B doesn't tell you anything about A.

***A misconception is that two disjoint sets are independent. This is NOT true since knowing one event automatically tells you the other is impossible.

Some trivial expository examples on independence.

6.2 Conditional Independence

Definition 6.2. Conditional Independence is defined as

$$P(A \cap B|C) - P(A|C) \cdot P(B|C) \tag{14}$$

that is, A and B might not be independent by themselves, but within C, they are.

Proof. LHS $P(A \cap B|C) = \frac{P(A \cap B \cap C)}{P(C)}$

RHS: $P(A|C)P(B|C) = \frac{P(A \cap C)P(B \cap C)}{P(C)P(C)}$

By the definition of conditional independence,

$$\begin{aligned}
\frac{P(A \cap B \cap C)}{P(C)} &= \frac{P(A \cap C)P(B \cap C)}{P(C)^2} \\
\implies \frac{P(A \cap B \cap C)}{P(B \cap C)} &= \frac{P(A \cap C)}{P(C)} \implies P(A|B \cap C) &= P(A|C) && \text{(true by cond. ind.)}
\end{aligned}$$

□

Example 6.1. For example, if you are given that

$$\begin{aligned}
&H_1 : 1^{\text{st}} \text{ flip heads}\\
&H_2 : 2^{\text{nd}} \text{ flip heads}\\
&D : 2 \text{ tosses diff. results}
\end{aligned}$$

we can figure out probabilities

$$\begin{aligned}
P(H_1) &= \frac{1}{2}\\
P(H_2) &= \frac{1}{2}\\
P(H_1 \cap H_2) &= \frac{1}{4} && (H_1, H_2 \text{ are ind.})
\end{aligned}$$

but,

$$\begin{aligned}
P(H_1|D) &= \frac{1}{2}\\
P(H_2|D) &= \frac{1}{2}\\
P(H_1 \cap H_2|D) &= 0 && (H_1, H_2 \text{ are } \textit{dep.})
\end{aligned}$$

6.3 Mutual Independence

Definition 6.3. Mutual Independence of events $\{A_i\}_{i=1}^n$ is

$$P\left(\bigcap_{i\in S} A_i\right) = \prod_{i\in S} P(A_i) \quad \forall S \subseteq \{A_i\}_{i=1}^n \tag{15}$$

7 2/7/2017: Lecture VI, Random Variables

7.1 Random Variables

Random variables are

1. **numbers** we care about
2. tags each outcome in an experiment with a number

Definition 7.1. A **random variable** is a function from the outcomes of an experiment to the real numbers, i.e. $\{\text{outcomes}\} \mapsto \mathbb{R}$

Example 7.1. In 3 coin flips, we can have random variables,

- X: number of heads $\{0, 1, 2, 3\}$
- Y: whether the first flip is a head $\{0, 1\}$
- Z: number of heads $-$ number of tails $\{-3, -1, 1, 3\}$

Example 7.2. Indicator RVs are random variables that are defined to be 1 if A happens and 0 otherwise.

When we say

$$\{X = x\},$$

we have all the outcomes where RV X takes on the value of x.

7.2 Probability Mass Functions (PMFs)

Definition 7.2. We define the probability of an event $\{X = x\}$ as

$$p_X(x) = P(\{X = x\}) \tag{16}$$

p here is known as the probability mass function (PMF).

Example 7.3. What if we try to consider a continuous event? For example if it is equally probable for $x \in [0, 1]$?

$$P(\Omega) = 1 = \sum_{x \in [0,1]} P(\{X = x\})$$

however, since there are infinite possible x, we run into a problem. Thus, PMFs are only defined over **discrete** random variables.

7.3 Properties of PMFs

1. $0 \leq p_X(x) \leq 1$
2. $p_X(x) = 0$ for a countable number of x since X is discrete
3. $\sum_{\forall x} p_X(x) = 1$ by normalization

Example 7.4. Bernoulli random variable is defined to be

$$p_X(0) = 1 - p$$
$$p_X(1) = p$$

Example 7.5. Discrete uniform RV is defined to be

$$p_X(x) = \frac{1}{b - a + 1}, \quad a \leq x \leq b | x \in \mathbb{Z}$$

8 2/9/2017: Lecture VII, Probability Mass Functions (PMFs)

Definition 8.1. A **random variable** X assigns a value to each possible outcome of an experiment.

For example, $X = \{0, 1, 2, 3\}$

8.1 Bernoulli Random Variable

Definition 8.2. A **Bernoulli Random Variable** can take on the values 0 or 1.

Definition 8.3. A **Bernoulli Process** is a series of Bernoulli trials, X_i, where each trial is **identical** and **independent** of each other.

For example, a Bernoulli process could be,

- Sequence of coin flips
- Digital inverter
- Free throws

Problem 8.1. What's the probability of getting a sequence of k successes in n trials of a Bernoulli random variable?

Solution 8.1. The probability of k successes is p^k and the probability of $n-k$ failures is $(1-p)^{n-k}$. Since there are $\binom{n}{k}$ ways to make such a sequence, our answer is,

$$p_X(k) = \binom{n}{k} p^k (1-p)^{n-k} \tag{17}$$

We are interested to see if normalization applies to our answer. It should, since it is a probability. We start with

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

Observing 17 we expect the most probable outcome to be np successes, because of independent trials and linearity. If we graph $p_X(k)$ against $k = 0$ to $k = n$, the peak will be at np, and the distribution looks a lot like a normal distribution.

A quick definition for the next problem,

Definition 8.4. In the **Erdős-Rényi** random graph model, there are n nodes, and, for each pair of notes, a probability p that there is a link between them.

Problem 8.2. If you choose a random node from such a graph, what is the probability that it is linked to exactly d other nodes?

Solution 8.2. We need d successful connections, and $(n-1)-d$ "failures" (non-connections). Therefore, our answer is

$$p_D(d) = \binom{n-1}{d} p^d (1-p)^{n-d-1}$$

8.2 Geometric Random Variable

Suppose now we want to know Y, the **trial number** of the first success. E.g.

$$TTTTH \qquad (Y = 5)$$

Then

$$p_Y(k) = (1-p)^{k-1} p, \quad k \in \mathbb{Z}, 1 \le k \le \infty$$

We should check the normalization of p_Y.

$$\begin{aligned}\sum_{k=1}^{\infty} p_Y(k) &= \sum_{k=1}^{\infty} (1-p)^{k-1} p \\ &= p \sum_{k=1}^{\infty} (1-p)^{k-1} \\ &= p \sum_{j=0}^{\infty} (1-p)^j \\ &= p \left(\frac{1}{1-(1-p)} \right) \\ &= p \left(\frac{1}{p} \right) \\ &= 1.\end{aligned}$$

We notice that because $(1-p) < 1$,

$$p_X(k) < p_X(k+1),$$

so for $p > 0$, $p_X(k)$ is strictly *decreasing* in k. The geometric distribution is exponentially decaying.

Problem 8.3. Monty needs to find a date for Valentine's day. He asks one person a day and has probability p of successfully asking someone out on any day. Suppose today is 2/10/17 (he starts asking this day). What is the probability that he cops a date by 2/14/17?

a) Suppose that $p = 0.1$. What is the probability that he finds a date **on** 2/14/17?

b) Suppose that $p = 0.1$. What is the probability that he finds a date **by** 2/14/17?

Solution 8.3. a) $(0.9)^4(0.1) = \boxed{0.07}$

b) Add up the probability for each day before also,

$$\begin{aligned}\sum_{k=1}^{5} p_X(k) &\approx 0.1 + 0.09 + 0.08 + 0.07 + 0.07 \\ &= \boxed{0.41}\end{aligned}$$

Looking good, Monty.

9 2/14/17: Lecture VIII, The Poisson PMF

9.1 Bernoulli Process in Time

We can choose different Δt for coin flip trials.

How many heads do we get in an interval T?

$$\begin{aligned}\text{\# of heads in } n \text{ flips: } &\sim np \\ \text{\# of flips in time } T: &\sim \frac{T}{\Delta t} \\ \Longrightarrow \text{\# of heads in time } T: &\sim p\left(\frac{T}{\Delta t}\right)\end{aligned}$$

As $\Delta t \to 0$, the number of heads in time T goes to ∞. This is not good because we want the distribution to be roughly the same over different time intervals.

We want to keep the number of heads in time T fixed. So we try to change p as we change Δt. In particular,

$$p = \frac{\lambda \Delta t}{T}$$

Now, the number of heads is

$$\text{\# heads in time } T: \frac{\lambda \Delta t}{T} \cdot \frac{T}{\Delta t} = \lambda$$

As we take $\Delta t \to 0$, we get λ heads still in time T. This is called the **Poisson Process**. The Poisson Process happens in *continuous time*, and can be thought of as the limit of the Bernoulli Process.

Examples of Poisson Processes:

- Radioactive Decay
- Server Failure
- Traffic to a webpage

Let X be the number of events (heads) in a time interval of length T in a Poisson process. Then

$$p_X(k) = e^{-\lambda}\frac{\lambda^k}{k!}, k \geq 0, k \in \mathbb{Z}. \tag{18}$$

In $p_X(k)$, the λ exponential grows quickly, but not as fast as $\frac{1}{k!}$ is decreasing.

Let's check that Equation (18) is actually a PMF.

If we try summing $p_X(k)$,

$$\begin{aligned}\sum_{k=0}^{\infty} e^{-\lambda}\frac{\lambda^k}{k!} &= e^{-\lambda}\sum_{k=0}^{\infty}\frac{\lambda^k}{k!}\\ &= e^{-\lambda}\left(e^{\lambda}\right)\\ &= 1.\end{aligned}$$

So the normalization is good.

When,

- $\lambda = 0$, can't be normalized
- $\lambda < 0$, non-negativity is violated

Problem 9.1. Assume N, the number of photons hitting a pixel during an exposure can be modeled with Poisson, with $\lambda = 1$.

Solution 9.1.

$$p_N(0) = e^{-1}\frac{(1)^0}{0} = \boxed{0.368} \tag{19}$$

$$p_N(1) = e^{-1}\frac{(1)^1}{1} = \boxed{0.368} \tag{20}$$

$$P(N > 1) = 1 - p_N(0) - p_N(1) = \boxed{0.26}. \tag{21}$$

Using binomial, the probability of one photon with $p = 0.1$ in 10 trials is

$$\binom{10}{1}(0.1)(0.9)^9 = \boxed{0.387},$$

and with $p = 0.01$ in 100 trials,

$$\binom{100}{1}(0.01)(0.99)^9 9 = \boxed{0.370}.$$

9.2 Poisson RV from Binomial

Start with the binomial PMF,

$$p_B k = \binom{n}{k}p^k(1-p)^{n-k} \tag{22}$$

Replace p with $\lambda = np$

$$\begin{aligned}
p_Bk &= \binom{n}{k}\left(\frac{\lambda}{n}\right)^k\left(1-\left(\frac{\lambda}{n}\right)\right)^{n-k} \\
&= \frac{\lambda^k}{k!}\frac{n!}{n^k(n-k)!}\left(1-\left(\frac{\lambda}{n}\right)\right)^{n-k} \\
&= \left(\frac{\lambda^k}{k!}\right)\left(\frac{n}{n}\cdot\frac{n-1}{n}\cdot\frac{n-2}{n}\cdots\frac{n-k+1}{n}\right)\left(1-\left(\frac{\lambda}{n}\right)\right)^{n-k} \\
&= \left(\frac{\lambda^k}{k!}\right)\left(1-\left(\frac{\lambda}{n}\right)\right)^{n-k} && \text{(assume large } n\text{)} \\
&= \left(\frac{\lambda^k}{k!}\right)\exp\left((n-k)\log\left(1-\left(\frac{\lambda}{n}\right)\right)\right) \\
&\approx \left(\frac{\lambda^k}{k!}\right)\exp\left((n-k)\left(-\frac{\lambda}{n}\right)\right) && (\log(1+x)\approx x, \quad |x|<<1) \\
&= \left(\frac{\lambda^k}{k!}\right)\exp\left(-\lambda+k\frac{\lambda}{n}\right) \\
&= \left(\frac{\lambda^k}{k!}\right)\exp(-\lambda+kp) && (\lambda = np) \\
&\approx \left(\frac{\lambda^k}{k!}\right)\left(e^{-\lambda}\right) && (p \text{ is small}) \\
\implies p_B(k) &= \boxed{\frac{\lambda^k}{k!}e^{-\lambda}} && \text{(Poisson PMF)}
\end{aligned}$$

When

$$n >> 1, p << 1, \quad \text{for } k \sim \lambda = np.$$

Problem 9.2. Find the answers to the following Poisson distributions, wth $n = 100$, and $p = 0.4$.

Solution 9.2. Poisson:

- $p_X(40) = \frac{40^40}{40!}e^{-40} = \boxed{0.0629}$.
- $p_X(60) = \frac{40^60}{60!}e^{-40} = \boxed{0.000679}$
- $p_X(120) = \frac{40^120}{120!}e^{-40} = \boxed{1.122 \cdot 10^{-24}}$

Binomial:

- $p_X(40) = \frac{40^40}{40!}e^{-40} = \boxed{0.0629}$.
- $p_X(60) = \frac{40^60}{60!}e^{-40} = \boxed{0.000679}$
- $p_X(120) = \frac{40^120}{120!}e^{-40} = \boxed{1.122 \cdot 10^{-24}}$

10 2/16/17: Midterm 1

Didn't do that well...most upset about the last problem because of **overcounting.**

10.1 Problem 6 (20 points)

There are 54 cards in this standard deck with a black and red joker. We deal 18 cards to three people, Anna, B and C.

1. Suppose Anna tells you she has one joker in her hand. What is the probability she has both jokers?

   Solution

   So we use conditional probability, (this is the part where I screwwed up because I calculated the number of ways to get one joker wrong :( )

$$\begin{aligned} P(2J|J) &= \frac{P(2J \cap J)}{P(J)} \\ &= \frac{\binom{52}{16}}{\binom{54}{18} - \binom{52}{18}} && \text{(Notice the denom. is using comp. counting)} \\ &= \frac{\frac{52!}{16!\cdot 36!}}{\frac{54!}{18!\cdot 36!} - \frac{52!}{18!\cdot 34!}} \\ &= \frac{52! \cdot 18 \cdot 17}{54! - 52! \cdot 35 \cdot 36} \\ &= \frac{18 \cdot 17}{54 \cdot 53 - 36 \cdot 35} \\ &= \frac{17}{89}. \end{aligned}$$

2. Now, Anna tells you she has a *red* joker specifically. What is the probability she has both jokers?

   Solution

   We use a similar approach,

$$\begin{aligned} P(2J|J_r) &= \frac{P(2J \cap J_r)}{P(J_r)} \\ &= \frac{\binom{52}{16}}{\binom{54}{18} - \binom{53}{18}} && \text{(because } \tbinom{53}{18} \text{ ways \textbf{not} to pick the red)} \\ &= \frac{\frac{52!}{16!\cdot 36!}}{\frac{54!}{18!\cdot 36!} - \frac{53!}{18!\cdot 35!}} \\ &= \frac{52! \cdot 18 \cdot 17}{54! - 53! \cdot 36} \\ &= \frac{18 \cdot 17}{54 \cdot 53 - 53 \cdot 36} \\ &= \frac{17}{53}. \end{aligned}$$

3. Suppose now Anna says she *doesn't* have the red joker. What is the probability she doesn't have any jokers?

Again, conditional probability,

$$\begin{aligned}
P(2J^c|J_r^c) &= \frac{P(2J^c \cap J_r^c)}{P(J_r^c)} \\
&= \frac{\binom{52}{18}}{\binom{54}{18} - \binom{53}{17}} && \text{(because } \tbinom{53}{17} \text{ ways to pick the red)} \\
&= \frac{\frac{52!}{18! \cdot 34!}}{\frac{54!}{18! \cdot 36!} - \frac{53!}{17! \cdot 36!}} \\
&= \frac{52! \cdot 18 \cdot 17}{54! - 53! \cdot 18} \\
&= \frac{18 \cdot 17}{54 \cdot 53 - 53 \cdot 18} \\
&= \frac{17}{106}.
\end{aligned}$$

Actually, I don't think that was the problem but I'm just doing it for practice.

