STAT410

Hari

Contents

Chapter 1

Introduction to Counting and Probability

1.1 Introduction

Office hours: MTH4107, Tuesday 3–4:30 (zoom), Friday 3–4:30 (in person)

1.2 Basic Counting

Definition 1 (Permutation)**.** A permutation of an n-element set is a rearrangement of the (distinct) elements in a specific order.

Definition 2 (k-Permutation)**.** A k-permutation is an arrangement of k elements from the n element set. The total number of k-permutations of an n-element set is

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

Example (Permutation)**.** How many ways can we seat 10 people in a row?

$$P(10,10) = \boxed{10!}$$

Example (k-Permutation)**.** How many ways can you create a committee with a president, vice president, and secretary (exactly 1 role)?

$$P(10,3) = 10 \cdot 9 \cdot 8$$

Definition 3 (Combination). The total number of ways to create a k-size subset from an n-element set is

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

This is also known as a binomial coefficient.

Theorem 1 (Binomial). For any positive integer n,

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

Proof 1. To get the coefficient of $x^k y^{n-k}$, we must "choose" k x terms and $n-k$ y terms. We pick the number of ways to have an x term, giving $\binom{n}{k}$. However, note that we will just pick the rest of the values for y, giving us $\binom{n-k}{n-k} = 1$. Therefore, coefficient of $x^k y^{n-k}$ is $\binom{n}{k}\binom{n-k}{n-k} = \binom{n}{k}$, and we see that the theorem is true. □

Example (Binomial). There are 10 cows, 9 pigs, 8 horses.

- How many ways can we pick 5 animals?
  Picking 5 animals gives $\binom{27}{5}$.
- How many ways can we pick 5 animals such that either we have 3 cows and 2 pigs or 2 cows and 3 pigs?
  We can either count cows and pigs, or pigs and cows. So in total we have $\binom{10}{3}\binom{9}{2} + \binom{10}{2}\binom{9}{3}$

Example (Permutation with Repetition). There are 10 kids and 4 types of candy. How many ways can we distribute the candy so exactly 3 children get type 1, 2 get type 2, 5 get type 3? The answer is $\binom{10}{3}\binom{7}{2}\binom{5}{5}$. We can think of it as permuting 3 As, 2 Bs, and 5 Cs. This is a permutation **with repetition**.

Definition 4 (Permutation with repetition). Suppose there are k distinct objects and object i occurs a_i times. Assume $a_1 + a_2 \dots a_k = n$. Then the total number of rearrangements of the n objects is

$$\frac{n!}{a_1! a_2! \dots a_k!} = \binom{n}{a_1 a_2 \dots a_k}$$

Theorem 2 (Multinomial: Binomial extension). For any positive integer n,

$$(x_1 + x_2 + \dots x_k)^n = \sum_{a_1+a_2+\dots a_k} \binom{n}{a_1 a_2 \dots a_k} x_1^{a_1} x_2^{a_2} \dots x_k^{a_k}$$

Example (Multinomials). How many rearrangements of AAABBCC are there?
Solution is: $\binom{7}{3,2,2}$

Example. What if the Bs must all be together, but none of the Cs can be together?
We can consider the "B"s as one object X. So we are trying to find arrangements of "AAXCC". We can first permute the "AAX", and then insert the "C"s into the slots in between. Therefore, it will be

$$\binom{4}{3,1} \cdot \binom{5}{2} = \frac{4!}{3!1!}\binom{5}{2}$$

1.3 Axioms of Probability

Remark. This is in Chapter 2.1 of the book.

Definition 5 (Sample Space). The set of all possible outcomes of an experiment is called the **sample space** S.

Definition 6 (Event). A subset of S is an **event**. The **null event**, or empty set, denoted $\varnothing$ or $\emptyset$, is the set with no elements.

Definition 7 (Union, Intersection). For $A, B \subseteq S$, the **union** and **intersection** is defined respectively as

- $A \cup B = \{x | x \in A \text{ or } x \in B\}$
- $A \cap B = \{x | x \in A \text{ and } x \in B\}$

Definition 8 (Disjoint). We say A, B are disjoint or mutually exclusive if $A \cap B = \varnothing$

Definition 9 (Complement). The complement of A is denoted A^c and is defined as

$$\{x | x \in S, x \notin A\}$$

Definition 10 (Bracket notation). We denote the set $S = 1, 2, \dots n = [n]$.

Theorem 3 (Demorgan's laws for Sets). For $A_1, A_2, \dots A_n \subseteq S$,

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

Proof 2. For number 2 (informal).

We prove by set inclusion. Let $x \in (\bigcup A_i)^c$. This means that $x \notin A_i$. Therefore, $x \in A_i^c$ for all A_i. So $x \in \bigcap A_i^c$. Therefore, $(\bigcup A_i) \subseteq \bigcap A_i^c$.

Let $x \in \bigcap A_i^c$. Then $x \in A_i^c$ for all A_i. So $x \notin A_i$. Therefore, $x \notin \bigcup A_i$. So $x \in (\bigcup A_i)^c$ □

Definition 11 (σ-algebra). Let F be a family of subsets of S (a subset of the powerset of S) We say F is a σ-algebra on a set S if

- $S \in F$
- If $A \in F$, then $A^c \in F$ ("closed under complements")
- If $A_1, A_2, \dots A_n \in F$, then
$$\bigcup_{i=1}^{n} A_i \in F$$
This is the same as being ("closed under countable unions")

Remark. By Demorgan's Law, by points 2 and 3 we have closure under intersections as well.

Example (σ-algebra). If $S = [6]$,

$$F = \{\varnothing, S, \{1, 2, 3\}, \{4, 5, 6\}\}$$

We want the idea of union subsets to be in the family for probability, so we can state things like "this event occurs or this event occurs".

Definition 12 (Probability Axioms). A probability function $P : F \to \mathbb{R}$ from a σ-algebra F on a set S satisfies

1. $P(A) \in [0,1]$ for all $A \in F$
2. $P(S) = 1$
3. If $A_1, A_2, \ldots$ are pairwise disjoint, then

$$P\left(\bigcup_{i\geq 1} A_i\right) = \sum_{i\geq 1} P\left(A_i\right)$$

Note. Property 3 can be extended to more sets.
For 3 sets,

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

This is known as the principle of inclusion and exclusion. See below.

Example (Probability function). You flip a two sided coin, and the probability of heads (H) is $\frac{1}{2}$. Here, $S = \{H, T\}$, $F = \{\varnothing, S, \{H\}, \{T\}\}$ By axiom 2, $P(S) = 1 = P(\{H, T\})$. So we can rewrite this as

$$P(\{H\} \cup \{T\}) = P(\{H\}) + P(\{T\})$$

Since we know the probability of heads is $\frac{1}{2}$, then

$$1 = P(\{H\}) + P(\{T\}) = \frac{1}{2} + P(\{T\})$$

So

$$P(\{T\}) = \frac{1}{2}$$

Theorem 4 (Properties of the probability function). Let A, B be events from a sample space S.

1. $P\left(A^c\right) = 1 - P\left(A\right)$
2. $P(\varnothing) = 0$
3. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
4. if $A \subseteq B$, then $P(A) \leq P(B)$

Proof 3. Proof for 2.

$$P(\varnothing) = P(\varnothing \cup \varnothing)$$

By axiom 3, this can be split up.

$$P(\varnothing) + P(\varnothing) = 2P(\varnothing)$$

Therefore,

$$P(\varnothing) = 2 \cdot P(\varnothing)$$

and so $P(\varnothing) = 0$. □

Proof 4. Proof for 3. Let $A \backslash B$ denote the set of elements in A, but not in B. We can rewrite

$$P(A \cup B) = P(A \backslash B \cup (A \cap B) \cup B \backslash A)$$

By axiom 3, since these three sets are disjoint, we can split it into

$$P(A \backslash B) + P(A \cap B) + P(B \backslash A)$$

We can now do a little trick:

$$P(A \backslash B) + P(A \cap B) + P(B \backslash A) + P(A \cap B) - P(A \cap B)$$

And now apply axiom 3 again, to get

$$P(A) + P(B) - P(A \cap B)$$

□

Theorem 5 (Inclusion-Exclusion). If $A_1, A_2, \ldots A_n \subseteq S$, then

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

$$+ \ldots (-1)^{n-1} P\left(\bigcap_{i=1} A_i\right)$$

Proof left as exercise, prove by induction.

Remark. Let $I \subseteq [n]$, Denote $A_I = \bigcap_{i \in I} A_i$, with $A_\varnothing = S$. Then

$$P\left(\bigcup_{i=1}^{n} A_i\right) = 1 - \sum_{I \subseteq [n]} (-1)^{|I|} P\left(A_I\right)$$

$$P\left(\bigcap_{i=1}^{n} A_i^c\right) = \sum_{I \subseteq [n]} (-1)^{|I|} P(A_I)$$

This sum is iterating over the powerset of $[n]$. This is a nicer expression in terms of the complement of our expression (see the $1 - \sum$).

1.4 Probability

We often focus on teh case where the sample space S is finite. In the case that every element has equal chance, we have for any event $E \in S$,

$$P(E) = \frac{|E|}{|S|} = \frac{\text{probability all ways event can occur}}{\text{size of sample space}}$$

Example. A coin is flipped 10 times. What is the probability of getting exactly 3 heads? The answer is

$$\frac{\binom{10}{3}}{2^{10}}$$

Example. There are 10 cats, 7 dogs, and 5 mice. Four animals are chosen at once. What is the probability we get at most one mouse. We want the cases where there are 0 mice or one mouse.

The number of ways of getting 0 mice is $\binom{17}{4}$.
The number of ways of getting one mouse is $\binom{5}{1}\binom{17}{3}$
The sample space size is $\binom{22}{4}$.
The probability is then

$$\frac{\binom{17}{4} + \binom{5}{1}\binom{17}{3}}{\binom{22}{4}}$$

Example. In some cases, S may be uncountable (no bisection from the naturals). For example, $S = [0, 1]$. Consider $E = [a, b] \subseteq [0, 1]$. Then

$$P(E) = \frac{b - a}{1 - 0} = b - a$$

Example. If n labeled balls are placed in n labeled boxes, what is the probability exactly one box is empty?

The number of ways to choose which box is empty is n.
We then choose the number of ways to put two balls together, which is $\binom{n}{2}$.
Finally, we arrange those $n - 1$ balls in the boxes, which is $(n - 1)!$

Therefore, the probability is

$$\frac{n\binom{n}{2}(n-1)!}{n^n}$$

Theorem 6. Let $A_1 \subseteq A_2 \subseteq A_3 \subseteq \dots$ be events in S. Then

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \lim_{n\to\infty} P(A_n)$$

Remark. An analog of this is

$$\dots A_3 \subseteq A_2 \subseteq A_1$$

then

$$P\left(\bigcap_{i=1}^{\infty} A_i\right) = \lim_{n\to\infty} P(A_n)$$

Proof 5. Let $B_1 = A_1, B_2 = A_2 \setminus A_1, \dots B_i = A_i \setminus A_{i-1}$. Then

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

$$= \lim_{n\to\infty} \sum_{i=1}^{n} P(B_i) = \lim_{n\to\infty} P\left(\bigcup_{i=1}^{n} B_i\right) = \lim_{n\to\infty} P(A_n)$$

Note in this proof we constructed the B_i to be disjoint and apply axiom 3. We then took the sum of $P(B_i)$ and consolidated it as they are disjoint. Also note that the union of A_i covers everything in the last A_i. If we union all the B_i, we get the same thing. □

Example. At 11.59 balls #1 to 10 are put into a box. One ball is removed. At 11:59:30 ... #11-20 ... and 1 ball is removed. At 11:59:45 ... #21-30 ... 1 ball is removed. At 12:00 how many balls are in the box?

The answer is none. What???? This is actually fake news.

Anyways, let A_i be the event 1 ball survives step i. Then

$$P(A_1) = \frac{9}{10}$$

$$P(A_2) = \frac{9}{10} \cdot \frac{18}{19}$$

$$P(A_3) = \frac{9}{10} \cdot \frac{18}{19} \cdot \frac{27}{28}$$

$$A = \bigcap_{i=1}^{\infty} A_i$$

This is the event 1 ball survives until 12:00. Note that $\ldots A_3 \subseteq A_2 \subseteq A_1$. By the theorem, we take the limit and see that it goes to 0.

Chapter 2

Conditional Probability

2.1 Introduction to Conditional Probability

Definition 13 (Conditional Probability). Let $A, B \in S$, $P(B) \neq 0$. The **conditional probability** of A given B ("probability of A given B") is

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

This is the probability of A, restricting S by knowing B.

Theorem 7 (Properties of conditional probability).

1. $0 \leq P(A|B) \leq 1$
2. $P(S|B) = 1$
3. if $A_1, A_2, \ldots$ are pairwise disjoint, then

$$P\left(\bigcup_{i=1}^{\infty} A_i \middle| B\right) = \sum_{i=1}^{\infty} P(A_i|B)$$

Proof 6. Proof of (3).

$$\begin{aligned} P\left(\bigcup_{i=1}^{\infty} A_i \middle| B\right) &= \frac{P\left(\left(\bigcup_{i=1}^{\infty} A_i\right) \cap B\right)}{P(B)} \\ &= \frac{P\left(\bigcup_{i=1}^{\infty} (A_i \cap B)\right)}{P(B)} \\ &= \sum_{i=1}^{\infty} \frac{P(A_i \cap B)}{P(B)} \\ &= \sum_{i=1}^{\infty} P(A_i|B) \end{aligned}$$

□

Example. A loaded 6-sided die has an odd number occurring twice as likely as even. Determine the probability that the number is a perfect square, given the value is larger than 3.
Let A be the even of a perfect square ($\{1, 4\}$), and B be the event of getting larger than 3 ($\{4, 5, 6\}$). Then $A \cap B = \{4\}$.
Let x be the probability of rolling an even number. We know then $2x+x+2x+x+2x+x = 1$, so $x = \frac{1}{9}$. So we have

$$\begin{aligned} P(A|B) &= \frac{P(A \cap B)}{P(B)} \\ &= \frac{P(\{4\})}{P(\{4, 5, 6\})} \\ &= \frac{\frac{1}{9}}{\frac{1}{9} + \frac{2}{9} + \frac{1}{9}} \\ &= \frac{1}{4} \end{aligned}$$

Logically we we see that "given B" restricts the sample size to $\{4, 5, 6\}$ which yields a $\frac{1}{4}$ chance to roll a 4.

Example. There are 7 black socks and 5 white socks. The socks are distinct, and we take two, one at a time ,without replacement. What is P(both socks are black)?
Let A_i be the event that the first sock is black, and A_2 be the event that the second sock is black. We want $P(A_1 \cap A_2) = \frac{7}{12} \cdot \frac{6}{11} = P(A_1) \cdot P(A_2|A_1)$.

Theorem 8 (Multiplication Rule). Let $A_1, A_2, \dots A_n \subseteq S$ and $P(A_1 \cap A_2 \dots A_n) > 0$. Then

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

Definition 14 (Independent). We say A and B are **independent** if

$$P(A \cap B) = P(A)P(B)$$

Otherwise, the events are dependent. In particular, if $P(B) \neq 0$, then if A, B are independent, we see that

$$P(A|B) = P(A)$$

Or that the conditioning of B does not affect A.

Example. We draw 2 cards from a deck of cards without replacement. If A is the event of drawing a spade on the first draw and B is the event of drawing a spade on the second draw, are A and B independent?
$P(A) = \frac{1}{4}$ However, $P(B) = P(\text{no spade}, \text{spade}) + P(\text{spade}, \text{spade}) = \frac{39}{52} \cdot \frac{13}{51} + \frac{13}{52} \cdot \frac{12}{51} = \frac{1}{4}$. Then

$$P(A \cap B) = P(A)P(B|A) = \frac{1}{4} \cdot \frac{12}{51} = \frac{1}{17}$$

Therefore, since $P(A \cap B) \neq P(A)P(B)$, these events are dependent.

Example. Toss a coin 3 times. Let A be the event that a head occurs on toss 1 and 2, and let B be the event that a tails occurs on toss 3.
Here $A = \{HHT, HHH\}$, $B = \{HHT, HTT, THT, TTT\}$ Therefore, $P(A) = \frac{2}{8}$ and $P(B) = \frac{4}{8}$. Note the intersection only contains one element, so $P(A \cap B) = \frac{1}{8}$. We see

$$P(A \cap B) = P(A)P(B)$$

so these two events are independent. Observe that even though $P(A \cap B) \neq \varnothing$, independence still holds!

Definition 15 (Independent (2)). Events $A_1, A_2, \dots A_k$ are independent if

$$P\left(\bigcap_{i \in J} A_i\right) = \prod_{i \in J} P(A_i)$$

where J is any subset of $\{1, 2, \dots k\}$.

Definition 16 (Pairwise Independence)**.** Events $A_1, \dots A_k$ are pairwise independent if

$$P(A_i \cap A_j) = P(A_i)P(A_j)$$

for all $1 \leq i \neq j \leq k$.

Note that if we know events are independent, then we know that they are pairwise independent.

Example (Pairwise independent)**.** Consider 4 labeled balls, pick one. Let $A_1 = \{1, 2\}$, $A_2 = \{1, 3\}$, $A_3 = \{1, 4\}$. Then we see that $P(A_1 \cap A_2) = P(\{1\}) = \frac{1}{4} = P(A_1)P(A_2)$. Note that this is true for any pair. Therefore, these events are pairwise independent. However, note that $P(A_1 \cap A_2 \cap A_3) \neq P(A_1)P(A_2)P(A_3)$, so these events are not independent.

Remark. While independence implies pairwise independence, note that pairwise independence does not imply independence in general.

Theorem 9 (Compl Independence)**.** Let A, B be events. Then A, B are independent if and only if A and B^c are independent.

Proof 7. (One direction). If A, B are independent, we know that $P(A \cap B) = P(A)P(B)$.

$$\begin{aligned} P(A \cap B^c) &= P(A) - P(A \cap B) \\ &= P(A) - P(A)P(B) \\ &= P(A)\left(1 - P(B)\right) \\ &= P(A)P(B^c) \end{aligned}$$

□

Example. Alice and Bob can solve 75% and 70% of the problems in a book, respectively. Assume the events are independent. If a problem is selected at random from the book, what is the probability it will be solved?
Let A = Alice solves the problem and B = Bob solves the problem. We want the value $P(A \cup B)$. By inclusion exclusion, $P(A \cup B) = P(A) + P(B) - P(A \cap B)$, and by independence this is equal to $P(A) + P(B) - P(A)P(B) = 0.925$.

2.2 Bayes Formula

Definition 17 (Law of total probability)**.** Let $A, B_1, B_2, \dots B_n$ be events such that $B_i \cap B_j = \varnothing$

for $1 \leq i \neq j \leq n$, and

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

In essence, the B events partition the sample space and do not intersect.

We can see that now we can separate the probabilities of A into pieces from the "cuts" of B.

$$P(A) = P\left(\bigcup_{i=1}^{n}(A \cap B_i)\right) = P(A \cap B_1) + P(A \cap B_2) + \cdots + P(A \cap B_n)$$

By the previous section on conditioning, note that this is the same as

$$P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + \ldots P(A|B_n)P(B_n)$$

Theorem 10 (Bayes Formula). Let $B_1, \ldots B_n$ be a partition of S ($B_i \cap B_j = \varnothing$, $\bigcup B_i = S$). Then

$$\begin{aligned} P(B_i|A) &= \frac{P(B_i \cap A)}{P(A)} \\ &= \frac{P(A|B_i)P(B_i)}{\sum_{n=1}^{k} P(A|B_k)P(B_k)} \end{aligned}$$

So given $P(A|B_k)$, we can find $P(B_i|A)$.

Example (Bayes). Alice is the main goalie of a soccer team, and Bob is a back-up. If Alice plays, there is a 75% chance the team wins. If Bob plays there is a 40% chance the team wins. The team doctor says there is a 70% chance Alice can play. If you read in the newspaper that the team won, what is the probability Bob played?

Let A be the even that the team wins, B_1 be the event that Alice plays, and B_2 be the event that Bob plays. Then note that

$$P(B_2|A) = \frac{P(B_2 \cap A)}{P(A)} = \frac{P(A|B_2) \cdot P(B_2)}{P(B_1)P(A|B_1) + P(B_2)P(A|B_2)}$$

We know $P(A|B_1) = 0.75$, $P(A|B_2) = 0.4$, $P(B_1) = 0.7$, $P(B_2) = 1 - 0.7 = 0.3$. Plugging all of this in, we get 0.186.

Example. A multiple choice question has m choices. Alice knows the answer with probability p. Otherwise she guesses with probability $1 - p$. What is the probability Alice knew the answer, given that she answered correctly?

So we have that A is the event Alice answers correctly, B_1 the event that she knows the

answer, and B_2 is the event that she guesses. We want to find

$$P(B_1|A) = \frac{P(B_1 \cap A)}{P(A)} = \frac{P(A|B_1) \cdot P(B_1)}{P(B_1)P(A|B_1) + P(B_2)P(A|B_2)}$$

We know $P(A|B_1) = 1$, $P(B_1) = p$, $P(B_2) = 1 - p$, $P(A|B_2) = \frac{1}{m}$. So then we get that

$$P(B_1|A) = \frac{1 \cdot p}{1 \cdot p + \frac{1}{m} \cdot (1 - p)} = \frac{mp}{1 + (m - 1)p}$$

Example. A rare disease affects 1% of a population. There is a test to check. If you have the disease, the probability that you test negative is 1%. The probability of a false positive is 2%. Given a positive test, what is the probability that you have the disease? What is the probability of a positive test?

We want $P(+)$, and we know $P(+|D^c) = 0.02$, $P(+^c|D) = 0.01$, and $P(D) = 0.01$. Therefore, we also know that $P(+|D) = 0.99$, and $P(D^c) = 0.99$.

