Probabilistic Methods

Lewis McConkey

April 1, 2024

Abstract

Mathematical methods of probability arose in the investigations first of Gerolamo Cardano in the 1560s however he has nothing to do with the cryptocurrency cardano! This book is intended for anyone interested in probability but has a focus on undergraduate probability and the methods, techniques and theories used to describe how likely events are to occur. We focus on discrete and continuous random variables for most of the book independently and go into depth with univariate and bivariate distributions. While we also touch on some extra topics in probability like covariance and correlation, limit theorems and characteristic functions i have tried to include some history in the book about when certain theorems came about and who invented them. I want people to understand the techniques used in this book so they can fulfil their passion or pass their classes but I also want people to see where probability topics started and i have also included some cool example/theorems in the last chapter too. I hope you enjoy!

Contents

1 Probability

Probability is the topic in mathematics that describes how likely an event is to occur. This book covers degree level probability starting from the basics all the way to transformations of random variables and some important limit theorems and finishes on some extra things i think are cool in probability. We start by covering the foundations of probability and denote some key terminology, which will use throughout the rest of the book. Some set theory, calculus and some simple linear algebra techniques are required prerequisites to this book.

1.1 The Foundations of Probability

We start with the following defintions:

- Ω is the sample space and is the set of all possible outcomes of an experiment (possibly infinite or uncountable)
- $\omega \in \Omega$ are the sample outcomes/realisations
- Subsets of Ω are called events

Example:
Take the sample space $\Omega = \{1, 2, 3, 4, 5, 6\}$, the set of all possible outcomes on a normal 6 sided die. The event that you throw an even number on this said die is A = $\{2, 4, 6\} \in \Omega$, other events include the event that you throw a number divisible by 3, the event you throw an odd number and the event you throw a number that is a factor of 12 (Try these yourself!).

We denote the probability of A by $\mathbb{P}$(A). The axioms of probability (Kolmogorov axioms) were first introduced by Russian mathematician Andrey Kolmogorov in 1933. A function $\mathbb{P}$ that assigns a real number $\mathbb{P}$(A) to each event A is a probability distribution if it satisfies the fundamental axioms of probability which are:

- **Axiom 1:** $\mathbb{P}(\text{A}) \geq 0 \; \forall A \subseteq \Omega$ (Probability of an event is a non-negative real number)
- **Axiom 2:** $\mathbb{P}(\Omega) = 1$ (Probability of the entire sample space will be 1)
- **Axiom 3:** if $A_1, A_2, \ldots\ldots$ are disjoint then:

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

(Any countable sequence of disjoint sets satisfies this equation)

Proofs from the axioms:

1. If $A \subset B$ then $\mathbb{P}(A) \leq \mathbb{P}(B)$ (Monotocity) *Proof.*
   - $(B \cap A^c)$ and A are disjoint and $B = (B \cap A^c) \cup A$
   - Now by axiom 3: $\mathbb{P}(B) = \mathbb{P}(B \cap A^c) + \mathbb{P}(A)$
   - By axiom 1: $\mathbb{P}(B \cap A^c) \geq 0$
   - Hence $\mathbb{P}(B) \geq \mathbb{P}(A)$ $\square$
2. $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$ (The law of complementary events) *Proof.*
   - $\Omega = A \cup A^c$ (These events are exhaustive*)
   - By axiom 2: $1 = \mathbb{P}(\Omega) = \mathbb{P}(A \cup A^c)$
   - By axiom 3: $\mathbb{P}(A \cup A^c) = \mathbb{P}(A) + \mathbb{P}(A^c)$ (These events are exclusive**)
   - So $1 = \mathbb{P}(A) + \mathbb{P}(A^c)$
   - Hence $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$ because $\mathbb{P}(A)$ is finite $\square$

Other examples of results that can be proved from the axioms include:

- $\mathbb{P}(\emptyset) = 0$
- $0 \leq \mathbb{P}(A) \leq 1$
- $A \bigcap B = \emptyset \implies \mathbb{P}(A \bigcup B) = \mathbb{P}(A) + \mathbb{P}(B)$ (Try these yourself!)

We will now define two terms used above that are actually very important properties of the set of outcomes in a sample space

- *Exhaustive: All possible outcomes in a sample space are listed
- **Exclusive: No two outcomes in the sample space can both occur

Addition Law

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

The proof is left to the reader, here is an example using this law:
Take the set of outcomes $\Omega = \{1, 2, 3, 4, 5, 6\}$, the outcomes from a throw of a normal six sided die. Now take event A to be the set of outcomes of throwing an even number and the set B to be the set of outcomes of throwing a number that is a factor of 18. So we have $A = \{2, 4, 6\}$ and $B = \{1, 2, 3, 6\}$ which are subsets of the sample space Ω. Quite simply we can deduce that $\mathbb{P}(A) = \frac{1}{2}$ and $\mathbb{P}(B) = \frac{2}{3}$ by using the fact that $\mathbb{P}(A) = \frac{|A|}{|\Omega|}$. Now the probability of A and B can be deduced by looking at the set of outcomes $A \cap B$, in this case $A \cap B = \{2, 6\}$, in other words this is the set of all outcomes that are common in both A and B and the probability is $\frac{1}{3}$. Using the addition law we get $\mathbb{P}(A \cup B) = \frac{1}{2} + \frac{2}{3} - \frac{1}{3} = \frac{5}{6}$. Alternatively this could also be done more easily by just looking at the set $A \cup B$.

Independent Events

Two events are independent if the occurrence of one does not affect the probability of occurrence of the other one. An example of this would be if you toss a coin the probability of tossing a head on throw one is $\frac{1}{2}$ and the probability of getting a head or tail on the next throw is still $\frac{1}{2}$. It does not matter what you throw on the first toss the probability of any throw on the second go will always remain $\frac{1}{2}$ as long as the coin is a normal unbiased coin.

If two events A and B are independent (often written as $A \perp B$) then:

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$$

In the coin example this becomes $\mathbb{P}(A \cap B) = (\frac{1}{2})(\frac{1}{2}) = \frac{1}{4}$. One might contemplate what we might find if we extend this to more than two events. Think of tossing a coin 3 times now, we still have probability $\frac{1}{2}$ no matter if we throw a head or a tail on each successive attempt. If we think of tossing three successive heads in a row we can quickly deduce that the probability of this is $\frac{1}{8}$ by multiplying $\frac{1}{2}$ three times. What we have found is $\mathbb{P}(A \cap B \cap C)$ where A is the event of a head on the first throw, B is the event of a head on the second throw and C is the event of a head on the third throw. We multiplied the probability of the events happening, this works because they are independent (Each event is not affected by the previous event). We can now extend this by thinking about what happens with 4,5, 6,...n events. Well just like before to get the probability of n **Independent** events all happening you would need to multiply the probabilities of the events individually (Which would amount to n probabilities being multiplied together). This theory gives rise to this result:

A set of events $\{A_i : i \in I\}$ is indepenent if

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

for every finite subset J of I

It can also be deduced that if A and B are disjoint events each with positive probability then they are not independent. This is because $\mathbb{P}(A)\mathbb{P}(B) > 0$ but $\mathbb{P}(A \cap B) = 0$ (Because they are disjoint).

1.2 Conditional Probability

Early discussions of conditional probability goes back to the analysis of Pascal and Fermat (1654) of the problem of points. Not so long after (1665) Christiaan Huygens and John Hudde also wrote about the difference between conditional

and unconditional probabilities. It wasn't until 1931 however that the notation for conditional probability that we still use today was introduced by Harold Jeffreys in his publication called "scientific inference".

Conditional probability is a measure of the probability of an event occurring given that another event is already known to have occurred. If A and B are two events such that $\mathbb{P}(B) > 0$ then the conditional probability of A given B is:

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

It can be show quite quickly that the first two axioms of probability hold for this probability however to show the third axiom holds it requires slightly more reasoning. Show that the first two axioms hold and for the third consider two disjoint events A_1 and A_2 and use the law of total probability mentioned further on in this section to prove that all three axioms are satisfied.

Let's take S to be the probability that a person is sick and say it is 0.05 and the probability that a person is coughing is 0.3 and denoted as C. Then if these two events are assumed to be independent then what is the probability that a person is sick given that they are coughing?

$$\mathbb{P}(S|C) = \frac{\mathbb{P}(S \cap C)}{\mathbb{P}(C)} = \frac{0.3 \times 0.05}{0.3} = 0.05 = \mathbb{P}(S)$$

Now we noticed something cool happening here and that because of our assumption that the two events are independent then the probability that someone is sick given that they are coughing is equivalent to the probability they are sick. This might be a weird concept at first but is really simple when thought about using the following result:

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

We have seen that if the two events are independent then the first equation holds true so now the top of the fraction in the conditional probability equation becomes $\mathbb{P}(A)\mathbb{P}(B)$ and therefore the B probabilities cancel out to leave the $\mathbb{P}(A)$. This makes even more sense when thought of that the independence means that one event doesn't affect the occurrence of another and so the probability of that event occurring given the other has happened is the same as if the event never happened in the first place and so it just equals the probability of the second event happening.

The Law of Total Probability

$$\mathbb{P}(A) = \mathbb{P}(A|B)\mathbb{P}(B) + \mathbb{P}(A|B^c)\mathbb{P}(B^c)$$

Proof

$$\begin{aligned} \mathbb{P}(A) &= \mathbb{P}(A \cap B) + \mathbb{P}(A \cap B^c) \text{ (This is the partition law)} \\ &= \mathbb{P}(A|B)\mathbb{P}(B) + \mathbb{P}(A|B^c)\mathbb{P}(B^c) \text{ (By conditional probability)} \end{aligned}$$

By thinking of B and B^c as a partition of the sample space Ω we can expand the law of total probability to think what if we partition the sample space in to an increasing amount of spaces similar to what we did when we thought about having n independent events. These thoughts lead to the following result, let $B_1, B_2, ..., B_k$ be a partition of Ω. Then for any event A,

$$\mathbb{P}(A) = \sum_{i=1}^{k} \mathbb{P}(A|B_i)\mathbb{P}(A_i)$$

Example:
We are told that if it rains then a football match tomorrow will likely not be played but if it is not raining then it will be played. The probabilities that the game is played despite it raining is 0.05 and the probability it rains is 0.4. Determine the probability that the game is played.

First of all let us denote the event that it rains as R and the probability the game is played as P. Now from the question we get $\mathbb{P}(R) = 0.4$ and so $\mathbb{P}(R^c) = 0.6$. We also get $\mathbb{P}(P|R) = 0.5$ and $\mathbb{P}(P|R^c) = 1$. Now we can use the law of total probability:

$$\begin{aligned}\mathbb{P}(P) &= \mathbb{P}(P|R)\mathbb{P}(R) + \mathbb{P}(P|R^c)\mathbb{P}(R^c) \\ &= (0.5)(0.4) + (1)(0.6) \\ &= 0.8\end{aligned}$$

1.3 Bayes' Theorem

Bayes' theorem was discovered by the English mathematician Thomas Bayes and was eventually published by Richard Price in 1763 after Thomas Bayes's death. The published work was called "An Essay towards solving a Problem in the Doctrine of Chances" and contained the theorem. The theorem describes the probability of an event based on prior knowledge of conditions that might be related to the event and one of the biggest applications of the theorem is a particular approach to statistical inference called Bayesian inference. Here is the theorem:
If A and B are events in the sample space with $\mathbb{P}(A)$, $\mathbb{P}(B) > 0$ then:

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

Using the law of total probability we can also express Bayes' theorem as:

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(A|B)\mathbb{P}(B)}{\mathbb{P}(A|B)\mathbb{P}(B)+\mathbb{P}(A|B^c)\mathbb{P}(B^c)}$$

Like in previous discussions we may think what happens if we have more than two possibilities (What happens if we partition the sample space into $B_1, B_2, B_3, ..., B_k$ instead of just B and B^c). Again these thoughts leads us to this result:

Let $B_1, ..., B_k$ be a partition of the sample space Ω such that $\mathbb{P}(A_i) > 0$ for each i. if $\mathbb{P}(B) > 0$ then, for each $i, ..., k$,

$$\mathbb{P}(B_i|A) = \frac{\mathbb{P}(A|B_i)\mathbb{P}(B_i)}{\sum_j \mathbb{P}(A|B_j)\mathbb{P}(B_j)}$$

Where $\mathbb{P}(B_i)$ is called the prior probability of B and $\mathbb{P}(B_i|A)$ is called the posterior probability of B.

Proof

$$\mathbb{P}(B_i|A) = \frac{\mathbb{P}(B_i \cap A)}{\mathbb{P}(A)} = \frac{\mathbb{P}(A|B_i)\mathbb{P}(B_i)}{\mathbb{P}(A)} = \frac{\mathbb{P}(A|B_i)\mathbb{P}(B_i)}{\sum_j \mathbb{P}(A|B_j)\mathbb{P}(B_j)}$$

We used the conditional probability definition twice and the law of total probability

Example
Say that we know 30% of people at a party are alcoholics and 80% of the people at the party are drinking (There are 40 people at the party). We also know that of the people drinking 8 of them are alcoholics. What is the probability that a person at the party is drinking given that they are an alcoholic?

First we denote the event that a person is an alcoholic as A and denote the event that a person is drinking as B. We have $\mathbb{P}(A) = 0.3$ and $\mathbb{P}(B) = 0.8$. $\frac{8}{28} = \frac{2}{7}$ is $\mathbb{P}(A|B)$ (The probability that a person is an alcoholic given that they are drinking). Now we can use Bayes' theorem:

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(A|B)\mathbb{P}(B)}{\mathbb{P}(A)} = \frac{(\frac{2}{7})(0.8)}{0.3} = 0.762 \text{ (to 3 d.p.)}$$

This is the probability that a person at the party is drinking given that they are an alcohlic.

This is the end of the 1st chapter. All things covered so far are standard probability theory. Some aspects of this chapter are high school/college level probability and some are taught/retaught in a 1st year university probability class. The chapter has covered the basics of probability and is set as an introduction to probability and to give a basis to work from for the rest of the book. The only required prerequisite to this chapter was some set theory. Hope you enjoyed it and get ready for the rest of the book because the content only gets harder from here!!

2 Random Variables

This chapter is on random variables. We will first look at discrete random variables, probability mass functions and cumulative distribution functions. We will then go about looking at specific types of discrete random variables like uniform and binomial. After that we will look at the continuous case and look at specific types of continuous random variables like Poisson and normal. We will finish the chapter by looking at more properties of random variables like expectation and variance and take into account the discrete and continuous cases individually.

2.1 Discrete Random Variables

The concept of random variables was first introduced by Pafnuty Chebyshev in the mid-nineteenth century however the modern understanding of random variables didn't arrive until work by Andrey Kolmogorov was introduced in 1933. Here is the definition we use today:

A random variable is a function

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

that assigns a real number $X(\Omega)$ to each outcome ω

Example

Say that you throw a normal 6 sided dice 10 times and let X(ω) be the number of 6's thrown in the sequence ω. If $\omega = 1, 4, 5, 4, 3, 1, 6, 6, 2, 1$, then X($\omega$) = 2.

Every time this experiment is conducted, one value (realisation) of the random variable is observed. You would have to throw the dice another 10 times and count up the number of times a 6 occurs to get another realisation of the random variable.

The induced sample space is the range of values taken by the random variable X defined on Ω, that is $\{X(\omega) : \omega \in \Omega\}$. We will also denote this as H for future use

Now that we have defined a random variable and provided all other necessary explanation for the basic set ups of random variables we can move on to the title of this section of the chapter which is discrete random variables. Discrete random variables are random variables where the function is finite or countable. We will provide examples next and cover the discrete case in more detail, we will then move on to continuous random variables to distinguish the differences between them. For now we can just say that continuous random variables are random variables where the function is uncountable and leave the rest until later on in the chapter. First let us focus on discrete random variables.

Discrete random variables can arise from many different experiments, examples include:

- The amount of cars that pass your window in 10 minutes
- The outcomes of throwing a 6 sided dice

Example
A normal 6 sided dice is thrown. The sample space is $\Omega = \{1, 2, 3, 4, 5, 6\}$. Define a rv (random variable) for the outcomes that can occur with this dice throw.
Simply if we denote the rv as $X(\omega)$:

$$X(1) = X(2) = X(3) = X(4) = X(5) = X(6) = \tfrac{1}{6}$$

This is a very simple example of a random variable whose probabilities all equal each other. The induced sample space is $X(\omega) = \{\frac{1}{6}\}$. We can also calculate the expectation and the variance of this rv relatively easily in this case because it is a simple rv but we will leave that until later in the chapter when we discover how to calculate these.

Example
Take a normal deck of cards (With 52 cards). Denote hearts as 1, spades as 2, diamonds as 3 and clubs as 4. Define the rv to be the number of cards belonging to each suit.

1. Find the induced sample space for X (The range of values taken by the rv).
2. Now say we are playing a game and the rv is now the number of cards belonging to each suit in everyone's hands. Find the induced sample space in this new case when person 1 has 2 clubs, 5 diamonds and a heart, person 2 has 4 spades, 2 hearts and a club and person 3 has 3 clubs, 4 hearts and 2 diamonds.

(Try these yourself!). This is what you should get:

1. $X(\omega) = \{\frac{1}{13}\}$
2. $X(\omega) = \{\frac{4}{23}, \frac{6}{23}, \frac{7}{23}\}$

Now we will move on to what are called probability mass functions (pdf). A pdf is a function that gives the probability that a discrete random variable is exactly equal to some value. We will now look at this in more detail.

Probability Mass Functions

The probability mass function (pmf) of a discrete random variable, X is defined by:

$$p_X(x) = \mathbb{P}(X = x)$$

for all $x \in H$. It is a function $p : \mathbb{R} \to [0, 1]$

Properties of pmf's:

1. $P_X(x) \geq 0 \; \forall x$
2. $\sum_{x \in H} p_X(x) = 1$

These can also be proved by the axioms of probability. For the first property use axiom 1 and state why you can use this. For the second property use axiom 2 and then axiom 3. Try this yourself!

Example
Find the pmf of the number of cards belonging to each suit in a normal deck of cards (52 cards). Denote hearts as 1, spades as 2, diamonds as 3 and clubs as 4.
Quite simply due to there being 13 cards in each suit:

$$P_X(1) = P(X = 1) = \tfrac{1}{13}$$

$$P_X(2) = P(X = 2) = \tfrac{1}{13}$$

$$P_X(3) = P(X = 3) = \tfrac{1}{13}$$

$$P_X(4) = P(X = 4) = \tfrac{1}{13}$$

$$P_X(x) = 0 \text{ for } x \notin \{1, 2, 3, 4\}$$

Example
If a pmf is specified by $p_X(x) = t$ for $x = 0, 1, ..., m$ and $p_X(x) = 0$ otherwise, where t is constant, determine the value of t and hence the value of m. What is the value of t when m is 10?

$$1 = \sum_{x=0}^{m} p_X(x) = (m + 1)t$$

$$\text{Hence } t = \tfrac{1}{m+1} = \tfrac{1}{10+1} = \tfrac{1}{11} \text{ (When m = 10)}$$

$$\text{Also } m = \tfrac{1}{t} - 1$$

Example
If a pmf is specified as by $p_X(x) = mx^2$ for $x = 2, 4, 6, 8, 10$ and $p_X(x) = 0$ otherwise, where m is a constant, determine the value of m.

$$\begin{aligned} 1 &= p_X(2) + p_X(4) + p_X(6) + p_X(8) + p_X(10) \\ &= 4m + 16m + 36m + 64m + 100m \\ &= 220m \end{aligned}$$

$$\text{Hence } m = \tfrac{1}{220}$$

Now what happens if you change the pmf to $p_X(x) = m^2x$ but keep everything else the same. Calculate m.

$$\begin{aligned} 1 &= p_X(2) + p_X(4) + p_X(6) + p_X(8) + p_X(10) \\ &= 2m^2 + 4m^2 + 6m^2 + 8m^2 + 10m^2 \\ &= 30m^2 \end{aligned}$$

$$\text{Hence } m^2 = \tfrac{1}{30}$$

$$\text{and } m = \tfrac{1}{\sqrt{30}} \text{ or } m = \tfrac{-1}{\sqrt{30}}$$

You might think at first that the second solution isn't valid because it is negative and probabilities have to be non-negative but when you put the value back into the pmf you would square the m value and make it positive. From these examples we have gained a better understanding of the pmf and how to use it in questions but also have been shown how important the properties of the pmf can be in answering some problems.

This helps us move on to another result we can use related to pmf's. That is if we want to find the probability of an event occurring for an rv we can do this using pmf's.

Let $E \subseteq H$ be an event in the induced sample space. The probability of E is given by:

$$\mathbb{P}(X \in E) = \sum_{x \in E} p_X(x)$$

Proof. Write $E = x_1, ..., x_k \subseteq H$. Then:

$$\begin{aligned} \mathbb{P}(X \in E) &= \mathbb{P}(\{X = x_1\} \cup \{X = x_2\} \cup \cdots \cup \{X = x_k\}) \\ &= \mathbb{P}(X = x_1) + \mathbb{P}(X = x_2) + \cdots + \mathbb{P}(X = x_k) \\ &= p_X(x_1) + p_X(x_2) + \cdots p_X(x_k) \\ &= \sum_{x \in E} p_X(x) \end{aligned}$$

Example
Using this result for both of the pmf's in the previous example find the following probabilities:

1. $\mathbb{P}(X \leq 6)$
2. $\mathbb{P}(4 \leq X \leq 8)$
3. $\mathbb{P}(X = 10)$

First for $p_X(x) = mx^2 = \frac{1}{220}x^2$:

1. $\mathbb{P}(X \leq 6) = \mathbb{P}(X = 2) + \mathbb{P}(X = 4) + \mathbb{P}(X = 6) = \frac{14}{55}$
2. $\mathbb{P}(4 \leq X \leq 8) = \mathbb{P}(X = 4) + \mathbb{P}(X = 6) + \mathbb{P}(X = 8) = \frac{29}{55}$
3. $\mathbb{P}(X = 10) = \frac{5}{11}$

Next for $p_X(x) = m^2x = \frac{1}{30}x$:

1. $\mathbb{P}(X \leq 6) = \mathbb{P}(X = 2) + \mathbb{P}(X = 4) + \mathbb{P}(X = 6) = \frac{2}{5}$
2. $\mathbb{P}(4 \leq X \leq 8) = \mathbb{P}(X = 4) + \mathbb{P}(X = 6) + \mathbb{P}(X = 8) = \frac{3}{5}$
3. $\mathbb{P}(X = 10) = \frac{1}{3}$

Check if you can do this and get to the same answers!