4. Now, suppose Anna says she *doesn't* have the red joker. What is the probability she has the other joker?

$$\begin{aligned}
P(J_b|J_r^c) &= \frac{P(J_b \cap J_r^c)}{P(J_r^c)} \\
&= \frac{\binom{52}{17}}{\binom{54}{18} - \binom{53}{17}} && \text{(because } \tbinom{53}{17} \text{ ways to pick the red)} \\
&= \frac{\frac{52!}{17! \cdot 35!}}{\frac{54!}{18! \cdot 36!} - \frac{53!}{17! \cdot 36!}} \\
&= \frac{52! \cdot 18 \cdot 36}{54! - 53! \cdot 18} \\
&= \frac{18 \cdot 36}{54 \cdot 53 - 53 \cdot 18} \\
&= \frac{18}{53}.
\end{aligned}$$

Once again, just practice since I don't recall this answer either.

5. Ok, time for the whammy. Suppose Anna tells you she has a red joker and then reveals n of her non-joker cards. What is the probability she has both jokers after her n reveals?

   Let A_n denote the event of n reveals. I think I screwed up because once she tells you her n reveals, THOSE CARDS ARE DEFINED, and cannot be done whatever $\binom{52}{n}$ ways. They are the n cards exactly. Therefore, the numerator will have to change.

$$\begin{aligned}
P(2J|J_r \cap A_n) &= \frac{P(2J \cap J_r \cap A_n)}{P(J_r \cap A_n)} \\
&= \frac{\binom{52-n}{16-n}}{1 \cdot \binom{52}{n}\binom{53-n}{17-n}} \\
&= \frac{\frac{(52-n)!}{(16-n)! \cdot 36!}}{\frac{52!}{(n)! \cdot (52-n)!} \cdot \frac{(53-n)!}{(17-n)! \cdot 36!}} \\
&= \frac{(52-n)!}{(16-n)! \cdot 36!} \cdot \frac{(n)! \cdot (52-n)!}{52!} \cdot \frac{(17-n)! \cdot 36!}{(53-n)!} \\
&= \frac{(52-n)!}{1} \cdot \frac{(n)!}{52!} \cdot \frac{17}{53} \\
&= \boxed{\frac{17}{53 \cdot \binom{52}{n}}}
\end{aligned}$$

Not sure if this the answer...

Chris got

$$\frac{17-n}{52-n},$$

which makes a lot of sense.

10.2 Problem 1 (15 points)

This problem went a little better, but I shouldn't have failed the proof for this until so late.

1. Given $P(A|B) = 1$, prove $P(B^c|A^c) = 1$.

   Solution

$$\begin{aligned}
P(A|B) &= 1 \\
\frac{P(A \cap B}{P(B)} &= 1 \\
P(A \cap B) &= P(B)
\end{aligned}$$

From here, we know that we must have $A \cap B = B$. Continuing with this fact,

$$\begin{aligned}
A \cap B &= B \\
A^c \cup B^c &= B^c && \text{(DeMorgan's Laws)}
\end{aligned}$$

We also know that

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

and

$$\begin{aligned}
(A^c \cap B) \cap (A^c \cap B^c) &= B \cap B^c \cap A^c && \text{(def. intersection)} \\
&= \emptyset && \text{(def. complement)} \\
\implies P(A^c) &= P(A^c \cap B) + P(A^c \cap B^c)
\end{aligned}$$

similarly,

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

So

$$\begin{aligned}
A^c \cup B^c &= (A^c \cap B) \cup (A^c \cap B^c) \cup (B^c \cap A) \\
P(A^c \cup B^c) &= P(A^c \cap B) + P(A^c \cap B^c) + P(B^c \cap A) \\
P(A^c \cup B^c) + P(B^c \cap A^c) &= (P(A^c \cap B) + P(A^c \cap B^c)) + (P(B^c \cap A) + P(B^c \cap A^c)) \\
P(A^c \cup B^c) + P(B^c \cap A^c) &= [P(A^c \cap B) + P(A^c \cap B^c)] + [P(B^c \cap A) + P(B^c \cap A^c)] \\
P(A^c \cup B^c) + P(B^c \cap A^c) &= P(A^c) + P(B^c) \\
\implies P(A^c \cup B^c) &= P(A^c) + P(B^c) - P(B^c \cap A^c)
\end{aligned}$$

From earlier that $A^c \cup B^c = B^c$, we have

$$\begin{aligned}
P(A^c \cup B^c) &= P(B^c) \\
P(A^c) + P(B^c) - P(B^c \cap A^c) &= P(B^c) \\
P(A^c) &= P(B^c \cap A^c) \\
1 &= \frac{P(B^c \cap A^c)}{P(A^c)} \\
1 &= P(B^c | A^c)
\end{aligned}$$

11 2/21/17: Lecture IX, Expected Value I

In a Poisson random variable, λ represents how many successes you expect to get in a time period.

We are trying to talk about **averages** today. Once we have PMFs, which gives us numbers, we can talk in terms of numbers about how the PMF behaves—we can make **weighted averages**.

11.1 Expected Value

Definition 11.1.

$$\mathbb{E}[\underbrace{X}_{\text{function } X}] = \sum_{\underbrace{x}_{\text{runs over values of } x}} \underbrace{x}_{\text{thing we're averaging}} \underbrace{p_X(x)}_{\text{weights}} \tag{23}$$

For some notational things, $\mathbb{E}$ is NOT a function. The brackets [] represent that it is a **functional**, meaning that it takes in not a number, but a function as its argument.

An analogy of the expected value is center of mass. We can imagine a bunch of values on a number line.

Problem 11.1. You have two questions to answer, a and b. You can only answer the second question if you get the first question right. You will get a right with probability p_a and b right with p_b. You get a prize with value V_a for getting question a right and V_b for getting question b right.

Which question should you attempt first?

Solution 11.1. We first try to look at the sample space and figure out the random variables of interest.

Questions we can answer correctly	Payment if we answer a first	Payment if we answer b first
Neither	0	0
Only a	V_a	0
Only b	0	V_b
Both	$V_a + V_b$	$V_a + V_b$

The PMF of A is

$$\begin{aligned} p_A(0) &= P(\text{None}) + P(\text{only } B) = 1 - p_a \\ p_A(V_a) &= P(\text{only } a) = p_a(1-p_b) \\ p_A(V_a+V_b) &= P(\text{both}) = p_a p_b \end{aligned}$$

Then the expected value of A is,

$$\begin{aligned} \mathbb{E}[A] &= \sum_x x p_A(x) && \text{(Expected Value Definition)} \\ &= 0\cdot(1-p_a) + V_a\cdot(p_a(1-p_b)) + (V_a+V_b)\cdot(p_a p_b) \\ &= V_a p_a + V_b p_a p_b \end{aligned}$$

Now let's d $\mathbb{E}[A]$.
The PMF of B is

$$\begin{aligned} p_B(0) &= 1 - p_b \\ p_B(V_b) &= p_b(1-p_a) \\ p_B(V_b+V_a) &= p_b p_a \end{aligned}$$

Then the expected value of B is,

$$\begin{aligned} \mathbb{E}[B] &= \sum_x x p_B(x) \\ &= 0\cdot(1-p_b) + V_b\cdot(p_b(1-p_a)) + (V_b+V_a)\cdot(p_b p_a) \\ &= V_b p_b + V_a p_b p_a \end{aligned}$$

Now we want to make a decision. Based on our expected values,

$$\begin{aligned} \mathbb{E}[A] &= V_a p_a + V_b p_a p_b \\ \mathbb{E}[B] &= V_b p_b + V_a p_b p_a \end{aligned}$$

When is it better to answer A first?
When $\mathbb{E}[A] > \mathbb{E}[B]$, it is better to **answer a first**. We can simplify this inequality,

$$\begin{aligned} V_a p_a + V_b p_a p_b &> V_b p_b + V_a p_b p_a \\ V_a p_a - V_b p_b &> (V_a - V_b) p_a p_b \\ \frac{V_a}{p_b} - \frac{V_b}{p_a} &> V_a - V_b \end{aligned}$$

Problem 11.2. There are three small cars. There are 36 clowns in the first car, 40 clowns in the second car, and 44 clowns in the third. Suppose you choose one clown from the 120 total (with equal probability of picking each).

Solution 11.2. What is $\mathbb{E}[X]$? Where X denotes the number of clowns in the car of the randomly chosen clown.

We find the PMF,

$$\begin{aligned} p_X(36) &= \frac{36}{120} \\ p_X(40) &= \frac{40}{120} \\ p_X(44) &= \frac{44}{120} \end{aligned}$$

Thus

$$\begin{aligned}
\mathbb{E}[X] &= \sum_x x \cdot p_X(x) \\
&= 36 \cdot \frac{36}{120} + 40 \cdot \frac{40}{120} + 44 \cdot \frac{44}{120} \\
&= \frac{1}{120}\left(36^2 + 40^2 + 44^2\right) \\
&= \frac{4832}{120} = \frac{604}{15} = \boxed{40.27}.
\end{aligned}$$

This is slightly higher than the average number of clowns in the cars, which is 40. The reason is because the probability of finding a clown in a car with a higher number of cars is higher than a car with less clowns. Thus, the expected value will be skewed towards a higher than average number. (O'Connell: the bigger car is getting more weight, and the smaller car is getting a smaller weight) The application of this example is that there can be a bias when you sample and don't be careful about distributions of people like this. You might end up with the wrong average.

11.2 $\mathbb{E}[X]$ of Canonical PMFs

Bernoulli

$$\mathbb{E}[X] = 0 \cdot (1-p) + 1 \cdot p = \boxed{p}.$$

12 2/23/17: Lecture X, Expected Value II

We are trying to find the expected value of the binomial.

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

Using the property that

$$\binom{n}{k} = \frac{n}{k}\binom{n-1}{k-1},$$

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{k=1}^{n} n \binom{n-1}{k-1} p^k (1-p)^{n-k} && \text{(Applying property)} \\
&= np \sum_{k=1}^{n} \binom{n-1}{k-1} p^{k-1} (1-p)^{(n-1)-(k-1)} && \text{(Factoring out)} \\
&= np \sum_{j=0}^{n-1} \binom{n-1}{j} p^{j} (1-p)^{(n-1)-j} && \text{(Changing dummy)} \\
&= np \cdot 1^* = \boxed{np}
\end{aligned}$$

where at $*$ we notice that we are just summing over the entire Binomial RV so we know

$$\sum_j p_x(j) = 1.$$

Problem 12.1. There are N plastic balls in a bucket. K are white and the remaining $N-K$ are red. Suppose you draw n balls from the bucket without putting any back. If X is the number of balls you draw from the bucket, the PMF for X is **hypergeometric**,

$$p_X(k) = \frac{\binom{K}{k}\binom{N-k}{n-k}}{\binom{N}{n}}. \tag{24}$$

Find the $\mathbb{E}[X]$. Assume $n < K$.

Solution 12.1. Using the definition of expected value,

$$\begin{aligned}
\mathbb{E}[X] &= \sum_x x \cdot p_X(x) \\
&= \sum_{k=0}^{n} k \frac{\binom{K}{k}\binom{N-k}{n-k}}{\binom{N}{n}} \\
&= \sum_{k=1}^{n} k \frac{\frac{K}{k}\binom{K-1}{k-1}\binom{(N-1)-(k-1)}{(n-1)-(k-1)}}{\frac{N}{n}\binom{N-1}{n-1}} && \text{(Try to cancel out the k and produce $V-1$'s)} \\
&= \frac{Kn}{N} \sum_{k=1}^{n} \frac{\binom{K-1}{k-1}\binom{(N-1)-(k-1)}{(n-1)-(k-1)}}{\binom{N-1}{n-1}} \\
&= \frac{Kn}{N} \sum_{j=0}^{n-1} \frac{\binom{K-1}{j}\binom{(N-1)-j)}{(n-1)-j}}{\binom{N-1}{n-1}} \\
&= \frac{Kn}{N} \underbrace{\sum_{j=0}^{n'} \frac{\binom{K'}{j}\binom{(N')-j)}{(n')-j}}{\binom{N'}{n'}}}_{\text{Sum of Hypergeometric PMF}} \\
&= \boxed{\frac{Kn}{N}}
\end{aligned}$$

12.1 Convergence

Theorem 4. *If a random variable X covers a finite number of values, $\mathbb{E}[X]$ is finite.*

If X can take on a countably infinite number of values, we are not guaranteed that the expected value is finite. What this means is that

1. Individual outcomes: finite
2. Naïve average of a finite # of outcomes: finite
3. When **divergent**, the averages do not approach a single value.

Example 12.1. Given the PMF,

$$p_X(2^k) = 2^{-k}, \quad k \in \mathbb{N}, \tag{25}$$

We first find the normalization,

$$\begin{aligned}
\sum_{k=1}^{\infty} p_X(2^k) &= \sum_{k=1}^{\infty} 2^{-k} \\
&= \left(\sum_{k=0}^{\infty} 2^{-k}\right) - 2^0 \\
&= 2 - 1 = \boxed{1}
\end{aligned}$$

We have a valid PMF then.

The expected value is

$$\begin{aligned}\mathbb{E}[X] &= \sum_{k=1}^{\infty} 2^k 2^{-k} \\ &= \sum_{k=1}^{\infty} 1 = \boxed{\infty}\end{aligned}$$

.

So the PMF is perfectly valid but the expected value still *diverges*.

12.2 Functions of Random Variables

Definition 12.1. We define a function of a random variable as

$$Y = g(X), \tag{26}$$

where if X takes the value x, Y takes the value $g(x)$.

If each value of X is mapped to a unique value of $Y = g(X)$, the PMF of Y is,

$$p_Y(g(x)) = p_x(X)$$

If this isn't true,

$$p_Y(y) = \sum_{x|g(x)=y} p_X(x),$$

so tries to find all values such that it maps into the Y space (??).

Suppose the PMF for X is,

then the PMF of $Y = |X|$ is,

Example 12.2. Let $Y = g(x)$, do we need the PMF for Y if we want to find the expected value?

$$\begin{aligned}\mathbb{E}[Y] &= \sum_y y p_Y(y) \\ &= \sum_y y \sum_{x|g(x)=y} p_X(x) \\ &= \sum_y \sum_{x|g(x)=y} g(x) p_X(x) && \text{(mapping from x to y)} \\ &= \sum_x g(x) p_X(x) && \text{(we are just summing over the values of x)}\end{aligned}$$

In the double sum, we needed to make sure that each value of x is represented exactly once.

Our result shows us that just knowing $p_X(x)$ is sufficient to compute $\mathbb{E}[g(x)]$.

