MTH 202: Probability and Statistics

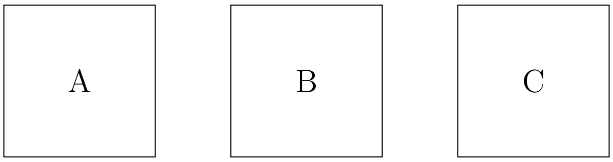
Semester 2, 2023-2024

Prahlad Vaidyanathan

Contents

Two Experiments

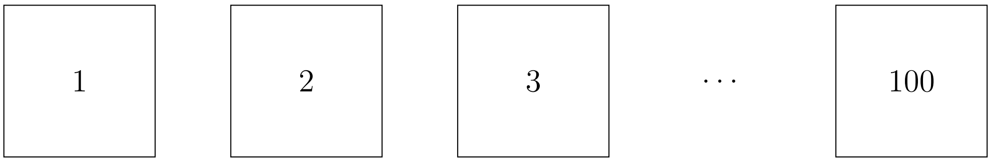
Remark 0.1 (The Monty Hall Problem)**.** Imagine you are on a game show. There are three doors, one with a prize behind it.



You're allowed to pick any door, so you choose the first one at random, door A.

Before opening Door A, the rules of the game require the host to open one of the other doors and let you switch your choice if you want. Because the host doesn't want to give away the game, they always open an empty door. In your case, the host opens door C: no prize, as expected. "Do you want to switch to door B?" the host asks.

Remark 0.2. Now the same problem but with 100 doors.



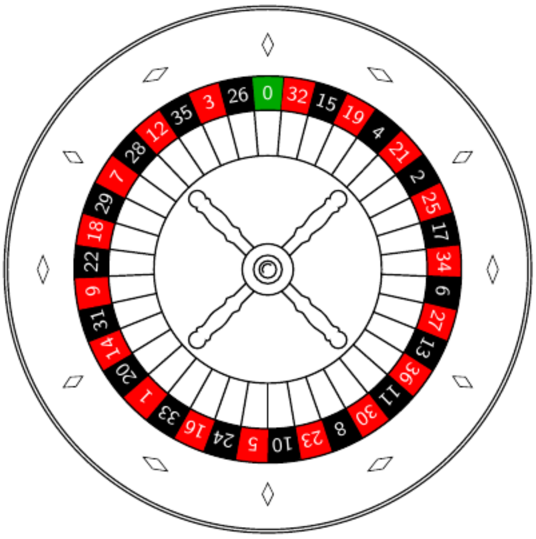
You again pick Door 1. The host now opens all doors except Door 72. Now should you switch to Door 72?

Here, you have to ask yourself: Why did they not open Door 72? Almost certainly because that's where the prize is hidden! Maybe you got really lucky and picked right with the first door at the beginning. But it's way more likely you didn't, and now Door 72 must be the right one!

Solution: In the Monty Hall Problem, it is always correct to switch doors as the chance of your success increases. Indeed, in the first problem, the chance of your success is double if you switch!

We will see why shortly (See Example I.4.7). □

Remark 0.3 (The Gambler's Fallacy)**.** In a game of Roulette, a wheel is spun in one direction and a ball placed on it is spun in the other direction. The ball eventually stops and lands in one of 37 slots on the edge of the wheel. The slot is coloured either red or black (The zero slot is coloured green).



When you reach the table, you are told that the ball has landed in a black slot 26 times in a row. So is the next one likely to be a red?

- Answer 1: Yes, it is likely to be red. 27 blacks in a row is an extremely unlikely outcome.
- Answer 2: No, the next one is likely to be black. The game must be rigged to only give blacks!
- Answer 3: No, the next is equally like to be black or red. Each roll of the wheel is a purely random event, similar to a coin flip. The 26 blacks so far is merely a coincidence.

Solution: Indeed, Answer 3 is correct (See Example I.5.4). □

(End of Day 1)

I. Probability Spaces

1. Examples of Random Phenomena

Definition 1.1. Probability Theory is the study of *random phenomena.*

(i) A phenomenon is an experiment, with some fixed inputs. A deterministic phenomenon is one where the outcome of the experiment can be exactly predicted from the inputs. A random phenomenon is one that is not deterministic.

(ii) Examples:

(a) Addition of two numbers is deterministic.

(b) If a ball is dropped from a height d metres in a vacuum, the time taken to reach the bottom is deterministic.

(c) Flipping a coin is a random phenomenon.

(d) The outcome of an election is random.

(e) Other random phenomena: Card games, DRS system in cricket, effects of climate change (in practice), Expansion of a species in a region, etc.

(iii) A random phenomenon cannot be predicted exactly, but many such phenomena show *statistical regularity.* i.e. If you repeat the experiment many many times, the relative frequency with which a single outcome occurs may be predicted.

(iv) Example: If you toss a coin, the outcome is either H or T. If you toss the coin n times, the numbers

$$N_n(H) := \frac{\text{Number of } H}{n} \text{ and } N_n(T) := \frac{\text{Number of } T}{n}$$

are called the relative frequencies. We know from experience, that

$$N_n(H) \to \frac{1}{2} \text{ as } n \to \infty.$$

The number $\frac{1}{2}$ is thus called the probability that H occurs.

Example 1.2. A twenty-sided die (a $D20$, an icosahedron) is rolled repeatedly. Each time it is rolled, the number is noted down (this is the outcome of the experiment).

(i) This is a random phenomenon, so we we repeat the experiment n times and write the relative frequency of seeing a k (for $1 \leq k \leq 20$) as

$$N_n(k) := \frac{\text{Number of times we see a } k}{n}.$$

For instance, if $n = 10$, we may get

$$1, 5, 7, 18, 11, 12, 14, 17, 1, 18.$$

Then,

$$\begin{aligned} N_{10}(1) &= \frac{2}{10} \\ N_{10}(2) &= 0 \\ &\vdots \\ N_{10}(17) &= \frac{1}{10} \\ N_{10}(18) &= \frac{2}{10} \\ &\vdots \end{aligned}$$

For each $1 \leq k \leq 20$, let $p_k = \frac{1}{20}$. Then we expect that

$$\lim_{n\to\infty} N_n(k) = p_k.$$

Observe that

- $0 \leq p_k \leq 1$.
- $p_1 + p_2 + \ldots + p_{20} = 1$.

(ii) Suppose we now want more information from this experiment. For instance, we may want to know how many even numbers have been rolled. Once again, we may take the relative frequency

$$N_n(\text{even}) := \frac{\text{Number of times an even number is rolled}}{n}.$$

And we may consider

$$p_{\text{even}} := \lim_{n\to\infty} N_n(\text{even}).$$

However, it is clear that

$$\begin{aligned} N_n(\text{even}) &= N_n(2) + N_n(4) + \ldots + N_n(20) \text{ and therefore} \\ p_{\text{even}} &= p_2 + p_4 + \ldots + p_{20} \end{aligned}$$

(iii) Suppose the faces are coloured red (1,2,3,4,8,9,10), blue (5,6,7,11,18,19,20) and yellow (the rest). We may now want to know p_{yellow}. Again, it is clear that

$$p_{\text{yellow}} = \sum_{k \text{ is yellow}} p_k.$$

(iv) In order to make all these experiments more systematic, we introduce the idea the set

$$\Omega := \{w_1, w_2, \dots, w_{20}\}.$$

where each w_k corresponds to the number $k \in \{1, 2, \dots, 20\}$. We now assign

$$w_k \mapsto p_k.$$

Then, for any subset $A \subset \Omega$, we define

$$P(A) := \sum_{k \in A} p_k.$$

This set A might be {the set of even numbers} or {the numbers painted yellow} or anything else. This function

$$P : A \to P(A)$$

is now defined for all subsets of Ω. Therefore, if

$$\begin{aligned} A &= \{\text{the set of even numbers}\} \\ B &= \{\text{the numbers painted yellow}\}, \text{ then} \\ A \cap B &= \{\text{even numbers painted yellow}\}. \end{aligned}$$

We may then ask: What is $P(A \cap B)$? Similarly,

$$B^c = \{\text{the numbers not painted yellow}\}$$

and we can ask: What is $P(B^c)$?

Therefore, this approach lets us ask (and answer) more questions from the same experiments.

(v) Observe that

- $P(\emptyset) = 0$ and $P(\Omega) = 1$.
- If $A \cap B = \emptyset$, then $P(A \cup B) = P(A) + P(B)$.

Example 1.3. Consider a pointer that is free to spin about the centre of a circle of radius 1. If the pointer is spun, it comes to rest at an angle θ (in radians) from the X axis. This is the outcome of the experiment.

(i) Again, this is a random phenomenon. For each $t \in [0, 2\pi)$, θ can take the value t, so we may set

$$\Omega := [0, 2\pi)$$

(ii) Given a point, say $t = 1$, what it the probability that a random spin will come to rest at t? Here, the idea of *relative frequency* is problematic because there are infinitely many choices. Instead we may ask different questions.

(iii) What is the probability that a random spin will come to rest in such a way that θ lies in the upper half-circle? Answer: 1/2 of course. So if

$$A = \{t \in [0, 2\pi) : 0 \leq t \leq \pi\} \Rightarrow P(A) = 1/2$$

Similarly,

$$A = \{t \in [0, 2\pi) : 0 \leq t \leq \pi/2\} \Rightarrow P(A) = 1/4.$$

More generally, if $s \in [0, 2\pi)$ and

$$A_s = \{t \in [0, 2\pi) : 0 \leq t \leq s\} \Rightarrow P(A_s) = \frac{s}{2\pi}$$

The same is true for any sector whose angle is at the origin is s. Hence, if

$$A_{[r,s]} = \{t \in [0, 2\pi) : r \leq t \leq s\} \Rightarrow P(A_{[r,s]}) = \frac{s - r}{2\pi}.$$

(iv) Given a subset $A \subset \Omega$, can we measure $P(A)$? Answer: Not always, but it is possible for many subsets.

(v) If $s \in [0, 2\pi)$ is fixed and $A = \{s\}$, what is $P(A)$?

Solution: Consider

$$A_n = \{t : s - 1/n \leq t \leq s\} \Rightarrow P(A_n) = \frac{1}{2n\pi}.$$

Now, $\{s\} \subset A_n$ so by a theorem we will prove later, it will follow that

$$0 \leq P(\{s\}) = P(A_n).$$

This is true for each $n \in \mathbb{N}$ so $P(\{s\}) = 0$. Hence, the probability of hitting any one angle is zero! □

2. Probability Spaces

Given a random phenomenon, we first need a set Ω consisting of all possible outcomes. For some subsets $A \subset \Omega$, we must assign a probability

$$A \mapsto P(A).$$

Such a subset is called an event. We say that an event A *occurs* if a given outcome w belongs to A. The function P must satisfy three conditions:

- $P(\emptyset) = 0, P(\Omega) = 1$.
- $0 \leq P(A) \leq 1$ for all events A.
- If $A \cap B = \emptyset$, then $P(A \cup B) = P(A) + P(B)$.

It may not be possible to say that

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

for each subset $A \subset \Omega$.

Definition 2.1. Let Ω be a set. A collection $\mathcal{A}$ of subsets of Ω is called a σ-field (or σ-algebra) if the following conditions hold:

(i) $\emptyset \in \mathcal{A}$.

(ii) If $A \in \mathcal{A}$, then $A^c = \Omega \setminus A \in \mathcal{A}$. (Hence, $\Omega \in \mathcal{A}$).

(iii) If $\{A_1, A_2, \ldots\}$ are a countable collection in $\mathcal{A}$, then

$$\bigcup_{n=1}^{\infty} A_n \in \mathcal{A}.$$

Example 2.2.

(i) If Ω is any set, $\mathcal{A} := \{\emptyset, \Omega\}$ is a σ-algebra.

(ii) If Ω is any set, $\mathcal{A} = \mathcal{P}(\Omega)$, the power set, is a σ-algebra.

(iii) If $\Omega = [a, b]$, there is a 'nice' σ-algebra $\mathcal{L}$ that contains all intervals in Ω. This is called the Lebesgue σ-algebra.

Definition 2.3. Let Ω be a set and $\mathcal{A}$ be a σ-algebra on Ω. A probability measure on $(\Omega, \mathcal{A})$ is a function

$$P : \mathcal{A} \to [0, \infty)$$

satisfying the following conditions:

(i) $P(\Omega) = 1$.

(ii) If $\{A_1, A_2, \ldots\} \subset \mathcal{A}$ are mutually disjoint, then

$$P\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} P(A_n).$$

The triple $(\Omega, \mathcal{A}, P)$ is called a probability space.

Example 2.4.

(i) Let Ω be a finite set, $\mathcal{A} = \mathcal{P}(\Omega)$, then we may define

$$P(A) := \frac{|A|}{|\Omega|}$$

where $|\cdot|$ denotes the cardinality. This probability space is called a symmetric probability space.

(ii) Let $\Omega = \{H, T\}$ and $\mathcal{A} = \mathcal{P}(\Omega) = \{\emptyset, \{H\}, \{T\}, \Omega\}$. Define $P : \mathcal{A} \to [0, \infty)$ by

$$P(\emptyset) = 0, P(\{H\}) = \frac{1}{2}, P(\{T\}) = \frac{1}{2}, P(\Omega) = 1.$$

This is an example of a symmetric probability space.

(End of Day 2)

(iii) In the previous example, we may also define

$$P(\emptyset) = 0, P(\{H\}) = \frac{1}{3}, P(\{T\}) = \frac{2}{3}, P(\Omega) = 1.$$

This is a probability space, but is not symmetric.

(iv) Let $\Omega = [a, b]$ and $\mathcal{A} = \mathcal{L}$. There is a function $\mu : \mathcal{L} \to [0, \infty)$ such that

$$\mu((c, d)) = \mu((c, d]) = \mu([c, d)) = |d - c|$$

for all $a \leq c < d \leq b$. The function $P : \mathcal{L} \to [0, \infty)$ given by

$$P(A) := \mu(A)/(b - a)$$

is a probability measure on Ω. This probability space is called the uniform probability space.

(v) If $\Omega \subset \mathbb{R}^d$ is a 'nice' set, we may define a similar σ-algebra $\mathcal{L}$ that contains all rectangles of the form

$$\prod_{i=1}^{d} (a_i, b_i).$$

This is also called the Lebesgue σ-algebra and it also carries a Lebesgue measure μ so that

$$\mu\left(\prod_{i=1}^{d} (a_i, b_i)\right) = \prod_{i=1}^{d} (b_i - a_i).$$

Again, we may define $P : \mathcal{L} \to [0, \infty)$ by

$$P(A) = \frac{\mu(A)}{\mu(\Omega)}.$$

This defines a probability measure and the space is also called a uniform probability space.

Remark 2.5.

(i) If Ω is a finite set, we will typically take $\mathcal{A} = \mathcal{P}(\Omega)$.