Now that we have discussed probability density functions in detail and gave examples of how to use them, their properties and how to find probabilities using them we can now move on to cumulative distribution functions. As the name suggests these are functions that give the cumulative probabilities which make it useful when dealing with probabilities with ranges like in the last example.

Cumulative Distribution Function

The cumulative distribution function (cdf) of a random variable X is a function $F_X : \mathbb{R} \to \mathbb{R}$ given by:

$$F_X(x) = \mathbb{P}(X \leq x)$$

Specifically for a discrete random variable X the cdf is:

$$F_X(x) = \mathbb{P}(X \leq x) = \sum_{x_i \leq x} \mathbb{P}(X = x_i) = \sum_{x_i \leq x} p_X(x_i)$$

Properties of cdf's:

1. $0 \leq F_X(x) \leq 1$
2. $\lim_{x\to-\infty} F_X(x) = F_X(-\infty) = 0$ and $\lim_{x\to\infty} F_X(x) = F_X(\infty) = 1$
3. $F_X(x)$ is a non-decreasing function of x

We will now look at an example of a cdf in action.

Example
Take the following cdf:

$$F_X(x) = \begin{cases} 0 & \text{for } x < 0 \\ \frac{1}{6} & \text{for } 0 \leq x < 1 \\ \frac{1}{3} & \text{for } 1 \leq x < 2 \\ \frac{1}{2} & \text{for } 2 \leq x < 3 \\ \frac{2}{3} & \text{for } 3 \leq x < 4 \\ \frac{5}{6} & \text{for } 4 \leq x < 5 \\ 1 & \text{for } x \geq 5 \end{cases}$$

Find:

1. $\mathbb{P}(X \leq 3)$
2. $\mathbb{P}(X > 2)$
3. $\mathbb{P}(1 < X \leq 4)$

We use the definition of the cdf to answer these questions:

1. $\mathbb{P}(X \leq 3) = F_X(3) = \frac{1}{2}$
2. $\mathbb{P}(X > 2) = 1 - \mathbb{P}(X \leq 2) = 1 - F_X(2) = \frac{1}{2}$
3. $\mathbb{P}(1 < X \leq 4) = \mathbb{P}(X \leq 4) - \mathbb{P}(X \leq 1) = F_X(4) - F_X(1) = \frac{5}{6} - \frac{1}{6} = \frac{2}{3}$

Notice how we get $\mathbb{P}(X \leq 3) = \mathbb{P}(X > 2) = \frac{1}{2}$. This can also be shown by the symmetry of the cdf given. Notice also in 2. we had $\mathbb{P}(X > 2) = 1 - \mathbb{P}(X \leq 2) = 1 - F_X(2)$ and in 3. we had $\mathbb{P}(1 < X \leq 4) = \mathbb{P}(X \leq 4) - \mathbb{P}(X \leq 1) = F_X(4) - F_X(1)$. This takes us on to the next two results.

The Survivor Function

The survivor function is defined by:

$$S_X(x) = \mathbb{P}(X > x) = 1 - \mathbb{P}(X \leq x) = 1 - F_X(x)$$

This shows the relationship stated for part 2 of the last example and how to relate probabilities that are greater than a value to probabilities that are less than or equal to a value. It also shows how to find the probability of greater than a value in terms of the cdf.

Probabilities of Intervals

$$\mathbb{P}(a < X \leq b) = \mathbb{P}(X \leq b) - \mathbb{P}(X \leq a) = F_X(b) - F_X(a)$$

This works by the law of total probability. This also shows the relationship stated for part 3 of the last example and how to find the probabilities of intervals in terms of the cdf.

Another result that can be quite useful is:

For any $x \in \mathbb{R}$ we have:

$$\mathbb{P}(X = x) = F_X(x) - \lim_{i \to \infty} F(x - \tfrac{1}{i})$$

One question we might want to move onto now is how to calculate the cdf given the pdf and vice-versa. Here we will show how to calculate the cdf from the pmf and we will leave the other case for an example later on in the chapter.

Example

Take the pdf of $p_X(x) = \frac{1}{220}x^2$ from an earlier example. Now find the cdf of this:

$$\begin{aligned}
p_X(2) &= \frac{1}{55} \\
p_X(2) + p_X(4) &= \frac{1}{11} \\
p_X(2) + p_X(4) + p_X(6) &= \frac{14}{55} \\
p_X(2) + p_X(4) + p_X(6) + p_X(8) &= \frac{6}{11} \\
p_X(2) + p_X(4) + p_X(6) + p_X(8) + p_X(10) &= 1
\end{aligned}$$

Hence the cdf is:

$$F_X(x) = \begin{cases} 0 & \text{for } x < 0 \\ \frac{1}{55} & \text{for } 0 \leq x < 2 \\ \frac{1}{11} & \text{for } 2 \leq x < 4 \\ \frac{14}{55} & \text{for } 4 \leq x < 6 \\ \frac{6}{11} & \text{for } 6 \leq x < 8 \\ 1 & \text{for } x \geq 5 \end{cases}$$

Now we have seen how to find the cdf from the pmf try to verify that this is a valid cdf (use the properties of cdf's). Also try to find the cdf from the pdf of $p_X(x) = \frac{1}{30}x$ from an earlier example.

Now we have gone through the basics of discrete random variables and know how to set them up, what the pmf and cdf are and how to convert between them we are ready to move on to continuous random variables. These are random variables that can take any value or an interval of values along the real line instead of just distinct points. From this point onwards calculus is a required prerequisite and it should be clear to those who have taken calculus before why it is needed for continuous random variables and not discrete random variables particularly why we need to use differentiation and integration.

2.2 Continuous Random Variables

As said previously continuous random variables take on values or intervals of values across the real line and have an infinite number of possible values (uncountable). Examples include height, weight and the time taken to run a mile.

In the discrete case we used this result:

$$\mathbb{P}(X = x) = F_X(x) - \lim_{i \to \infty} F(x - \tfrac{1}{i})$$

But since $F_X(x)$ is now continuous everywhere the result is now equal to 0 for all x. So the probability of being equal to a particular value $x \in X$ in the continuous case is always 0 and we also have:

$$\mathbb{P}(X \leq x) = \mathbb{P}(X < x)$$

When we considered the discrete case we saw the pmf however in the continuous case it is called the probability density function (pdf)

Probability Density Function

The Probability Density Function (pdf) of a continuous random variable X is:

$$f_X(x) = \frac{d}{dx} F_X(x)$$

Properties of pdf's:

1. $f_X(x) \geq 0 \ \forall x$ (Positivity)
2. $\int_{-\infty}^{\infty} f_X(x)dx = 1$ (Unit-integrability)

These are very similar to the properties of pmf's but include calculus to take care of the continuous case.

Cumulative distribution function

The Cumulative Distribution Function (cdf) of a continuous random variable X is:

$$F_X(x) = \int_{-\infty}^{x} f_X(s)ds$$

This was quite clear from the definition of the pmf from above and the variable s is called the dummy variable.

Properties of $F_X(x)$:

1. $0 \leq F_X(x) \leq 1$ with $\lim_{x\to-\infty} F_X(x) = 0$ and $\lim_{x\to\infty} F_X(x) = 1$
2. $F_X(x)$ is non-decreasing function of x

These properties are the same in both the discrete and continuous case.

Probabilities of Intervals

$$\begin{aligned}\mathbb{P}(a < X \leq b) &= F_X(b) - F_X(a) \\ &= \int_{-\infty}^{b} f_X(s)ds - \int_{-\infty}^{a} f_X(s)ds \\ &= \int_{a}^{b} f_X(s)ds\end{aligned}$$

Example

Take the following pdf:

$$f_X(x) = \begin{cases} \frac{1}{2} & \text{for } 0 < x \leq 2 \\ 0 & \text{otherwise} \end{cases}$$

Find:

1. $\mathbb{P}(X \leq 1.4)$
2. $\mathbb{P}(0.6 < X \leq 1.3)$
3. the cdf

We use the definition of the cdf to answer these questions:

1. $\mathbb{P}(X \leq 1.4)$

$$\begin{aligned} F_X(1.4) &= \int_{-\infty}^{1.4} f_X(s)\, ds \\ &= \int_0^{1.4} \frac{1}{2}\, ds \\ &= \left[\frac{1}{2}s\right]_0^{1.4} \\ &= 0.7 \\ &= \mathbb{P}(X \leq 1.4) \end{aligned}$$

2. $\mathbb{P}(0.6 < X \leq 1.3)$

$$\begin{aligned} F_X(1.3) - F_X(0.6) &= \int_{-\infty}^{1.3} f_X(s)\, ds - \int_{-\infty}^{0.6} f_X(s)\, ds \\ &= \int_0^{1.3} \frac{1}{2}\, ds - \int_0^{0.6} \frac{1}{2}\, ds \\ &= \left[\frac{1}{2}s\right]_0^{1.3} - \left[\frac{1}{2}s\right]_0^{0.6} \\ &= 0.35 \\ &= \mathbb{P}(0.6 < X \leq 1.3) \end{aligned}$$

3. the cdf
   For $0 < x \leq 2$:

$$\begin{aligned} F_X(x) &= \int_{-\infty}^{x} f_X(s) ds \\ &= \left[\frac{1}{2}s\right]_0^{x} \\ &= \frac{1}{2}x \end{aligned}$$

   For $x \leq 0$ and $x > 2$ we need to think about the properties of the cdf. For the properties to hold we need 0 for ≤ 0 and 1 for > 2 (See if you can spot this and understand why from the properties).

   Hence the cdf is:

$$F_X(x) = \begin{cases} 0 & x \leq 0 \\ \frac{1}{2}x & \text{for } 0 < x \leq 2 \\ 1 & x > 2 \end{cases}$$

Notice from the questions we understand that $\mathbb{P}(X \leq x) = F_X(x)$. This was stated in the discrete section but is the same in the continuous case except we use the definition for the cdf in the continuous case instead. Also notice that we could have answered the questions in the reverse order by finding the cdf first and then simply inputting the values of the probabilities we want into it. This is a simpler way to work them out if we have many probabilities we want to find then it is practical to find the cdf first.

Example

Take the following cdf:

$$F_X(x) = \begin{cases} 0 & x \leq 0 \\ \frac{x^2}{4} & \text{for } 0 < x \leq 2 \\ 1 & x > 2 \end{cases}$$

Find:

1. $\mathbb{P}(X \leq 0.8)$
2. $\mathbb{P}(1.2 < X \leq 1.8)$
3. the pdf

We use the definition of the cdf and pdf to answer these questions:

1. $\mathbb{P}(X \leq 0.8)$

$$F_X(0.8) = \frac{(0.8)^2}{4} = 0.16 = \mathbb{P}(X \leq 0.8)$$

2. $\mathbb{P}(1.2 < X \leq 1.8)$

$$F_X(1.8) - F_X(1.2) = \frac{(1.8)^2}{4} - \frac{(1.2)^2}{4} = 0.45 = \mathbb{P}(1.2 < X \leq 1.8)$$

3. the pdf
   For $0 < x \leq 2$ we use:

$$f_X(x) = \frac{d}{dx} F_X(x)$$

   Hence the pdf is:

$$f_X(x) \begin{cases} \frac{x}{2} & \text{for } 0 < x \leq 2 \\ 0 & \text{otherwise} \end{cases}$$

We can see now how it is easier to work out the probabilities using the cdf than the pdf. This ends the small section on continuous random variables however they still play a huge part in the rest of the chapter and the rest of the book. So let's hope you enjoyed the basics because there is more to come!

2.3 Expectation, Variance and Quantiles

We now move onto finding the expectation, variance and quantiles of random variables. We look at the discrete case first and then move onto the continuous case after that.

Discrete case

The expected value originated in the middle of the 17th century from the study of the "problem of points" which is a classical problem in probability theory that led Blaise Pascal to the first reasoning about what today is known as an expected value.

The expectation or expected value of a discrete random variable X is:

$$\begin{aligned} E(X) &= \sum_{x \in H} x\, p_X(x) \\ &= \sum_{x \in H} x \sum_{\omega : X\{\omega\} = x} P(\{\omega\}) \\ &= \sum_{x \in H} \sum_{\omega : X\{\omega\} = x} X(\{\omega\}) P(\{\omega\}) \\ &= \sum_{\omega \in \Omega} X(\{\omega\}) P(\{\omega\}) \end{aligned}$$

This gives two definitions for the expected value of a discrete random variable. The top one and the last line are equal to each other and so are both fine to be used for the definition and for working out the expected value.

Example

Take the example with the 6 sided dice mentioned previously. In this example we had 1,2,3,4,5,6 as the sample space all with a probability of $\frac{1}{6}$. We want to find the expectation of this.

$$\begin{aligned} E(X) &= \sum_{x=1}^{6} x\, p_X(x) \\ &= \frac{1}{6}(1+2+3+4+5+6) \\ &= 3.5 \end{aligned}$$

Example

Take the example with the pmf of $p_X(x) = \frac{1}{220}x^2$. We want to find the expectation of this.

$$\begin{aligned} E(X) &= \sum_{x \in H} x\, p_X(x) \\ &= 2\left(\frac{4}{220}\right) + 4\left(\frac{16}{220}\right) + 6\left(\frac{36}{220}\right) + 8\left(\frac{64}{220}\right) + 10\left(\frac{100}{220}\right) \\ &= \frac{90}{11} \end{aligned}$$

The expected value of a function g of a discrete random variable X is:

$$E(g(X)) = \sum_{x \in H} g(x)\, p_X(x)$$

Try to prove this (it is similar to the proof of the two definitions of the expected value of a discrete random variable).

Example

Take the example with the 6 sided dice mentioned previously. In this example we had 1,2,3,4,5,6 as the sample space all with a probability of $\frac{1}{6}$. We want to find the expectation of $E(X^3)$.

$$\begin{aligned} E(X^3) &= \sum_{x=1}^{6} x^3\, p_X(x) \\ &= \frac{1}{6}(1^3 + 2^3 + 3^3 + 4^3 + 5^3 + 6^3) \\ &= 73.5 \end{aligned}$$

Example

Take $p_X(x) = \frac{1}{4}$ for $r = 1, 2, 3, 4$, find $E(X^2)$

$$\begin{aligned} E(X^2) &= \sum_{x=1}^{4} x^2\, p_X(x) \\ &= \frac{1}{4}(1^2 + 2^2 + 3^2 + 4^2) \\ &= 7.5 \end{aligned}$$

Linearity of Expectation

For arbitrary functions g and h, and constant c:

$$E(g(X) + h(X)) = E(g(X)) + E(h(X))$$
$$E(cg(X)) = cE(g(X))$$

Note also that $E[c] = c$. Try to prove these results by using the definition of expectation of a discrete random variable.

Example

Find $E(X^3 + X + 1)$ of the discrete random variable of the 6 sided dice example.

$$\begin{aligned} E(X^3 + X + 1) &= E(X^3) + E(X) + E(1) \\ &= 73.5 + 3.5 + 1 \\ &= 78 \end{aligned}$$

This is from the linearity of expectation and from previous calculations of $E(X^3)$, $E(X)$ and also using $E[c] = c$.

Example

Find $E(X^3)$ if $E(X(X^2 + 1) - 4) = 5$ and $E(X) = 3$

$$\begin{aligned} 5 = E(X(X^2 + 1) - 4) &= E(X^3 + X - 4) \\ &= E(X^3) + E(X) - E(4) \\ &= E(X^3) + 3 - 4 \end{aligned}$$

Hence $E(X^3) = 6$. Again we used the linearity of expectation.

While expectation is a measure of the location of the pmf, the variance is a measure of the spread of a random variable. Variance was first introduced by Ronald Fisher in 1918 in his article on theoretical population genetics.

The variance of a random variable X, Var(X), is defined as:

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

Like with the two definitions of the expectation of a discrete random variable there are also two definitions of the variance of a discrete random variable as shown above. The proof is left to the reader (Hint use $m = E(X)$ along with the linearity of expectation).

The standard deviation of X, s.d.(X), is defined to be the square root of the variance.

Example

Take the example of the 6 sided dice. We want to find the variance and standard deviation.

We already have that $E(X) = 3.5$, but need $E(X^2)$

$$\begin{aligned} E(X^2) &= \sum_{x=1}^{6} x^2 \, p_X(x) \\ &= \frac{1}{6}(1^2 + 2^2 + 3^2 + 4^2 + 5^2 + 6^2) \\ &= \frac{91}{6} \end{aligned}$$

$$\text{Hence } \mathrm{Var}(X) = E[X]^2 - (E[X^2])^2 = \tfrac{91}{6} - 3.5^2 = \tfrac{35}{12}$$

Also s.d.$(X) = \sqrt{\frac{35}{12}} = \frac{\sqrt{105}}{6}$

Example

Take $p_X(x) = \frac{1}{4}$ for $r = 1, 2, 3, 4$, find $\mathrm{Var}(X)$ and s.d.(X)

We already have that $E(X^2) = 7.5$, but we need $E(X)$

$$\begin{aligned} E(X) &= \sum_{x=1}^{4} x \, p_X(x) \\ &= \frac{1}{4}(1 + 2 + 3 + 4) \\ &= 2.5 \end{aligned}$$

$$\text{Hence } \mathrm{Var}(X) = E[X]^2 - (E[X^2])^2 = 7.5 - 2.5^2 = 1.25$$

Also s.d.$(X) = \sqrt{1.25} = \frac{\sqrt{5}}{2}$

When we discussed expectation we saw that we had the linearity of expectation. There is a similar result for variance.

For constants a and b and random variable X:

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

This uses the definition of variance and linearity of expectation.

Example

Take the example of the 6 sided dice. We want to find $\text{Var}(5X-2)$ and $\text{Var}(2X+4)$.

We already have that $\text{Var}(X) = \frac{35}{12}$

$$\text{Var}(5X-2) = 5^2\text{Var}(X) = 25\left(\frac{35}{12}\right) = \frac{875}{12}$$
$$\text{Var}(2X+4) = 2^2\text{Var}(X) = 4\left(\frac{35}{12}\right) = \frac{35}{3}$$

Finally the corresponding for standard deviation is $\text{s.d.}(aX+b) = |a|\text{s.d.}(X)$. Again for random variable X and constants a and b.

Now we have covered all things expectation and variance in the discrete case it is time to look at the continuous. It follows on nicely from the continuous random variables section and using the discrete case one could start to think about what we might see in this section particularly the definitions of expectation and variance etc.

Continuous Case

The expected value of a continuous random variable X is:

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

This could've been "logically" deduced by thinking about continuous random variables and the definition of the discrete random variable expectation and applying to the continuous case. That is the same with most of what is covered

in this section including this result:

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

Example

Take the following pdf:

$$f_X(x) = \begin{cases} \frac{1}{2} & \text{for } 0 < x \leq 2 \\ 0 & \text{otherwise} \end{cases}$$

Find:

1. $E(X)$
2. $E(2X)$
3. $E(X^2 + 1)$

We use the definition of expectation of a continuous random variable and the result just stated involving a function of X to answer these:

1. $E(X)$

$$\begin{aligned} E(X) &= \int_{-\infty}^{\infty} x f_X(x)\ dx \\ &= \int_0^2 \frac{x}{2}\ dx \\ &= \left[\frac{x^2}{4}\right]_0^2 \\ &= 1 \end{aligned}$$

2. $E(2X)$

$$\begin{aligned} E(X) &= \int_{-\infty}^{\infty} 2x f_X(x)\ dx \\ &= \int_0^2 x\ dx \\ &= \left[\frac{x^2}{2}\right]_0^2 \\ &= 2 \end{aligned}$$

3. $E(X^2+1)$

$$\begin{aligned} E(X) &= \int_{-\infty}^{\infty} (x^2+1) f_X(x)\, dx \\ &= \int_0^2 \frac{x^2}{2} + \frac{1}{2}\, dx \\ &= \left[\frac{x^3}{6} + \frac{x}{2}\right]_0^2 \\ &= \frac{7}{3} \end{aligned}$$

Notice that $E(X) = 1$, this was quite obvious without even solving the integral as you could've done it by symmetry. This is because the pdf of $\frac{1}{2}$ is symmetrical about the point 1 since half of the distribution is between 0 and 1 and half is between 1 and 2. Notice also that $E(2X) = 2 = 2E(X)$, this was also quite obvious from the linearity of expectation however we will go over this again for the continuous case next. The same can be said for part 3 as in $E(X^2+1) = E(X^2) + 1$, so we could have found just the integral of $x^2 f_X(x)$ and then added 1 after (Try this, you should get the same answer!).