Tip 12.1. IF YOU WANT TO FIND THE EXPECTED VALUE, DO NOT USE

$$g(p_X(x)), p_X(g(x))$$

or anything like that. Use the **original PMF**. We are just changing the random variable.

e.g., if you want to find $g(x) = \sqrt{x}$, do NOT do

$$\sum_x \sqrt{x}\sqrt{p_X(x)}.$$

13 2/28/17: Lecture XI, Variance

Example 13.1. We have the following probabilities at a casino,

$p(X)$	Winnings
$\frac{2}{3}$	\$0
$\frac{1}{6}$	\$15
$\frac{1}{6}$	\$30

For high rollers, they are allowed to bet twice as much \$10 instead \$5, but get twice the earnings. There is linearity in this problem which makes it possible to do the 3 random variables, which are

$$W(\text{ winnings}), NW_1(\text{ normal }), W_2(\text{ high roller}),$$

We calculate the expected values for the following,

$$\begin{aligned}\mathbb{E}[W] &= \frac{2}{3}\cdot 0 + \frac{1}{6}\cdot 15 + \frac{1}{6}\cdot 30 = \frac{15}{2} \\ \mathbb{E}[W] &= W - 10 + 5 = \frac{5}{2} \\ \mathbb{E}[W] &= 2W - 20 + 5 = 0\end{aligned}$$

13.1 Variance

Knowing which family the distribution comes from **does not** tell you too much about how similar they are. E.g. we look at three distributions, the two binomial ones are very different but the Poisson and binomial are similar. We notice that qualitatively, the **width** is what we should look at.

A *narrow* distribution are closer to $\mathbb{E}[X]$ and *wide* ones are more scattered.

Definition 13.1. The **variance** is defined qualitatively as how scattered points are in a distribution. Quantitatively,

$$\text{Var}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2] \tag{27}$$

If we want to find the variance, we just use **the PMF of** x. Another formula for the variance, using PMF.

$$\text{Var}[X] = \sum_x (x - \mathbb{E}[X])^2 p_X(x) \tag{28}$$

To find another formula for variance,

$$\begin{aligned}\text{Var}[X] &= \mathbb{E}[(X - \mathbb{E}[X])^2] \\ &= \mathbb{E}[X^2 - 2X\mathbb{E}[X] + \mathbb{E}[X]^2] \\ &= \mathbb{E}[X^2] - 2\mathbb{E}[X\mathbb{E}[X]] + \mathbb{E}[\mathbb{E}[X]^2] \\ &= \mathbb{E}[X^2] - 2\mathbb{E}[X\mathbb{E}[X]] + \mathbb{E}[X]^2 && \text{(Because thing inside } \mathbb{E} \text{ is constant)} \\ &= \mathbb{E}[X^2] - 2\mathbb{E}[X]^2 + \mathbb{E}[X]^2 && \text{(Because } \mathbb{E} \text{ is constant and linearity of } \mathbb{E}\text{)} \\ &= \boxed{\mathbb{E}[X^2] - \mathbb{E}[X]^2}\end{aligned}$$

Example 13.2. Uniformly distributed with

$$p_X(k) = \frac{1}{2a+1},$$

where k ranges from $-a$ to a, and $k \in \mathbb{Z}$. We first do the calculation for $a = 3$.

To find $\mathbb{E}[X]$, we notice that the distribution is symmetric and uniform, and thus the expected value must be $\boxed{0}$.

To find $\mathbb{E}[X^2]$,

$$\begin{aligned}\mathbb{E}[X^2] &= \frac{1}{2\cdot 3+1}\Big((-1)^2+(-2)^2+(-3)^2+0^2+1^2+2^2+3^2\Big)\\ &= \frac{1}{7}(28)\\ &= \boxed{4}.\end{aligned}$$

To find $\text{Var}[X]$,

$$\text{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X] = \boxed{4}.$$

We are a little worried about our variance because it is greater than the maximum value in our distribution. However, we remind ourselves that we are ranging from -3 to 3, so the answer is not all that unreasonable.

13.2 Standard Deviation

There happens to be another measure of distribution scatter,

Definition 13.2. The **standard deviation** is defined to be

$$\sigma_X = \sqrt{\text{Var}[X]}. \tag{29}$$

The standard deviation is useful because it has the same units as $\mathbb{E}[X]$. So what does

$$\mathbb{E}[X] + \sigma_X$$

mean? In our case, this is a *width* of the distribution.

Example 13.3. Find the expected value and variance of the Poisson distribution.

Recall the Poisson PMF is,

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

O'Connell just gives us,

$$\mathbb{E}[X] = \lambda$$

In order to compute the variance, we need

$$\mathbb{E}[X^2] = \sum_{k=0}^{\infty} k^2 \cdot e^{-\lambda}\frac{\lambda^k}{k!}$$

which is absolutely terrible. So we try to do a different, but similar sum,

$$\begin{aligned}\mathbb{E}[X(X-1)] &= \sum_{k=0}^{\infty} k(k-1)\cdot e^{-\lambda}\frac{\lambda^k}{k!}\\ &= \sum_{k=2}^{\infty} k(k-1)\cdot e^{-\lambda}\frac{\lambda^k}{k!} && \text{(because } k=0,1 \text{ contribute 0)}\\ &= \sum_{k=2}^{\infty} e^{-\lambda}\frac{\lambda^k}{(k-2)!}\\ &= \lambda^2\sum_{k=2}^{\infty} e^{-\lambda}\frac{\lambda^{k-2}}{(k-2)!}\\ &= \lambda^2\sum_{j=0}^{\infty} e^{-\lambda}\frac{\lambda^j}{(j)!}\\ &= \lambda^2 && \text{(Poisson PMF is normalized)}\end{aligned}$$

With this result, we can say,

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

Thus we can get the variance,

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

Since we got λ for both $\mathbb{E}[X]$ and $\text{Var}[X]$, we must conclude that X is **dimensionless**, since it is not possible for expected value and variance to have the same units.

13.3 Samples Statistics

So when we have a PMF, we can compute $\mathbb{E}[X], \text{Var}[X], \sigma_X$. But when we don't, and have finite data, we can use **sample statistics,**

- Sample Mean

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

- Sample Variance

$$\sigma_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

Differences between sample statistics and $\mathbb{E}[X]$, *etc.* (ideal quantities,

Sample Statistics	**Ideal**
Almost always informative about the actual experiment	Always incorporates modeling assumptions about the experiment
Actual numbers chnage from trial to trial	Compute once and does not change
Always finite (even when they “shouldn't” be)	Sometimes diverge

14 3/2/2017: Lecture XII, PMFs: Joint, Marginal, Conditional

Problem 14.1. You have a bucket of 6 balls.

- 1 red
- 2 white
- 3 blue

You draw 3 balls. Let X be the number of red balls you draw and Y be the number of white balls you draw. What is the joint PMF of X and Y?

Solution 14.1. We know that the total number of outcomes is $\binom{6}{3} = 20$.

Calculating all the combinations,

- $X = 0, Y = 0 : \binom{3}{3} = 1$
- $X = 0, Y = 1 : \binom{2}{1}\binom{3}{2} = 6$
- $X = 0, Y = 2 : \binom{2}{2}\binom{3}{1} = 3$
- $X = 1, Y = 0 : \binom{1}{1}\binom{3}{2} = 3$
- $X = 1, Y = 1 : \binom{1}{1}\binom{2}{1}\binom{3}{1} = 6$
- $X = 1, Y = 2 : \binom{1}{1}\binom{2}{2} = 1$

We then draw a table for $p_X, Y(x, y)$,

$p_{X,Y}(x,y)$	0	1	2
0	$\frac{1}{20}$	$\frac{6}{20}$	$\frac{3}{20}$
1	$\frac{3}{20}$	$\frac{6}{20}$	$\frac{1}{20}$

Notice the probabilities add up to 1.

14.1 Marginal PMFs

Definition 14.1. The marginal PMF is defined to be

$$p_X(x) = \sum_y p_{X,Y}(x,y) \tag{30}$$

$$p_Y(y) = \sum_x p_{X,Y}(x,y) \tag{31}$$

because the probability that $Y = y$ *ignores* the value of X.

Example 14.1. Consider the PMF from Problem 14.1, then

$p_{X,Y}(x,y)$	0	1	2	
0	$\frac{1}{20}$	$\frac{6}{20}$	$\frac{3}{20}$	$P_X(0) = \frac{10}{20}$
1	$\frac{3}{20}$	$\frac{6}{20}$	$\frac{1}{20}$	$P_X(1) = \frac{10}{20}$
	$P_Y(0) = \frac{4}{20}$	$P_Y(1) = \frac{12}{20}$	$P_Y(2) = \frac{4}{20}$	

Notice we satisfy

$$\sum_y p_Y(y) = 1$$

$$\sum_x p_X(x) = 1$$

So we just went from the joint PMF to the marginal PMF. Can we go the other way?

In general, it will **not** be possible. Suppose that X can take n different values and Y can take on m different values. We need $nm = 1$ numbers in order to specify $p_{X,Y}(x,y)$ because having any two empty can be ambiguous.

If we want to specify $p_X(x)$ and $p_Y(y)$, it takes

$$X : n - 1$$

$$Y : m - 1$$

which is a total of $n + m - 2$ numbers. Since these numbers are unequal in general, we cannot guarantee that we find the joint PMF from the marginal PMF; the joint PMF contains *wayy more information* (in general) than the marginal PMF.

14.2 Functions of more than one random variable

Example 14.2. Suppose $Q = g(X, Y)$. The PMF for Q is,

$$p_Q(q) = \sum_{x,y|g(x,y)=q} p_{X,Y}(x,y)$$

If follows that $\mathbb{E}[Q]$ is,

$$\begin{aligned}
\mathbb{E}[Q] &= \sum_q q \cdot p_Q(q) \\
&= \sum_q q \sum_{x,y|g(x,y)=q} p_{X,Y}(x,y) \\
&= \sum_{X,Y} g(x,y) p_{X,Y}(x,y) && \text{(because } q = g(x,y)\text{)}
\end{aligned}$$

14.3 Linearity of expectation

Suppose that $Z = aX + bY$. What is $\mathbb{E}[Z]$?

$$\begin{aligned}
\mathbb{E}[Z] &= \sum_{x,y} (ax + by) p_{X,Y}(x,y) \\
&= \left(\sum_{x,y} ax p_{X,Y}(x,y) \right) + \left(\sum_{x,y} by p_{X,Y}(x,y) \right) \\
&= \left(a \sum_x x \sum_y p_{X,Y}(x,y) \right) + \left(b \sum_y y \sum_x p_{X,Y}(x,y) \right) \\
&= \left(a \sum_x x p_X(x) \right) + \left(b \sum_y y p_Y(y) \right) && \text{(We spot marginal PMFs)} \\
&= a\mathbb{E}[X] + b\mathbb{E}[Y] && \text{(Exp. val. definition)}
\end{aligned}$$

This is a *powerful* result because we did not make any assumptions about $p_{X,Y}(x,y)$.

To generalize,

Definition 14.2. If random variable Z is a linear combination of some other random variables,

$$Z = \sum_{i=1}^{n} c_i X_i,$$

then

$$\mathbb{E}[Z] = \sum_{i=1}^{n} c_i \mathbb{E}[X_i]. \tag{32}$$

The result is known as **linearity of expectations**.

Example 14.3. Suppose we are performing n independent Bernoulli trials, each with probability p of success. Let X_i be an *indicator* variable for success on the i^{th} trial. $X_i = 1$ for a success, $X_i = 0$ for a failure. $\mathbb{E}[X_i] = p$ since all trials have success with probability p.

We know X, the total number of success in terms of X_i is,

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

So

$$\begin{aligned}\mathbb{E}[X] &= \mathbb{E}\left[\sum_{i=1}^{n} X_i\right] \\ &= \sum_{i=1}^{n} \mathbb{E}[X_i] \\ &= \sum_{i=1}^{n} p \\ &= \boxed{np}\end{aligned}$$

Problem 14.2. There are n people in a party. Each person pulls out a hat without replacement. Let X be the number of people who get their own hat back. What is $\mathbb{E}[X]$?

Solution 14.2. We know that

$$\mathbb{E}[X_i] = \frac{1}{n}$$

$$\begin{aligned}\mathbb{E}[X] &= \sum_{i=1}^{n} \mathbb{E}[X_i] \\ &= \sum_{i=1}^{n} \frac{1}{n} \\ &= n \cdot \frac{1}{n} = \boxed{1}.\end{aligned}$$

Alternative:
The sample space is $n!$ permutations. $(n-1)!$ permutations with one hat fixed.

$$\mathbb{E}[X_i] = \frac{(n-1)!}{n!} = \frac{1}{n}$$

So when we apply linearity of expectations we have $\sum_{i=1}^{n} \mathbb{E}[X_i] = \boxed{1}$. The problem is *much easier* because we were able to split up the problem into multiple separate cases.

15 3/7/2017: Lecture XIII, PMFs: Conditioning and Independence

15.1 Conditional PMF

Has the analogous definition to **conditional probability**.

Definition 15.1. The **conditional PMF** is defined as

$$p_{X|Y}(x \mid y) = \frac{p_{X,Y}(x,y)}{p(Y)} \tag{33}$$

Example 15.1. Trying to send a message with probability p each time. After n trials your program gives up. What is the PMF for number of attempts?

Without a limit of trials, we know that the PMF is

$$p_X(k) = (1-p)^{k-1}p$$

Let A be the event that the number of attempts is n or fewer.

$$P(A) = \sum_{k=1}^{n} p_X(k) = \sum_{k=1}^{n} (1-p)^{k-1}p.$$

So we can find

$$\begin{aligned} p_{X|A}(x) &= \frac{P(\{X = x\} \cap A}{P(A)} \\ &= \frac{(1-p)^{k-1}p}{\sum_{m=1}^{n}(1-p)^{m-1}p} \quad (k \leq n) \end{aligned}$$

Normalization:

$$\sum_{k=1}^{n} p_{X|A}(k) = \sum_{k=1}^{n} \frac{(1-p)^{k-1}p}{\sum_{m=1}^{n}(1-p)^{m-1}p} = 1$$

15.2 Conditional Expectation

Since conditional PMFs are still PMFs, computing *conditional expected values are still the same.*

Definition 15.2. The **conditional expected value** given an event A is defined to be

$$\mathbb{E}[X \mid A] = \sum_{x} x p_{X|A}(x) \tag{34}$$

Given a value of Y is defined to be

$$\mathbb{E}[X \mid Y = y] = \sum_{x} x p_{X|Y}(x \mid y) \tag{35}$$

Use a vertical line inside the PMF function so you know it's conditional.

Problem 15.1. Two rolls of a fair 4-sided die. Let X be the sum of the two rolls. Find the PMF and expected values of X.

Now suppose that you are told the sum is less than 5. Call this event A. What are the PMF and expected value of X given A?

x	$p_X(x)$
2	$\frac{1}{16}$
3	$\frac{1}{8}$
4	$\frac{3}{16}$
5	$\frac{1}{4}$
6	$\frac{3}{16}$
7	$\frac{1}{8}$
8	$\frac{1}{16}$

Table 1: Expected value for two rolls

Solution 15.1. The expected value is then

$$\mathbb{E}[X] = \frac{2 + 6 + 12 + 20 + 18 + 14 + 8}{16} = \boxed{5}.$$

We could have also noticed that the PMF was summetric, so we know the expected value is 5.

