Probability Theory

Aamod Varma

MATH - 3235, Fall 2025

Contents

Chapter 1

Introduction

Example. What is the probability that two people among N people have the same birthday. ◇

Example. What is the probability that all people have different birthday
We have,

$$\begin{aligned} q_1 &= 1 \\ q_2 &= \big(1 - \frac{1}{365}\big) \\ q_3 &= q_3\big(1 - \frac{2}{365}\big) \\ &\vdots \\ q_n &= \prod_{i=1}^{n-1}\big(1 - \frac{i}{365}\big) \end{aligned}$$

We get $q_n = 0.14$ which gives us 0.86 for the previous example.

◇

Note. We assume certain assumptions like the following to make this work,

1. Uniformity

2. Independence

Here we have a probability model and deduced the probability of an event,

Example. Say there is a test for a disease,

1. P(positive | sick) = 1

2. P(positive | not sick) = 0.01

Need to find P(sick | positive) which would be P(positive | sick) P(sick) / P(positive)
We test everybody, we have Assume 100 S and 100 NS,
100 P from the S, 99 P from the NS
So we have 199 P of which only 100 S which gives around .5

◇

1.1 Probability Theory

Experiment whose outcome is not determined. We define the following,

1. Ω : Sample space, set of possible outcomes

Example. (a) Throw a die,
$\Omega = \{1,2,3,4,5,6\} \rightarrow$ finite

(b) Flip a coin till heads,
$\Omega = \{1,2,3,\dots\} = \mathbb{N} \rightarrow$ countably infinite

(c) Time to wait till next bus arrival,
$\Omega = \mathbb{R}^+ \rightarrow$ uncountabaly infinite

◇

2. F : Family of events, $A, B, \dots$

   Something that may or may not happen

   Example. (a) For a die we can ask,

   - Is the outcome even?
   - Is the outcome ≤ 3?

   Here an event $A \subseteq \Omega$ and $|\Omega| = 6$ so $|2^\Omega| = 64$
   We have $F =$ family of events $= 2^\Omega$

   (b) Here we have,
   $\Omega = \mathbb{N}$ so $F = 2^\mathbb{N}$

   (c) In this case our sample space is $R^+ = (0, \infty)$. But we cannot take $2^\mathbb{R}$. So we axiomatically define F as noted below. Under this definition F is the smallest family that contains all open intervals of R

   ◇

3. P : How likely an event is

Definition (Axiomatic definition of F)**.** So here we define F to be a family of events of Ω if,

1. not empty
2. if $A \in F \Rightarrow A^c \in F$ ($A^c = \Omega \setminus A$)
3. for any two $A, B \in F$ then $A \cup B \in F$
4. If A_i for $i = 1, \dots, \infty$ are events, then $\bigcup_{i=1}^{\infty} A_i$ is an event

Note. Here, countable closure $\Rightarrow$ finite closure (proof just involves adding infinite ϕ to our finite sets $A_1, \dots, A_n$)

Note. Using this definition we have,

1. $A \in F \Rightarrow A^c \in F, \Rightarrow A \cup A^c = \Omega \in F$ and $\phi = \Omega^c \in F$

   So every event space has Ω, ϕ

2. $(A \cup B)^c = A^c \cap B^c \in F$ so,

   If $A_i, i = 1, 2, \dots$ are events then we have,
   $(\bigcap_{i=1}^{\infty} A_i)^c \in F = \bigcup_{i=1}^{\infty} A_i^c \in F$

1.2 Probability

Definition (Axiomatic definition of Probability). A probability is a function $\mathbb{P} : F \to [0,1]$ with the following probabilities, We want the following properties,

1. $\mathbb{P}(A) \geq 0$
2. $\mathbb{P}(\Omega) = 1$ and $\mathbb{P}(\phi) = 0$
3. If A & B are events, they are mutually exclusive if $A \cup B = \phi$ so it should have,

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

If A_i for $i = 1, 2, 3, \dots$ are events with $A_i \cap A_j$ where $i \neq j$ then,

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

Example. (a). For the die, we have $\mathbb{P}(\{i\})$ for $i \in \{1, \dots, 6\}$. So if Ω is finite, then the probability is completely defined by $\mathbb{P}(\omega)$ for $\omega \in \Omega$, here $\{\omega\}$ is called in atomic event. If A is an event then we have,

$$\mathbb{P}(A) = \sum_{\omega \in A} \mathbb{P}(w)$$

In particular, $\mathbb{P}$ is called uniform if,

$$\mathbb{P}(\omega) = \frac{1}{|\Omega|}$$

(b). Coin flip.
We have our sample space as $\mathbb{N}$. First, let's say that $\mathbb{P}(H) = p$ and $\mathbb{P}(T) = q = 1 - p$. Let x be the number of flips to get first head and $x \in \mathbb{N}$.

$$\begin{aligned} P(1) &= p \\ P(2) &= (1-p)p \\ &\dots \\ P(n) &= (1-p)^{n-1}p \end{aligned}$$

We have,

$$\begin{aligned} \sum_{n=1}^{\infty} (1-p)^{n-1}p = p \sum_{m=0}^{\infty} (1-p)^m \\ &= p\frac{1}{1-(1-p)} = \frac{p}{p} \\ &= 1 \end{aligned}$$

Note. This is true, $\sum_{n=0}^{\infty} x^n = \frac{1}{1-x}$ if $|x| < 1$
So we have $\mathbb{P}(A) = \sum_{n \in A} \mathbb{P}(n)$
(c). Consider $[A, B] \subset R$, if we take, $(x, y) \subset [A, B]$ so we have,

$$\mathbb{P}([x, y]) = k(y - x)$$

and

$$\mathbb{P}([A, B]) = 1$$

this means that $k = \frac{1}{B-A}$ so,

$$\mathbb{P}([x, y]) = \frac{y - x}{B - A}$$

◇

Definition (Probability Space). The probability space is defined by $(\Omega, \mathbb{F}, \mathbb{P})$ where Ω is a sample space, $\mathbb{F}$ is a family of events and $\mathbb{P}$ is a probability on $\mathbb{F}$

Some consequence are,
1. $\Omega = A \cup A^c$ and $A \cap A^c = \phi$. So,

$$\mathbb{P}(\Omega) = 1 = \mathbb{P}(A) + \mathbb{P}(A^c)$$

which gives us,

$$\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$$

2. As $\phi = \Omega^c \Rightarrow$ if $\mathbb{P}(\Omega) = 1 \Rightarrow \mathbb{P}(\phi) = 0$
3. Given A, B as events,

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

Proof. We know that $A = A \setminus B \cup (A \cap B)$ and $B = B \setminus A \cup (A \cap B)$

$$\begin{aligned}\mathbb{P}(A) &= \mathbb{P}(A \setminus B) + \mathbb{P}(A \cap B)\\ \mathbb{P}(B) &= \mathbb{P}(B \setminus A) + \mathbb{P}(A \cap B)\end{aligned}$$

We can write,

$$A \cup B = (A \setminus B) \cup (B \setminus A) \cup (A \cap B)$$

This gives us,

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

So get,

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

□

1.3 Conditional Probability

Given A, B what is the probability of B if I know that A happened?

Theorem 1.1. Given B with $\mathbb{P}(B) > 0$ let $\mathbb{Q}(A) = \mathbb{P}(A|B)$. $\mathbb{Q}$ is a probability.

Proof. 1. $\mathbb{Q}(A) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \geq 0$ so $\mathbb{Q}(A) \geq 0$
2. $\mathbb{Q}(\omega) = \frac{\mathbb{P}(\Omega \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B)}{\mathbb{P}(B)} = 1$
3.

$$\begin{aligned}\mathbb{Q}(\bigcup_{i=1}^{\infty} A_i) &= \frac{\mathbb{P}((\bigcup_{i=1}^{\infty} A_i) \cap B)}{\mathbb{P}(B)}\\ &= \frac{\mathbb{P}((\bigcup_{i=1}^{\infty} A_i \cap B))}{\mathbb{P}(B)}\\ &= \frac{\sum_{i=1}^{\infty} \mathbb{P}(A_i \cap B)}{\mathbb{P}(B)}\end{aligned}$$

□

$\mathbb{P}(A|B) = \mathbb{P}(A)$ then A is independent from B, this implies that, $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B) \Rightarrow \mathbb{P}(B|A) = \mathbb{P}(B)$

Definition. A and B are independent if $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$

Note. This implies that $\mathbb{P}(A|B) = \mathbb{P}(A)$

Example. A and B are independent iff A and B^c are independent.
We can write $A = (A \cap B) \cup (A \cap B^c)$. So we have,

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

Now we can write,

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

◇

Consider if we have three events A, B, C. Then if we have,

$$\begin{aligned} \mathbb{P}(A \cap C) &= \mathbb{P}(A)\mathbb{P}(C) \\ \mathbb{P}(A \cap B) &= \mathbb{P}(A)\mathbb{P}(B) \\ \mathbb{P}(B \cap C) &= \mathbb{P}(B)\mathbb{P}(C) \end{aligned}$$

This is called mutually independent (not a good definition for independence)

Example. Let four possible outcomes be $\{1, 2, 3, 4\}$. Now if we have $A = \{1, 2\}, B = \{1, 3\}, C = \{2, 3\}$. This gives us,

$$\begin{aligned} \mathbb{P}(A \cap B) &= \frac{1}{4} \\ \mathbb{P}(A) = \mathbb{P}(B) &= \frac{1}{2} \end{aligned}$$

Now $\mathbb{P}(A \cap B \cap C) = \mathbb{P}(\phi) = 0 \neq \mathbb{P}(A)\mathbb{P}(B \cap C)$
So if we want that $\mathbb{P}(A|B \cap C) = \mathbb{P}(A)$ then $\mathbb{P}(A \cap B \cap C) = \mathbb{P}(A)\mathbb{P}(B \cap C) = \mathbb{P}(A)\mathbb{P}(B)sw_{,0(}$a1) ! store value of c at the address in $a1\mathbb{P}(C)$

◇

Exercise. A, B, C are independent then $\mathbb{P}(A|B \cup C) = \mathbb{P}(A)$. We can write $B \cup C = (B \cap C^c) \cup (B \cap C) \cup (B^c \cap C)$

Proposition 1.2. In general, $A_i, i \in I$ of events. A_i are independent if $\forall J \subset I$ then,

$$\mathbb{P}(\bigcap_{j \in J} A_j) = \prod_{j \in J} \mathbb{P}(A_j)$$

Note. This implies that if $J_1, J_2 \subset I$ with $J_1 \cap J_2 = \phi$. Then any combination of $A_i, i \in J_1$ is independent to any combination of $A_i, i \in J_2$

Definition (Parition)**.** Assume a family of events A_i. We call it a partition if $\bigcup_i A_i = \Omega$ and $A_i \cap A_j = \phi, \forall i \neq j$.

Theorem 1.3. If B is an event and A_i is a partition, then

$$\mathbb{P}(B) = \sum_i \mathbb{P}(B|A_i)\mathbb{P}(A_i)$$

Proof. We write,

$$\begin{aligned} B &= \bigcup_i (B \cap A_i) \\ \mathbb{P}(B) &= \sum_i \mathbb{P}(B \cap A_i) \\ &= \sum_i \mathbb{P}(B|A_i)\mathbb{P}(A_i) \end{aligned}$$

□

Example. Consider two production lines,

1. 1000 items, 0.01 defective
2. 500 items, 0.02 defective

If all items are collected and pick one at random, what is the probability that that it is defective. If D is the event that the item is defective so we need to find $P(D)$ if we have I and II as both the production lines then we have,

$$\mathbb{P}(D) = \mathbb{P}(D|I)\mathbb{P}(I) + \mathbb{P}(D|II)\mathbb{P}(II) = 0.01 \times \frac{2}{3} + 0.02 \times \frac{1}{3} = \frac{0.04}{3}$$

We can also ask if an item is picked and it's defective, what is the probability that it is from line I. So we need to find $\mathbb{P}(I|D)$.

$$\begin{aligned} \mathbb{P}(I|D) = \frac{\mathbb{P}(I \cap D)}{\mathbb{P}(D)} &= \frac{\mathbb{P}(D|I)\mathbb{P}(I)}{\mathbb{P}(D)} \\ &= \frac{\mathbb{P}(D|I)\mathbb{P}(I)}{\mathbb{P}(D|I)\mathbb{P}(I) + \mathbb{P}(D|II)\mathbb{P}(II)} \\ &= \frac{0.01 \times \frac{2}{3}}{\frac{0.04}{3}} \\ &= \frac{1}{2} \end{aligned}$$

◇

Theorem 1.4 (Bayes Thoerem)**.** If A_i is a partition and B is an event. Then,

$$\mathbb{P}(A_i|B) = \frac{\mathbb{P}(B|A_i)\mathbb{P}(A_i)}{\sum_j \mathbb{P}(B|A_j)\mathbb{P}(A_j)}$$

Proof.