For constants a and b and continuous random variable X and arbitrary functions g and h:

$$E[ag(X) + bh(X)] = aE[g(X)] + bE[h(X)]$$

Proof

$$\begin{aligned} E[ag(X) + bh(X)] &= \int_{-\infty}^{\infty} [ag(x) + bh(x)] f_X(x)\, dx \\ &= a\int_{-\infty}^{\infty} g(x)) f_X(x)\, dx + b\int_{-\infty}^{\infty} h(x)) f_X(x)\, dx \\ &= aE[g(X)] + bE[h(X)] \end{aligned}$$

Example

Take the following pdf:

$$f_X(x) = \begin{cases} \frac{x}{2} & \text{for } 0 < x \leq 2 \\ 0 & \text{otherwise} \end{cases}$$

Find:

1. $E(4X^3 - X + 2)$
2. $E(2X^2 + 4X - 5)$

We use the linearity of expectation to solve these:

1. $E(4X^3 - X + 2)$

$$\begin{aligned} E(4X^3 - X + 2) &= 4E(X^3) - E(X) + 2 \\ &= 4\int_0^2 \frac{x^4}{2}\,dx - \int_0^2 \frac{x^2}{2}\,dx + 2\int_0^2 dx \\ &= \left[\frac{4x^5}{10} - \frac{x^3}{6} + 2x\right]_0^2 \\ &= \frac{232}{15} \end{aligned}$$

2. $E(2X^2 + 4X - 5)$

$$\begin{aligned} E(2X^2 + 4X - 5) &= 2E(X^2) = 4E(X) + 2 \\ &= 2\int_0^2 \frac{x^3}{2}\,dx + 4\int_0^2 \frac{x^2}{2}\,dx - 5\int_0^2 dx \\ &= \left[\frac{2x^4}{8} + \frac{x^3}{6} - 5x\right]_0^2 \\ &= -\frac{14}{3} \end{aligned}$$

Now solve them directly from the expectation definition, you should get the same answers.

We remember the definition of variance from the discrete case section.

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

Example

Take the following pdf:

$$f_X(x) = \begin{cases} \frac{1}{2} & \text{for } 0 < x \leq 2 \\ 0 & \text{otherwise} \end{cases}$$

Find:

1. $\mathrm{Var}(X)$
2. s.d.(X)
3. $\mathrm{Var}(2X + 1)$

We already have $E(X) = 1$ and $E(X^2) = \frac{4}{3}$ from previous examples.

1. $\text{Var}(X)$

$$\text{Var}(X) = E(X^2) - (E(X))^2 = \frac{4}{3} - 1^2 = \frac{1}{3}$$

2. s.d.(X)

$$\text{s.d.}(X) = \sqrt{Var(X)} = \sqrt{\frac{1}{3}} = \frac{1}{\sqrt{3}}$$

3. $\text{Var}(2X + 1)$

$$\text{Var}(2X + 1) = 4\text{Var}(X) = \frac{4}{3}$$

Last in this section is quantiles. Quantiles output the value of a random variable such that its probability is less than or equal to an input probability value. Quantiles with x_p the $100p\%$ quantile are defined by:

$$F_X(x_p) = p$$

Special types of quantiles we might be interested in:

1. Median

   The median is the 50% quantile, $x_{0.5}$ and $F_X(x_{0.5}) = 0.5$. It is the middle of a distribution as half the values are less than this value and half are greater.

2. Quartiles

   The quartiles consist of the lower quartile $x_{0.25}$, the median $x_{0.5}$ and the upper quartile $x_{0.75}$ and they split the distribution into four sections.

$$\begin{aligned}\mathbb{P}(X < x_{0.25}) &= \mathbb{P}(x_{0.25} < X < x_{0.5}) \\ &= \mathbb{P}(x_{0.5} < X < x_{0.75}) \\ &= \mathbb{P}(X > x_{0.75}) \\ &= 0.25\end{aligned}$$

3. Inter-quartile range

   The inter-quartile range is the difference between the upper quartile and the lower quartile. Which is:

$$x_{0.75} - x_{0.25}$$

Example

Take the following pdf:

$$f_X(x) = \begin{cases} \frac{1}{2} & \text{for } 0 < x \leq 2 \\ 0 & \text{otherwise} \end{cases}$$

Find:

1. The Median
2. The upper and lower quartile
3. The interquartile range

For this example it looks like we would need to go through the integral and find the value of each quantile we want to find. If we did this we would get the right answer however there is an easier way to do it if you can spot it. The pdf is symmetrical about 1, if you draw the distribution you would see this. Therefore we can easily deduce that the median is 1, the lower quartile is 0.5, the upper quartile is 1.5 and the inter-quartile range is $1.5 - 0.5 = 1$. This is a cool trick to get away with doing the integral, the integrals we have mentioned in this so far have been pretty trivial but it is good to spot these things to prevent us having to do any unnecessary calculations.

Example

Take the following pdf:

$$f_X(x) = \begin{cases} \frac{x}{2} & \text{for } 0 < x \leq 2 \\ 0 & \text{otherwise} \end{cases}$$

Find:

1. The Median
2. The $x_{0.3}$ quantile
3. The upper quartile

The last example could be done by simply spotting that the distribution was symmetrical however this one isn't and so we would have to go through the integration. However we have already seen what the cdf is for this pdf in a previous example and so we will use this:

1. The Median

$$\begin{aligned} 0.5 &= \mathbb{P}(X \leq x_{0.5}) \\ &= F_X(x_{0.5}) \\ &= \frac{x^2}{4} \end{aligned}$$

Hence:

$$\begin{aligned} 0.5 &= \frac{x^2}{4} \\ x^2 &= 2 \\ x &= \sqrt{2} \end{aligned}$$

The solution of $-\sqrt{2}$ is dropped because it is outside the range of the distribution and so it doesn't make sense that this would the median and would be wrong if we stated this.

2. The $x_{0.3}$ quantile

$$\begin{aligned} 0.3 &= \mathbb{P}(X \leq x_{0.3}) \\ &= F_X(x_{0.3}) \\ &= \frac{x^2}{4} \end{aligned}$$

Hence:

$$\begin{aligned} 0.3 &= \frac{x^2}{4} \\ x^2 &= 1.2 \\ x &= \frac{\sqrt{30}}{5} \end{aligned}$$

3. The upper quartile

$$\begin{aligned} 0.75 &= \mathbb{P}(X \leq x_{0.75}) \\ &= F_X(x_{0.75}) \\ &= \frac{x^2}{4} \end{aligned}$$

Hence:

$$\begin{aligned} 0.75 &= \frac{x^2}{4} \\ x^2 &= 3 \\ x &= \sqrt{3} \end{aligned}$$

The negative solutions have all been omitted as stated previously because they don't make sense and would be wrong as they are not in the range. Notice how for the x_p quantile we use $x_p = F_X(x_p)$ and remember how this relates to the pdf. Use this to go from the pdf and work through this example again, take care with the limits and integrate like we have previously.

This is the end of this section and we will now move onto special types of random variables, first going through the discrete ones and then moving onto the continuous ones after that.

2.4 Special Types of Discrete Random Variables

We now move onto special types of discrete random variables which are used to model a wide variety of real world scenarios. These are well known discrete random variables that have many uses and advantages and includes Uniform, Bernoulli, Binomial, Geometric and Poisson. First we will go through the Uniform distribution.

Discrete Uniform Random Variables

The term uniform distribution is believed to be used first by J. V. Uspensky in "introduction to mathematical probability" however it is also believed that similar forms of this under different names were used by other mathematicians prior to this date most notably by Thomas Bayes in the 18th century. What we use today though is credited back to J. V. Uspensky in 1937.

The discrete uniform random variable is a symmetrical probability distribution where values $\{0, 1, ..., m\}$ have equal probability and we denote this as $U(0, m)$. The pmf is:

$$p_X(x) = \begin{cases} \frac{1}{m+1} & \text{for } x = 0, 1, ..., m \\ 0 & \text{otherwise} \end{cases}$$

$$E(X) = \frac{m}{2}$$
$$\text{Var}(X) = \frac{m(m+2)}{12}$$

Example

Using the definition of the expectation of a discrete random variable find the expectation of a discrete uniform random variable.

$$\begin{aligned} E(X) &= \sum_{x=0}^{\infty} x\, p_X(x) \\ &= \sum_{x=0}^{m} x \frac{1}{m+1} \\ &= \frac{1}{m+1} \sum_{x=0}^{m} x \\ &= \frac{1}{m+1} \frac{1}{2} m(m+1) = \frac{m}{2} \end{aligned}$$

Note that we used the standard result $\sum_{x=0}^{m} x = \frac{1}{2} m(m+1)$

Now see if you can find the variance of the discrete uniform random variable using the definition of variance and the fact that $\sum_{x=0}^{m} x^2 = \frac{1}{6}m(m+1)(2m+1)$.

Bernoulli Random Variables

The Bernoulli random variable comes from the Swiss mathematician Jacob Bernoulli in work that ended up being published in 1713 which was 8 years after his death. Jacob was part of a mathematically gifted family in which 8 of them went onto contribute substantially to the development of mathematics and physics, even him by himself published many works over his 50 years of life although Leonard Euler was able to solve a famous convergence problem in 1737 that even Jacob couldn't solve.

The Bernoulli random variable has a sample space of $\{0, 1\}$ and most of the time 0 is for failure and 1 is for success in an event. Using this we can see that $p_X(0) = 1 - \theta$ and $p_X(1) = \theta$ and hence the pmf is:

$$p_X(x) = \begin{cases} \theta^x(1-\theta)^{1-x} & \text{for } x = 0, 1 \\ 0 & \text{otherwise} \end{cases}$$

$$E(X) = \theta$$
$$\text{Var}(X) = \theta(1-\theta)$$

See if you can prove the expectation and variance formulas using the definitions for discrete random variables (Because there are only 2 outcomes for Bernoulli it should be easier than for the uniform).

Binomial Random Variables

The binomial random variable is a model for the outcomes of experiments which count the number of 1 values in a sequence of n independent Bernoulli trials. Stating this makes sense from where the binomial distribution comes from as it also comes from Jacob Bernoulli's published work in 1713 called Ars Conjectandi. We denote the binomial distribution $B(n, \theta)$ and the pmf is:

$$p_X(x) = \begin{cases} \binom{n}{x}\theta^x(1-\theta)^{n-x} & \text{for } x = 0, 1, ..., n \\ 0 & \text{otherwise} \end{cases}$$

$$E(X) = n\theta$$
$$\text{Var}(X) = n\theta(1-\theta)$$

Example

Say we toss a fair coin 20 times and denote X as the number of heads thrown in the 20 tosses.

Find:

1. $\mathbb{P}(X = 12)$
2. $\mathbb{P}(X \leq 3)$
3. $E(X)$ and $\text{Var}(X)$

We use the pmf of the binomial to solve these:

1. $\mathbb{P}(X = 12)$

$$\begin{aligned}\mathbb{P}(X = 12) &= \binom{20}{12}\left(\frac{1}{2}\right)^{12}\left(1-\frac{1}{2}\right)^{20-12} \\ &= 0.120 \text{ (to 3 d.p.)}\end{aligned}$$

2. $\mathbb{P}(X \leq 3)$

$$\begin{aligned}\mathbb{P}(X \leq 3) &= \mathbb{P}(X = 0) + \mathbb{P}(X = 1) + \mathbb{P}(X = 2) \\ &= \binom{20}{0}\left(\frac{1}{2}\right)^{0}\left(1-\frac{1}{2}\right)^{20-0} \\ &+ \binom{20}{1}\left(\frac{1}{2}\right)^{1}\left(1-\frac{1}{2}\right)^{20-1} \\ &+ \binom{20}{2}\left(\frac{1}{2}\right)^{2}\left(1-\frac{1}{2}\right)^{20-2} \\ &= 0.0002 \text{ (to 4 d.p.)}\end{aligned}$$

3. $E(X)$ and $\text{Var}(X)$

$$E(X) = n\theta = 20\left(\frac{1}{2}\right) = 10$$

$$Var(X) = n\theta(1-\theta) = 20\left(\frac{1}{2}\right)\left(1-\frac{1}{2}\right) = 5$$

Geometric Random Variables

It is hard to pinpoint the exact moment when the Geometric distribution is first used however Jacob Bernoulli's work on Bernoulli trials and Laplace's work in probability and distributions definitely helped pave the way for it as probability became more systematic in the 19th century.

The Geometric distribution is one that counts the number of 0 values before

the first 1 in a sequence of independent Bernoulli trials and is denoted as $G(\theta)$ and the pmf is:

$$p_X(x) = \begin{cases} (1-\theta)^x\theta & \text{for } x = 0, 1, ..., n \\ 0 & \text{otherwise} \end{cases}$$

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

Example

Let 1-p be the probability of a win in a football game. p is the probability of a loss since it is a knockout tournament and you cannot draw the game. You start in the round of 32 game and so there are 5 games to win the whole tournament. Let X denote the number of games team A win before being knocked out of the tournament.

Find:

1. in terms of p, the probability that team A gets knocked out in the semi final
2. in terms of p, the probability that team A gets to the semi final
3. p, if the probability of team A getting to the quarter final is 0.6
4. using this p, find the probability that team A wins the tournament
5. using this p, what stage of the tournament are team A expected to get to

We will use the pdf of the Geometric distribution to solve these:

1. in terms of p, the probability that team A gets knocked out in the semi final
$$\mathbb{P}(X = 3) = (1-p)^3 p$$
2. in terms of p, the probability that team A gets to the semi final
$$\mathbb{P}(X > 3) = 1 - \mathbb{P}(X \leq 3) = 1 - (1 - (1-p)^3) = (1-p)^3$$
3. p, if the probability of team A getting to the quarter final is 0.6

   Getting to the quarter final:
$$\begin{aligned} \mathbb{P}(X > 2) &= 1 - \mathbb{P}(X \leq 3) \\ &= 1 - (1 - (1-p)^2) \\ &= (1-p)^2 \end{aligned}$$

Hence:

$$\begin{aligned}(1-p)^2 &= 0.6\\ 1-2p+p^2 &= 0.6\\ p^2-2p+0.4 &= 0\\ p &= \frac{5-\sqrt{15}}{5}\end{aligned}$$

Here we omit the solution $\frac{5+\sqrt{15}}{5}$ as this is greater than 1 and by the axioms of probability this cannot be a probability.

4. using this p, find the probability that team A wins the tournament

$$\begin{aligned}\mathbb{P}(X>5) &= 1-\mathbb{P}(X\leq 5)\\ &= 1-(1-(1-p)^5)\\ &= (1-p)^5\\ &= \left(1-\frac{5-\sqrt{15}}{5}\right)^5\\ &= 0.279 \text{ (to 3 d.p.)}\end{aligned}$$

5. using this p, what stage of the tournament are team A expected to get to

$$\begin{aligned}E(X) &= \frac{1-p}{p}\\ &= 3.437 \text{ (to 3 d.p.)}\end{aligned}$$

So team A are expected to at least win 3 games and get to the semi final of the tournament but might struggle in reaching any rounds after that.

Poisson Random Variables

The Poisson distribution was first introduced by Siméon Denis Poisson in 1837 where he published some work on probability theory.

The Poisson distribution expresses the probability of a given number of events occurring in a fixed interval of time. These events are independent with a known constant mean. We denote the Poisson distribution Pois(λ) and the pmf is:

$$p_X(x) = \begin{cases}\frac{\lambda^x e^{-\lambda}}{x!} & \text{for } x = 0, 1, ...,\\ 0 & \text{otherwise}\end{cases}$$

$$E(X) = \text{Var}(X) = \lambda$$

Example

In a football tournament on average 3.2 goals are scored in each game. Use the Poisson distribution to find:

1. $\mathbb{P}(X = 2)$ and $\mathbb{P}(X = 5)$
2. $\mathbb{P}(X \leq 2)$
3. state what would happen as $x \to \infty$

We use the pmf of the Poisson distribution to solve these:

1. $\mathbb{P}(X = 2)$ and $\mathbb{P}(X = 5)$

$$\begin{aligned}\mathbb{P}(X = 2) &= \frac{3.2^2 e^{-3.2}}{2!} \\ &= 0.209 \text{ (to 3 d.p.)}\end{aligned}$$

$$\begin{aligned}\mathbb{P}(X = 5) &= \frac{3.2^5 e^{-3.2}}{5!} \\ &= 0.114 \text{ (to 3 d.p.)}\end{aligned}$$

2. $\mathbb{P}(X \leq 2)$

$$\begin{aligned}\mathbb{P}(X \leq 2) &= \mathbb{P}(X = 0) + \mathbb{P}(X = 1) + \mathbb{P}(X = 2) \\ &= \frac{3.2^0 e^{-3.2}}{0!} + \frac{3.2^1 e^{-3.2}}{1!} + \frac{3.2^2 e^{-3.2}}{2!} \\ &= 0.380 \text{ (to 3 d.p.)}\end{aligned}$$

3. state what would happen as $x \to \infty$

   As $x \to \infty$ we are calculating the probability of more and more goals being scored in each game and logically we would expect the probability to decrease when the expectation is 3.2 goals. If we are calculating the probability of scoring 50 goals for example even for this small number in the greater picture we would get a small probability of scoring 50 goals in a game. Hence as can be seen by the distribution and by our logical thinking we would see that the probabilities quite rapidly tend to 0 as $x \to \infty$.

We have now gone through the main special types of discrete random variables. We will now move on to the special types of continuous random variables.

2.5 Special Types of Continuous Random Variables

Uniform Distribution

A continuous random variable for which all outcomes in a given range have equal chance of occurring is said to be uniformly distributed. The uniform distribution over the interval (a, b) is denoted $U(a, b)$ and the pdf and cdf are:

$$f_X(x) = \begin{cases} \frac{1}{b-a} & \text{for } a < x < b \\ 0 & \text{otherwise} \end{cases}$$

$$F_X(x) = \begin{cases} 0 & x \leq a \\ \frac{x-a}{b-a} & a < x < b \\ 1 & x \geq b \end{cases}$$

$$E(X) = \frac{b+a}{2}$$
$$\text{Var}(X) = \frac{(b-a)^2}{12}$$

See if you can show that the cdf is what it is by using the pdf and the relationship between continuous pdf and cdf.

Exponential Distribution

The Exponential distribution is a probability distribution of time between events in the Poisson point process. It is denoted by $\text{Exp}(\beta)$ and the pdf and cdf are:

$$f_X(x) = \begin{cases} \beta e^{-\beta x} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

$$F_X(x) = \begin{cases} 1 - e^{-\beta x} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

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

Example

Find:

1. The median of the exponential distribution when $\beta = 3$
2. The variance of the exponential distribution in terms of β

We will use the cdf to answer 1. and the pdf and the expectation to answer 2.

1. The median of the exponential distribution when $\beta = 3$

$$\begin{aligned} 0.5 &= \mathbb{P}(X \leq x_{0.5}) \\ &= F_X(x_{0.5}) \\ &= 1 - e^{-3x} \end{aligned}$$

   Hence:

$$\begin{aligned} 0.5 &= e^{-3x} \\ \ln(0.5) &= -3x \\ x &= \frac{-\ln(0.5)}{3} = 0.231 \text{ (to 3 d.p.)} \end{aligned}$$

   While we have solved the median for a particular β we can also find what it is in terms of β and it can be shown that this is $\frac{\ln(2)}{\beta}$. Try to show this in a similar way to the question.

2. The variance of the exponential distribution in terms of β

$$\begin{aligned} E(X^2) &= \int_{-\infty}^{\infty} x^2 f_X(x)\, dx \\ &= \int_0^{\infty} \beta x^2 e^{-\beta x}\, dx \\ &= \frac{2}{\beta^2} \end{aligned}$$

   This was done using integration by parts twice, which is a prerequisite to this book, make sure you can do it to solve this integral.

$$\begin{aligned} \text{Var}(X) &= E(X^2) - (E(X))^2 \\ &= \frac{2}{\beta^2} - \left(\frac{1}{\beta}\right)^2 \\ &= \frac{2}{\beta^2} - \frac{1}{\beta^2} \\ &= \frac{1}{\beta^2} \end{aligned}$$

The memoryless property

The exponential distribution has a property called the memoryless property and is the only distribution to have this property. It shows it's lack of memory and a random variable satisfies this property if:

$$\mathbb{P}(X > s + t | X > t) = \mathbb{P}(X > s)$$

This is the probability that a random variable exceeds s + t given that it has already exceeded t is equal to the probability that it exceeds s. It has no memory of how large it is already which is quite a unique and cool result related to the exponential distribution and can be quite useful. The proof is left for the reader with the hint that you should use the formula for conditional probability and then the pdf of the exponential distribution.

Gamma Distribution

The Gamma distribution was first defined as the distribution of a "precision constant" by Laplace in 1836.

A random variable X has a Gamma distribution with shape parameter α and rate parameter β if its pdf is given by:

$$f_X(x) = \begin{cases} \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

We denote the Gamma distribution as Gamma(α, β)

It has expectation and variance of:

$$E(X) = \frac{\alpha}{\beta}$$
$$\text{Var}(X) = \frac{\alpha}{\beta^2}$$

When we look at the pdf of the Gamma we see it contains the Gamma function $\Gamma(x)$. This is quite an important function in probability and statistics and is not only used in the Gamma distribution but also other distributions throughout probability including some of them we will go through in this book. It makes sense now to show what the Gamma function is and give it some explanation:

Gamma function

The Gamma function $\Gamma(s)$ is:

$$\Gamma(s) = \int_0^\infty s^{s-1}e^{-s}\, ds$$

Things to note about the Gamma function:

- $\Gamma(1) = \int_0^\infty e^{-s}\, ds = [-e^{-s}]_0^\infty = 1$
- $\Gamma(s+1) = s\Gamma(s)$ for $s > 0$
- $\Gamma(s) = (s-1)!$ for positive integers s

Example

Suppose we have a Gamma distribution of:

$$f_X(x) = \begin{cases} tx^{\alpha-1}e^{-\beta x} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

Find t in terms of β and α and then determine what t would be if $\alpha = 3$ and $\beta = 5$

$$\begin{aligned} \int_0^1 tx^{\alpha-1}e^{-\beta x}\, dx &= t\frac{\Gamma(\alpha)}{\beta^\alpha} \times \frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^1 x^{\alpha-1}e^{-\beta x}\, dx \\ &= t\frac{\Gamma(\alpha)}{\beta^\alpha} \times 1 \\ &= 1 \end{aligned}$$

Hence:

$$\begin{aligned} t &= \frac{\beta^\alpha}{\Gamma(\alpha)} \\ &= \frac{5^3}{\Gamma(3)} \\ &= 62.5 \text{ (From } \alpha = 3 \text{ and } \beta = 5) \end{aligned}$$

We used the fact that everything after the multiplication on the first line is the pdf of a Gamma with parameters α and β and so over the full range of the pdf it integrates to 1 (This is a property of the pdf if you remember!). We used the same logic from the distribution we were given and the last line equalling 1. The value of t could also have been easily deduced from the pdf of the Gamma.