(ii) If Ω is infinite, it will almost always occur as a subset of $\mathbb{R}^d$ for $d \in \{1,2,3\}$. In that case, we will usually use some modification of the Lebesgue measure as above.

3. Properties of Probabilities

Lemma 3.1. *Suppose $(\Omega, \mathcal{A}, P)$ is a probability space. Then*

(i) For any $A, B \in \mathcal{A}$,

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

(ii) For any $A \in \mathcal{A}$,

$$P(A^c) = 1 - P(A).$$

(iii) $P(\emptyset) = 0$.

(iv) If $A, B \in \mathcal{A}$ are such that $A \subset B$, then

$$P(A) \leq P(B).$$

(v) If $B \subset A$, then

$$P(A) - P(B) = P(A \setminus B)$$

where $A \setminus B = A \cap B^c$.

(vi) If $\{A_1, A_2, \ldots\}$ is any sequence of sets, then

$$P\left(\bigcup_{n=1}^{\infty}\right) = 1 - P\left(\bigcap_{n=1}^{\infty} A_n^c\right)$$

Proof.

(i) Note that

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

and these two sets are disjoint. Therefore,

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

(ii) Since $A \cup A^c = \Omega$ and $A \cap A^c = \emptyset$, we have

$$P(A) + P(A^c) = P(\Omega) = 1.$$

(iii) Since $\emptyset^c = \Omega$, we have

$$P(\emptyset) + 1 = 1 \Rightarrow P(\emptyset) = 0.$$

(iv) If $A \subset B$, then $A \cap B = A$ so by part (i),

$$P(B) = P(A) + P(B \cap A^c) \geq P(A)$$

because $P(B \cap A^c) \geq 0$.

(v) If $C := A \setminus B$, then $C \cap B = \emptyset$ and $C \sqcup B = A$. Hence,

$$P(A) = P(B) + P(C)$$

as desired.

(vi) This follows from part (ii) and the fact that

$$\left(\bigcup_{n=1}^{\infty} A_n\right)^c = \bigcap_{n=1}^{\infty} A_n^c.$$

by De Morgan's laws.

□

Remark 3.2. Consider part (vi) above: Each A_n represents an event, so if $B := \bigcup_{n=1}^{\infty} A_n$, then

$$P(B) = \text{probability that at least one of these events occur.}$$

Hence,

$$P\left(\bigcap_{n=1}^{\infty} A_n^c\right) = \text{probability that none of these events occur.}$$

Example 3.3. Suppose three perfectly balanced and identical coins are tossed. Find the probability that at least one of them lands heads.

Solution: Each coin C_1, C_2, C_3 lands with two possibilities H, T. So there are 8 possible outcomes of this experiment. In other words, $|\Omega| = 8$. Let A_1 be the event that C_1 lands heads. Let A_2 and A_3 be defined analogously. We are then looking for

$$P(A_1 \cup A_2 \cup A_3).$$

However, $D := A_1^c \cap A_2^c \cap A_3^c$ is the event that none of them lands heads. In other words, each of them lands tails. So, $|D| = 1$ and

$$P(D) = \frac{1}{8}.$$

Thus, $P(A_1 \cup A_2 \cup A_3) = 7/8$. □

(End of Day 3)

From now onwards, $(\Omega, \mathcal{A}, P)$ will denote a fixed probability space. We also write $\sqcup$ to denote disjoint union (implicitly implying that the sets in question are mutually disjoint).

Lemma 3.4.

(i) If A, B are two events (not necessarily mutually disjoint), then

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

(ii) Let $A_1, A_2, \ldots, A_n$ be a finite collection of events (not necessarily mutually disjoint). Then,

$$P\left(\bigcup_{i=1}^n A_i\right) \leq \sum_{i=1}^n P(A_i).$$

Proof.

(i) Write $A = (A \cap B) \cup (A \cap B^c)$, then

$$\begin{aligned} P(A) &= P(A \cap B) + P(A \cap B^c) \\ \Rightarrow P(A) - P(A \cap B) &= P(A \cap B^c) \\ \Rightarrow P(A) + P(B) - P(A \cap B) &= P(B) + P(A \cap B^c). \end{aligned}$$

However, $A \cup B = B \sqcup (A \cap B^c)$, so the result follows.

(ii) We prove this by induction on n.

(a) Suppose $n = 2$, then by part (i),

$$P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2) \leq P(A_1) + P(A_2).$$

(b) Now suppose the result is true for $k = (n-1)$, and consider n sets as above. Write $B_1 := A_1 \cup A_2 \cup \ldots \cup A_{n-1}$ and $B_2 := A_n$. Then,

$$\begin{aligned} P\left(\bigcup_{i=1}^n A_i\right) &= P(B_1 \cup B_2) \\ &\leq P(B_1) + P(B_2) \\ &= P\left(\bigcup_{i=1}^{n-1} A_i\right) + P(A_n) \\ &\leq \sum_{i=1}^{n-1} P(A_i) + P(A_n). \end{aligned}$$

where the last inequality follows by induction. Hence the result.

□

Theorem 3.5. *Let $\{A_1, A_2, \ldots\}$ be a countable collection of events.*

(i) If $A_1 \subset A_2 \subset \ldots$ and $A = \bigcup_{n=1}^{\infty} A_n$, then

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

(ii) If $A_1 \supset A_2 \supset \ldots$ and $A = \bigcap_{n=1}^{\infty} A_n$, then

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

Proof.

(i) Let $B_1 := A_1$ and for $n \geq 2$, set $B_n := A_n \cap A_{n-1}^c$. Then, $\{B_1, B_2, \ldots\}$ are mutually disjoint, so

$$P\left(\bigcup_{n=1}^{\infty} B_n\right) = \sum_{n=1}^{\infty} P(B_n).$$

However, for any fixed $n \in \mathbb{N}$,

$$A_n = \bigsqcup_{i=1}^{n} B_i \Rightarrow P(A_n) = \sum_{i=1}^{n} P(B_i).$$

This last term is the partial sum of the series. Therefore,

$$P(A) = \lim_{n\to\infty} \sum_{i=1}^{n} P(B_n) = \lim_{n\to\infty} P(A_n).$$

(ii) Let $B_n := A_n^c$, then $B_1 \subset B_2 \subset \ldots$. So if $B := \bigcup_{n=1}^{\infty} B_n$, then by part (i),

$$P(B) = \lim_{n\to\infty} P(B_n).$$

However,

$$B^c = \bigcap_{n=1}^{\infty} B_n^c = \bigcap_{n=1}^{\infty} A_n = A.$$

Hence,

$$P(A) = P(B^c) = 1 - P(B) = 1 - \lim_{n\to\infty} P(B_n) = 1 - \lim_{n\to\infty} (1 - P(A_n)) = \lim_{n\to\infty} P(A_n).$$

□

4. Conditional Probability

Example 4.1. Suppose a box has r red balls labelled $1, 2, \ldots, r$, and b black balls labelled $1, 2, \ldots, b$. Assume that the probability of choosing any one ball is

$$\frac{1}{(b+r)}.$$

(In other words, this is a uniform probability). Suppose that a ball is drawn from the box, and that it is known to be red. What is the probability that it is labelled 1?

Solution: The probability space is $\Omega = \{r1, r2, \ldots, rr, b1, b2, \ldots, bb\}$. Consider two events:

$$\begin{aligned} A &:= \text{chosen ball is red} = \{r1, r2, \ldots, rr\} \\ B &:= \text{chosen ball is number 1} = \{r1, b1\} \end{aligned}$$

We are *not* looking for $P(A \cap B)$. Indeed, we are given that A has occurred, and we wish to find the probability that B will occur. □

Definition 4.2. Suppose $P(A) > 0$. The conditional probability of B given A is

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

If $P(A) = 0$, the conditional probability of B given A is undefined.

Remark 4.3. Explanation of Definition 4.2 using relative frequency: Suppose n trials of an experiment are conducted, let

$$\begin{aligned} N_n(A) &:= \text{the number of times } A \text{ occurs} \\ N_n(B) &:= \text{the number of times } B \text{ occurs} \\ N_n(A \cap B) &:= \text{the number of times } A \text{ and } B \text{ both occur} \end{aligned}$$

Then, we expect that

$$\begin{aligned} P(A) &= \lim_{n\to\infty} \frac{N_n(A)}{n} \\ P(B) &= \lim_{n\to\infty} \frac{N_n(B)}{n} \\ P(A \cap B) &= \lim_{n\to\infty} \frac{N_n(A \cap B)}{n} \end{aligned}$$

We are only interested in those experiments where A occurs. Out of these $N_n(A)$ experiments, we wish to count the number of experiments where B also occurs. In other words, we care about the ratio

$$\frac{N_n(A \cap B)}{N_n(A)}$$

Notice that

$$\lim_{n\to\infty} \frac{N_n(A\cap B)}{N_n(A)} = \frac{P(A\cap B)}{P(A)}.$$

Hence the definition.

Example 4.4. In Example 4.1, we have

$$P(B|A) = \frac{P(A\cap B)}{P(A)} = \frac{|A\cap B|}{|A|} = \frac{1}{r}$$

Note that the unconditional probability of B is

$$P(B) = \frac{2}{(b+r)}$$

(End of Day 4)

Note that, by definition,

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

We use this formula repeatedly.

Example 4.5. Suppose that the population of a city is 40% male and 60% female. Also, 50% of the males smoke and 30% of the females smoke. Find the probability that a smoker is male.

Solution: Here, Ω is the set of all people in the city. Consider the following events:

$$\begin{aligned} M &:= \{\text{the person is male}\} \\ F &:= \{\text{the person is female}\} \\ S &:= \{\text{the person smokes}\} \end{aligned}$$

The given information is that

$$\begin{aligned} P(M) &:= 0.4 \\ P(F) &:= 0.6 \\ P(S|M) &:= 0.5 \\ P(S|F) &:= 0.3 \end{aligned}$$

Our question asks us to find

$$P(M|S) = \frac{P(M\cap S)}{P(S)}$$

Note that

$$\begin{aligned} P(M \cap S) &= P(S|M)P(M) = 0.5 \times 0.4 = 0.2 \\ P(S) &= P(S \cap M) + P(S \cap F) \\ P(S \cap F) &= P(S|F)P(F) = (0.6)(0.3) = 0.18 \\ \Rightarrow P(S) &= 0.2 + 0.18 = 0.38 \\ \Rightarrow P(M|S) &= \frac{0.2}{0.38} \end{aligned}$$

□

Observe that

$$P(M|S) = \frac{P(M)P(S|M)}{P(F)P(S|F) + P(M)P(S|M)}$$

Theorem 4.6 (Bayes' Rule)**.** *Let $A_1, A_2, \ldots, A_n$ be n mutually disjoint events with $\Omega = \bigsqcup_{i=1}^{n} A_i$. Let B be an event with $P(B) > 0$. Then,*

$$P(A_i|B) = \frac{P(A_i)P(B|A_i)}{\sum_{k=1}^{n} P(A_k)P(B|A_k)}$$

Proof. We have

$$B = B \cap \left(\bigsqcup_{k=1}^{n} A_k \right) = \bigsqcup_{k=1}^{n} (A_k \cap B)$$

Hence,

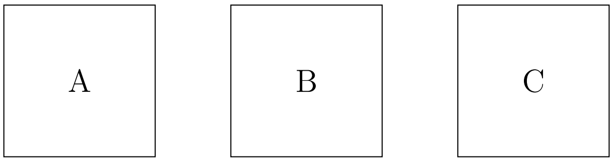
$$P(B) = \sum_{k=1}^{n} P(A_k \cap B) = \sum_{k=1}^{n} P(A_k)P(B|A_k).$$

Moreover,

$$P(A_i \cap B) = P(A_i)P(B|A_i).$$

Taking a ratio gives us Bayes' rule. □

Example 4.7. Consider the Monty Hall Problem from Remark .0.1: Imagine you are on a game show. There are three doors, one with a prize behind it.



You're allowed to pick any door, so you choose the first one at random, door A.

Before opening Door A, the rules of the game require the host (Monty Hall) to open one of the other doors and let you switch your choice if you want. Because the host doesn't want to give away the game, they always open an empty door. In your case, the host opens door C: no prize, as expected. "Do you want to switch to door B?" the host asks.

Solution: Once you have opened Door A, there are four possible outcomes:

$$\begin{aligned} Ab &:= \text{the prize is behind Door A and Monty opened Door B} \\ Ac &:= \text{the prize is behind Door A and Monty opened Door C} \\ Bc &:= \text{the prize is behind Door B and Monty opened Door C} \\ Cb &:= \text{the prize is behind Door C and Monty opened Door b} \end{aligned}$$

Define

$$\begin{aligned} A &:= \text{the event that the prize is behind Door A} = \{Ab, Ac\} \\ B &:= \text{the event that the prize is behind Door B} = \{Bc\} \\ C &:= \text{the event that the prize is behind Door C} = \{Cb\} \end{aligned}$$

Then, it is clear that

$$P(A) = P(B) = P(C) = \frac{1}{3}$$

Moreover,

$$P(\{Ab\}) = P(\{Ac\}) = \frac{P(A)}{2} = \frac{1}{6}.$$

Since Monty opened Door C, the event $D := \{Ac, Bc\}$ has occured. In order to determine whether to switch doors or not, we wish to compute two conditional probabilities:

$$\begin{aligned} P(A|D) &= \frac{P(A \cap D)}{P(D)} \\ &= \frac{P(\{Ac\})}{P(\{Ac\}) + P(\{Bc\})} \\ &= \frac{1/6}{1/6 + 1/3} \\ &= \frac{1}{3} \\ P(B|D) &= \frac{P(\{Bc\})}{1/6 + 1/3} \\ &= \frac{2}{3}. \end{aligned}$$

Hence, it is correct to switch to Door B! □

5. Independence

Definition 5.1. Two events A and B are said to be independent if

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

Remark 5.2.

(i) Note that if A and B are independent, then

$$P(A|B) = P(A) \text{ and } P(B|A) = P(B).$$

Hence, the two events do not affect each other.

(End of Day 5)

(ii) We say that events $\{A_1, A_2, \ldots, A_n\}$ are mutually independent if for any $1 \leq i_1 \leq i_2 \leq \ldots \leq i_k \leq n$, we have

$$P(A_{i_1} \cap A_{i_2} \cap \ldots \cap A_{i_k}) = P(A_{i_1})P(A_{i_2}) \ldots P(A_{i_k}).$$

(iii) Consider $\Omega = \{1, 2, 3, 4\}$ with each point having probability $1/4$. Let

$$A = \{1, 2\}, B = \{1, 3\} \text{ and } C = \{1, 4\}.$$

Then,

$$\begin{aligned} P(A \cap B) &= P(\{1\}) = \frac{1}{4} = P(A)P(B) \\ P(B \cap C) &= P(B)P(C) \\ P(A \cap C) &= P(A)P(C) \\ P(A \cap B \cap C) &= \frac{1}{4} \neq P(A)P(B)P(C). \end{aligned}$$

So the events $\{A, B, C\}$ are pairwise independent but not mutually independent.