To find the conditional PMF, we know that our denominator is now $\frac{3}{8}$ because we only have $X = 2, 3, 4$. This is the value of $P(A)$, the event we get 2, 3, or 4 as the roll.

x	$p_{X\|A}(x)$
2	$\frac{1}{16} \cdot \frac{8}{3} = \frac{1}{6}$
3	$\frac{1}{8} \cdot \frac{8}{3} = \frac{1}{3} =$
4	$\frac{3}{16} \cdot \frac{8}{3} = \frac{1}{2}$

Table 2: Expected value for two rolls given A

Now the expected value,

$$\mathbb{E}[X|A] = 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{3} + 4 \cdot \frac{1}{2} = \boxed{\frac{10}{3}}.$$

15.3 Independence

Definition 15.3. Independent PMFs are defined as follows:

- For a random variable and an event,

$$P(\{X = x\} \cap A) = p_X(x)P(A) \tag{36}$$

$$p_{X|A}(x) = p_X(x) \tag{37}$$

  these conditions have to hold for *all* x, which is much stronger than our original independence relationship.

- For two random variables,

$$p_{X,Y}(x, y) = p_X(x)p_Y(y) \tag{38}$$

$$p_{X|Y}(x|y) = p_X(x) \tag{39}$$

  must hold for *all* pairs of x and y.

15.4 Multiple random variables

Reminder, if we have 3 random variables, X, Y, Z, they are mutually independent if

$$p_{X,Y,Z}(x, y, z) = p_X(x)p_Y(y)p_Z(z)$$

Let's look at $p_{X,Y}(x, y)$,

$$\begin{aligned} p_{X,Y}(x, y) &= \sum_z p_{X,Y,Z}(x, y, z) \\ &= \sum_z p_X(x)p_Y(y)p_Z(z) \\ &= p_X(x)p_Y(y) \sum_z p_Z(z) \\ &= p_X(x)p_Y(y) && \text{(Since } \textstyle\sum_z p_Z(z) = 1\text{)} \end{aligned}$$

so we have that triple independence implies $p_{X,Y}(x, y) = p_X(x)p_Y(y)$.

So why is this different that we get this stronger result? It is because our original condition is *stronger*.

Example 15.2. Assume that X and Y are 2 independent RV's. Then,

$$\begin{aligned}
\mathbb{E}[XY] &= \sum_{x,y} xyp_{X,Y}(x,y) \\
&= \sum_{x,y} xyp_X(x)p_Y(y) && \text{(By independence)} \\
&= \left[\sum_x xp_X(x)\right]\left[\sum_y yp_Y(y)\right] \\
&= \boxed{\mathbb{E}[X]\mathbb{E}[Y]}
\end{aligned}$$

Example 15.3. Suppose $X.Y$ are 2 independent RV's. Find var$[X+Y]$.

We know by definition that

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

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

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

So

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

This looks like variance is linear but IT IS NOT. Just a coincidence from the independence of X, Y.

Problem 15.2. Suppose X is a binomial RV with parameters n and p. We know that $\mathbb{E}[X] = np$. What is var$[X]$?

Solution 15.2. We start with the variance definition,

$$\text{var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$$

And then we realize that calculating $\mathbb{E}[X^2]$ is basically the worst thing ever.

We notice that all the trials in a binomial distribution are independent. Then, the strategy is to introduce a set of Bernoulli RVs X_i that are 1 if the trial i is successful and 0 otherwise.

What is the variance of Bernoulli RV? We have for the PMF of Bernoulli that

$$\begin{aligned}
p_{X_i}(1) &= p \\
p_{X_i}(1) &= (1-p)
\end{aligned}$$

and thus

$$\text{var}[X_i] = \mathbb{E}[X_i^2] - \mathbb{E}[X_i]^2 = p - p^2 = p(1-p)$$

$$\begin{aligned}\text{var}[X] &= \text{var}\left[\sum_{i=1}^{n} X_i\right] \\ &= \sum_{i=1}^{n} \text{var}[X_i] \\ &= \sum_{i=1}^{n}\Big(\mathbb{E}[X_i^2] - \mathbb{E}[X_i]^2\Big) \\ &= \sum_{i=1}^{n}\Big(p - p^2\Big) \\ &= \boxed{np(1-p)}.\end{aligned}$$

16 3/9/2017: Lecture XIV, Total Expectation and Iterated Expectation

Definition 16.1. Suppose we repeat an experiment n times and record the outcome of each experiment in a discrete random variable X_i. If we assume that the X_i are independent,

$$S_n = \frac{\sum_{i=1}^{n} X_i}{n} \tag{40}$$

Example 16.1. The expected value of S_n,

$$\begin{aligned}\mathbb{E}[S_n] &= \mathbb{E}\left[\frac{\sum_{i=1}^{n} X_i}{n}\right] \\ &= \frac{1}{n}\cdot\big(n\mathbb{E}[X_i]\big) \\ &= \boxed{\mathbb{E}[X_i]}\end{aligned}$$

Example 16.2. The variance of S_n,

$$\begin{aligned}\text{Var}[S_n] &= \text{Var}\frac{1}{n}\sum_{i=1}^{n} X_i \\ &= \sum_{i=1}^{n} \text{Var}[\frac{1}{n}X_i] && \text{(By independence)} \\ &= \frac{1}{n^2}\sum_{i=1}^{n} \text{Var}[X_i] && \text{(Scaling of Variance)} \\ &= \frac{n\text{Var}[X_i]}{n^2} \\ &= \boxed{\frac{\text{Var}[X_i]}{n}}\end{aligned}$$

Notice that because variance is *not linear*, we had to square when we pulled the constant out of the variance.

16.1 Total Probability

- General case: given a partition $\{A_i\}_{i=1}^n$,

$$P(B) = \sum_{i=1}^{n} P(A_i)P(B|A_i) \tag{41}$$

- RV conditioned on events: From general, make $B = \{X = x\}$

$$p_X(x) = \sum_{i=1}^{n} P(A_i)p_{X|A_i}(x) \tag{42}$$

- RV conditioned on another RV: from RV conditioned on events, $A_i = \{Y = y\}$

$$p_X(x) = \sum_{y} p_Y(y)p_{X|Y}(x|y) \tag{43}$$

Theorem 5. *The Total Expectation Theorem states that if $A_i, i = 1, 2, \ldots, n$ partition the sample space, total probability implies*

$$\begin{aligned}
p_X(x) &= \sum_{i=1}^{n} P(A_i)p_{X|A}(x) \\
\sum_x xp_X(x) &= \sum_x x \sum_{i=1}^{n} P(A_i)p_{X|A}(x) \\
\mathbb{E}[X] &= \sum_x \sum_{i=1}^{n} xP(A_i)p_{X|A}(x) \\
&= \sum_{i=1}^{n} \sum_x xP(A_i)p_{X|A}(x) \\
&= \sum_{i=1}^{n} P(A_i) \underbrace{\sum_x xp_{X|A}(x)}_{conditional\ exp.} \\
&= \sum_{i=1}^{n} P(A_i)\mathbb{E}[X|A_i]
\end{aligned}$$

we could have also had

$$\mathbb{E}[X] = \sum_y p_Y(y)\mathbb{E}[X|Y = y] \tag{44}$$

Example 16.3.

$$\mathbb{E}[X] = \sum_{i=1}^{3} P(A_i)\mathbb{E}[X|A_i]$$

- $A_1 = A, P(A_1) = \frac{1}{2}$, $E[X|A] = 0.05$
- $A_2 = B, P(A_2) = \frac{3}{10}$, $E[X|B] = 0.1$
- $A_3 = C, P(A_3) = \frac{1}{5}$, $E[X|C] = 0.3$

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{Y=A,B,C} P(Y)E[X|Y] \\
&= \text{plug in values}
\end{aligned}$$

Example 16.4. We will compute the mean of the geometric distribution. Intuition, when we fail, the rest of the experiment can be described with a new geometric RV.

What is $\mathbb{E}[X|X > 1]$?

$$\begin{aligned} p_{X|X>1}(k) &= \frac{P(\{X = x\} \cap \{X > 1\})}{P(\{X > 1\})} \\ &= \begin{cases} (1-p)^{k-2}p, k = 2, 3, 4, \dots, \infty \\ 0, \text{otherwise} \end{cases} \end{aligned}$$

We now try to find the expected value,

$$\begin{aligned} \mathbb{E}[X|X > 1] &= \sum_{k=2}^{\infty} k(1-p)^{k-2}p \\ &= \sum_{j=1}^{\infty} (j+1)(1-p)^{j-1}p \end{aligned}$$

$$\mathbb{E}[X|X > 1] = \mathbb{E}[X + 1] = 1 + \mathbb{E}[X] \tag{45}$$

Now, apply total expectation,

$$\begin{aligned} \mathbb{E}[X] &= P(X = 1)E[X|X = 1] + P(X > 1)E[X|X > 1] \\ &= p \cdot 1 + (1-p) \cdot (1 + \mathbb{E}[X]) \\ &= p + 1 - p + \mathbb{E}[X] - p\mathbb{E}[X] \\ \implies \mathbb{E}[X] &= \frac{1}{p} \end{aligned}$$

Definition 16.2. The **conditional independence** is defined for X, Y conditioned on A as

$$p_{X,Y|A}(x, y) = p_{X|A}(x)p_{Y|A}(y) \tag{46}$$

this is equivalent to $p_{X|Y,A}(x|y) = p_{X|A}(x)$

Notice that like conditional independence we defined before, conditional independence *does not* imply unconditional independence and vice versa. They describe different spaces.

16.2 Conditional Independence and Modeling

Suppose we're building up a model. It might be easier to build up probabilities conditioned on an event A and its complement.

We can use total probability to build up the joint PMF,

$$p_{X,Y}(x, y) = P(A)p_{X|A}(x)p_{Y|A}(y) + (1 - P(A))p_{X|A^c}(x)p_{Y|A^c}(y)$$

the marginal PMFs are

$$\begin{aligned} p_X(x) &= P(A)p_{X|A}(x) + (1 - P(A))p_{X|A^c}(x) \\ p_Y(y) &= P(A)p_{Y|A}(y) + (1 - P(A))p_{Y|A^c}(y) \end{aligned}$$

Generally speaking, X and Y are NOT independent.

16.3 Iterated Expression

Earlier we defined

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

however, we can think of $\mathbb{E}[X|Y]$ as a *random variable*, i.e.

$$\mathbb{E}[X|Y] = E[X|Y=y], \qquad \text{when } Y=y$$

if $\mathbb{E}[X|Y=y] = f(y)$, then $\mathbb{E}[X|Y] = f(Y)$

If we think of $\mathbb{E}[X|Y]$ as a random variable, what is its expected value?

$$\begin{aligned}\mathbb{E}[\mathbb{E}[X|Y]] &= \sum_y \mathbb{E}[X|Y=y]p_Y(y) \\ &= \mathbb{E}[X] && \text{(by total expectation)}\end{aligned}$$

this is useful when we know how to split a problem up using a RV *and* we're mostly interested in an expected value.

Example 16.5. Suppose the number of free throws a team attempts in a basketball game is a Poisson random variable, with parameter λ. Suppose each free throw is independent and successful with probability p. If Y is the number of successful free throws in a game, what is $\mathbb{E}[Y]$?

The expected number of successes given a fixed number of attempts.

$$\mathbb{E}[Y|X] = Xp$$

we can iterate expectations to remove the condition,

$$\begin{aligned}\mathbb{E}[Y] &= \mathbb{E}[\mathbb{E}[Y|X]] \\ &= \mathbb{E}[Xp] && = \boxed{p\lambda}\end{aligned}$$

17 3/21/2017: Lecture XV, Continuous Random Variables

We are interested in continuous random variables because they are more pragmatic than doing discrete experiments when the number of outcomes is too high.

Example 17.1. We drop a needle on a table and record the angle it makes with a horizontal line.

The point of this experiment is that there are infinite values that the angle can be, thus it is **continuous**. Now, we ask ourselves what is the probability that $\theta = \pi/4$. We know that this should be probability $\frac{1}{N}$, where N is the number of outcomes. But wait...there are *infinite* N's...so the probability is just 0. We see that it doesn't make sense to talk about the probability of individual outcomes.

However, we see that taking a range of values, such as $0 \le \theta \le \pi/4$ over $[0, \pi/2]$, makes sense. The probability of this θ is $\frac{1}{2}$, since it is half of the total area.

So now, we decide to introduce a new concept, a **probability density function**. This function assigns a *number* (**not a probability**) to each outcome. Integrating the PDF over a range of outcomes gives a probability. The PDF **cannot** be negative because then bad things happen

Definition 17.1. A **probability density function** has to satisfy

1. $P(a \le X \le b) = \int_a^b f_X(x)\,\mathrm{d}x$, where $f_X(x)$ is the PDF of X
2. $P(a \le X \le a + \delta \approx f_X(x) \cdot \delta$
3. Normalization: $\int_{-\infty}^{\infty} f_X(x)\,\mathrm{d}x = 1$

   Note that we often adjust the range when $f_X(x) = 0$ for parts of the domain, because we don't have to do the integral in those places.

Problem 17.1. We are dealing with the uniform PDF of θ from 0 to $\pi/2$. We want to find the PDF of this experiment.

Solution 17.1. Since the PDF is uniform, $\forall\theta, f_X(\theta) = k$. We can integrate this PDF for normalization,

$$\begin{aligned}
\int_0^{\frac{pi}{2}} k \,\mathrm{d}\theta &= k\theta\Big|_0^{\frac{\pi}{2}} \\
&= k \cdot \frac{\pi}{2} - 0 \\
&= k \cdot \frac{\pi}{2} = 1 \\
\Longrightarrow &\boxed{k = \frac{2}{\pi}}
\end{aligned}$$

Therefore, we have

$$\boxed{f_X(\theta) = \begin{cases} \frac{2}{\pi}, & 0 \le \theta \le \pi/2 \\ 0, & \text{Otherwise} \end{cases}}$$

Problem 17.2. Given the PDF

$$f_X(x) = \frac{c}{\sqrt{x}}$$

- sketch the PDF
- compute c
- find $f_X(\frac{1}{16})$, $f_X(0)$
- find $P(X \le \delta)$ where $\delta \in (0, 1)$. What happens as $\delta \to 0$?

Solution 17.2.

-

$$\begin{aligned}
\int_0^1 \frac{c}{\sqrt{x}} \,\mathrm{d}x &= c\, 2\sqrt{x}\Big|_0^1 \\
&= c(2 - 0) \\
&= 2c = 1 \\
\Longrightarrow &\boxed{c = \frac{1}{2}}.
\end{aligned}$$

- $f_X(\frac{1}{16}) = \frac{1}{2}\frac{1}{\sqrt{1/16}} = \boxed{2}$. $f_X(0)$ is undefined.
- $\int_0^\delta \frac{c}{\sqrt{x}} \,\mathrm{d}x = 2\sqrt{\delta}$. The probability of $\lim_{\delta\to 0} P(X \le \delta) = \lim_{\delta\to 0} 2\sqrt{\delta} = 0$.

Left early for major declaration, finish rest later.