Note that

$$\begin{aligned} P(+) &= P(+|D) \cdot P(D) + P(+|D^c) \cdot P(D^c) \\ &= 0.01 \cdot 0.99 + 0.99 \cdot 0.02 \\ &= 0.0297 \end{aligned}$$

Next, we want $P(D|+)$. We can then expand this to get

$$P(D|+) = \frac{P(D \cap +)}{P(+)} = \frac{P(+|D)P(D)}{P(+)} = \frac{0.01 \cdot 0.99}{0.0297} \approx 0.33$$

So even if you test positive, the test is somewhat counter intuitive to being accurate. The rareness of the disease is the "stranger force".

Example. In tennis, "deuce" is a score of 40–40. To win the game, one person must score 2 points in a row. If each player scores 1, it resets to deuce.

Let B_1 be the event that she wins the next two points, B_2 be the event that she loses the next two points, and B_3 be the event that Alice has one win and one loss in the next two points. Then

$$\begin{aligned} P(\text{Win}) &= P(\text{Win}|B_1)P(B_1) + P(\text{Win}|B_2)P(B_2) + P(\text{Win}|B_3)P(B_3) \\ &= 1 \cdot 0.6^2 + 0 \cdot 0.4^2 + P(\text{Win}) \cdot (0.6 \cdot 0.4 + 0.4 \cdot 0.6) \end{aligned}$$

Solving for $P(\text{Win})$, we get $\frac{9}{13}$.

Chapter 3

Discrete Random Variables

3.1 Random Variables

Definition 18 (Random Variable). A real valued function on a sample space S is a random variable

$$X : S \to \mathbb{R}$$

Example (Socks). There are 100 red, 100 blue, 100 white socks. Three are chosen at once. Let X be the total red socks. So X takes on $0, 1, 2, 3$.

Definition 19 (Discrete Random Variable). If x takes on a **finite** or countable number of values, we say X is a **discrete random variable**.

Definition 20 (Probability Mass Function). For a discrete random variable X, the probability mass function (or "distribution function" or PMF) of X is $p : \mathbb{R} \to \mathbb{R}$ (may be $f : \mathbb{R} \to \mathbb{R}$) such that

$$p(x) = \text{probability } X \text{ takes on } x = P(X = x)$$

In the discrete case, we have $p(x_i) \geq 0$ for $i = 1, 2, \ldots$ and $p(x) = 0$ for all other numbers. It follows then that

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

Example. Toss a coin twice, let X be the number of tails.

$$p(0) = p(\{HH\}) = \frac{1}{4}$$

$$p(1) = p(\{HT, TH\}) = \frac{1}{2}$$
$$p(2) = p(\{TT\}) = \frac{1}{4}$$
$$\frac{1}{4} + \frac{1}{2} + \frac{1}{4} = 1$$

Example. Let $S = [5]$. Assume 1 value is chosen with equal chance. Let

$$X = X(i) = (i-2)^2 - 1$$

We want to find the PMF, meaning that we should see what values X is taking on. For example,

$$i = 1, 3 \implies X = 0$$
$$i = 2 \implies X = -1$$
$$i = 4 \implies X = 3$$
$$i = 5 \implies X = 8$$

We can see that

$$p(x) = \begin{cases} \frac{1}{5} & \text{if } x = -1 \\ \frac{2}{5} & \text{if } x = 0 \\ \frac{1}{5} & \text{if } x = 3 \\ \frac{1}{5} & \text{if } x = 8 \end{cases}$$

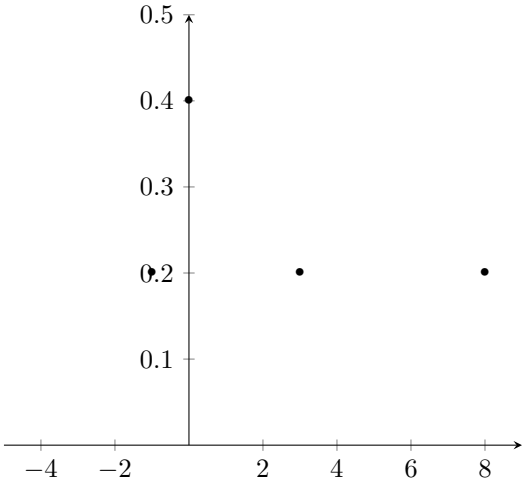

Figure 3.1: PMF

Definition 21 (Cumulative Distribution Function). A cumulative distribution function (cdf) of a random variable X is $F_x(x) = P(X \leq x)$

Example. From the previous PMF, we can see now that

$$F_x(x) = \begin{cases} 0 & \text{if } -\infty < x < -1 \\ \frac{1}{5} & \text{if } -1 \leq x < 0 \\ \frac{3}{5} & \text{if } 0 \leq x < 3 \\ \frac{4}{5} & \text{if } 3 \leq x < 8 \\ 1 & \text{if } 8 \leq x < \infty \end{cases}$$

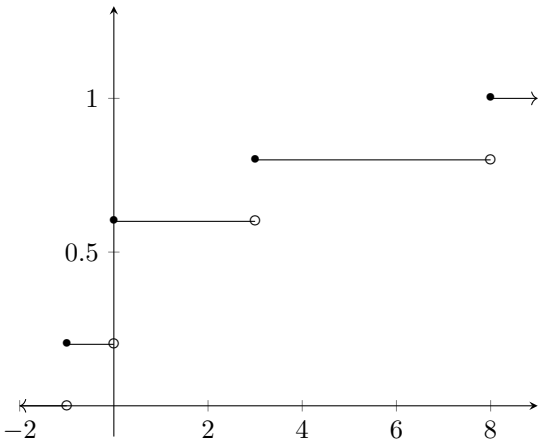

Figure 3.2: CDF

Example. Let the CDF of a discrete RV be

$$F_x(x) = \begin{cases} 0 & x < 1 \\ \frac{3}{10} & 1 \leq x < 2 \\ \frac{6}{10} & 2 \leq x < 3 \\ \frac{8}{10} & 3 \leq x < 5 \\ 1 & x \geq 5 \end{cases}$$

We can then do

$$P(2 < X \leq 3) = F_X(3) - F_X(2) = \frac{8}{10} - \frac{6}{10} = \frac{2}{10}$$

$$P(3 \leq X \leq 5) = F_X(5) - F_X(3) + P(X = 3)$$

Note that **in the discrete case**, $P(X = 3) = F_X(3) - F_X(2)$. Therefore,

$$P(3 \leq X \leq 5) = F_X(5) - F_X(2) = 1 - \frac{6}{10} = \frac{4}{10}$$

Theorem 11 (CDF). If a discrete random variable takes on $x_1 < x_2 < x_3 < \cdots < x_n$, we can find the probability of each value through $P(X = x_1) = F_X(x_1)$ and $P(X = x_i) = F_X(x_i) - F_X(x_{i-1})$ for $i = 2, \ldots n$. Moreover,

(a) $P(a < x \leq b) = F_X(b) - F_X(a)$

(b) $P(a < x < b) = F_X(b) - F_X(a) - P(X = b)$

(c) $P(a \leq X \leq b)$ left as exercise.

(d) $P(a \leq X < b)$ left as exercise.

Remark. You can actually think of the PDF as a combination of scaled dirac delta functions. Then, if we integrate it, we get the Heaviside function giving us this nice step.

Theorem 12 (CDF). A function $F_X(x)$ is a CDF if and only if

1. $0 \leq F_X(x) \leq 1$ for all x that X takes on.
2. $\lim_{x \to \infty} F_X(x) = 1$
3. $\lim_{x \to -\infty} F_X(x) = 0$
4. $F_X(x)$ is non-decreasing
5. $F_X(x)$ is right continuous for all $x \in R$. In other words,

$$\lim_{x \to x_0^+} F_X(x) = F_X(x_0)$$

Example (Bernoulli Random Variable). A Bernoulli random variable is the following

$$X = \begin{cases} 1 & \text{success} \\ 0 & \text{fail} \end{cases}$$

where the probability of success is p and fail is $1 - p$.

Example (Binomial Random Variable). The Binomial random variable is denoted as

$$X \sim \text{Bin}(n, p)$$

where a Bernoulli experiment is done n times, each with probability of success p. If X is the number of successes, then X takes on $0, 1, 2, \ldots n$. Here, the probability mass function will satisfy

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

Observe that

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

This satisfies the condition of a PMF.

Also, we see that the CDF is

$$\begin{aligned} F_X(i) &= P(X \leq i) \\ &= \sum_{j=1}^{i} \binom{i}{j} p^j (i-p)^{i-j} \end{aligned}$$

3.2 Expectation

Definition 22 (Expected Value). If X is a discrete random variable with PMF $p(x)$, then the expected value of X is

$$\begin{aligned} E(X) &= \sum_x x p(x) \\ &= \sum_x x P(X = x) \end{aligned}$$

provided the sum $\sum_x |x| p(x)$ does not diverge.

The expected value of a function $x(X)$ is

$$\begin{aligned} E(g(X)) &= \sum_x g(x) p(x) \\ &= \sum_x g(x) P(X = x) \end{aligned}$$

provided that the sum $\sum_x |g(x)|p(x)$ does not diverge.

Remark. If $g(x) = x^n$, then

$$E(g(x)) = E(x^n)$$

is the **nth moment of** X.

Example. A coin is tossed until you see a tails. On toss i, you win $(-2)^{i-1}$ dollars. What is the expected winnings?

Let X be the possible winnings. Then X takes on $1, -2, 4, -8, \ldots$. The probability we end on round i is $\left(\frac{1}{2}\right)^i$. So

$$E(X) = 1\left(\frac{1}{2}\right) - 2\left(\frac{1}{4}\right) + 4\left(\frac{1}{8}\right) - 8\left(\frac{1}{16}\right)\ldots$$

This is equal to $\frac{1}{2} - \frac{1}{2} + \frac{1}{2} - \frac{1}{2}\ldots$, which diverges.

Example. Same problem as before, but we win i dollars on round i. Then we have

$$E(X) = \sum_{i=0}^{\infty} i\left(\frac{1}{2}\right)^i$$

We denote $S = E(X)$. Then

$$\frac{1}{2}S = 0 + 1\left(\frac{1}{2}\right)^2 + 2\left(\frac{1}{2}\right)^3 + 3\cdot\left(\frac{1}{2}\right)^4\ldots$$
$$S - \frac{1}{2}S = \frac{1}{2} + \frac{1}{2}^2 + \frac{1}{2}^3 + \frac{1}{2}^4 + \ldots$$

By the geometric formula for sums, we can evaluate this, giving us

$$\frac{1}{2}S = \frac{1}{1-\frac{1}{2}} - 1$$

So therefore,

$$S = 2$$

Definition 23 (Variance)**.** The **variance** of a random variable X is

$$\mathrm{Var}\,(X) = E\left((X - E(X))^2\right)$$

Theorem 13 (Expectation Properties). For a random variable X, we have

1. $E(aX+b) = aE(X)+b$ for $a, b \in \mathbb{R}$.
2. For $g_1(x), g_2(x)$, $E(g_1(x)+g_2(x)) = E(g_1(x)) + E(g_2(x))$ (generalizes to n functions).
3. $\text{Var}(X) = E(X^2) - (E(X))^2$
4. $\text{Var}(aX+b) = a^2 \text{Var}(X)$ (note that $\text{Var}(b) = 0$).

Proof 8. (Proof of 1).

$$\begin{aligned} E(aX+b) &= \sum_x (ax+b)p(x) \\ &= a\sum_x xp(x) + b\sum_x p(x) \\ &= aE(x) + b \cdot 1 \\ &= aE(x) + b \end{aligned}$$

(Proof of 2).

$$\begin{aligned} E(g_1(x)+g_2(x)) &= \sum_x (g_1(x)+g_2(x))\,p(x) \\ &= \sum_x g_1(x)p(x) + \sum_x g_2(x)p(x) \\ &= E(g_1(x)) + E(g_2(x)) \end{aligned}$$

(Proof of 3).

$$\begin{aligned} \text{Var}(X) &= E\left((X-E(X))^2\right) \\ &= E\left(X^2 + 2XE(X) + E(X)^2\right) \end{aligned}$$

Note here that $E(X)$ is constant. Therefore, we get $X^2 + KX + K^2$. Therefore, when we distribute the expected value, it only applies to X^2 and X.

$$\begin{aligned} \text{Var}(X) &= E(X^2) - 2E(X)E(X) + (E(X))^2 \\ &= E(X^2) - E(X)^2 \end{aligned}$$

(Proof of 4).

$$\begin{aligned}
\operatorname{Var}(aX+b) &= E((aX+b)^2) - E(aX+b)^2 \\
&= E(a^2X^2 + 2abX + b^2) - (aE(X)+b)^2 \\
&= E(a^2X^2 + 2abX + b^2) - a^2E(X)^2 - 2abE(X) - b^2 \\
&= a^2E(X^2) + 2abE(X) - a^2E(X)^2 - 2abE(X) \\
&= a^2E(X^2) - a^2E(X)^2 \\
&= a^2\left(E(X^2) - E(X)^2\right) \\
&= a^2\left(\operatorname{Var} X\right)
\end{aligned}$$

□

Example. Consider rolling a 4-sided die once. Let X be the value the die takes on. Find $V(X)$.

$$E(X) = 1\cdot\frac{1}{4} + 2\cdot\frac{1}{4} + 3\cdot\frac{1}{4} + 4\cdot\frac{1}{4} = \frac{5}{2}$$

$$E(X^2) = 1\cdot\frac{1}{4} + 4\cdot\frac{1}{4} + 9\cdot\frac{1}{4} + 16\cdot\frac{1}{4} = \frac{30}{4}$$

$$\operatorname{Var} X = E(X^2) - (E(X))^2 = \frac{30}{4} - (\frac{5}{2})^2 = \frac{5}{4}$$

Alternatively, compute $X - E(X)$ for $X = 1, 2, 3, 4$. Then $\operatorname{Var} X = E\left((X - E(X))^2\right)$.

Note. By definition $\operatorname{Var} X = E\left((X - E(X))^2\right)$. The intuition is the expected value of the squares of the difference. We are trying to find the "average" of this difference. If the variance is large, then we see that the values are farther away from the average. If it is large, then likely so was $X - E(X)$, so the values X takes on are far away from $E(X)$.

3.3 Discrete Random Variable Examples

Definition 24 (Uniform Discrete Random Variable)**.** A random variable X has a discrete uniform distribution with parameter n if and only if the PMF is

$$p(x) = \begin{cases} \frac{1}{n} & x = 1, 2, \ldots n \\ 0 & \text{otherwise} \end{cases}$$

and the uniform discrete random variable has the following properties:

(a)
$$E(X) = \frac{n+1}{2}$$

(b)
$$\text{Var}(X) = \frac{(n+1)(n-1)}{12}$$

Proof 9. We see that

$$\begin{aligned} E(X) &= \frac{1}{n}\sum_{i=1}^{n} i \\ &= \frac{1}{n}\frac{n(n+1)}{2} \\ &= \frac{n+1}{2} \end{aligned}$$

We can then find the variance:

$$\begin{aligned} \text{Var}(X) &= E(X^2) - E(X)^2 \\ &= \sum_{i=1}^{n} \frac{i^2}{n} - \left(\frac{n+1}{2}\right)^2 \\ &= \frac{1}{n}\frac{n(n+1)(2n+1)}{6} - \left(\frac{n+1}{2}\right)^2 \\ &= \frac{(n+1)(n-1)}{12} \end{aligned}$$

□

Definition 25 (Bernoulli Distribution)**.** We write $X \sim \text{Bernoulli}(p)$, p is the probability of success if

$$X = \begin{cases} 1 & \text{success with prob } p \\ 0 & \text{failure with prob } 1-p \end{cases}$$

The Bernoulli distribution satisfies the following properties:

(a)
$$E(X) = p$$

(b)
$$\text{Var}(X) = p(1-p)$$

Proof 10. We see that the PMF is $p(1) = p$, $p(0) = 1 - p$. Therefore,

$$E(X) = 0(1-p) + 1(p) = p$$

$$\text{Var}(X) = E(X^2) - (E(X))^2 = p - p^2 = p(1-p)$$

□

Definition 26 (Geometric Definition). We write $X \sim \text{Geom}(p)$. This is seen as the probability of the first success after x trials of Bernoulli events with probability p for a success). Therefore, the probability mass function for $x = 0, 1, \ldots$ is

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

Observe

$$\sum_{x=1}^{\infty} p(1-p)^{x-1} = p\sum_{x=1}^{\infty}(1-p)^{x-1}$$

By the Geometric Series, we see that this is the same as

$$p \cdot \frac{1}{1-(1-p)} = 1$$

Therefore, this is a valid PMF. Note the following properties:

(a)

$$E(X) = \frac{1}{p}$$

(b)

$$V(X) = \frac{1-p}{p^2}$$

Proof 11.

$$\begin{aligned} E(X) &= \sum_{x=1}^{\infty} xp(1-p)^{x-1} \\ &= p\sum_{x=1}^{\infty} x(1-p)^{x-1} \end{aligned}$$

Note here that the inside looks like a derivative of $(1-p)^x$. Let $q = 1 - p$. Therefore (needs

some rigor but this is not analysis class)

$$\begin{aligned}
&= p\frac{\mathrm{d}}{\mathrm{d}q}\sum_{x=0}^{\infty} q^x \\
&= p\frac{\mathrm{d}}{\mathrm{d}q}\left(\frac{1}{1-q}\right) \\
&= p\frac{1}{(1-q)^2} \\
&= p\left(\frac{1}{p^2}\right) \\
&= \frac{1}{p}
\end{aligned}$$

Next, we see that

$$\begin{aligned}
E(X^2) &= \sum_x x^2 p(x) \\
&= \sum_{x=1}^{\infty} x^2 p q^{x-1} \\
&= \sum_{x=1}^{\infty} (x(x-1)+x) q^{x-1} p \\
&= \sum_{x=1}^{\infty} (x(x-1)) q^{x-1} p + \sum_{x=1}^{\infty} x q^{x-1} p \\
&= pq\sum_{x=1}^{\infty} (x(x-1)) q^{x-2} + E(X) \\
&= pq\sum_{x=1}^{\infty} (x(x-1)) q^{x-2} + \frac{1}{p} \\
&= pq\left(\frac{\mathrm{d}^2}{\mathrm{d}q^2}\left(\sum_{x=0}^{\infty} q^x\right)\right) + \frac{1}{p} \\
&= pq\left(\frac{\mathrm{d}^2}{\mathrm{d}q^2}\frac{1}{1-q}\right) + \frac{1}{p} \\
&= pq\left(\frac{2}{(1-q)^3}\right) + \frac{1}{p} \\
&= \frac{2-p}{p^2}
\end{aligned}$$

Therefore,

$$V(X) = E(X^2) - E(X)^2 = \frac{2-p}{p^2} - \frac{1}{p^2} = \frac{1-p}{p^2}$$

□

Definition 27 (Binomial Distribution)**.** We write $X \sim \text{Binomial}(n, p)$, where X is the number of successes, and the probability of success is p for each of n independent trials. The PMF is

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

where $x = 0, 1, \ldots n$.

The Binomial Distribution has the following properties:

(a)

$$E(X) = np$$

(b)

$$V(X) = np(1-p)$$

Proof 12.

$$\begin{aligned}
E(X) &= \sum_{x=0}^{n} x \binom{n}{k} p^x (1-p)^{n-x} \\
&= \sum_{x=1}^{n} x \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} \\
&= np \sum_{x=1}^{n} \frac{(n-1)!}{(x-1)!(n-x)!} p^{x-1} (1-p)^{n-x} \\
&= np \sum_{x=1}^{n} \binom{n-1}{x-1} p^{x-1} (1-p)^{(n-1)-(x-1)} \\
&= np(p + (1-p))^{n-1} \\
&= np
\end{aligned}$$

We also see that

$$\begin{aligned}
E(X^2) &= \sum_{x=0}^{n} x^2 \binom{n}{k} p^x (1-p)^{n-x} \\
&= \sum_{x=0}^{n} x \cdot \frac{n!}{(x-1)!(n-x)!} p^x (1-p)^{n-x} \\
&= np \sum_{x=1}^{n} x \cdot \frac{(n-1)!}{(x-1)!(n-x)!} p^{x-1} (1-p)^{n-x} \\
&= np \left(\sum_{x=1}^{n} (x-1) \cdot \binom{n-1}{x-1} p^{x-1} (1-p)^{n-x} + \sum_{x=1}^{n} \binom{n-1}{x-1} p^{x-1} (1-p)^{n-x} \right) \\
&= np \left(E(\text{Binomial}(n-1,p)) + ((1-p)+p)^{n-1} \right) \\
&= np \left((n-1)p + 1 \right)
\end{aligned}$$

Therefore,

$$V(X) = E(X^2) - (E(X))^2 = np(1-p)$$

□

Definition 28 (Hypergeometric Distribution)**.** We write $X \sim \text{Hypergeometric}(n, m, k)$, with PMF

$$p(x) = \frac{\binom{m}{x}\binom{n-m}{k-x}}{\binom{n}{k}}$$

We see this as n total elements, with m successes. Then $p(x)$ measures the probability of drawing exactly x successes without replacement if we draw from the n k times, where $x = 0 \ldots k$. Think of the PMF as stating "out of all k draws from n, we pick the ones where x of them are from m and the other $k - x$ are from the failures".