If A and B are independent, then A and B^c are also independent because

$$\begin{aligned} P(A)P(B^c) &= P(A)[1 - P(B)] \\ &= P(A) - P(A)P(B) \\ &= P(A) - P(A \cap B) \\ &= P(A \setminus (A \cap B)) \text{ by part (iv) of Lemma 3.1)} \\ &= P(A \cap B^c) \end{aligned}$$

Example 5.3. Consider an experiment where a coin is tossed n times with the condition that

$$P(\{H\}) = p \text{ and } P(\{T\}) = 1 - p.$$

for some fixed $0 \leq p \leq 1$. Construct the corresponding probability space.

Solution:

(i) The experiment has 2^n possible outcomes, so we may take

$$\Omega = \{(x_1, x_2, \ldots, x_n) : x_i \in \{0, 1\} \text{ for all } 1 \leq i \leq n\}.$$

Here, if $x_i = 1$, we think of it as 'Heads' and if $x_i = 0$ we think of it as 'Tails'. The σ-algebra is $\mathcal{A} = \mathcal{P}(\Omega)$. We now wish to assign $P(\cdot)$ to sets of the form

$$\{\overline{x}\}$$

where $\overline{x}$ has k copies of 1 and $(n-k)$ copies of 0. Assume without loss of generality that

$$\overline{x} = (\underbrace{1, 1, \ldots, 1}_{k \text{ times}}, \underbrace{0, 0, \ldots, 0}_{(n-k) \text{ times}})$$

(ii) Let A_i be the event that the i^{th} toss yields a H. Then, the events $\{A_1, A_2, \ldots, A_n\}$ are mutually independent and

$$P(A_i) = p$$

for all $1 \leq i \leq n$. Hence,

$$\begin{aligned} P(\{x\}) &= P(A_1 \cap A_2 \cap \ldots \cap A_k \cap A_{k+1}^c \cap A_{k+2}^c \ldots A_n^c) \\ &= P(A_1)P(A_2) \ldots P(A_k)P(A_{k+1}^c) \ldots P(A_n^c) \\ &= p^k(1-p)^{n-k} \end{aligned}$$

(iii) Let B_k be the event that a H occurs precisely k times, then B_k is made up of all vectors of the same type. Hence,

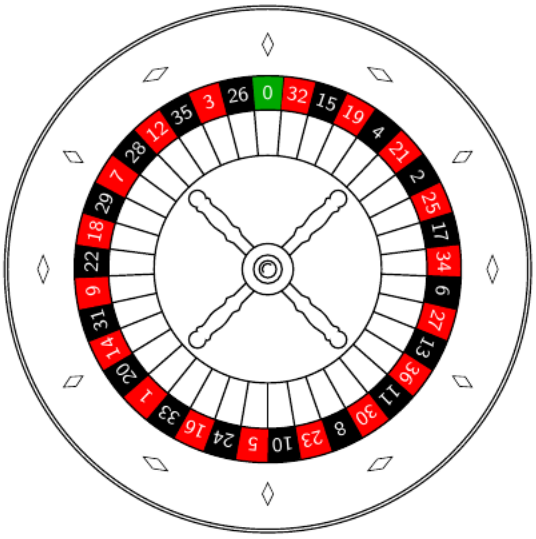
$$P(B_k) = \binom{n}{k} p^k (1-p)^{n-k}.$$

(iv) The events $\{B_0, B_1, \ldots, B_n\}$ are all mutually disjoint and their union is Ω. Therefore, we get

$$1 = P(\Omega) = \sum_{k=0}^{n} P(B_k) = \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k}.$$

□

Example 5.4. Consider the Gambler's Fallacy from Remark .0.3: In a game of Roulette, a wheel is spun in one direction and a ball placed on it is spun in the other direction. The ball eventually stops and lands in one of 37 slots on the edge of the wheel. The slot is coloured either red or black (The zero slot is coloured green).



When you reach the table, you are told that the ball has landed in a black slot 26 times in a row. So is the next one likely to be a red?

- Answer 1: Yes, it is likely to be red. 27 blacks in a row is an extremely unlikely outcome.
- Answer 2: No, the next one is likely to be black. The game must be rigged to only give blacks!
- Answer 3: No, the next is equally like to be black or red. Each roll of the wheel is a purely random event, similar to a coin flip. The 26 blacks so far is merely a coincidence.

Solution: Let A_i be the event that the i^{th} ball lands in a black slot. Then, the A_i are mutually independent and

$$P(A_i) = \frac{18}{37}.$$

Hence,

$$P(A_1 \cap A_2 \cap \ldots \cap A_{27}) = \left(\frac{18}{37}\right)^{27} \approx 3.5 \times 10^{-9}$$

However, this is *not* the probability we are interested in! We are interested in

$$P(A_{27} | \cap_{i=1}^{26} A_i) = P(A_{27}) = \frac{18}{37} \approx 0.48.$$

In other words, Answer 3 is correct. □

II. Combinatorial Analysis

Remark 0.1. Consider the following types of counting problems:

(i) A multiple choice form has 20 questions; each question has 3 choices. In how many possible ways can the exam be completed?

(ii) Consider a horse race with 10 horses. How many ways are there to gamble on the placings (1st, 2nd, 3rd).

(iii) How many different throws are possible with 3 dice? (A 'throw' is an unordered set of the form $\{1, 4, 5\}$).

(iv) Veena has a collection of 20 CDs, she wants to take 3 of them to work. How many possibilities does she have?

Each of these counting problems can be cast as a special case of a single question: Consider an Urn (a bowl) with n different labelled balls in it. Suppose that k balls are to be chosen from this urn and the numbers noted. This can be done in a few different ways:

- Balls can be drawn one-by-one or all k can be drawn together. In the first case, the order in which the balls are chosen can be noted. In the second case, this is not possible.
- Once a ball is drawn, it may be put back into the urn before the second ball is drawn (or it may not be put back). This is called drawing with or without replacement.

(End of Day 6)

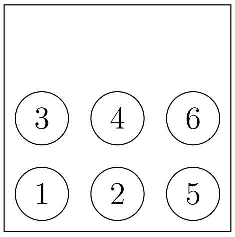


Figure II.1.: Urn with 6 balls

In each case, the k chosen balls are called a sample. We will now write

$$(1,2,3)$$

for an ordered sample, and

$$\{1,2,3\}$$

for an unordered sample. There are now four different experiments, which we enumerate. The examples given above fall into the following categories:

(i) Drawing 20 balls from an urn containing 3 balls, noting the order, with replacement.

(ii) Drawing 3 balls from an urn containing 10 balls, noting the order, without replacement.

(iii) Drawing 3 balls from an urn containing 6 balls, without noting the order, with replacement.

(iv) Drawing 3 balls from an urn containing 20 balls, without noting the order, without replacement.

1. Ordered Sample with Replacement

If there are n balls and a sample of size k is chosen, then we have

$$n^k$$

possible outcomes. In fact, we may take

$$\Omega := \{(x_1, x_2, \dots, x_k) : x_i \in \{1,2,\dots,n\} \text{ for all } 1 \leq i \leq k\}$$

Example 1.1. If $n = 4$ and $k = 3$, the possible outcomes may be given as

$$\begin{gathered}(1,1,1)\\(1,1,2)\\(1,1,3)\\(1,1,4)\\(1,2,1)\\\vdots\\(4,4,1)\\(4,4,2)\\(4,4,3)\\(4,4,4)\end{gathered}$$

There are $4 \times 4 \times 4 = 64$ possible outcomes. Here, Ω would consist of all these 64 points.

2. Ordered Sample without Replacement (Permutations)

Here, each outcome is an *arrangement* of the numbers in a given order. Such an arrangement is called a permutation.

Example 2.1. For instance, if $n = 4$ and $k = 3$, the possible outcomes are:

$$\begin{gathered}(1, 2, 3)\\(1, 2, 4)\\\vdots\end{gathered}$$

Note that $(1, 1, 1)$ is no longer a valid option.

The number of permutations of size k from $\{1, 2, \dots, n\}$ is

$${}^nP_k := n \times (n-1) \times \dots (n-k+1) = \frac{n!}{(n-k)!}$$

3. Unordered Sample without Replacement (Combinations)

Here, each outcome is an unordered subset of $\{1, 2, \dots, n\}$. Such an set is called a combination.

Example 3.1. If $n = 4$ and $k = 3$, the possible outcomes are:

$$\begin{gathered}\{1, 2, 3\}\\\{1, 2, 4\}\\\{1, 3, 4\}\\\{2, 3, 4\}\end{gathered}$$

The number of combinations of size k from $\{1, 2, \dots, n\}$ is

$${}^nC_k = \frac{{}^nP_k}{k!} = \frac{n!}{k!(n-k)!}.$$

4. Unordered Sample with Replacement

Here, each outcome is a *multiset*; a ‘subset’ of $\{1, 2, \dots, n\}$ in which repetitions are allowed.

Definition 4.1. A multiset is a set of the form

$$M = \{(a, m(a)) : a \in A\}$$

where A is a set and $m : A \to \mathbb{N}_0$ is a function. We denote the multiset in the form

$$M = \{1 \cdot a, 3 \cdot b, 6 \cdot c\}$$

Example 4.2. If $n = 4$ and $k = 3$, the possible outcomes are all multisets of the form

$$\{(1, d_1), (2, d_2), (3, d_3), (4, d_4)\}$$

where $d_1 + d_2 + d_3 + d_4 = 3$.

Theorem 4.3. *The total number of distinct k samples from an n-element set such that repetition is allowed and ordering does not matter is*

$$\binom{n+k-1}{k} = \binom{n+k-1}{n-1}$$

Proof. The number we are after is the same as the number of distinct solutions to the equation

$$d_1 + d_2 + \ldots + d_n = k$$

where $d_i \in \{0, 1, 2, \ldots\}$. Given such a tuple $(d_1, d_2, \ldots, d_n)$, we associate to it a list of the form

$$\underbrace{+, +, \ldots, +}_{d_1 \text{ times}}, -, \underbrace{+, +, \ldots, +}_{d_2 \text{ times}}, -, \ldots, -, \underbrace{+, +, \ldots, +}_{d_n \text{ times}}$$

which has precisely $(n-1)$ minus signs. We can think of this problem as having $n+k-1$ positions to fill, of which $(n-1)$ must be '$-$' signs and the remaining are '$+$' signs. There are a total of

$$\binom{n+k-1}{n-1}$$

such solutions. □

(End of Day 7)

We summarize this section in a table: For a sample of size k to be chosen from an n-element set, we have

Ordered Sample with Replacement	n^k
Ordered Sample without Replacement	${}^nP_k$
Unordered Sample with Replacement	$\binom{n+k-1}{k-1}$
Unordered Sample without Replacement	$\binom{n}{k}$

In our examples at the beginning of the section, we have:

(i) A multiple choice form has 20 questions; each question has 3 choices. In how many possible ways can the exam be completed?

Solution: 3^{20}. □

(ii) Consider a horse race with 10 horses. How many ways are there to gamble on the placings (1st, 2nd, 3rd).

Solution: ${}^{10}P_3 = \frac{8!}{5!}$. □

(iii) How many different throws are possible with 3 dice?

Solution: $\binom{6+3-1}{3} = \binom{8}{3}$. □

(iv) Veena has a collection of 20 CDs, she wants to take 3 of them to work. How many possibilities does she have?

Solution: $\binom{20}{3}$. □

5. Examples

Example 5.1 (The Birthday Problem)**.** Assume that people's birthdays occur with equal probability on each of the 365 days of the year. Given a group of n people, find the probability that no two people in the group have the same birthday.

Solution: Here, the k^{th} person has a birthday $b_k \in \{1, 2, \ldots, 365\} =: S$. Hence, the ordered set of all birthdays is given by a tuple

$$(b_1, b_2, \ldots, b_{365})$$

Since birthdays can repeat, the set Ω consists of all such ordered n-samples with replacements, so

$$|\Omega| = 365^n.$$

Let A be the event that no two b_k are equal, then $|A|$ coincides with an ordered n-sample of S without replacement. Hence,

$$|A| = {}^{365}P_n = 365 \times 364 \times \ldots \times (365 - n + 1)$$

Hence,

$$\begin{aligned} P(A) = \frac{|A|}{|\Omega|} &= \frac{365 \times 364 \times \ldots \times (365 - n + 1)}{365 \times 365 \times \ldots \times 365} \\ &= \left(1 - \frac{1}{365}\right)\left(1 - \frac{2}{365}\right) \ldots \left(1 - \frac{n-1}{365}\right) \end{aligned}$$

We know that if $n \geq 366$, then two people *must* share a birthday. However, this tells us that even if $n = 56$,

$$P(A) \approx 0.01,$$

so it is very likely that two people in the group will share a birthday. □

Example 5.2. A deck of playing cards consists of 52 cards - 13 cards in four suits (clubs, spades, hearts, diamonds). The thirteen cards are $2, 3, 4, 5, 6, 7, 8, 9, 10, J, Q, K, A$ where J is Jack, Q is Queen, K is King and A is Ace. Many games are played with these cards; one example is poker, where each player is dealt five cards at random.

(i) How many poker hands are there?

Solution: A hand is sample of 5 cards is an unordered sample chosen without replacement, so the number of hands is

$$\binom{52}{5}.$$

□

(ii) The face value of a card is its number. So, 2 of Spades and 2 of clubs have the same face value. A hand of poker is said to be *four of a kind* if it has four cards of a single face value (and one card is forced to have another face value). What is the probability of a four of a kind?

Solution: Here, Ω consists of $\binom{52}{5}$ poker hands. If A is the event that a hand is a four of a kind, we wish to find

$$P(A) = \frac{|A|}{|\Omega|}.$$

To find $|A|$, there are 13 choices for the face value making up the four, and 48 remaining choices for the fifth card. Therefore,

$$P(A) = \frac{13 \times 48}{\binom{52}{5}}.$$

□

(iii) What is the probability that a poker hand has exactly three clubs?