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

> And we have from the partition theorem that $\mathbb{P}(B) = \sum_j \mathbb{P}(B|A_j)\mathbb{P}(A_j)$. Plugging this back in gives us the theorem. □

Given P_1, P_2 positive at the first and second test. Then what is $\mathbb{P}(P_1 \cap P_2|NS)$

1.4 Examples

Example. Coin flip arrival of the first H
Let A be the event that $n > 10$ and B be the event that $n > 13$ and C is the even that $n > 13$.
We show that $\mathbb{P}(B|A) = \mathbb{P}(C)$
We compute $\mathbb{P}(A)$ first. So,

$$\begin{aligned}\mathbb{P}(A) = \sum_{n=11}^{\infty} \mathbb{P}(\{n\}) = \sum_{n=11}^{\infty} p(1-p)^{n-1} &= (1-p)^{10} \sum_{n=11}^{\infty} p(1-p)^{n-11} \\ &= (1-p)^{10}\end{aligned}$$

This is the probability that the first 10 flips are tails. Similarly $\mathbb{P}(B) = (1-p)^{13}$ and $\mathbb{P}(C) = (1-p)^3$
So $\mathbb{P}(B|A) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)}$
We have $B \subseteq A$ so it's the same as $\mathbb{P}(B)$ so we have,

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(B)}{\mathbb{P}(A)} = \frac{(1-p)^{13}}{(1-p)^3} = (1-p)^{10} = \mathbb{P}(C)$$

◇

Example. Given a box with 9 balls and 3 are blue and 6 are red.
Let R_1 be the first draw being red so we have $\mathbb{P}(R_1) = 2/3$. If we don't reinsert the ball we have $\mathbb{P}(R_2|R_1) = 5/8$ and $\mathbb{P}(R_2|B_1) = 6/8$.
Suppose if you pick a blue ball you win 10 and a red will give you 0.

$$\mathbb{P}(R_2) = \frac{2}{3}\frac{5}{8} + \frac{6}{8}\frac{1}{3} = \frac{5}{12} + \frac{3}{12} = \frac{8}{12} = \frac{2}{3}$$

◇

Remark. The point here is before starting drawing, the probability of the first, second, etc draw of reds or blues are the same. The probability will change given information on what the previous draw is but in the beginning it shouldn't make a difference.

Example. Consider N flips of a fair coin. Sample space would be $\{H, T\}^N$. Here $H_i = \{\text{i'th flip is } H\}$. We have $|\Omega| = 2^N$. Some $\omega \in \Omega$ is a sequence $\omega = \{w_1, \ldots, w_N\}$. We can show that H_i is independent from H_j if $i \neq j$. We have,

$$\mathbb{P}(H_i) = \frac{2^{n-1}}{2^n} = \frac{1}{2}$$

Similarly

$$\mathbb{P}(H_j) = \frac{1}{2}$$

And $\mathbb{P}(H_i \cap H_j) = \frac{1}{4}$
So we see that their independent. So the collection $H_i, i = 1, 2, \ldots, N$ is an independent collection of events. ◇

Remark. More general we have, given two subsets I_1 and $I_2 \subset \{1, \ldots, N\}$ such that $I_1 \cap I_2 = \phi$ then for events A and B are some specific outcomes for $i \in I_1$ and $i \in I_2$ respectively. Then both the events A, B are independent.
Here A, B are called cylinder sets.

Example. If you flip the coin infinitely many times we have, $= \{0, 1\}^N$ which is uncountable. So we consider $I \subset N$ and take $\sigma_i \in \{0, 1\}, i \in I$. So our cylinder set would be,

$$\{\omega : \omega_i = \sigma_i\}$$

so here $\omega = \{\omega_1, \omega_2, \dots, \omega_n, \dots\}$

◇

Example. What is the probability of the outcome of getting only zeroes - would be zero. But this isn't an event in our family of events. But we can consider an event $A_n = \{\text{The outcome of the first } n \text{ flips is } 0\}$. For a fair coin this is 2^{-n}. We can say that,

$$A = \bigcap_{n=1}^{\infty} A_n \qquad \text{as we have } A_{n+1} \subset A_n$$

◇

Theorem 1.5. If A_n is a decreasing collection s.t. $A_{n+1} \subset A_n$ and we have $A = \bigcap_{n=1}^{\infty}$ then we have,

$$\mathbb{P}(A) = \lim_{n\to\infty} \mathbb{P}(A)$$

Theorem 1.6. If A_n is increasing so $A_n \subset A_{n+1}$ and $A = \bigcup_{n=1}^{\infty} A_n$ then we have,

$$\mathbb{P}(A) = \lim_{n\to\infty} \mathbb{P}(A_n)$$

Proof. We have,

$$\begin{aligned}
A_1 \cup A_2 &= A_1 \cup (A_2 \setminus A_1) \\
A_1 \cup A_2 \cup A_3 &= A \cup (A_2 A_1) \cup (A_3 A_2) \\
\bigcup_{i=1}^{N} A_i &= A_1 \bigcup_{i=2}^{N} B_i \qquad B_i = A_i A_{i-1} \\
\mathbb{P}(\bigcup_{i=1}^{\infty} A_i) &= \mathbb{P}(A_1) + \sum_{i=2}^{\infty} \mathbb{P}(B_i) \\
&= \mathbb{P}(A_1) + \sum_{i=2}^{\infty} (\mathbb{P}(A_i) - \mathbb{P}(A_{i-1})) \\
&= \mathbb{P}(A_1) + \lim_{n\to\infty} \sum_{i=2}^{n} (\mathbb{P}(A_i) - \mathbb{P}(A_{i-1})) \\
&= \mathbb{P}(A_i) = \lim_{n\to\infty} (\mathbb{P}(A_n) - \mathbb{P}(A_1))
\end{aligned}$$

□

Chapter 2

Random Vairables

2.1 Introduction

Definition. If we have $(\Omega, \mathscr{F}, \mathbb{P})$ then $X : \Omega \to \mathbb{R}$ is a discrete random variable if,

1. $X(\Omega)$ is finite or countable.
2. $\forall a \in \mathbb{R}$ $(\forall a \in X(\Omega))$
$$\{\omega : X(\omega) = a\} \in \mathscr{F}$$

Remark. Here Ω can be uncountable but the point is that the mapping in $\mathbb{R}, X(\Omega)$ has to be countable.

Remark. The point of the second condition is that for any value in R, there is some event in $\mathscr{F}$ which maps to that value. So this point guarantees measurability.

Example. Consider $Y =$ number of H in T tosses of a coin, possible values are clearly $\{0, 1, \ldots, N\}$. We have,

$$\begin{aligned}
\mathbb{P}(Y = 0) &= (1-p)^N \\
\mathbb{P}(Y = 1) &= Np(1-p)^{N-1} \\
\mathbb{P}(Y = 2) &= \binom{N}{2} p^2 (1-p)^{N-2} \\
\mathbb{P}(Y = y) &= \binom{N}{y} p^y (1-p)^{N-y}
\end{aligned}$$

◇

Remark. Here $p_Y(y) = \mathbb{P}(Y = y)$ is called the probability mass function of Y.

Definition. If X is a discrete random variable then if,

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

Then $p(x)$ is called the probability mass function (pmf)

Remark. We also have,

$$\begin{aligned}
\sum_{x \in X(\Omega)} p(x) &= \sum_{x \in X(\Omega)} \mathbb{P}(X = x) \\
&= \mathbb{P}\Bigg(\bigcup_{x \in X(\Omega)} \{\omega \in \Omega : X(\omega) = x\} \Bigg) \\
&= \mathbb{P}(\Omega) = 1
\end{aligned}$$

2.2 Examples

Example. A random variable that takes only 2 values (conventionally represented with 0, 1) is called a **Bernoulli r.v**. The pmf of a Bernoulli r.v. is completely given by $p(1) = \mathbb{P}(X = 1)$ and $p(0) = 1 - p(1)$ ◇

Example. If a random variable that takes values $0, \dots, N$ with p.m.f,

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

is called a ***binomial r.v.*** This is a 2 parameter pmf (p, N).

◇

Remark. ***Binomial can be thought of as a sum of*** N ***independent Bernoulli r.v. with parameter*** p. In addition, it can be thought of as the number of successes in N independent Bernoulli trials with success probability p.

Example. I have a bowl with N blue balls and M red balls. If we extract n balls without reinsertion. Then, what is the probability of x blue.

Here the total possible outcomes is $\binom{M+N}{n}$. We want to select x blue balls and $n - x$ red balls. So possibility for blue is $\binom{N}{x}$ and for red is $\binom{M}{n-x}$. So,

$$p(x) = \frac{\binom{N}{x}\binom{M}{n-x}}{\binom{M+N}{n}}$$

This is called a **hypergeometric r.v.** which is a 3 parameter pmf. ◇

Note. If $x \ll N$ then $\binom{N}{x} \sim N^x$

Remark. If n and x are fixed and take $N, M \to \infty$ then the hypergeometric pmf converges to a binomial.

Example. Geometric r.v.. Take $\Omega = \{\underline{\omega} = (\omega_1, \dots, \omega_n, \dots)\}$ where $\omega_i \in \{0, 1\}$

$X(\underline{\omega}) =$ the position of the first 1 in $\underline{\omega}$

$\{\underline{\omega} \mid X(\underline{\omega}) = n\} = A_n$
But A_n is the set of all $\underline{\omega}$ such that $\omega_1 = \omega_2 = \cdots = \omega_{n-1} = 0$
Here $X(\Omega) = \mathbb{N}$ which means its countable and $X^{-1}(n) \in \mathscr{F}$ so it means it's a random variable.

$p(n) = \mathbb{P}(X = x) = p(1-p)^{n-1}$
If X has the p.m.f $p(x)$ above then it's called a **Geometric r.v.** with parameter p.

◇

Remark. ***Geometric can be thought of as the number of trials until the first success in a sequence of independent Bernoulli trials each with success probability*** p***.***

Remark. Here A_n is a cylinder set as we're fixing the value of ω on a finite number of points (in this case from $1, \dots, n-1$)

Remark. Sometimes q is taken as $1 - p$ so this is also correct, $p(x) = pq^{x-1}$

Example. Suppose you have a binomial with large N but $pN = \lambda$ is finite. So,

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

From here we have $p = \frac{\lambda}{n}$ so we have,

$$p(x) = \frac{N!}{x!(N-x)!} \left(\frac{\lambda}{N}\right)^x \left(1 - \frac{\lambda}{N}\right)^{N-x}$$

But we have $\frac{N!}{(N-x)!}$ is around N^x for $x \ll N$. We also have $\lim_{N\to\infty}(1-\frac{\lambda}{N})^N = e^{-\lambda}$ so is $\lim_{N\to\infty}(1-\frac{\lambda}{N})^{N-x} = e^{-\lambda}$. This gives us,

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

This is called a **Poisson distribution** ◇

Note. Poisson is from $0 \to \infty$

If $p(x)$ is the p.m.f of a random variable, then, we need $p(x) \geq 0, \forall x$ and $\sum_x p(x) = 1$

Example. For binomial we have,

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

Now,

$$\begin{aligned}\sum_{x=0}^{N} p(x) &= \sum_{x=0}^{N} \binom{N}{x} p^x (1-p)^{N-x}\\ &= (p + (1-p))^N\\ &= 1^N = 1\end{aligned}$$

◇

Example. For Poisson we have,

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

◇

2.3 Mean of a random variable

Also called average, expectation etc. Here $\mathbb{E}(x)$ is the expectation of X.

Example. With probability p say you get 1 and with $(1-p)$ you get 0. Here the average after N flips is,

$$1 \cdot \frac{\#1}{N} + 0 \cdot \frac{\#0}{N} = 1 \cdot p(1) + 0p(0) = p$$

◇

Definition. If X is a discrete random variable with p.m.f $p(x)$ then,

$$\mathbb{E}(X) = \sum_x x \cdot p(x)$$

Note. Here x is all possible values of X.