18 3/23/2017: Lecture XVI, Canonical Continuous Random Variables

18.1 Discrete vs. Continuous

- Specifying Probabilities:

$$\text{Discrete: } P(\{X = x\}) = p_X(x)$$

$$\text{Continuous: } P(\{a \le x \le b\}) = \int_a^b f_X(x) \,\mathrm{d}x\,,$$

where $f_X(x)$ is the PDF, *not the probability*, and $\mathrm{d}x$ is the probability

- Expected Values:

$$\text{Discrete: } \mathbb{E}[g(x)] = \sum_x g(x|P_X(x))$$

DOREST

$$\text{Continuous: } P(\{a \le x \le b\}) = \int_a^b f_X(x)\,\mathrm{d}x\,,$$

18.2 An Example: Uniform PDF

Definition 18.1. The Uniform PDF with range (a, b) is defined to be

$$f_X(x) = \begin{cases} \frac{1}{b-a}, & a \le x \le b \\ 0, & \text{otherwise} \end{cases} \tag{47}$$

Let us compute the **expected value** of the Uniform PDF.

$$\begin{aligned}
\mathbb{E}[X] &= \int_{-\infty}^{\infty} x f_X(x)\,\mathrm{d}x \\
&= \int_a^b x \frac{1}{b-a}\,\mathrm{d}x \\
&= \frac{1}{b-a}\int_a^b x\,\mathrm{d}x \\
&= \frac{1}{b-a}\left.\frac{x^2}{2}\right|_a^b \\
&= \frac{1}{2(b-a)}(b^2 - a^2) \\
&= \boxed{\frac{b+a}{2}}.
\end{aligned}$$

Let us now compute the variance for the Uniform PDF.

$$\begin{aligned}
\mathbb{E}[X^2] &= \int_{-\infty}^{\infty} x^2 f_X(x)\,\mathrm{d}x \\
&= \int_a^b x^2 \frac{1}{b-a}\,\mathrm{d}x \\
&= \frac{1}{b-a}\int_a^b x^2\,\mathrm{d}x \\
&= \frac{1}{b-a}\left.\frac{x^3}{3}\right|_a^b \\
&= \frac{1}{3(b-a)}(b^3 - a^3) \\
&= \frac{b^2 + ab + a^2}{3}. && \text{(Since } b^3 - a^3 = (b-a)(b^2+ab+a^2)\text{)}
\end{aligned}$$

$$\begin{aligned}
\mathrm{Var}[X^2] &= \mathbb{E}[X^2] - \mathbb{E}[X]^2 \\
&= \frac{b^2+ab+a^2}{3} - \frac{b^2+2ab+a^2}{4} \\
&= \frac{4b^2+4ab+4a^2}{12} - \frac{3b^2+6ab+3a^2}{12} \\
&= \frac{b^2+a^2-2ab}{12} \\
&= \boxed{\frac{(b-a)^2}{12}}
\end{aligned}$$

We can then calculate the standard deviation as well

$$\sigma_X = \sqrt{\mathrm{Var}[X]} = \boxed{\frac{b-a}{\sqrt{12}}}$$

An observation that we look at here is that the location of the distribution does not affect the variance, because $b-a$ is constant no matter where we slide the distribution, whereas the expected value $b+a$ will change as we slide the distribution around.

18.3 Exponential Random Variable

The motivation for this is that between Bernoulli and Poisson, we have

1. Binomial → $\lambda = \frac{p}{\Delta t}$ → Poisson
2. Geometric → ??? (Exponential)

The exponential random variable can be thought of *time* until success.

In order to derive this, we are going to start with the geomemtric random variable. Let X be the geometric random variable for trials until next success.

The PMF for geometric,

$$p_X(k) = (1-p)^{k-1}p$$

To find the PMF for T, we notice $T = X\Delta t$. So $k\Delta t = t$ in the equation above,

$$p_T(t) = (1-p)^{\frac{t}{\Delta t}-1}p$$

We need to write p in terms of Δt as we send $p, \Delta t \to 0$. We want to find a rate for the success...

$$\beta = \frac{p}{\Delta t} \longleftrightarrow p = \beta\Delta t$$

Motivation is that we expect to see p events in a Δt interval. Now, we substitute in for p,

$$p_T(t) = (1-\beta\Delta t)^{\frac{t}{\Delta t}-1}\beta\Delta t,$$

and take the limit $\Delta t \to 0$. But we want to compare probabilities first,

$$\begin{aligned}
f_T(t)\Delta t &= (1-\beta\Delta t)^{\frac{t}{\Delta t}-1}\beta\Delta t \\
\implies f_T(t) &= (1-\beta\Delta t)^{\frac{t}{\Delta t}-1}\beta \\
f_T(t) &= \lim_{\Delta t\to 0}\beta(1-\beta\Delta t)^{\frac{t}{\Delta t}-1} \\
&= \lim_{\Delta t\to 0}\beta\left[\left(\frac{t}{\Delta t}-1\right)\log(1-\beta\Delta t)\right] \\
&= \lim_{\Delta t\to 0}\beta\left[\left(\frac{t}{\Delta t}-1\right)(-\beta\Delta t)\right] \\
&= \lim_{\Delta t\to 0}\beta\left[-\beta t+\beta\Delta t\right] \\
&= \boxed{\beta e^{-\beta t}}
\end{aligned}$$

We see that the exponent forces the units of β to be $\frac{1}{\text{time}}$. This means the PDF is also $\frac{1}{\text{time}}$ units. These are the right units because integrating the PDF over time gives probabilities, which means the PDF has to be units $\frac{1}{t}$. *The PDF is a density* (with units $\frac{1}{t}$ in this case).

We want to check normalization for this PDF,

$$\begin{aligned}
1 &= \int_{-\infty}^{\infty} f_T(t)\,\mathrm{d}t \\
&= \int_0^{\infty} \beta e^{-\beta t}\,\mathrm{d}t \\
&= -e^{-\beta t}\Big|_0^{\infty} \\
&= 0-(-1) = \boxed{1}.
\end{aligned}$$

18.4 A Calculus Trick

For calculations like expected value and variance, we need to calculate integrals like,

$$\int t^a e^{-\beta t}\,\mathrm{d}t$$

We start from an easy integral,

$$\int_0^{\infty} e^{-\beta t}\,\mathrm{d}t = \frac{1}{\beta}$$

We are going to differentiate both sides with respect to β,

$$\begin{aligned}
\frac{\partial}{\partial\beta}\left(\int_0^{\infty} e^{-\beta t}\,\mathrm{d}t\right) &= \frac{\partial}{\partial\beta}\left(\frac{1}{\beta}\right) \\
\int_0^{\infty} -te^{-\beta t}\,\mathrm{d}t &= -\frac{1}{\beta^2} \\
\int_0^{\infty} te^{-\beta t}\,\mathrm{d}t &= \frac{1}{\beta^2}
\end{aligned}$$

Let's do it again!

$$\begin{aligned}
\frac{\partial}{\partial\beta}\left(\int_0^{\infty} te^{-\beta t}\,\mathrm{d}t\right) &= \frac{\partial}{\partial\beta}\left(\frac{1}{\beta^2}\right) \\
\int_0^{\infty} -t^2e^{-\beta t}\,\mathrm{d}t &= -\frac{2}{\beta^3} \\
\int_0^{\infty} t^2e^{-\beta t}\,\mathrm{d}t &= \frac{2}{\beta^3}
\end{aligned}$$

Computing the expected value and variance,

- Expected Value:
$$\mathbb{E}[T] = \int_0^\infty t \cdot \beta e^{-\beta t}\,\mathrm{d}t = \beta \frac{1}{\beta^2} = \frac{1}{\beta}$$
- Variance: First do
$$\mathbb{E}[T^2] = \int_0^\infty t^2 \cdot \beta e^{-\beta t}\,\mathrm{d}t = \beta \frac{2}{\beta^3} = \frac{2}{\beta^2}$$
then,
$$\mathrm{Var}[T] = \mathbb{E}[T^2] - \mathbb{E}[T] = \frac{2}{\beta^2} - \frac{1}{\beta^2} = \frac{1}{\beta^2}$$

Problem 18.1. Your server is able to run on average for 10 days without crashing. You just turned it on, what is the probability that it runs for at least one day, but not for more than two days?

Solution 18.1. We have
$$\mathbb{E}[T] = \frac{1}{\beta} = 10 \Longrightarrow \beta = \frac{1}{10}$$

We calculate,
$$\begin{aligned}\int_1^2 \frac{1}{10} e^{-\frac{t}{10}}\,\mathrm{d}t &= \frac{1}{10}\left(-10e^{-\frac{t}{10}}\right)\Big|_1^2 \\ &= \frac{1}{10}\left(-10e^{-\frac{1}{5}} + 10e^{-\frac{1}{10}}\right) \\ &= e^{-\frac{1}{10}} - e^{-\frac{1}{5}} \\ &\approx \boxed{8.61\%}.\end{aligned}$$

Just some notes...we have that $\beta = 0.1\,\mathrm{d}^{-1}$

19 3/28/17: Lecture XVII, Cumulative Distributive Functions

19.1 Normal Distribution

We start with the PDF,
$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right], \tag{48}$$
where
$$\mathbb{E}[X] = \mu, \qquad \mathrm{Var}[X] = \sigma^2.$$

Notice that we can read $\mathbb{E}, \mathrm{Var}$ directly from the PDF.

Some useful things, (YOU SHOULD LEARN THESE)

- $P(\mu - \sigma < x < \mu + \sigma) \approx 0.68$
- $P(\mu - 2\sigma < x < \mu + 2\sigma) \approx 0.95$

19.2 Cumulative Distributive Function (CDF)

Definition 19.1.
$$F_X(x) = P(X \leq x), \text{ defined } \forall x \in \mathbb{R} \tag{49}$$

- **Continuous**: $F_X(x) = \int_{-\infty}^{x} f_X(x')\,\mathrm{d}x'$
- **Discrete**: $F_X(x) = \sum_{x' \leq x} p_X(x)$

Notice that the CDF is a *probability*, and contains the same info as the PDF/PMF. For continuous RV, the CDF is the area under the PDF up to x.

CDF Properties

- **Nondecreasing**: If $x \leq y$, then $F_X(x) \leq F_X(y)$. For continuous variables, we have $F_X(y) - F_X(x) = \int_x^y f_X(x')\,\mathrm{d}x'$, $f_X(x') \geq 0$
- **Range**: As $x \longrightarrow -\infty$, $F_X(x) \longrightarrow 0$, and as $x \longrightarrow \infty$, $F_X(x) \longrightarrow 1$

19.3 PMF/PDFs from CDFs

- PDF:
$$f_X(x) = \frac{\mathrm{d}F_X(x)}{\mathrm{d}x}$$
- PMF:
$$\lim_{\epsilon \to 0^+} \big(F_X(x) - F_X(x - \epsilon)\big)$$

Example 19.1. The uniform RV:

- PDF: $f_X(x) = \frac{1}{b-a}, \quad a \leq x \leq b$
- CDF: $F_X(x) = \int_a^x \frac{1}{b-a}\,\mathrm{d}x' = \frac{x-a}{b-a}$

Problem 19.1. Find the CDF for exponential RV.

Solution 19.1.

$$\begin{aligned}
F_X(x) &= \int_{-\infty}^x \beta e^{-\beta t}\,\mathrm{d}t \\
&= \int_0^x \beta e^{-\beta t}\,\mathrm{d}t && \text{(Make sure to redefine the domain)} \\
&= -e^{-\beta t}\Big|_0^x \\
&= \boxed{1 - e^{-\beta x}}
\end{aligned}$$

Example 19.2. Let's find the CDF of the geometric PDF.

$$F_X(k) = \sum_{k' \leq k} (1-p)^{k'-1} p$$

We notice $P(X > k) = (1-p)^k$, because you have to fail k times at least in order to have $X > k$, and thus by complement, $P(X \leq k) = 1 - (1-p)^k \implies$ definition of CDF!

$$F_X(k) = 1 - (1-p)^{\lfloor k \rfloor}, \quad k \geq 0$$

We observe the geometric and exponential CDFs, and see that they converge as $\Delta t \longrightarrow \infty$.

19.4 More about the Normal Distribution

We have *bad news*, the normal CDF doesn't have a closed form. We can get numerical approximations, but no closed forms.

However, we can **standardize** normal distributions!

Consider $Y = aX + b$, with X normal. We notice

- $\mathbb{E}[Y] = a\mathbb{E}[X] + b$. If we choose the right b, then $\mathbb{E}[Y] = 0$.

- $\mathrm{Var}[Y] = a^2 \mathrm{Var}[X]$. We can choose a such that $\mathrm{Var}[Y] = 1$.

so we can fix the spot of the normal distribution. We did a change of variables for Y, so let's find the PDF for Y, (the big idea is that we are trying to match probabilities rather than the densities)

$$f_Y(y = ax + b)\,\mathrm{d}y = f_X(x)\,\mathrm{d}x\,,$$

the differentials turn the densities into probabilities. Using $\mathrm{d}y = a\,\mathrm{d}x$ and $x = \frac{y-b}{a}$,

$$\begin{aligned} f_Y(y) &= \frac{1}{a} f_X\left(\frac{y-b}{a}\right) \\ &= \frac{1}{a\sigma\sqrt{2\pi}} \exp\left[-\frac{\left(\frac{y-b}{a} - \mu\right)^2}{2\sigma^2}\right] \\ &= \frac{1}{(a\sigma)\sqrt{2\pi}} \exp\left[-\frac{\big(y - (a\mu + b)\big)^2}{2(a\sigma)^2}\right] \end{aligned}$$

our result is a *normal distribution.* The expected values and variance match so Y is normal.

Definition 19.2. CDF for a general normal RV. Given a normal RV X with mean μ and variance σ^2, we can relate it to a standard normal variable Y with mean 0 and variance 1 via

$$Y = \frac{X - \mu}{\sigma} \tag{50}$$

The CDF for the standard normal variable is denoted $\Phi(x)$

20 4/4/17: Lecture XVIII, Joint Distribution Functions

Example 20.1. We can relate a normal RV with mean μ and variance σ^2 and relate it to **standard normal** Y with mean 0 and variance 1 via

$$Y = \frac{X - \mu}{\sigma}$$

20.1 CDF for General Normal RV

Definition 20.1. The standard normal CDF, denoted

$$\Phi(x),$$

is a normal CDF with parameters

$$\mu = 0, \quad \sigma = 1.$$

Because this is the standard CDF for normal, it can be found in many sources.

We can map the general random variable back to a standard random variable to find

$$\begin{aligned} P(X \le x) &= P\left(\frac{X-\mu}{\sigma} \le \frac{x-\mu}{\sigma}\right) \\ &= P\left(Y \le \frac{x-\mu}{\sigma}\right) \\ &= F_Y\left(\frac{x-\mu}{\sigma}\right) \\ &= \Phi\left(\frac{x-\mu}{\sigma}\right) \end{aligned}$$

Memorize $\boxed{\Phi(1) = 0.84, \Phi(2) = 0.977}$!!!