Solution: Again, we wish to find $|A|$ where A is the event that a hand has exactly three clubs. There are 13 clubs and we wish to choose 3 of them. The remaining two cards are chosen from the remaining $39 = 52 - 13$ cards, so we have

$$|A| = \binom{13}{3} \times \binom{39}{2}.$$

and

$$P(A) = \frac{|A|}{\binom{52}{5}}.$$

□

III. Discrete Random Variables

1. Definitions

Example 1.1. Consider an experiment of tossing a coin 3 times, where $p := P(\{H\})$ and $(1-p) = P(\{T\})$. For each toss, if it lands H, you get ₹1 and if it lands T, you lose ₹1. We wish to know the possible values for the amount of money you would earn, which we denote by X:

w	$X(w)$	$P(\{w\})$
(H,H,H)	3	p^3
(H,H,T)	1	$p^2(1-p)$
(H,T,H)	1	$p^2(1-p)$
(H,T,T)	-1	$p(1-p)^2$
(T,H,H)	1	$p^2(1-p)$
(T,H,T)	-1	$p(1-p)^2$
(T,T,H)	-1	$p(1-p)^2$
(T,T,T)	-3	$(1-p)^3$

Here, we may wish to know the probability that we earn exactly ₹1 or at least ₹1. These events may be described using X as

$$\begin{aligned} A &:= \{w \in \Omega : X(w) = 1\}, \text{ and} \\ B &:= \{w \in \Omega : X(w) \geq 1\}. \end{aligned}$$

Note that

$$\begin{aligned} A &= \{(H,H,T),(H,T,H),(T,H,H)\} \Rightarrow P(A) = 3p^2(1-p) \\ B &= \{(H,H,T),(H,T,H),(THH,HHH\} \Rightarrow P(B) = p^3 + 3p^2(1-p). \end{aligned}$$

We write these events as

$$\{X = 1\} \text{ and } \{X \geq 1\}$$

respectively, and we think of X as a *measurement*.

Definition 1.2. Let $(\Omega, \mathcal{A}, P)$ be a probability space. A discrete random variable is a function

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

such that

(i) The range of X contains either finite or countably many elements.

(ii) For each $x \in \mathbb{R}$, the set $\{w \in \Omega : X(w) = x\}$ is an event (it belongs to $\mathcal{A}$).

(End of Day 8)

Remark 1.3.

(i) For a random variable X, we will often write

$$P(X = x) := P(\{w \in \Omega : X(w) = x\}).$$

(ii) If Ω is a finite set and $\mathcal{A} = \mathcal{P}(\Omega)$, then if $X : \Omega \to \mathbb{R}$ is *any* function,

(a) The range of X is necessarily finite.

(b) For any $x \in \mathbb{R}$,

$$\{w \in \Omega : X(w) = x\} \in \mathcal{A}$$

Hence, in this case, any function $X : \Omega \to \mathbb{R}$ is a discrete random variable.

Definition 1.4. The function $f : \mathbb{R} \to \mathbb{R}$ defined by

$$f(x) := P(X = x)$$

is called the discrete density function of X (also called probability density function or probability mass function). For clarity, this is sometimes denoted f_X as well.

Example 1.5. Consider the random variable in Example 1.1 with $p = 0.4$. Then,

$$\begin{aligned}
f(-3) &= P(X = -3) = P(\{TTT\}) = (1-p)^3 = 0.216 \\
f(-1) &= P(X = -1) = P(\{TTH, THT, HTT\}) = 3p(1-p)^2 = 0.432 \\
f(1) &= P(X = 1) = P(\{HHT, HTH, THH\}) = 3p^2(1-p) = 0.288 \\
f(3) &= P(X = 3) = P(\{HHH\}) = p^3 = 0.064 \\
f(x) &= 0 \text{ if } x \notin \{3, 1, -1, -3\}.
\end{aligned}$$

Example 1.6 (Bernoulli Density)**.** Suppose an experiment is performed whose outcome is classified either as a *success* or as a *failure*. We may take

$$\Omega = \{0, 1\}$$

where 0 denotes failure and 1 denotes success. We may take $\mathcal{A} = \mathcal{P}(\Omega)$. Fix $p \in [0, 1]$ and define $P : \mathcal{A} \to \mathbb{R}$ by

$$P(\{1\}) = p \text{ and } P(\{0\}) = (1-p).$$

Then $(\Omega, \mathcal{A}, P)$ is a probability space.

Let $X : \Omega \to \mathbb{R}$ be the function $X(0) = 0$ And $X(1) = 1$. Then X is a random variable (by part (ii) of Remark 1.3) and

$$\begin{aligned}
&P(X = 1) = p \\
&P(X = 0) = (1 - p) \\
&P(X = x) = 0 \text{ if } x \notin \{0, 1\}.
\end{aligned}$$

Thus, the density function is given by

$$f(x) = \begin{cases} p & : \text{ if } x = 1 \\ (1-p) & : \text{ if } x = 0 \\ 0 & : \text{ otherwise.} \end{cases}$$

This is called the Bernoulli density function or Bernoulli distribution (we will discuss distribution functions later) with parameter p. We will write

$$X \sim \text{Bern}(p)$$

if X is a random variable with probability density function as above.

Example 1.7 (Binomial Density)**.** Consider a success/failure experiment as above with probability of success being p ($p \in [0, 1]$ is fixed). Fix $n \in \mathbb{N}$ and suppose the experiment is repeated n times (as in Example 1.1). Here, we may take

$$\Omega = \{(x_1, x_2, \ldots, x_n) : x_i \in \{0, 1\}\}$$

Moreover, we may take $\mathcal{A} = \mathcal{P}(\Omega)$ and $P : \mathcal{A} \to [0, \infty)$ may be defined on singleton sets as

$$P(\{(x_1, x_2, \ldots, x_n)\}) = p^k(1 - p)^{n-k}$$

where k denotes the number of times 1 occurs in the tuple. Then, $(\Omega, \mathcal{A}, P)$ is a probability space.

Let $X : \Omega \to \mathbb{R}$ denote the random variable

$$X(x_1, x_2, \ldots, x_n) = k$$

where k is as above. Then, X is a random variable (by part (ii) of Remark 1.3) and takes values in $\{0, 1, \ldots, n\}$. Now, for each $0 \leq k \leq n$,

$$|\{(x_1, x_2, \ldots, x_n) : 1 \text{ occurs exactly } k \text{ times}\}| = \binom{n}{k}$$

by section 3. Hence,

$$\begin{aligned}
P(X = k) &= P(\{(x_1, x_2, \ldots, x_n) : 1 \text{ occurs exactly } k \text{ times}\}) \\
&= \binom{n}{k} p^k (1 - p)^{n-k}.
\end{aligned}$$

Hence, the corresponding density function is

$$f(x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & : \text{ if } x \in \{0, 1, \dots, n\} \\ 0 & : \text{ otherwise.} \end{cases}$$

This is called the Binomial density with parameters n and p. Again, we write

$$X \sim B(n, p)$$

when X is a random variable with this density function. Note that

$$\text{Bern}(p) = B(1, p).$$

Example 1.8 (Simplest form of Mendelian inheritance)**.** Suppose that eye colour (brown or blue) is determined by a pair of genes. Suppose that the allele for brown eyes (B) is dominant, while the allele for blue eyes (b) is recessive. So an individual can have one of three possible pairs: purely dominant (BB), hybrid (Bb), or purely recessive (bb). In the first two cases, they will have brown eyes, and if they have *bb* then they will have blue eyes.

Suppose that two hybrid parents have four children. What is the probability that exactly 3 out of 4 children will have brown eyes?

Solution:

(i) For a single child, there are three possible pairs from the set $S = \{BB, Bb, bb\}$. Since the parents are hybrid, the probabilities are

$$P(\{BB\}) = \frac{1}{4}, P(\{Bb\}) = \frac{1}{2} \text{ and } P(\{bb\}) = \frac{1}{4}.$$

Now,

$$\begin{aligned} A &:= \{\text{brown eyes}\} = \{BB, Bb\} \Rightarrow P(A) = \frac{3}{4} \\ B &:= \{\text{blue eyes}\} = \{bb\} \Rightarrow P(B) = \frac{1}{4}. \end{aligned}$$

Since we only care about the colour, we have an experiment whose 'success' probability is 3/4 and 'failure' probability is 1/4.

(ii) In our problem, there are four children, so

$$X \sim B(4, 3/4).$$

Hence,

$$P(X = 3) = \binom{4}{3} \left(\frac{3}{4}\right)^3 \left(\frac{1}{4}\right) = \frac{27}{64}.$$

□

(End of Day 9)

Example 1.9. Consider a population of N balls, of which r are red and $b = (N - r)$ are black. Suppose a random sample of $n \leq N$ is chosen. Thus,

$$|\Omega| = \binom{N}{n}.$$

Let X denote the number of red balls drawn. Then, X takes values $0, 1, \ldots, n$. Then,

$$P(X = k) = \frac{\binom{r}{k}\binom{N-r}{n-k}}{\binom{N}{n}}$$

whenever $0 \leq k \leq n, 0 \leq n - k \leq n$ and 0 otherwise. Therefore,

$$f(x) = \begin{cases} \frac{\binom{r}{x}\binom{N-r}{n-x}}{\binom{N}{n}} & : \text{ if } 0 \leq x \leq n, 0 \leq n - x \leq n \\ 0 & : \text{ otherwise} \end{cases}$$

This is the Hypergeometric density and is denoted

$$\text{Hyp}(N, r, n)$$

Example 1.10 (Constant Random Variable)**.** Given any probability space $(\Omega, \mathcal{A}, P)$ and any constant $c \in \mathbb{R}$, we may define the constant function

$$X : \Omega \to \mathbb{R} \text{ given by } X(w) = c \text{ for all } w \in \Omega.$$

Then, the corresponding density function is

$$f(x) = \begin{cases} 1 & : \text{ if } x = c \\ 0 & : \text{ if } x \neq c \end{cases}$$

Example 1.11 (Indicator Random Variable - Bernoulli revisited)**.** Lett $(\Omega, \mathcal{A}, P)$ be a probability space and $A \in \mathcal{A}$ be a fixed event. Define $X : \Omega \to \mathbb{R}$ by

$$X(w) = \begin{cases} 1 & : \text{ if } w \in A \\ 0 & : \text{ if } w \notin A. \end{cases}$$

Then, X is a discrete random variable with density function given by

$$f(1) = P(X = 1) = P(A), \text{ and } f(0) = P(X = 0) = 1 - P(A).$$

Therefore, if $p := P(A)$, then

$$f(x) = \begin{cases} p & : \text{ if } x = 1 \\ 1 - p & : \text{ if } x = 0 \\ 0 & : \text{otherwise.} \end{cases}$$

Hence, $X \sim \text{Bern}(p)$ from Example 1.6.

Lemma 1.12. *Let X be a discrete random variable and $f : \mathbb{R} \to \mathbb{R}$ denote its density function. We denote the range of X by by $S := \{x_1, x_2, \dots\}$ as it is either finite or countably infinite. Then,*

(i) $f(x) \geq 0$ for all $x \in \mathbb{R}$.

(ii) The set $\{x \in \mathbb{R} : f(x) \neq 0\}$ is either finite or countably infinite.

(iii) $\sum_{i=1}^{\infty} f(x_i) = 1$.

Proof.

(i) If $x \in \mathbb{R}$, then $f(x) = P(X = x) \geq 0$.

(ii) If $f(x) \neq 0$, then $P(X = x) \neq 0$, so x must be in the range of X. Hence,

$$\{x \in \mathbb{R} : f(x) \neq 0\} \subset S.$$

(iii) Consider the set $\{x_1, x_2, \dots\}$, then consider the events

$$A_i = \{w \in \Omega : X(w) = x_i\}$$

If $i \neq j$, then $A_i \cap A_j = \emptyset$. Moreover, if $\omega \in \Omega$, then $X(w) = x_j$ for some $j \in \mathbb{N}$. Hence, $w \in A_j$. Thus,

$$\Omega = \bigsqcup_{i=1}^{\infty} A_i$$

Hence,

$$1 = P(\Omega) = \sum_{i=1}^{\infty} P(A_i) = \sum_{i=1}^{\infty} f(x_i).$$

□

Remark 1.13 (Density function determines the Random Variable)**.** Suppose $f : \mathbb{R} \to \mathbb{R}$ is a function that satisfies the conditions of Lemma 1.12. Then, define

$$\Omega = \{x \in \mathbb{R} : f(x) \neq 0\} = \{x_1, x_2, \dots\}$$

Let $\mathcal{A} = \mathcal{P}(\Omega)$ and define $P : \mathcal{A} \to [0, \infty)$ by

$$P(\{x_i\}) = f(x_i).$$

Then, $(\Omega, \mathcal{A}, P)$ is a probability space. Define $X : \Omega \to \mathbb{R}$ by

$$X(x_i) = x_i.$$

Then, X is a random variable and f is its density function.

Definition 1.14. A function $f : \mathbb{R} \to \mathbb{R}$ is called a discrete density function if it satisfies the three conditions of Lemma 1.12.

Remark 1.15.

(i) The conclusion of Lemma 1.12 and Remark 1.13 is that a discrete random variable X may be used to construct a discrete density function f, and conversely, every discrete density function f may be used to construct a discrete random variable X in such a way that $f_X = f$.

(ii) Also, if we want to describe an experiment with finite (or countably infinite) outcomes $\{x_1, x_2, \ldots\}$, we may simply be given a discrete density function $f : \mathbb{R} \to \mathbb{R}$ whose possible values are $\{x_1, x_2, \ldots\}$. Then, we may take

$$\begin{aligned}\Omega &:= \{x_1, x_2, \ldots\}\\ \mathcal{A} &:= \mathcal{P}(\Omega)\\ P(\{x_i\}) &:= f(x_i).\end{aligned}$$

This defines a probability space $(\Omega, \mathcal{A}, P)$. Moreover, the random variable $X : \Omega \to \mathbb{R}$ given by inclusion

$$X(x_i) = x_i$$

is naturally associated to f. Thus, the probability space is somewhat *unnecessary* if you already have a density function.

(End of Day 10)

Example 1.16 (Uniform Density)**.** Consider the experiment of picking a point at random from a finite set $S = \{a_1, a_2, \ldots, a_n\} \subset \mathbb{R}$ in a way that each point has equal probability. Here, there is a natural density function

$$f(x) = \begin{cases} \frac{1}{n} & : \text{ if } x \in S \\ 0 & : \text{ otherwise.} \end{cases}$$

This density function describes an experiment and a random variable, called the Uniform density which we denote by $U(n)$.

Example 1.17 (Geometric Density)**.** Consider an experiment with two outcomes as in Example 1.6. We repeat this experiment and each repetition is independent. Let X be the random variable that counts the number of failures before the first success. Then

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

and so on. Therefore, if $f : \mathbb{R} \to \mathbb{R}$ denotes the corresponding density function, then

$$f(x) = \begin{cases} p(1-p)^{x-1} & : \text{ if } x \in \mathbb{N} \\ 0 & : \text{ otherwise.} \end{cases}$$

This is called the Geometric Density with parameter p. We write

$$X \sim \text{Geom}(p).$$

Example 1.18 (Poisson Density)**.** Let $\lambda > 0$ be fixed.

$$f(x) = \begin{cases} e^{-\lambda}\frac{\lambda^x}{x!} & : \text{ if } x \in \{0,1,2,\ldots\} \\ 0 & : \text{ otherwise.} \end{cases}$$

Note that f satisfies the following conditions:

(i) $f(x) \geq 0$ for all $x \in \mathbb{R}$.