Example. For Bernoulli r.v.,

$$\mathbb{E}(X) = p$$

◇

Example. For Binomial N, p we have,

$$\mathbb{E}(X) = \sum_{x=0}^{N} x \binom{N}{x} p^x (1-p)^{N-x}$$

First we have $x\binom{N}{x} = \frac{N!}{(x-1)!(N-x!)} = \frac{(N-1)!(N)}{(x-1)!(N-x)!} = N\binom{N-1}{x-1}$. So,

$$N\sum_{x=1}^{N}\binom{N-1}{x-1}p^x(1-p)^{N-x} = Np\sum_{x=1}^{N}\binom{N-1}{x-1}p^{x-1}(1-p)^{N-x}$$

Taking $y = x-1$ we have,

$$Np\sum_{y=0}^{N-1}\binom{N-1}{y}p^y(1-p)^{N-1-y} = Np$$

◇

Remark. Intuitively if $Y_1, \dots, Y_n$ are independent Bernoulli r.v. with parameter p. Then $\sum_{i=1}^{N} Y_i = X$ binomial and $\mathbb{E}(\sum_{i=1}^{N} Y_i) = \sum_{i=1}^{N}\mathbb{E}(Y_i) = Np$

Example. For Geometric r.v. we have,

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

So,

$$\sum_{x=0}^{\infty} xp(x) = p\sum_{x=0}^{\infty} x(1-p)^{x-1}$$

We see that $x(1-p)^{x-1} = \frac{-d}{dp}(1-p)^x$ so,

$$\begin{aligned}\sum_{x=1}^{\infty} x(1-p)^{x-1} &= -\sum_{x=1}\frac{d}{dp}(1-p)^x \\ &= -\frac{d}{dp}\sum_{x=1}^{\infty}(1-p)^x \\ &= -(1-p)\frac{d}{dp}\sum_{x=0}^{\infty}(1-p)^x \\ &= -\frac{d}{dp}(\frac{1}{p}-1) \\ &= -1 \times -\frac{1}{p^2} = \frac{1}{p^2}\end{aligned}$$

So we have,

$$\mathbb{E}(X) = p\sum_{n=1}^{\infty} x(1-p)^{x-1} = p\frac{1}{p^2} = \frac{1}{p}$$

◇

Example. For Poisson we have,

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

◇

Example. For Hypergeometric we have,

$$p(x) = \frac{\binom{N}{x}\binom{M}{n-x}}{\binom{M+N}{n}}$$

In this example we have N red balls and M blue balls and X is the r.v. of number of reds by taking x balls. Intuitively we get $\mathbb{E}(X) = np$ where $p = \frac{N}{N+M}$ ◇

Remark. The point here is that extracting the red ball at the third extraction and the first extraction will be the same (at the beginning that is, even though given information about the first 2 extractions the third will have different probability). But in the beginning we don't have any more information and there is no reason to think that the first or second extraction is better than the later ones.

2.4 Review

1. A discrete r.v. $X : \Omega \to \mathbb{R}$ where $X(\Omega)$ is finite or countable and we have $X^{-1}(x) \in \mathscr{F}$.
2. The probability mass function (pmf) is defined as,

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

Example. We defined the following random variables,

1. Bernoulli r.v. $X \sim (p)$, $p_X(1) = p$, $p_X(0) = 1 - p$.
2. Binomial r.v. $X \sim (n, p)$, $p_X(k) = \binom{n}{k}p^k(1-p)^{n-k}$, $k = 0, 1, \dots, n$
3. Hypergeometric r.v. $X \sim (N, K, n)$, $p_X(k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$, $k = 0, 1, \dots, \min\{n, K\}$.
4. Geometric r.v. $X \sim (p)$, $p_X(k) = (1-p)^{k-1}p$, $k = 1, 2, \dots$
5. Poisson r.v. $X \sim (\lambda)$, $p_X(k) = e^{-\lambda}\frac{\lambda^k}{k!}$, $k = 0, 1, 2, \dots$

◇

We also have $\mathbb{E}(X) = \sum_x x p_X(x)$ and got,

Example. For Bernoulli, $\mathbb{E}(x) = p$,
For Binomial, $\mathbb{E}(X) = np$,
For Hypergeometric, $\mathbb{E}(X) = \frac{N}{N+M}n$
For Geometric, $\mathbb{E}(X) = \frac{1}{p}$,
For Poisson, $\mathbb{E}(X) = \lambda$. ◇

Consider a r.v. Y such that $ImY = \{a, b\}$ and,

$$\mathbb{P}(X = b) = p$$

And have,

$$X = \frac{Y - a}{b - a}$$

Then X takes two values 0 and 1 with probabilities $1 - p$ and p. So,

$$\mathbb{P}(X = 1) = \mathbb{P}(Y = b) = p$$

So X is a bernoulli r.v. with parameter p and,

$$X = \frac{Y - a}{b - a} \Rightarrow Y = a + (b - a)X$$

So any function that takes two values can be written as a linear functions of a Bernoulli r.v. And,

$$\begin{aligned}\mathbb{E}(Y) &= (1-p)a + pb \\ &= a + (b-a)p\end{aligned}$$

We can also say that,

$$\begin{aligned}\mathbb{E}(Y) &= \mathbb{E}(a + (b-a)X) \\ &= a + (b-a)\mathbb{E}(X) \\ &= a + (b-a)p\end{aligned}$$

Theorem 2.1. So if X is a r.v. and $a, b \in \mathbb{R}$ then,

$$\mathbb{E}(a + bX) = a + bE(X)$$

Proof. We have $Y = a+bX$. The possible values of Y are $\{y : y = a+bx, x \text{ is a possible value for } x\}$ Here,

$$p_y(y) = p_x(x) \text{ where } y = a + bx$$

So,

$$\begin{aligned}\mathbb{E}(Y) &= \sum_y y p_Y(y) \\ &= \sum_x (a+bx) p_X(x) \\ &= a \sum_x p_X(x) + b \sum_x x p_X(x) \\ &= a + bE(X)\end{aligned}$$

□

Example. If X is a r.v. then $Y = X^2$. We have,

$$\mathbb{P}(Y = y) = \mathbb{P}(X = \pm\sqrt{y})$$

so,

$$p_Y(y) = p_X(\sqrt{y}) + p_X(-\sqrt{y})$$

Then we have,

$$\begin{aligned}E(Y) &= \sum_y y p_Y(y) \\ &= \sum_y y(p_X(\sqrt{y}) + p_X(-\sqrt{y})) \\ &= \sum_x x^2 p_X(x) \\ &= E(X^2)\end{aligned}$$

Thus we have $\mathbb{E}(X^2) = \sum_x x^2 p_X(x)$ ◇

Theorem 2.2. Let X be a r.v. and $f : \mathbb{R} \to \mathbb{R}$. Then,

$$Y = f(x) \text{ is a r.v.}$$

$$\mathbb{P}(Y = y) = \sum_{x: f(x)=y} \mathbb{P}(X = x)$$

Moreover,

$$\mathbb{E}(f(x)) = \mathbb{E}(Y) = \sum_x f(x) p_X(x)$$

Remark. If $A_x = \{\omega | X(\omega) = x\}$ we have $A_x \in \mathscr{F}$ then,

$$\{\omega | Y(\omega) = y\} = \bigcup_{x: f(x)=y} A_x$$

Example. Take $\mathbb{E}(X^n) = m_n(X)$ is the $n'th$ moment of X.

We would also like to know what the width of the distribution is. A naive approach is $\mathbb{E}(X - \mathbb{E}(X)) = 0$ which is useless. But we can define,

$$V(X) = \mathbb{E}((X - \mathbb{E}(X))^2)$$

which is the average squared distance from the average.
And we call,

$$\sigma_X^2 = V(X) \text{ where } \sigma \text{ is the standard deviation}$$

◇

Example. Consider for bernoulli,

$$\mathbb{E}(X^2) = 1p + 0(1 - p) = p$$

But,

$$\begin{aligned} \mathbb{E}((X - \mathbb{E}(X))^2) = \mathbb{E}((X - p)^2) &= (1 - p)^2 p + (0 - p)^2 (1 - p) \\ &= p(1 - p) \end{aligned}$$

◇

Remark. Take $\mathbb{E}(X) = m$ then we can simplify $V(X)$ as follows,

$$\begin{aligned} V(X) &= \mathbb{E}((X - m)^2) \\ &= \mathbb{E}(X^2 - 2mX + m^2) \\ &= \mathbb{E}(X^2) - 2m\mathbb{E}(X) + m^2 \\ &= \mathbb{E}(X^2) - 2m^2 + m^2 \\ &= \mathbb{E}(X^2) - m^2 \\ &= \mathbb{E}(X^2) - (\mathbb{E}(X))^2 \end{aligned}$$

So using this formula for bernoulli, we have,

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

Theorem 2.3. If X is a random variable then $\mathbb{E}(X^2) \geq \mathbb{E}(X)^2$.

Proof. We have $V(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2 \geq 0$. Thus, $\mathbb{E}(X^2) \geq (\mathbb{E}(X))^2$. □

2.5 Variance

Definition. If X is a discrete r.v. then, variance of X is defined as

$$Var(X) = E[(X - E[X])^2]$$

Remark. But we can also write it as,

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

Example. If we have a **bernoulli r.v**. we have,

$$\begin{aligned} E(X) &= p \\ Var(X) &= p(1-p) \end{aligned}$$

We have $V(X)$ is maximal if $p = \frac{1}{2}$ and zero if $p = 0$ or $p = 1$ ◇

Example. For **Binomial r.v.**, we have,

$$\begin{aligned} E(X) &= np \\ Var(X) &= np(1-p) \end{aligned}$$

We have,

$$\begin{aligned} E(X(X-1)) &= \sum_{n=0}^{\infty} x(x-1)\binom{n}{x} p^x (1-p)^{n-x} \\ &= n(n-1)p^2 \end{aligned}$$

Now we have $Var(X) = E(X(X-1)) + E(X) - E(X)^2$ which is,

$$= np(1-p)$$

◇

Example. Consider **Hypergeometric r.v.**, with parameters N, M, n. We have,

$$p(x) = \frac{\binom{N}{x}\binom{M}{n-x}}{\binom{M+N}{n}}$$

We know that,

$$E(X) = np$$

and that

$$V(X) = np(1-p)\frac{N+M-n}{N+M-1}$$

◇

Example. For **Poisson r.v.**, we have,

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

and we can do,

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

Now we have $Var(X) = E(X(X-1)) + E(X) - E(X)^2$ which is,

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

◇

Remark. Poisson variance is also similar to binomial variance, where $n \to \infty$.
Consider,

Example.

$$p(x) = \frac{C}{x^2} \qquad \text{for } x \geq \text{N}$$

We know that $\sum_{x=1}^{\infty} \frac{1}{x^2}$ converges to some k. So we have,

$$C = \frac{1}{k}$$

But we have the $E(X)$ doesn't exist as,

$$\begin{aligned} E(X) &= \sum_{x=N}^{\infty} x \frac{C}{x^2} \\ &= C \sum_{x=N}^{\infty} \frac{1}{x} \end{aligned}$$

But the sum is not divergent so we can say that $E(X) = +\infty$

Here for X the expected value is dominated by large values of X which are rare but have a large contribution to the expected value. ◇

2.6 Conditional Expectation

Definition. We know that if A is an event with $\mathbb{P}(A) > 0$ then $\mathbb{Q}(B) = \mathbb{P}(B \mid A)$ is a probability. So if X is a random variable then,

$$\begin{aligned} E_Q(X) &= \sum_x \mathbb{P}(X = x \mid A) \\ &= E(X \mid A) \end{aligned}$$

which is called te conditional expectation of X given A.