Problem 20.1. Given $\mu = 60$, $\sigma = 20$ for normal RV, find the

$$P(X \geq 80)$$

Solution 20.1.

$$\begin{aligned} P(X \geq 80) &= 1 - P(X \leq 80) \\ &= 1 - \Phi\left(\frac{80-60}{20}\right) \\ &= 1 - 0.84 \\ &= \boxed{0.16}. \end{aligned}$$

20.2 Joint Distributions

- **Joint PDF:** Imagine a space of X and Y, what is the probability that the variable will fall in some range of X and Y.

$$P((X,Y) \in R) = \iint_{(x,y)\in R} \underbrace{f_{X,Y}(x,y)}_{\text{joint PDF}} \,\mathrm{d}x \,\mathrm{d}y \tag{51}$$

- **Joint CDF:**

$$\begin{aligned} F_{X,Y}(x,y) &= P(X \leq x \text{ and } Y \leq y) \\ &= \int_{-\infty}^{x} \mathrm{d}x' \int_{-\infty}^{y} f_{X,Y}(x',y')\,\mathrm{d}y' \\ f_{X,Y}(x,y) &= \frac{\partial^2 F_{X,Y}(x,y)}{\partial x \partial y} \end{aligned} \tag{52}$$

Example 20.2. Let's consider the product of two normal distributions. Call these two normal distributions X and Y.

Define the joint PDF as

$$\begin{aligned} f_{X,Y}(x,y) &= f_X(x) f_Y(y) \\ &= \frac{1}{2\pi\sigma_x\sigma_y} \exp\left[-\frac{1}{2}\left(\frac{(x-\mu_x)^2}{\sigma_x^2} + \frac{(y-\mu_y)^2}{\sigma_y^2}\right)\right] \end{aligned}$$

We see that the 3D visualization shows a hump near the μ_X, μ_Y. We can also visualize using **level curves**, where contours are drawn for each $f_{X,Y}(x,y)$ value.

Let us consider the marginal PDF, which is defined similar to the discrete marginal PDF,

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,\mathrm{d}y \tag{53}$$

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,\mathrm{d}x \tag{54}$$

Let us define the other discrete functions we had in the continuous land. Expected Value,

$$\mathbb{E}[g(X,Y)] = \int_{-\infty}^{\infty} \mathrm{d}x \int_{-\infty}^{\infty} \mathrm{d}y\, g(x,y) F_{X,Y}(x,y)$$

We still have linearity of expectations,

$$\mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbb{E}[X_i] \tag{55}$$

Example 20.3. Suppose X and Y are uniformly distributed for $X \in [a, b]$ and $Y \in [c, d]$, so that

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

In order to find q, we normalize,

$$\begin{aligned}
1 &= \int_{-\infty}^{\infty} \mathrm{d}x \int_{-\infty}^{\infty} \mathrm{d}y \, f_{X,Y}(x, y) \\
&= \int_{a}^{b} \mathrm{d}x \int_{c}^{d} \mathrm{d}y \, f_{X,Y}(x, y) \\
&= \int_{-\infty}^{\infty} \mathrm{d}x \int_{-\infty}^{\infty} \mathrm{d}y \, q \\
&= \int_{-\infty}^{\infty} \mathrm{d}x \, (b - a)q \\
&= (b - a)(d - c)q
\end{aligned}$$

so we have that

$$q = \frac{1}{(b - a)(d - c)} = \frac{1}{\text{Area}}$$

Example 20.4. Find the marginal PDF of the joint PDF. We will do this for $f_X(x)$

$$\begin{aligned}
f_X(x) &= \int_{-\infty}^{\infty} f_{X,Y}(x, y) \, \mathrm{d}y \\
&= \int_{c}^{d} q \, \mathrm{d}y \\
&= q(d - c) \\
&= \frac{1}{(b - a)(d - c)}(d - c) \\
&= \frac{1}{b - a}
\end{aligned}$$

Example 20.5. Find the expected value of uniform $\mathbb{E}[X]$

$$\begin{aligned}
\mathbb{E}[X] &= \int_{-\infty}^{\infty} \mathrm{d}x \int_{-\infty}^{\infty} \mathrm{d}y \, x \cdot f_{X,Y}(x, y) \\
&= \int_{a}^{b} (d - c)xq \, \mathrm{d}x \\
&= (d - c)q \left[\frac{1}{2}\right]_{a}^{b} \\
&= \frac{a + b}{2}
\end{aligned}$$

We can reason that

$$\mathbb{Y} = \frac{c + d}{2}.$$

Problem 20.2. The joint PDF is

X	Y	$f_{X,Y}(x, y)$
$1 \le x \le 2$	$1 \le y \le 4$	$\frac{1}{4}$
$2 \le x \le 3$	$2 \le y \le 3$	$\frac{1}{4}$

Rip typing. Did the problems by hand.

21 4/6/17: Lecture XIX, Conditioning and Continuous RVs

Definition 21.1. The **Expected Value** for a PDF is

$$\mathbb{E}[g(X,Y)] = \int_{-\infty}^{\infty} \mathrm{d}x \int_{-\infty}^{\infty} \mathrm{d}y\, g(x,y) f_{X,Y}(x,y) \tag{56}$$

Example 21.1. Some examples of $g(X,Y)$

$$\begin{aligned} g_1(X,Y) &= cX + bY \\ g_2(X,Y) &= kXY \\ g_3(X,Y) &= e^{X+Y} \end{aligned}$$

A crucial point here is that the **PDF stays the *same***, despite the function g that we use to act on our RVs.

Example 21.2. Linearity of Expectations still holds,

$$\mathbb{E}\left[\sum_{i=0}^{n} c_i X_i\right] = \sum_{i=0}^{n} c_i \mathbb{E}[X_i]$$

21.1 Conditioning on an Event

Suppose we have an event A with $P(A) > 0$. Then we define the **conditional PDF**

$$\int_B f_{X|A}(x)\,\mathrm{d}x = P(X \in B|A) \tag{57}$$

These PDFs must abide by the normalization condition.

Now suppose that $A = \{X \in S\}$,

$$\begin{aligned} P((X \in B)|(X \in S)) &= \frac{P((X \in B) \cap (X \in S))}{P(X \in S)} \\ &= \frac{\int_{S \cap B} f_X(x)\,\mathrm{d}x}{P(X \in S)} \end{aligned}$$

Then we can say

$$f_{X|A}(x) = \begin{cases} \frac{f_X(x)}{P(X \in S)} \\ 0, \text{otherwise} \end{cases} \tag{58}$$

notice this is the same idea as the other conditional definition—we scale based on the region we are working in.

Problem 21.1. Suppose we have the exponential PDF. Let $A = \{T > y\}$.

- Compute $P(A)$
- Compute $f_{T|A}(t+y)$ and compare it to the original PDF.

Solution 21.1.

$$\begin{aligned} P(A) &= \int_y^{\infty} \beta e^{-\beta t}\,\mathrm{d}t \\ &= -e^{-\beta t}\Big|_y^{\infty} \\ &= \boxed{e^{-\beta y}} \end{aligned}$$

$$\begin{aligned} f_{T|A}(t+y) &= \frac{f_T(t+y)}{e^{-\beta}} \\ &= \frac{\beta e^{-\beta(t+y)}}{e^{-\beta y}} \\ &= \boxed{\beta e^{-\beta t}} \end{aligned}$$

which is surprisingly the same PDF as we originally started with.

So about the result that we got from the previous example,

$$f_{T|A}(t+y) = f_T(t), \quad t \geq 0$$

We call this result a "memoryless" result of the exponential RV.

Example 21.3. Suppose we have the probability of a server crashing to be an exponential RV. Suppose we turn the server on at $T = 0$, then we expect a crash at $\frac{1}{\beta}$.

If the server is still ok after 4 days, we expect a crash at $4 + \frac{1}{\beta}$.

21.2 Conditioning on a Random Variable

Start with two RVs X and Y, and consider

$$\begin{aligned} P(x \leq X \leq x+\delta | y \leq Y \leq y+\delta) &= \frac{P(x \leq X \leq x+\delta \cap y \leq Y \leq y+\delta)}{P(y \leq Y \leq y+\delta)} \\ &= \frac{f_{X,Y}(x,y)\delta x \delta y}{f_Y(y)\delta y} \\ &= \frac{f_{X,Y}(x,y)\delta x}{f_Y(y)} \\ &= f_{X|Y}(x|y)\delta x \\ \implies & \boxed{f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}} \end{aligned}$$

there are very small slivers that we are finding the joint PDF of. Notice that our result was in the form,

$$\text{Conditional} = \frac{\text{Joint}}{\text{Marginal}}$$

You can condition on multiple RVs. Similar to discrete variables, we can get joint distributions like the following,

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

NOTICE THIS IS DIFFERENT FROM,

$$f_X(x) = \int_{\infty}^{\infty} f_{X|Y}(x|y) f_Y(y) \, \mathrm{d}y$$

You can interpret $f_{X,Y}(x,y)$ as a 3-D figure, with two dimensions x, y representing the range the variables can go, and then f being a third dimension representing the actual PDF value. This helpls you figure out the Marginal PDFs pretty well.

22 4/11/17: Lecture XX, Continuous RVs: Independence and Bayes' Theorem

Qualitatively, the conditional PDF is "higher" than the original PDF because you still need normalization to hold. Thus, over a smaller interval of the original PDF, you still need the area to add up to 1.

For 2-D, you can imagine "walking along the line of the given condition".

22.1 Conditional Expectations

Since conditional PDFs are also PDFs, we can just compute the conditional expectation the same way.

$$\begin{aligned}\mathbb{E}[X|Y=y] &= \int_{-\infty}^{\infty} x \cdot f_{X|Y}(x|y)\,\mathrm{d}x \\ \mathbb{E}[X|A] &= \int_{-\infty}^{\infty} x \cdot f_{X|A}(x)\,\mathrm{d}x\end{aligned}$$

22.2 Total Probability

We are going to use the same idea that we used for discrete RVs.

We start from the definition of conditional probability,

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

and then integrate to get the marginal probability,

$$\begin{aligned}f_X(x) &= \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,\mathrm{d}x \\ &= \int_{-\infty}^{\infty} f_{X|Y}(x|y) f_Y(y)\,\mathrm{d}x\end{aligned}$$

22.3 Total Expectation

$$\begin{aligned}\mathbb{E}[X] &= \int_{-\infty}^{\infty} \mathbb{E}[X|Y=y] \cdot f_Y(y)\,\mathrm{d}y \\ &= \int_{-\infty}^{\infty} \mathrm{d}y \left[\int_{-\infty}^{\infty} \mathrm{d}x\, x \cdot f_{X|Y}(x|y)\right] f_Y(y) \\ &= \int_{-\infty}^{\infty} \mathrm{d}y \int_{-\infty}^{\infty} \mathrm{d}x\,(x f_{X,Y}(x,y)) \\ &= \mathbb{E}[X].\end{aligned}$$

which is a nicer way to compute $\mathbb{E}[X]$.

Example 22.1. Suppose

$$f_{T|B}(t,\beta) = \beta e^{-\beta t}$$

First, we want to find $f_T(t)$.

$$\begin{aligned}f_T(t) &= \int_{-\infty}^{\infty} f_{T|B} t | b f_B(\beta)\,\mathrm{d}\beta \\ &= \int_{\beta_1}^{\beta_2} \beta e^{-\beta t} \cdot \frac{1}{\beta_2 - \beta_1}\,\mathrm{d}\beta\end{aligned}$$

notice this is a very nontrivial integral.

We want to find

$$\mathbb{E}[T]$$

which we should do without using $f_T(t)$ since we can't even compute the integral.

So we do

$$\begin{aligned}\mathbb{E}[T] &= \int_{-\infty}^{\infty} \mathbb{E}[T|B=\beta] f_B(\beta)\,\mathrm{d}\beta \\ &= \int_{\beta_1}^{\beta_2} \frac{1}{\beta}\frac{1}{\beta_2-\beta_1}\,\mathrm{d}\beta \\ &= \frac{1}{\beta_2-\beta_1}\int_{\beta_1}^{\beta_2} \frac{1}{\beta}\,\mathrm{d}\beta \\ &= \boxed{\frac{1}{\beta_2-\beta_1}\log\left(\frac{\beta_2}{\beta_1}\right)}\end{aligned}$$

22.4 Independence

22.4.1 PDFs

In the past, we have done this with PMFs. In terms of the joint PDF:

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

and in terms of the condition PDF

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

or that "conditional = marginal"

22.4.2 CDFs

In terms of the joint PDF,

$$\begin{aligned}P((X \le x)\cap(Y \le y)) &= \int_{-\infty}^{x}\mathrm{d}x'\int_{-\infty}^{y}\mathrm{d}y'\, f_{X,Y}(x',y') \\ &= \int_{-\infty}^{x}\mathrm{d}x'\int_{-\infty}^{y}\mathrm{d}y'\, f_X(x')f_Y(y') \\ &= \int_{-\infty}^{x} f_X(x')\,\mathrm{d}x'\int_{-\infty}^{y} f_Y(y')\,\mathrm{d}y' \\ &= P(X \le x)P(Y \le y)\end{aligned}$$

and in terms of the condition PDF

$$F_{X,Y}(x,y) = F_X(x)F_Y(y)$$

22.4.3 Expected Value and Variance

The expected value,

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

and variance,

$$\mathrm{Var}[X+Y] = \mathrm{Var}[X] + \mathrm{Var}[Y]$$

22.5 Bayes' Theorem

Reminder of Bayes' theorem,

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

We have that

- $P(A|B)$: posterior probability

- $P(B|A)$: likelihood
- $P(A)$: prior
- $P(B)$: evidence

So the application to inference is that

1. We start with a prior on A
2. We then observe B
3. We compute the likelihood $P(B|A)$ of seeing B given A
4. Compute $P(B)$ (use total probability)
5. The result is $P(A|B)$, the "updated" probability for A

22.6 Continuous Bayes' Theorem

We start with our expressions for conditional probabilities,

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

Strategy is

- **Compute** the likelihood $f_X|Y(x,y)$
- **Choose** a prior $f_Y(y)$
- **Integrate** to find the evidence,
$$f_X(x) = \int_{-\infty}^{\infty} f_{X|Y}(x|y)f_Y(y)\,\mathrm{d}y$$

Example 22.2. We have

- Prior $f_B(\beta) = \begin{cases} 2, & 1 \le \beta \le \frac{3}{2} \\ 0, & \text{otherwise} \end{cases}$
- Likelihood $f_{T|B}t|\beta = \beta e^{-\beta t}$
- Evidence
$$f_T(t) = \int_{-\infty}^{\infty} f_{T|B}(t|\beta)f_B(\beta)\,\mathrm{d}\beta$$
$$= 2\int_1^{\frac{3}{2}} \beta e^{-\beta t}\,\mathrm{d}\beta$$
the point here is that the evidence does not event on β
- Apply Bayes' theorem to find the posterior
$$f_{B|T}(\beta|t) = \frac{f_{T|B}(t|\beta)f_B(\beta)}{f_T(t)}$$

We want to find the mode of $f_{B|T}(\beta|t)$. We do this through differentiation,

$$\frac{d}{d\beta}f_{B|T}(\beta|t) = \frac{d}{d\beta}$$
$$= -\frac{2}{f_T(t)}(e^{-\beta t} - \beta t e^{-\beta t})$$

23 4/13/17: Lecture XXI, New Random Variables

The idea is that we want to create new RVs based on a function of other RVs.

23.1 Sum of 2 RVs

Suppose we have X and Y as two independent random variables, and that X is independent of Y.

Suppose we have a new RV Z defined as,

$$Z = X + Y.$$

We consider,

$$f_{Z|Y}(z|y) = f_X(z-y)$$

this because if we are given a Y, we just need to figure out what X has to be in order to satisfy the equation; $x = z - y$.