The Hypergeometric Distribution has the following properties:

(a)
$$E(X) = \frac{km}{n}$$

(b)
$$V(X) = \frac{km(n-m)(n-k)}{n^2(n-1)}$$

Lemma 1. We introduce the following lemma to prove the expected value.

$$\sum_{i=0}^{k} \binom{a}{i}\binom{b}{k-i} = \binom{a+b}{k}$$

Proof 13. We see that

$$\begin{aligned}
E(X) &= \sum_{x=0}^{k} x \cdot \frac{\binom{m}{x}\binom{n-m}{k-x}}{\binom{n}{k}} \\
&= m\sum_{x=0}^{k} \frac{(m-1)!}{(x-1)!(m-x)!} \cdot \frac{\binom{n-m}{k-x}}{\binom{n}{k}} \\
&= \frac{m}{\binom{n}{k}} \sum_{x=0}^{k} \binom{m-1}{x-1}\binom{n-m}{k-x} \\
&= \frac{m}{\binom{n}{k}} \cdot \binom{n-1}{k-1} \\
&= \frac{km}{n}
\end{aligned}$$

Proof for variance left as exercise (it is much more algebra intensive). □

Remark. Let p, q be the proportions of successes and failures, respectively ($p = \frac{m}{n}$). Then the PMF is

$$\begin{aligned}
p(x) &= \binom{np}{x}\binom{nq}{k-x} \\
&= \binom{k}{x} p\left(p-\frac{1}{n}\right)\left(p-\frac{2}{n}\right)\cdots\left(p-\frac{x-1}{n}\right) q\left(q-\frac{1}{n}\right)\cdots\left(q-\frac{k-x-1}{n}\right)
\end{aligned}$$

and in the limit, we have

$$\lim_{n\to\infty} p(x) = \binom{k}{x} p^x q^{k-x}$$

which is the binomial distribution. This is stating that for large n, when drawing without replacement, we can approximate this by drawing without replacement (or doing each trial independently, as in the binomial distribution).

Definition 29 (Negative Binomial). We write $X \sim \text{NegativeBinomial}(r, p)$, with PMF

$$p(x) = \binom{x-1}{r-1} p^r (1-p)^{x-r}$$

For $x = r, r+1, \ldots$

This is seen as the probability of getting r-th success after x trials (probability p of success). Note the following properties proceed from the binomial:

(a)
$$E(X) = \frac{r}{p}$$

(b)
$$V(X) = r\left(\frac{1-p}{p^2}\right)$$

If the PMF is $p(k)$ where k is the number of failures, we have

(a)
$$E(X) = \frac{r}{p} - r$$

(b)
$$V(X) = r\left(\frac{1-p}{p^2}\right)$$

Example. If a person is exposed to a disease, 30% show symptoms. What is the probability that the 100-th ($x = 100$) person exposed is the 7-th ($r = 7$) person to show symptoms?

Let $k = x - r =$ number of failures. Then

$$p(k) = \binom{k+r-1}{r-1}(1-p)^k p^r$$

Note. We have

$$\binom{k+r-1}{r-1} = \frac{(k+r-1)(k+r-2)\ldots(r)}{k!} = \frac{(-1)^k(-r)(-r-1)(-r-2)\ldots(-r-k+1)}{k!}$$

This is equal to

$$(-1)^k \binom{-r}{k}$$

This is where the phrase "negative binomial" comes from.

Proof 14. We prove the negative binomial is a PMF, and omit expected value and variance.
We have that

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

so therefore,

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

□

Consider a time period of 1 week, and X is the number of car accidents. We break up the time interval into n subintervals such that at most 1 accident occurs per subinterval. In any subinterval, there is a probability p of exactly 1 accident happening. If each subinterval is independent, this is a binomial distribution.

We expect as the number of subintervals n gets larger (smaller time frame), p gets smaller. We assume here $np = \lambda$ is constant.

Observe that the binomial random variable had

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

Substituting in $p = \frac{\lambda}{n}$, we get

$$\begin{aligned} p(x) &= \binom{n}{x} \left(\frac{\lambda}{n}\right)^x \left(1 - \frac{\lambda}{n}\right)^{n-x} \\ &= \frac{(n)(n-1)\dots(n-x+1)}{x!} \left(\frac{\lambda}{n}\right)^x \left(1 - \frac{\lambda}{n}\right)^{n-x} \\ &= \frac{1\left(1-\frac{1}{n}\right)\dots\left(1-\frac{x-1}{n}\right)}{x!} \lambda^x \left(\left(1 - \frac{\lambda}{n}\right)^{-\frac{n}{\lambda}}\right)^{-\lambda} \left(\left(1 - \frac{\lambda}{n}\right)^{-x}\right) \end{aligned}$$

Therefore, as we take $\lim_{n\to\infty} p(x)$, we get the Poisson distribution. We remember that

$$\lim_{n\to\infty}\left(1-\frac{1}{n}\right)^{-n}=e$$

Therefore, we get

$$\lim_{n\to\infty} p(x)=\frac{1}{x!}\lambda^x e^{-\lambda}$$

Definition 30 (Poisson Distribution). We say that $X\sim \text{Poisson}(\lambda)$ has PMF

$$p(x)=\frac{\lambda^x}{x!}e^{-\lambda}$$

for $x=0,1,\ldots$

(a)
$$E(X)=\lambda$$

(b)
$$V(X)=\lambda$$

Proof 15. Note that

$$\begin{aligned}
E(X)&=\sum_x xp(x)\\
&=\sum_{x=0}^{\infty} x\cdot\frac{\lambda^x}{x!}e^{-\lambda}\\
&=\sum_{x=1}^{\infty}\frac{\lambda^x}{(x-1)!}e^{-\lambda}\\
&=\lambda e^{-\lambda}\sum_{x=1}^{\infty}\frac{\lambda^{x-1}}{(x-1)!}\\
&=\lambda e^{-\lambda}e^{\lambda}\\
&=\lambda
\end{aligned}$$

We see that

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

Let $x = i + 1$. Then

$$\begin{aligned} \lambda e^{-\lambda} \sum_{x=1}^{\infty} x \frac{\lambda^{x-1}}{(x-1)!} &= \lambda e^{-\lambda} \sum_{i=0}^{\infty} (i+1) \frac{\lambda^i}{i!} \\ &= \lambda e^{-\lambda} \left(\sum_{i=1}^{\infty} \frac{\lambda^i}{(i-1)!} + \sum_{i=0}^{\infty} \frac{\lambda^i}{i!} \right) \\ &= \lambda e^{-\lambda} \left(\lambda e^{\lambda} + e^{\lambda} \right) \\ &= \lambda^2 + \lambda \end{aligned}$$

Therefore,

$$V(X) = E(X^2) - (E(X))^2 = (\lambda^2 + \lambda) - \lambda^2 = \lambda$$

□

Chapter 4

Continuous Random Variables

4.1 Continuous Random Variables

Definition 31 (Continuous Random Variable)**.** (Informal). A continuous random variable $X : S \to \mathbb{R}$ is a function that takes on an uncountable many elements (no bijection to the natural numbers). Alternatively, X is continuous if there is a non negative function $f(x)$ such that $f(x)$ is defined for all $\mathbb{R}$ and

$$P(X \in B) = \int_B f(x)\,\mathrm{d}x$$

The function $f(x)$ is called the probability density function (PDF) or "distribution function".

Theorem 14 (Properties of PDF)**.** Here are some properties of the PDF.

1. $$f(x) \geq 0$$
2. $$\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = 1$$
3. $$P(a \leq X \leq b) = \int_a^b f(x)\,\mathrm{d}x$$
4. $$P(X = a) = \int_a^a f(x)\,\mathrm{d}x = 0$$
5. $$P(A \leq X \leq b) = P(a < X < b)$$

Example. Let X have a PDF

$$f(x) = \begin{cases} ke^{-3x} & x > 0 \\ 0 & x \leq 0 \end{cases}$$

Find k and determine $P(0.5 \leq X \leq 1)$.

Since a PDF must sum to 1, we have

$$\begin{aligned} 1 &= \int_{-\infty}^{\infty} f(x)\,\mathrm{d}x \\ &= \int_{0}^{\infty} ke^{-3x}\,\mathrm{d}x \\ &= k \lim_{t\to\infty} \int_0^t e^{-3x}\,\mathrm{d}x \\ &= k \lim_{t\to\infty} \frac{e^{-3x}}{3}\Big|_0^t \\ &= \frac{k}{3} \end{aligned}$$

Therefore, k is 3. Then

$$\begin{aligned} P(0.5 \leq X \leq 1) &= \int_{0.5}^{1} 3e^{-3x}\,\mathrm{d}x \\ &= -e^{-3} + e^{-1.5} \\ &\approx 0.173 \end{aligned}$$

Definition 32 (CDF). If X is a continuous random variable with pdf $f(t)$, then the cumulative distribution function or CDF is

$$F_X(x) = F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)\,\mathrm{d}t$$

CDF properties in the discrete case follow similarly. For example,

1. $F(\infty) = 1$
2. $F(-\infty) = 0$

Theorem 15 (FTC). If $f(x), F(x)$ are the PDF and CDF of a continuous random variable X, then $P(a \leq X \leq b) = F(b) - F(a)$. Moreover,

$$f(x) = \frac{\mathrm{d}}{\mathrm{d}x} F(x)$$

Given the CDF, we can take a derivative to get the PDF, assuming derivative exists.

Example. Find the CDF from the last example

$$f(x) = \begin{cases} 3e^{-3x} & x > 0 \\ 0 & x \leq 0 \end{cases}$$

We get

$$\begin{aligned} F(x) &= \int_{-\infty}^{x} f(t)\,\mathrm{d}t \\ &= \int_{0}^{x} 3e^{-3t}\,\mathrm{d}t \\ &= e^{3t}\Big|_{0}^{x} \\ &= 1 - e^{-3x} \end{aligned}$$

So

$$F(x) = \begin{cases} 1 - e^{-3x} & x > 0 \\ 0 & x \leq 0 \end{cases}$$

Note by the theorem that $P(0.5 \leq X \leq 1) = F(1) - F(0.5) = -e^{-3} + e^{-1.5}$.

Definition 33 (Expected Value). If X is a continuous random variable with PDF $f(x)$, then the expected value is

$$E(X) = \int_{-\infty}^{\infty} x f(x)\,\mathrm{d}x$$

Definition 34 (Variance). If X is a continuous random variable with PDF $f(x)$, then the variance is

$$V(X) = \int_{-\infty}^{\infty} (x - E(X))^2 f(x)\,\mathrm{d}x$$

However, we will often use the form

$$\begin{aligned} V(X) &= E(X^2) - (E(X))^2 \\ &= \int_{-\infty}^{\infty} x^2 f(x)\,\mathrm{d}x - \left(\int_{-\infty}^{\infty} x f(x)\,\mathrm{d}x\right)^2 \end{aligned}$$

Note the properties for the discrete case follow in the same manner.

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

(b) $V(aX+b) = a^2V(X)$

(c) $E(g(x)) = \int_{-\infty}^{\infty} g(x)f(x)\,\mathrm{d}x$

Example. If X has PDF

$$f(x) = \begin{cases} e^{-x} & x > 0 \\ 0 & x \leq 0 \end{cases}$$

find $E(e^{\frac{3x}{4}})$.

$$\begin{aligned} E(e^{\frac{3x}{4}}) &= \int_{-\infty}^{\infty} g(x)f(x)\,\mathrm{d}x \\ &= \int_{0}^{\infty} e^{\frac{3x}{4}} e^{-x}\,\mathrm{d}x \\ &= \int_{0}^{\infty} e^{-\frac{x}{4}}\,\mathrm{d}x \\ &= 4 \end{aligned}$$

Example. If X has PDF

$$f(x) = \begin{cases} \frac{4}{1+x^2} & 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

find $V(X)$.

$$\begin{aligned} E(X) &= \int_{0}^{1} x\frac{4}{1+x^2}\,\mathrm{d}x \\ &= \frac{1}{2}4\int_{1}^{2} \frac{1}{u}\,\mathrm{d}u = 2\ln 2 \end{aligned}$$

$$\begin{aligned}
E(X^2) &= \int_0^1 x^2 \left(\frac{4}{1+x^2}\right) \mathrm{d}x \\
&= 4\int_0^1 \frac{x^2}{1+x^2}\,\mathrm{d}x \\
&= 4\int_0^1 \frac{x^2+1-1}{1+x^2}\,\mathrm{d}x \\
&= 4\int_0^1 1 - \frac{1}{1+x^2}\,\mathrm{d}x \\
&= 4\left(x - \arctan x\right)\Big|_0^1 \\
&= 4-\pi
\end{aligned}$$

Therefore,

$$V(X) = E(X^2) - (E(X))^2 = (4-\pi) - (2\ln 2)^2$$

4.2 Continuous Random Variable Examples

Definition 35 (Uniform). Random variable X has a uniform distribution, denoted $X \sim U(\alpha, \beta)$, over an interval (α, β) if the PMF is

$$f(x;\alpha,\beta) = \begin{cases} \dfrac{1}{\beta-\alpha} & \alpha < x < \beta \\ 0 & \text{otherwise} \end{cases}$$

Moreover, we have the CDF for $U(\alpha, \beta)$ is

$$F(x) = \begin{cases} 0 & x \le \alpha \\ \frac{x-\alpha}{\beta-\alpha} & \alpha < x < \beta \\ 1 & x \ge \beta \end{cases}$$

Properties:

(a)

$$E(X) = \frac{\alpha+\beta}{2}$$

(b)

$$V(X) = \frac{(\beta-\alpha)^2}{12}$$

Proof 16. We have

$$\begin{aligned} E(X) &= \int_{-\infty}^{\infty} xf(x)\,\mathrm{d}x \\ &= \int_{\alpha}^{\beta} \frac{x}{\beta-\alpha}\,\mathrm{d}x \\ &= \frac{\alpha+\beta}{2} \end{aligned}$$

Then

$$\begin{aligned} E(X^2) &= \int_{-\infty}^{\infty} x^2 f(x)\,\mathrm{d}x \\ &= \int_{\alpha}^{\beta} \frac{x^2}{\beta-\alpha}\,\mathrm{d}x \\ &= \frac{\alpha^2+2\alpha\beta+\beta^2}{3} \end{aligned}$$

Therefore,

$$V(X) = E(X^2) - (E(X))^2 = \frac{(\beta-\alpha)^2}{12}$$

□

Definition 36 (Normal Distribution)**.** A random variable X has a normal distribution denoted $X \sim N(\mu, \sigma)$ if the PMF is

$$f(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$$

Properties:

(a)
$$E(X) = \mu$$

(b)
$$V(X) = \sigma^2$$

The standard deviation here is σ. We can verify that $f(x)$ satisfies the definition of a PMF.

Proof 17. (Verification of PMF). Let $k = \int_{-\infty}^{\infty} f(x)\,\mathrm{d}x$, and $z = \dfrac{x-\mu}{\sigma}$. Then note that

$$\mathrm{d}z = \frac{1}{\sigma}\,\mathrm{d}x$$

We have

$$\begin{aligned} k &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}\,\mathrm{d}z \\ k\sqrt{2\pi} &= \int_{-\infty}^{\infty} e^{-\frac{1}{2}z^2}\,\mathrm{d}z \\ 2\pi k^2 &= \int_{-\infty}^{\infty} e^{-\frac{1}{2}w^2}\,\mathrm{d}w \int_{-\infty}^{\infty} e^{-\frac{1}{2}z^2}\,\mathrm{d}z \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-\frac{w^2+z^2}{2}}\,\mathrm{d}w\,\mathrm{d}z \end{aligned}$$

Using polar coordinates with $w = r\cos\theta$ and $z = r\sin\theta$, we see the bounds are equivalent to $0 \le \theta \le 2\pi$ and $0 < r < \infty$. Therefore, we can rewrite this integral as

$$\begin{aligned} k^2 2\pi &= \int_0^{2\pi}\int_0^{\infty} e^{-\frac{r^2}{2}} r\,\mathrm{d}r\,\mathrm{d}\theta \qquad\qquad = \int_0^{2\pi}\mathrm{d}\theta \int_0^{\infty} e^{-\frac{r^2}{2}} r\,\mathrm{d}r \\ &= 2\pi \lim_{t\to\infty} \int_0^{t} e^{-\frac{r^2}{2}} r\,\mathrm{d}r \\ &= 2\pi \lim_{t\to\infty} -\left(e^{-\frac{r^2}{2}}\right)\Big|_0^t \\ &= 2\pi \end{aligned}$$

Since $f(x) \ge 0$, we find $k = 1$, which verifies that f is in fact a PMF. □

Definition 37 (Standard Normal Distribution)**.** For $X \sim N(0,1)$, we call this the standard normal distribution.

Proof 18. (Properties of Standard Normal Distribution). Here, the PMF and CDF are

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

$$F_X(x) - \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}\,\mathrm{d}t$$

One can show that $\int_0^\infty f(x)\,\mathrm{d}x$ exists, so $E(X) = \int_{-\infty}^\infty xf(x)\,\mathrm{d}x = 0$ since $xf(x)$ is odd. Then

$$\begin{aligned} E(X^2) &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} x^2 e^{-\frac{x^2}{2}}\,\mathrm{d}x \\ &= 2\int_0^{\infty} \frac{1}{\sqrt{2\pi}} x^2 e^{-\frac{x^2}{2}} \end{aligned}$$

Let $u = x$, $\mathrm{d}u = \mathrm{d}x$, $\mathrm{d}v = xe^{-\frac{x^2}{2}}$, $v = -e^{-\frac{x^2}{2}}$. Then if we use integration by parts, we get

$$\begin{aligned} E(X^2) &= \frac{2}{\sqrt{2\pi}}\left(uv - \int_0^\infty v\,\mathrm{d}u\right) \\ &= \frac{2}{\sqrt{2\pi}}\left(\lim_{t\to\infty} -xe^{-\frac{x^2}{2}}\Big|_0^t + \int_0^\infty e^{-\frac{x^2}{2}}\,\mathrm{d}x\right) \\ &= \frac{2}{\sqrt{2\pi}}\int_0^\infty e^{-\frac{x^2}{2}}\,\mathrm{d}x \\ &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^\infty e^{-\frac{x^2}{2}}\,\mathrm{d}x \\ &= \int_{-\infty}^\infty f(x)\,\mathrm{d}x \\ &= 1 \end{aligned}$$

Therefore,

$$V(X) = E(X^2) - E(X)^2 = 1 - 0 = 1$$

□

Proof 19. (Properties for Normal Distribution).

Recall that for $Z \sim N(0,1)$, $E(Z) = 0$ and $V(Z) = 1$. Thus, for $z = \frac{x-\mu}{\sigma}$, we have $X = \mu + \sigma Z$.

Therefore, by linearity, we have

$$\begin{aligned} E(X) &= E(\mu + \sigma Z) \\ &= E(\mu) + \sigma E(Z) \\ &= \mu \end{aligned}$$

By our variance rules, we have

$$\begin{aligned} V(X) &= V(\mu + \sigma Z) \\ &= V(\mu) + V(\sigma Z) \\ &= 0 + \sigma^2 \cdot 1 \\ &= \sigma^2 \end{aligned}$$

□

Theorem 16 (Z-score). If $X \sim N(\mu, \sigma)$, then the random variable $z = \frac{x-\mu}{\sigma}$ has the standard normal distribution.

Proof 20. Use a u-substitution with $z = \frac{x-\mu}{\sigma}$. □

Note. Since the PMF of the standard normal distribution is symmetric about 0, the CDF can be written as

$$F_X(x) = \frac{1}{2} + \int_0^t \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}$$

Example. The height of students obey a normal distribution with $\mu = 67$ inches, and $\sigma = 3$ inches. What percent of students have heights between 64 and 70 inches?

Let $z = \frac{x-\mu}{\sigma}$. Then z follows $N(0, 1)$. So,

$$\begin{aligned} P(64 < x < 70) &= P(-1 < z < 1) \\ &= F(1) - F(-1) \\ &\approx 0.6826 \end{aligned}$$

Remark. The CDF for $X \sim N(0, 1)$ is often denoted $\Phi(x)$. From the above example, we generally have that for $X \sim N(\mu, \sigma)$,

$$\begin{aligned} P(a < X < b) &= P\left(\frac{a-\mu}{\sigma} < z < \frac{b-\mu}{\sigma}\right) \\ &= \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right) \end{aligned}$$

Also, for $Z \sim N(0, 1)$, $\Phi(-z) = 1 - \Phi(z)$.

Example. Find the probabilty that $X \sim N(0,1)$ takes on a value less than -0.88. We can find this by

$$\begin{aligned}\Phi(-0.88) &= \int_{-\infty}^{-0.88} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \\ &= 1 - \int_{-0.88}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \\ &= 1 - \int_{-\infty}^{0.88} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \\ &= 1 - \Phi(0.88) \\ &\approx 0.1894\end{aligned}$$

Example. Radiation exposure in an area takes on a normal distribution with mean $\mu = 4.35$ mrem and standard deviation $\sigma = 0.59$ mrem. What is the probability that a person is exposed to more than 5.2 mrem?

Converting to the standard normal distribution, we get

$$z = \frac{x - \mu}{\sigma} = 1.44 \implies 1 - \Phi(1.44) = 0.0749$$

Definition 38 (Gamma Function)**.** The gamma function is given by

$$\Gamma(\alpha) = \int_0^{\infty} t^{\alpha - 1} e^{-t} \, \mathrm{d}t$$

Theorem 17 (Properties of the Gamma Function)**.**

1. For $\alpha > 0$, $\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1)$, and in particular if α is a positive integer, $\Gamma(\alpha) = (\alpha - 1)!$.
2. We have $\Gamma(\frac{1}{2}) = \sqrt{\pi}$.