Theorem 2.4. Let A_i be a partition of Ω such that $\mathbb{P}(A_i) > 0$ for all i. Then,

$$E(X) = \sum_i \mathbb{E}(X \mid A_i)\mathbb{P}(A_i)$$

Proof.

$$\begin{aligned}E(X) &= \sum_i x\mathbb{P}(X = x)\\ &= \sum_i x \sum_j \mathbb{P}(X = x \mid A_j)\mathbb{P}(A_j)\\ &= \sum_j \mathbb{P}(A_j) \sum_x x\mathbb{P}(X = x \mid A_j)\\ &= \sum_j \mathbb{E}(X \mid A_j)\mathbb{P}(A_j)\end{aligned}$$

□

Example. Consider if 10 percent of the population is sick and 90 percent is not. Now take n individual and X is the r.v denoting the number of sick people. As N is large, we can approximate X by a binomial r.v. with parameters $n, p = 0.1$. Now we do a test and consider,

$$\mathbb{P}(P \mid Notsick) = 0.1, \mathbb{P}(NP \mid Sick) = 0$$

Take the n extracted people and test them. Let Y be the number of positive tests. Here Y is also binomial. Now we have $\mathbb{P}(P) = \mathbb{P}(P \mid S)\mathbb{P}(S) + \mathbb{P}(P \mid NS)\mathbb{P}(NS) = 0.1 \times 0.9 + 1 \times 0.1 = 0.19$. Now we have,

$$E(Y) = n \times 0.19$$

◇

Example. Flip a coin and look at the initial string of heads. Let X be the length of this string. We have arrival of first head as $\frac{1}{p} - 1$. Now we have, for tails at first flip $\frac{1}{q} - 1$ so we have,

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

Another way to do this is,

$$\begin{aligned}E(X) &= E(X \mid H)\mathbb{P}(H) + E(X \mid T)\mathbb{P}(T)\\ &= p^{x-1}qp + q^{x-1}pq\end{aligned}$$

which gives us the same answer.

◇

Remark. The logic in the first method is that arrival of first head means that the rest before it is tails i.e. a continuous run of tails. So as $1/p$ includes the heads we do $\frac{1}{p} - 1$ which represent the average length of a run of tails. We get $1/p$ as that is the mean of the geometric random variable.

Another way of doing this is to define X as the number of continuous run of heads. So $\mathbb{P}(X = x) = p^x q$ and finding the expected value of this gives us the desired answer.

Chapter 3

Multivariate discrete distribution and Independence

Example. Roll 2 dice (blue and red). So,

$$\Omega = \{(x,y) \mid 1 \leq x, y \leq 6\}$$

We have X outcome of the first dice and Y the outcome of the second, with

$$X(\Omega) = Y(\Omega) = \{1, \dots, 6\}$$

and,

$$p(x,y) = \mathbb{P}(X = x, Y = y) = \frac{1}{36}$$

So here $p(x,y)$ is called the *joint probability mass function* of the r.v. X and Y.

We can say $\mathbb{P}(X = x) = \sum_y \mathbb{P}(X = x \mid Y = y) = p_X(x)$ which is called the marginal over X of $p(x,y)$
Similarly, $\mathbb{P}(Y = y) = \sum_x \mathbb{P}(Y = y \mid X = x) = p_Y(y$ which is the marginal over Y of $p(x,y)$

In this case of the dice we have,

$$p_X(x) = p_Y(y) = \frac{1}{6} \quad \forall x, y$$

◇

Example. Consider the above example but now let $Z_+ = X + Y$ and $Z_- = X - Y$. So we have,

$$Z_+ = \{2, \dots, 12\} \quad \text{and} \quad Z_- = \{-5, \dots, 5\}$$

Now we can say that,

$$\begin{aligned} Z_+ = 2 &\Rightarrow Z_- = 0 \\ Z_+ = 3 &\Rightarrow Z_- = \{1, -2\} \\ Z_+ = 4 &\Rightarrow Z_- = \{-2, 0, 2\} \\ &\vdots \end{aligned}$$

We can ask $\mathbb{P}(Z_+ = 4 \text{ and } Z_- = -2) = \frac{1}{36}$ as there is only one possibility which is when $X = 1, Y = 3$.

Notice that the set of possible values of Z_- is like a rhomboid i.e. it increases linearly until 7 and then decreases linearly.

We can ask $\mathbb{P}(Z_+ = 4) = \sum_z \mathbb{P}(Z_+ = 4 \text{ and } Z_i = z) = \frac{3}{36} = \frac{1}{12}$

◇

Definition. If x and y are r.v. over Ω then,

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

is called the joint p.m.f of X and Y. And we must have,

$$\begin{aligned} p(x,y) &\geq 0 \\ \sum_{x,y} p(x,y) &= 1 \end{aligned}$$

3.1 Expected Values

Given X, Y we want $\mathbb{E}(X)$. We have,

$$\begin{aligned} \mathbb{E}(X) &= \sum_x x p_X(x) \\ &= \sum_x x \sum_y \mathbb{P}(X = x \text{ and } Y = y) \\ &= \sum_{x,y} x \mathbb{P}(X = x \text{ and } Y = y) \end{aligned}$$

Similarly,

$$\mathbb{E}(Y) = \sum_{x,y} y \mathbb{P}(X = x \text{ and } Y = y)$$

Given (X, Y), I can think of X as a function of (X, Y). Take a less simple function,

$$Z = aX + bY \quad a, b \in \mathbb{R}$$

We can ask $\mathbb{E}(Z)$ we have,

$$\begin{aligned} \mathbb{E}(Z) &= \sum_z \mathbb{P}(Z = z) \\ &= \sum_{x,y} (ax + by) \mathbb{P}(X = x \text{ and } Y = y) \\ &= a \sum_{x,y} x \mathbb{P}(X = x \text{ and } Y = y) + b \sum_{x,y} y \mathbb{P}(X = x \text{ and } Y = y) \\ &= a\mathbb{E}(X) + b\mathbb{E}(Y) \end{aligned}$$

Corollary 3.1.

$$\mathbb{E}(aX + bY + c) = a\mathbb{E}(X) + b\mathbb{E}(Y) + c$$

Consider r.v X, Y and $g : \mathbb{R}^2 \to \mathbb{R}$ so that,

$$Z = g(X, Y)$$

Since X, Y are r.v we can say,

$$\{\omega \mid X(\omega) = x \text{ and } Y(\omega) = y\} = \{\omega \mid X(\omega) = x\} \cap \{\omega \mid Y(\omega) = y\} \in \mathscr{F}$$

We have,

$$\begin{aligned} \mathbb{E}(Z) &= \mathbb{E}(g(X,Y)) \\ &= \sum_{x,y} g(x,y)\mathbb{P}(X = x \text{ and } Y = y) \end{aligned}$$

3.2 Covariance

Definition. Covariance is defined as,

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

Note. The covariance of X with itself is just the variance of X.

Note. We can used this to measure how two r.v are correlated.

We can define the correlatoin coefficient as,

$$\rho = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

Where $\sigma_X^2 = Var(X)$ and $\sigma_Y^2 = Var(Y)$ and,

$$-1 \leq \rho_{x,y} \leq 1$$

We also have,

$$\begin{aligned} \rho_{X,Y} &= 1 \quad X = aY \quad a > 0 \\ \rho_{X,Y} &= -1 \quad X = bY \quad b < 0 \\ \rho_{X,Y} &= 0 \quad \text{we say they are uncorrelated} \end{aligned}$$

3.3 Independence

Definition. If X, Y are independent then if I know the value of X it gives me no information on the value of Y. So we have,

$$\mathbb{P}(Y = y \mid X = x) = \mathbb{P}(Y = y)$$

or that,

$$\mathbb{P}(Y = y \text{ and } X = x) = \mathbb{P}(X = x)\mathbb{P}(Y = y) \quad \forall x, y$$

Theorem 3.2. If X, Y are independent then we have $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$

Proof. We have,

$$\begin{aligned}\mathbb{E}(XY) &= \sum_{x,y} xy\mathbb{P}(X = x \text{ and } Y = y)\\ &= \sum_{x,y} xy\mathbb{P}(X = x)\mathbb{P}(y = y)\\ &= \sum_{x,y} x\mathbb{P}(X = x)\sum_{y} y\mathbb{P}(y = y)\\ &= \mathbb{E}(X)\mathbb{E}(Y)\end{aligned}$$

□

Remark. Which also means that we have $(XY) = 0$. However,

$$Cov(X, Y) = 0 \not\Rightarrow \text{ independence}$$

For instance consider we have possible values only $\{(0, 1), (1, 0), (-1, 0), (0, -1)\}$. So X can take value -1 with probability $\frac{1}{4}$, 0 with $\frac{1}{2}$ and 1 with $\frac{1}{4}$. We have $\mathbb{E}(X) = 0, \mathbb{E}(Y) = 0$ and $XY = 0$. So we get $Cov(X, Y) = 0$. But if we know that value of X then we know the possible value of Y. More specifically we have,

$$\mathbb{P}(X = 0) = \mathbb{P}(Y = 0) = \frac{1}{2} \quad \text{but} \quad \mathbb{P}(X = 0 \text{ and } Y = 0) = 0$$

Here X, Y are not independent but $Cov(X, Y) = 0$.

Remark. So if X, Y are independent and g, f are functions then we get,

$$\mathbb{E}(g(X)f(Y)) = \mathbb{E}(g(X))\mathbb{E}(f(Y))$$

Definition. In general given $X_1, \ldots, X_n$ then the joint p.m.f is,

$$\mathbb{P}(X_1 = x_1 \ldots X_n = x_n)$$

Remark. Independence means that $\mathbb{P}(X_1 = x_1 \ \& \ \ldots \ \& \ X_n = x_n) = \prod_{i=1}^{n} \mathbb{P}(X_i = x_i)$

We can further show this is true for any subset of X by summing over X_i to exclude X_i.

We can recognize independence if, $p(x, y) = f(x)g(y)$. For instnace,

$$\begin{aligned}p(x, y) &= \frac{e^{-\lambda-\mu}}{x!y!}\mu^x\lambda^y\\ &= \left(\frac{e^{-\lambda-\mu}}{x!}\mu^x\right)\left(\frac{\lambda^y}{y!}\right)\end{aligned}$$

So we're able to write $p(x, y)$ as the product of two functions dependent on x, y.

We know that $V(X) = \mathbb{E}((X - E(X)^2)$. So we look at $V(X + Y)$ and get,

Theorem 3.3. If X and Y are independent then we have,

$$V(X + Y) = V(X) + V(Y)$$

Proof.

$$\begin{aligned}
V(X+Y) &= \mathbb{E}(((X+Y)-E(X+y))^2) \\
&= \mathbb{E}((X+Y-\mu_x-\mu_y)^2)) \\
&= \mathbb{E}((X-\mu_x)^2+(Y-\mu_y)^2+2(X-\mu_x)(Y-\mu_x)) \\
&= \mathbb{E}((X-\mu_x)^2)+\mathbb{E}((Y-\mu_y)^2)+\mathbb{E}(2(X-\mu_x)(Y-\mu_x))) \\
&= Var(X)+(Y)+2Cov(X,Y)
\end{aligned}$$

So we have $V(X+Y)=V(X)+V(Y)$ if we have $Cov(X,Y)=0$. But if X,Y are independent then we have $Cov(X,Y)=0$ □

Remark. Note that $V(X+Y)=V(X)+V(Y)$ does **NOT** mean that X,Y are independent.

Remark. Let $X_1,\dots,X_n$ be independent Bernoulli r.v. of parameter p. Consider n independent coin flips so,

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

Y is binomial n,p. So we have,

$$\mathbb{E}(Y)=\sum_i \mathbb{E}(X_i)=np$$

Now for variance of Y we have,

$$V(Y)=\sum_i V(X_i)=np(1-p)$$

as each X_i is independent from each other and variance of X_i is $(1-p)$

Given $V(aX)$ we have $V(aX)=\mathbb{E}(a^2X^2)-\mathbb{E}(aX)^2=a^2V(X)$. This also gives us standard deviation $\sigma_{aX}=|a|\sigma_X$.

We also have $V(X+a)=V(X)$ as,

$$\begin{aligned}
\mathbb{E}(X+a) &= \mathbb{E}(X)+a \\
(X+a)-\mathbb{E}(X+a) &= X-\mathbb{E}(X)
\end{aligned}$$

Remark. If variance gives an idea of spread of the distribution then adding a constant value to a r.v. X should not change the spread.

Remark. We can also see it by looking at a a constant as independent from X and then use linearity of variance when independence is true to separate it out.

3.4 Sum of random variables

Take X,Y with $p(x,y)$. Now consider,

$$Z=X+Y$$

We have,

$$\begin{aligned}
\mathbb{P}(Z=z) &= \sum_x \mathbb{P}(X=x \;\&\; Y=z-x) \\
&= \sum_x p(x,z-x)
\end{aligned}$$

If X and Y are independent then,

$$\mathbb{P}(Z=z)=\sum_x p_X(x)p_Y(z-x)$$

This is called the convolution.