Use total probability,

$$f_Z(z) = \int_{-\infty}^{\infty} f_{Z|Y}(z|y) f_Y(y)\,\mathrm{d}y$$

Substitute in our previous result,

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(z-y) f_Y(y)\,\mathrm{d}y \tag{59}$$

this is the **convolution integral**. We are "tying together" two functions.

Example 23.1. Suppose X is uniformly distributed from 0 to a, and Y is uniformly distributed from 0 to b. Assume $a < b$ (WLOG).

PDFs:

- $f_X(x) = \frac{1}{a}, \quad 0 \le x \le a$
- $f_Y(y) = \frac{1}{b}, \quad 0 \le y \le b$

Let $Z = X + Y$. Z is nonzero for $0 \le Z \le a + b$.

Let us find the PDF of Z using the convolution integral from equation 59.

$$\begin{aligned} f_Z(z) &= \int_{-\infty}^{\infty} \underbrace{f_X(z-y)}_{0 \le z-y \le a} \underbrace{f_Y(y)}_{0 \le y \le b}\,\mathrm{d}y && \text{(the first condition is } z-a \le y \le z\text{)} \\ &= \int_{\max(0,z-a)}^{\min(b,z)} \frac{1}{a}\cdot\frac{1}{b}\,\mathrm{d}y \\ &= \boxed{\frac{1}{ab}\left[\min(b,z) - \max(0,z-a)\right]}. \end{aligned}$$

kinda disgusting tbh.

To finish our example, let us consider our PDF in three regions,

1. $0 \le z \le a$: $\frac{1}{ab}[z-0] = \frac{z}{ab}$. Some values, $f_Z(0) = 0, f_Z(a) = \frac{1}{b}$.
2. $a \le z \le b$: $\frac{1}{ab}\big[z - (z-a)\big] = \frac{a}{ab} = \frac{1}{b}$. PDF is constant in this region.
3. $b \le z \le a+b$: $\frac{1}{ab}\big[b - (z-a)\big] = \frac{a+b-z}{ab}$. Some values, $f_Z(b) = \frac{1}{b}, f_Z(a+b) = 0$.

If we graph this, it is a plateau (trapezoid) that is symmetric.

Example 23.2. Let us find $Z = X + Y$ where X and Y are exponential distributions with parameters $\beta_x, beta_y$ respectively.

Then to find our PDF,

$$\begin{aligned}
f_Z(z) &= \int \beta_x \beta_y e^{-\beta_x(z-y)} e^{-\beta_y y} \, \mathrm{d}y \\
&= \beta_x \beta_y e^{-\beta_x z} \int_0^z e^{-(\beta_y - \beta_x)y} \, \mathrm{d}y \\
&= \beta_x \beta_y e^{-\beta_x z} \left[-\frac{e^{-(\beta_y - \beta_x)y}}{\beta_y - \beta_x} \Bigg|_0^z \right] \\
&= \frac{\beta_x \beta_y}{\beta_y - \beta_x} e^{-\beta_x z} \left[1 - e^{-(\beta_y - \beta_x)z} \right] \\
&= \boxed{\frac{\beta_x \beta_y}{\beta_y - \beta_x} \left[e^{-\beta_x z} - e^{-\beta_y z} \right]}
\end{aligned}$$

You can just use the fact that $\mathbb{E}[\beta e^{-\beta X}] = \frac{1}{\beta}$, so then our expected value is just (simplifying first)

$$\begin{aligned}
\frac{\beta_x \beta_y}{\beta_y - \beta_x} \left[e^{-\beta_x z} - e^{-\beta_y z} \right] &= \frac{1}{\beta_y - \beta_x} \left[\beta_x \beta_y e^{-\beta_x z} - \beta_x \beta_y e^{-\beta_y z} \right] \\
&\implies \frac{1}{\beta_y - \beta_x} \left[\beta_y \frac{1}{\beta_x} - \beta_x \frac{1}{\beta_y} \right] \\
&= \boxed{\frac{1}{\beta_y} + \frac{1}{\beta_x}}
\end{aligned}$$

23.2 Functions of a Random Variable

We notice that it is easier to consider probabilities than the actual densities.

Example 23.3. $Y = e^X$, where X, Y are RVs.

Relate the CDF for Y to the CDF for X,

$$\begin{aligned}
F_Y(y) &= P(Y \le y) \\
&= P(e^X \le y) \\
&= P(X \le \log(y)) \\
&= F_X(\log(y))
\end{aligned}$$

We differentiate the CDF to get the PDF,

$$\begin{aligned}
f_Y(y) &= \frac{d}{dy} F_Y(y) \\
&= \frac{d}{dy} F_X(\log(y)) \\
&= F_x'(\log(y)) \cdot \frac{1}{y} \\
&= \frac{f_X(\log(y))}{y}
\end{aligned}$$

at the last step, $\frac{1}{y}$ is the *factor*, which changes depending on which change of variables we make.

Example 23.4. Let's apply our result above to an exponential random variable X, and define $Y = e^X$.

$$\begin{aligned} f_Y(y) &= f_X(\log(y)) \cdot \frac{1}{y} \\ &= \beta e^{-\beta \log(y)} \cdot \frac{1}{y} \\ &= \frac{\beta}{y} y^{-\beta} \\ &= \beta y^{-\beta-1} \end{aligned}$$

What is the range of Y? $x \geq 0, y = e^x, \Longrightarrow y \geq 1$.

So

$$f_Y(y) = \beta$$

24 4/18/2017

24.1 Moment Generating Function

Definition 24.1. We define the **Moment Generating Function (MGF)** as

$$M_X(t) = 1 + t\mathbb{E}[X] + \frac{t^2}{2}\mathbb{E}[X^2] + \cdots = \boxed{\sum_{i=0}^{\infty} \frac{t^i}{i!}\mathbb{E}[X^i]} \tag{60}$$

Notice that this is also just

$$M_X(t) = \mathbb{E}\left[e^{Xt}\right]$$

If we differentiate w.r.t t several times,

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} M_X(t)\bigg|_{t=0} &= \mathbb{E}[X] \\ \frac{\mathrm{d}^2}{\mathrm{d}t^2} M_X(t)\bigg|_{t=0} &= \mathbb{E}[X^2] \end{aligned}$$

***If two distributions have the same MGF, then they're actually the same distribution.

Example 24.1. Suppose we are given an MGF,

$$M_X(t) = 1 - p + pe^t,$$

Let's compute,

- Mean: $\mathbb{E}[X] = \frac{\mathrm{d}}{\mathrm{d}t} M_X(t)\Big|_{t=0} = pe^t\big|_{t=0} = \boxed{p}$
- Variance: $\mathbb{E}[X^2] = \frac{\mathrm{d}^2}{\mathrm{d}t^2} M_X(t)\Big|_{t=0} = pe^t\big|_{t=0} = p$

  So $\mathrm{Var}[X] = E[X^2] - E[X]^2 = p - p^2 = \boxed{p(1-p)}$.

We suspect this is the MGF for the Bernoulli RV?

Problem 24.1. Compute the mean and variancce from the following MGF,

$$M_X(t) = e^{\lambda(e^t - 1)}$$

Solution 24.1. Rip doing derivatives in head,

- Mean: $\mathbb{E}[X] = \frac{\mathrm{d}}{\mathrm{d}t} M_X(t)\Big|_{t=0} = e^{\lambda(e^t-1)} \cdot (\lambda e^t)\Big|_{t=0} = \boxed{\lambda}$
- Variance: $\mathbb{E}[X^2] = \frac{\mathrm{d}^2}{\mathrm{d}t^2} M_X(t)\Big|_{t=0} = \lambda e^{\lambda(e^t-1)+t} \cdot (\lambda e^t + 1)\Big|_{t=0} = \lambda(\lambda+1)$

  So $\mathrm{Var}[X] = E[X^2] - E[X]^2 = \lambda^2 + \lambda - \lambda^2 = \boxed{\lambda}$.

We should suspect this is the MGF for Poisson.

24.2 Computing the MGF

Although I noticed this earlier let's just do it again. Observe that,

$$\begin{aligned} M_X(t) &= 1 + t\mathbb{E}[X] + \frac{t^2}{2!}\mathbb{E}[X^2] + \cdots \\ &= \sum_{i=0}^{\infty} \frac{t^i}{i!}\mathbb{E}[X^i] \\ &= \boxed{\mathbb{E}\left[e^{tX}\right]} \end{aligned}$$

now that we have this, we can find $\mathbb{E}[X^n]$ in general.

For a discrete RV, we see that

$$M_X(t) = \mathbb{E}\left[e^{tX}\right] = \sum_k p_X(k)e^{tk}$$

and for continuous,

$$M_X(t) = \mathbb{E}\left[e^{tX}\right] = \int_{-\infty}^{\infty} f_X(x)e^{tx}\,\mathrm{d}x$$

However, we know that most integrals cannot be computed...so what do we do?

One strategy that we have implemented a couple of times is doing a differentiation trick.

Example 24.2. Let us do the Poisson MGF. We will start with the def. of MGF,

$$\begin{aligned} M_X(t) &= \mathbb{E}\left[e^{tX}\right] \\ &= \sum_{k=0}^{\infty} e^{-\lambda}\frac{\lambda^k}{k!}e^{tk} \\ &= e^{-\lambda}\sum_{k=0}^{\infty} \frac{(\lambda e^t)^k}{k!} \\ &= e^{-\lambda}e^{\lambda e^t} \\ &= \boxed{e^{\lambda(e^t-1)}}. \end{aligned}$$

Problem 24.2. Compute the MGF for the exponential PDF.

Solution 24.2.

$$\begin{aligned}
M_X(t) &= \mathbb{E}\left[e^{tX}\right] \\
&= \int_{-\infty}^{\infty} f_X(x)e^{tx}\,\mathrm{d}x \\
&= \int_{0}^{\infty} (\lambda e^{-\lambda x})e^{tx}\,\mathrm{d}x \\
&= \lambda \int_{0}^{\infty} e^{(t-\lambda)x}\,\mathrm{d}x \\
&= \frac{\lambda}{t-\lambda}\left[e^{(t-\lambda)x}\right]\Big|_0^{\infty} \\
&= \boxed{\begin{cases} \frac{\lambda}{\lambda-t}, & t < \lambda \\ \infty, & t \geq \lambda \end{cases}}
\end{aligned}$$

The MGF is not defined everywhere.

24.3 Properties of MGF

- Even though some MGFs are not well-defined everywhere, as long as they are defined from some $t \in (-\delta, \delta)$, it is ok, since we only need to be able to take derivatives at $t = 0$.
- The MGF determines all the moments of an RV. This is *almost* equivalent to determining the PDF of the RV.
- Some distributions don't have a finite MGF.

24.4 Manipulating MGFs

Example 24.3. We are going to rescale an RV. $Z = aY$.

$$\begin{aligned}
M_Z(t) &= \mathbb{E}\left[e^{tZ}\right] \\
&= \mathbb{E}\left[e^{atY}\right] \\
&= M_Y(at) && \text{(Grouping together } at\text{)}
\end{aligned}$$

Example 24.4. Consider $Z = X + Y$.

$$\begin{aligned}
M_Z(t) &= \mathbb{E}\left[e^{tZ}\right] \\
&= \mathbb{E}\left[e^{t(X+Y)}\right] \\
&= \mathbb{E}\left[e^{tX}e^{tY}\right]
\end{aligned}$$

If X, Y are independent, then

$$\begin{aligned}
M_Z(t) &= \mathbb{E}\left[e^{tX}e^{tY}\right] \\
&= \mathbb{E}\left[e^{tX}\right]\mathbb{E}\left[e^{tY}\right] \\
&= M_X(t)M_Y(t)
\end{aligned}$$

So adding two RVs is the same as multiplying MGFs.

To illustrate why the adding property of MGFs is so important,

Example 24.5. Let us consider the Bernoulli RV. Since they are independent, we can find Binomial by,

$$\begin{aligned}\text{Binomial RV} &= \sum_{i=0}^{n} X_i \\ &\Longrightarrow \prod_{i=0}^{n} M_X(t) \\ &= \boxed{(1-p+pe^t)^n}\end{aligned}$$

Recall that last lecture, we saw that we can do convolution integrals for the sum of two RVs. However, this is difficult since there are more integrals to compute. Instead, we can use MGFs that are much easier.

25 4/25/17: Lecture XIII, The Central Limit Theorem

25.1 Motivation 1: Sample Statistics

Definition 25.1. The **sample mean** is defined to be,

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_n \tag{61}$$

The

- $\mathbb{E}[X] = \frac{1}{n}\mathbb{E}[X_i]$
- $\text{Var}[X] = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$

The PDF of X is a

25.2 The Normal Distribution?

It looks like as we increase the number of samples, the PDF becomes closer and closer to the Normal distribution.

25.3 Motivation 2: Binomial Distribution

Recall that the binomial RV is a sum of n independent and identically distributed Bernoulli RVs. What happens if we fix p and take $n \to \infty$?

It turns out, we get something that looks like the normal distribution...again.

Problem 25.1. We are considering $X_1, X_2, \ldots, X_n$ as independent and identically distributed RVs.

	Sample Mean	**Sum**
Expected Value	$\mathbb{E}[\overline{X}] = \mathbb{E}[X_i]$	$E[\sum X_i] = nE[X_i]$
Variance	$\text{Var}[\overline{X}] = \frac{\text{Var}[X_i]}{n}$	$\text{Var}[\sum X_i] = n\,\text{Var}[X_i]$

Table 3: Expected values and variance

Theorem 6. *(The Central Limit Theorem). Let $X_1, X_2, \ldots, X_n$ be a sequence of independent identically distributed (i.i.d.) RVs with finite mean and finite variance σ^2. Introduce,*

$$Z_n = \frac{\sum_{i=1}^{n}(X_i - \mu)}{\sigma\sqrt{n}}$$

In the limit as $n \to \infty$, the CDF for Z_n approaches the CDF of the standard normal RV.

Example 25.1. What is the expected value and variance of Z_n as stated above in the Central Limit Theorem?

$$\begin{aligned}
\mathbb{E}[Z_n] &= \frac{1}{\sigma\sqrt{n}}\big(n\mathbb{E}[X_i] - n\mu\big) \\
&= \frac{\sqrt{n}(\mathbb{E}[X_i] - \mu)}{\sigma} \\
&= \boxed{0} && (\mathbb{E}[X] = \mu)
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{Var}[Z_n] &= \frac{1}{\sigma^2 n}(n\,\mathrm{Var}[X_i]) \\
&= \frac{\mathrm{Var}[X_i]}{\sigma^2} \\
&= \boxed{1} && (\mathrm{Var}[X] = \sigma^2)
\end{aligned}$$

If X_i are normal, then Z_n is a sum of normals, which means Z_n itself is normal. From our results above, we then know that Z_n is indeed a standard normal distribution.

25.4 Proof of Central Limit Theorem with MGFs

Our goal is to compare the MGF of Z_n with the MGF for the standard normal RV. If their MGFs match, then we can conclude the distributions match.