Proof 21. (1). Use integration of parts with $u = e^{-t}$, $\mathrm{d}v = t^{\alpha - 1}$. Then

$$\begin{aligned}\Gamma(\alpha) &= (\alpha - 1)\Gamma(\alpha - 1) & &= (\alpha - 1)(\alpha - 2)\Gamma(\alpha - 2) \\ &= (\alpha - 1) \dots (2)\Gamma(1)\end{aligned}$$

Note that

$$\Gamma(1) = \int_0^\infty e^{-t}\,\mathrm{d}t = 1$$

Therefore, $\Gamma(\alpha) = (\alpha-1)!, \alpha \in \mathbb{Z}^+$. □

Proof 22. (2). We do a u substitution with $t = u^2$ and $\mathrm{d}t = 2u\,\mathrm{d}u$.

$$\begin{aligned}\Gamma\left(\frac{1}{2}\right) &= \int_0^\infty t^{-\frac{1}{2}} e^{-t}\,\mathrm{d}t \\ &= \int_0^\infty u^{-1} e^{-u^2} 2u\,\mathrm{d}u \\ &= \int_0^\infty 2e^{-u^2}\,\mathrm{d}u \\ &= \sqrt{\pi}\end{aligned}$$

(See the normal distribution PMF verification proof for details on the final step). □

Definition 39 (Gamma Distribution)**.** A random variable X has a gamma distribution, denoted $X \sim \Gamma(\alpha, \beta) = \text{Gamma}(\alpha, \beta)$, if the PMF is

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

for $x, \alpha, \beta > 0$, where $\Gamma(\alpha)$ is the gamma function, α is the shape parameter, and β is the rate parameter.

Properties:

(a)

$$E(X) = \frac{\alpha}{\beta}$$

(b)

$$V(X) = \frac{\alpha}{\beta^2}$$

Proof 23.

$$\begin{aligned} E(X) &= \frac{1}{\Gamma(\alpha)} \int_0^\infty x\beta^\alpha x^{\alpha-1} e^{-\beta x}\,\mathrm{d}x \\ &= \frac{1}{\Gamma(\alpha)} \int_0^\infty \beta^\alpha x^\alpha e^{-\beta x}\,\mathrm{d}x \end{aligned}$$

Now let $y = \beta x$ and $\mathrm{d}y = \beta\,\mathrm{d}x$.

$$\frac{1}{\Gamma(\alpha)} \int_0^\infty \beta^\alpha x^\alpha e^{-\beta x}\,\mathrm{d}x = \frac{1}{\Gamma(\alpha)} \int_0^\infty y^\alpha e^{-y} \frac{1}{\beta}\,\mathrm{d}y$$

Note that inside the integral, we have $\Gamma(\alpha+1)$. Therefore,

$$\frac{1}{\beta}\frac{\Gamma(\alpha+1)}{\Gamma(\alpha)} = \frac{\alpha\Gamma(\alpha)}{\beta\Gamma(\alpha)} = \frac{\alpha}{\beta}$$

For $E(X^2)$, we can similarly proceed.

$$\begin{aligned} E(X^2) &= \frac{1}{\Gamma(\alpha)} \int_0^\infty \beta^\alpha x^{\alpha+1} e^{-\beta x}\,\mathrm{d}x \\ &= \frac{1}{\Gamma(\alpha)} \frac{1}{\beta} \int_0^\infty (\beta x)^{\alpha+1} e^{-\beta x}\,\mathrm{d}x \\ &= \frac{1}{\Gamma(\alpha)} \frac{1}{\beta^2} \int_0^\infty y^{\alpha+1} e^{-y}\,\mathrm{d}y \\ &= \frac{\Gamma(\alpha+2)}{\Gamma(\alpha)\cdot\beta^2} \\ &= \frac{(\alpha+1)\cdot(\alpha)\cdot\Gamma(\alpha)}{\Gamma(\alpha)\beta^2} \\ &= \frac{\alpha^2+\alpha}{\beta^2} \end{aligned}$$

Therefore, the variance is

$$V(X) = E(X^2) - E(X)^2 = \frac{\alpha^2+\alpha}{\beta^2} - \frac{\alpha^2}{\beta^2} = \frac{\alpha}{\beta^2}$$

□

Example (Intuition: Relating Poisson with Gamma).
Recall the probability of exactly x successes for $X \sim$ Poisson was $p(x) = \frac{\lambda^x}{x!}e^{-\lambda}$, where we

"approximated" with $\lambda = np$ was a constant. This is seen as "rate per unit time". Now for example, suppose

$$X = \#\text{ pulses in a 30 second interval}$$

Assume pulses occur on average 6 per minute. We see that

$$\lambda = 6 \cdot \frac{1}{2} = \text{expect 3 pulses per 30 second interval}$$

Generally, if x successes occur with average rate δ per unit time, the probability of x successes in an interval of t units with $\lambda = \delta t$ is

$$p(x) = \frac{1}{x!}(\delta t)^x e^{-\delta t}$$

How does this relate to the Gamma distribution? Define the random variable $X =$ wait time for α successes. We have that

$$\begin{aligned} CDF = F(x) &= P(\text{wait time for } \alpha \text{ successes} \leq X) \\ &= 1 - P(\text{fewer than } \alpha \text{ successes in } [0, x]) \\ &= 1 - P(\text{exactly 0 successes or exactly 1 success} \ldots \text{or } \alpha - 1 \text{ successes}) \\ &= 1 - \sum_{k=0}^{\alpha-1} \text{Poisson with parameter } \lambda x \end{aligned}$$

We can differentiate $F(x)$ to get the PDF, so we get

$$\begin{aligned} f(x) &= \text{prob of } \alpha \text{ successes in time interval } x = F'(x) \\ &= \frac{\lambda e^{-\lambda x}(\lambda x)^{\alpha-1}}{(\alpha-1)!} \end{aligned}$$

And this is the PDF of the Gamma distribution with $\lambda = \beta$.

Definition 40 (Exponential Distribution)**.** If Random variable X has PMF

$$f(x; \beta) = \beta e^{-\beta x}$$

This is known as the exponential distribution, which is a special case of the Gamma distribution, where $\alpha = 1$, or $X \sim \Gamma(1, \beta)$.

Example. Suppose the total cars exceeding the speed limit in half an hour is a random variable with a Poisson distribution with $\lambda = 8.4$. What is the probability the wait time is at most 5 minutes between cars exceeding the speed limit?

Here we define $t = 1$ which is one unit of 30 minutes. Then

$$8.4 = \lambda = \delta t = \delta \cdot 1 = \delta$$

We want $P(\text{wait time at most 5 min}) =$

$$\begin{aligned} P(\text{wait time at most 5 min}) &= \int_0^{\frac{1}{6}} 8.4e^{-8.4}\,\mathrm{d}x \\ &\approx 0.75 \end{aligned}$$

Remark. It follows by the earlier results, if Y is an exponential distribution,

$$E(Y) = \frac{1}{\beta}$$

and

$$V(Y) = \frac{1}{\beta^2}$$

Definition 41 (Chi Squared)**.** If a random variable X has PMF

$$f(x;\nu) = \frac{1}{2^{\frac{\nu}{2}}\Gamma\left(\frac{\nu}{2}\right)} x^{\frac{\nu-2}{2}} e^{-\frac{x}{2}}$$

This is a special case of the Gamma distribution, where $X \sim \Gamma\left(\frac{\nu}{2}, \frac{1}{2}\right)$.

Remark. This distribution is used a lot in sampling theory.

Remark. It follows from earlier results that

$$E(X) = \nu$$

$$V(X) = 2\nu$$

Definition 42 (Cauchy Distribution)**.** If a random variable X has PMF

$$f(x;\theta) = \frac{1}{\pi}\left(\frac{1}{(x-\theta)^2 + 1}\right)$$

then it follows the Cauchy distribution.

Properties:

(a)
$$E(X) = \text{undefined}$$

(b)
$$V(X) = \text{undefined}$$

This can be thought of as similar to a "slowly decreasing" normal distribution. For $\theta = 0$, we can verify that it is a PMF.

$$\begin{aligned}\frac{1}{\pi}\int_{-\infty}^{\infty}\frac{1}{x^2+1}\,\mathrm{d}x &= \frac{1}{\pi}\left(\lim_{t\to\infty}\arctan t - \lim_{s\to-\infty}\arctan s\right)\\ &= \frac{1}{\pi}\left(\frac{\pi}{2}-\left(-\frac{\pi}{2}\right)\right)\\ &= 1\end{aligned}$$

We have

$$\begin{aligned}E(X) &= \frac{1}{\pi}\int_{-\infty}^{\infty}\frac{x}{1+x^2}\,\mathrm{d}x\\ &= \frac{1}{\pi}\left(\int_{-\infty}^{0}\frac{x}{1+x^2}\,\mathrm{d}x + \int_{0}^{\infty}\frac{x}{1+x^2}\,\mathrm{d}x\right)\\ &= \frac{2}{\pi}\int_{0}^{\infty}\frac{x}{1+x^2}\,\mathrm{d}x\\ &= \frac{1}{\pi}\lim_{t\to\infty}\ln(1+u)\Big|_0^t\end{aligned}$$

This diverges, so the expected value is undefined. Similarly, the variance of the Cauchy distribution is undefined as well.

Definition 43 (Beta Distribution)**.** We say $X \sim \text{Beta}(\alpha, \beta)$ if X is a random variable with PMF

$$f(x;\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1}$$

for $0 < x < 1$ and $\alpha, \beta > 0$.

Properties:

(a)
$$E(X) = \frac{\alpha}{\alpha+\beta}$$

(b)
$$V(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

Proof 24. (Verification of PMF).
Using $\Gamma(\alpha) = \int_0^\infty t^{\alpha-1}e^{-t}\,\mathrm{d}t$ with $\alpha = a+1, \beta = b+1$, we have

$$\begin{aligned}\Gamma(a+1)\Gamma(b+1) &= \Gamma(\alpha)\Gamma(\beta)\\ &= \int_0^\infty x^{\alpha-1}e^{-x}\,\mathrm{d}x \int_0^\infty y^{\beta-1}e^{-y}\,\mathrm{d}y\\ &= \int_0^\infty \int_0^\infty x^a y^b e^{-x-y}\,\mathrm{d}x\,\mathrm{d}y\end{aligned}$$

Now we can do a substitution with $x = uv$ and $y = (1-u)v$. Therefore, $u = \frac{x}{x+y}$ and $v = \frac{y}{1-u}$. Note that as $x \to \infty$, $u \to 1$. This equals

$$\begin{aligned}\Gamma(a+1)\Gamma(b+1) &= \int_0^\infty \int_0^1 u^a(1-u)^b v^{a+b+1}e^{-v}\,\mathrm{d}u\,\mathrm{d}v\\ &= \Gamma(a+b+2)\int_0^1 u^a(1-u)^b\,\mathrm{d}u\end{aligned}$$

Therefore, we see that since this is equal to $\Gamma(a+1)\Gamma(b+1)$, we get

$$\begin{aligned}1 &= \frac{\Gamma(a+b+2)}{\Gamma(a+1)\Gamma(b+1)}\int_0^1 u^a(1-u)^b\,\mathrm{d}u\\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\int_0^1 x^{\alpha-1}(1-x)^{\beta-1}\,\mathrm{d}x\\ &= \int_0^1 f(x;\alpha,\beta)\,\mathrm{d}x\end{aligned}$$

So we see that f is a PMF. □

Proof 25. (Properties).

$$
\begin{aligned}
E(X) &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_0^1 x^{\alpha}(1-x)^{\beta-1}\,\mathrm{d}x \\
&= \left(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\right) \frac{\Gamma(\alpha+1)\Gamma(\beta)}{\Gamma(\alpha+1+\beta)} \frac{\Gamma(\alpha+1+\beta)}{\Gamma(\alpha+1)\Gamma(\beta)} \int_0^1 x^{\alpha}(1-x)^{\beta-1}\,\mathrm{d}x \\
&= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \cdot \frac{\Gamma(\alpha+1)\Gamma(\beta)}{\Gamma(\alpha+1+\beta)} \\
&= \frac{\Gamma(\alpha+\beta)\alpha\Gamma(\alpha)}{\Gamma(\alpha)(\alpha+\beta)\Gamma(\alpha+\beta)} \\
&= \frac{\alpha}{\alpha+\beta}
\end{aligned}
$$

For $E(X)^2$, a similar argument can be made.

$$E(X^2) = \frac{(\alpha+1)\alpha}{(\alpha+\beta+1)(\alpha+\beta)}$$

$$
\begin{aligned}
V(X) &= E(X^2) - (E(X))^2 \\
&= \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}
\end{aligned}
$$

□

Chapter 5

Transforming Random Variables

Given X, how can we find the distribution of $Y = g(X)$? In the discrete case, as long as $g(x)$ is one to one, we can simply substitute it into the random variable function.

Example. If $X \sim \text{Binomial}(n, p)$, find the PMF of $Y = 2X + 3$.
We have $f_X(x) = \binom{n}{x} p^x (1-p)^{n-x}$ for $x = 0, 1, \ldots, n$. Therefore,

$$P(Y = y) = P(2X + 3 = y) = P\left(X = \frac{y-3}{2}\right)$$

If $\frac{y-3}{2} \notin \{0, 1, \ldots, n\}$, then $P(Y = y) = 0$. Therefore, the PMF of Y is

$$f_Y(y) = \begin{cases} \binom{n}{(y-3)/2} p^{\frac{y-3}{2}} (1-p)^{n-\frac{y-3}{2}} & y = 3, 5, \ldots, 2n+3 \\ 0 & \text{otherwise} \end{cases}$$

5.1 Transforming Continuous Random Variables

Theorem 18. Let $f_X(x)$ be the PMF of a continuous random variable. Suppose $y = g(x)$ is strictly increasing or decreasing. Then the PMF of $Y = g(X)$ is

$$F_Y(y) = \begin{cases} f_X(g^{-1}(y)) \left| \frac{\mathrm{d}}{\mathrm{d}y} \left(g^{-1}(y)\right) \right| & y = g(x) \text{ for some } x \\ 0 & \text{otherwise} \end{cases}$$

provided that $\frac{\mathrm{d}}{\mathrm{d}y} \left(g^{-1}(y)\right)$ exists.

Proof 26. Assume $y = g(x)$ is increasing. Then

$$\begin{aligned} P(a < Y < b) &= P\left(g^{-1}(a) < X < g^{-1}(b)\right) \\ &= \int_{g^{-1}(a)}^{g^{-1}(b)} f_X(x)\,\mathrm{d}x \\ &= \int_a^b f_X(g^{-1}(y)) \frac{\mathrm{d}}{\mathrm{d}y}\left(g^{-1}(y)\right)\mathrm{d}y \end{aligned}$$

Therefore, we see that $f_X(g^{-1}(y))\frac{\mathrm{d}}{\mathrm{d}y}\left(g^{-1}(y)\right)$ is the PMF of Y. □

Example. Let the PMF of X be

$$f_X(x) = \begin{cases} e^{-x} & x > 0 \\ 0 & x \le 0 \end{cases}$$

We want to find the PMF of $Y = \sqrt{X}$, which we note is strictly increasing. Then

$$\begin{aligned} y &= \sqrt{x} \\ x &= y^2 \\ \frac{\mathrm{d}}{\mathrm{d}y}\left(g^{-1}(y)\right) &= 2y \\ f_X(g^{-1}(y)) = f_X(y^2) &= e^{-y^2} \end{aligned}$$

Therefore, the PMF $f_Y(y)$ is

$$f_Y(y) = \begin{cases} 2ye^{-y^2} & y > 0 \\ 0 & y \le 0 \end{cases}$$

Example. Let $X \sim \text{Uniform}(0, 4)$. Find the PMF of $Y = \sqrt{X}$.

We have

$$f_X(x) = \begin{cases} \frac{1}{4} & 0 \le x \le 4 \\ 0 & \text{otherwise} \end{cases}$$

Therefore,

$$\begin{aligned} F_Y(y) &= P(Y \le y) \\ &= P(\sqrt{X} \le y) \\ &= P(X \le y^2) \\ &= \int_{-\infty}^{y^2} f_X(x)\,\mathrm{d}x \\ &= \int_{-\infty}^{y^2} \frac{1}{4}\,\mathrm{d}x \\ &= \frac{1}{4}y^2 \end{aligned}$$

Differentiating this gives us the PMF

$$f_Y(y) = \begin{cases} \frac{y}{2} & 0 \le y \le 2 \\ 0 & \text{otherwise} \end{cases}$$

Example. Let $X \sim N(0,1)$ be the standard normal distribution. Find the PMF of $Y = X^2$.

Using the CDF of Y, we have

$$\begin{aligned} F_Y(y) &= P(Y \le y) \\ &= P(X^2 \le y) \\ &= P(|X| \le \sqrt{y}) \\ &= P(-\sqrt{y} \le X \le \sqrt{y}) \\ &= F_X(\sqrt{y}) - F_X(-\sqrt{y}) \end{aligned}$$

Differentiating gives us the PMF as

$$\begin{aligned} f_Y(y) &= f_X(\sqrt{y})\left(\frac{1}{2}y^{-\frac{1}{2}}\right) - f_X(-\sqrt{y})\left(-\frac{1}{2}y^{-\frac{1}{2}}\right) \\ &= \begin{cases} \frac{1}{\sqrt{2\pi}}e^{-\frac{y}{2}}y^{-\frac{1}{2}} & y > 0 \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

Chapter 6

Joint Distributions

Consider two random variables over a joint sample space. For random variables X, Y, we are interested in $P(X = x, Y = y)$.

6.1 Discrete Joint Distributions

Definition 44 (Joint Probability Distribution)**.** If X and Y are discrete random variables, the function

$$f(x, y) = P(X = x, Y = y) = P(X = x \text{ and } Y = y)$$

for each pair of values (x, y) that X, Y take on is the joint probability distribution of X and Y, called the joint PDF.

Theorem 19 (Properties of the Joint PMF)**.**

1. $f(x, y) \geq 0$
2. $\sum_x \sum_y f(x, y) = 1$

Definition 45 (Joint Cumulative Distribution)**.** If X, Y are discrete random variables, the function

$$F(X, Y) = P(X \leq x, Y \leq y) = \sum_{s \leq x} \sum_{t \leq y} f(s, t)$$

where $f(s, t)$ is the PMF at (s, t) is the joint cumulative distribution, or joint CDF.

Theorem 20 (Properties of the Joint CDF)**.**

1. $F(-\infty, -\infty) = 0$

2. $F(\infty, \infty) = 1$
3. If $a < b$ and $c < d$, then $F(a, c) \leq F(b, d)$

6.2 Continuous Joint Distributions

Definition 46 (Joint Probability Distribution). If X and Y are random variables and $f(x, y)$ is defined over $\mathbb{R}^2$, then $f(x, y)$ is the joint probability density function (joint PDF) of X and Y if

$$P(X, Y) = \iint_R f(x, y)\,\mathrm{d}A$$

where R is a region on the xy-plane.

Theorem 21 (Properties of the Joint PDF).

1. $f(x, y) \geq 0$ for all $x, y \in \mathbb{R}$.
2. $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)\,\mathrm{d}y\,\mathrm{d}x = 1$

Definition 47 (Joint Cumulative Distribution). If X and Y are continuous random variables then

$$\begin{aligned} F(x, y) &= P(X \leq x, Y \leq y) \\ &= \int_{-\infty}^{y} \int_{-\infty}^{x} f(s, t)\,\mathrm{d}s\,\mathrm{d}t \end{aligned}$$

where $f(s, t)$ is the joint PDF of X and Y, is the joint cumulative distribution function, or joint CDF of X and Y.

Similarly to the one variable case, we have

$$\begin{aligned} f(x, y) &= \frac{\partial^2}{\partial x \partial y} F(x, y) \\ &= F_{yx} \\ &= \frac{\partial^2}{\partial x \partial y} F(x, y) \\ &= F_{xy} \end{aligned}$$

provided the partials exist and are continuous.

Example. Let

$$f(x,y) = \begin{cases} x+y & 0<x<1, 0<y<1 \\ 0 & \text{otherwise} \end{cases}$$

be the joint PDF of X and Y. Find the joint CDF.

Clearly, if $x, y \leq 0$, then $F(x,y) = 0$. For $0 < x < 1, 0 < y < 1$, we have

$$\begin{aligned} F(x,y) &= \int_0^y \int_0^x (s+t)\,\mathrm{d}s\,\mathrm{d}t \\ &= \frac{1}{2}xy(x+y) \end{aligned}$$

If $x \geq 1$, $0 < y < 1$, we have

$$\begin{aligned} F(x,y) &= P(X \leq x, Y \leq y) \\ &= \int_0^y \int_0^1 (s+t)\,\mathrm{d}s\,\mathrm{d}t \\ &= \frac{1}{2}(y)(y+1) \end{aligned}$$

Similarly, if $y \geq 1$, $0 < x < 1$, we have $F(x,y) = \frac{1}{2}(x)(x+1)$, and for $x, y \geq 1$, we have $F(x,y) = 1$. So,

$$F(x,y) = \begin{cases} 0 & x, y \leq 0 \\ \frac{1}{2}xy(x+y) & 0<x<1, 0<y<1 \\ \frac{1}{2}y(y+1) & x \geq 1, 0<y<1 \\ \frac{1}{2}x(x+1) & y \geq 1, 0<x<1 \\ 1 & x, y \geq 1 \end{cases}$$

Example. Given the joint PDF

$$f(x) = \begin{cases} \frac{x^2+y}{60} & 0 \leq x \leq 3, 0 \leq y \leq 4 \\ 0 & \text{otherwise} \end{cases}$$

find $P\left(|X-1| \leq \frac{1}{2}\right)$.

We have

$$\begin{aligned} P\left(|X-1| \leq \frac{1}{2}\right) &= P\left(\frac{1}{2} \leq X \leq \frac{3}{2}\right) \\ &= \int_0^4 \int_{\frac{1}{2}}^{\frac{3}{2}} \frac{x^2+y}{60}\,\mathrm{d}x\,\mathrm{d}y \\ &= \frac{37}{180} \end{aligned}$$

Example. Given the joint CDF

$$F(x,y) = \begin{cases} (1-e^{-x})(1-e^{-y}) & x,y>0 \\ 0 & \text{otherwise} \end{cases}$$

Find $P(1<X<3, 1<Y<2)$.