Example. If X and Y are Poisson, then with μ, ν we have,

$$p_X(x) = \frac{\mu^x}{x!}e^{-\mu} \quad p_Y(y) = \frac{\nu^y}{y!}e^{-nu}$$

Now consider $Z = X + Y$ then,

$$\begin{aligned}
\mathbb{P}(Z = z) &= \sum_{x=0}^{z} p_X(x)p_Y(z-x) \\
&= \sum_{x=0}^{z} \frac{\mu^x \nu^{z-x}}{x!(z-x)!}e^{-(\mu+\nu)} \\
&= \sum_{x=0}^{z} \frac{z!}{z!}\frac{\mu^x \nu^{z-x}}{x!(z-x)!}e^{-(\mu+\nu)} \\
&= \frac{1}{z!}\sum_{x=0}^{z}\binom{z}{x}\mu^x\nu^{z-x}e^{-(\mu+\nu)} \\
&= \frac{1}{z!}(\mu+\nu)^z e^{-(\mu+\nu)}
\end{aligned}$$

◇

Exercise. If X is binomial n, p and Y is binomial m, p then $X + Y$ is binomial $n + m, p$

Remark. If X is binomial n, p_1 and Y is binomial n, p_2 where $p_1 \neq p_2$. Then $X+Y$ is not binomial. But we have,

$$\mathbb{E}(X+Y) = np_1 + mp_2$$

and,

$$V(X+Y) = np_1(1-p_1) + mp_2(1-p_2)$$

3.5 Indicator Function

Definition. Indicator function is a r.v $1_A(\omega)$ such that,

$$1_A(\omega) = \begin{cases} 1 & \omega \in A \\ 0 & \omega \notin A \end{cases}$$

We have the follwoing,

$$\begin{aligned}
1_A^c(\omega) &= 1 - 1_A(\omega) \\
1_{A\cup B}(\omega) &= 1_A(\omega)1_B(\omega) \\
1_{A\cap B}(\omega) &= 1_{(A^c\cap B^c)^c}(\omega) = 1 - (1 - 1_A(\omega))(1 - 1_B(\omega)) \\
&= 1_A(\omega) + 1_B(\omega) - 1_A(\omega)1_B(\omega)
\end{aligned}$$

Now we have,

$$\mathbb{E}(1_{A\cup B}) = \mathbb{E}(1_A) + \mathbb{E}(1_B) - \mathbb{E}(1_A 1_B)$$

this tells us that,

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

In general,

$$1_{\bigcup_i A_i}(\omega) = 1 - \prod_i (1 - 1_{Ai}(\omega))$$

We can write this as,

$$\sum_{I \in \{1,\ldots,n\}} (-1)^{|I|+1}\mathbb{P}\left(\bigcap_{i\in I} A_i\right)$$

Example. Consider a gas station on a highway and N_1 a poisson r.v λ_1 representing a car that stops and needs service on top of gas. Let N_2 be poisson λ_2, for a car that stops and do not need service.
So we have,

$$N_1 + N_2 = \text{ Poisson with } \lambda_1 + \lambda_2$$

We can describe it as follows as well,

N is a Poisson with prob λ for every car that arrives with probability p of requiring service and $1-p$ of not requiring service.

Now let N_1 is the number of car that requires service. We have $N_1$is a Poisson r.v with parameter $p\lambda$. We can compute this by taking $\mathbb{P}(N_1 = n) = \sum_{m \geq n} \mathbb{P}(N_1 = n \mid N = m)\mathbb{P}(N \mid m)$
The first term is binomial with m, p and second term is a poisson. So this will give us,

$$\sum_{n \geq m} \binom{m}{n} p^n (1-p)^{m-n} \frac{\lambda^m}{m!} e^{-\lambda} = \frac{(\lambda p)^n}{n!} e^{-\lambda p}$$

But an easier way to do this is by assuming that N is actually a binomial with large n and $p = \frac{\lambda}{n}$. But if we have a binomial then at each time step a car arrives with probability $\frac{\lambda}{n}$ and probability is p that car needs service so we have $\frac{p\lambda}{n}$ as probability of car arriving and needing service at each time step.

So consider a binomial with n flips and $\frac{p\lambda}{n}$ which is a Poisson with parameter $p\lambda$

So if N_2 is the number of cars that do not need service it is a Poisson with $(1-p)\lambda$ ◇

Remark. Point here is that sum of two independent Poisson is Poisson.

Chapter 4

Generating Functions

Given a sequence $a_0, a_1, \ldots, a_n$, an infinite sequence of numbers.
A generating function is of the form,

$$f(s) = \sum_{n=0}^{\infty} a_n s^n$$

or,

$$f(s) = \sum_{n=0}^{\infty} \frac{a_n}{n!} s^n$$

Example. If we take $a_n = 2^n$ then we have,

$$f(s) = \sum_{n=0}^{\infty} 2^n s^n = \frac{1}{1-2s} \qquad \text{when } |s| < \frac{1}{2}$$

◇

Example. We have $a_n = n!$ then,

$$\sum_n a_n s^n \qquad \text{does not exist for any } s = 0$$

◇

Theorem 4.1. If $f(s)$ and $g(s)$ are generating function for a sequence a_n and b_n,

$$f(s) = \sum_n a_n s^n \qquad g(s) = \sum_n b_n s^n$$

Then,

1. If $a_n = b_n, \forall n \Rightarrow f(s) = g(s)$
2. If $\exists s_0, f(s) = g(s)$ for $s \leq s_0 \Rightarrow a_n = b_n \forall n$

Proof. We have $a_0 = f(0) = g(0) = b_0$. Now,

$$f'(s) = \sum_{n=0}^{\infty} n a_n s^{a-1} = \sum_{n=0}^{\infty} (n+1) a_{n+1} s^n$$
$$g'(s) = \sum_{n=0}^{\infty} n b_n s^{a-1} = \sum_{n=0}^{\infty} (n+1) b_{n+1} s^n$$

So we have $g'(0) = b_1, f'(0) = a_1$. So if $f(s) = g(s), |s| < 0$ then we have, $f'(0) = g'(0)$ then we have $a_1 = b_1$

So we have $f^{(n)}(0) = n!a_n$ and $g^{(n)}(0) = n!b_n$ so if $f(s) = g(s), |s| < 0$ then we have $f^{(n)} = g^{(n)}(0)$ and hence $a_n = b_n, \forall n$. □

4.1 Random variables with values in $\mathbb{N}$

Let X be a r.v such that $Im(X) = \mathbb{N}$. Now we have,

$$\begin{aligned} G_X(s) &= \sum_{x=0}^{\infty} \mathbb{P}(X = x)s^n \\ &= \sum_{x=0}^{\infty} p_X(x)s^n \end{aligned}$$

for $0 < p_X(x) < 1$

So,

$$\left| \sum_{x=0}^{\infty} p_X(s)s^x \right| \leq \sum_{x=0}^{\infty} |s|^x < \infty \quad \text{if } s < 1$$

In particular if modulus of X is bounded then $Im(X) \subset \{0, \ldots, N\}$ then $G_X(s)$ is a polynomial and exists for every s.

Example. Bernoulli

We have 0 w $1 - p$ and 1 w p. So,

$$G_X(s) = (1 - p)s^0 + ps^1 + sp = (1 - p) = (s - 1)p + 1$$

◇

Example. Binomial

We have $\mathbb{P}(X = x) = \binom{N}{x} p^x (1 - p)^{N-x}$. So our generating function is,

$$\begin{aligned} G_X(s) &= \sum_{x=0}^{\infty} \binom{N}{x} p^x (1 - p)^{N-x} s^x \\ &= \sum_{x=0}^{\infty} \binom{N}{x} ps^x (1 - p)^{N-x} \\ &= ((1 - p) + sp)^n \end{aligned}$$

◇

Example. Poisson

We have,

$$\begin{aligned} G_X(s)) &= \sum_{x}^{\infty} \frac{\lambda^x}{x!} e^{-\lambda} s^x \\ &= e^{-\lambda} \sum_{x}^{\infty} \frac{(\lambda s)^s}{x!} \\ &= e^{-\lambda} e^{\lambda s} = e^{-\lambda(1-s)} \end{aligned}$$

◇

Applications

We have

$$G_X(s) = \sum_{x=0}^{\infty} s^x \mathbb{P}(X = s) = \mathbb{E}(s^x)$$

So,

$$\begin{aligned}
G_X(1) &= \mathbb{E}(1) = 1 \\
\frac{d}{ds}G_s(s) &= \mathbb{E}(xs^{x-1}) \\
G'_X(1) &= \mathbb{E}(X) \\
G''_X(s) &= \mathbb{E}(X(X-1)s^{X-2}) \\
G''_X(1) &= \mathbb{E}(X(X-1)) \\
G^{(n)}(1) &= \mathbb{E}(X(X-1)\dots(X-n+1))
\end{aligned}$$

And we have,

$$\mathbb{E}(X^2) = \mathbb{E}(X(X-1)) + E(X) = G''_X(1) + G'_X(1)$$

So this gives us,

$$V(X) = G''_X(1) + G'_X(1) - (G'_X(1))^2$$

Example. Geometric r.v.

We have $p_X(x) = p(1-p)^{x-1}$ so,

$$\begin{aligned}
G_X(s) &= p\sum_{x=1}^{\infty}(1-p)^{x-1}s^x \\
&= ps\sum_{x=1}^{\infty}(s(1-p))^{x-1} \\
&= ps\left(\frac{1}{1-s(1-p)}\right) = \frac{ps}{1-sq}
\end{aligned}$$

And we see that $G_X(1) = \frac{p}{1-q} = 1$ and,

$$\begin{aligned}
G'_X(s) &= \frac{p}{1-sq} + \frac{psq}{(1-sq)^2} \\
G'_X(1) &= \frac{p}{1-q} + \frac{pq}{p^2} = 1 + \frac{q}{p} = \frac{1}{p}
\end{aligned}$$

◇

So $G^n(0) = n!\mathbb{P}(X = n)$ and $G^n_X(1) = \mathbb{E}(X(X-1)\dots(X-n+1))$.

4.2 Sum of random variables

Let X and Y be independent r.v with values in n then we have,

$$G_{X+Y}(s) = \mathbb{E}(s^{X+Y}) = \mathbb{E}(s^X s^Y) = \mathbb{E}(s^X)\mathbb{E}(s^Y)$$

as functions of X and Y are independent if X and Y are independent.

So we have,

$$G_{X+Y}(s) = G_X(s)G_Y(s)$$

Example. X and Y are Poisson with λ, μ. Let $Z = X + Y$. We see,

$$\begin{aligned} G_X(s) &= e^{-\lambda(1-s)} \\ G_Y(s) &= e^{-\mu(1-s)} \\ G_{X+Y}(s) &= e^{-(\lambda+\mu)(1-s)} \end{aligned}$$

and hence $X + Y$ is a Poisson r.v with parameter $\lambda + \mu$ $\diamond$

Chapter 5

Distribution and density functions

5.1 Distribution functions

Similar to discrete r.v we can define continuous r.v as X on $(\sigma, F, \mathbb{P})$ is a mapping $\sigma \to \mathbb{R}$ such that $\forall x \in \mathbb{R}$ we have,

$$\{\omega \in \Omega : X(\omega) \leq x\} \in F$$

Definition. The distribution function F_X of a r.v is $F_X : \mathbb{R} \to [0, 1]$ defined as,

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

Remark. Here F is called the cumulative distribution function (CDF).

Properties of distribution functions,

1. For any $x \leq y$ we have $F_X(x) \leq F_X(y)$. So F has to be monotonic non-decreasing.
2. We have $\lim_{x\to\infty} F_X(x) = 1$ and $\lim_{x\to-\infty} F_X(x) = 0$.
3. F has to be right-continuous which means we have $F(X + \varepsilon) \to F(X)$ for any $\varepsilon \to 0$ from the right side. So $\varepsilon > 0$.
4. We have $P(a < X \leq b) = P(X \leq b) - P(X \leq a) = F(b) - F(a)$

Exercise (5.11)**.** Let X be a random variable taking integer values such that P(X = k) = pk for k = . . . , 1, 0, 1, ... Show that the distribution function of X satisfies FX(b) FX(a) = pa+1 + pa+2 + · · · + pb for all integers a, b with a < b.