(ii) The set $\{x \in \mathbb{R} : f(x) \neq 0\}$ is countable.

(iii) Finally,

$$\sum_{x=0}^{\infty} f(x) = e^{-\lambda}\sum_{x=0}^{\infty}\frac{\lambda^x}{x!} = e^{-\lambda}e^{\lambda} = 1.$$

By Lemma 1.12, f is a density function, called the Poisson density with parameter λ. We write

$$X \sim \mathrm{Po}(\lambda)$$

Poisson random variables are important in applications (such as understanding any system with a 'queue'). We will discuss these later.

2. Computations with Densities

Remark 2.1. Let X be a discrete random variable on a probability space $(\Omega, \mathcal{A}, P)$. The density function of X is given by the formula

$$f(x) := P(X = x)$$

In other words, it calculates the probability of the event $B = \{w \in \Omega : X(w) = x\}$. However, we may be interested in other similar events. For instance, suppose X takes values in $\{0,1,2,\ldots\}$, then we may be interested in

$$P(\{w \in \Omega : X(w) \leq 5\}) = \sum_{i=0}^{5} P(X = i) = \sum_{i=0}^{5} f(i).$$

In general, it is such events that we are often interested in.

Definition 2.2. The distribution function of a discrete random variable X is given by

$$F(t) = P(X \leq t) = \sum_{x \leq t} f(x)$$

Remark 2.3. Some immediate observations:

(i) If $s \leq t$, then $\{w \in \Omega : X(w) \leq s\} \subset \{w \in \Omega : X(w) \leq t\}$. Hence,

$$F(s) \leq F(t)$$

so F is *non-decreasing.*

(ii) Suppose the density function f takes finitely many values $\{x_1, x_2, \ldots, x_k\}$ and we list then in increasing order $x_1 < x_2 < \ldots < x_k$. Then,

$$F(t) = \begin{cases} 0 & : \text{ if } t < x_1. \\ f(x_1) & : \text{ if } x_1 \leq t < x_2 \\ f(x_1) + f(x_2) & : \text{ if } x_2 \leq t < x_3 \\ \vdots & \vdots \\ f(x_1) + f(x_2) + \ldots + f(x_{k-1}) & : \text{ if } x_{k-1} \leq t < x_k \\ 1 & : \text{ if } x_k \leq t. \end{cases}$$

Hence, for each interval $[x_i, x_{i+1})$, F is constant, and there is a *jump* of $f(x_j)$ are the point x_j.

(iii) If $(a, b]$ is an interval in $\mathbb{R}$, then

$$\{w \in \Omega : X(w) \leq b\} = \{w \in \Omega : X(w) \leq a\} \sqcup \{w \in \Omega : a < X(w) \leq b\}$$

Hence, by part (v) of Lemma I.3.1,

$$P(\{w \in \Omega : a < X(w) \leq b\}) = P(a < X \leq b) = F(b) - F(a).$$

Example 2.4. Let X be the number rolled on a 20-sided dice. What is the probability that the number rolled is at least 10?

Solution: Let $\Omega = \{1, 2, \ldots, 20\}$ and $X : \Omega \to \mathbb{R}$ be uniformly distributed on Ω. i.e., the density function is given by

$$f(x) = \begin{cases} \frac{1}{20} & : x \in \{1, 2, \ldots, 20\} \\ 0 & : \text{ otherwise.} \end{cases}$$

The distribution function F is given by

$$F(t) = \sum_{x=0}^{[t]} \frac{1}{20} = \frac{[t]}{20}.$$

Hence,

$$P(X \leq 9) = F(9) = \frac{9}{20} \Rightarrow P(X \geq 10) = \frac{11}{20}.$$

□

(End of Day 11)

Example 2.5. Suppose $X \sim \text{Geom}(p)$ (see Example 1.17), so its density function is given by

$$f(x) = \begin{cases} p(1-p)^{x-1} & : x \in \{1, \ldots\} \\ 0 & : \text{ otherwise} \end{cases}$$

(i) We compute the distribution function for X: If $t < 1$, then $F(t) = 0$, and if $t \geq 1$, then

$$F(t) = \sum_{x=1}^{[t]} p(1-p)^{x-1} = p\frac{1-(1-p)^{[t]}}{1-(1-p)} = 1-(1-p)^{[t]}.$$

(ii) We compute $P(X > n)$ when $n \in \mathbb{N}$: Observe that

$$P(X > n) = 1 - P(X \leq n) = 1 - F(n) = (1-p)^n$$

(iii) For each $n \in \mathbb{N}$, consider

$$A_n = \{w \in \Omega : X(w) > n\}.$$

Then, $P(A_n) = P(X > n)$ is the probability that the one fails at least n times before the first success. Note that $A_n \subset A_{n-1}$ in general. Now fix $n, m \in \mathbb{N}$ and consider the conditional probability

$$\begin{aligned} P(X > n+m | X > n) = P(A_{n+m}|A_n) &= \frac{P(A_{n+m} \cap A_n)}{P(A_n)} \\ &= \frac{P(A_{n+m})}{P(A_n)} \\ &= \frac{P(X > n+m)}{P(X > n)} \\ &= \frac{(1-p)^{n+m}}{(1-p)^n} \\ &= (1-p)^m \\ &= P(X > m) \end{aligned}$$

This has a practical implication: If you know that the machine has failed n times, then the probability that it fails another m times, is the same as the unconditional probability that it fails m times. In other words, the machine has *no memory*. The property

$$P(X > n+m|X > n) = P(X > m)$$

is called the memoryless property of the Geometric density.

3. Discrete Random Vectors

Example 3.1. For a single experiment, one may be interested in many different random variables at the same time. For instance, if an urn has n labelled balls, both black and red, and you select k of them. We may be interested in

$$\begin{aligned}
X &:= \{\text{the number of red balls chosen}\} \\
Y &:= \{\text{the minimum number (label) on the balls chosen}\} \\
Z &:= \{\text{the maximum number (label) on the balls chosen}\}.
\end{aligned}$$

In some cases, we may wish to compare random variables, which leads to the next definition.

Definition 3.2. Let $(\Omega, \mathcal{A}, P)$ be a probability space, and let $X_1, X_2, \ldots, X_r$ be r discrete random variables on it. Define $Y : \Omega \to \mathbb{R}^r$ by

$$Y(w) := (X_1(w), X_2(w), \ldots, X_r(w))$$

Such a function is called an <u>r-dimensional random vector</u>.

Definition 3.3. Given such a random vector, the associated <u>density function</u> is given by $f : \mathbb{R}^r \to \mathbb{R}$ as

$$\begin{aligned}
f(\overline{x}) &= P(\{w \in \Omega : X_1(w) = x_1, X_2(w) = x_2, \ldots, X_r(w) = x_r\} = P(X_1 = x_1, X_2 = x_2, \ldots, X_r = x_r) \\
&= P(Y = \overline{x})
\end{aligned}$$

Example 3.4. Suppose two six-sided dice are thrown simultaneously.

(i) Let X denote the number on the first die, and Y the number on the second die. Then,

$$\Omega = \{(i, j) : 1 \leq i, j \leq 6\}$$

Now,

$$P(X = 1, Y = 1) = \frac{|\{(1,1)\}|}{|\Omega|} = \frac{1}{36}.$$

Thus, if $f : \mathbb{R}^2 \to \mathbb{R}$ denotes the joint density function, then

$$f(x, y) = \begin{cases} \frac{1}{36} & : \text{ if } 1 \leq x \leq 6 \text{ and } 1 \leq y \leq 6 \\ 0 & : \text{ otherwise.} \end{cases}$$

(ii) Let X be the number on the first die, and Z be the larger of the two numbers. Then,

$$P(X = 1, Z = 1) = \frac{1}{36} \text{ but } P(X = 2, Z = 1) = 0.$$

Hence, the joint density function $f : \mathbb{R}^2 \to \mathbb{R}$ is given the following table (values are multiplied by 36):

z,x	1	2	3	4	5	6
1	1	0	0	0	0	0
2	1	2	0	0	0	0
3	1	1	3	0	0	0
4	1	1	1	4	0	0
5	1	1	1	1	5	0
6	1	1	1	1	1	6

In other words,

$$f(x,z) = \begin{cases} \frac{1}{36} & : \text{if } 1 \leq x \leq 6, x < z \leq 6 \\ \frac{x}{36} & : \text{if } 1 \leq x \leq 6, z = x \\ 0 & : \text{otherwise} \end{cases}$$

(End of Day 12)

Remark 3.5.

(i) The density function defined as above has the following properties:

(a) $f(\overline{x}) \geq 0$ for all $\overline{x} \in \mathbb{R}^r$.

(b) $S := \{\overline{x} \in \mathbb{R}^r : f(\overline{x}) \neq 0\}$ is finite or countably infinite.

(c) $\sum_{\overline{x} \in S} f(\overline{x}) = 1$.

(ii) Any function $f : \mathbb{R}^r \to \mathbb{R}$ satisfying these three conditions is called a discrete r-dimensional density It is also called the joint density function of the variables $(X_1, X_2, \ldots, X_r)$.

(iii) As Remark 1.13, any such function is the density function of some r dimensional random vector.

(iv) The 1-dimensional density function associated to the random variable X_i is then called the marginal density function and is denoted by f_{X_i}.

(v) Given two random variables X, Y, we write

$$\{X = x, Y = y\} := \{w \in \Omega : X(w) = x \text{ and } Y(w) = y\}$$

Lemma 3.6. *Let X and Y be two random variables with joint density function $f : \mathbb{R}^2 \to \mathbb{R}$. If f_X and f_Y denote the marginal density functions of X and Y respectively, then*

$$f_X(x) = \sum_{y \in \mathbb{R}} f(x,y)$$
$$f_Y(y) = \sum_{x \in \mathbb{R}} f(x,y).$$

In both cases, the sum is over a finite or countable set.

Proof. Suppose X takes values $\{x_1, x_2, \dots\}$, then for any $y \in \mathbb{R}$

$$\begin{aligned}
P(Y = y) &= P(\{w \in \Omega : Y(w) = y\}) \\
&= P(\bigsqcup_{i=1}^{\infty} \{w \in \Omega : X(w) = x_i \text{ and } Y(w) = y\}) \\
&= \sum_{i=1}^{\infty} P(X = x_i, Y = y).
\end{aligned}$$

Hence, if f is the joint density function of X and Y, then the marginal density of Y is given by

$$f_Y(y) = \sum_{i=1}^{\infty} f(x_i, y)$$

Similarly, if Y takes values $\{y_1, y_2, \dots\}$, then the marginal density of X is given by

$$f_X(x) = \sum_{j=1}^{\infty} f(x, y_j)$$

□

Example 3.7. Consider part (ii) Example 3.4 where X is the number on the first die while Z is the maximum of the two numbers. Then, we can represent the probabilities (multiplied by 36) as a table:

z,x	1	2	3	4	5	6
1	1	0	0	0	0	0
2	1	2	0	0	0	0
3	1	1	3	0	0	0
4	1	1	1	4	0	0
5	1	1	1	1	5	0
6	1	1	1	1	1	6

The marginal density of X is given by

$$\begin{aligned}
f_X(1) &= \sum_{z=1}^{6} f(1, z) = \frac{6}{36} = \frac{1}{6} \\
f_X(2) &= \sum_{z=1}^{6} f(2, z) = \frac{6}{36} = \frac{1}{6}
\end{aligned}$$

and so on. Therefore,

$$f_X(x) = \begin{cases} \frac{1}{6} & : \text{ if } 1 \leq x \leq 6 \\ 0 & : \text{ otherwise.} \end{cases}$$

For Z, the marginal density is given by

$$f_Z(z) = \begin{cases} \frac{1}{36} & : \text{ if } z = 1 \\ \frac{3}{36} & : \text{ if } z = 2 \\ \frac{5}{36} & : \text{ if } z = 3 \\ \frac{7}{36} & : \text{ if } z = 4 \\ \frac{9}{36} & : \text{ if } z = 5 \\ \frac{11}{36} & : \text{ if } z = 6 \\ 0 & : \text{ otherwise} \end{cases}$$

4. Independent Random Variables

Definition 4.1.

(i) Two random variables X and Y are said to be independent if for any $x, y \in \mathbb{R}$

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

In other words, the joint density function is given by

$$f(x, y) = f_X(x)f_Y(y).$$

(ii) A collection $X_1, X_2, \dots, X_r$ of random variables are said to be mutually independent if the joint density function is given by

$$f(x_1, x_2, \dots, x_r) = f_{X_1}(x_1)f_{X_2}(x_2) \dots f_{X_r}(x_r).$$

Example 4.2. Suppose two six-sided dice are rolled. Let X denote the number on the first die, Y the number on the second die, and Z the maximum of the two numbers.

(i) For any $1 \leq x, y \leq 6$,

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

This is also true if $x \notin \{1, 2, \dots, 6\}$ or $y \notin \{1, 2, \dots, 6\}$, so X and Y are independent.

(ii) However,

$$P(X = 1, Z = 2) = \frac{2}{36} \text{ while } P(X = 1) = \frac{1}{6} \text{ and } P(Z = 2) = \frac{3}{36}.$$

Hence, X and Z are not independent.

Remark 4.3.

(i) If X is a random variable with density function f, and if $A \subset \mathbb{R}$ then

$$P(X \in A) = P(\{w \in \Omega : X(w) \in A\}) = \sum_{x \in A} P(X = x) = \sum_{x \in A} f(x)$$

(ii) If X and Y are two random variables with joint density function $f_{X,Y}$ and if $A, B \subset \mathbb{R}$ are two sets, then

$$\begin{aligned} P(X \in A, Y \in B) &= P(\{w \in \Omega : X(w) \in A \text{ and } Y(w) \in B\}) \\ &= \sum_{x \in A} \sum_{y \in B} f_{X,Y}(x, y) \end{aligned}$$

Lemma 4.4. *If X, Y are independent random variables and A and B are two subsets of $\mathbb{R}$, then*

$$P(X \in A, Y \in B) = P(X \in A)P(Y \in B).$$

Proof. Suppose $f_{X,Y}$ denotes the joint density and f_X and f_Y denote the marginal densities. Note that

$$\begin{aligned} P(X \in A, Y \in B) &= \sum_{x \in A} \sum_{y \in B} f_{X,Y}(x, y) \\ &= \sum_{x \in A} \sum_{y \in B} f_X(x) f_Y(y) \\ &= \left(\sum_{x \in A} f_X(x) \right) \left(\sum_{y \in B} f_Y(y) \right) \\ &= P(X \in A)P(Y \in B). \end{aligned}$$

□

Remark 4.5.