First, we need to find $f(x,y)$. Taking the partial derivatives, we have $F_{xy}(x,y) = e^{-(x+y)}$, so

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

which means that

$$P(1<X<3, 1<Y<2) = \int_1^2 \int_1^3 e^{-(x+y)}\,\mathrm{d}x\,\mathrm{d}y \approx 0.074$$

Example. In a study, the hours X spent using their phones, and hours Y spent on the job is approximated by the joint PDF

$$f(x,y) = xye^{-(x+y)}$$

for $x,y \geq 0$. What is the probability that a person spends at least twice as much time on their phone than doing their job?

We want $P(X \geq 2Y)$, which is given by

$$P(X \geq 2Y) = \int_{x=0}^{x=\infty} \int_{y=0}^{y=\frac{x}{2}} xye^{-(x+y)}\,\mathrm{d}y\,\mathrm{d}x$$

This is the region R where $x \geq 2y$.

6.3 Conditional Distributions and Independence

Definition 48 (Discrete Marginal Distribution)**.** If $f(x,y)$ is the joint PMF of a discrete random

variable, then the function $g_X(x) = \sum_y f(x,y)$ for each x that X takes on is the **marginal distribution of** X.

Definition 49 (Continuous Marginal Distribution)**.** If $f(x,y)$ is the joint PDF of a continuous random variable, then the function $g_X(x) = \int_{-\infty}^{\infty} f(x,y)\,\mathrm{d}y$ for $-\infty < x < \infty$ is the **marginal distribution of** X.

Definition 50 (Conditional Distribution)**.** If $f(x,y)$ is the joint PMF or PDF of a random variable X and Y and $g_Y(y)$ is the marginal distribution of Y, then

$$f(x|y) = \frac{f(x,y)}{g_Y(y)}$$

with $g_Y(y) \neq 0$ for each x in X is the **conditional distribution** or conditional PMF of X, given $Y = y$.

Similarly, the conditional CDF of X given $Y = y$ is

$$F(x|y) = P(X \leq x|Y = y) = \sum_{a \leq x} f(a|y)$$

Example. Let the joint PDF of X and Y be

$$f(x,y) = \begin{cases} \frac{2}{3}(x+2y) & 0 < x, y < 1 \\ 0 & \text{otherwise} \end{cases}$$

We want to find $P\left(X \leq \frac{1}{2} \middle| Y = \frac{1}{2}\right)$.

Therefore, we want to find

$$\int_0^{\frac{1}{2}} f\left(x|y = \frac{1}{2}\right) \mathrm{d}x$$

First, the marginal PDF is

$$\begin{aligned} g_Y(y) &= \int_{-\infty}^{\infty} f(x,y)\,\mathrm{d}x \\ &= \int_0^1 f(x,y)\,\mathrm{d}x \\ &= \frac{1}{3}(1+4y) \end{aligned}$$

Then we see that

$$\begin{aligned} f(x|y) &= \frac{f(x,y)}{g_Y(y)} \\ &= \frac{\frac{2}{3}(x+2y)}{\frac{1}{3}(1+4y)} \\ f\left(x\middle|\frac{1}{2}\right) &= \frac{2x+2}{3} \\ P\left(X \le \frac{1}{2}\middle|Y=\frac{1}{2}\right) &= \int_0^{\frac{1}{2}} \frac{2x+2}{3}\,\mathrm{d}x \\ &= \frac{5}{12} \end{aligned}$$

Definition 51 (Independence). Let $g_i(x)$ be the marginal distribution of a random variable X_i. We say that the random variables $X_1, \dots X_n$ are independent if and only if for all $(x_1, \dots x_n)$ that $X_1, \dots X_n$ takes on,we have the joint PMF or PDF is

$$f(x_1, \dots x_n) = g_1(x_1)g_2(x_2)\dots g_n(x_n)$$

Example. Given the joint PDF of X, Y as

$$f(x,y) = \begin{cases} 12xy(1-y) & 0 < x, y < 1 \\ 0 & \text{otherwise} \end{cases}$$

We have $g_X(x) = \int_0^1 f(x,y)\,\mathrm{d}y = 2x$ and $g_Y(y) = \int_0^1 f(x,y)\,\mathrm{d}x = 6(1-y)y$.

Note that $g_X(x)g_Y(y) = f(x,y)$, so X and Y are independent.

Example. Let $X \sim \text{Binom}(n,p)$ and $Y \sim \text{Binom}(m,p)$. Assume X and Y are independent. Define $Z = X + Y$, so $Z \sim \text{Binom}(n+m,p)$ (why?). Determine the PMF of $X|Z$.

First we note that

$$\begin{aligned} P(X+Y=k) &= \sum_{i=0}^{\infty} P(X=i, Y=k-i) \\ &= \sum_{i=0}^{k} P(X=i)P(Y=k-i) \\ &= \sum \text{product of PMFs} \\ &= \sum \binom{n+m}{k} p^k (1-p)^{n+m-k} \end{aligned}$$

Now

$$\begin{aligned} f(x|z) &= \frac{f(x,z)}{g_Z(z)} \\ &= \frac{P(X=x, Z=x+y)}{g_Z(z)} \\ &= \frac{P(X=x, Y=z-x)}{g_Z(z)} \\ &= \frac{P(X=x)P(Y=z-x)}{g_Z(z)} \\ &= \frac{g_X(x)g_Y(y)}{g_Z(z)} \\ &= \frac{\binom{n}{k}p^x(1-p)^{n-x}\binom{m}{z-x}p^{z-x}(1-p)^{m-z+x}}{\binom{n+m}{z}p^z(1-p)^{n+m-z}} \\ &= \frac{\binom{n}{x}\binom{m}{z-x}}{\binom{n+m}{z}} \end{aligned}$$

We see that this is the Hypergeometric distribution. Thus, $X|Z$ is the Hypergeometric distribution while $X \sim \text{Binom}(n,p)$, so X, Z are NOT independent. This is because our conditional distribution $X|Z = \frac{f(x,z)}{g_Z(z)}$ should be equal to $g_X(x)$ if they are independent.

Theorem 22 (Sum of Variables: Convolution)**.** Let X, Y be independent random variables with PMF or PDF $f_X(x), f_Y(y)$. Define $Z = X + Y$. Then (for the continuous case)

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

which is called the **convolution** of the two functions.

Example. Let X, Y be independent random variables and assume that

$$f_X(t) = f_Y(t) = \begin{cases} e^{-t} & t > 0 \\ 0 & \text{otherwise} \end{cases}$$

Then the PDF of $Z = X + Y$ is

$$\begin{aligned} f_Z(z) &= \int_{-\infty}^{\infty} f_X(x) f_Y(z-x)\,\mathrm{d}x \\ &= \int_0^z e^{-x} e^{-(z-x)}\,\mathrm{d}x \\ &= ze^{-z} \end{aligned}$$

Note that this distribution is NOT exponential. In general, $X+Y$ may NOT stay as the same distribution as X and Y.

6.4 Transforming Joint Random Variables

Recall that if $f_X(x)$ is the PDF for X and $y = g(x)$, then the PDF for $Y = g(x)$ is

$$f_Y(y) = f_X(g^{-1}(y)) \frac{\mathrm{d}}{\mathrm{d}y} g^{-1}(y)$$

Now given $f_{X,Y}(x, y)$ and $u = h_1(x, y)$, $v = h_2(x, y)$. What is the joint PDF of U and V?

Theorem 23 (Transforming Joint Distributions)**.** If random variables X, Y and $U = h_1(X, Y)$, $V = h_2(X, Y)$, and $f_{X,Y}(x, y)$ is the joint PDF of X and Y, then the joint PDF of U and V is

$$f_{U,V}(u, v) = f_{X,Y}\left(x(u, v), y(u, v)\right) |J(u, v)|$$

Where $J(u, v) = \det \begin{bmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{bmatrix}$ is the Jacobian.

Remark. One can show that with $U = h_1(X, Y)$, $V = h_2(X, Y)$, we have

$$J^{-1} = \frac{1}{\det \begin{bmatrix} \frac{\partial h_1}{\partial x} & \frac{\partial h_1}{\partial y} \\ \frac{\partial h_2}{\partial x} & \frac{\partial h_2}{\partial y} \end{bmatrix}}$$

though this isn't always useful as we may still need to solve for x and y in terms of u and v.

Example. Let the joint PDF of X and Y be

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

Find the joint PDF of $U = X+Y$, $V = \frac{X}{X+Y}$.

We have that $u = x+y$, $v = \frac{x}{x+y}$. We see that $v = \frac{x}{u}$, so $x = uv$. We also have $u = uv+y$, so $y = u-uv$.

This allows us to calculate

$$\begin{aligned} J(u,v) &= \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix} \\ &= \begin{vmatrix} v & u \\ 1-v & -u \end{vmatrix} \\ &= -u \end{aligned}$$

So the joint PDF is

$$\begin{aligned} f_{U,V}(u,v) &= f_{X,Y}(x(u,v),y(u,v))|J(u,v)| \\ &= f_{X,Y}(uv, u-uv)\cdot u \\ &= e^{-(uv+u-uv)}\cdot u \\ &= \begin{cases} ue^{-u} & u>0, 0<v<1 \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

Example. Let $X \sim N(\mu_1, \sigma_1^2)$, $Y \sim N(\mu_2, \sigma_2^2)$, and assume X, Y are independent. Find the joint PDF of $U = X+Y$, $V = X-Y$.

Being independent , we have $f_{X,Y} = f_X(x)\cdot f_Y(y)$. Note that $x = \frac{u+v}{2}$, $y = \frac{u-v}{2}$, and so

$$J = \begin{vmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} \end{vmatrix} = -\frac{1}{2}$$

Therefore, we get

$$\begin{aligned} f_{U,V}(u,v) &= f_{X,Y}\left(\frac{u+v}{2}, \frac{u-v}{2}\right) \cdot \frac{1}{2} \\ &= \frac{1}{2} f_X\left(\frac{u+v}{2}\right) f_Y\left(\frac{u-v}{2}\right) \\ &= \frac{1}{4\pi\sigma_1\sigma_2} \exp\left(\frac{\left(\frac{u+v}{2}-\mu_1\right)^2}{2\sigma_1^2} + \frac{\left(\frac{u-v}{2}-\mu_2\right)^2}{2\sigma_2^2}\right) \end{aligned}$$

Example. Let X, Y have joint PDF

$$f(x,y) = \begin{cases} \frac{1}{96}xy & 0 < x < 4, 1 < y < 5 \\ 0 & \text{otherwise} \end{cases}$$

Find the PDF of $U = X + 2Y$.

Let $u = x + 2y$ and $v = x$. Then $x = v$ and $y = \frac{u-v}{2}$. Note that this solves a different problem! We need to do further manipulation afterwards.

We have $J(u,v) = \begin{vmatrix} 0 & 1 \\ \frac{1}{2} & -\frac{1}{2} \end{vmatrix} = -\frac{1}{2}$. Our new bounds are $0 < v < 4$ and $1 < \frac{u-v}{2} < 5$, so $v + 2 < u < v + 10$.

The joint PDF is

$$\begin{aligned} f_{U,V}(u,v) &= f_{X,Y}\left(v, \frac{u-v}{2}\right) \cdot \left|-\frac{1}{2}\right| \\ &= \begin{cases} \frac{1}{96} v \left(\frac{u-v}{2}\right)\left(\frac{1}{2}\right) & 0 < v < 4, v+2 < u < v+10 \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

However, we want the distribution of just U, which requires us to find the marginal distribution $g_U(u)$. We find that

$$g_U(u) = \begin{cases} \int_{v=0}^{v=u-2} f_{U,V}(u,v)\,\mathrm{d}v & 2 < u < 6 \\ \int_{v=0}^{v=4} f_{U,V}(u,v)\,\mathrm{d}v & 6 < u < 10 \\ \int_{v=u-10}^{v=4} f_{U,V}(u,v)\,\mathrm{d}v & 10 < u < 14 \end{cases}$$

Draw the picture! The distribution forms a parallelogram on the u, v space.

We can also try to do this by using the CDF.

$$\begin{aligned} F_U(u) &= P(X + 2Y \le u) \\ &= \iint_{x+2y\le u} f(x,y)\,\mathrm{d}A \end{aligned}$$

For $2 < u < 6$, we have

$$P(X + 2Y \le u) = \int_{x=0}^{x=u-2} \int_{y=1}^{y=-\frac{x}{2}+\frac{u}{2}} f(x,y)\,\mathrm{d}y\,\mathrm{d}x$$

as the CDF for $2 < u < 6$. Then to get the PDF, we must differentiate. This yields $\frac{(u-2)^2(u+4)}{2304}$ for $2 < u < 6$, which equals the integral we got before.

6.5 Expected Value and Variance of Joint Variables

Theorem 24 (Expected Value)**.** Let X, Y be random variables. If $f(x, y)$ is the joint PMF or PDF, then

- In the discrete case,
$$E(g(X,Y)) = \sum_x \sum_y g(x,y) f(x,y)$$
- In the continuous case,
$$E(g(X,Y)) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y) f(x,y)\,\mathrm{d}y\,\mathrm{d}x$$

Corollary (Linearity of expectation)**.** Let $c_1, \dots c_n$ be constants, and $X_1, \dots X_k$ be random variables. Then

$$E\left(\sum_{i=1}^{n} c_i g_i\left(X_1, \dots X_k\right)\right) = \sum_{i=1}^{n} c_i E\left(g_i\left(X_1, \dots X_k\right)\right)$$

In particular, for X, Y and $c_1 = c_2 = 1$,

$$E(X + Y) = E(X) + E(Y)$$

Example. Let the PDF of X and Y be

$$f(x,y) = \begin{cases} \frac{2}{7}(x + 2y) & 0 < x < 1, 1 < y < 2 \\ 0 & \text{otherwise} \end{cases}$$

Then we see that

$$\begin{aligned} E\left(\frac{X}{Y^3}\right) &= \int_1^2 \int_0^1 \left(\frac{x}{y^3}\right) \frac{2}{7}(x+2y)\,\mathrm{d}x\,\mathrm{d}y \\ &= \frac{10}{56} \end{aligned}$$

Theorem 25. If X and Y are independent, then $E(XY) = E(X)E(Y)$.

Proof 27. Recall that X and Y are independent if and only if $f_{X,Y}(x,y) = f_X(x)f_Y(y)$. So if X and Y are independent, we find that

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

□

Example. Consider the joint PMF of X and Y given by

Table 6.1: Joint PMF

Y, X	-1	0	1
0	1	$\frac{1}{6}$	$\frac{1}{12}$
1	$\frac{1}{4}$	0	$\frac{1}{2}$

Then

$$\begin{aligned} E(X) &= -1 \cdot \frac{1}{4} + 0 + 1 \cdot \left(\frac{1}{2} + \frac{1}{2}\right) \\ &= \frac{1}{3} \\ E(Y) &= 0 + 1 \cdot \left(\frac{1}{4} + \frac{1}{2}\right) \\ &= \frac{3}{4} \\ E(XY) &= 1 \cdot -1 \cdot \left(\frac{1}{4}\right) + 1 \cdot 1 \cdot \frac{1}{2} \\ &= \frac{1}{4} \end{aligned}$$

so $E(XY) = E(X)E(Y)$. Now

$$\begin{aligned} P(X = -1|Y = 1) &= \frac{P(X = -1, Y = 1)}{P(Y = 1)} \\ &= \frac{\frac{1}{4}}{\frac{1}{4} + \frac{1}{2}} \\ &= \frac{1}{3} \end{aligned}$$

However, $P(X = -1) = \frac{1}{4}$, which does not equal what we calculated above. So $E(XY) = E(X)E(Y)$ DOES NOT imply independence, and the converse is false.

Definition 52 (Covariance)**.** The **covariance** of X and Y is

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

Remark. This measures a relationship with X and Y when they may not necessarily be independent.

Theorem 26.

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

Proof 28.

$$\begin{aligned}\mathrm{Cov}(X,Y) &= E(XY - E(X)Y - XE(Y) + E(X)E(Y)) \\ &= E(XY) - E(X)E(Y) - E(X)E(Y) + E(X)E(Y) \\ &= E(XY) - E(X)E(Y)\end{aligned}$$

□

Corollary. If X and Y are independent, then $\mathrm{Cov}(X,Y) = 0$ ("uncorrelated").

Theorem 27 (Properties of Covariance)**.**

- $\mathrm{Cov}(X,X) = V(X)$
- $\mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,X)$
- $\mathrm{Cov}(kX,Y) = k\,\mathrm{Cov}(Y,X)$

Note. We note that $\mathrm{Cov}(X,Y)$ is a measure of the linear relationship of X and Y, and the slope of the line depends on $\mathrm{Cov}(X,Y)$.

Definition 53 (Correlation Coefficient)**.** The **correlation coefficient** of X and Y is

$$\rho_{X,Y} = \frac{\mathrm{Cov}(X,Y)}{\sqrt{V(X)V(Y)}}$$

Theorem 28 (Correlation Coefficient Properties)**.** For random variables X, Y,

- $-1 \le \rho_{X,Y} \le 1$
- $\rho_{X,Y} = 0$ if X, Y are independent.
- $\rho_{X,Y} = \pm 1$ if and only if $Y = \alpha X + \beta$ for some $\alpha, \beta \in \mathbb{R}$

Example. Let the PDF of X, Y be

$$f(x,y) = \begin{cases} 2 & x > 0, y > 0, x + y < 1 \\ 0 & \text{otherwise} \end{cases}$$

Find $\mathrm{Cov}(X,Y)$.

We have

$$\begin{aligned} E(X) &= \int_0^1 \int_0^{1-x} x(2)\,\mathrm{d}y\,\mathrm{d}x \\ &= \frac{1}{3} \\ E(Y) &= \int_0^1 \int_0^{1-x} y(2)\,\mathrm{d}y\,\mathrm{d}x \\ &= \frac{1}{3} \\ E(XY) &= \int_0^1 \int_0^{1-x} xy(2)\,\mathrm{d}y\,\mathrm{d}x \\ &= \frac{1}{12} \end{aligned}$$

So $\mathrm{Cov}(X,Y) = E(XY) - E(X)E(Y) = \frac{1}{18}$. Moreover, can find

$$\begin{aligned} V(X) &= E(X^2) - (E(X))^2 \\ &= \frac{1}{18} \\ V(Y) &= \frac{1}{18} \\ \rho_{X,Y} &= \frac{\mathrm{Cov}(X,Y)}{\sqrt{V(X)V(Y)}} \\ &= 1 \end{aligned}$$

Theorem 29. Let $X_1, \dots X_n$ be random variables, and assume

$$Y_1 = \sum_{i=1}^{n} a_i X_i$$

$$Y_2 = \sum_{i=1}^{n} b_i X_i$$

Then

$$\mathrm{Cov}(Y_1, Y_2) = \sum_{i=1} a_i b_i V(X_i) + 2\sum_{i<j} a_i b_j \,\mathrm{Cov}(X_i, X_j)$$

Proof 29. "tedious algebra" □

Corollary. If $Y = X_1 + \ldots + X_n$, then

$$\begin{aligned}\operatorname{Cov}(Y,Y) &= V(Y)\\ &= \sum_{i=1}^{n} V(X_i) + 2\sum_{i<j} \operatorname{Cov}(X_i, X_j)\end{aligned}$$

Example. Let the PDF of X, Y be

$$f(x,y) = \begin{cases} \frac{1}{3}(x+y) & 0 < x < 1, 0 < y < 2 \\ 0 & \text{otherwise} \end{cases}$$

Find $V(Z)$, where $Z = 3X + 4Y - 5$.

Since the variance of a constant is 0 and $\operatorname{Cov}(X, 5) = \operatorname{Cov}(Y, 5) = 0$, we want

$$V(3X + 4Y) = 9V(X) + 16V(Y) + 2 \cdot (3 \cdot 4 \operatorname{Cov}(X,Y))$$

We find

$$\begin{aligned}E(X) &= \int_0^1 x f(x,y)\,\mathrm{d}y\,\mathrm{d}x\\ &= \frac{5}{9}\\ E(X^2) &= \int_0^1 \int_0^2 x^2 f(x,y)\,\mathrm{d}y\,\mathrm{d}x\\ &= \frac{7}{18}\\ V(X) &= E(X^2) - (E(X))^2\\ &= \frac{13}{162}\\ E(Y) &= \frac{4}{9}\\ E(Y^2) &= \frac{16}{9}\\ V(Y) &= \frac{23}{81}\end{aligned}$$

and

$$E(XY) = \int_0^1 \int_0^2 xy f(x,y)\,\mathrm{d}y\,\mathrm{d}x = \frac{2}{3}$$

so $\operatorname{Cov}(X,Y) = E(XY) - E(X)E(Y) = -\frac{1}{81}$. Thus, $V(3X + 4Y) = \frac{805}{162}$.

Example. Suppose we have 10 chips, 4 red and 6 blue. Let Y be the number of red chips drawn, where each draw has equal probability of getting a red chip with replacement. Find $E(Y)$ in 3 trials.

Let

$$X_i = \begin{cases} 1 & \text{red chip on trial } i \\ 0 & \text{otherwise} \end{cases}$$

be our indicator random variable. We want

$$\begin{aligned} E(Y) &= E\left(\sum_{i=1}^{3} X_i\right) \\ &= \sum_{i=1}^{3} E(X_i) \\ &= 3\left(1 \cdot P(X_1 = 1) + 0 \cdot P(X_1 = 0)\right) \\ &= 3P(X_1 = 1) \\ &= \frac{6}{5} \end{aligned}$$

Also,

$$\begin{aligned} V(Y) &= \sum_{i=1}^{3} V(X_i) \\ &= \sum_{i=1}^{3} E(X^2) - (E(X))^2 \\ &= 3\left(\frac{4}{10} - \left(\frac{4}{10}\right)^2\right) \\ &= \frac{18}{25} \end{aligned}$$

Example. Recall that $Y \sim \text{Hypergeometric}(n, m, k)$ measured the probability of m successes when drawing k elements without replacement from a set of n elements.