Solution. As X is a discrete r.v taking on integers we have $P(x - 1 < X \leq x) = P(x) = F(x) - F(x - 1)$. Now we can write

$$\begin{aligned} F(b) - F(a) &= F(b) - F(b-1) + F(b+1) + \cdots - F(a) \\ &= (F(b) - F(b-1)) + (F(b+1) + F(b-2)) + \cdots - F(a) \\ &= p_{a+1} + p_{a+2} + \cdots + p_b \end{aligned}$$

5.2 Examples of Distribution functions

Example (Uniform Distribution)**.** We have,

$$F(x) = \begin{cases} 0 & \text{if } x < a, \\ \frac{x-a}{b-a} onthe & \text{if } a \leq x \leq a, \\ 1 & \text{if } x > b, \end{cases}$$

A r.v with this distribution function is said to have the uniform distribution on the interval (a, b)
$\diamond$

Example (Exponential Distribution). For $\lambda > 0$, F is given by,

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

$\diamond$

Exercise (5.18). We have $F(y) = \alpha F_1(y) + (1-\alpha)F_2(y)$ and $F(x) = \alpha F_1(x) + (1-\alpha)F_2(x)$ and if $y > x$ we get $F(y) - F(x) = \alpha F_1(y) + (1-\alpha)F_2(y) - \alpha F_1(x) + (1-\alpha)F_2(x) = \alpha(F_1(y) - F_1(x)) + (1-\alpha)(F_2(y) - F_2(x))$

Now as F_1 and F_2 are distribution functions they both are monotonic non decreasing and hence $F_2(y) - F_2(x) > 0$ and $F_1(y) - F_1(x) > 0$. This gives us,

$$\begin{aligned} F(y) - F(x) &= \alpha(F_1(y) - F_1(x)) + (1-\alpha)(F_2(y) - F_2(x)) \\ &> 0 \end{aligned}$$

which means F is monotonic non decreasing

Now we need to show that as $x \to \infty$ and $x \to -\infty$ we have $F(x) = 1$ and 0 respectively. We have,

$$\lim_{x\to\infty} F_1(x) = 1, \lim_{x\to\infty} F_2(x) = 1$$

So, $\lim_{x\to\infty} \alpha F_1(x) + (1-\alpha)F_2(x) = \alpha + (1-\alpha) = 1$ and similarly as,

$$\lim_{x\to-\infty} F_1(x) = 0, \lim_{x\to-\infty} F_2(x) = 0$$

we have, $\lim_{x\to-\infty} \alpha F_1(x) + (1-\alpha)F_2(x) = \alpha 0 + (1-\alpha)0 = 0$

Now as both F_1 and F_2 are continuous from the right as F is a linear combination of those functions we have that F is continuous from the right as well.

Lastly we have for any $a < b$ that $P(X \leq b) = F(b) = \alpha F_1(b) + (1-\alpha)F_2(b)$ and similarly $P(x \leq a) = F(a) = \alpha F_1(a) + (1-\alpha)F_2(a)$. If $F_1(k) = P_1(X \leq k)$ and $F_2(k) = P_2(X \leq k)$ we get,

$$\begin{aligned} F(b) - F(a) &= \alpha(F_1(b) - F_1(a)) + (1-\alpha)(F_2(b) - F_2(a)) \\ &= \alpha(P_1(a < X \leq b)) + (1-\alpha)(P_2(a < X \leq b)) \\ &= P(a < X \leq b) \end{aligned}$$

Exercise (5.19). We have $F(x) = c\int_{-\infty}^{x} e^{-|u|}du$ for $x \in \mathbb{R}$. We need to answer for what value of c is F a distribution function.

We need some c such that $\lim_{x\to\infty} F(x) = 1$ so we need,

$$\begin{aligned} 1 &= c\int_{-\infty}^{\infty} e^{-|u|}du \\ &= 2c\int_{0}^{\infty} e^{-u}du \\ &= 2c[-e^{-u}]_0^\infty \\ &= 2c \end{aligned}$$

So $c = 1/2$

5.3 Continuous random variables

Discrete r.v only take on countable many values and their distribution functions look like step functions. The r.v for which the distribution functions are smooth are called continuous r.v

Definition. A r.v is continuous if its distribution function F_X can be written as,

$$F_X(x) = \mathbb{P}(X \leq x) = \int_{-\infty}^{x} f(x)dx \quad \text{for } x \in \mathbb{R}$$

Remark. Here $f(x)$ is called the probability density function (PDF).

Example. Consider $\mathbb{P}(X = 1) = p$ and $\mathbb{P}(X = 0) = 1 - p$ where X is arrival in a second. Now if we want to consider arrival in a smaller intervals say N we have, $\frac{p}{N}$ as probability of arrival in $\frac{1}{N}$ time and $1 - \frac{p}{N}$ the probability of not arriving.

Taking $N \to \infty$ we have,

$$\mathbb{P}(X > t) = \lim_{n \to \infty} (1 - \frac{p}{n})^{nt} = e^{-pt}$$

So we have,

$$\begin{aligned} F(t) &= \mathbb{P}(X \leq x) = 1 - e^{-pt} \\ f(t) &= pe^{-pt} \end{aligned}$$

Which is written as $\lambda e^{-\lambda t}$ and,

$$f(x) = \begin{cases} 0 & x < 0 \\ \lambda e^{-\lambda} x & x \geq 0 \end{cases}$$

$$F(x) = \begin{cases} 0 & x < 0 \\ 1 - e^{-\lambda x} & x \geq 0 \end{cases}$$

◇

If X is a cont r.v then we can find the density function generally by,

$$f_X(x) = \begin{cases} \frac{d}{dx} F_x(x) & \text{if the derivative exists at } x \\ 0 & \text{otherwise} \end{cases}$$

A density function is in some way the analogous of the probability mass function in the discrete case. For instance we have,

$$\begin{aligned} f_X(x) \geq 0 \text{ for } x \in \mathbb{R} &\quad p_Y(x) \geq 0 \\ \int_{-\infty}^{\infty} f_X(x)dx = 1 &\quad \sum_x p_Y(x) = 1 \end{aligned}$$

However, f unlike p doesn't actually give us the probability. For instance we can have $f_X(x)$ be greater than 1. However a more analogous version to p would be if we consider a small δ and take $P(x < X \leq x + \delta) = F(x + \delta) - F(x) = \int_x^{x+\delta} f_X(u)du = f_X(x)\delta x$.

Theorem 5.1. If X is cont w PDF f_X then,

$$\mathbb{P}(X = x) = 0$$

$$\mathbb{P}(a \leq X \leq b) = \int_a^b f_X(u)du$$

A r.v can also be neither discrete nor continuous.

5.4 Common Density Functions

In general if f satisfies,

$$f(x) \geq 0 \quad \text{for } x \in \mathbb{R}$$

and

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

then f is a density function of some $r.v$ w CDF,

$$F(x) = \int_{-\infty}^{x} f(u)du$$

Example (Uniform distribution)**.** The density function is,

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

◇

Example (Exponential distribution)**.** With parameter $\lambda > 0$ we have,

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

◇

Example. Normal Standard
We have $f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ which gives us,

$$\begin{aligned} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-x^2/2} &= 1 \\ \int_{-\infty}^{\infty} e^{-x^2/2} &= \sqrt{2\pi} \end{aligned}$$

So we have $f(x)$ integrates to 1 and is positive and is a p.d.f

The distribution function is $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-y^2/2} dy$.

This is called the probability integral.

The general normal is of the form,

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

◇

5.5 Functions of random variables

Let X be a r.v and $Y = g(X)$ be a function of X. If g is well-behaved then Y is also a r.v.

Theorem 5.2. If X is a cont r.v w PDF f_x and g is a increasing and differentiable function, then $Y = g(X)$ has density function,

$$f_Y(y) = f_X(g^{-1}(y))\frac{d}{dy}[g^{-1}(y)] \text{ for } y \in \mathbb{R}$$

Proof.

$$\begin{aligned}\mathbb{P}(Y \leq y) &= \mathbb{P}(g(x) \leq y) \\ &= \mathbb{P}(X \leq g^{-1}(y)) \\ &= F_X(g^{-1}(y))\end{aligned}$$

Now we know the density function is the derivative of the CDF so we have,

$$\begin{aligned}f_Y &= \frac{d}{dy}F_Y(y) \\ &= \frac{d}{dy}F_X(g^{-1}(y)) \\ &= f_X(g^{-1}(y))\frac{d}{dy}g^{-1}(y)\end{aligned}$$

□

5.6 Expectations of continuous random variable

Example. Expected of uniform in $[A, B]$ is,

$$\begin{aligned}\int_{-\infty}^{\infty} xf(x)dx &= \int_A^B \frac{x}{B-A}dx \\ &= \frac{1}{B-A}\frac{x^2}{2}\bigg|_A^B = \frac{A+B}{2}\end{aligned}$$

◇

Example. Expected of exponential with pdf $\lambda e^{-\lambda x}$

$$E(X) = \int_0^{\infty} \lambda x e^{-\lambda x}dx$$

$$\lambda e^{-\lambda x} = -\frac{d}{dx}e^{-\lambda x}$$

so,

$$\begin{aligned}E(X) &= -xe^{-\lambda x}\big|_0^{\infty} + \int_0^{\infty} e^{-\lambda x}dx \\ &= \int_0^{\infty} e^{-\lambda x}dx \\ &= -e^{-\lambda x}\frac{1}{\lambda}\bigg|_0^{\infty} \\ &= \frac{1}{\lambda}\end{aligned}$$

so if X is exponential λ we have $E(X) = \frac{1}{\lambda}$

If we write $f(x) = \frac{1}{\mu}e^{-\frac{x}{\mu}}$ then $E(X) = \mu$ ◇

Example. For gaussian we have,

$$E(X) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} xe^{-x^2/2} dx = 0$$

as it is an odd function. ◇

Chapter 6

Multivariate distributions and independence

6.1 Random vectors and independence

Definition. The joint distribution function of two random variables X, Y is,

$$F_{X,Y}(x, y) = \mathbb{P}(X \leq x, Y \leq y)$$

Remark. We have similar properties to ordinary distributions like,

$$\lim_{x,y \to -\infty} F(x, y) = 0$$

$$\lim_{x,y \to \infty} F(x, y) = 1$$

$$F(x_1, y_1) \leq F(x_2, y_2) \quad \text{if } x_1 \leq x_2, y_1 \leq y_2$$

Given the joint distribution of two r.v we can find that of a single variable as follows,

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

More precisely we have,

$$F_X(x) = \lim_{y \to \infty} F(x, y) \qquad F_Y(y) = \lim_{x \to \infty} F(x, y)$$

These are called the marginal distribution functions of X and Y.

Definition. We call X, Y independent if

$$F(X, Y) = F(X)F(Y)$$

or

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

6.2 Joint density functions

For single variables we have X is cont if the distribution function can be written as,

$$F_X(x) = \mathbb{P}(X \leq x) = \int_{-\infty}^{x} f(x)dx \quad \text{for } x \in \mathbb{R}$$

Definition. X, Y are called jointly continuous r.v if the distribution function can be expressed as,

$$F_{X,Y}(x,y) = \mathbb{P}(X \leq x, Y \leq y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{X,Y}(u,v) du dv$$

We can similar find f from F by finding the second order partial derivative of F as follows,

$$f_{X,Y}(x,y) = \begin{cases} \frac{\partial^2}{\partial x \partial y} F_{X,Y}(x,y) & \text{if it exists} \\ 0 & \text{otherwise} \end{cases}$$

We have similar properties to single variable such as,

$$f_{X,Y}(x,y) \geq 0$$

and,

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(u,v) du\, dv = 1$$

In general if a function f satisfies the above it is a joint density function of some pair of r.v X, Y.

We see that f in this case is also not directly analogous to the joint probability mass function in the discrete case. As we can $f(x,y) \geq 1$ in some cases. But a more analogous equivalent is to consider a small $\delta x, \delta y$ rectangle as follows,

$$\mathbb{P}(x \leq X \leq x + \delta x, y \leq Y \leq y + \delta y) \approx f_{X,Y}(x,y)\, \delta x\, \delta y$$

6.3 Marginal density functions and Independence

We can find the marginal density function of a r.v as follows,

$$\begin{aligned} f_X(x) &= \frac{d}{dx} \mathbb{P}(X \leq x, Y \leq \infty) \\ &= \frac{d}{dx} \int_{-\infty}^{x} \int_{-\infty}^{\infty} f_{X,Y}(u,v) du\, dv \\ &= \int_{-\infty}^{\infty} f_{X,Y}(x,v) dv \end{aligned}$$

If X, Y are independent then we also have,

$$\begin{aligned} F(x,y) &= F(x)F(y) \\ f(x,y) &= f(x)f(y) \end{aligned}$$

Theorem 6.1. Two jointly continuous variables are independent if and only if their joint density function can be written as,

$$f_{X,Y}(x,y) = g(x)h(y) \quad \text{for } x, y \in \mathbb{R}$$

i.e. as a product of a function of the first variable and another of the second variable.