The MGF for the standard normal RV is,

$$\begin{aligned}
M_N(t) &= \mathbb{E}\left[e^{tX}\right] \\
&= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{tx}\cdot e^{-x^2/2}\,\mathrm{d}x \\
&= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\frac{1}{2}(x^2-2tx)}\,\mathrm{d}x \\
&= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\frac{1}{2}(x-t)^2}e^{\frac{t^2}{2}}\,\mathrm{d}x \\
&= \frac{1}{\sqrt{2\pi}}e^{\frac{t^2}{2}}\int_{-\infty}^{\infty} e^{-\frac{1}{2}(x-t)^2}\,\mathrm{d}x \\
&= \frac{1}{\sqrt{2\pi}}e^{\frac{t^2}{2}}\cdot(\sqrt{2\pi}) && \text{(Do u-sub and scaling sub.)} \\
&= \boxed{e^{\frac{t^2}{2}}}
\end{aligned}$$

Now the idea is to standardize all the X_i by introducing Y_i,

$$Y_i = \frac{X_i - \mu}{\sigma}, \quad Z_n = \sum_{i=1}^{n}\frac{Y_i}{\sqrt{n}}$$

Now we try to relate the MGF of Z_n to the MGF of Y_i,

$$\begin{aligned}
M_{Z_n}(t) &= \prod_{i=1}^{n} M_{Y_i/\sqrt{n}}(t) \\
&= \left[M_{Y_1}\left(\frac{t}{\sqrt{n}}\right)\right]^n && \text{(MGF prod-sum prop., constant rule)} \\
\implies \log\big(M_{Z_n}(T)\big) &= n\log\left(M_{Y_1}\left(\frac{t}{\sqrt{n}}\right)\right) && \text{(take log of both sides)}
\end{aligned}$$

Ok now all we have to do is show that as $n \to \infty$, the generating function for $M_{Z_n}(t)$ becomes the standard normal.

$$\lim_{n\to\infty} \log\big(M_{Z_n}(t)\big) = \lim_{n\to\infty} \underbrace{n}_{\text{goes to } \infty \text{ :(}} \log\left(M_{Y_1}\left(\frac{t}{\sqrt{n}}\right)\right)$$

Quite unfortunate...let's try l'Hopital's rule,

$$\begin{aligned}\lim_{n\to\infty} \log\big(M_{Z_n}(t)\big) &= \lim_{n\to\infty} \frac{\partial n \log\big(M_{Y_1}(t/\sqrt{n})\big)}{\partial n (n^{-1})} \\ &= \lim_{n\to\infty} \frac{\big(M_{Y_1}(t/\sqrt{n})\big)^{-1} M'_{Y_1}(t/\sqrt{n}) \left[-\frac{1}{2} t n^{-3/2}\right]}{-n^{-2}} \\ &= \lim_{n\to\infty} \frac{t}{2} \frac{M'_{Y_1}(t/\sqrt{n})}{n^{-1/2}} = \frac{0}{0}\end{aligned}$$

>: (
Not looking good.
Let's try l'Hopital AGAIN,

$$\begin{aligned}\lim_{n\to\infty} \log\big(M_{Z_n}(t)\big) &= \frac{t}{2} \lim_{n\to\infty} \frac{\partial n M'_{Y_1}(t/\sqrt{n})}{\partial n n^{-1/2}} \\ &= \frac{t}{2} \lim_{n\to\infty} \frac{M''_{Y_1}(t/\sqrt{n}) \left[-\frac{1}{2} t n^{-3/2}\right]}{-\frac{1}{2} n^{-3/2}} \\ &= \frac{t^2}{2} \lim_{n\to\infty} M''_{Y_1}(t/\sqrt{n}) \\ &= \frac{t^2}{2} & (\mathbb{E}[Y_1^2] = 1 \text{ from standard normal})\end{aligned}$$

thus we have that

$$M_{Z_n}(t) = e^{t^2/2} = M_N(t)$$

which proves the central limit theorem.

26 4/27/17: Lecture XIV, The CLT: Applications and Failures

26.1 Recap of the CLT

Let $X_1, X_2, \ldots, X_n$ be a sequence of independent, identically distributed RVs with finite mean and variance.

We defined

$$Z_n = \frac{\sum_{i=1}^n (X_i - \mu)}{\sigma\sqrt{n}} \tag{62}$$

As $n \to \infty$, the distribution of Z_n approaches a standard normal distribution.

The average of the X_i's,

$$\overline{X_n} = \frac{1}{n}\sum_{i=1}^n X_i = \frac{\sigma}{\sqrt{n}} Z_n + \mu$$

We notice that as $n \to \infty$, the distribution of $\overline{X_n}$ is normal, and

$$\mathbb{E}[\overline{X_n}] = \mu, \quad \mathrm{Var}[\overline{X_n}] = \frac{\sigma^2}{n}$$

And if we consider the sum of the X_i's,

$$Y_n = \sum_{i=1}^{n} X_i = \sigma\sqrt{n} Z_n + n\mu$$

as $n \to \infty$, the distribution for Y_n is normal,

$$\mathbb{E}[\overline{Y_n}] = n\mu, \quad \mathrm{Var}[\overline{Y_n}] = n\sigma^2$$

26.2 When can we use the CLT?

- It requires that the distribution for X_i has **finite mean** and **finite variance**
- It describes the distribution for **sums** and **averages**.
- It's useful when the number of things being combined is **large**.
- An example: "noise" is the sum of many different random effects, each with their own peculiar distribution. But if the noise is the **sum** of those effects, the noise will be approximately normally distributed.

26.3 How can we use the CLT?

- The CLT is most useful when we need the full distribution. It doesn't tell us anything new about the mean and variance.
- Examples of questions the require the distribution: what is the probability that the sum is less than x, what is the probability that the average is between y and z.
- We can answer those kinds of questions with the standard normal CDF, $\Phi(x)$.

Example 26.1. We are loading 100 boxes into a plane. Assuming the weights of the boxes are distributed uniformly between 5 and 50 pounds and that the capcity of the plane is 3000 pounds, what is the probability that we exceed the plane's capacity.

For this example, it is hard to get the exact distribution of the sum of the weights, but easy to get the mean and variance of the weight of one box (using uniform PDF definitions),

$$\mathbb{E}[X] = \frac{50+5}{2} = 27.5, \quad \mathrm{Var}[X] = \frac{(50-5)^2}{12} \approx 170.$$

The weights are i.i.d., and have finite mean and variance. We can approximate the PDF for their sum, Y, as

$$f_Y(y) \text{ is normal}, \quad \mu = n\mathbb{E}[X] = n \cdot 27.5, \quad \sigma_Y^2 = n\,\mathrm{Var}[X] \approx n \cdot 170,$$

We can then approximate,

$$\begin{aligned} P(Y \leq 3000) &= P\left(\frac{Y - \mathbb{E}[Y]}{\sigma_Y} \leq \frac{3000 - 100 \cdot 27.5}{\sqrt{100 \cdot 170}}\right) \\ &\approx \Phi(1.92) \\ &\approx \boxed{0.97} \qquad \text{(3\% chance off overloading the plane)} \end{aligned}$$

Problem 26.1. We are given

$$n = 25, \quad \mathbb{E}[X] = 20\,\text{min}, \quad \sigma_X = 4\,\text{min},$$

We have a block of 450 min, and we need to grade 25 exams. Can we do it?

Solution 26.1. We assume a normal distribution. Let

$$Y = \sum_{i=1}^{25} X_i$$

We calculate,

$$\begin{aligned} \mu = \mathbb{E}[Y] = 25 \cdot \mathbb{E}[X_i] = 500 \\ \sigma_Y^2 = \text{Var}[Y] = 25 \cdot \text{Var}[X_i] = 400 \end{aligned}$$

Now we need

$$\begin{aligned} P(Y \leq 450\,\text{min}) &= P\left(\frac{Y - \mathbb{E}[Y]}{\sigma_Y} \leq \frac{450 - 25 \cdot 20}{\sqrt{25} \cdot 4}\right) \\ &= P\left(\frac{Y - \mathbb{E}[Y]}{\sigma_Y} \leq \frac{-50}{20}\right) \\ &= P\left(\frac{Y - \mathbb{E}[Y]}{\sigma_Y} \leq \frac{-50}{20}\right) \\ &= \Phi(-2.5) \\ &\approx \boxed{0.6\%}. \end{aligned}$$

26.4 Rules of thumb

- **When can you use CLT?** It turns out, almost always! With a little work we can relax the requirement that the RVs are identically distributed. The requirements of finite mean and variance are met by most real world applications.
- You can't use CLT when you have infinite mean and variance.

Example 26.2. The power distribution,

- PDF: $f_X(x) = (\alpha - 1)x^{-\alpha}$
- Mean: $\mathbb{E}[X] = \begin{cases} \frac{\alpha-1}{\alpha-2}, & \alpha \geq 2 \\ \infty, & \text{otherwise} \end{cases}$
- Second Moment: $\mathbb{E}[X^2] = \begin{cases} \frac{\alpha-1}{\alpha-3}, & \alpha > 3 \\ \infty, & \text{otherwise} \end{cases}$

We look at the power law distributions for $\alpha = 3.5$ (finite mean and var.) and $\alpha = 2.5$ (finite mean, infinite var.) We see that for the infinite variance distribution, it remains skewed no matter how big n is.

26.5 The Law of Large Numbers

Given a series of iid RVs X_i with finite $\mathbb{E}[X_i]$,

$$P\left(\lim_{n\to\infty} \overline{X_n} = \mu\right) = 1$$

This is the strong law of large numbers.

27 5/2/17: Lecture XXV, Generating Samples

27.1 Rejection Sampling

We want to pick points at random in a box bounding a PDF, and keep the ones that fall under the curve.

1. Identify the range for x and the range for $f_X(x)$
2. Draw two sets of n points. X_i are uniformly distributed on (a, b), Y_i are uniformly distributed on $(0, c)$
3. For each pair of points (X_i, Y_i), check whether $Y_i \leq f_X(X_i)$. When this is true, keep X_i, otherwise throw it out
4. The X_i we keep should follow $f_X(x)$

Example 27.1. Pseudocode for doing rejection sampling,

```
xVals = (rangeX) * rand(#samples)
yVals = (rangeY) * rand(#samples)
keepers = (yVal <= f_X(xVal)) #boolean to filter out bad ones,
                              #use bitwise to do more
result = xVals[keepers]
```

Example 27.2. The code for a complete rejection sampling problem.

```
from numpy.random import *
import matplotlib.pyplot as plt

numSamples = 10000
xVals = 2*rand(numSamples)
yVals = 2*rand(numSamples)

#filter out the ones you want, keepers contains corresponding bools
keepers = (((yVals<=xVals) & (xVals>=0) & (xVals<1)) |
          ((yVals<=(2-xVals)) & (xVals>=1) & (xVals<=2)))

#plot random values
plt.subplot(311)
plt.scatter(xVals,yVals)

#plot filtered values
plt.subplot(312)
plt.scatter(xVals[keepers], yVals[keepers])

#plot histogram
plt.subplot(313)
plt.hist(xVals[keepers], range=(0,2), bins = 20, normed="True")
plt.show()
```

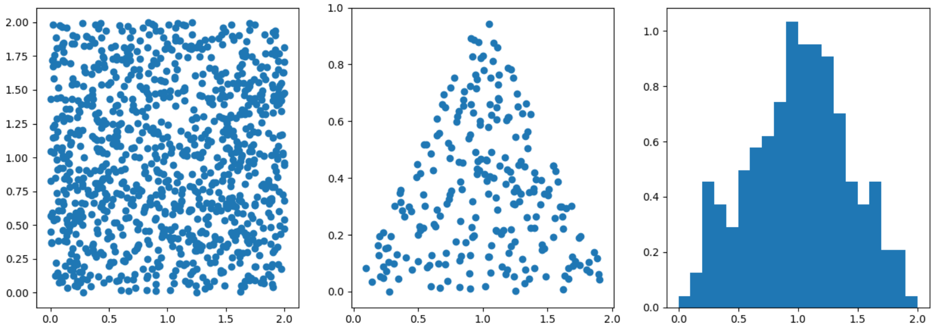

Figure 1: Random gen, filter, and histogram for thsi example

27.2 Inversion Sampling

A limitation of rejection sampling is that you need the PDF to fit in a box. This is not always the case. Another sampling we can use is inversion sampling, which is supported even for a PDF that extends to $\pm\infty$.

We need to introduce,

Definition 27.1. The **Quantile Function** is the inverse of the CDF,

$$Q_X(F_X(x)) = x, \quad F_X(Q_X(q)) = q \tag{63}$$

Intuitively, the quantile function tells you where you should go in order to cut out a particular fraction of the distribution.

Example 27.3. Q_X for an exponential distribution. Consider,

$$F_X(x) = 1 - e^{-\beta x}$$

then

$$\begin{aligned} q &= 1 - e^{-\beta x} \\ \implies e^{-\beta x} &= 1 - q \\ -\beta x &= \log(1-q) \\ x &= \boxed{-\frac{1}{\beta}\log(1-q) = Q_X(q)} \end{aligned}$$

1. Evaluate the CDF $F_X(x)$ of the distribution
2. Invert the CDF to find $Q_X(q)$
3. Generate random numbers uniformly distributed between 0 and 1
4. Apply Q_X to the uniform sample

So why does this work?

$$\begin{aligned} P(Q_X(q) \leq x) &= P(q \leq F_X(x)) && \text{(apply } F_X \text{ to both sides)} \\ &= F_X(x) && (P(q \leq y) = y \text{ if } q \text{ is uniformly distributed } [0,1)) \end{aligned}$$

so the CDF for $Q_X(q)$ matches $F_X(x)$ for the target distribution.

Example 27.4. Find the PDF of the power law distribution using the quantile method.

The power law is

$$f_X(x) = (\alpha - 1)x^{-\alpha}$$

We compute the CDF,

$$\begin{aligned} F_X(x) &= \int_1^x (\alpha - 1)y^{-\alpha}\,\mathrm{d}y \\ &= \frac{\alpha - 1}{1 - \alpha}\left[y^{-\alpha+1}\right]\Big|_1^x \\ &= 1 - x^{-\alpha+1} \end{aligned}$$

We can find the quantile function by

$$\begin{aligned} q &= 1 - x^{-\alpha+1} \\ \implies \boxed{(1-q)^{1/(1-\alpha)}} &= x = Q_X(q) \end{aligned}$$

Now we can do the inversion sampling,

```
from numpy.random import *
import matplotlib.pyplot as plt

numSamples = 10000
qVals = rand(numSamples) #from 0 to 1

QVals = (1-qVals)**(-1/1.5)

#plot CDF
plt.subplot(121)
plt.scatter(QVals,qVals)
plt.xlim([0,10])

#plot histogram
plt.subplot(122)
plt.hist(QVals, range=(0,10), bins = 20, normed=”True”)

plt.show()
```

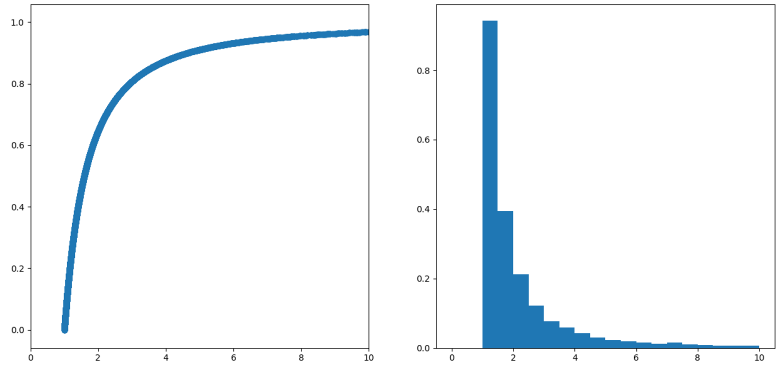

Figure 2: The CDF and PDF for this example