In our example, let $n = 10$, $m = 4$, $k = 3$, and let

$$X_i = \begin{cases} 1 & \text{red chip on trial } i \\ 0 & \text{otherwise} \end{cases}$$

We have $E(X_1) = \frac{4}{10}$. For X_2, we have

$$\begin{aligned} E(X_2) &= 1 \cdot P(X_2 = 1) + 0 \cdot P(X_2 = 0) \\ &= P(X_2 = 1) \\ &= \left(\frac{4}{10}\right)\left(\frac{3}{9}\right) + \left(\frac{6}{10}\right)\left(\frac{4}{9}\right) \\ &= \frac{4}{10} \end{aligned}$$

which we notice is the same as $E(X_1)$. This generalizes for every X_i, so

$$\begin{aligned} E(Y) &= E\left(\sum_{i=1}^{3} X_i\right) \\ &= 3\left(\frac{4}{10}\right) \\ &= \frac{6}{5} \end{aligned}$$

We have that the expected values are the same with or without replacement. We can compute the variance as well:

$$\begin{aligned} V(Y) &= \sum_{i=1}^{3} V(X_i) + 2 \sum_{1 \le i < j < 3} \text{Cov}(X_i, X_j) \\ &= \frac{14}{25} \end{aligned}$$

This is a smaller variance than the previous example. This is reasonable, since when we draw without replacement, the sample size gets smaller each time.

Chapter 7

Moments and Moment Generating Functions

7.1 Univariate Moment Generating Functions

Definition 54 (Moment)**.** Let X be a random variable. The **r'th moment of X about the origin** is

$$\begin{aligned}\mu'_r &= E(X^r) \\ &= \sum_x x^r f(x) \quad \text{(discrete case)} \\ &= \int_{-\infty}^{\infty} x^r f(x)\,\mathrm{d}x \quad \text{(continuous case)}\end{aligned}$$

for any non-negative integer r.

Note. Note that $\mu'_1 = E(X)$.

Definition 55 (Moment about the mean)**.** Let X be a random variable. The **r'th moment**

about the mean μ is

$$\begin{aligned}\mu_r &= E((X-\mu)^r) \\ &= \sum_x (x-\mu)^r f(x) \ \text{(discrete case)} \\ &= \int_{-\infty}^{\infty} (x-\mu)^r f(x)\,\mathrm{d}x \ \text{(continuous case)}\end{aligned}$$

for any non-negative integer r.

Note. Note that $\mu_2 = E((X-\mu)^2) = V(X)$.

Definition 56 (Moment Generating Function)**.** The **moment generating function** or MGF of a random variable X is

$$\begin{aligned}M_X(t) &= E\left(e^{tX}\right) \\ &= \sum_x e^{tx} f(x) \ \text{(discrete case)} \\ &= \int_{-\infty}^{\infty} e^{tx} f(x)\,\mathrm{d}x \ \text{(continuous case)}\end{aligned}$$

Why do we define this? Note that expanding

$$\begin{aligned}e^{tx} &= 1 + tx + \frac{(tx)^2}{2!} + \frac{(tx)^3}{3!}\dots \\ \int_{-\infty}^{\infty} e^{tx} f(x)\,\mathrm{d}x &= \int_{-\infty}^{\infty} \left(1 + tx + \frac{(tx)^2}{2!} + \dots\right) f(x)\,\mathrm{d}x \\ M_X(t) &= \int_{-\infty}^{\infty} f(x)\,\mathrm{d}x + t\int_{-\infty}^{\infty} x f(x)\,\mathrm{d}x + \frac{t^2}{2!}\int_{-\infty}^{\infty} x^2 f(x)\,\mathrm{d}x \dots \\ &= 1 + \mu_1' t + \mu_2' \frac{t^2}{2} + \dots + \dots \mu_r' \frac{t^r}{r!} + \dots \\ &= \sum_{n=0}^{\infty} \mu_n' \frac{t^n}{n!}\end{aligned}$$

So the coefficient of $\frac{t^n}{n!}$ of e^{tx} is μ_n', the n'th moment.

Theorem 30. We have

$$M_X^{(n)}(0) = \frac{\mathrm{d}^n}{\mathrm{d}t^n}\left(M_X(n)\right)\Big|_{t=0} = \mu_n'$$

Proof 30. By differentiating n times, all lower powers on n will vanish, but higher powers will still have t involved. The theorem follows from the fact that

$$\frac{\mathrm{d}^n}{\mathrm{d}t^n}\left(\mu_n'\frac{t^n}{n!}\right) = \mu_n'$$

□

Example. Let the PDF of X be

$$f(x) = \begin{cases} e^{-x} & x > 0 \\ 0 & \text{otherwise} \end{cases}$$

Find the MGF of $f(x)$ and a simple expression for μ_r'.

We have

$$\begin{aligned} M_X(t) &= E(e^{tX}) \\ &= \int_0^\infty e^{tx}e^{-x}\,\mathrm{d}x \\ &= \int_0^\infty e^{-x(1-t)}\,\mathrm{d}x \\ &= \frac{1}{1-t} \qquad |t| < 1 \\ &= \sum_{n=0}^\infty t^n \\ &= \sum_{n=0}^\infty n!\frac{t^n}{n!} \end{aligned}$$

So we see that $\mu_r' = r!$.

Alternatively, we can evaluate the derivatives and find a pattern:

$$\begin{aligned} M_X'(t) &= \frac{1}{(1-t)^2} \\ M_X''(t) &= \frac{2}{(1-t)^3} \\ \cdots M_X^{(r)}(t) &= \frac{1 \cdot 2 \cdot 3 \cdot \ldots r}{(1-t)^{r+1}} \end{aligned}$$

So $\mu_n' = M_X^{(n)}(0) = n!$.

Example. Let random variable X have PMF of

$$f(x) = \begin{cases} \frac{1}{8}\binom{3}{x} & x = 0, 1, 2, 3 \\ 0 & \text{otherwise} \end{cases}$$

Find the MGF of X, and determine the second moment about the origin.

We have

$$\begin{aligned} M_X(t) &= E\left(e^{tX}\right) \\ &= \frac{1}{8}\sum_{x=0}^{3} e^{tx}\binom{3}{x} \\ &= \frac{1}{8}\left(1 + 3e^t + 3e^{2t} + e^{3t}\right) \end{aligned}$$

So $\mu_2' = M_X''(0) = 3$.

7.2 Moment Generating Functions of Some Discrete Distributions

Definition 57 (MGF of the Uniform Distribution)**.** Let X be a random variable with PMF

$$p(x) = \frac{1}{n} \text{ for } x = 1, 2, \ldots, n$$

Then

$$M_X(t) = \frac{e^t}{n}\left(\frac{1 - e^{tn}}{1 - e^t}\right)$$

Proof 31. We have

$$\begin{aligned} M_X(t) &= E(e^{tX}) \\ &= \sum_{x=1}^{n} e^{tx} \frac{1}{n} \\ &= \sum_{x=0}^{n-1} e^{t(x+1)} \frac{1}{n} \\ &= \frac{e^t}{n} \sum_{x=0}^{n-1} \left(e^t\right)^x \\ &= \frac{e^t}{n} \left(\frac{1 - e^{tn}}{1 - e^t} \right) \end{aligned}$$

□

Definition 58 (MGF of the Binomial Distribution)**.** Let $X \sim \text{Binom}(n, p)$. Then we have

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

Proof 32. The PMF is $f(x; n, p) = \binom{n}{x} p^x (1-p)^{n-x}$ for $x = 0, \ldots n$. We have

$$\begin{aligned} M_X(t) &= E(e^{tX}) \\ &= \sum_{x=0}^{n} e^{tx} \binom{n}{x} p^x (1-p)^{n-x} \\ &= \sum_{x=0}^{n} \binom{n}{x} (pe^t)^x (1-p)^{n-x} \\ &= (pe^t + (1-p))^n \end{aligned}$$

□

Definition 59 (MGF of the Geometric Distribution)**.** Let $X \sim \text{Geom}(p)$. Then we have

$$M_X(t) = \frac{(pe^t)^r}{(1 - (1-p)e^t)^r}$$

for $t < -\ln(1-p)$

Proof 33. We see that the PMF is $p(x) = (1-p)^{x-1}p$ for $x = 1, 2, \ldots$. We have

$$\begin{aligned} M_X(t) &= E(e^{tX}) \\ &= \sum_{x=0}^{\infty} e^{tx}(1-p)^x p \\ &= p\sum_{x=0}^{\infty}((1-p)e^t)^x \\ &= \frac{p}{1-(1-p)e^t} \end{aligned}$$

This only holds when $t < -\ln(1-p)$, as otherwise we cannot use the geometric series to evaluate this. □

Definition 60 (MGF of the Negative Binomial)**.** Let $X \sim \text{NegativeBinomial}(r, p)$. Then

$$M_X(t) = \frac{(pe^t)^r}{(1-(1-p)e^t)^r}$$

for $t < -\ln(1-p)$

Proof 34. We see the PMF is $f(x; r, p) = \binom{x-1}{r-1}p^r(1-p)^{x-r}$ for $x = r, r+1, \ldots$. This is similar to the geometric distribution, and following a similar argument we get

$$M_X(t) = \frac{(pe^t)^r}{(1-(1-p)e^t)^r}$$

for $t < -\ln(1-p)$ □

Definition 61 (MGF of the Hypergeometric)**.** Get pranked there is no easy formula :)

Definition 62 (MGF of the Poisson Distribution)**.** Let $X \sim \text{Poisson}(\lambda)$. Then

$$M_X(t) = \exp\left(\lambda(e^t - 1)\right)$$

Proof 35. The Poisson PMF is $f(x;\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}$. We have

$$\begin{aligned} M_X(t) &= E(e^{tX}) \\ &= \sum_{x=0}^{\infty} e^{tx} \frac{\lambda^x e^{-\lambda}}{x!} \\ &= \sum_{x=0}^{\infty} \frac{(e^t\lambda)^x e^{-\lambda}}{x!} \\ &= e^{-\lambda} e^{e^t\lambda} \\ &= \exp\big(\lambda(e^t - 1)\big) \end{aligned}$$

□

Remark. Observe that we can use the MGF to find $E(X)$ and $V(X)$. For example, for the Poisson distribution, we have $M_X'(t) = e^{\lambda(e^t-1)}(\lambda e^t)$ and $M_X'(0) = \lambda = E(X)$. Moreover, $M_X''(0) = E(X^2)$, which allows us to find $V(X)$.

Lemma 2. If $Y = aX + b$, then

$$M_Y(t) = e^{bt} M_X(at)$$

Proof 36.

$$\begin{aligned} M_Y(t) &= E(e^{tY}) \\ &= E\left(e^{t(aX+b)}\right) \\ &= E\left(e^{tb}e^{atX}\right) \\ &= e^{tb}E(e^{atX}) \\ &= e^{bt}M_X(at) \end{aligned}$$

□

7.3 Moment Generating Functions of Some Continuous Distributions

Definition 63 (MGF of the Uniform Distribution)**.** For $X \sim U(\alpha, \beta)$, we have

$$M_X(t) = \frac{1}{t(\beta - \alpha)} \left(e^{\beta t} - e^{\alpha t}\right)$$

Proof 37. We know that the PDF is $f(x; \alpha, \beta) = \frac{1}{\beta - \alpha}$ for $\alpha < x < \beta$. We have

$$\begin{aligned} M_X(t) &= E(e^{tX}) \\ &= \int_\alpha^\beta \frac{1}{\beta - \alpha} e^{tx} \, \mathrm{d}x \\ &= \frac{1}{t(\beta - \alpha)} \left(e^{\beta t} - e^{\alpha t}\right) \end{aligned}$$

□

Definition 64 (MGF of the Normal Distribution)**.** If $X \sim N(\mu, \sigma^2)$, we have

$$M_X(t) = \exp\left(\mu t + \frac{(\sigma t)^2}{2}\right)$$

Proof 38. Note that if $Z \sim N(0, 1)$, then $X = \mu + \sigma Z$. The PDF of Z is

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

Then

$$\begin{aligned}
M_Z(t) &= E(e^{tZ}) \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{tz} e^{-z^2/2} \, \mathrm{d}z \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-z^2/2+tz} \, \mathrm{d}z \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(z^2-2tz)/2} \, \mathrm{d}z \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-((z-t)^2-t^2)/2} \, \mathrm{d}z \\
&= e^{t^2/2} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(z-t)^2/2} \, \mathrm{d}z \\
&= e^{t^2/2} \int_{-\infty}^{\infty} N(\mu = t, \sigma^2 = 1) \, \mathrm{d}z \\
&= e^{t^2/2}
\end{aligned}$$

So the MGF of X is

$$\begin{aligned}
M_X(t) &= M_{\mu+\sigma Z}(t) \\
&= e^{\mu t} M_Z(\sigma t) \\
&= e^{\mu t} e^{(\sigma t)^2/2} \\
&= \exp\left(\mu t + \frac{(\sigma t)^2}{2}\right)
\end{aligned}$$

□

Definition 65 (MGF of the Gamma Distribution). For $X \sim \text{Gamma}(\alpha, \beta)$, we have

$$M_X(t) = \left(1 - \frac{t}{\beta}\right)^{-\alpha}$$

Proof 39. We see that

$$\begin{aligned}
M_X(t) &= \int_0^\infty e^{tx} f(x)\,\mathrm{d}x \\
&= \int_0^\infty \frac{\beta^\alpha x^{\alpha-1}}{\Gamma(a)} e^{-(\beta-t)x}\,\mathrm{d}x \\
&= \frac{\beta^\alpha}{\Gamma(\alpha)(\beta-t)} \int_0^\infty \left(\frac{u}{\beta-t}\right)^{\alpha-1} e^{-u}\,\mathrm{d}u \text{ where we u-sub with } u=(\beta-t)x \\
&= \frac{\beta^\alpha}{\Gamma(\alpha)(\beta-t)^\alpha} \int_0^\infty u^{\alpha-1} e^{-u}\,\mathrm{d}u \\
&= \frac{\beta^\alpha}{\Gamma(\alpha)(\beta-t)^\alpha}\Gamma(\alpha) \\
&= \left(1-\frac{t}{\beta}\right)^{-\alpha}
\end{aligned}$$

□

Definition 66 (MGF of the Beta Distribution)**.** Pranked, no easy formula.

7.4 Joint Moments, Sums, and Products

Definition 67 (Joint MGF)**.** Let $X_1, \dots X_n$ be random variables and define $\boldsymbol{X} = (X_1, \dots X_n)$, and $\boldsymbol{t} = (t_1, \dots t_n)$. The function

$$\begin{aligned}
M_{\boldsymbol{X}}(t_1, \dots t_n) &= M(\boldsymbol{t}) \\
&= E(\exp(t_1 x_1 + \dots + t_n x_n)) \\
&= E(\exp(\boldsymbol{t} \cdot \boldsymbol{x}))
\end{aligned}$$

is the Joint MGF of $X_1, \dots X_n$.

Lemma 3. Let $\boldsymbol{a} = (a_1, \dots a_n)$ and $\boldsymbol{b} = (b_1, \dots b_n)$. If $\boldsymbol{Y} = \boldsymbol{a} \cdot \boldsymbol{X} + \boldsymbol{b} \cdot \boldsymbol{1} = (a_1 X_1 + b_1) + (a_2 X_2 + b_2) + \dots$, then

$$M_{\boldsymbol{Y}}(\boldsymbol{t}) = e^{\boldsymbol{t}\cdot\boldsymbol{b}} M_{\boldsymbol{X}}(\boldsymbol{a} \cdot \boldsymbol{t})$$

Note. Note that $M_{\boldsymbol{X}}(0, 0, \dots, t_i, \dots, 0, 0) = E(e^{t_i X_i}) = M_{X_i}(t_i)$.

Theorem 31 (Uniqueness). Two random variables X and Y have the same distribution if and only if $M_X(t) = M_Y(t)$.

Theorem 32 (Independence). If $X_1, \ldots, X_n$ are independent, then the joint MGF is

$$M_{\boldsymbol{X}}(\boldsymbol{t}) = \prod_{i=1}^{n} M_{X_i}(t_i)$$

Theorem 33 (Expectation from Moments). For random variable $\boldsymbol{X}$,

$$\frac{\partial^2}{\partial t_i \partial t_j} M_{\boldsymbol{X}}(\boldsymbol{t})\Big|_{t_1=t_2=\ldots t_n=0} = E(X_i X_j)$$

Remark. From the product of moments, if we look at the univariate case, say $Z = X + Y$ where X, Y are independent, we get

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

Example. Suppose $X, Y \sim U(-1, 1)$. From earlier, we find that

$$M_X(t) = \frac{1}{2t}\left(e^t - e^{-t}\right)$$

If X and Y are independent, then

$$\begin{aligned} M_{X+Y}(t) &= \left(\frac{e^t - e^{-t}}{2t}\right)^2 \\ &= \frac{e^{2t} + 2e^{-2t}}{4t^2} \end{aligned}$$

which does NOT remain uniform.

Example. Let $X_1, \dots X_n$ be independent, and assume

$$X_i \sim \text{Gamma}(\alpha_i, \beta)$$

To obtain the MGF of $Y = \sum_{i=1}^{n} X_i$, we get

$$\begin{aligned} M_Y(t) &= \prod_{i=1}^{n} M_{X_i}(t) \\ &= \prod_{i=1}^{n} \left(1 - \frac{t}{\beta}\right)^{-\alpha_i} \\ &= \left(1 - \frac{t}{\beta}\right)^{-\sum_{i=1}^{n} \alpha_i} \end{aligned}$$

so $M_Y(t)$ is the MGF of Gamma $\left(\sum_{i=1}^{n} \alpha_i, \beta\right)$. Therefore, the sum of independent Gamma random variables remains Gamma, provided that β is the same for all random variables.

Example. Let the joint PDF of X and Y be

$$f_{X,Y}(x,y) = \begin{cases} e^{-y} & 0 < x < y < \infty \\ 0 & \text{otherwise} \end{cases}$$

Find the joint MGF.

We have

$$\begin{aligned} M_{X,Y}(t_0, t_1) &= E\left(e^{t_0 X + t_1 Y}\right) \\ &= \int_{x=0}^{x=\infty} \int_{y=x}^{y=\infty} e^{t_0 x + t_1 y}(e^{-y}) \, \mathrm{d}y \, \mathrm{d}x \\ &= \int_{x=0}^{x=\infty} \int_{y=x}^{y=\infty} e^{t_0 x} e^{(t_1 - 1)y} \, \mathrm{d}y \, \mathrm{d}x \\ &= \int_{0}^{\infty} \frac{1}{t_1 - 1} \left(-e^{t_0 x} e^{(t_1 - 1)x}\right) \mathrm{d}x \\ &= \int_{0}^{\infty} \frac{1}{t_1 - 1} \left(-e^{x(t_0 + t_1 - 1)}\right) \mathrm{d}x \\ &= \frac{1}{t_1 - 1} \cdot \frac{1}{t_0 + t_1 - 1} \end{aligned}$$

for $t_0 + t_1 < 1, t < 1$. We can get $M_X(t)$ as $M_{X,Y}(t, 0) = \frac{1}{1-t}$.

Example. Let the joint PDF of X, Y be

$$f_{X,Y}(x,y) = \begin{cases} \frac{1}{x!(y-x)!} & y = 0, 1, \ldots, x = 0, 1, \ldots y \\ 0 & \text{otherwise} \end{cases}$$

Find the joint MGF of X, Y.

We have

$$\begin{aligned} M_{X,Y}(s,t) &= E\left(e^{sX+tY}\right) \\ &= \sum_{y=0}^{\infty} \sum_{x=0}^{y} \left(e^{sx+ty}\right) \frac{1}{x!(y-x)!} \\ &= \sum_{y=0}^{\infty} e^{ty} \sum_{x=0}^{y} \frac{e^{sx}}{x!(y-x)!} \\ &= \sum_{y=0}^{\infty} \frac{e^{ty}}{y!} \sum_{x=0}^{y} \frac{y!}{x!(y-x)!} e^{sx} \\ &= \sum_{y=0}^{\infty} \frac{e^{ty}}{y!} \sum_{x=0}^{y} \binom{y}{x} e^{sx} \\ &= \sum_{y=0}^{\infty} \frac{e^{ty}}{y!} (1+e^s)^y \\ &= \sum_{y=0}^{\infty} \frac{(e^t(1+e^s))^y}{y!} \\ &= \exp\left(e^t(1+e^s)\right) \end{aligned}$$

Chapter 8

Conditional Expectation

8.1 Conditional Expectation

Definition 68 (Conditional Expectation)**.** If X is a random variable, then the conditional expectation $g(x)$ given $Y = y$ is

- $E(g(x)|y) = \sum_x g(x)f(x|y)$ in the discrete case
- $E(g(x)|y) = \int_{-\infty}^{\infty} g(x)f(x|y)\,\mathrm{d}x$ in the continuous case

In particular, if $g(x) = x^i$, the conditional i'th moment is $E(X^i|Y = i)$. Additionally, $E(X|Y = y)$ is the conditional mean and $V(X|Y = y) = E(X^2|Y = y) - (E(X|Y = y))^2$ is the conditional variance.