Example. Say we have,

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

Then we have the marginals as,

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)dy = \int_{-\infty}^{\infty} e^{-x-y}dy = e^{-x} \text{ if x} > 0\text{, 0 otherwise}$$

Symmetrically we have $f_Y(y) = e^{-y}$ if $y > 0$ and we can also conclude that X and Y are independent ◇

Exercise. We have,

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

We know the density function double integrates to 1 so we have,

$$\begin{aligned} 1 &= \int_0^1 \int_y^1 cx\, dx\, dy \\ &= \int_0^1 [\frac{cx^2}{2}]_y^1\, dy \\ &= \int_0^1 \frac{c}{2} - \frac{cy^2}{2}\, dy \\ &= [\frac{cy}{2} - \frac{cy^3}{6}]_0^1 \\ &= \frac{c}{2} - \frac{c}{6} = \frac{c}{3} \end{aligned}$$

So $c = 3$
Now to find the marginals we need to integrate over each of the variables so we have,

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} f(x,y)\, dy \\ &= \int_0^x 3x\, dy \\ &= 3x^2 \end{aligned}$$

Marginal over y is,

$$\begin{aligned} f_Y(y) &= \int_{-\infty}^{\infty} f(x,y)\, dx \\ &= \int_y^1 3x\, dx \\ &= [\frac{3x^2}{2}]_y^1 = \frac{3}{2} - \frac{3y^2}{2} \end{aligned}$$

Clearly we do not have $f(x,y) = f(x)f(y)$ so hence they are dependent.

Exercise. Need to find if var are independent and $\mathbb{P}(X > Y)$ and $\mathbb{P}(Y > Z)$ for,

$$f(x,y,z) = \begin{cases} 8xyz & \text{if } 0 < x, y, z < 1 \\ 0 & \text{otherwise} \end{cases}$$

Solution. We have,

$$\begin{aligned} f_X(x) &= \int_0^1 \int_0^1 f(x,y,z)\, dy\, dz \\ &= \int_0^1 \int_0^1 8xyz\, dy\, dz \\ &= 4x\frac{z^2}{2} \\ &= 2x \end{aligned}$$

Because of symmetric we have marginal over y and z are $2y$ and $2z$ respectively. Now we also have $f(x, y, z) = f(x)f(y)f(z)$ hence we have independence for all x, y, z.

To find $\mathbb{P}(X > Y)$ we have,

$$\begin{aligned}
\mathbb{P}(X > Y) &= \int_0^1 \int_y^1 \int_0^1 8xyz\ dz\ dx\ dy \\
&= \int_0^1 \int_y^1 4xydx\ dy \\
&= \int_0^1 4y \int_y^1 xdx\ dy \\
&= \int_0^1 4y \frac{x^2}{2}]_y^1\ dy \\
&= \int_0^1 4y \left(\frac{1}{2} - \frac{y^2}{2}\right)\ dy \\
&= \int_0^1 2y - 2y^3\ dy \\
&= [y^2 - \frac{y^4}{2}]_0^1 \\
&= \frac{1}{2}
\end{aligned}$$

6.4 Sums of continuous random variables

If X, Y have joint density function $f_{X,Y}$ then,

$$\mathbb{P}(Z \leq z) = \mathbb{P}(X + Y \leq z) = \int \int_A f_{X,Y}(x, y)\ dx\ dy$$

Here $A = \{(x, y) \in \mathbb{R}^2 : x + y \leq z\}$. So this is equivalent to,

$$\begin{aligned}
\mathbb{P}(Z \leq z) &= \int_{-\infty}^{\infty} \int_{-\infty}^{z-y} f(x, y)\ dx\ dy \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{z} f(x, y - x)\ dy\ dx
\end{aligned}$$

Differentiating with respect to z we have t he density function of Z which is,

$$f_Z(z) = \int_{-\infty}^{\infty} f(x, z - x)\ dx$$

Theorem 6.2. If X, Y are independent and continuous, then $Z = X + Y$ is,

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

6.5 Conditional density functions

Say X, Y are jointly continuous r.v. Then we have,

$$f_Y(y) = \int_{-\infty}^{\infty} f(x, y)\ dx$$

Now when considering conditional density, for instance $\mathbb{P}(Y \leq y \mid X = x)$ the issue that arises is that we have $\mathbb{P}(X = x) = 0$ unlike int the discrete case. So we have to consider the even $x \leq X \leq x + \delta x$. So we have,

$$\mathbb{P}(Y \leq y \mid x \leq X \leq x + \delta x) = \frac{\mathbb{P}(Y \leq y, x \leq X \leq x + \delta x)}{\mathbb{P}(x \leq X \leq x + \delta x)}$$

Now for the numerator we have,

$$\begin{aligned}\mathbb{P}(Y \leq y, x \leq X \leq x + \delta x) &= \int_{-\infty}^{y} \int_{x}^{x+\delta x} f(x, y)\ dx\ dy \\ &= \int_{-\infty}^{y} f_{X,Y}(x, y)\delta x\ dy\end{aligned}$$

And similar,

$$\begin{aligned}\mathbb{P}(Y \leq \infty, x \leq X \leq x + \delta x) &= \int_{-\infty}^{\infty} \int_{x}^{x+\delta x} f(x, y)\ dx \\ &= \int_{x}^{x+\delta x} f_X(x)\ dx \\ &= f_X(x)\delta x\end{aligned}$$

So we have,

$$\begin{aligned}\mathbb{P}(Y \leq y, x \leq X \leq x + \delta x) &= \frac{\int_{-\infty}^{y} f_{X,Y}\ dy\ \delta x}{f_X(x)\delta x} \\ &= \int_{-\infty}^{y} \frac{f_{X,Y}(x, v)}{f_X(x)}\ dv \\ &= G(y)\end{aligned}$$

So we have G a distribution function with density function $g(y) = \frac{f_{X,Y}(x,y)}{f_X(x)}$.

Theorem 6.3. Conditional density of Y given $X = x$ is denoted $f_{Y|X}(. \mid x)$ defined by,

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

6.6 Expectations of continuous r.v

Theorem 6.4. We have,

$$\mathbb{E}(g(X, Y)) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f_{X,Y}(x, y)\ dx\ dy$$

Other properties such as the following hold,

1. $\mathbb{E}(aX + bY) = a\mathbb{E}(X) + b\mathbb{E}(Y)$
2. $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$ if X, Y are independent (note that the converse is false, it is only true if $\mathbb{E}(g(X)h(Y)) = \mathbb{E}(g(X))\mathbb{E}(h(Y))$ is true for all functions $g, h : \mathbb{R} \to \mathbb{R}$)

Definition (Conditional expectation)**.** The conditional expectation of Y given $X = x$, written by $\mathbb{E}(Y \mid X = x)$ is the mean of the conditional density function,

$$\mathbb{E}(Y \mid X = x) = \int_{-\infty}^{\infty} y f_{Y|X}(y \mid x)\ dy = \int_{-\infty}^{\infty} y \frac{f_{X,Y}(x, y)}{f_X(x)}\ dy$$

We have another useful theorem from this which is, If X, Y are jointly cont r.v then we have,

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

Remark. One way to look at this is to calculate expected value of Y, we may first fix a value for x then then compute the expected value and then average over all x later.

6.7 Normal distribution

Univariate Normal

The normal r.v is called $\phi(z) = \mathbb{P}(Z = z)$ where $z \approx N(0, 1)$ where,

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

$$\phi(z) = \int_{-\infty}^{z} e^{-\frac{1}{2}x^2} \frac{1}{\sqrt{2\pi}} dx$$

So $z \in N(0, 1)$ and $\mathbb{P}(a \leq Z \leq b) = \phi(b) - \phi(a)$

If we have $X \in N(\mu, \sigma^2)$ then that means the density is,

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

If X is normal then $aX + b$ is normal. So we have $E(X) = \mu$ and $Var(X) = \sigma^2$

If you take $z = \frac{x-\mu}{\sigma}$ then we have $E(z) = 0$ and $Var(z) = 1$. So $z = N(0, 1)$.

When we take $\mathbb{P}(X \leq a)$ we can write it as,

Let X from $N(1, 2)$ then find x' such that we have,

$$\mathbb{P}(X \geq x') = 0.02$$

we can write this as,

$$\mathbb{P}(X \leq x') = 0.98$$

Now,

$$\mathbb{P}(\frac{x-1}{\sqrt{2}} \leq \frac{x'-1}{\sqrt{2}}) = 0.98$$
$$\mathbb{P}(Z \leq z') = 0.98$$

So $\frac{x'-1}{\sqrt{2}} = z'$ which gives us $x' = 1 + z'\sqrt{2}$

6.7.1 Bivariate Normal

We have,

$$f(x, y) = \frac{1}{2\pi\sqrt{1-\rho^2}} exp\left(-\frac{1}{2(1-\rho^2)}(x^2 - 2\rho xy + y^2)\right)$$

Take $x = \begin{bmatrix} x \\ y \end{bmatrix}$. We can take A as $A = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$
Then you can write,

$$f(x, y) = \frac{1}{2\pi\sqrt{\det A}} e^{-\frac{x^T Ax}{2 \det A}}$$

Note. This is the exponential of a quadratic form

Now this can be generalized for any A if A is positive definite so we need $x^T Ax > 0, \forall x \neq 0$ which is the same as saying all eigenvalues of A are positive.

Now if A is symmetric then $A = U^T DU$ where D is a diagonal matrix. So we have,

$$X^T AX = (UX)^T D(UX)$$

If we write $UX = Y$ then we have,

$$Y^T DY = \sum_i \lambda_i y_i^2$$

Remark. Here as U is unitary the Jacobian is 1 and the change of variables is simplified.

Example. If X is a r.v with distribution function $F(x)$?

Let U be a uniform in $[0, 1]$. Claim that $F^{-1}(U)$ has d.f F as $\mathbb{P}(F^{-1}(U) \leq x) = \mathbb{P}(U \leq F(x)) - F(x)$ ◇

Example. If U is uniform in [0, 1] then $-\log(1 - U)$ is exponential with parameter 1. ◇

Example. If U is uniform $[0, 1]$ then,

$$\phi^{-1}(U) \text{ is } N(0, 1)$$

◇

Chapter 7

Moments, and moment generating functions

7.1 Moments

Definition. We define the $k'th$ moment of a r.v X as $\mathbb{E}(X^k)$

Note. The exponential distribution has moments of all orders (so for any $k \geq 1 \in \mathbb{N}$). But for instance the Cuachy distribution does **NOT** have moments for $k \geq 1$.

Theorem 7.1. If all the moments $\mathbb{E}(X), \mathbb{E}(X^2), \ldots$ and the series,

$$\sum_{n=0}^{\infty} \frac{1}{k!} t^k \mathbb{E}(X^k)$$

is absolutely convergent for some $t > 0$. Then the sequence of moments uniquely determines the distribution of X.

7.2 Variance and Covariance

Variance is just $Var(X) = \mathbb{E}([X - \mu]^2)$ where $\mu = \mathbb{E}(X)$. Now this gives us some notion of dispersion from the mean μ.

Note. We can technically "quantify" dispersion in other ways as well like $|X - \mu|$ or $[X - \mu]^3$ but the square is convenient.

Note that we have $\mathbb{E}(X^2) = 0 \Leftrightarrow \mathbb{P}(X = 0) = 1$. This is because we have $\mathbb{E}(X^2) = \sum_0^\infty x^2 \mathbb{P}(X = x)$. But note that x^2 and $\mathbb{P}(X = x)$ is always non-negative. So for $x > 0$ it is necessarily a non-negative product. But if we have $\mathbb{E}(X^2) = 0$ then we need for all $x \geq 1$ the probability to be zero which gives us $\mathbb{P}(X = 0) = 1$ as probability should still add up to 1.

We have,

$$Var(X) = \mathbb{E}([X - \mu]^2) = \mathbb{E}([X - \mu][X - \mu]) = \mathbb{E}(X^2 - 2X\mu + \mu^2) = \mathbb{E}(X^2) - \mu^2 \tag{7.1}$$

Further,

$$Var(aX + b) = a^2 Var(X) \tag{7.2}$$

Now consider the sum of two r.v X, Y we have,

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

We now call the term in the middle as the covariance.