(i) Let X, Y be two random variables on a probability space $(\Omega, \mathcal{A}, P)$. Then, we may define a number of new random variables:

(a) Define $Z : \Omega \to \mathbb{R}$ by

$$Z(w) := X(w) + Y(w).$$

Predictably, we write $(X + Y)$ for this random variable.

(b) Define $Z : \Omega \to \mathbb{R}$ by

$$Z(w) := \min\{X(w), Y(w)\}.$$

Again, we write $\min\{X, Y\}$ for this random variable.

(c) Similarly, we define $\max\{X, Y\}$ as well.

In general, given a 'nice' function $g : \mathbb{R}^2 \to \mathbb{R}$, we may define $Z : \Omega \to \mathbb{R}$ by

$$Z(w) := g(X(w), Y(w)).$$

This is a discrete random variable, and is denoted $g(X, Y)$.

(ii) More generally, given r random variables $X_1, X_2, \ldots, X_r$ and a 'nice' function $g : \mathbb{R}^r \to \mathbb{R}$, we may define $Z : \Omega \to \mathbb{R}$ by

$$Z(w) := g(X_1(w), X_2(w), \ldots, X_r(w)).$$

This is a discrete random variable and is written as $g(X_1, X_2, \ldots, X_r)$.

(End of Day 13)

5. Sums of Independent Random Variables

Remark 5.1. Suppose X and Y are independent random variables and $Z = X + Y$. Let $\{x_1, x_2, \ldots\}$ be the range of X, then for any $z \in \mathbb{R}$,m

$$\{w \in \Omega : Z(w) = z\} = \bigsqcup_{i=1}^{\infty} \{w \in \Omega : X(w) = x_i, \text{ and } Y(w) = z - x_i\}$$

Since X and Y are independent,

$$P(Z = z) = \sum_{i=1}^{\infty} P(X = x_i)P(Y = z - x_i) = \sum_{i=1}^{\infty} f_X(x_i) f_Y(z - x_i).$$

Hence,

$$f_{X+Y}(z) = \sum f_X(x) f_Y(z - x).$$

This is called the convolution of the two functions f_X and f_Y. In particular, if X and Y both take only non-negative integer values, then $X + Y$ only takes non-negative integer values and for $z \geq 0$,

$$f_{X+Y}(z) = \sum_{x=0}^{z} f_X(x) f_Y(z - x).$$

The next definition only makes sense for integer valued random variables.

Definition 5.2. Let X be an integer valued discrete random variable. The <u>probability generating functi</u> of X is $\Phi_X : [-1, 1] \to \mathbb{R}$ given by

$$\Phi_X(t) := \sum_{x=0}^{\infty} f_X(x) t^x$$

Remark 5.3.

(i) Note that the series converges absolutely because $\sum_{x=0}^{\infty} f_X(x) = 1$. Therefore, Φ_X is differentiable and

$$\Phi_X'(t) = \sum_{x=1}^{\infty} x f_X(x) t^{x-1}$$

This can be done repeatedly.

(ii) Observe that

$$\begin{aligned}\Phi_X(0) &= P(X = 0)\\ \Phi_X'(0) &= f_X(1) = P(X = 1)\\ \Phi_X''(0) &= 2f_X(2) = 2P(X = 2)\end{aligned}$$

and so on. In general,

$$P(X = k) = \frac{\Phi_X^{(k)}(0)}{k!}.$$

(iii) If X and Y are two random variables as above such that $\Phi_X = \Phi_Y$ as functions, then

$$P(X = 0) = \Phi_X(0) = \Phi_Y(0) = P(Y = 0).$$

Differentiating once, we see that

$$P(X = 1) = P(Y = 1).$$

And more generally, $P(X = k) = P(Y = k)$ for all $k \in \{0, 1, 2, \ldots\}$. Hence, X and Y are equivalent random variables. In particular, X and Y have the same distribution function.

Example 5.4. If $X \sim B(n, p)$, then

$$f_X(x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & : x \in \{0, 1, 2, \ldots, n\} \\ 0 & : \text{ otherwise.} \end{cases}$$

Hence,

$$\begin{aligned}\Phi_X(t) &= \sum_{x=0}^{n} \binom{n}{x} (pt)^x (1-p)^{n-x}\\ &= (tp + 1 - p)^n\end{aligned}$$

Example 5.5. If $X \sim \text{Po}(\lambda)$ for some $\lambda > 0$, then

$$f_X(x) = \begin{cases} e^{-\lambda} \frac{\lambda^x}{x!} & : x \in \{0, 1, 2, \ldots\} \\ 0 & : \text{ otherwise} \end{cases}$$

Hence,

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

Lemma 5.6. *If $X_1, X_2, \ldots, X_r$ are mutually independent, integer-valued random variables, then*

$$\Phi_{X_1+X_2+\ldots+X_r}(t) = \Phi_{X_1}(t)\Phi_{X_2}(t)\ldots\Phi_{X_r}(t)$$

for all $t \in [-1, 1]$.

Proof. Assume $r = 2$ and $X = X_1$ and $Y = X_2$. Fix $t \in [-1, 1]$. Then

$$\begin{aligned}\Phi_{X+Y}(t) &= \sum_{z=0}^{\infty} f_{X+Y}(z)t^z \\ &= \sum_{z=0}^{\infty}\sum_{x=0}^{z} f_X(x)f_Y(z-x)t^z \\ &= \sum_{x=0}^{\infty} f_X(x)t^x \sum_{z=x}^{\infty} f_Y(z-x)t^{z-x} \\ &= \left(\sum_{x=0}^{\infty} f_X(x)t^x\right)\left(\sum_{y=0}^{\infty} f_Y(y)t^y\right) \\ &= \Phi_X(t)\Phi_Y(t).\end{aligned}$$

□

Theorem 5.7. *Let $X_1, X_2, \ldots, X_r$ be independent random variables.*

(i) If $X_i \sim B(n_i, p)$ for some $n_1, n_2, \ldots, n_r \in \mathbb{N}$ and $0 \le p \le 1$, then

$$X_1 + X_2 + \ldots + X_r \sim B(n_1 + n_2 + \ldots + n_r, p).$$

(ii) If $X_i \sim Po(\lambda_i)$ for some $\lambda_1, \lambda_2, \ldots, \lambda_r > 0$, then

$$X_1 + X_2 + \ldots + X_r \sim Po(\lambda_1 + \lambda_2 + \ldots + \lambda_r).$$

Proof.

(i) If $X_i \sim B(n_i, p)$, then

$$\Phi_{X_i}(t) = (pt + 1 - p)^{n_i}.$$

By Lemma 5.6,

$$\Phi_{X_1+X_2+\ldots+X_r}(t) = (pt + 1 - p)^{n_1+n_2+\ldots+n_r}.$$

By part (iii) of Remark 5.3,

$$X_1 + X_2 + \ldots + X_r \sim B(n_1 + n_2 + \ldots + n_r, p).$$

(ii) Similar (check!)

□

(End of Day 14)

IV. Expectation of Discrete Random Variables

1. Definition of Expectation

Definition 1.1. Let X be a discrete random variable on a probability space $(\Omega, \mathcal{A}, P)$ with range $\{x_1, x_2, \ldots\}$ and density function f. Suppose that

$$\sum_{j=1}^{\infty} |x_j| f(x_j) < \infty$$

Then, the expectation or mean of X is defined as

$$EX := \sum_{j=1}^{\infty} x_j f(x_j).$$

Note that if $\sum_{j=1}^{\infty} |x_j| f(x_j) = +\infty$, then the expectation is not defined.

We think of this as a weighted average of all the values X takes.

Example 1.2. Suppose $\Omega = \{0, 1\}$ and $X \sim \text{Bern}(p)$ for some $0 \leq p \leq 1$. Then, X takes two values $\{0, 1\}$ and

$$EX = 0P(X = 0) + 1P(X = 1) = P(X = 1) = p.$$

Example 1.3. Suppose $\Omega = \{(x_1, x_2, \ldots, x_n) : x_i \in \{0, 1\}\}$ and $X : \Omega \to \mathbb{R}$ is $X \sim B(n, p)$ for some $0 \leq p \leq 1$. Then the density function is given by

$$f(x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & : x \in \{0, 1, \ldots, n\} \\ 0 & : \text{ otherwise.} \end{cases}$$

Therefore,

$$EX = \sum_{x=0}^{n} x \binom{n}{x} p^x (1-p)^{n-x}$$

We use the fact that for all $j \geq 1$,

$$j \binom{n}{j} = n \binom{n-1}{j-1}.$$

Therefore,

$$
\begin{aligned}
EX &= \sum_{x=1}^{n} n\binom{n-1}{x-1} p^x (1-p)^{n-x} \\
&= n\sum_{i=0}^{n-1} \binom{n-1}{i} p^{i+1}(1-p)^{n-i-1} \\
&= np(p+1-p)^{n-1} \\
&= np.
\end{aligned}
$$

Example 1.4. Suppose $X \sim U(n)$. In other words, the density function of X is

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

Then,

$$
EX = \sum_{i=1}^{n} \frac{i}{n} = \frac{n+1}{2}.
$$

Example 1.5. Suppose $X \sim \text{Geom}(p)$ for some $0 < p < 1$, then

$$
f(x) = \begin{cases} p(1-p)^{x-1} & : \text{ if } x \in \{1, 2, \ldots\} \\ 0 & : \text{ otherwise.} \end{cases}
$$

We verify that

$$
\sum_{i=1}^{\infty} if(i) < \infty
$$

To see this, observe that

$$
\begin{aligned}
\sum_{i=1}^{\infty} if(i) &= \sum_{i=1}^{\infty} ip(1-p)^{i-1} \\
&= p\sum_{i=1}^{\infty} i(1-p)^{i-1} \\
&= -p\left(\frac{d}{dx}\sum_{i=1}^{\infty}(1-x)^i|_{x=p} - 1\right) \\
&= -p\left(\frac{d}{dx}\frac{1}{x}|_{x=p} - 1\right) \\
&= -p\left(\frac{-1}{p^2} - 1\right) \\
&= \frac{1-p}{p}.
\end{aligned}
$$

Hence the series converges absolutely and $EX = \frac{1-p}{p}$.

Example 1.6. There is an example of a random variable (density function) whose expectation does not exist because the series not absolutely convergent. See [HPS, Example 4, Section 4.1].

(End of Day 15)

2. Properties of Expectation

Lemma 2.1. *Let $\overline{X} = (X_1, X_2, \ldots, X_r)$ be an r-dimensional random vector with joint density function $f_{\overline{X}}$. Let $\varphi : \mathbb{R}^r \to \mathbb{R}$ be a function and define*

$$Z := \varphi(\overline{X}).$$

Then, Z has finite expectation if

$$\sum_{\overline{x} \in \mathbb{R}^r} |\varphi(\overline{x})| f_{\overline{X}}(\overline{x}) < \infty.$$

In that case,

$$EZ = \sum_{\overline{x} \in \mathbb{R}^r} \varphi(\overline{x}) f_{\overline{X}}(\overline{x}).$$

Proof. Omitted. See [HPS, Section 4.1, Theorem 1]. □

Theorem 2.2. *Let X and Y be two discrete random variables with finite expectation.*

(i) If $c \in \mathbb{R}$ and $P(X = c) = 1$, then $EX = c$.

(ii) If $c \in \mathbb{R}$, then cX has finite expectation and $E(cX) = cEX$.

(iii) $X + Y$ has finite expectation and

$$E(X + Y) = EX + EY.$$

(iv) If $P(X \geq Y) = 1$, then $EX \geq EY$. Moreover, if $P(X = Y) = 1$, then $EX = EY$.

(v) $|EX| \leq E(|X|)$ where $|X|$ is the random variable $w \mapsto |X(w)|$.

Proof. Assume for simplicity that X and Y both have finite ranges, say X has range $\{x_1, x_2, \ldots, x_n\}$ and Y has range $\{y_1, y_2, \ldots, y_m\}$.

(i) If $P(X = c) = 1$, then the density function of X is given by

$$f(x) = \begin{cases} 1 & : \text{ if } x = c \\ 0 & : \text{ otherwise.} \end{cases}$$

Therefore, $EX = \sum_{x \in \mathbb{R}} x f(x) = c f(c) = c$.

(ii) Clearly, cX has range $\{cx_1, cx_2, \ldots, cx_n\}$ and

$$P(cX = cx_i) = P(X = x_i).$$

Therefore, the density function for cX is

$$g(x) = \begin{cases} f(x_i) & : \text{ if } x = cx_i \\ 0 & : \text{ otherwise} \end{cases}$$

Therefore,

$$E(cX) = \sum_{i=1}^{n} cx_i f(x_i) = cE(X).$$

(iii) Clearly, $X + Y$ has range $\{x_i + y_j\}$. Let f denote the joint density function and f_X and f_Y denote the marginal density functions. Then,

$$f_X(x_i) = \sum_{j=1}^{m} f(x_i, y_j) \text{ and } f_Y(y_j) = \sum_{i=1}^{n} f(x_i, y_j).$$

Therefore,

$$\begin{aligned} E(X+Y) &= \sum_{i=1}^{n}\sum_{j=1}^{m}(x_i + y_j) f(x_i, y_j) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{m} x_i f(x_i, y_j) + \sum_{i=1}^{n}\sum_{j=1}^{m} y_j f(x_i, y_j) \\ &= \sum_{i=1}^{n} x_i f_X(x_i) + \sum_{j=1}^{m} y_j f_Y(y_j) \\ &= EX + EY. \end{aligned}$$

(iv) Omitted. See the book.

(v) Note that $-|X| \le X \le |X|$ so by part (iv),

$$-E(|X|) \le E(X) \le E(|X|).$$

Hence, $|E(X)| \le E(|X|)$.

□

(End of Day 16)

Example 2.3. Let $X_1, X_2, \ldots, X_n$ each independent Bernoulli random variables with common parameter p. If

$$X := X_1 + X_2 + \ldots + X_n$$

then by Theorem III.5.7, $X \sim B(n,p)$. Moreover,

$$EX = EX_1 + EX_2 + \ldots + EX_n = np.$$

as in Example 1.3.