Theorem 34. We have

1. $E(X^i) = E(E(X^i|Y = y))$
2. $V(X) = V(E(X|Y)) + E(V(X|Y))$

Proof 40. (Proof of 1). We have

$$\begin{aligned} E(X^i|Y = y) &= \int_{-\infty}^{\infty} x^i f(x|y)\,\mathrm{d}x \\ &= \int_{-\infty}^{\infty} x^i \frac{f(x,y)}{g_Y(y)}\,\mathrm{d}x \\ &= h(Y) \end{aligned}$$

So then we get

$$\begin{aligned}
E(h(Y)) &= \int_{-\infty}^{\infty} h(y) g_Y(y)\,\mathrm{d}y \\
&= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} x^i \frac{f(x,y)}{g_Y(y)}\,\mathrm{d}x \right) g_Y(y)\,\mathrm{d}y \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x^i f(x,y)\,\mathrm{d}x\,\mathrm{d}y \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x^i g_X(x) f(y|x)\,\mathrm{d}x\,\mathrm{d}y \\
&= \int_{-\infty}^{\infty} x^i g_X(x) \int_{-\infty}^{\infty} f(y|x)\,\mathrm{d}y\,\mathrm{d}x \\
&= \int_{-\infty}^{\infty} x^i g_X(x)\,\mathrm{d}x \\
&= E(X^i)
\end{aligned}$$

□

Proof 41. (Proof of 2). We have

$$\begin{aligned}
V(E(X|Y)) + E(V(X|Y)) &= E((E(X|Y))^2) - (E(E(X|Y)))^2 + E\left(E(X^2|Y) - (E(X|Y))^2\right) \\
&= E(E(X|Y))^2) - E(X)^2 + E(E(X^2|Y)) - E(E(X|Y))^2
\end{aligned}$$

By (1), we have that $E(E(X^2|Y)) = E(E(X^2)) = E(X^2)$. Therefore, we get

$$E(E(X|Y)^2) - (E(X))^2 + E(X^2) - E(E(X|Y))^2$$

The first and last cancel, and we get

$$E(X^2) - (E(X))^2 = V(X)$$

□

Example. Suppose $P(Y=1) = \frac{1}{8}$, $P(Y=2) = \frac{7}{8}$. Let $Z = X|Y$. Define $P(Z = 2Y) = \frac{3}{4}$, $P(Z = 3Y) = \frac{1}{4}$.

Note if $y = 1$, then

$$X|(Y=1) = \begin{cases} 2 & \text{with prob } \frac{3}{4} \\ 3 & \text{with prob } \frac{1}{4} \end{cases}$$

Then $E(X|Y=1) = 2 \cdot \frac{3}{4} + 3 \cdot \frac{1}{4} = \frac{9}{4}$. Similarly, $E(X|Y=2) = 4 \cdot \frac{3}{4} + 6 \cdot \frac{1}{4} = \frac{18}{4}$. Therefore,

$$E(X|Y) = \begin{cases} \frac{9}{4} & \text{if } y = 1, \text{ prob } \frac{1}{8} \\ \frac{18}{4} & \text{if } y = 2, \text{ prob } \frac{7}{8} \end{cases}$$

Then we see that

$$E(E(X|Y)) = \frac{9}{4} \cdot \frac{1}{8} + \frac{18}{4} \cdot \frac{7}{8}$$

Theorem 35 (Law of Total Expectation)**.** We have

$$E(X) = E(E(X|Y)) = \sum_y E(X|Y=y)P(Y=y)$$

Moreover, if $A_1, \dots A_n$ partition the sample space, then

$$E(X) = \sum_{i=1}^{n} E(X|A_i)P(A_i)$$

Note. Recall if $B_1, \dots B_n$ partition S, then $P(A) = P(A|B_1)P(B_1) + \dots P(A|B_n)P(B_n)$.

Example. Let the joint PDF of X, Y be

$$f_{X,Y}(x,y) = \begin{cases} x^2 e^{-x(y+1)} & x, y > 0 \\ 0 & \text{otherwise} \end{cases}$$

Find $E(X|Y)$.

$$\begin{aligned} E(X|Y) &= \int_{-\infty}^{\infty} x f(x|y)\,\mathrm{d}x \\ &= \int_0^{\infty} x \frac{x^2 e^{-x(y+1)}}{g_Y(y)}\,\mathrm{d}x \end{aligned}$$

We need $g_Y(y)$. Note

$$g_Y(y) = \int_{-\infty}^{\infty} x^2 e^{-x(y+1)}\,\mathrm{d}x$$

Integrating this by parts twice, we get

$$g_Y(y) = \frac{2}{(y+1)^3}$$

Then we get

$$E(X|Y) = \int_0^\infty \frac{1}{2}(y+1)^3 x^2 e^{-x(y+1)}\,\mathrm{d}x$$

Let $u = x(y+1)$. Then $\mathrm{d}u = y+1$.

$$\begin{aligned} E(X|Y) &= \frac{1}{2(y+1)} \int_0^\infty u^3 e^{-u}\,\mathrm{d}u \\ &= \frac{1}{2(y+1)}\Gamma(4) \\ &= \frac{3!}{2(y+1)} \\ &= \frac{6}{2(y+1)} \end{aligned}$$

Example. A man is stuck in a cave. There are 3 tunnel exists. One tunnel takes 2 hours to get out. Tunnel 2 takes 5 hours but returns to the starting location. Tunnel 3 takes 7 hours but returns to the starting location. If he picks a tunnel at random each time, what is the expected time until the man gets out of the cave?

Let X be the total hours to get out and Y be which tunnel he begins. We want

$$\begin{aligned} E(X) &= \sum_{y=1}^{3} E(X|Y=y)P(Y=y) \\ &= E(X|Y=1)P(Y=1) + E(X|Y=2)P(Y=2) + E(X|Y=3)P(Y=3) \\ &= 2\cdot\frac{1}{3} + (5+E(X))\cdot\frac{1}{3} + (7+E(X))\,\frac{1}{3} \end{aligned}$$

Now solving for $E(X)$, we get $E(X) = 14$.

Example. A biased coin has $P(\text{heads}) = p$. We toss the coin until we get 2 consecutive heads. What is the expected number of tosses.

Let A be the event of the sequence H, H, A_2 be the event of H, T, A_3 be the event of just T. Note that A_1, A_2, A_3 partition the sample space of toss sequences. If $X =$ the number of

tosses to get 2 consecutive heads,

$$\begin{aligned} E(X) &= E(X|A_1)P(A_1) + E(X|A_2)P(A_2) + E(X|A_3)P(A_3) \\ &= 2 \cdot p^2 + (2 + E(X)) \cdot p(1-p) + (1 + E(X)) \cdot (1-p) \end{aligned}$$

Solving this for $E(X)$ gives us

$$E(X) = \frac{p+1}{p^2}$$

Chapter 9

Bounds

9.1 Hölder and Minkowski's Inequality

Sometimes, we may not know the distribution. Estimations are made for various parameters, such as $E(X)$. Moreover what if we do not know the joint distribution, but do we know the distributions of X and Y?

Theorem 36 (Hölder's Inequality)**.** Let X and Y be random variables, and $p > 1$ such that $\frac{1}{p} + \frac{1}{q} = 1$. Then

$$\begin{aligned} |E(XY)| &\leq E(|XY|) \\ &\leq (E(|X|^p))^{1/p}(E(|Y|^q))^{1/q} \end{aligned}$$

Proof 42. (First Inequality). We have

$$\begin{aligned} E(XY) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf_{XY}(x,y)\,\mathrm{d}y\,\mathrm{d}x \\ &\leq \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |xy|f_{XY}(x,y)\,\mathrm{d}y\,\mathrm{d}x \\ &= E(|XY|) \end{aligned}$$

$$\begin{aligned} E(XY) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf_{XY}(x,y)\,\mathrm{d}y\,\mathrm{d}x \\ &\geq \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} -|xy|f_{XY}(x,y)\,\mathrm{d}y\,\mathrm{d}x \\ &= -E(|XY|) \end{aligned}$$

So $|E(XY)| \leq E(|XY|)$. □

Remark. If $p = 2$, then

$$|E(XY)| \leq \left(E(|X|^2)\right)^{1/2} \left(E(|Y|^2)\right)^{1/2}$$

which is the Cauchy-Schwarz inequality with respect to the inner product $\langle X, Y \rangle = E(XY)$.

Also note that if $X = |Z|^r$ for $1 < r < p$ and $Y = 1$, then

$$\begin{aligned} |E(XY)| &\leq E(|Z|^r) \\ &\leq (E(|Z|^{rp}))^{1/p} \end{aligned}$$

Letting $s = pr$, we find that

$$\begin{aligned} E(|Z|^r) &\leq (E(|Z|^s))^{r/s} \\ (E(|Z|^r))^{1/r} &\leq (E(|Z|^s))^{1/s} \end{aligned}$$

for $1 < r < s$, which is called Lyapunov's Inequality.

Example. Suppose $Z_1 = X - E(X)$ and $Z_2 = Y - E(Y)$. By the Cauchy-Schawrz inequality, we have

$$\begin{aligned} |E(z_1 z_2)| &= |E((X - E(X))(Y - E(Y)))| \\ &\leq E\left(|X - E(X)|^2\right)^{1/2} E\left(|Y - E(Y)|^2\right)^{1/2} \\ |\operatorname{Cov}(X, Y)| &\leq \sqrt{V(X)}\sqrt{V(Y)} \\ \left|\frac{|\operatorname{Cov}(X, Y)|}{\sqrt{V(X)}\sqrt{V(Y)}}\right| &\leq 1 \end{aligned}$$

which shows that the correlation coefficient satisfies $|\rho_{X,Y}| \leq 1$.

Example. Let $X \sim N(\mu_1, 1)$ and $Y \sim N(\mu_2, 1)$. We want to find $E(XY)$, but we don't have the joint PDF. Then we have

$$\begin{aligned} V(X) &= E(X^2) - (E(X))^2 \\ E(X^2) &= 1 + (E(X))^2 \end{aligned}$$

By the Cauchy-Schwarz inequality, we have

$$\begin{aligned}|E(XY)| &\leq \left(E(|X|^2)\right)^{1/2}\left(E(|Y|)^2\right)^{1/2}\\ &= \left(1+(E(X))^2\right)^{1/2}\left(1+(E(X))^2\right)^{1/2}\end{aligned}$$

which gives a bound on $E(XY)$ without the joint PDF.

Theorem 37 (Minkowski's Inequality)**.** Let $p \geq 1$. Then

$$E(|X+Y|^p) \leq (E(|X|^p))^{1/p} + (E(|Y|^p)^{1/p}$$

Proof 43. We have

$$\begin{aligned}E(|X+Y|^p) = E(|X+Y||X+Y|^{p-1}) &\\ \leq E\left((|X|+|Y|)(|X+Y|^{p-1})\right)&\\ = E\left(|X||X+Y|^{p-1}+|Y||X+Y|^{p-1}\right)&\\ = E\left(|X||X+Y|^{p-1}\right)+E\left(|Y||X+Y|^{p-1}\right)&\end{aligned}$$

Note that

$$\begin{aligned}E(|X||X+Y|^{p-1}) &\leq E(|X|^p)^{1/p}E\left(|X+Y|^{q(p-1)}\right)^{1/q}\\ &= E(|X|^p)^{1/p}E(|X+Y|^p)^{1/q}\\ E(|Y||X+Y|^{p-1}) &\leq E(|Y|^p)^{1/p}E\left(|X+Y|^{q(p-1)}\right)^{1/q}\\ &= E(|Y|^p)^{1/p}E(|X+Y|^p)^{1/q}\end{aligned}$$

So

$$\begin{aligned}E(|X+Y|^p) &\leq \left(E(|X|^p)^{1/p}+E(|Y|^p)^{1/p}\right)E(|X+Y|^p)^{1/q}\\ E(|X+Y|^p)^{1-1/q} &\leq E(|X|^p)^{1/p}+E(|Y|^p)^{1/p}\\ E(|X+Y|^p)^{1/p} &\leq E(|X|^p)^{1/p}+E(|Y|^p)^{1/p}\end{aligned}$$

□

Example. Let $X, Y \sim \text{Gamma}(\alpha, \beta)$. We have $V(X) = \alpha\beta^2$ and $E(X) = \alpha\beta$.

The left hand side of Minkowski's inequality for $p = 2$ is

$$E(|X+Y|^2)^{1/2} = E(|X|^2 + 2|XY| + |Y|^2)^{1/2}$$

which needs the Joint PDF. Instead, we can use the inequality to get

$$E(|X+Y|^2)^{1/2} \leq 2\left(\alpha\beta^2 + (\alpha\beta)^2\right)^{1/2}$$

since $E(X^2) = V(X) + E(X)^2 = \alpha\beta^2 + (\alpha\beta)^2$.

9.2 Markov and Chebyshev's Inequality

Theorem 38 (Markov)**.** Let X be a random variable that takes on non-negative values. Then for any $a > 0$,

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

Proof 44. We have

$$\begin{aligned} E(X) &= \sum_{x \geq a} xP(X = x) + \sum_{x < a} xP(X = x) \\ &\geq \sum_{x \geq a} xP(X = x) \\ &\geq \sum_{x \geq a} aP(X = x) \\ &= aP(X \geq a) \end{aligned}$$

So $P(X \geq a) \leq \frac{E(X)}{a}$. □

Example. An exam average is 75%. If we let $a = 80$, and X be the score on the exam, then

$$P(X \geq 80) \leq \frac{75}{80} \approx 0.9375$$

So the percentage of students scoring at least 80% can be at most 93.75%.

Example. Let $X \sim \text{Binomial}(10, \frac{1}{2})$, so $E(X) = 5$. If $a = 11$, then we get $P(X \geq 11) \leq \frac{5}{11}$, though the probability is clearly 0.

Theorem 39 (Chebyshev)**.** Let X be a random variable, and define $\mu = E(X)$, $\sigma^2 = V(x)$. Here, σ is the standard deviation. Then for any $k > 0$,

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

In the case $k = c\sigma$, where c is a positive constant, we can write

$$P(|X - \mu| \leq c\sigma) \geq 1 - \frac{1}{c^2}$$

In other words, the probability that X takes on values c standard deviations is at least $1 - \frac{1}{c^2}$.

Proof 45. We have

$$\begin{aligned}
\sigma^2 &= V(X) \\
&= E((X-\mu)^2) \\
&= \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\,\mathrm{d}x \\
&= \int_{-\infty}^{\mu-k} (x-\mu)^2 f(x)\,\mathrm{d}x + \int_{\mu-k}^{\mu+k} (x-\mu)^2 f(x)\,\mathrm{d}x + \int_{\mu+k}^{\infty} (x-\mu)^2 f(x)\,\mathrm{d}x \\
&\geq \int_{-\infty}^{\mu-k} (x-\mu)^2 f(x)\,\mathrm{d}x + \int_{\mu+k}^{\infty} (x-\mu)^2 f(x)\,\mathrm{d}x
\end{aligned}$$

If $x - \mu \leq -k$ or $x - \mu \geq k$, then $(x-\mu)^2 \geq k^2$. Therefore,

$$\begin{aligned}
\sigma^2 &\geq \int_{-\infty}^{\mu-k} (x-\mu)^2 f(x)\,\mathrm{d}x + \int_{\mu+k}^{\infty} (x-\mu)^2 f(x)\,\mathrm{d}x \\
&\geq \int_{-\infty}^{\mu-k} k^2 f(x)\,\mathrm{d}x + \int_{\mu+k}^{\infty} k^2 f(x)\,\mathrm{d}x
\end{aligned}$$

So then

$$\begin{aligned}
\frac{\sigma^2}{k^2} &\geq \int_{-\infty}^{\mu-k} f(x)\,\mathrm{d}x + \int_{\mu+k}^{\infty} f(x)\,\mathrm{d}x \\
&= P(|X - \mu| \geq k)
\end{aligned}$$

□

Example. Let the PDF of X be

$$f(x) = \begin{cases} 630x^4(1-x)^4 & 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

Integrating yields $\mu = E(X) = \frac{1}{2}$, and $\sigma = \sqrt{\frac{1}{44}} \approx 0.15$.

By Chebyshev, the probability that X will take on values within 2 standard deviations of the mean yields

$$\begin{aligned} P(0.2 < X < 0.8) &= P\left(\left|X - \frac{1}{2}\right| < 0.3\right) \\ &\geq 1 - \frac{1}{2^2} \\ &= \frac{3}{4} \end{aligned}$$

Example. An IQ test has mean 100, standard deviation 16. Show that the probability that a person has an IQ of at least 148 or at most 52 is at most $\frac{1}{9}$ using Chebyshev's theorem.

Here, $X \geq 148$, $X \leq 52$, so

$$\begin{aligned} |X - 100| &\geq k \\ P(|X - \mu| \geq 48) &\leq \frac{\sigma^2}{k^2} \\ &= \frac{16^2}{48^2} \\ &= \frac{1}{9} \end{aligned}$$

Chapter 10

The Weak and Strong Law of Large Numbers

10.1 Convergence and Law of Large Numbers

Definition 69 (Convergence in probability)**.** Let $X_1, X_2, \ldots$ be an infinite sequence of random variables defined on a sample space S. Then the sequence **converges in probability** to a random variable X if for all $\epsilon > 0$,

$$\lim_{n\to\infty} P(|X_n - X| < \epsilon) = 1$$

We write $X_n \xrightarrow{P} X$ if this occurs.
Denote $C_{n,\epsilon} = \{s \in S | |X_n(s) - X(s)| < \epsilon\}$. If the sequence converges in probability, then

$$\lim_{n\to\infty} P(C_{n,\epsilon}) = 1$$

for all $\epsilon > 0$.

Example. Let $S = [0, \infty)$ and $f(x)$ be a continuous PDF. Define

$$X_n(s) = \begin{cases} 1 & (n, \infty) \\ 0 & \text{otherwise} \end{cases}$$

We claim that X_n converges in probability to the zero random variable. Here, we have $C_{n,\epsilon} = \{s \in S | |X_n(s) - 0| < \epsilon\} = [0, n]$. So

$$\lim_{n\to\infty} P(C_{n,\epsilon}) = \lim_{n\to\infty} \int_0^n f(x)\,\mathrm{d}x$$

Definition 70 (Almost surely convergence). Let $X_1, X_2, \ldots$ be an infinite sequence of random variables. We say the sequence **converges almost surely** if for all $\epsilon > 0$,

$$P\left(\lim_{n\to\infty} |X_n - X| < \epsilon\right) = 1$$

We write $X_n \overset{a.s.}{\to} X$.

Remark. Almost surely convergence implies convergence in probability. The proof involves real analysis, and uses infs and sups.

Example. Define a random variable X_n that takes on values 1 and 2, where 2 occurs more often as n gets larger. Suppose the limiting random variable is $X = 2$. We need

$$\lim_{n\to\infty} P(|X_n - X| < \epsilon) = 1$$

Here, $P(|X_n - 2| < \epsilon)$ means we have nearly all values $X_n = 2$ as $n \to \infty$, so the chance that $|X_n - 2| < \epsilon$ approaches 100%. So,

$$\lim_{n\to\infty} P(|X_n - 2 < \epsilon) = 1$$

and so $X_n \overset{P}{\to} X$.

However, $\lim_{n\to\infty} |X_n - X| < \epsilon$ will NOT hold since there will always be some $X_n = 1$, meaning $|X_n - 2|$ will not be less than ϵ.

Therefore, $X_n \overset{P}{\to} X$ does NOT imply $X_n \overset{a.s.}{\to} X$.

Theorem 40 (Weak Law of Large Numbers). Let $X_1, X_2, \ldots$ be a sequence of random variables which are independent and identically distributed (i.i.d.). Assume $E(X_i) = \mu$ and $V(X_i) = \sigma^2$. Denote

$$\overline{X_n} = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

Then $\overline{X_n} \overset{P}{\to} \mu$.

Note. Observe that

$$E(\overline{X_n}) = \frac{1}{n}(E(X_1) + \ldots + E(X_n)) = \frac{1}{n}(n\mu) = \mu$$

However,

$$V(\overline{X_n}) = \frac{1}{n^2}(V(X_1) + \ldots V(X_n)) = \frac{1}{n^2}(n\sigma^2) = \frac{\sigma^2}{n}$$

which varies with n.

Proof 46. We show that $\lim_{n\to\infty} P(|\overline{X}_n - \mu| > \epsilon) = 0$. By Markov's inequality, we have

$$\begin{aligned}
\lim_{n\to\infty} P(|\overline{X}_n - \mu| > \epsilon) &= \lim_{n\to\infty} P\left(|\overline{X}_n - \mu|^2 > \epsilon^2\right) \\
&\leq \lim_{n\to\infty} \frac{E(|\overline{X}_n - \mu|^2)}{\epsilon^2} \\
&= \lim_{n\to\infty} \frac{V(\overline{X}_n)}{\epsilon^2} \\
&= \lim_{n\to\infty} \frac{\sigma^2}{n} \cdot \frac{1}{\epsilon^2} \\
&= 0
\end{aligned}$$

□

Note. In the big picture, the observed mean converges to the expected value as the number of trials increases.

Example. Flip a fair coin. If it is heads, place a blue chip in a box. Otherwise, place a red chip. Let

$$X_i = \begin{cases} 1 & \text{heads} \\ 0 & \text{tails} \end{cases}$$

count the number of blue chips on the i th trial. Clearly, $\mu = \frac{1}{2}$.

Example. Let

$$X = \begin{cases} 1 & \text{success} \\ 0 & \text{failure} \end{cases}$$

Let X_i be the value of X at trial i. The average number of succcesses is

$$\overline{X}_n = \frac{X_1 + X_2 + \ldots X_n}{n}$$

By the weak law of large numbers, $\overline{X}_n \xrightarrow{P} E(X)$. If we know $X \sim \text{Bernoulli}(p)$, then $E(X) = p$.

Example (Bernstein's Theorem). If $f(x)$ is continuous on $[a, b]$, then for any $\epsilon > 0$, there exists a polynomial $h(x)$ such that

$$|f(t) - h(t)| < \epsilon$$

for all $t \in [a, b]$. This proof involves the weak law of large numbers (see Wikipedia).

Theorem 41 (Strong Law of Large Numbers)**.** If $X_1, X_2, \ldots$ are independently and identically distributed, then

$$P\left(\lim_{n\to\infty} |\overline{X}_n - \mu| < \epsilon\right) = 1$$

That is,

$$\overline{X}_n \overset{a.s.}{\to} \mu$$

Remark. The Strong Law of Large Numbers implies the Weak Law of Large Numbers.