Beta Distribution

The origin of the Beta distribution has been traced back to 1676 when it was thought to be first used by Sir Isaac Newton in a letter he wrote to Henry Oldenberg.

The Beta distribution is a family of continuous probability distributions defined on the interval $[0, 1]$ in terms of positive parameters α_1 and α_2. It is denoted by $\text{Beta}(\alpha_1, \alpha_2)$ and has pdf of:

$$f_X(x) = \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} x^{\alpha_1 - 1}(1 - x)^{\alpha_2 - 1}$$

$$\begin{aligned} E(X) &= \frac{\alpha_1}{\alpha_1 + \alpha_2} \\ \text{Var}(X) &= \frac{\alpha_1 \alpha_2}{(\alpha_1 + \alpha_2)^2(\alpha_1 + \alpha_2 + 1)} \end{aligned}$$

The range of values for this pdf are $0 \leq x \leq 1$

Example

Prove that the expectation of a Beta random variable is what it is as stated above:

$$\begin{aligned} E(X) &= \int_0^1 s \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} s^{\alpha_1 - 1}(1 - s)^{\alpha_2 - 1} \, ds \\ &= \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \int_0^1 s^{\alpha_1 + 1 - 1}(1 - s)^{\alpha_2 - 1} \, ds \\ &= \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \frac{\Gamma(\alpha_1 + 1)\Gamma(\alpha_2)}{\Gamma(\alpha_1 + 1 + \alpha_2)} \times \frac{\Gamma(\alpha_1 + 1 + \alpha_2)}{\Gamma(\alpha_1 + 1)\Gamma(\alpha_2)} \int_0^1 s^{\alpha_1 + 1 - 1}(1 - s)^{\alpha_2 - 1} \, ds \\ &= \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \frac{\Gamma(\alpha_1 + 1)\Gamma(\alpha_2)}{\Gamma(\alpha_1 + 1 + \alpha_2)} \times 1 \\ &= \frac{\alpha_1}{\alpha_1 + \alpha_2} \end{aligned}$$

We used the fact that on the 3rd line everything after the multiplication is a pdf of a Gamma with parameters $\alpha_1 + 2$ and α_2 and so it will integrate to 1 over the full range of the pdf (This is a property of the pdf if you remember!). We then used the fact that $\Gamma(s + 1) = s\Gamma(s)$ and cancelled out all the Gamma functions (This is a property of the Gamma function if you remember!).

Weibull Distribution

The Weibull distribution was first introduced by Maurice René Fréchet and first applied by Rosin and Rammler to describe a particle size distribution in 1933. Although this was the case the distribution ended up being named after Waloddi Weibull, a Swedish mathematician who described the distribution in detail in 1939.

The distribution is mostly used to model failure times and is a generalisation of the exponential distribution seen early on in this section. It has a shape parameter α and rate parameter β, is denoted by Weib(α, β) and had pdf and cdf of:

$$\begin{aligned} f_X(x) &= \alpha\beta^\alpha x^{\alpha-1} e^{-(\beta x)^\alpha} \text{ for } 0 < x < \infty \\ F_X(x) &= 1 - e^{-(\beta x)^\alpha} \\ E(X) &= \frac{\Gamma\left(1+\frac{1}{\alpha}\right)}{\beta} \\ \text{Var}(X) &= \frac{\Gamma\left(1+\frac{2}{\alpha}\right) - \Gamma\left(1+\frac{1}{\alpha}\right)^2}{\beta^2} \end{aligned}$$

Normal Distribution

The normal distribution was developed as an approximation to the binomial distribution by de Moivre in 1773 and was later used by Laplace in 1783 to study measurement errors and also by Gauss in 1809 in the analysis of astronomical data. It is arguably the most famous and most used distribution of them all largely due to its symmetrical bell shaped curve that makes it useful to model a large variety of different real world scenarios. IQ scores, classroom grades and weights to name a few.

A random variable X has a normal distribution if its pdf is given by:

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\left\{\frac{-(x-\mu)^2}{2\sigma^2}\right\}} \text{for } -\infty < x < \infty$$

We denote the normal distribution as $N(\mu, \sigma^2)$. μ is the mean and σ is the standard deviation. The curve is symmetrical around μ and the width is controlled by σ^2 (bigger σ^2 results in a wider distribution). Look up the curve online and view the images for it to gain a better understanding and a visual representation of the theory behind the normal distribution.

We might want to compare normal distributions against each other or actually just calculate probabilities of one distribution however the pdf is complicated and so calculations are hard. It is often better to standardised our distribution,

this way we can perform calculations much more easily. This is because when we standardised we get a normal distribution with a mean of 0 and a standard deviation of 1 which gives of a way of comparing normal distributions much easier than before and also gives an easier pdf to work with which is:

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{\left\{\frac{-z^2}{2}\right\}} \quad -\infty < z < \infty$$

Hence the cdf is:

$$\Phi(z) = \mathbb{P}(Z \leq z) = \int_{-\infty}^{z} \phi(s)ds = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{\left\{\frac{-s^2}{2}\right\}} \, ds$$

$\phi(z)$ is called the standard normal distribution. With $\Phi(z)$ being the cdf this would give us the probabilities we would want to find out in problems. Hence we will use this in examples, this can be used alongside a table of standard normal probabilities, coding langauges like r or by using some calculators however in this book we will work out any examples in terms of $\Phi(z)$.

We standardised our distribution using the follow:

If $X \sim \mathrm{N}(\mu, \sigma^2)$, then the random variable:

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

So to standardise the distribution X we need to take away the mean and divide by the standard deviation to get a standard normal distribution.

Example

We have a normal distribution X with mean $\mu = 110$ and variance of $\sigma^2 = 16$. Find the probability that X is greater than 137.

$$\begin{aligned}
\mathbb{P}(X > 137) &= \mathbb{P}\left(\frac{X - 110}{\sqrt{16}} > \frac{137 - 110}{\sqrt{16}}\right) \\
&= \mathbb{P}\left(Z > \frac{27}{4}\right) \\
&= 1 - \mathbb{P}\left(Z \leq \frac{27}{4}\right) \\
&= 1 - \Phi(6.75)
\end{aligned}$$

Example

Suppose we have a normal distribution $X \sim \mathrm{N}(\mu, \sigma^2)$ with $\mathbb{P}(X < 74) = 0.22$, $\mathbb{P}(X > 89) = 0.4$, $\Phi^{-1}(0.22) = -0.772$ and $\Phi^{-1}(0.6) = 0.253$. Find μ and σ^2

We have:

$$-0.772 = \frac{74-\mu}{\sigma}$$
$$0.253 = \frac{89-\mu}{\sigma}$$

This gives:

$$\mu = 74 + 0.772\sigma$$
$$\mu = 89 - 0.253\sigma$$

We can solve for σ by equating both the right hand sides of the equations since they both equal μ and then we can solve for μ by plugging the σ value back into one of the equations. When we do this we get:

$$\mu = 85.298$$
$$\sigma = 14.634$$
$$\sigma^2 = 214.158$$

These are all to 3 decimal places. Note that $\Phi^{-1}(z)$ is the inverse normal and it finds the x value corresponding to the probability z. In other words it finds x such that $\mathbb{P}(X \leq x) = z$. The input value of the inverse normal is the area to the left of the normal curve (That is why we used 0.6 rather than 0.4 for the second equation!).

Cauchy Distribution

The Cauchy distribution was first studied by French mathematician Augustin-Louis Cauchy in the 19th century and later applied by Dutch physicist Hendrik Lorentz to explain forced resonance or vibrations.

The Cauchy distribution has pdf and cdf of:

$$f_X(x) = \frac{1}{\pi(1+x^2)} \text{ for } -\infty < x < \infty$$
$$F_X(x) = \frac{1}{\pi}\arctan(x) + \frac{1}{2}$$
$$E(X) \text{ is not defined}$$

Example

Suppose we have $X \sim$ Cauchy. Find:

1. $\mathbb{P}(X \leq b)$
2. $\mathbb{P}(X \leq 4)$

3. The median of X

We will use the cdf of the Cauchy to answer these:

1. $\mathbb{P}(X \leq b)$

$$\begin{aligned}\mathbb{P}(X \leq b) &= F_X(b) \\ &= \frac{1}{\pi}\arctan(b) + \frac{1}{2}\end{aligned}$$

2. $\mathbb{P}(X \leq 4)$

$$\begin{aligned}\mathbb{P}(X \leq 4) &= F_X(4) \\ &= \frac{1}{\pi}\arctan(4) + \frac{1}{2} \\ &= 0.922 \text{ to 3 d.p.}\end{aligned}$$

3. The median of X

$$\begin{aligned}&F_X(x) = 0.5 \\ &\frac{1}{\pi}\arctan(x) + \frac{1}{2} = 0.5 \\ &\arctan(x) = 0\end{aligned}$$

The solutions to this equation are:

$$x = k\pi \text{ for } k \in \mathbb{Z}$$

The solutions to 3, are in radians. However if you look up the graph of a Cauchy it is symmetrical about the point 0 and it can be deduced easily that the median is 0 (Which is the case of k = 0 in our solution). Try to plot the cdf graph on graphing software you can also see the median is 0 because $F_X(0) = 0.5$. Try to do these questions again but by integrating the pdf, use the fact that $\int \frac{1}{1+x^2}\, dx = \arctan(x)$.

χ^2 Distribution

The chi-squared distribution was first introduced by Friedrich Robert Helmert in papers in 1875/76 where he computed the sampling distribution of the sample variance of a normal population. It was initially named the Helmert distribution in memory of the German statistician however it was rediscovered by Karl Pearson in the context of the goodness of fit in 1900 where he developed the Pearson's chi-squared test. The name chi comes from the Greek letter χ and is pronounced like the name Kai.

The distribution has pdf of (with $v > 0$):

$$\begin{aligned} f_X(x) &= \frac{1}{2^{\frac{v}{2}}\Gamma(\frac{v}{2})} x^{\frac{v}{2}-1} e^{\left(\frac{-x}{2}\right)} \text{ for } -\infty < x < \infty \\ E(X) &= v \\ \text{Var}(X) &= 2v \end{aligned}$$

We denote the chi-squared distribution as $X \sim \chi^2_v$ with v degrees of freedom. Another cool thing about the distribution is that the χ^2_v distribution is the Gamma$(\frac{v}{2}, \frac{1}{2})$.

t Distribution

The t distribution was first derived as a posterior distribution (A type of distribution used in Bayesian statistics) by Friedrich Robert Helmert and Jacob Lüroth in 1876. The t distribution is used a lot in statistics and in hypothesis testing when calclulating something called a t statistic by performing a t test.

The t distribution's pdf with $v > 0$ degrees of freedom is:

$$\begin{aligned} f_X(x) &= \frac{\Gamma(\frac{v+1}{2})}{\sqrt{v}\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{v}{2}\right)} \left(1 + \frac{x^2}{v}\right)^{-\frac{v+1}{2}} \text{ for } -\infty < x < \infty \\ E(X) &= 0 \text{ when } v > 1 \text{ (Otherwise not defined)} \\ \text{Var}(X) &= \frac{v}{v-2} \text{ when } v > 2 \end{aligned}$$

F Distribution
The F distribution was introduced by Sir Ronald Fisher in 1922. It is used in statistics as the distribution of the test statistic in analysis of variance (ANOVA).

The F distribution with $v_1 > 0$ and $v_2 > 0$ degrees of freedom is denoted as $X \sim F_{v_1,v_2}$ has pdf of:

$$\begin{aligned} f_X(x) &= \frac{\Gamma\left(\frac{v_1+v_2}{2}\right)}{\Gamma\left(\frac{v_1}{2}\right)\Gamma\left(\frac{v_2}{2}\right)} \frac{v_1^{\frac{v_1}{2}} v_2^{\frac{v_2}{2}} x^{\frac{v_1}{2}-1}}{(v_1 x + v_2)^{\left(\frac{v_1+v_2}{2}\right)}} \text{ for } -\infty < x < \infty \\ E(X) &= \frac{v_2}{v_2 - 2} \text{ when } v_2 > 2 \\ \text{Var}(X) &= \frac{2v_2^2(v_1 + v_2 - 2)}{v_1(v_2-2)^2(v_2-4)} \text{ when } v_2 > 4 \end{aligned}$$

log-Normal Distribution

The log-Normal distribution first appeared in 1879 when Donald McAlister and Francis Galton gave a comprehensive view of the distribution including the median, mode, variance and certain quantiles.

The log-Normal with $\xi \in \mathbb{R}$ and $\sigma^2 > 0$ has pdf of:

$$\begin{aligned} f_X(x) &= \frac{1}{\sqrt{2\pi}} \frac{1}{x\sigma} e^{\left\{ \frac{(\log x - \xi)^2}{2\sigma^2} \right\}} \text{ for } 0 < x < \infty \\ E(X) &= e^{\left(\frac{\sigma^2}{2} + \xi \right)} \\ \text{Var}(X) &= \left(e^{(\sigma^2)} - 1 \right) e^{(\sigma^2 + 2\xi)} \end{aligned}$$

Gumbel Distribution

The Gumbel distribution has been named after Julius Gumbel since 1935 where he introduced the distribution but it is also called the extreme value distribution. This is because it is used to study the maximum of random variables and to model extreme events in statistics. The distribution is also used in machine learning in the Gumbel-max trick and also comes up in number theory and has many more applications.

The Gumbel distribution has location parameter $\alpha \in \mathbb{R}$ and scale parameter $\beta > 0$, is denoted by $\text{GEV}(\alpha, \beta, 0)$ and has pdf and cdf of:

$$\begin{aligned} f_X(x) &= \frac{1}{\beta} e^{\left\{ \frac{-(x-\alpha)}{\beta} \right\}} e^{-e^{\left\{ \frac{-(x-\alpha)}{\beta} \right\}}} \text{ for } -\infty < x < \infty \\ F_X(x) &= e^{-e^{\left\{ \frac{-(x-\alpha)}{\beta} \right\}}} \\ E(X) &= \alpha + \beta\gamma \text{ where } \gamma \approx 0.5772 \text{ is Euler's constant} \\ \text{Var}(X) &= \frac{\beta^2 \pi^2}{6} \end{aligned}$$

This ends the discussion and this section on special types of continuous random variables. We only cover a limited number of distributions in this book and there are many more special types of discrete and continuous random variables out there (maybe try to research more of them if you are interested in learning more!). We could also come up with an infinite amount of other distributions as long as they satisfy the properties of the pdf and cdf for continuous random variables or the pmf and cdf for discrete random variables. The chapter is finished on random variables, i hope you enjoyed it! Next we move onto bivariate distributions which can be tricky at first especially the continuous case as it involves partial differentiation and double integrals however the book will break it down and explain it in detail. Enjoy the next chapter of the book!

3 Bivariate Distributions

So far we have looked at probability distributions of only one random variable and these are called univariate distributions. We now want to look at distributions that take two random variables into account, these are called bivariate distributions. Using such distributions allows us to gain insights into relationships between real life events. Such examples include:

- Correlations between measurements of biological characteristics within a population such as height, weight and blood pressure and the relationship between these traits
- Correlations between environmental factors such as temperature, humidity and pollution levels at different locations to capture dependencies between these factors.
- The relationship between two biological markers in the brain amyloid-β and tau and figuring out the levels of these proteins together to help in diagnosing conditions relating to cognitive function and brain health

So as can be seen some important real life examples can be modelled by bivariate distributions showing how important they can be. In the first and second example more than two factors were given, if we wanted to study between two of these we would use a bivariate distribution but if we wanted to study between more than two of these then we would use a multivariate distribution.

3.1 Discrete Random Variables

Again like in previous sections when considering a new topic in probability we need to discuss the discrete and continuous cases separately. We will go through the continuous case in the next section but first will go through the discrete case and define effectively the equivalent to the probability mass function for univariate distributions but now for bivariate distributions.

Suppose we have the discrete random variables X and Y. The joint probability mass function is:

$$p_{XY}(x, y) = \mathbb{P}(X = x, Y = y)$$

Here "," means "and". This is the probability that two events occur at the same time from two different discrete random variables. Like with the pmf in the univariate case we also have properties that need to be satisfied by the joint pmf which are:

1. $0 \leq p_{XY}(x, y) \leq 1 \, \forall x$ and y
2. $\Sigma_{all\, x,y}\, p_{XY}(x, y) = \Sigma_{x=-\infty}^{\infty} \Sigma_{y=-\infty}^{\infty} p_{XY}(x, y) = 1$
3. $\mathbb{P}((X, Y) \in A) = \Sigma_{all\, x,y \in A}\, p_{XY}(x, y)$

Example

Suppose the joint pmf of X and Y is:

$$p_{X,Y}(x,y) = \frac{xy}{36}$$

for x, y = 1,2,3

1. Calculate the joint probabilities for all x and y
2. Show that this is a valid pmf
3. Find $\mathbb{P}(X = 3)$ and $\mathbb{P}(X \leq Y)$

We will use the definition of the joint pmf to answer these along with its properties:

1. Calculate the joint probabilities for all x and y

$$\mathbb{P}(X=1,Y=1) = \frac{1}{36} \quad \mathbb{P}(X=1,Y=2) = \frac{2}{36} \quad \mathbb{P}(X=1,Y=3) = \frac{3}{36}$$
$$\mathbb{P}(X=2,Y=1) = \frac{2}{36} \quad \mathbb{P}(X=2,Y=2) = \frac{4}{36} \quad \mathbb{P}(X=2,Y=3) = \frac{6}{36}$$
$$\mathbb{P}(X=3,Y=1) = \frac{3}{36} \quad \mathbb{P}(X=3,Y=2) = \frac{6}{36} \quad \mathbb{P}(X=3,Y=3) = \frac{9}{36}$$

2. Show that this is a valid pmf

$$p_{X,Y}(x,y) \geq 0 \text{ for all } x \text{ and } y \text{ and } \sum_{all\ x,y} p_{X,Y}(x,y) = 1$$

3. Find $\mathbb{P}(X = 3)$ and $\mathbb{P}(X \leq Y)$

$$\begin{aligned}\mathbb{P}(X=3) &= \mathbb{P}(X=3,Y=1) + \mathbb{P}(X=3,Y=2) + \mathbb{P}(X=3,Y=3) \\ &= \frac{12}{36} = \frac{1}{3}\end{aligned}$$

$$\begin{aligned}\mathbb{P}(X \leq Y) &= \mathbb{P}(X=1,Y=1) + \mathbb{P}(X=1,Y=2) + \mathbb{P}(X=1,Y=3) \\ &\quad + \mathbb{P}(X=2,Y=2) + \mathbb{P}(X=2,Y=3) + \mathbb{P}(X=3,Y=3) \\ &= \frac{1}{36} + \frac{2}{36} + \frac{3}{36} + \frac{4}{36} + \frac{6}{36} + \frac{9}{36} \\ &= \frac{25}{36}\end{aligned}$$

Try these yourself and see if you can get the same answers! we have now seen how to set up the joint pmf, show how it is a valid joint pmf and how to calculate probabilities from the joint pmf. We will now move onto another example using the properties of the joint pmf.

Example

Suppose the joint pmf of X and Y is:

$$p_{X,Y}(x,y) = \frac{x+y}{c}$$

for x, y = 2,4,6

Find the value of c such that this a valid joint pmf:

$$\mathbb{P}(X=2,Y=2) = \frac{4}{c} \;\; \mathbb{P}(X=2,Y=4) = \frac{6}{c} \;\; \mathbb{P}(X=2,Y=6) = \frac{8}{c}$$
$$\mathbb{P}(X=4,Y=2) = \frac{6}{c} \;\; \mathbb{P}(X=4,Y=4) = \frac{8}{c} \;\; \mathbb{P}(X=4,Y=6) = \frac{10}{c}$$
$$\mathbb{P}(X=6,Y=2) = \frac{8}{c} \;\; \mathbb{P}(X=6,Y=4) = \frac{10}{c} \;\; \mathbb{P}(X=6,Y=6) = \frac{12}{c}$$

Since we need:

$$\sum_{x=-\infty}^{\infty} \sum_{y=-\infty}^{\infty} p_{XY}(x,y) = 1$$

We have:

$$\frac{72}{c} = 1$$
$$c = 72$$

We used the second property of the joint pmf to answer this.