Example 2.4. Suppose there are N balls in an urn of which r are red and $b = (N - r)$ aer black. If a random sample of $n \leq N$ is chosen and X denotes the number of red balls, then

$$X \sim \text{Hyp}(N, r, n)$$

with density function

$$f(x) = \begin{cases} \frac{\binom{r}{x}\binom{N-r}{n-x}}{\binom{N}{n}} & : \text{ if } 0 \leq x \leq n, 0 \leq n - x \leq n \\ 0 & : \text{ otherwise} \end{cases}$$

To calculate EX, we may consider n random variables $X_1, X_2, \ldots, X_n$ where $X_i(w) = 1$ if and only if the i^{th} ball drawn is red. Then,

$$EX_i = P(X_i = 1) = \frac{r}{N}.$$

Note that the X_i are not independent. However, $X = X_1 + X_2 + \ldots + X_n$ and therefore

$$EX = \frac{nr}{N}.$$

Theorem 2.5. *If X and Y are independent, then $E(XY) = E(X)E(Y)$.*

Proof. The joint density for X and Y is given by $f(x,y) = f_X(x)f_Y(y)$. Thus by Lemma 2.1,

$$\begin{aligned} E(XY) &= \sum_{x,y} xyf(x,y) \\ &= \sum_{x,y} xf_X(x)yf_Y(y) \\ &= \left(\sum_x xf_X(x)\right)\left(\sum_y yf_Y(y)\right) = E(X)E(Y). \end{aligned}$$

□

3. Moments

Definition 3.1. Let X be a discrete random variable and $r \geq 0$.

(i) We say that X has a moment of order r if X^r has finite expectation. In that case, we define the r^{th} moment of X to be

$$E(X^r).$$

(ii) If $\mu = E(X)$ is the mean, then the r^{th} central mean is defined as

$$E(X-\mu)^r$$

(iii) The variance of X is

$$\text{Var}X := E(X-\mu)^2 = E(X^2-2\mu X+\mu^2) = E(X^2)-2\mu E(X)+\mu^2 = E(X^2)-(EX)^2.$$

(iv) The standard deviation of X is $\sigma := \sqrt{\text{Var}X}$.

Remark 3.2.

(i) If μ is the mean of X, then $E(X-\mu) = E(X) - \mu E(1) = 0$. The higher central means can be non-zero though.

(ii) By Lemma 2.1,

$$E(X^r) = \sum_x x^r f(x).$$

(iii) If X has a moment of order $r \geq 0$, then it has moments of order k for each $1 \leq k \leq r$.

(iv) The standard deviation of X is a measure of how much X varies from the mean.

(v) $\text{Var}(X) = 0$ if and only if X is constant.

Proof. If $X = c$, then $P(X = c) = 1$ so

$$\mu = cP(X = c) = c.$$

Moreover, $\text{Var}X = E(X^2) - (EX)^2 = c^2 - c^2 = 0$.

Conversely, if $\text{Var}X = 0$, then $E(X-\mu)^2 = 0$. In other words, if f denotes the density function of X, then

$$\sum_x (x-\mu)^2 f(x) = 0 \Rightarrow f(x) = 0 \text{ unless } x = \mu.$$

Hence, $P(X = \mu) = 1$ so $X = \mu$. □

(vi) Suppose we wish to minimize the function $g : \mathbb{R} \to \mathbb{R}$ given by

$$g(a) := E(X-a)^2,$$

then we write

$$g(a) = E(X)^2 - 2aE(X) + a^2$$

Differentiating with respect to a gives $g'(a) = -2E(X) + 2a$ so the extremum occurs at $a = E(X)$. Moreover,

$$g''(a) = 2 > 0$$

so this is a minimum. Hence, Var$X$ is the minimum value that g takes.

(vii) If X is integer valued with probability generating function $\Phi_X : [-1,1] \to \mathbb{R}$. Suppose that there is a $t_0 > 1$ so that

$$\sum_{x=0}^{\infty} f_X(x) t_0^x < \infty$$

(This need not be the case, but it is if X is finitely valued). Then, we differentiate with respect to t to get

$$\Phi_X'(t) = \sum_{x=1}^{\infty} x f_X(x) t^{x-1}.$$

Hence, $\Phi_X'(1) = E(X)$. Similarly,

$$\Phi_X''(t) = \sum_{x=2}^{\infty} x(x-1) f_X(x) t^{x-1}$$

so that $\Phi_X''(1) = E(X(X-1)) = E(X^2 - X) = E(X^2) - E(X)$. Hence,

$$\text{Var}X = \Phi_X''(1) + \Phi_X'(1) - (\Phi_X'(1))^2.$$

Such formulae are also applicable for moments of higher order.

Example 3.3. Suppose $X \sim \text{Po}(\lambda)$, then the probability generating function of X is given by

$$\Phi_X(t) := e^{\lambda(t-1)}.$$

Hence, $\Phi_X'(t) = \lambda e^{\lambda(t-1)}$ and $\Phi_X''(t) = \lambda^2 e^{\lambda(t-1)}$. At $t = 1$, we get

$$\Phi_X'(1) = \lambda \text{ and } \Phi_X''(1) = \lambda^2.$$

Hence,

$$EX = \lambda$$

and

$$\text{Var}X = \lambda^2 + \lambda - \lambda^2 = \lambda.$$

(End of Day 17)

4. Variance of a Sum

Remark 4.1. If X and Y are two discrete random variables, then

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

Definition 4.2. The covariance of X and Y is

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

Remark 4.3.

(i) Hence,

$$\text{Var}(X+Y)=\text{Var}X+\text{Var}Y+2\text{Cov}(X,Y).$$

(ii) By Theorem 2.5, $\text{Cov}(X,Y)=0$ whenever X and Y are independent. In that case,

$$\text{Var}(X+Y)=\text{Var}X+\text{Var}Y.$$

Lemma 4.4.

(i) If $X_1,X_2,\ldots,X_n$ are discrete random variables with finite variance, then

$$Var(X_1+X_2+\ldots+X_n)=\sum_{i=1}^n Var(X_i)+2\sum_{i=1}^n\sum_{j=i+1}^n Cov(X_i,X_j).$$

(ii) If $X_1,X_2,\ldots,X_n$ are mutually independent with common variance σ^2, then

$$Var(X_1+X_2+\ldots+X_n)=n\sigma^2.$$

Example 4.5. Let $X_1,X_2,\ldots,X_n$ be independent and such that $X_i\sim\text{Bern}(p)$ for all $1\le i\le n$. If

$$X:=X_1+X_2\ldots+X_n$$

then $X\sim B(n,p)$ by Theorem III.5.7. By Lemma 4.4,

$$\text{Var}(X)=\sum_{i=1}^n\text{Var}(X_i).$$

Now, $E(X_i)=p$ from Example 1.2. Since $X_i=X_i^2$, we have $E(X_i^2)=p$ as well. Hence,

$$\text{Var}(X_i)=p-p^2=p(1-p).$$

Hence, $\text{Var}(X)=np(1-p)$.

Example 4.6. Suppose there are N balls in an urn of which r are red and $b = (N - r)$ are black (as in Example 2.4). If a random sample of $n \leq N$ is chosen and X denotes the number of red balls, then

$$X \sim \text{Hyp}(N, r, n)$$

Consider n random variables $X_1, X_2, \dots, X_n$ where $X_i(w) = 1$ if and only if the i^{th} ball drawn is red. Then,

$$EX_i = P(X_i = 1) = \frac{r}{N}.$$

Note that the X_i are not independent and $X = X_1 + X_2 + \dots + X_n$. Therefore,

$$EX = \frac{nr}{N}.$$

and

$$\text{Var}(X) = \sum_{i=1}^{n} \text{Var}(X_i) + 2 \sum_{i=1}^{n} \sum_{j=i+1}^{n} \text{Cov}(X_i, X_j).$$

Now consider

$$E(X_i X_j) = P(X_i = 1, X_j = 1) = \frac{r}{N}\frac{r-1}{N-1}$$

$$\begin{aligned}\text{Cov}(X_i, X_j) &= E(X_i X_j) - E(X_i)E(X_j) \\ &= \frac{r(r-1)}{N(N-1)} - \frac{r^2}{N^2}\end{aligned}$$

Hence,

$$\text{Var}(X) = n\frac{r}{N}\left(1 - \frac{r}{N}\right)\left(1 - \frac{n-1}{N-1}\right)$$

5. Correlation Coefficient

Definition 5.1. Suppose X and Y have finite non-zero variance. The correlation coefficient of X and Y is given by

$$\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}$$

X and Y are said to be uncorrelated if $\rho(X, Y) = 0$.

Remark 5.2. If X and Y are independent, then by Theorem 2.5, X and Y are uncorrelated.

Theorem 5.3 (Schwarz Inequality). *If X and Y have finite variance, then*

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

Moreover, equality holds if and only if $Y \sim 0$ or $X \sim aY$ for some $a \in \mathbb{R}$.

Proof.

(i) If $Y \sim 0$ then $P(Y = 0) = 1$. Then, $P(XY = 0) = 1$ so $E(XY) = 0$. Then, the inequality holds trivially and indeed, equality holds because $E(Y^2) = 0$.

(ii) If $X \sim aY$ for some $a \in \mathbb{R}$, then $P(X = aY) = 1$. Then,

$$E(XY) = E(aY^2) = aE(Y^2).$$

Moreover,

$$E(X^2)E(Y^2) = E(a^2Y^2)E(Y^2) = a^2[E(Y^2)]^2.$$

Hence equality holds.

(iii) Now assume that $P(Y = 0) < 1$ so that $E(Y^2) > 0$. Now note that for any $\lambda \in \mathbb{R}$,

$$0 \leq E(X - \lambda Y)^2 = \lambda^2 E(Y^2) - 2\lambda E(XY) + E(X^2) =: g(\lambda).$$

Now,

$$g'(\lambda) = 2\lambda E(Y^2) - 2E(XY) \text{ and } g''(\lambda) = 2E(Y^2) > 0.$$

Hence g attains its minimum at $a = \frac{E(XY)}{E(Y^2)}$. Now

$$0 \leq g(a) = E(X^2) - \frac{[E(XY)]^2}{E(Y^2)}.$$

This proves the inequality. Moreover, equality holds if and only if

$$g(a) = E(X - aY)^2 = 0.$$

This happens if and only if $X \sim aY$.

□

(End of Day 18)

Corollary 5.4. *For any two X and Y as above,*

$$-1 \leq \rho(X, Y) \leq 1$$

and $|\rho(X, Y)| = 1$ if and only if $X \sim aY$ for some $a \in \mathbb{R}$.

Proof. By Theorem 5.3 applied to $(X - EX)$ and $(Y - EY)$, we have

$$E[(X - EX)(Y - EY)]^2 \leq E(X - EX)^2 E(Y - EY)^2.$$

In other words,

$$\text{Cov}(X, Y)^2 \leq \text{Var}(X)\text{Var}(Y).$$

Hence, $|\rho(X, Y)| \leq 1$. The case of equality can also be verified similarly (Check!). □

6. Chebyshev's Inequality

Remark 6.1. Suppose X is a nonnegative random variable with finite expectation, and let $t > 0$. Define

$$Y(w) = \begin{cases} 0 & : \text{ if } X(w) < t \\ t & : \text{ if } X(w) \geq t \end{cases}$$

Then, Y is a discrete random variable with

$$P(Y = 0) = P(X < t) \text{ and } P(Y = 1) = P(X \geq t).$$

Hence,

$$E(Y) = 0P(Y =) + tP(Y = t) = tP(X \geq t).$$

Moreover, $X \geq Y$ so by Theorem 2.2, $EX \geq EY$. Hence,

$$P(X \geq t) \leq \frac{EX}{t}.$$

Theorem 6.2 (Chebyshev's Inequality)**.** *Let X be a random variable with mean μ and variance σ^2. Then, for any $t > 0$,*

$$P(|X - \mu| \geq t) \leq \frac{\sigma^2}{t^2}.$$

Proof. Consider the random variable $Z := (X - \mu)^2$ and apply the previous remark. Then,

$$P(Z \geq t^2) \leq \frac{EZ}{t^2} = \frac{E(X-\mu)^2}{t^2} = \frac{\sigma^2}{t^2}.$$

Now observe that $Z \geq t^2 \Leftrightarrow |X - \mu| \geq t$. □

Remark 6.3. Let $X_1, X_2, \ldots, X_n$ be independent random variables with a common distribution with finite mean μ and variance σ^2.

(i) We think of of these X_i as n independent measurements of the same quantity, and is called a random sample of size n from its common distribution.

(ii) Define

$$Y_n := \frac{X_1 + X_2 + \ldots + X_n}{n}.$$

We would expect that Y_n would be close to μ as $n \to \infty$. Since the X_i are independent,

$$E(Y_n) = \frac{n\mu}{n} = \mu$$

$$\text{Var}(Y_n) = \frac{1}{n^2}\sum_{i=1}^n \text{Var}(X_i) = \frac{\sigma^2}{n}$$

(iii) For a fixed $\delta > 0$, by Chebyshev's inequality

$$\begin{aligned} P\left(|Y_n - \mu| \geq \delta\right) &\leq \frac{\sigma^2}{n\delta^2} \\ \Rightarrow \lim_{n\to\infty} P\left(|Y_n - \mu| \geq \delta\right) &= 0. \\ \Rightarrow \lim_{n\to\infty} P(|Y_n - \mu| < \delta) &= 1. \end{aligned}$$

Theorem 6.4 (Weak Law of Large Numbers)**.** *Let $X_1, X_2, \ldots$ be independent random variables having a common distribution with finite mean μ. Set*

$$S_n := X_1 + X_2 + \ldots + X_n$$

Then, for any $\delta > 0$,

$$\lim_{n\to\infty} P\left(\left|\frac{S_n}{n} - \mu\right| \geq \delta\right) = 0.$$

Example 6.5. Suppose $X_1, X_2, \ldots, X_n$ are independent random variables with $X_i \sim \text{Bern}(p)$ for some $0 \leq p \leq 1$. Then,

$$\mu = p \text{ and } \sigma^2 = p(1-p).$$

By Theorem 6.4,

$$P\left(\left|\frac{S_n}{n} - \mu\right| \geq \delta\right) \leq \frac{p(1-p)}{n\delta^2}$$

Now, the function $p \mapsto p(1-p)$ has its maximum value on $[0,1]$ at $p = 1/2$, so

$$p(1-p) \leq \frac{1}{4}$$

for all $0 < p < 1$. Hence,

$$P\left(\left|\frac{S_n}{n} - \mu\right| \geq \delta\right) \leq \frac{1}{4n\delta^2}$$

Hence, given $\epsilon > 0$, if we choose $n \in \mathbb{N}$ so that

$$\frac{1}{4n\delta^2} < \epsilon$$

then we can ensure that after n trials,

$$P\left(\left|\frac{S_n}{n} - \mu\right| \geq \delta\right) < \epsilon.$$

(End of Day 19)

V. Continuous Random Variables

1. Random Variables and their distribution functions

Definition 1.1. Let $(\Omega, \mathcal{A}, P)$ be a probability space. A random variable is a function

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

such that, for each $x \in \mathbb{R}$, the set $\{w \in \Omega : X(w) \leq x\}$ belongs to $\mathcal{A}$.

Example 1.2.

(i) Note that, for a random variable, the range of X need not be countable. Typically, measurements such as time, distance, weight, etc. are random variables. Every discrete random variable is also a random variable.