Chapter 11

Sampling and the Central Limit Theorem

11.1 Sampling

Definition 71 (Convergence in Distribution)**.** A sequence of random variables $X_1, X_2, \ldots$ **converges in distribution** to a random variable X (denoted $X_n \xrightarrow{d} X$) if the CDFs satisfy

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$$

for all x where $F_X(x)$ is continuous.

Example. Let

$$X_n(s) = \begin{cases} 1 & s = 1 \\ 0 & s = 0 \end{cases}$$

and let

$$X(s) = \begin{cases} 1 & s = 0 \\ 0 & s = 1 \end{cases}$$

Now $\lim_{n\to\infty} P(|X_n - X| < \epsilon) \neq 1$, so X_n does not converge in probability to X. However,

$$F_{X_n} = \begin{cases} \frac{1}{2} & 0 \leq x < 1 \\ 1 & x \geq 1 \end{cases}$$

Note that this is the same as F_X, so this sequence converges in distribution.

Definition 72 (Random Sample)**.** If $X_1, \ldots X_n$ are independently and identically distributed

random variables, we say they constitute a **random sample from an infinite population.**

Definition 73 (Statistic)**.** A **statistic** is a value calculated from the random sample (not the population). This process is called **statistical inference**.

Definition 74 (Sample Mean and Variance)**.** If $X_1, X_2, \dots X_n$ is a random sample, then the **sample mean** is

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

and the sample variance is

$$s^2 = \frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)}{n-1}$$

The factor of $\frac{1}{n-1}$ is known as **Bessel's correction**, and compensates for the fact that $X_i - \overline{X}$ is smaller than $X_i - \mu$.

Theorem 42. If $X_1, \dots X_n$ is a random sample from an infinite population with mean μ and variance σ^2, then

$$E(\overline{X}) = \mu$$

$$V(\overline{X}) = \frac{\sigma^2}{n}$$

Now given a statistic from a sample of size n, we can naturally define a function on the random variables $X_1, \dots X_n$. Then

$$Y = g(X_1, \dots X_n)$$

can be seen as a random variable. We now ask, what is the distribution of Y?

Example. Let $X_i \sim \text{Gamma}(\alpha, \beta)$. Consider the statistic

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

Then we showed before that

$$T \sim \text{Gamma}(n\alpha, \beta)$$

Moreover, to find the distribution of $\overline{X} = \frac{T}{n}$, we find

$$M_{\overline{X}}(t) = M_T\left(\frac{1}{n}t\right)$$

(Previously we had $M_{aX}(t) = M_X(at)$). Therefore,

$$M_T\left(\frac{t}{n}\right) = \left(\frac{1}{1-\frac{\beta}{n}t}\right)^{n\alpha} \implies \overline{X} \sim \text{Gamma}\left(n\alpha, \frac{\beta}{n}\right)$$

But what about $Y = \overline{X} + a$? We can find

$$M_Y(t) = e^{at}\left(\frac{1}{1-\frac{\beta}{n}t}\right)^{n\alpha}$$

This does not correspond to a known distribution. The Central Limit Theorem provides a way to approximate some of these distributions.

11.2 Central Limit Theorem

Theorem 43 (Central Limit Theorem). Let $X_1, X_2, \ldots X_n$ be a sequence of i.i.d. random variables with mean μ, variance σ^2 (finite), and MGF $M_X(t)$. Then if

$$\overline{X_n} = \frac{X_1 + \ldots + X_n}{n}$$

Then

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

Equivalently,

$$\lim_{n\to\infty} P\left(\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} < a\right) = \int_{-\infty}^{a} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\,\mathrm{d}x$$

Remark.

1. Idea of the CLT: Test samples many times. Compute the average of each sample. If the sample size is "large enough", then the distribution of these averages will look like a normal distribution.
2. The CLT can be proven without using $M_X(t)$, and there are generalizations where the random variables are not i.i.d.
3. CLT states the limit in terms of the average: however we can also approximate the sum by multiplying the top and bottom by n.
4. As in the previous point, note that

$$\frac{\overline{X} - \mu}{\sigma\sqrt{n}} = \frac{X_1 + X_2 + \ldots - n\mu}{\sigma\sqrt{n}}$$

Example. From earlier, $Y = \overline{X} + a$ had MGF

$$M_Y(t) = e^{at}\left(\frac{1}{1-\frac{\beta}{n}t}\right)^{n\alpha}$$

with the fact if $X \sim N(\mu, \sigma)$, then $Z = \frac{x-\mu}{\sigma}$ has $Z \sim N(0,1)$, then one finds $\overline{X} + a$ is approximately

$$N\left(\mu + a, \frac{\sigma^2}{n}\right)$$

So

$$Z = \frac{\overline{X} + a - (\mu + a)}{\sigma/\sqrt{n}}$$

Example. Consider the PMF

$$p(x) = \begin{cases} \frac{1}{6} & x = 1,2,3,4,5,6 \\ 0 & \text{else} \end{cases}$$

Consider picking a sample of size 2 (2 dice), with replacement and ordering.

Example. The amount of nicotine in a cigarette is a random variable with $\mu = 0.8$ mg and $\sigma = 0.1$ mg. If a person smokes 5 packs of cigarettes (100 cigarettes) a week, what is the probability that the total nicotine consumed in a week is at least 82 mg?

Let X_i be the amount of nicotine in cigarette i. If $X = \sum_{n=1}^{100} X_i$, we want $P(X \geq 82)$. Let $Z \sim N(0,1)$ and let Φ be the CDF of $N(0,1)$. We have

$$\begin{aligned} P(X \geq 82) &= P\left(\frac{X - 100 \cdot 0.8}{0.1 \cdot \sqrt{100}} \geq \frac{82 - 100 \cdot 0.8}{0.1 \cdot \sqrt{100}}\right) \\ &= P(Z \geq 2) \\ &= 1 - \Phi(2) \\ &\approx 0.02275 \end{aligned}$$

Example. In a sample of 25 people, their height is measured. We find $\overline{X} = 67.64$ inches. Suppose $\sigma^2 = 9$ for the population. Use the CLT to find the probability that μ exceeds 70 inches.

Here, if $\mu = 70$ then $Z = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} = \frac{67.64-70}{3/\sqrt{25}} \approx -3.266$. Note that we expect Z to become

more negative for larger values than 70. Therefore, we want

$$P(Z < -3.266) = \Phi(-3.266) \approx 0.00056$$

Observe that in many applications, we begin with a discrete random variable and approximate it as $N(0,1)$, a continuous random variable. We must use a **continuity correction** where

$$P(X = i) \approx P\left(i - \frac{1}{2} < X < i + \frac{1}{2}\right)$$

which is commonly used when X takes on integer values.

Example. 10 dice are rolled. Determine the probability that the sum is between 30 and 40, inclusively, using the CLT.

Let X_i be the value rolled on die i. We find that $\mu = \frac{7}{2}$, and $\sigma^2 = \frac{35}{12}$. If $X = \sum_{i=1}^{10} X_i$, using the continuity correction, we want

$$\begin{aligned} P(29.5 \le X \le 40.5) &= P\left(\frac{29.5 - 35}{\sqrt{\frac{35}{12}}\sqrt{10}} \le \frac{X - 35}{\sqrt{\frac{35}{12}}\sqrt{10}} \le \frac{40.5 - 35}{\sqrt{\frac{35}{12}}\sqrt{10}}\right) \\ &= P(-1.0184 \le Z \le 1.0184) \\ &= \Phi(1.0184) - \Phi(-1.0184) \\ &\approx 0.7 \end{aligned}$$

11.3 Proof of the Central Limit Theorem

Theorem 44. Let $F(x), F_1(x), F_2(x), \ldots$ be the CDFs of random variables $X_1, X_2, X_3, \ldots$ with moment generating functions $M(t), M_{X_1}(t), M_{X_2}(t), \ldots$. Then if

$$\lim_{n\to\infty} M_{X_n}(t) = M_X(t)$$

then $X_n \xrightarrow{d} X$.

Theorem 45. Let $P_n(x)$ be the n 'th degree Taylor polynomial centered at a:

$$P_n(x) = \sum_{k=0}^{n} \frac{f^{(k)}(a)}{k!}(x-a)^k$$

Then there is a $c \in (a, x)$ such that

$$f(x) = P_n(x) + \frac{f^{(n+1)}(c)}{(n+1)!}(x-c)^{n+1}$$

Proof 47. To prove the CLT, by our first theorem, if $Z_n = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$, it suffices to show that $M_{Z_n}(t) \xrightarrow{d} e^{-t^2/2}$, which is the MGF of the normal distribution. Let $Y_i = \frac{X_i-\mu}{\sigma}$. We have

$$\begin{aligned} Z_n &= \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \\ &= \frac{\frac{1}{n}\left(\sum_{i=1}^{n} X_i\right) - \mu}{\sigma/\sqrt{n}} \\ &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{X_i - \mu}{\sigma} \\ &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Y_i \end{aligned}$$

Observe that

$$E(Y_i) = E\left(\frac{X_i-\mu}{\sigma}\right) = \frac{1}{\sigma} E\left(X_i - \mu\right) = 0$$

$$V(Y_i) = V\left(\frac{Y_i-\mu}{\sigma}\right) = \frac{1}{\sigma^2} V(X_i) = 1$$

Recall that if $M_x(t)$ is an MGF, $M_X^{(r)}(0) = \mu_r' = E(X^r)$. So,

$$M_{Y_i}(0) = E(1) = 1$$

$$M_{Y_i}'(0) = E(Y_i) = 0$$

$$M_{Y_i}''(0) = V(Y_i) = 1$$

Using Taylor's Theorem for $n = 1$ on $M_{Y_i}(t)$, we get that for some $0 < c < t$, we have

$$\begin{aligned} M_{Y_i}(t) &= M_{Y_i}(0) + M_{Y_i}'(0) \cdot t + M_{Y_i}''(c) \cdot \frac{t^2}{2!} \\ &= 1 + M_{Y_i}''(c) \cdot \frac{t^2}{2} \\ &= 1 + \frac{t^2}{2} + \frac{1}{2}\left(M_{Y_i}''(c) - 1\right) t^2 \end{aligned}$$

We can then compute the MGF of Z_n as the product of the MGFs of $\frac{1}{\sqrt{n}}Y_i$, which is

$$\begin{aligned} M_{Z_n}(t) &= \left(M_{Y_i/\sqrt{n}}(t)\right)^n \\ &= \left(M_{Y_i}\left(\frac{t}{\sqrt{n}}\right)\right)^n \\ &= \left(1 + \frac{1}{2}\left(\frac{t}{\sqrt{n}}\right)^2 + \frac{1}{2}(M''_{Y_i}(c) - 1)\left(\frac{t}{\sqrt{n}}\right)^2\right)^n \end{aligned}$$

for some $0 < c < \frac{t}{\sqrt{n}}$. As $n \to \infty$, we get

$$\lim_{n\to\infty} M''_{Y_i}(c) - 1 = M''_{Y_i}(0) - 1 = 0$$

So,

$$\begin{aligned} \lim_{n\to\infty} M_{Z_n}(t) &= \lim_{n\to\infty}\left(1 + \frac{1}{2}\left(\frac{t}{\sqrt{n}}\right)^2 + \frac{1}{2}(M''_{Y_i}(c) - 1)\left(\frac{t}{\sqrt{n}}\right)^2\right)^n \\ &= \lim_{n\to\infty}\left(1 + \frac{t^2/2}{n}\right)^n \\ &= e^{t^2/2} \end{aligned}$$

Which is the MGF of $N(0, 1)$, as desired. □

Chapter 12

More Bounds

12.1 Contelli and Jensen's Inequality

Recall that Markov's inequality gave us $P(X \geq a) \leq \frac{E(X)}{a}$.

Theorem 46 (Contelli; One-sided Chebyshev)**.** If $a > 0$, then

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

Proof 48. Let $b > 0$. Then

$$\begin{aligned} P(X - E(X) \geq a) &= P(X \geq a + E(X)) \\ &= P(X + b \geq (a + b) + E(X)) \\ &= P((X + b)^2 \geq ((a + b) + E(X))^2) \\ &\leq \frac{E((X + b)^2)}{(a + b + E(X))^2} \\ &\leq \frac{V((X + b)^2) + (E(X + b))^2}{(a + b + E(X))^2} \\ &= \frac{\sigma^2 + (b + \mu)^2}{(a + b + \mu)^2} \end{aligned}$$

Minimizing this gives $b = \frac{\sigma^2}{a} - \mu$, which yield the desired bound. □

Example. An exam had average 75% and variance 8. Find an upper bound on the probability a student scored at least an 80%.

We have

$$\begin{aligned} P(X \geq 80) &= P(X - 75 \geq 5) \\ &\leq \frac{8^2}{8^2 + 5^2} \\ &\approx 0.72 \end{aligned}$$

Definition 75 (Convex, Concave)**.** A twice differentiable function $g(x)$ is **convex** if $g''(x) \geq 0$. Equivalently, for all $0 \leq t \leq 1$ and x_1, x_2 in the domain, we have

$$g(tx_1 + (1-t)x_2) \leq tg(x_1) + (1-t)g(x_2)$$

We say that $g(x)$ is **concave** if $-g(x)$ is convex.

Theorem 47 (Jensen)**.** If $g(x)$ is a convex function, then

$$E(g(X)) \geq g(E(X))$$

Proof 49. We use Taylor's theorem, centered at $\mu = E(X)$. Assume $g''(x) \geq 0$. For $c \in (\mu, x)$, we get

$$\begin{aligned} g(x) &= g(\mu) + g'(\mu)(x-\mu) + \frac{1}{2}g''(c)(x-\mu)^2 \\ &\geq g(\mu) + g'(\mu)(x-\mu) \\ E(g(X)) &\geq E(g(\mu) + g'(\mu)(x-\mu)) \\ &= g(\mu) + g'(\mu)E(x-\mu) \\ &= g(\mu) \\ &= g(E(X)) \end{aligned}$$

□

Example. Let $g(x) = \sqrt{x}$. Then

$$g''(x) = -\frac{1}{4}x^{-\frac{3}{2}} < 0$$

for all $x > 0$. This is concave, and hence $h(x) = -\sqrt{x}$ is convex. By Jensen's Inequality,

$$E(-\sqrt{X}) \geq -\sqrt{E(x)}$$
$$-E(\sqrt{X}) \geq -\sqrt{E(X)}$$
$$E(\sqrt{X}) \leq \sqrt{E(X)}$$

In general, if $g(X)$ is concave, then $E(g(X)) \leq g(E(X))$.

Example. Let $g(x) = x^2$. We find that $g(x)$ is convex, so by Jensen's Inequality,

$$E(X^2) \geq (E(X))^2$$
$$E(X^2) - (E(X))^2 \geq 0$$
$$V(X) \geq 0$$

Definition 76 (AM, GM, HM). Let $a_i > 0$ be a sequence. Then

- $\frac{1}{n}\sum_{i=1}^{n} a_i$ is the **arithmetic mean**.
- $(\prod_{i=1}^{n} a_i)^{\frac{1}{n}}$ is the **geometric mean**.
- $\frac{n}{\sum_{i=1}^{n} \frac{1}{a_i}}$ is the **harmonic mean**.

Example. A car drives 60 mph for 3 hours one way, and then 20 mph for 3 hours. The average speed is

$$\frac{\text{total distance}}{\text{total time}} = \frac{3(60) + 3(20)}{3 + 3} = 40$$

Suppose a car travels 60 miles each way. On the way there, it goes 60 mph. On the way back, it goes 20 mph. The average speed is

$$\frac{\text{total distance}}{\text{total time}} = \frac{2(60)}{\frac{60}{60} + \frac{60}{20}} = 30$$

Theorem 48 (AM-GM-HM).

$$\text{HM} \leq \text{GM} \leq \text{AM}$$

Proof 50. Define the PMF

$$p(x) = \begin{cases} \frac{1}{n} & x = a_1, a_2, \dots \\ 0 & \text{else} \end{cases}$$

Let $g(x) = \ln x$. We see that $g''(x) = -\frac{1}{x^2} < 0$ for $x > 0$.
We have

$$E(X) = \frac{1}{n}\sum_{i=1}^{n} a_i$$

is the arithmetic mean. We also have

$$\begin{aligned} E(g(X)) &= E(\ln X) \\ &= \frac{1}{n}\ln a_1 + \frac{1}{n}\ln a_2 + \cdots + \frac{1}{n}a_n \\ &= \frac{1}{n}\sum_{i=1}^{n} \ln a_i \\ &= \frac{1}{n}\ln\left(\prod_{i=1}^{n} a_i\right) \\ &= \ln\left(\left(\prod_{i=1}^{n} a_i\right)^{\frac{1}{n}}\right) \\ &= \ln \text{GM} \end{aligned}$$

By Jensen's inequality,

$$\begin{aligned} \ln \text{GM} &= E(g(X)) \\ &\leq g(E(X)) \\ &= \ln \text{AM} \end{aligned}$$

so GM $\leq$ AM. To show HM $\leq$ GM, define the PMF

$$p(y) = \begin{cases} \frac{1}{n} & \frac{1}{a_1}, \frac{1}{a_2}, \dots \frac{1}{a_n} \\ 0 & \text{else} \end{cases}$$

Let $g(y) = \ln y$. Then

$$\begin{aligned} E(Y) &= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{a_i} \\ &= \frac{1}{\text{HM}} \\ g(E(Y)) &= \ln\frac{1}{\text{HM}} \end{aligned}$$

Then $E(g(Y)) = E(\ln Y) = \frac{1}{n}\sum_{i=1}^{n}\ln\frac{1}{a_i} = \ln\left(\prod_{i=1}^{n}\frac{1}{a_i^{1/n}}\right) = \ln\frac{1}{\text{GM}}$. By Jensen's Inequality,

$$\begin{aligned} \ln\frac{1}{\text{GM}} &= E(g(Y)) \\ &\leq g(E(Y)) \\ &= \ln\frac{1}{\text{HM}} \end{aligned}$$

Therefore, HM $\leq$ GM. □

Example. Let X be a random variable. Assume $E(X) = 100$. Apply Jensen's Inequality to obtain a bound on $E\left(\frac{1}{1+x}\right)$. We check

$$\begin{aligned} g(x) &= \frac{1}{1+x} \\ g''(x) &= \frac{2}{(x+1)^3} \end{aligned}$$

which is positive for $x > -1$. Using Jensen's gives

$$\begin{aligned} E(g(X)) &= E\left(\frac{1}{1+X}\right) \\ &\geq g(E(X)) \\ &= \frac{1}{E(X)+1} \\ &= \frac{1}{101} \end{aligned}$$

Example. Consider the PMF

$$p(x) = \begin{cases} p_i & x = a_i, 1 \leq i \leq n \\ 0 & \text{else} \end{cases}$$

Applying Jensen's Inequality with $g(x) = \ln x$, we find

$$\begin{aligned} E(g(X)) &= E(\ln X) \\ &= \sum_{i=1}^{n} p_i \ln a_i \\ &\leq g(E(X)) \\ &= \ln E(X) \\ &= \ln \left(\sum_{i=1}^{n} p_i a_i \right) \end{aligned}$$

Thus,

$$\begin{aligned} \sum_{i=1}^{n} \ln a_i^{p_i} &\leq \ln \left(\sum_{i=1}^{n} p_i a_i \right) \\ \ln \left(a_1^{p_1} a_2^{p_2} \dots a_n^{p_n} \right) &\leq \ln \left(\sum_{i=1}^{n} p_i a_i \right) \\ a_1^{p_1} a_2^{p_2} \dots a_n^{p_n} &\leq p_1 a_1 + p_2 a_2 + \dots + p_n a_n \end{aligned}$$

Now let $n = 2$, and define $p_1 = \frac{1}{p}$ and $p_2 = \frac{1}{q}$, and let $a_1 = c^p$ and $a_2 = d^q$, where c and d are non negative. We get

$$\begin{aligned} a_1^{p_1} a_2^{p_2} &= (c^p)^{\frac{1}{p}} (d^q)^{\frac{1}{q}} \\ &= cd \\ &\leq p_1 a_1 + p_2 a_2 \\ &= \frac{c^p}{p} + \frac{d^q}{q} \end{aligned}$$

where $\frac{1}{p} + \frac{1}{q} = 1$, and $p, q > 1$. This is **Young's Inequality**.

Choose $c = \frac{|X|}{E(|X|^p)^{1/p}}$ and $d = \frac{|Y|}{E(|Y|^q)^{1/q}}$. Using Young's Inequality, we get

$$\frac{|X||Y|}{E(|X|^p)^{1/p}E(|Y|^q)^{1/q}} \leq \frac{|X|^p}{pE(|X|^p} + \frac{|Y|^q}{qE(|Y|^q)}$$
$$E\left(\frac{|X||Y|}{E(|X|^p)^{1/p}E(|Y|^q)^{1/q}}\right) \leq E\left(\frac{|X|^p}{pE(|X|^p} + \frac{|Y|^q}{qE(|Y|^q)}\right)$$

Now we can simplify the right hand side:

$$\begin{aligned} E\left(\frac{|X|^p}{pE(|X|^p} + \frac{|Y|^q}{qE(|Y|^q)}\right) &= \frac{E(|X|^p)}{pE(|X|^p)} + \frac{E(|Y|^q)}{qE(|Y|^q)} \\ &= \frac{1}{p} + \frac{1}{q} \\ &= 1 \end{aligned}$$

Therefore,

$$E\left(\frac{|X||Y|}{E(|X|^p)^{1/p}E(|Y|^q)^{1/q}}\right) \leq 1$$
$$E(|X||Y|) \leq E(|X|^p)^{1/p}E(|Y|^q)^{1/q}$$

which is just Holder's Inequality.