Although we are now talking about the joint distribution of X and Y it is important to note that X and Y still have their own distribution. In the case of joint pmf's these individual distributions are called marginal pmf's. If X and Y are discrete random variables then their marginal pmf's are:

$$p_X(x) = \sum_{y=-\infty}^{\infty} p_{XY}(x,y)$$
$$p_Y(y) = \sum_{x=-\infty}^{\infty} p_{XY}(x,y)$$

Example

Find the marginal pmf's of X and Y when the joint pmf is:

$$p_{X,Y}(x,y) = \frac{xy}{36}$$

for x, y = 1,2,3

For the marginal of X:

$$p_X(1) = \frac{1}{36} + \frac{2}{36} + \frac{3}{36} = \frac{1}{6}$$
$$p_X(2) = \frac{2}{36} + \frac{4}{36} + \frac{6}{36} = \frac{1}{3}$$
$$p_X(3) = \frac{3}{36} + \frac{6}{36} + \frac{9}{36} = \frac{1}{2}$$

For the marginal of Y:

$$p_Y(1) = \frac{1}{36} + \frac{2}{36} + \frac{3}{36} = \frac{1}{6}$$
$$p_Y(2) = \frac{2}{36} + \frac{4}{36} + \frac{6}{36} = \frac{1}{3}$$
$$p_Y(3) = \frac{3}{36} + \frac{6}{36} + \frac{9}{36} = \frac{1}{2}$$

In this example the marginal pmf's are the exact same. Notice how the sum over all outcomes of the marginals still comes to 1 since these are pmf's in their own right. Obviously both marginals are 0 anywhere else where it is not 1,2 or 3.

If X and Y are random variables, the conditional pmf's are:

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

Example

Suppose the joint pmf of X and Y is:

$$p_{X,Y}(x,y) = \frac{xy}{36}$$

for x, y = 1,2,3

1. Find the conditional pmf of X given Y = 3
2. Find the conditional pmf of Y given X = 1

We will use the definition of the conditional pmf and the answers we got from previous examples relating to this particular pmf given.

1. Find the conditional pmf of X given Y = 3

$$p_{X|Y}(x=1|y=3) = \frac{p_{XY}(x=1,y=3)}{p_Y(3)} = \frac{\frac{3}{36}}{\frac{1}{2}} = \frac{1}{6}$$
$$p_{X|Y}(x=2|y=3) = \frac{p_{XY}(x=2,y=3)}{p_Y(3)} = \frac{\frac{6}{36}}{\frac{1}{2}} = \frac{1}{3}$$
$$p_{X|Y}(x=3|y=3) = \frac{p_{XY}(x=3,y=3)}{p_Y(3)} = \frac{\frac{9}{36}}{\frac{1}{2}} = \frac{1}{2}$$

   Hence the conditional pmf is:

$$p_{X|Y}(1|3) = \frac{1}{6}$$
$$p_{X|Y}(2|3) = \frac{1}{3}$$
$$p_{X|Y}(3|3) = \frac{1}{2}$$

   Again this is 0 otherwise. Also notice how we have the probabilities adding to 1 again as this is a pmf in it's own right so satisfies the properties of pmf's. The same can be said about the next question once we answer it.

2. Find the conditional pmf of Y given X = 1

$$p_{Y|X}(y=1|x=1) = \frac{p_{XY}(y=1,x=1)}{p_X(1)} = \frac{\frac{1}{36}}{\frac{1}{6}} = \frac{1}{6}$$
$$p_{Y|X}(y=2|x=1) = \frac{p_{XY}(y=2,x=1)}{p_X(1)} = \frac{\frac{2}{36}}{\frac{1}{6}} = \frac{1}{3}$$
$$p_{Y|X}(y=3|x=1) = \frac{p_{XY}(y=3,x=1)}{p_X(1)} = \frac{\frac{3}{36}}{\frac{1}{6}} = \frac{1}{2}$$

   Hence the conditional pmf is:

$$p_{Y|X}(1|1) = \frac{1}{6}$$
$$p_{Y|X}(2|1) = \frac{1}{3}$$
$$p_{Y|X}(3|1) = \frac{1}{2}$$

We now move onto independence as the last step in this discrete bivariate distributions section before we move onto the continuous case. Independence in terms of bivariate distributions means that one random variable doesn't have

any affect on the other and knowing the value of one random variable doesn't give any information away about the other.

Two random variables X and Y are independent if the events $\{X \in A\}$ and $\{Y \in B\}$ are independent $\forall$ sets A and B:

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

for all sets A and B.

In the case of discrete random variables, two random variables are independent if an only if:

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

for all x and y. See if you can show whether or not the two examples discussed in this section are independent or not based on this formula above.

Using the conditional pmf's formula along with independence it can be shown that:

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

Which also shows that the random variable X has no effect on the random variable Y and vice versa. This ends this section, we now look at the continuous case.

3.2 Continuous Bivariate Random Variables

We now cover the continuous case and this time we start with the joint probability density functions which is the equivalent of the joint probability mass function but now for continuous random variables.

Joint probability density function

If X and Y are both continuous random variables then their joint probability density function is:

$$f_{XY}(x, y) = \frac{\partial^2 F_{XY}(x, y)}{\partial x \partial y} = \frac{\partial^2 F_{XY}(x, y)}{\partial y \partial x}$$

Properties of $f_{XY}(x, y)$:

1. $f_{XY}(x, y) \geq 0 \; \forall (x, y)$ (Positivity)
2. $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{XY}(s, t) \; ds \; dt = 1$ (Summability)
3. $\mathbb{P}((X, Y) \in A) = \int \int_A f_{XY}(s, t) \; ds \; dt$

Like with the univariate case we also have the equivalent of the cumulative distribution function but for bivariate distributions which is called the joint cumulative distribution function and is defined as:

$$\begin{aligned} F_{XY}(x,y) &= \mathbb{P}(X \leq x, Y \leq y) \\ &= \int_{-\infty}^{y} \int_{-\infty}^{x} f_{XY}(s,t)\, ds\, dt \\ &= \int_{-\infty}^{x} \int_{-\infty}^{y} f_{XY}(s,t)\, dt\, ds \end{aligned}$$

Properties of $F_{XY}(x,y)$:

1. $0 \leq F_{XY}(x,y) \leq 1\ \forall (x,y), F_{XY}(-\infty, y) = 0,$
   $F_{XY}(x,-\infty) = 0, F_{XY}(\infty,\infty) = 1$
2. $F_{XY}(x,\infty) = F_X(x), F_{XY}(\infty,y) = F_Y(y)$
3. $F_{XY}(x,y)$ is non-decreasing in both x and y

With the univariate case we could think of the cdf as calculating the area under the function in question. In the bivariate case because we now have two parameters X and Y we can think of the joint cdf as finding the volume under the function (A density surface) in question.

Example

$$f_{XY}(x,y) = \begin{cases} x + \frac{3}{2}y^2 & 0 \leq x \leq 1, 0 \leq y \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

1. Find the joint cdf
2. Find $\mathbb{P}(X < 0.5, Y < 0.25)$

We will use the definition of the joint cdf to answer these.

1. Find the joint cdf

$$\begin{aligned} F_{XY}(x,y) &= \int_{-\infty}^{y} \int_{-\infty}^{x} f_{XY}(s,t)\, ds\, dt \\ &= \int_{0}^{y} \int_{0}^{x} s + \frac{3}{2}t^2\, ds\, dt \\ &= \int_{0}^{y} \frac{x^2}{2} + \frac{3}{2}t^2 x\, dt \\ &= \frac{yx^2}{2} + \frac{y^3 x}{2} \text{ for } 0 \leq x \leq 1, 0 \leq y \leq 1 \end{aligned}$$

2. Find $\mathbb{P}(X < 0.5, Y < 0.25)$

$$\begin{aligned} F_{XY}(x < 0.5, y < 0.25) &= \int_{-\infty}^{0.25} \int_{-\infty}^{0.5} f_{XY}(s,t)\, ds\, dt \\ &= \int_{0}^{0.25} \int_{0}^{0.5} s + \frac{3}{2}t^2\, ds\, dt \\ &= \int_{0}^{y} \frac{3t^2}{4} + \frac{1}{8}\, dt \\ &= \frac{9}{256} \end{aligned}$$

We integrated with respect to each variable in turn in these examples. For the second question we could have also just inputted $x = 0.25$ and $y = 0.5$ into the joint cdf that was calculated in the first question.Try to go backwards in the first question and show the pdf by using partial derivatives on the cdf.

Remember we went through independence in the discrete case, well in the continuous we have the same results except two continuous random variables X and Y are independent if and only if:

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

This means it is relatively simple to show that two continuous random variables are independent using the joint pdf. There are two conditions that two continuous random variables need to satisfy in order to be independent which are:

1. $f_{XY}(x, y) = g(x)h(y)$ (Factorisation)
2. The range of X does not depend on the range of Y and vice-versa

If the two continuous random variables satisfy the second condition we say that they are variationally independent. To show that two continuous random variables are independent it is fairly simple to just show that they satisfy the above conditions however showing that two continuous random variables are not independent may be harder. We can still show they are variationally independent quite simply by checking their ranges however we have two methods to disprove the factorisation condition. Method 1 is to show that a conditional distribution is not the same as a marginal distribution however we don't cover that until later sections in this chapter so we will leave that until then. The second method uses the fact that (The first condition) f_{XY} can be factorised as a function of x times a function of y if and only if for x_1, x_2, y_1, y_2:

$$f_{XY}(x_1, y_1) f_{XY}(x_2, y_2) = f_{XY}(x_1, y_2) f_{XY}(x_2, y_1)$$

Hence if we show this is not satisfied then the first condition (factorisation) does not work either and therefore the two continuous random variables would not

be independent.

Example

$$f_{XY}(x,y) = \begin{cases} x + \frac{3}{2}y^2 & 0 \leq x \leq 1, 0 \leq y \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Decide whether this joint pdf is variationally independent and whether it satisfies the factorisation condition. Hence decide whether it is independent.

It is variationally independent because the range of X does not depend on the range of Y and vice-versa. To show whether it satisfies the first condition we will test the second method we discussed previously. Let $x_1 = y_1 = \frac{1}{2}$ and $x_2 = y_2 = \frac{3}{4}$

$$\begin{aligned} f_{XY}(x_1,y_1)f_{XY}(x_2,y_2) &= \left(\frac{1}{2} + \frac{3}{2}\left(\frac{1}{2}\right)^2\right)\left(\frac{3}{4} + \frac{3}{2}\left(\frac{3}{4}\right)^2\right) \\ &= \frac{357}{256} \end{aligned}$$

$$\begin{aligned} f_{XY}(x_1,y_2)f_{XY}(x_2,y_1) &= \left(\frac{1}{2} + \frac{3}{2}\left(\frac{3}{4}\right)^2\right)\left(\frac{3}{4} + \frac{3}{2}\left(\frac{1}{2}\right)^2\right) \\ &= \frac{387}{256} \end{aligned}$$

Here we have shown $f_{XY}(x_1,y_1)f_{XY}(x_2,y_2) = \frac{357}{256} \neq \frac{387}{256} = f_{XY}(x_1,y_2)f_{XY}(x_2,y_1)$. Hence the joint pdf is unable to be factorised and so it is not independent.

3.3 Marginal Distributions

When we discussed the discrete case for bivariate distributions we talked about the marginal distributions. Like we said previously just because we are considering the joint distribution of X and Y we cannot forget that the individual distributions of X and Y still exist in their own right. We will now go through this but in the continuous case.

Finding the marginal cdf given the joint cdf (For X and Y continuous random variables):

$$\begin{aligned} F_X(x) &= \mathbb{P}(X \leq x) = \mathbb{P}(X \leq x, Y < \infty) = F_{XY}(x,\infty) \\ F_Y(x) &= \mathbb{P}(Y \leq y) = \mathbb{P}(X \leq \infty, Y < y) = F_{XY}(\infty,y) \end{aligned}$$

Finding the marginal pdf given the joint pdf (For X and Y continuous random variables):

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x,t)\ dt$$
$$f_Y(y) = \int_{-\infty}^{\infty} f_{XY}(s,y)\ ds$$

Example

$$f_{XY}(x,y) = \begin{cases} x + \frac{3}{2}y^2 & 0 \le x \le 1, 0 \le y \le 1 \\ 0 & \text{otherwise} \end{cases}$$

1. Find the pdf and cdf of X
2. Find the pdf and cdf of Y

We will use the formulas for the marginal distribution to solve these and do both by using the joint cdf and then by using the joint pdf.

Recall that the joint cdf of this joint pdf is:

$$f_{XY}(x,y) = \begin{cases} \frac{yx^2}{2} + \frac{y^3x}{2} & 0 \le x \le 1, 0 \le y \le 1 \\ 0 & \text{otherwise} \end{cases}$$

1. Find the pdf and cdf of X

   (a) Finding the cdf:

   $$F_X(x) = F_{XY}(x,1) = \frac{1}{2}x(x+1) \text{ for } 0 \le x \le 1$$

   (b) Finding the pdf:

   $$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} f_{XY}(x,t)\ dt \\ &= \int_0^1 x + \frac{3}{2}t^2\ dt \\ &= x + \frac{1}{2} \text{ for } 0 \le x \le 1 \end{aligned}$$

2. Find the pdf and cdf of Y

(a) Finding the cdf:

$$F_Y(y) = F_{XY}(1, y) = \frac{1}{2}y(y^2 + 1) \text{ for } 0 \leq y \leq 1$$

(b) Finding the pdf:

$$\begin{aligned} f_Y(y) &= \int_{-\infty}^{\infty} f_{XY}(s, y)\ ds \\ &= \int_0^1 s + \frac{3}{2}y^2\ ds \\ &= \frac{3}{2}y^2 + \frac{1}{2} \text{ for } 0 \leq y \leq 1 \end{aligned}$$

3.4 Conditional Distributions

We might be interested in the value of one random variable given another. Similar to the section on conditional probability in the 1st chapter it is often good to consider this when we want to calculate probabilities based on past events occurring. Where in the conditional probability case we were looking at the probability an event occurs based on another event that has already occurred, in the conditional distribution case we are looking at one random variable given information about another random variable. This discussion leads on to the mathematical topic called Bayesian inference which is not covered in this book however i will say that the overall idea of the topic is that we use a prior distribution which incorporates the existing knowledge and multiply it by a likelihood to get a posterior distribution. The posterior reflects the updated beliefs in light of new evidence and we can use this to make predictions about unknown parameters.

If X and Y are random variables, then the conditional pdf's:

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

Using the conditional pdf's formula with independence it can be shown that:

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

The prove of this is simple to deduce from the conditional pdf's formula and then using independence and cancelling out (Try it!).

Example

$$f_{XY}(x,y) = \begin{cases} x + \frac{3}{2}y^2 & 0 \leq x \leq 1, 0 \leq y \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Find

1. $f_{X|Y}(X|Y)$

$$\begin{aligned} f_{X|Y}(x|y) &= \frac{f_{XY}(x,y)}{f_Y(y)} \\ &= \frac{x + \frac{3}{2}y^2}{\frac{3}{2}y^2 + \frac{1}{2}} \\ &= \frac{3y^2 + 2x}{3y^2 + 1} \end{aligned}$$

2. $\mathbb{P}(X < b|Y = a)$ for $0 \leq a \leq 1, 0 \leq b \leq 1$

$$\begin{aligned} f_{X|Y}(X|Y = a) &= \frac{x + \frac{3}{2}a^2}{\frac{3}{2}a^2 + \frac{1}{2}} \\ \mathbb{P}(X < b|Y = a) &= \int_0^b \frac{s + \frac{3}{2}a^2}{\frac{3}{2}a^2 + \frac{1}{2}}\, ds \\ &= \frac{1}{\frac{3}{2}a^2 + \frac{1}{2}} \int_0^b s + \frac{3}{2}a^2\, ds \\ &= \frac{1}{\frac{3}{2}a^2 + \frac{1}{2}} \left[\frac{s^2}{2} + \frac{3}{2}a^2 s\right]_0^b \\ &= \frac{\frac{b^2}{2} + \frac{3}{2}a^2 b}{\frac{3}{2}a^2 + \frac{1}{2}} \\ &= \frac{b(3a^2 + b)}{3a^2 + 1} \end{aligned}$$

3. $\mathbb{P}(X < 0.5|Y = 0.25)$

$$\begin{aligned} \mathbb{P}(X < 0.5|Y = 0.25) &= \frac{0.5(3(0.25)^2 + 0.5)}{3(0.25)^2 + 1} \\ &= 0.289 \text{ (To 3 d.p.)} \end{aligned}$$

4. $f_{Y|X}(Y|X)$

$$\begin{aligned} f_{Y|X}(y|x) &= \frac{f_{XY}(x,y)}{f_X(x)} \\ &= \frac{x + \frac{3}{2}y^2}{x + \frac{1}{2}} \\ &= \frac{3y^2 + 2x}{2x + 1} \end{aligned}$$

Even though it is not generally true that:

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

It cannot be overseen in this case that the two are quite similar so a question we might want to think about is what values of x and y are they equal for. Let's set the 2 equal:

$$\begin{aligned} \frac{3y^2 + 2x}{3y^2 + 1} &= \frac{3y^2 + 2x}{2x + 1} \\ 3y^2 + 1 &= 2x + 1 \\ x &= \frac{3}{2}y^2 \end{aligned}$$

If you also consider the x and y values such that $f_X(x) = f_Y(y)$ happens you would also get the same conclusion, have a think why this might be the case by thinking of the formulas for the conditional distributions.

3.5 Expectation and variance

Now we talk about expectation and variance but for bivariate distributions. Again like in the univariate case we consider the discrete case first and then the continuous case after.

Discrete case

Given two discrete random variables X and Y some important results of expectations are:

$$\begin{aligned} E[g(X,Y)] &= \sum_{s=-\infty}^{\infty} \sum_{t=-\infty}^{\infty} g\,(s,t) p_{XY}(s,t) \\ E[X] &= \sum_{s=-\infty}^{\infty} \sum_{t=-\infty}^{\infty} s\, p_{XY}(s,t) = \sum_{s=-\infty}^{\infty} s\, p_{XY}(s,t) \\ E[Y] &= \sum_{s=-\infty}^{\infty} \sum_{t=-\infty}^{\infty} t\, p_{XY}(s,t) = \sum_{t=-\infty}^{\infty} t\, p_{XY}(s,t) \end{aligned}$$

Example

Suppose the joint pmf of X and Y is:

$$p_{XY}(x,y) = \frac{xy}{36}$$

for x,y = 1,2,3

1. Find Var(X)
2. Find Var(Y)

This is simpler than it would look as we already have the marginal distributions from a previous example.

1. Find Var(X)

$$\begin{aligned} E(X) &= \left(1 \times \frac{1}{6}\right) + \left(2 \times \frac{1}{3}\right) + \left(3 \times \frac{1}{2}\right) = \frac{7}{3} \\ E(X^2) &= \left(1^2 \times \frac{1}{6}\right) + \left(2^2 \times \frac{1}{3}\right) + \left(3^2 \times \frac{1}{2}\right) = 6 \\ \text{Var}(X) &= E(X^2) - [E(X)]^2 \\ &= 6 - \left(\frac{7}{3}\right)^2 \\ &= \frac{5}{9} \end{aligned}$$

2. Find Var(Y)

$$\text{Var}(Y) = \frac{5}{9}$$

   It is easily deduced that Var(X) = Var(Y) since X and Y have the same distribution.