Definition (Covariance)**.** The cov of two r.v. X, Y is $cov(X, Y)$ and given by,

$$cov(X, Y) = \mathbb{E}([X - \mathbb{E}(X)][Y - \mathbb{E}(Y)])$$

Note. We can expand this to get $cov(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$

Note. So if we have $cov(X, Y) = 0$ then we get $Var(X + Y) = Var(X) + Var(Y)$. In such cases we call X, Y to be uncorrelated (note this is **NOT** the same as being independent although being independent implies being uncorrelated)

Remark. Co variance is generally unscaled we can scale it to be between $-1, 1$ by divide by $\sqrt{Var(X), Var(Y)}$

Theorem 7.2 (Cauchy-Schwarz inequality)**.** For X, Y tow r.v we have,

$$[\mathbb{E}(XY)]^2 \leq \mathbb{E}(X^2)\mathbb{E}(Y^2)$$

Proof. Define $Z = sX + Y$ then as $Z^2 \geq 0$ we have,

$$\begin{aligned} 0 \leq \mathbb{E}(Z^2) &= \mathbb{E}(s^2X^2 + 2sXY + Y^2) \\ &= as^2 + bs + c \end{aligned}$$

if we define $a = \mathbb{E}(X^2), b = \mathbb{E}(2XY), c = E(Y^2)$. Now in the above quadratic equation $0 \leq g(s) = as^2 + bs + c$ we have $g(s) = 0$ at most 1 time and for $g(s) = 0$ has at most one real root so we get $b^2 - 4ac \leq 0$ or that,

$$[2\mathbb{E}(XY)]^2 - 4\mathbb{E}(X^2)\mathbb{E}(Y^2) \leq 0$$

□

Remark. Now given two r.v U, V if we set $X = U - \mathbb{E}(U)$ and $Y = V - \mathbb{E}(V)$ in the above equation then we get,

$$cov(U, V)^2 \leq Var(U)Var(V)$$

7.3 Moment Generating functions

We know the probability generating function of X as,

$$G_X(s) = \mathbb{E}(s^X) = \sum s^k \mathbb{P}(X = k)$$

Definition (MGF)**.** The mgf of a r.v X is,

$$M_X(t) = \mathbb{E}(e^{tX})$$

Note. This is basically the same as saying $M_X(t) = G_X(e^t)$

Note. Not every r.v has a moment generating function (for instance Cauchy r.v does not have one) but all r.v do have the characteristic function $\phi_X(t) = \mathbb{E}(e^{itX})$ which is a complex valued function.

Example. For normal dist we have,

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

◇

Example. We can write the mgf as follows,

$$\begin{aligned} M_X(t) = \mathbb{E}(e^{tX}) &= \mathbb{E}(\sum \frac{(tX)^n}{n!}) \\ &= 1 + t\mathbb{E}(X) + \frac{1}{2!}t^2\mathbb{E}(X^2) + \dots \end{aligned}$$

◇

This gives us the following,

Theorem 7.3. If $M_X(t)$ exists for some neighborhood of 0 then for $k = 1, 2\dots$ we have,

$$\mathbb{E}(X^k) = M_X^{(k)}(0)$$

For a linear function of a r.v X we have,

$$M_{aX+b} = \mathbb{E}(e^{t(aX+b)}) = \mathbb{E}(e^{atX}e^{tb}) = e^{tb}\mathbb{E}(e^{(at)X})$$

annd for sums of r.v we have,

Theorem 7.4. For X, Y independent r.v we have,

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

Proof. We have $M_{X+Y}(t) = \mathbb{E}(e^{t(X+Y)}) = \mathbb{E}(e^{tX}e^{tY})$ as X, Y are independent this gives us $\mathbb{E}(e^{tX})\mathbb{E}(e^{tY})$. □

Theorem 7.5. If the moment generating function is finite in some $\delta-$neighborhood of 0, then it uniquely determines a distribution. And we have,

$$M_X(t) = \sum_{k=0}^{\infty} \frac{1}{k!}t^k\mathbb{E}(X^k) \text{ for } |t| < \delta$$

Note. This is basically the Laplace inverse theorem as we have $M_X(t)$ is the Laplace transform of the density function $f_X(x)$.

7.4 Markov and Jenson's inequality

Theorem 7.6 (Markovs)**.** We have for non-negative r.v X,

$$\mathbb{P}(X > t) \leq \frac{\mathbb{E}(X)}{t} \text{ for } t > 0$$

Theorem 7.7 (Jensons Inequality)**.** For X a r.v taking values in (a, b) such that $\mathbb{E}(X)$ exists and $g : (a, b) \to \mathbb{R}$ is a convex function then we have,

$$\mathbb{E}(g(X)) \geq g(\mathbb{E}(X))$$

Note. An intuition for this is for a convex function (if you connect two points, the line segment lies above the graph of the function) the function evaluated at a weighted average is smaller than the weighted average of the function at those two points.

Characteristic functions

Definition. The characeristic function of X is,

$$\phi_X(t) = \mathbb{E}(e^{itX}) \text{ for } t \in \mathbb{R}$$

Remark. Unlike the MGF where we deal with exponential like e^{tX} which is unbounded we have that e^{itX} is bounded and more specifically we have $|e^{itX}| = 1$ which gives us $|\phi_X(t)| \leq 1$.

Example. For normal dist we have,

$$\phi_X(t) = M_X(it) = e^{-\frac{1}{2}t^2}$$

◇

We have similar properties to mgf when it comes to linear functions and sums of r.v.

Theorem 7.8. X, Y have the same distribution if and only if $\phi_X(t) = \phi_Y(t)$.

Theorem 7.9. Let X have characteristic function ϕ and density function f then,

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itx} \phi(t) \, dt$$

Note. Note this is a version of the Fourier inverse transform as the characteristic function is simply a Fourier transform in disguise.

Chapter 8

The main limit theorems

8.1 Law of averages

Consider we have $X_1, \dots, X_n$ that are i.i.d and consider,

$$S_n = \frac{1}{n}(X_1 + \cdots + X_n)$$

note that we have,

$$\mathbb{E}(S_n) = \frac{1}{n}\mathbb{E}(X_1 + \cdots + X_n) = \frac{1}{n}(n\mu) = \mu$$

and,

$$Var(S_n) = \frac{1}{n^2}Var(X_1 + \cdots + X_n) = \frac{\sigma^2}{n}$$

(note that we used independent of X_i to split the variance) and as $n \to \infty$ this goes to 0.

Definition. A sequence of r.v $Z_1, Z_2, \dots$ ***converges in mean square to*** Z if,

$$\mathbb{E}([Z_n - Z]^2) \to 0 \quad \text{as } n \to \infty$$

If this is true we write $Z_n \to Z$ in mean square as $n \to \infty$

Note. Note in the above example we showed the we have $Var(S_n) = \mathbb{E}([S_n - \mathbb{E}(S)]^2) \to 0$ as $n \to \infty$ so we can say that $S_n \to \mathbb{E}(S) = \mu$ in mean square as $n \to \infty$

This note gives us the theorem,

Theorem 8.1. If $X_1, X_2, \dots$ are independent r.v with mean μ and var σ^2 then we have,

$$\frac{1}{n}(X_1 + \cdots + X_n) \to \mu \quad \text{in mean square}$$

8.2 Weak law of large numbers

Definition. The sequence $Z_1, Z_2, \dots$ of r.v converges in probability to Z as $n \to \infty$ if,

$$\forall \varepsilon > 0, \quad \mathbb{P}(|Z_n - Z| > \varepsilon) \to \infty \text{ as } n \to \infty$$

Remark. Convergence in mean square is more powerful than convergence in probability and more specifically implies convergence in probability

Theorem 8.2 (Weak law of large numbers). If $X_1, X_2, \dots$ are independent r.v with μ, σ^2 then,

$$\frac{1}{n}(X_1 + \dots + X_n) \to \mu \quad \text{in probability}$$

Note. The proof is just in mean square and as mean square implies in probability we get the theorem.

8.3 CLT

Consider i.i.d r.v $X_1, X_2 \dots$ now define,

$$S_n = X_1 + \dots + X_n$$

But now consider a standardized version of S_n i.e,

$$\begin{aligned} Z_n &= \frac{S_n - \mathbb{E}(S_n)}{\sqrt{(S_n)}} \\ &= \frac{S_n - n\mu}{\sqrt{n}\sigma} \end{aligned}$$

Theorem 8.3. If $X_1, X_2, \dots$ are independent and identically distributed r.v, with μ, σ^2 then,

$$Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}}$$

where S_n is the sum of the r.v satisfies as $n \to \infty$,

$$\mathbb{P}(Z_n \leq x) \to \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}\, du$$

Note. This is essentially saying that the distribution of the standardized version point wise converges to a normal standard.

Remark. Note that as it's pointwise and not uniform for each x we have a different N where the convergence happens, so the number of samples for a given error bound changes with different values of x.

Theorem 8.4. Given $Z_1, Z_2, \dots$ a sequence or r.v with mgf's then if $n \to \infty$we have,

$$M_n(t) \to e^{\frac{1}{2}t^2} \quad \text{for } t \in \mathbb{R}$$

then we have,

$$\mathbb{P}(Z_n \leq x) \to \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}\, du$$

Note. This is basically saying if the mgf converges to the mgf of a normal standard then the underlying distribution function converges to that of a normal standard as well.

Proof. of CLT
We have,

$$Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} U_i$$

where $U_i = X_i - \mu$. Now note that we have $M_U(t) = \mathbb{E}(e^{t(X-\mu)} = e^{-t\mu}M_X(t)$
So now the mgf of Z_n is,

$$\begin{aligned} M_n(t) = \mathbb{E}(e^{tZ_n})) &= \mathbb{E}\left(e^{\frac{t}{\sigma\sqrt{n}}\sum_{i=1}^{n} U_i}\right) \\ &= \left[\mathbb{E}\left(e^{\frac{t}{\sigma\sqrt{n}}U_i}\right)\right]^n \\ &= \left[M_U\left(\frac{t}{\sigma\sqrt{n}}\right)\right]^n \end{aligned}$$

Now note that we can consider a power series around $x = 0$ to get,

$$\begin{aligned} M_U(x) &= 1 + x\mathbb{E}(U_1) + \frac{1}{2}x^2\mathbb{E}(U_1^2) + o(x^2) \\ &= 1 + \frac{1}{2}\sigma^2 x^2 + o(x^2) \end{aligned}$$

Here $o(x^2)$ means that the terms decrease faster than x^2 (note we're considering a power series about $x = 0$ so $|x| < 1$)

But now we have,

$$M_n(t) = \left[1 + \frac{t^2}{2n} + o\left(\frac{1}{n}\right)\right]^n \to e^{\frac{1}{2}t^2} \qquad \text{as } n \to \infty$$

which by the theorem above gives us that the distribution function converges to that of a normal standard. □

Example (Statistical sampling)**.** One of the main uses of CLT is in stats. Say a fraction of the population are categorized as $X = 1$ (could be say are men) then how many people do we need to survey to estimate p with error not exceeding 0.005. ◇

Solution. Let X_i be the indicated of the $i'th$ person being in the category i.e. $X = 1$. Then we have,

$$S_n = \sum^{n} X_i$$

Now we have the sample mean as $p' = \frac{1}{n}S_n$. So we basically need n large enough so that we have $|p' - p| \leq 0.005$. Now technically we can only do this in probability i.e. $\mathbb{P}(|p' - p| \leq 0.005) \geq 0.95$ or so. But now by CLT we can write,

$$\mathbb{P}(\left|\frac{S_n}{n} - p\right| \leq 0.005) = \mathbb{P}\left(|S_n - \mathbb{E}(S_n)|\frac{1}{\sqrt{var(S_n)}} \leq 0.005\sqrt{\frac{n}{p(1-p)}}\right)$$

But by clt the right side converges to distribution function of a normal standard. So we have approximately $\mathbb{P}(|N| \leq 0.005\sqrt{4n})$ where N is normal with mean 0 and variance 1. So our solution is,

$$\mathbb{P}(|p' - p| \leq 0.005) \geq \int_{-0.005\sqrt{4n}}^{0.005\sqrt{4n}} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2} du$$

Which is $2\phi(0.005\sqrt{4n}) - 1$. And for probabiltiy greater than 0.95 we ned $0.005\sqrt{4n} \geq 1.96$ or that $n \geq 40,000$