Some more important results are:

$$\begin{aligned} E[g(X) + h(Y)] &= \sum_{s=-\infty}^{\infty} \sum_{t=-\infty}^{\infty} [g(s) + h(t)]\, p_{XY}(s,t) \\ &= \sum_{s=-\infty}^{\infty} \sum_{t=-\infty}^{\infty} g(s)\, p_{XY}(s,t) + \sum_{s=-\infty}^{\infty} \sum_{t=-\infty}^{\infty} h(t)\, p_{XY}(s,t) \\ &= E[g(X)] + E[h(Y)] \end{aligned}$$

If X and Y are independent:

$$\begin{aligned} E[g(X)h(Y)] &= \sum_{s=-\infty}^{\infty} \sum_{t=-\infty}^{\infty} g(s)h(t)\ p_{XY}(s,t) \\ &= \sum_{s=-\infty}^{\infty} \sum_{t=-\infty}^{\infty} g(s)h(t)\ p_X(s)p_Y(t) \\ &= \sum_{s=-\infty}^{\infty} g(s)p_X(s) \sum_{t=-\infty}^{\infty} h(t)\ p_Y(t) \\ &= E[g(X)]E[h(Y)] \end{aligned}$$

Continuous case

For continuous random variables X and Y:

$$\begin{aligned} E[g(X,Y)] &= \int_{s=-\infty}^{\infty} \int_{t=-\infty}^{\infty} g(s,t)\ f_{XY}(s,t)\ dt\ ds \\ E[X] &= \int_{s=-\infty}^{\infty} \int_{t=-\infty}^{\infty} s\ f_{XY}(s,t)\ dt\ ds = \int_{-\infty}^{\infty} s\ f_X(s)\ ds \\ E[Y] &= \int_{s=-\infty}^{\infty} \int_{t=-\infty}^{\infty} t\ f_{XY}(s,t)\ dt\ ds = \int_{-\infty}^{\infty} t\ f_Y(t)\ dt \end{aligned}$$

Like in the discrete case we also have:

$$E[g(X) + h(Y)] = E[g(X)] + E[h(Y)]$$

And when X and Y are independent we again have:

$$E[g(X)h(Y)] = E[g(X)]E[h(Y)]$$

Example

$$f_{XY}(x,y) = \begin{cases} x + \frac{3}{2}y^2 & 0 \le x \le 1, 0 \le y \le 1 \\ 0 & \text{otherwise} \end{cases}$$

Find:

1. $E[X]$
2. $E[Y]$
3. $E[XY]$
4. $\text{Var}[X]$

5. $E[X^2(Y+1)]$

We will use the formulas stated above and we remember that $\text{Var}(X) = E(X^2) - (E(X))^2$ to answer these:

1. $E[X]$

$$\begin{aligned} E[X] &= \int_{s=-\infty}^{\infty} \int_{t=-\infty}^{\infty} s\ f_{XY}(s,t)\ dt\ ds \\ &= \int_0^1 \int_0^1 s\ (s + \frac{3}{2}t^2)\ dt\ ds \\ &= \int_0^1 \int_0^1\ s^2 + \frac{3}{2}t^2 s\ dt\ ds \\ &= \int_0^1 s^2 + \frac{s}{2}\ ds \\ &= \frac{7}{12} \end{aligned}$$

2. $E[Y]$

$$\begin{aligned} E[Y] &= \int_{t=-\infty}^{\infty} \int_{s=-\infty}^{\infty} t\ f_{XY}(s,t)\ ds\ dt \\ &= \int_0^1 \int_0^1 t\ (s + \frac{3}{2}t^2)\ ds\ dt \\ &= \int_0^1 \int_0^1\ st + \frac{3}{2}t^3\ ds\ dt \\ &= \int_0^1 \frac{1}{2}t + \frac{3}{2}t^3\ dt \\ &= \frac{5}{8} \end{aligned}$$

3. $E[XY]$

$$\begin{aligned} E[XY] &= \int_{s=-\infty}^{\infty} \int_{t=-\infty}^{\infty} st\ f_{XY}(s,t)\ dt\ ds \\ &= \int_0^1 \int_0^1 st\ (s + \frac{3}{2}t^2)\ dt\ ds \\ &= \int_0^1 \int_0^1\ s^2 t + \frac{3}{2}t^3 s\ dt\ ds \\ &= \int_0^1 \frac{s^2}{2} + \frac{3}{2}s\ ds \\ &= \frac{11}{12} \end{aligned}$$

4. $\text{Var}[X]$

$$\begin{aligned}
E[X^2] &= \int_{s=-\infty}^{\infty}\int_{t=-\infty}^{\infty} s^2\ f_{XY}(s,t)\ dt\ ds \\
&= \int_0^1\int_0^1 s^2\ (s+\frac{3}{2}t^2)\ dt\ ds \\
&= \int_0^1\int_0^1\ s^3+\frac{3}{2}t^2s^2\ dt\ ds \\
&= \int_0^1 s^3+\frac{1}{2}s^2\ ds \\
&= \frac{5}{12}
\end{aligned}$$

Hence:

$$\begin{aligned}
\text{Var}[X] &= \frac{5}{12}-\left(\frac{7}{12}\right)^2 \\
&= \frac{11}{144}
\end{aligned}$$

5. $E[X^2(Y+1)]$

$$\begin{aligned}
E[X^2(Y+1)] &= \int_{s=-\infty}^{\infty}\int_{t=-\infty}^{\infty} s^2(t+1)\ f_{XY}(s,t)\ dt\ ds \\
&= \int_0^1\int_0^1 s^2(t+1)\ (s+\frac{3}{2}t^2)\ dt\ ds \\
&= \int_0^1\int_0^1\ s^3t+\frac{3}{2}t^3s^2+s^3+\frac{3}{2}t^2s^2\ dt\ ds \\
&= \int_0^1 \frac{3}{2}s^3+\frac{7}{8}s^2\ ds \\
&= \frac{2}{3}
\end{aligned}$$

Notice how:

$$\frac{11}{12} = E[XY] \neq E[X]E[Y] = \left(\frac{7}{12}\right)\left(\frac{5}{8}\right) = \frac{35}{96}$$

This is because X and Y are not independent as we have shown earlier in the chapter. The result of:

$$E[g(X)h(Y)] = E[g(X)]E[h(Y)]$$

Only holds for all X and Y random variables if they are independent, one could think up an example to make it work for two dependent random variables however it is not generally true for all random variables X and Y.

We now move onto the expectations of conditional distributions.

The important results for expectations of conditional distributions are:

$$E[X|Y=y] = \int_{-\infty}^{\infty} s f_{X|Y}(s|y) ds$$
$$E[Y|X=x] = \int_{-\infty}^{\infty} t f_{Y|X}(t|x) dt$$
$$E[E[h(Y)|X]] = E[h(Y)]$$
$$E[g(X)h(Y)|X] = g(X)E[h(Y)|X]$$

Example

$$f_{XY}(x,y) = \begin{cases} x + \frac{3}{2}y^2 & 0 \le x \le 1, 0 \le y \le 1 \\ 0 & \text{otherwise} \end{cases}$$

Find:

1. $E[X|Y=b]$
2. $E[Y=b|X=0.3]$
3. $E[E[Y^2|X]]$
4. $E[X^2Y^2|X]$

We will use the results stated above this example and the conditional distributions of this pdf that we calculated previously to solve these:

1. $E[X|Y=b]$

$$\begin{aligned} E[X|Y=b] &= \int_{-\infty}^{\infty} s f_{X|Y}(s|y) ds \\ &= \int_0^1 \frac{s(3b^2+2s)}{3b^2+1}\, ds \\ &= \frac{1}{3b^2+1} \int_0^1 3b^2 s + 2s^2\, ds \\ &= \frac{3b^2}{6b^2+2} + \frac{2}{9b^2+3} \\ &= \frac{9b^2+4}{18b^2+6} \end{aligned}$$

2. $E[Y = b|X = 0.3]$

$$\begin{aligned} E[Y|X = 0.3] &= \int_{-\infty}^{\infty} t f_{Y|X}(t|x)dt \\ &= \int_0^1 \frac{t(3t^2 + 0.6)}{1.6}\ dt \\ &= \frac{5}{8}\int_0^1 3t^3 + 0.6t\ dt \\ &= \frac{21}{32} \end{aligned}$$

3. $E[E[Y^2|X]]$

$$\begin{aligned} E[E[Y^2|X]] &= E[Y^2] \\ &= \int_{-\infty}^{\infty} t^2 f_{XY}(x,t)dt \\ &= \int_0^1 t^2(x + \frac{3}{2}t^2)\ dt \\ &= \left[\frac{xt^3}{3} + \frac{3t^5}{10}\right]_0^1 \\ &= \frac{x}{3} + \frac{3}{10} \end{aligned}$$

4. $E[X^2Y^2|X]$

$$\begin{aligned} E[X^2Y^2|X] &= X^2E[Y^2|X] \\ &= x^2\int_{-\infty}^{\infty} t^2 f_{Y|X}(t|x)dt \\ &= x^2\int_0^1 \frac{t^2(3t^2 + 2x)}{2x+1}\ dt \\ &= \frac{x^2}{2x+1}\int_0^1 3t^4 + 2xt^2\ dt \\ &= \frac{x^2(10x+9)}{30x+15} \end{aligned}$$

Some more important results to do with conditonal distributions and expectation include:

$$\begin{aligned} \text{Var}[X|Y = y] &= E[X^2|Y = y] - E[X|Y = y]^2 \\ \text{Var}[Y|X = x] &= E[Y^2|X = x] - E[Y|X = x]^2 \end{aligned}$$

Also if X and Y are independent then:

$$E[X|Y = y] = E[X]$$
$$\text{Var}[X|Y = y] = \text{Var}[X]$$
$$E[Y|X = x] = E[Y]$$
$$\text{Var}[Y|X = x] = \text{Var}[Y]$$

Now we have finished this section and ultimately this chapter on bivariate distributions. In the next chapter we will look at transformations of random variables taking the discrete and continuous cases into account independently and now also splitting the chapter up into univariate transformations and bivariate transformations. Hope you enjoyed this chapter and hope you enjoy going through the rest of the book!

4 Transformations

In this chapter we consider transformations of random variables. Here we might be interested in a function of a random variable and could give real life examples of this use including the area of something in terms of its length or diameter etc. Using a transformation of a random variables could give us some useful knowledge about something else we might be interested in. We have already seen a transformation of a random variable in chapter 2 when we talked about standardisation with the normal distribution, the transformation in that case was:

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

We will deal with univariate transformations first and leave bivariate transformations for the next section.

4.1 Univariate Transformations

In the discrete case:

$$p_Y(y) = \sum_{x:g(X)=y} p_X(x)$$

This can be proved similarly too the proof of the two definitions of expected value being equal to each other.

Example

Suppose we have the pmf of $p_X(x) = \frac{1}{220}x^2$ for $x = 2, 4, 6, 8, 10$ and $p_X(x) = 0$ otherwise again and we wanted to find the pmfs of $Z = X^3$ and $T = X^2 + 3$

1. $Z = X^3$

   for X:

   $$p_X(2) = \frac{1}{55}\ p_X(4) = \frac{4}{55}\ p_X(6) = \frac{9}{55}\ p_X(8) = \frac{16}{55}\ p_X(10) = \frac{5}{11}$$

   Hence for $Z = X^3$ we have:

   $$p_X(8) = \frac{1}{55}\ p_X(64) = \frac{4}{55}\ p_X(216) = \frac{9}{55}\ p_X(512) = \frac{16}{55}\ p_X(1000) = \frac{5}{11}$$

2. $T = X^2 + 3$

   We already stated the pmf for X in the previous example and hence the pmf of $T = X^2 + 3$ is:

   $$p_X(7) = \frac{1}{55}\ p_X(19) = \frac{4}{55}\ p_X(39) = \frac{9}{55}\ p_X(67) = \frac{16}{55}\ p_X(103) = \frac{5}{11}$$

So as you can see it is quite simple to transform discrete random variables pmfs by just using the relationship between distributions to change the values taken up by the random variable and note the probabilities still stay the same. We can however not do this with continuous random variables since the probability that a continuous random variable is equal to a value is always 0 as we have said previously. We now move onto the distribution function method.

Distribution Function Method

The distribution function method is used to transform random variables using the cdf of the distribution and arises from:

$$F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(g(X) \leq y)$$

I think the best way to understand this method is to go through a couple of examples to show how it works.

Example

Let X be a continuous random variable with cdf of:

$$F_X(x) = x^3$$

for $0 < x < 1$. What is the cdf of $Y = X^2$ and the cdf of $Z = 3X^3$?

For $Y = X^2$:

$$\begin{aligned} F_Y(y) &= \mathbb{P}(Y \leq y) \\ &= \mathbb{P}(X^2 \leq y) \\ &= \mathbb{P}(X \leq y^{\frac{1}{2}}) \\ &= F_X(y^{\frac{1}{2}}) \\ &= y^{\frac{3}{2}} \end{aligned}$$

Clearly the range of y is $0 < y < 1$ as well from the relationship $Y = X^2$

For $Z = 3X^3$:

$$\begin{aligned} F_Z(z) &= \mathbb{P}(Z \leq z) \\ &= \mathbb{P}(3X^3 \leq z) \\ &= \mathbb{P}(X \leq \frac{1}{3} z^{\frac{1}{3}}) \\ &= F_X(\frac{1}{3} z^{\frac{1}{3}}) \\ &= \frac{1}{3} z \end{aligned}$$

The range here is $0 < z < 3$ from inputting the range of x into the relationship $Z = 3X^3$

Example

Let X be a continuous random variable with cdf of:

$$\begin{aligned} F_X(x) &= 0 \\ F_X(x) &= \frac{1}{2}x \\ F_X(x) &= 1 \end{aligned}$$

with ranges of $x \leq 0$, $0 < x \leq 2$, $x > 2$ respectively. Find the cdf and pdf of $T = 2X^2 + 1$

First the cdf:

$$\begin{aligned} F_T(t) &= \mathbb{P}(T \leq t) \\ &= \mathbb{P}(2X^2 + 1 \leq t) \\ &= \mathbb{P}\left(X \leq \sqrt{\frac{t-1}{2}}\right) \\ &= F_X\left(\sqrt{\frac{t-1}{2}}\right) \\ &= \frac{1}{2}\sqrt{\frac{t-1}{2}} \end{aligned}$$

The range here is $1 < t \leq 9$ from inputting the range of x into the relationship $T = 2X^2 + 1$. To find the pdf we differentiate using the chain rule and we get:

$$f_T(t) = \frac{1}{2^{\frac{5}{2}}\sqrt{t-1}}$$

For $1 < t \leq 9$ and 0 otherwise. We now move onto quite a cool thing in probability called the probability integral transform.

Probability Integral Transform

The probability integral transformation was first introduced by Ronald Fisher in 1932 in his book Statistical Methods for Research Workers. It is used to transform any continuous random variable to any other continuous random variable by repeated use of moving between Unif$(0, 1)$ distributed random variables and other random variables in both ways. This means we can transform any continuous random variable to the Unif$(0, 1)$ random variable and then transform that to any other continuous random variable. The transform works for all continuous random variables which shows how powerful and cool it can be however

it does not work for all discrete random variables. The transform is as follows:

Let Y be a continuous random variable with cdf $F(y)$ and inverse cdf F^{-1} and let U be a Uniform$(0,1)$ random variable. Then:

1. $F(Y)$ is a Uniform$(0,1)$ random variable
2. $F^{-1}(U)$ is a random variable with distribution function F

Proof.

Set $W = F(Y)$. Then for all $0 < w < 1$:

$$\begin{aligned}\mathbb{P}(W \leq w) &= \mathbb{P}(F(Y) \leq w) \\ &= \mathbb{P}(Y \leq F^{-1}(w)) \\ &= F(F^{-1}(w)) \\ &= w\end{aligned}$$

This proves the first part of the probability integral transform ($F(Y) \sim \text{Unif}(0,1)$)

Now for the second part set $V = F^{-1}(U)$. Then for all $-\infty < v < \infty$

$$\begin{aligned}\mathbb{P}(v \leq v) &= \mathbb{P}(F^{-1}(U) \leq v) \\ &= \mathbb{P}(U \leq F(v)) \\ &= F(v)\end{aligned}$$

since $0 \leq F(v) \leq 1$. So the cdf of V is F and the cdf of $F^{-1}(U)$ is F. This proves the second part.

Example

Use the probability integral transform to construct the transform $X \sim \text{Unif}(0,1)$ to $Y \sim \text{Exp}(\beta)$

By the PIT theorem if $X \sim \text{Unif}(0,1), F_Y(y)$ is the cdf of an Exp(β) random variable and $Y = F_Y^{-1}(X)$ then $Y \sim \text{Exp}(\beta)$

So we must find $F_Y^{-1}(x)$:

$$\begin{aligned}x &= F_Y(y) \\ &= 1 - e^{-\beta y}\end{aligned}$$

for $y > 0$, if and only if:

$$y = -\beta^{-1}\log(1-x) = F_Y^{-1}(x)$$

One-to-one Transformations

The following transformation only works for one-to-one transformations ($Y = g(X)$ so $X = g^{-1}(Y)$ exists)

If X has pdf $f_X(x)$ and $Y = g(X)$ defines a one-to-one transformation, then Y has pdf:

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

Where the modulus of $\frac{dy}{dx}$ is called the Jacobian of the transformation.

The proof of this contains some real analysis theory so i will omit this from this book for those who have not done it but you can find it with a quick google search if you are interested in it!

Example

Find the distribution of $Y = \frac{1}{X}$ when $X \sim$ Cauchy

Firstly $x = \frac{1}{y} = g^{-1}(y)$ and

$$\frac{dy}{dx} = \frac{-1}{x^2} = -y^2$$

$$\begin{aligned} f_Y(y) &= f_X(g^{-1}(y)) \left| \frac{dy}{dx} \right| \\ &= \frac{1}{\pi(1 + \frac{1}{y^2})} y^{-2} \\ &= \frac{1}{\pi(1 + y^2)} \end{aligned}$$

for $-\infty < y < \infty$. Hence $Y \sim$ Cauchy too.

A question you might ask is what transformation method should I use for each particular problem. Quite clearly when we don't have a one-to-one transformation then we have to use the distribution function method as the other does not work otherwise. Alternatively if we have a one-to-one transformation then both can be used and both will give the same answer. If we are given the cdf then we might want to use the distribution function method but if we are given the pdf and a straightforward Jacobian to work out then the other method may be the more preferred method.

4.2 Bivariate Transformations

We now deal with the bivariate case. There is a distribution function method for higher dimensions but is difficult to deal with in calculations so I will just focus on one-to-one transformations method.

One-to-one Transformations

Let X and Y be jointly continuous random variables with density function $f_{X,Y}$ and let g be a one-to-one transformation. Write $(U, V) = g(X, Y)$. The goal is to find the density (U, V). We can use a similar result as we did in the univariate case which is:

$$f_{U,V}(u, v) = f_{X,Y}(g^{-1}(u, v))|J(U, V)|$$

This is if g is a one-to-one linear transformation and $(U, V) = g(X, Y)$. This is also where $J(U, V)$ is:

$$J(U, V) = \det \begin{bmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{bmatrix}$$

where det means the determinant of the matrix.

Example

Let X and Y be independent standard normal random variables. Use the polar coordinates transformation to find $f_{R,\Theta}(r, \theta)$.

First for the joint pdf of X and Y we have:

$$f_{X,Y}(x, y) = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}} \frac{1}{\sqrt{2\pi}} e^{\frac{-y^2}{2}} = e^{\frac{-x^2-y^2}{2}}$$

Next for the polar coordinates transformation:

$$x = r\cos(\theta),\ y = r\sin(\theta)$$

The Jacobian is:

$$\begin{aligned} J(U, V) &= \det \begin{bmatrix} \frac{\partial x}{\partial r} & \frac{\partial x}{\partial \theta} \\ \frac{\partial y}{\partial r} & \frac{\partial y}{\partial \theta} \end{bmatrix} \\ &= \det \begin{bmatrix} \cos(\theta) & -r\sin(\theta) \\ \sin(\theta) & r\cos(\theta) \end{bmatrix} \\ &= r\cos^2(\theta) + r\sin^2(\theta) \\ &= r \end{aligned}$$

Hence:

$$f_{R,\Theta}(r, \theta) = \frac{1}{2\pi} r e^{\frac{-r^2}{2}}$$

We observe that R and Θ are independent and Θ is a uniform on $[0, 2\pi)$. The cdf of R is called the Rayleigh distribution $\left(F_R(r) = 1 - e^{\frac{-r^2}{2}}\right)$.

It is difficult to find when bivariate transformations in particular were first used but extensions of the techniques of Box and Cox in 1964 are proposed for obtaining data-based transformations of multivariate observations to enhance the normality of their distribution and to simplify the model. We now move onto some further probability methods.

5 Further Probability

In this chapter we will introduce covariance, correlation and the moment generating function. These are just some extra little bits in probability that can be quite useful as we will see. First we start with considering linear combinations of random variables and covariance.

5.1 Covariance and Correlation

We might be interested in linear combinations of random variables and their expectation and variance. If we consider the linear combination:

$$a_1X_1 + a_2X_2 + ... + a_nX_n = a^TX$$

We already know that the expectation of this is:

$$\begin{aligned} E[a_1X_1 + a_2X_2 + ... + a_nX_n] &= E[a_1X_1] + E[a_2X_2] + ...E[a_nX_n] \\ &= a_1E[X_1] + a_2E[X_2] + ...a_3E[X_n] \\ &= a^TE[X] \end{aligned}$$

This holds by the linearity of expectation but what happens if we want to find the variance of this linear combination of random variables. This is the motivation of this section of the chapter and we need to look at the covariance and correlation first before we see how to do this.

The covariance of X and Y is:

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

The covariance was first analysed by R. A. Fisher. Eden and Fisher in 1927 when they gave the decomposition of a sum of products. Sanders then became the first to analyse covariance for precision improvement in 1930 to increase the precision of a study and to remove a potential source of bias.

In probability theory the covariance is a measure of the joint variability of two random variables. It measures the total variation of two random variables from their expected values. The Covariance also has the following properties:

1. $\text{Cov}[aX, bY] = ab\text{Cov}[X, Y]$
2. $\text{Cov}[W + X, Y] = \text{Cov}[W, Y]+\text{Cov}[X, Y]$
3. $\text{Cov}[X, Y] = \text{Cov}[Y, X]$
4. $\text{Cov}[X, X] = \text{Var}[X]$

5. $\text{Cov}[X + a, Y + b] = \text{Cov}[X, Y]$

We are assuming that X, Y and W are random variables and a and b are constants. The first two properties together are called the bilinearity properties and the third property is the symmetry property. The first, third and fourth property can be proved straightaway by using the covariance formula and the linearity of expectation. The second and fifth property can be proved again by using the covariance formula and the linearity of expectation but require slightly more workings which we will now show:

1. $\text{Cov}[W + X, Y] = \text{Cov}[W, Y] + \text{Cov}[X, Y]$

   Proof

$$\begin{aligned}\text{Cov}[W + X, Y] &= E[(W + X)Y] - E[W + X]E[Y] \\ &= E[WY] + E[XY] - E[W]E[Y] - E[X]E[Y] \\ &= E[WY] - E[W]E[Y] + E[XY] - E[X]E[Y] \\ &= \text{Cov}[W, Y] + \text{Cov}[X, Y]\end{aligned}$$

2. $\text{Cov}[X + a, Y + b] = \text{Cov}[X, Y]$

   Proof

$$\begin{aligned}\text{Cov}[X + a, Y + b] &= E[(X + a)(Y + b)] - E[X + a]E[Y + b] \\ &= E[XY + bX + aY + ab] - (E[X] + a)(E[Y] + b) \\ &= E[XY] + bE[X] + aE[Y] + ab \\ &\quad - E[X]E[Y] - aE[X] - bE[Y] - ab \\ &= E[XY] - E[X]E[Y] \\ &= \text{Cov}[X, Y]\end{aligned}$$

Correlation was first spotted by Francis Galton when he recognised a common thread between three different scientific problems he was studying in 1888. Correlation is any statistical relationship between two random variables, more broadly it could be used for any association however in statistics it usually refers to the degree to which a pair of variables are linearly related. The correlation of X and Y is:

$$\rho = \text{Corr}[X, Y] = \frac{\text{Cov}[X, Y]}{\sqrt{\text{Var}[X]\text{Var}[Y]}}$$

The correlation has the following properties:

1. $\text{Corr}[aX + b, cY + d] = \text{sign}(ac)\text{Corr}[X, Y]$
2. $-1 \leq \rho \leq 1$

The first property is that correlation is invariant to location and scale changes. Here obviously a and c are constants and X and Y are random variables. The second property we will now prove:

Proof

$$\begin{aligned}
\text{Cov}[X,Y] &\leq \sqrt{\text{Var}(X)\text{Var}(Y)} \\
-\sqrt{\text{Var}(X)\text{Var}(Y)} \leq \text{Cov}[X,Y] &\leq \sqrt{\text{Var}(X)\text{Var}(Y)} \\
-1 \leq \frac{\text{Cov}[X,Y]}{\sqrt{\text{Var}(X)\text{Var}(Y)}} &\leq 1 \\
-1 \leq \rho &\leq 1
\end{aligned}$$

The first line is due to the Cauchy-Scharwz inequality and the rest follows on nicely.

The correlation is 1 when we have perfect positive linear association, -1 when we have perfect negative linear association and 0 when we have no relationship (no linear association) between variables. When a variable tends to increase when the other variable is increasing the correlation and covariance will be positive, when a variable tends to decrease when the other variable is increasing the correlation and covariance will be negative. The stronger the association between two variables the larger the correlation and covariance will be until it gets to perfect positive linear association and the weaker the association between two variables the smaller the correlation and covariance will be until it gets to perfect negative linear association.

When X and Y are independent then as we have seen from chapter 3:

$$E[g(X)h(Y)] = E[g(X)]E[h(Y)]$$

But more importantly:

$$E[XY] = E[X]E[Y]$$

Which consequently has a big impact on the correlation and covariance since:

$$\begin{aligned}
\text{Cov}[X,Y] &= E[XY] - E[X]E[Y] \\
&= E[X]E[Y] - E[X]E[Y] \\
&= 0
\end{aligned}$$

Which also means that the correlation is also 0. Therefore when X and Y are independent then the covariance and correlation are both 0 which when we think about the interpretation of the covariance and correlation it makes a lot of sense!

Now that we have gone through the covariance and correlation we can now

go onto the motivation of this section which is how to find the variance of linear combinations of random variables. For two random variables X and Y and constants a and b:

$$\begin{aligned}
\text{Var}[aX + bY] &= E[(aX + bY)^2] - E[aX + bY]^2 \\
&= E[a^2X^2 + 2abXY + b^2Y^2] - (aE[X] + bE[Y])^2 \\
&= a^2E[X^2] + 2abE[XY] + b^2E[Y^2] \\
&- (a^2E[X]^2 + 2abE[X]E[Y] + b^2E[Y]^2) \\
&= a^2E[X^2] - a^2E[X]^2 + b^2E[Y^2] - b^2E[Y]^2 \\
&+ 2ab(E[XY] - E[X]E[Y]) \\
&= a^2\text{Var}[X] + b^2\text{Var}[Y] + 2ab\text{Cov}[X, Y]
\end{aligned}$$

We can extend this result for a linear combination of n random variables:

$$\begin{aligned}
\text{Var}[a_1X_1 + ... + a_nX_n] &= \text{Var}\left(\sum_{i=1}^{n} a_iX_i\right) \\
&= \sum_{i=1}^{n} a_i^2\text{Var}[X_i] + 2\sum_{i=1}^{n}\sum_{j:j>i}^{n} a_ia_j\text{Cov}[X_i, X_j]
\end{aligned}$$

Example

Suppose $\text{Var}[X] = 1$ find an upper bound and lower bound for the covariance of random variables X and Y:

$$\begin{aligned}
-1 \leq \rho = \frac{\text{Cov}[X, Y]}{\sqrt{\text{Var}[X]\text{Var}[Y]}} \leq 1 \\
-1 \leq \frac{\text{Cov}[X, Y]}{\sqrt{\text{Var}[Y]}} \leq 1 \\
-\sqrt{\text{Var}[Y]} \leq \text{Cov}[X, Y] \leq \sqrt{\text{Var}[Y]}
\end{aligned}$$

We constructed a bound on the covariance due to the correlation having to be between -1 and 1. Now find the covariance when we have perfect positive linear association:

$$\text{Cov}[X, Y] = \sqrt{\text{Var}[Y]}$$

Since when we have perfect positive linear association $\rho = 1$ and so we have the covariance equal to our upper bound.

Example

Suppose we have two random variables X and Y with $E[X] = 4, E[Y] = 3$

and $E[XY] = 20$. Calculate the covariance of X and Y:

$$\begin{aligned}\text{Cov}[X, Y] &= E[XY] - E[X]E[Y] \\ &= 20 - (4)(3) \\ &= 12\end{aligned}$$

Now calculate the value of $\text{Var}[4X + 5Y]$ where $\text{Var}[X] = 2$ and $\text{Var}[Y] = 1$

$$\begin{aligned}\text{Var}[4X + 5Y] &= 16\text{Var}[X] + 25\text{Var}[Y] + 2(4)(3)\text{Cov}[X, Y] \\ &= 16(2) + 25(1) + 2(4)(3)(12) \\ &= 345\end{aligned}$$

Suppose we have a third random variable Z. Calculate $\text{Var}[4X + 5Y + aZ]$

$$\begin{aligned}\text{Var}[4X + 5Y + aZ] &= 16\text{Var}[X] + 25\text{Var}[Y] + a^2\text{Var}[Z] \\ &\quad + 2(4)(3)\text{Cov}[X, Y] + 16a^2\text{Cov}[X, Z] + 25a^2\text{Cov}[Y, Z] \\ &= 345 + a^2\text{Var}[Z] + 16a^2\text{Cov}[X, Z] + 25a^2\text{Cov}[Y, Z]\end{aligned}$$

Now suppose that Z is independent of both X and Y. Calculate $\text{Var}[4X+5Y+aZ]$

$$\begin{aligned}\text{Var}[4X + 5Y + aZ] &= 345 + a^2\text{Var}[Z] + 0 + 0 \\ &= 345 + a^2\text{Var}[Z]\end{aligned}$$

This is because when Z is independent of X and independent of Y then the covariance between the two is 0. A cool thing I started questioning when i was writing this book was what happens if you increase the amount of random variables you have and go about calculating the variance of linear combinations of these random variables. I thought about maybe there was a relationship between the amount of random variables you have and the amount of terms that would need to add together to calculate the variance of a linear combination of these random variables. I started to look into it and obviously for one random variable you would just be calculating the variance of that one random variable and would have 1 term to calculate. For two random variables you would need to calculate 3 terms (The variances of both random variables and the covariance of them both). For 3 random variables we have just seen that 6 terms need to be added together. If we keep increasing this we start to see a pattern as we think about the different combinations of covariances added to variances for higher and higher amounts of random variables. That pattern is that for n random variables the amount of terms we would need to add together in order to find the variance of a linear combination of these random variables is the sum of the first n positive integers, that is:

$$\frac{n(n+1)}{2}$$

So entering the amount of random variables we have into this gives us the amount of terms we would need to add together to calculate the variance of a

linear combination of these. As you can see this quickly increases the amount of terms so finding the variance of a linear combination of a large amount of random variables can become quite tedious and obviously as n approaches infinity this also tends to infinity.

We have now finished this section and move onto the moment generating function!

5.2 Moment Generating Functions

Moment generating functions were first introduced by Abraham De Moivre in 1730 in order to solve the general linear recurrence problem. The moment generating function (mgf) of a discrete random variable X is:

$$M_X(t) = E[e^{tX}] = \sum_i e^{ti} p_X(i)$$

If X is discrete rv with pmf $p_X(x)$ and $\forall$ real values of t for which the expectation exists.

Example

Find the moment generating function of the poisson distribution.

$$\begin{aligned} M_X(t) = E[e^{tX}] &= \sum_{i=0}^{\infty} e^{ti} p_X(i) \\ &= \sum_{i=0}^{\infty} e^{ti} \frac{\lambda^i e^{-\lambda}}{i!} \\ &= e^{-\lambda} \sum_{i=0}^{\infty} \frac{(\lambda e^t)^i}{i!} \\ &= e^{-\lambda} e^{\lambda e^t} \\ &= e^{\lambda(e^t - 1)} \end{aligned}$$

We used that $\sum_{i=0}^{\infty} \frac{a^i}{i!} = e^a$ to solve this.

The moment generating function (mgf) of a continuous random variable X is:

$$M_X(t) = E[e^{tX}] = \int_s e^{ts} f_X(s) ds$$

If X is continuous rv with pdf $f_X(x)$ and $\forall$ real values of t for which the expectation exists.

Example

Find the moment generating function of the Gamma distribution. Assume that t $< \beta$.

$$\begin{aligned}
M_X(t) = E[e^{tX}] &= \int_s e^{ts} f_X(s) ds \\
&= \frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^\infty x^{\alpha-1} e^{-(\beta-t)x} \, dx \\
&= \frac{\beta^\alpha}{\Gamma(\alpha)} \frac{\Gamma(\alpha)}{(\beta-t)^\alpha} \times \frac{(\beta-t)^\alpha}{\Gamma(\alpha)} \int_0^\infty x^{\alpha-1} e^{-(\beta-t)x} \, dx \\
&= \frac{\beta^\alpha}{\Gamma(\alpha)} \frac{\Gamma(\alpha)}{(\beta-t)^\alpha} \times 1 \\
&= \frac{\beta^\alpha}{(\beta-t)^\alpha} \\
&= \left(\frac{\beta}{\beta-t}\right)^\alpha
\end{aligned}$$

We used that the pdf over it's range integrates to 1 (The distribution was Gamma$(\alpha, \beta - t)$). We could also go about finding this mgf when $\beta = t$ and when $\beta < t$. When $\beta = t$ it can be shown that the mgf diverges and the integral tends to ∞ so the expectation does not exist. When $\beta < t$ it can be shown that the expectation does not exist as well. Try to work both cases out and show this for yourself using the same method from above! (The case when $\beta < t$ can be tricky to spot however as it does use some real analysis theory to get there).

If the mgf is defined in some neighbourhood of the origin, $|t| < t_0$, the following properties are satisfied:

1. If two random variables have the same mgf's then they have the same cdf
2. If $Z = a + bX$ then $M_Z(t) = e^{at} M_X(t)$ for a and b non zero real numbers
3. For a random variable X:

$$\begin{aligned}
M_X(0) &= E[X^0] = 1 \\
M_X^{'}(0) &= E[X] \\
M_X^{''}(0) &= E[X^2]
\end{aligned}$$

   etc. with the order of the differential of the moment generating function corresponding to the power of the random variable in the expectation.

4. Let X, Y be independent random variables with mgf's of $M_X(t)$ and $M_Y(t)$ then:

$$M_{X+Y}(t) = M_X(t) M_Y(t)$$

Another important result for $X_1, X_2, ..., X_n$ independent random variables is:

$$M_{X_1+X_2+....+X_n}(t) = M_{X_1}(t)M_{X_2}(t)....M_{X_n}(t)$$

Example

Find the variance of a gamma using the mgf:

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

We put the mgf into a better form to differentiate. We now find the expectations required.

$$\begin{aligned}
M_X^{'}(t) &= (-\alpha)\left(-\frac{1}{\beta}\right)\left(1 - \frac{t}{\beta}\right)^{-\alpha-1} \\
M_X^{'}(0) &= \frac{\alpha}{\beta} = E[X] \\
M_X^{''}(t) &= \left(\frac{\alpha}{\beta}\right)(-\alpha - 1)\left(-\frac{1}{\beta}\right)\left(1 - \frac{t}{\beta}\right)^{-\alpha-2} \\
M_X^{''}(0) &= \frac{\alpha(\alpha+1)}{\beta^2} = E[X^2]
\end{aligned}$$

Hence:

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

We can also use mgf's to prove that sums of random variables are a certain distribution. For example we can prove that the sum of Bernoulli distributions is binomial which we have already seen in chapter 3. We can prove this for a lot of distributions that summing them up will give a certain distribution if the mgf's are the same. We will only do it for the Bernoulli/binomial case though (See if you can find more cases and prove it using the mgf's):

We can take independent and identically distributed Bernoulli random variables $X_1, ..., X_n$. We will first find the mgf of the one of these random variables and then use the result that the mgf of the sum of these random variables is the product of the mgf's of each random variable. When we find the mgf of one of these random variables though we have the mgf of all of them since all of them are Bernoulli.

$$\begin{aligned}
M_{X_1}(t) &= \mathbb{P}(X_1 = 0)e^0 + \mathbb{P}(X_1 = 1)e^t \\
&= q + pe^t
\end{aligned}$$

Hence:

$$\begin{aligned} M_{X_1+X_2+....+X_n}(t) &= M_{X_1}(t)M_{X_2}(t)....M_{X_n}(t) \\ &= (q+pe^t)(q+pe^t)(q+pe^t).... \\ &= (q+pe^t)^n \end{aligned}$$

Now we will find the mgf of a binomial random variable and compare the two.

$$\begin{aligned} M_X(t) &= \sum_{i=0}^{n} e^{it} \frac{n!}{i!(n-i)!} p^i q^{n-i} \\ &= \sum_{i=0}^{n} (pe^t)^i \frac{n!}{i!(n-i)!} q^{n-i} \\ &= (q+pe^t)^n \end{aligned}$$

We used the fact that the second line is the expansion of binomial. Since the mgf of the sum of the bernoullis is the same as the mgf of the binomial they are the same distribution which proves what we needed too. This shows how important mgf's applications can be and why they are used throughout probability theory. For me mgf's are one of the coolest topics in undergraduate probability especially when they are used to prove what we have done above, i find there applications so fascinating from the "simple" results that come with them. This is also true for characteristic functions which we will touch on in chapter 7 on the extra topics i decided to add to the end of this book. Before that though we will go through some important limit theorems.

6 Limit Theorems

In this chapter we discuss different types of convergence involved in probability distributions and introduce two of the most important results in probability theory, the weak law of large numbers and the central limit theorem. First we will start with some definitions of convergence to build up to the two more important results later on.

6.1 Convergence

Convergence in probability

We say that a sequence of random variables, $X_1, ..., X_n$ converges in probability to a random variable X if:

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

The most famous example of convergence in probability is the weak law of large numbers which like we said previously we will discuss in a later section.

Convergence in distribution

A sequence of random variables $X_1, ..., X_n$ converges in distribution to a random variable X if:

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$$

for all x at which $F_X(x)$ is continuous.

Convergence in mean

A sequence of random variables $X_1, ..., X_n$ converges to a random variable X in mean if:

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

Provided that the sequence of random variables each have finite expectation.

Convergence in mean square

A sequence of random variables $X_1, ..., X_n$ converges to a random variable X in mean square if:

$$\lim_{n\to\infty} E(|X_n - X|^2) = 0$$

Provided that the random variables are square integrable.

There are more types of convergence in probability theory like the almost sure convergence along with others however we won't cover those in this book.

Some results that can be useful from the relationships between these types of convergence are:

1. convergence in mean square implies convergence in mean
2. convergence in mean implies convergence in probability
3. convergence in probability implies convergence in distribution

Average of first n variables

Given a sequence of independent and identically distributed random variables $X_1, ..., X_n$ the average of the first n variables is:

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

provided that $\mu < \infty$ and $\sigma^2 < \infty$ We will go about finding the expectation and variance of this random variable and lastly show what distribution it is.

Finding the expectation:

$$\begin{aligned} E\left[\overline{X}_n\right] &= E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] \\ &= \frac{1}{n}\sum_{i=1}^{n} E[X_i] \\ &= \frac{1}{n} n\mu \\ &= \mu \end{aligned}$$

Finding the variance:

$$\begin{aligned} \text{Var}\left[\overline{X}_n\right] &= \text{Var}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] \\ &= \frac{1}{n^2}\sum_{i=1}^{n} \text{Var}[X_i] \\ &= \frac{1}{n^2} n\sigma \\ &= \frac{\sigma^2}{n} \end{aligned}$$

This holds because the X_i are independent.

It can be shown that $\overline{X}_n$ has a normal distribution if each X_i are independent and identically distributed and are normally distributed. Because of our assumption of them being independent normal distributions then $X_1 + X_2$ is also normally distributed. By induction the sum of all of the random variables are then normally distributed, i.e.

$$S_n = \sum_{i=0}^{n} X_i$$

This is normally distributed and hence $\overline{X}_n$ is also normally distributed. Since we know what the expectation and variance of this is from previous calculations we can now say:

$$\overline{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

Now consider:

$$\overline{X}_n - \mu \sim N\left(0, \frac{\sigma^2}{n}\right)$$

This converges in distribution to $N(0, 0)$. When we go about rescaling this by dividing it by its standard deviation we actually get the standard normal distribution. Which makes sense when we think back to standardisation from chapter 3 since we did this exact thing but with variance σ^2 instead. We end up with:

$$\frac{\sqrt{n}\left(\overline{X}_n - \mu\right)}{\sigma} \sim N(0, 1)$$

and so:

$$\mathbb{P}\left(\frac{\sqrt{n}\left(\overline{X}_n - \mu\right)}{\sigma} \leq z\right) = \Phi(z)$$

This discussion leads us onto the central limit theorem but we will discuss further about it in that section but before that we cover the weak law of large numbers along with a few other results.

6.2 The Weak Law of Large Numbers

Markov's inequality

If V is a non-negative random variable then for any $a > 0$:

$$\mathbb{P}(V \geq a) \leq \frac{E[V]}{a}$$

This holds for both discrete and continuous random variables. We will prove this in the continuous case although the discrete case is proved in a very similar way.

Proof:

Let V have pdf $f(v)$, then for any $a > 0$:

$$\begin{aligned} E[V] &= \int_0^\infty t f(t) dt \\ &\geq \int_a^\infty t f(t) dt \\ &\geq \int_a^\infty a f(t) dt \\ &= a \int_a^\infty f(t) dt \\ &= a\mathbb{P}(V \geq a) \end{aligned}$$

Hence:

$$\mathbb{P}(V \geq a) \leq \frac{E[V]}{a}$$

Markov's inequality gives an upper bound on the probability that a random variable is greater than or equal to some positive constant. The inequality came up in Russian mathematician Pafnuty Chebyshev's work but was named after Andrey Markov who was one of Chebyshev's students.

Chebyshev's inequality

If Y is a random variable with expectation μ and variance $\sigma^2 < \infty$ then for any $\epsilon > 0$:

$$\mathbb{P}(|Y - \mu| \geq \epsilon) \leq \frac{\sigma^2}{\epsilon^2}$$

Proof:

We use Markov's inequality to prove this. For a random variable Y with $E[Y] = \mu$ and $\text{Var}[Y] = \sigma^2$, set $V = (Y - \mu)^2$ and $a = \epsilon^2$. Using Markov's inequality we now get:

$$\mathbb{P}((Y - \mu)^2 \geq \epsilon^2) \leq \frac{\sigma^2}{\epsilon^2}$$

and hence:

$$\mathbb{P}(|Y - \mu| \geq \epsilon) \leq \frac{\sigma^2}{\epsilon^2}$$

Chebyshev's inequality provides an upper bound on the probability of deviation of a random variable from its mean. The inequality was first used by Irénée-Jules Bienaymé in 1853 but was later proved by Pafnuty Chebyshev in 1867 and consequently was named after him.

The Weak Law of Large Numbers

Suppose $X_1, X_2, ...$ is a sequence of independent and identically distributed random variables with expectation μ and finite variance σ^2 and $\epsilon > 0$:

$$\lim_{n\to\infty} \mathbb{P}(|\overline{X}_n - \mu| < \epsilon) = 1$$

as $n \to \infty$

Effectively this law says that if you have a large enough sample size, there's a very good chance that the average of your observations will be quite close to what you expect it to be, as long as you're willing to accept a small difference between the observed average and the expected value. The law was introduced first in 1713 by Jacob Bernoulli although it's form has changed multiple times since.

The Strong Law of Large Numbers

Suppose $X_1, X_2, ...$ is a sequence of independent and identically distributed random variables with expectation μ and finite variance σ^2 and $\epsilon > 0$:

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

as $n \to \infty$

This law means that the probability that, as the number of trials n goes to infinity, the average of the observations converges to the expected value, is equal to one. The strong law of large numbers was proved by Kolmogorov in 1930. We now move onto the motivation of this chapter which is the central limit theorem which is one of the most crucial results in all of probability.

6.3 The Central Limit Theorem

The theory of characteristic functions (Theory we will go through in the next chapter) was developed in order to prove the central limit theorem with the theorem being introduced in 1733 by Abraham de Moivre. In probability theory, the central limit theorem states that the distribution of a normalised version of the sample mean converges to a standard normal distribution. The theorem is so powerful because it implies that probabilistic and statistical methods that work for normal distributions can be applicable to many problems involving other types of distributions. It also has a small amount of assumptions in that it does not even assume that the random variables are discrete or continuous as

both types work.

The Central Limit Theorem

Suppose $X_1, X_2, ...$ is a sequence of independent and identically distributed random variables with expectation μ and finite variance σ^2, then for any number $-\infty < x < \infty$

$$\mathbb{P}\left(\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \leq x\right) \to \Phi(x)$$

as n $\to \infty$. $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $\Phi(x)$ is the cdf for the standard Normal distribution evaluated at x.

Example

Suppose a restaurant claims that their burgers have a mean weight of 200 grams with a standard deviation of 10 grams. A health inspector randomly selects a sample of 36 burgers from the restaurant and finds that the mean weight of this sample is 195 grams.

We want to determine the probability of observing a sample mean as low as 195 grams, or even lower, if the restaurant's claim about the mean weight is true.

Let $X_1, ..., X_{36}$ be the weights of the burgers in grams. By the central limit theorem $\overline{X}_{36} = \sum_{i=1}^{36} X_i$ has distribution of:

$$N\left(200, \frac{10^2}{36}\right)$$

approximately. Hence:

$$\begin{aligned}\mathbb{P}(\overline{X}_{36} \leq 195) &= \mathbb{P}\left(\frac{\overline{X}_{36} - 200}{\sqrt{\frac{10^2}{36}}} \leq \frac{195 - 200}{\sqrt{\frac{10^2}{36}}}\right) \\ &= \Phi(-3) \\ &= 0.0013 \text{ (To 4 d.p.)}\end{aligned}$$

We could use some statistical software to show $\Phi(-3) = 0.0013$ like r or we could use a calculator or look it up in a normal distribution table. Hence there is only a 0.13% chance of observing such a low average weight of burger for 36 randomly selected burgers.

We have now finished the main theory in this book and what i was originally going to stop at but i decided to add some more cool stuff in another extra chapter. I hope you enjoy it! Some of the things i find fascinating about probability are in the next chapter and i hope you do too!

7 Some more cool stuff

Throughout this chapter i will touch on some extra cool things in probability through examples, results and then a small section on characteristic functions at the end.

Example

In this first example we will show how if we let X be the Chebyshev random variable with pdf down below. Then $Y = \sin^{-1}(X)$ is uniformly distributed on $[\frac{-\pi}{2}, \frac{\pi}{2}]$.

$$p(x) = \begin{cases} \frac{1}{\pi}\frac{1}{\sqrt{1-x^2}} & -1 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

We do this by using the distribution function method talked about in chapter 4 (Univariate transformations).

$$\begin{aligned} F_Y(y) &= \mathbb{P}(Y < y) \\ &= \mathbb{P}(\sin^{-1}(X) < y) \\ &= \mathbb{P}(X < \sin(y)) \\ &= \int_{-\infty}^{\sin(y)} \frac{1}{\pi}\frac{dx}{\sqrt{1-x^2}} \\ &= \left[\frac{1}{\pi}\sin^{-1}(x)\right]_{-1}^{\sin(y)} \\ &= \frac{1}{\pi}\left(y + \frac{\pi}{2}\right) \end{aligned}$$

Now if we differentiate with respect to y we get the pdf:

$$p(y) = \begin{cases} \frac{1}{\pi} & \frac{-\pi}{2} \leq y \leq \frac{\pi}{2} \\ 0 & \text{otherwise} \end{cases}$$

This is the pdf of a uniform random variable on $[\frac{-\pi}{2}, \frac{\pi}{2}]$. The distribution function method can be used in all sorts of cool ways like this to show transformations of some random variables are standard random variables. A classic one is that the square of a standard normal distribution is a $\chi^2(1)$ distribution, one would just have to go through the steps of the distribution function method similar to how we have done the above example. Try this for yourself (Hint: use $Y = X^2$ with X as a standard normal random variable).

One thing i always questioned about probability throughout my study in college and my first two years of university was about expectation of random variables.

I always would think that we can find the expectation of a random variable using the pdf however if we have the cdf of the distribution is there no way we can directly find the expectation without differentiating and using the pdf. The result in terms of the cdf to find expectation was finally introduced to me in my third year of university and after all that time i was thinking why and how did i not know this when it was such a simple result in practice and definitely easier to find the expectation given you have the cdf already. This is that result:

$$E(X) = \int_0^\infty (1 - F_X(x))dx$$

This holds provided that X is a non-negative random variable and the Riemann improper integral converges.

Proof.

Since $1 - F_X(x) = \mathbb{P}(X \geq x) = \int_x^\infty f_X(t)dt$,

$$\begin{aligned}
\int_0^\infty (1 - F_X(x))dx &= \int_0^\infty \mathbb{P}(X \geq x)dx \\
&= \int_0^\infty \int_x^\infty f_X(t)dt\ dx \\
&= \int_0^\infty \int_0^t f_X(t)dx\ dt \\
&= \int_0^\infty [xf_X(t)]_0^t\, dt \\
&= \int_0^\infty tf_X(t)dt \\
&= \int_0^\infty xf_X(x)dx \\
&= E(X)
\end{aligned}$$

The second to last line follows by a simple substitution $t = x$ and $dt = dx$.

Now we move onto a couple of results that i came across and thought they were rather cool. The first one we will prove by induction and it is known as Boole's inequality.

Boole's inequality

Boole's inequality was discovered by the English mathematician George Boole in the 1800s. In measure theory it comes from the fact that a measure is a σ-sub-additive. This is the inequality:

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

Proof. (By induction)

When $n = 1$:

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

$$\mathbb{P}(A_i) \leq \mathbb{P}(A_i)$$

Hence it holds true for $n = 1$.

We assume it holds for $n = k$:

$$\mathbb{P}\left(\bigcup_{i=1}^{k} A_i\right) \leq \sum_{i=1}^{k} \mathbb{P}(A_i)$$

Now we show it holds for $n = k + 1$:

$$\begin{aligned}
\mathbb{P}\left(\bigcup_{i=1}^{k+1} A_i\right) &= \mathbb{P}\left(\left(\bigcup_{i=1}^{k} A_k\right) \bigcup A_{k+1}\right) \\
&= \mathbb{P}\left(\bigcup_{i=1}^{k} A_k\right) + \mathbb{P}(A_{k+1}) - \mathbb{P}\left(\left(\bigcup_{i=1}^{k} A_k\right) \bigcap A_{k+1}\right) \\
&\leq \mathbb{P}\left(\bigcup_{i=1}^{k} A_k\right) + \mathbb{P}(A_{k+1}) \\
&\leq \sum_{i=1}^{k} \mathbb{P}(A_i) + \mathbb{P}(A_{k+1}) = \sum_{i=1}^{k+1} \mathbb{P}(A_i)
\end{aligned}$$

Hence the inequality holds for $n = k + 1$ and hence holds $\forall k \geq 1$ by induction. We used the addition law (in chapter 1) to get to the second line and then the first axiom of probability to get to the third line. We then used our assumption from $n = k$ to get to the last line.

The second result i came across is far more simpler to prove, this is the result:

$$\mathbb{P}\left(\bigcap_{j=1}^{n} A_j\right) \geq 1 - \sum_{j=1}^{n} \mathbb{P}(\Omega \setminus A_j)$$

Proof.

$$\begin{aligned}\mathbb{P}\left(\bigcap_{j=1}^{n} A_j\right) &= 1 - \mathbb{P}\left(\bigcup_{j=1}^{n} {A_j}^c\right) \\ &\geq 1 - \left(\sum_{j=1}^{n} \mathbb{P}({A_j}^c)\right) \\ &= 1 - \sum_{j=1}^{n} \mathbb{P}(\Omega \setminus A_j)\end{aligned}$$

This next example is a cool way to see how the Poisson distribution acts.

Example

Let X be a Poisson random variable that satisfies $\mathbb{P}(X = j) = \frac{\theta^j}{j!}e^{-\theta}$ for each integer $j \geq 0$. Calculate $E[X(X-1)...(X-k)]$:

$$\begin{aligned}E[X(X-1)...(X-k)] &= \sum_{n=0}^{\infty} n(n-1)...(n-k)\mathbb{P}(X=n) \\ &= \sum_{n=0}^{\infty} e^{-\theta}\frac{n(n-1)...(n-k)\theta^n}{n!} \\ &= \theta^{k+1}e^{-\theta}\sum_{n=k+1}^{\infty}\frac{\theta^{n-k-1}}{(n-k-1)!} \\ &= \theta^{k+1}e^{-\theta}\sum_{n=k+1}^{\infty}\frac{\theta^j}{j!} \\ &= \theta^{k+1}e^{-\theta}e^{\theta} \\ &= \theta^{k+1}\end{aligned}$$

This is quite cool as we have this simple relationship between the expectation and θ. Also when $k = 0$ we can see that this gives just the expectation of the Poisson distribution:

$$E[X] = \theta^{0+1} = \theta$$

We can also see if we use a combination of $k = 0$ and $k = 1$ we can get the variance of the Poisson distribution:

$$E[X(X-1)] = \theta^{1+1} = \theta^2 = E[X^2] - E[X]$$

To get $\text{Var}[X] = E[X^2] - (E[X])^2$ we need to add $E[X]$ and takeaway $(E[X])^2$. Hence:

$$\begin{aligned} E[X(X-1)] + E[X] - (E[X])^2 &= \theta^{1+1} + \theta^{0+1} - (\theta^{0+1})^2 \\ &= \theta^2 + \theta - \theta^2 \\ &= \theta \end{aligned}$$

I want to add this theorem in next because obviously i love the name of the theorem however the concept is actually quite fascinating and an example of when things are strange when dealing with infinity.

Infinite Monkey Theorem

The infinite monkey theorem states that if you have an infinite number of monkeys each hitting keys at random on typewriter keyboards, then the probability that one of them will type the complete works of William Shakespeare is 1.

This shows how weird things happen with infinity and how it can affect probability in cool ways. This can be proved by using what is called the second Borel-Cantelli lemma, see if you can find out what this is and find the proof yourself (Search the infinite monkey theorem lots of results come up for it).

Coupling Lemma

Let X and Y be random variables, then:

$$|\mathbb{P}(X \leq \lambda) - \mathbb{P}(Y \leq \lambda)| \leq \mathbb{P}(X \neq Y)$$

for any $\lambda \in \mathbb{R}$

Proof. For any E and F, we have:

$$|\mathbb{P}(E) - \mathbb{P}(F)| \leq \mathbb{P}(E \triangle F)$$

Where $\triangle$ is the symmetric difference. We apply this to the events $E = (X \leq \lambda)$ and $F = (Y \leq \lambda)$ with complements $E^c = (X > \lambda)$ and $F^c = (Y > \lambda)$ and we see that:

$$\begin{aligned} E \triangle F &= ((X \leq \lambda \text{ and } Y > \lambda) \text{ or } (X > \lambda \text{ and } Y \leq \lambda)) \\ &\subset (X \neq Y) \end{aligned}$$

Hence we have the coupling lemma:

$$|\mathbb{P}(X \leq \lambda) - \mathbb{P}(Y \leq \lambda)| \leq \mathbb{P}(X \neq Y)$$

This next example uses the convergence theory we went through in chapter 6 (convergence in probability and distribution).

Example

Let U be uniformly distributed on [0,1] and let U_j be mutually independent copies of U.

1. Find the cdf of U

$$F_U(t) = \begin{cases} 1, & t \geq 0 \\ t & 0 \leq t \leq 1 \\ 0 & t \leq 0 \end{cases}$$

2. Show that the cdf of the random variable $Z_n = \max\{U_1, U_2, ..., U_n\}$ satisfies $F_{Z_n} = F_U^n$ and hence find F_{Z_n}

$$\begin{aligned} F_{Z_n} &= \mathbb{P}(\max\{U_j : 1 \leq j \leq n\} \leq t) \\ &= \prod_{j=1}^{n} \mathbb{P}(U_j \leq t) \\ &= (\mathbb{P}(U \leq t))^n \\ &= F_U^n(t) \end{aligned}$$

This is by the independence of the $U_j's$ and because the $U_j's$ have the same distribution as U. Hence:

$$F_{Z_n}(t) = \begin{cases} 1, & t \geq 0 \\ t^n & 0 \leq t \leq 1 \\ 0 & t \leq 0 \end{cases}$$

3. Show that $Z_n \to 1$ in probability as $n \to \infty$

For $\epsilon > 0$ we have:

$$\begin{aligned} \mathbb{P}(|Z_n - 1| \leq \epsilon) &= \mathbb{P}(Z_n - 1 \geq \epsilon) + \mathbb{P}(Z_n \leq 1 - \epsilon) \\ &= 0 + F_{Z_n}(1 - \epsilon) \\ &= (1 - \epsilon)^n \\ &\to 0 \end{aligned}$$

as $n \to \infty$

4. Show that

$$\mathbb{P}(n(1 - Z_n) \leq t) \to 1 - e^{-t}$$

This is for $t > 0$ and 0 otherwise as $n \to \infty$.

$$\begin{aligned}\mathbb{P}(n(1-Z_n) \leq x) &= \mathbb{P}(Z_n \geq (1-x/n)) \\ &= 1 - \mathbb{P}(Z_n \leq (1-x/n)) \\ &= 1 - F_{Z_n}(1-x/n) \\ &= 1 - (1-x/n)^n \\ &\to 1 - e^{-x}\end{aligned}$$

as $n \to \infty$.

The next example again uses some theory from chapter 6, which is the central limit theorem this time.

Example

Let $X_1, X_2, ...$, be independent identically distributed random variables with Poisson distribution with parameter λ. Does:

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_{2i-1} - X_{2i})$$

converge in distribution to a normal distribution as $n \to \infty$

Consider random variables, $Y_i = X_{2i-1} - X_{2i}$ which are independent and identically distributed. We have:

$$\begin{aligned}E[Y_i] &= E[X_{2i-1}] - E[X_{2i}] = 0 \\ \text{Var}[Y_i] &= \text{Var}[X_{2i-1}] - \text{Var}[X_{2i}] = 2\lambda\end{aligned}$$

Where the second equation with variances works by independence of X_{2i-1} and X_{2i}. So by the central limit theorem:

$$\mathbb{P}\left\{ \frac{Y_1 + ... + Y_n}{\sqrt{2n\lambda}} < t \right\} \to \int_{-\infty}^{t} \frac{e^{-u^2/2}}{\sqrt{2\pi}} du$$

Hence:

$$\begin{aligned}\mathbb{P}\left\{ \frac{Y_1 + ... + Y_n}{\sqrt{n}} < t \right\} &\to \int_{-\infty}^{t/\sqrt{2\lambda}} \frac{e^{-u^2/2}}{\sqrt{2\pi}} du \\ &= \int_{-\infty}^{t} \frac{e^{-u^2/4\lambda}}{\sqrt{4\pi\lambda}}\end{aligned}$$

Which is a normal distribution with mean 0 and variance 2λ.

Now for the last section which is characteristic functions

Characteristic functions

The term characteristic function was first used by Poincaré in 1912 however this was the term given to what we now know to be the moment generating function which we covered in chapter 5. Some analysts find the term used strange as the function is just the Fourier transform of the probability measure (With i replaced by -i). The Fourier transform first appeared in English in 1923 however and so the probabilist's beat the analysts to call it the characteristic function! The theory of characteristic functions were also first developed in order to prove central limit theorems. The characteristic function is as follows:

Let X be a random variable and $t \in \mathbb{R}$, then e^{itX} is a bounded, complex valued random variable and hence has an expectation:

$$\varphi_X(t) = E[e^{itX}] = \int_{\mathbb{R}} e^{itX} f_X(x)dx = E[\cos(tX)] + iE[\sin(tX)]$$

that defines the characteristic function of X. A similar result holds in the discrete case.

This is very similar to the moment generating function (mgf) we discussed in chapter 5 however with the (mgf) it only exists for certain random variables with a restricted range of values of t but the characteristic function exists for all random variables. With the mgf we saw that it can have fascinating applications the same is true for the characteristic function as well. One such example is showing the poisson approximation to the binomial distribution holds by showing the characteristic function of binomial tends to the characteristic function of the poisson (When you let the probability of success, p, be $\frac{\theta}{n}$ in the binomial). This works since:

$$X \sim Y \Leftrightarrow \varphi_X(t) = \varphi_Y(t) \; \forall t \in \mathbb{R}$$

Properties of characteristic functions

Let φ be the characteristic function of a random variable X. Then φ: $\mathbb{R} \to \mathbb{C}$ is:

1. a bounded function satisfying $|\varphi(t)| \leq \varphi(0) = 1 \; \forall t \in \mathbb{R}$
2. a uniformly continuous function, that is:

$$\underset{t \in \mathbb{R}}{sup} |\varphi(t+h) - \varphi(t)| \to 0 \text{ as } h \to 0$$

Addition rule

If X and Y are independent random variables and $X + Y$ is their sum, then their respective characteristic functions satisfy:

$$\varphi_{X+Y}(t) = \varphi_X(t)\varphi_Y(t)$$

Proof.

$$\begin{aligned}\varphi_{X+Y}(t) &= E[e^{it(X+Y)}] \\ &= E[e^{itX}e^{itY}] \\ &= E[e^{itX}]E[e^{itY}] \\ &= \varphi_X(t)\varphi_Y(t)\end{aligned}$$

We can extend this but for mutually independent copies of a random variable too:

Extending the addition rule

Let $X_1, X_2, ..., X_n$ be mutually independent copies of a random variable X, and let

$$S_n = \alpha_n(X_1, X_2, ..., X_n)$$

be their sum, scaled by $\alpha_n > 0$. Then their respective characteristic functions satisfy:

$$\varphi_{S_n}(t) = \varphi_X^n(t) \ \forall t \in \mathbb{R}$$

The same as we did with mgf's we can do with characteristic functions which is that we can prove that sums of random variables are a certain distribution. We will show in a similar way that the sum of Bernoulli distributions is a binomial distribution but use the result just mentioned.

Example

For a Bernoulli random variable, X:

$$\begin{aligned}\varphi_X(t) &= e^{it0}\mathbb{P}(X = 0) + e^{it1}\mathbb{P}(X = 1) \\ &= 1 - \theta + e^{it}\theta\end{aligned}$$

We can now use the result mentioned before:

$$\varphi_{S_n}(t) = \varphi_X^n(t) = (1 - \theta + e^{it})^n$$

For a binomial random variable, Y:

$$\begin{aligned}\varphi_Y(t) &= \sum_{k=0}^{n} e^{itk}\mathbb{P}(X=k)\\ &= \sum_{k=0}^{n} e^{itk}\frac{n!}{k!(n-k)!}\theta^k(1-\theta)^{n-k}\\ &= \sum_{k=0}^{n}\binom{n}{k}(e^{it}\theta)^k(1-\theta)^{n-k}\\ &= (e^{it}\theta+1-\theta)^n\\ &= \varphi_{S_n}(t)\end{aligned}$$

We can also prove that sums of other mutually independent distributions are certain distributions, these include:

1. Sum of Gaussian random variables is also Gaussian
2. Sum of Poisson random variables is also Poisson
3. Sum of Cauchy random variables is also Cauchy
4. Sum of binomial random variables is also binomial
5. Sum of Gamma random variables is also Gamma
6. Sum of χ^2 random variables is also χ^2

See if you can prove any of these using characteristic functions!

Example

Let U be a random variable uniformly distributed on $[-1,1]$. Calculate its characteristic function:

$$\begin{aligned}\varphi_X(t) &= \frac{1}{2}\int_{-1}^{1} e^{itx}dx\\ &= \frac{1}{2}\left[\frac{e^{itx}}{it}\right]_{-1}^{1}\\ &= \frac{e^{it}-e^{-it}}{2it}\\ &= \frac{\sin(t)}{t}\end{aligned}$$

This is given that $t \neq 0$. We can also use characteristic functions to find expectations of a random variable and the variance similarly to how we did with moment generating functions.

Expectation

Let X be a random variable with finite expectation. Then the characteristic function $\varphi(t)$ of X is continuously differentiable with $\varphi'(0) = iE[X]$.

We can extend this result to find moments and ultimately the variance of random variables:

Moments

Suppose a random variable X has $(E[|X|^r]) < \infty$ for some integer $r \geq 1$. Then the characteristic function φ of X is r times continuously differentiable and:

$$\varphi^{(r)}(0) = i^r E(X^r)$$

Example

Find the expectation and variance of a Bernoulli random variable and a binomial random variable using characteristic functions:

We have already seen that the characteristic function of a Bernoulli random variable is:

$$\varphi(t) = 1 - \theta + e^{it}\theta$$

Hence:

$$\begin{aligned}\varphi'(0) &= i\theta e^{i0} = i\theta \\ \varphi''(0) &= i^2\theta^{i0} = i^2\theta\end{aligned}$$

Now using the expectation and moments theorems:

$$\begin{aligned} i\theta &= iE[X] \\ E[X] &= \theta \\ i^2E[X^2] &= i^2\theta \\ E[X^2] &= \theta \\ \text{Var}[X] &= \theta - \theta^2 \\ &= \theta(1-\theta)\end{aligned}$$

Now for the binomial distribution. We have already seen that the characteristic function of a binomial random variable is:

$$\varphi(t) = (1 - \theta + e^{it}\theta)^n$$

Hence:

$$\begin{aligned}\varphi'(0) &= in\theta e^{i0}(\theta e^{i0} - \theta + 1)^{n-1} = in\theta \\ \varphi''(0) &= -n\theta e^{i0}(\theta e^{i0} - \theta + 1)^{n-2}(n\theta e^{i0} - \theta + 1) \\ &= -n\theta(n\theta - \theta + 1)\end{aligned}$$

Now using the expectation and moments theorems, remembering we can sub in $i^2 = -1$:

$$\begin{aligned}in\theta &= iE[X] \\ E[X] &= n\theta \\ i^2E[X^2] &= i^2(-n\theta^2 + n^2\theta^2 + n\theta) \\ E[X^2] &= (-n\theta^2 + n^2\theta^2 + n\theta) \\ \text{Var}[X] &= (-n\theta^2 + n^2\theta^2 + n\theta) - n^2\theta^2 \\ &= n\theta(1 - \theta)\end{aligned}$$

This concludes the book unfortunately :(. Whilst I wish I could write an infinite book i have some studying to do for my university exams. I hope you enjoyed the book as much as I enjoyed writing it! I am doubting many people will read it let alone get to this point but if it helps even one person with their studies then i will be happy with that! I hope you succeed in your probability classes because of this book haha! Thank you to everyone who has read it :) it has been a pleasure!