(ii) Suppose $\Omega = [a, b]$ is a closed interval in $\mathbb{R}$, then there is a σ-algebra $\mathcal{A}$ called the Lebesgue σ-algebra that contains all open, closed and half-open intervals. On this σ-algebra, there is a probability measure P such that

$$P([x, y]) = \frac{|y - x|}{(b - a)} = P([x, y)) = P((x, y)).$$

The triple $(\Omega, \mathcal{A}, P)$ is a probability space, and many random variables and defined on such spaces.

(iii) Consider Example I.1.3: A pointer that is free to spin about the centre of a circle of radius 1. If the pointer is spun, it comes to rest at an angle (in radians) from the X axis. Define

$$\Omega = [0, 2\pi)$$

Recall from Example I.2.2 that there is a σ-algebra $\mathcal{A}$ on Ω, called the Lebesgue σ-algebra that contains all open (and half-open) intervals. There is also a probability measure P on $\mathcal{A}$ such that

$$P([a, b]) = \frac{[a, b]}{2\pi}.$$

This is the probability space $(\Omega, \mathcal{A}, P)$. Let X be the angle that the pointer comes to rest at. Then, for any $x \in \mathbb{R}$,

$$\{w \in \Omega : X(w) \leq x\} = [0, x] \cap \Omega \in \mathcal{A}.$$

Therefore, X is a random variable.

(iv) Consider a dart board of radius 1. A dart is thrown at the board, and we measure the distance of the dart from the origin, and denote it by X. Then,

$$\Omega := \{(a,b) \in \mathbb{R}^2 : a^2 + b^2 \leq 1\}.$$

and $\mathcal{A}$ is the Lebesgue σ-algebra which contains all open (and half-open) discs, and P be the Lebesgue measure (as above). Then, for any $x \in \mathbb{R}$,

$$\{w \in \Omega : X(w) \leq x\} = \{(a,b) \in \Omega : a^2 + b^2 \leq x^2\} \in \mathcal{A}.$$

Therefore, X is a random variable.

Definition 1.3. The distribution function of a random variable X is the function $F : \mathbb{R} \to \mathbb{R}$ given by

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

Example 1.4.

(i) In the spinning pointer example,

$$F(x) = \begin{cases} 0 & : x < 0 \\ \frac{x}{2\pi} & : 0 \leq x < 2\pi \\ 1 & : x \geq 2\pi. \end{cases}$$

(ii) In the dart board example, if $0 \leq x \leq 1$, we have

$$P(X \leq x) = \frac{1}{\pi}\text{Area}(\{(a,b) \in \Omega : a^2 + b^2 \leq x^2\}) = \frac{\pi x^2}{\pi} = x^2$$

Therefore,

$$F(x) = \begin{cases} 0 & : x < 0 \\ x^2 & : 0 \leq 1 \\ 1 & : x > 1. \end{cases}$$

Remark 1.5. This is a repeat of Remark III.2.3.

(i) For any $x \in \mathbb{R}$, $0 \leq F(x) \leq 1$.

(ii) If $s \leq t$, then

$$F(s) \leq F(t)$$

so F is *non-decreasing.*

(iii) If $(a,b]$ is an interval in $\mathbb{R}$, then

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

(End of Day 20)

(iv) Moreover, consider

$$F(-\infty) = \lim_{n\to-\infty} F(n) = \lim_{n\to-\infty} P(X \le n).$$

Now, $A_n := \{w \in \Omega : X(w) \le n\}$ form a decreasing sequence of sets with $\cap_{n=-1}^{-\infty} A_n = \emptyset$. Therefore, by Theorem I.3.5,

$$F(-\infty) = \lim_{n\to-\infty} P(A_n) = P(\emptyset) = 0.$$

Similarly,

$$F(+\infty) = 1.$$

(v) Fix $x \in \mathbb{R}$, and consider the right limit

$$F(x+) = \lim_{h\to0,h>0} F(x+h) = \lim_{n\to\infty} F(x+1/n)$$

(because F is non-decreasing and bounded, this limit exists). Now, set $B_n := \{w \in \Omega : X(w) \le x + 1/n\}$, then (B_n) is a decreasing sequence of sets with

$$B := \bigcap_{n=1}^{\infty} B_n = \{w \in \Omega : X(w) \le x\}$$

Again, by Theorem I.3.5,

$$F(x+) = \lim_{n\to\infty} P(X \le x + 1/n) = P(X \le x) = F(x).$$

(vi) Now fix $x \in \mathbb{R}$ and consider the left limit

$$F(x-) = \lim_{h\to0,h>0} F(x-h) = \lim_{n\to\infty} F(x-1/n).$$

Again, if $C_n := \{w \in \Omega : X(w) \le x - 1/n\}$, then (C_n) is an increasing family of sets with

$$C := \bigcup_{n=1}^{\infty} C_n = \{w \in \Omega : X(w) < x\}$$

Hence,

$$F(x-) = \lim_{n\to\infty} P(C_n) = P(C) = P(X < x).$$

Note that, in general, $F(x-) \ne F(x)$.

(vii) Observe that for any $x \in \mathbb{R}$,

$$F(x+) - F(x-) = P(X \le x) - P(X < x) = P(X = x).$$

Corollary 1.6. *The distribution function F is continuous if and only if*

$$P(X = x) = 0$$

for all $x \in \mathbb{R}$. In that case, X is called a continuous random variable.

Example 1.7.

(i) Discrete random variables are *not* continuous. For a continuous random variable, the notion of density function (as in Definition III.1.4) does not make sense.

(ii) In Example 1.2, both the spinning pointer and the dart board define continuous random variables.

Definition 1.8. A distribution function is a function $F : \mathbb{R} \to \mathbb{R}$ satisfying the following properties:

(i) $0 \leq F(x) \leq 1$ for all $x \in \mathbb{R}$.

(ii) F is a non-decreasing function

(iii) $F(-\infty) = 0$ and $F(+\infty) = 1$.

(iv) $F(x+) = F(x)$ for all $x \in \mathbb{R}$.

2. Densities of Continuous Random Variables

Definition 2.1. A density function is a non-negative function $f : \mathbb{R} \to \mathbb{R}$ such that

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

Example 2.2.

(i) Given real numbers $a < b$, define $f : \mathbb{R} \to \mathbb{R}$ by

$$f(t) = \begin{cases} 0 & : t < a \\ \frac{1}{(b-a)} & : a \leq t \leq b \\ 0 & : t > b \end{cases}$$

This is called the uniform density function on the interval $[a, b]$.

(ii) Fix $\lambda > 0$ and consider $f : \mathbb{R} \to \mathbb{R}$ given by

$$f(t) = \begin{cases} \lambda e^{-\lambda t} & : t > 0 \\ 0 & : t \leq 0. \end{cases}$$

Then, f is non-negative and for any $M > 0$,

$$\int_{-\infty}^{M} f(t)dt = -e^{-\lambda t}|_0^M = -e^{-\lambda M} + 1 \to 1 \text{ as } M \to \infty.$$

Therefore, f is a density function and is called the exponential density with parameter λ.

(iii) Define $f : \mathbb{R} \to \mathbb{R}$ by

$$f(t) = \frac{1}{\pi(1+x^2)}.$$

Then, for any $M > 0$,

$$\int_{-M}^{M} f(t)dt = \frac{1}{\pi}\arctan(t)|_{-M}^{M} = \frac{1}{\pi}(\arctan(M) - \arctan(-M)).$$

Hence,

$$\int_{-\infty}^{\infty} f(t)dt = \lim_{M\to\infty} \frac{1}{\pi}(\arctan(M) - \arctan(-M)) = \frac{1}{\pi}\left(\frac{\pi}{2} - \frac{-\pi}{2}\right) = 1.$$

Thus, f is a density function, called the Cauchy density.

Definition 2.3. Let $f : \mathbb{R} \to \mathbb{R}$ be a density function as above. Define $F : \mathbb{R} \to \mathbb{R}$ by

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

Then, F is a distribution function as per Definition 1.8. Moreover, if f is continuous at a point $x \in \mathbb{R}$, then

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

Most (but not all) distribution functions arise this way.

(End of Day 21)

Remark 2.4. If X is a continuous random variable with distribution function F and density function f, then

$$P(a \leq X \leq b) = P(a < X \leq b) = P(a \leq X < b) = \int_a^b f(t)dt = F(b) - F(a)$$

This quantity is represented by the area under the curve of f.

Example 2.5. If X is a continuous random variable with distribution function F and density function f, find the density function for $Y := X^2$.

Solution: Note that Y takes values in $\mathbb{R}_+$, so for any $y < 0$,

$$P(Y \leq y) = 0.$$

Now if $y \geq 0$, then

$$\begin{aligned} G(y) &:= P(Y \leq y) = P(-\sqrt{y} \leq X \leq \sqrt{y}) = F(\sqrt{y}) - F(-\sqrt{y}) \\ \Rightarrow g(y) &= G'(y) \\ &= \frac{1}{2\sqrt{y}}f(\sqrt{y}) + \frac{1}{2\sqrt{y}}f(-\sqrt{y}) \end{aligned}$$

So the density function of Y is

$$g(y) = \begin{cases} 0 & : \text{ if } y < 0 \\ \frac{1}{2\sqrt{y}} f(\sqrt{y}) + \frac{1}{2\sqrt{y}} f(-\sqrt{y}) & : \text{ if } y \geq 0 \end{cases}$$

□

Example 2.6. If X is uniformly distributed on $(0, 1)$ and $\lambda > 0$ is fixed, find the density function of $Y := -\lambda^{-1} \log(1 - X)$.

Solution: Note that the density function of X is

$$f(t) = \begin{cases} 0 & : \text{ if } t < 0 \text{ or } t > 1 \\ 1 & : \text{ if } 0 \leq t \leq 1. \end{cases}$$

So the distribution function of X is

$$F(x) = \begin{cases} 0 & : \text{ if } x < 0 \\ x & : \text{ if } 0 \leq x \leq 1 \\ 1 & : \text{ if } x > 1 \end{cases}$$

If G denotes the distribution function of Y, then $G(y) = 0$ if $y < 0$. Now if $y \geq 0$, then

$$\begin{aligned} G(y) = P(Y \leq y) &= P(\log(1 - X) \geq -\lambda) \\ &= P(1 - X \geq e^{-\lambda}) \\ &= P(X \leq 1 - e^{-\lambda}) \\ &= F(1 - e^{-\lambda}) \\ &= 1 - e^{-\lambda}. \end{aligned}$$

So the density function of Y is

$$g(y) = \begin{cases} 0 & : y < 0 \\ \lambda e^{-\lambda} & : y \geq 0 \end{cases}$$

This is the exponential density with parameter λ. □

3. Normal, Exponential and Gamma Densities

a. Normal Densities

Example 3.1. Let $g : \mathbb{R} \to \mathbb{R}$ be given by

$$g(x) = e^{-x^2/2}.$$

If $c := \int_{-\infty}^{\infty} g(x)dx$, then one can show (using polar coordinates) that

$$c^2 = 2\pi.$$

Therefore, if

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

then f is a density function called the standard normal density. We write $X \sim n(0,1)$.

Remark 3.2.

(i) Note that $f(t) = f(-t)$, so it is a *symmetric* random variable.

(ii) Let Φ denote the corresponding distribution function, then

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} dt.$$

This does not have a closed form solution.

(iii) However, we may observe some properties such as

$$\begin{aligned}\Phi(x) + \Phi(-x) &= \frac{1}{\sqrt{2\pi}} \left(\int_{-\infty}^{x} e^{-t^2/2} dt + \int_{-\infty}^{-x} e^{-t^2/2} dt \right). \\ &= \frac{1}{\sqrt{2\pi}} \left(\int_{-\infty}^{x} e^{-t^2/2} dt + \int_{x}^{\infty} e^{-t^2/2} dt \right) \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-t^2/2} dt \\ &= 1.\end{aligned}$$

Therefore,

$$\Phi(-x) = 1 - \Phi(x).$$

This is useful because we only need to determine values of $\Phi(x)$ when $x \geq 0$.

(iv) From the above calculation, we know that

$$\Phi(0) = \frac{1}{2}$$

(v) The other values of Φ are given at the end of the textbook. For instance, the table tells us that

$$\Phi(3) = 0.9987 \text{ and therefore } \Phi(-3) = 0.0013.$$

Therefore, the majority of the area of the curve lies between $[-3, 3]$. For a number $z \in \mathbb{R}$, the value $\Phi(z)$ is often called the z-score.

(vi) Suppose $X \sim n(0, 1)$ and $\mu, \sigma \in \mathbb{R}$ with $\sigma > 0$, then define

$$Y := \mu + \sigma X.$$

Then, Y is also a continuous random variable, and its distribution function is given by

$$\begin{aligned} G(y) = P(Y \leq y) &= P\left(X \leq \frac{y-\mu}{\sigma}\right) \\ &= \Phi\left(\frac{y-\mu}{\sigma}\right) \end{aligned}$$

So the density function of Y is $g : \mathbb{R} \to \mathbb{R}$ given by

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

This is a normal density with mean μ and variance σ^2, and is denoted $Y \sim n(\mu, \sigma^2)$. Note that

$$P(a \leq Y \leq b) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$$

This can be used to make calculations for all normal densities using the table.

(vii) For instance, if $X \sim n(1, 4)$, then

$$\begin{aligned} P(0 \leq Y \leq 3) &= \Phi\left(\frac{3-1}{2}\right) - \Phi\left(\frac{0-1}{2}\right) \\ &= \Phi(1) - (1 - \Phi(1/2)) \\ &= \Phi(1) + \Phi(0.5) - 1 \\ &= 0.5328 \text{ (from the table)} \end{aligned}$$

b. Exponential Densities

Recall, the density function for an exponential random variable with parameter $\lambda > 0$ is given by

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

and the distribution function is given by

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

Remark 3.3. Note that for any $a \geq 0$, we have

$$P(X > a) = 1 - P(X \leq a) = 1 - F(a) = e^{-\lambda a}$$

Hence, if $b \geq 0$, then

$$P(X > a)P(X > b) = P(X > a + b)$$

Equivalently,

$$P(X > a + b | X > a) = P(X > b).$$

This is the continuous analog of the *memoryless* property of the geometric random variable.

c. Gamma Densities

Definition 3.4. For $\alpha > 0$, define

$$\Gamma(\alpha) := \int_0^\infty x^{\alpha-1}e^{-x}dx$$

This integral is finite but does not have a closed form expression. This is called the Gamma function.

Definition 3.5. For a given $\alpha > 0$ and $\lambda > 0$, define the gamma density by

$$g(x) := \begin{cases} \frac{\lambda^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\lambda x} & : x \geq 0 \\ 0 & : x < 0 \end{cases}$$

This is denoted by $\Gamma(x; \alpha, \lambda)$.

Remark 3.6.

(i) Note that $g(x) \geq 0$ and by the definition of the Gamma function,

$$\int_{-\infty}^\infty g(x)dx = \frac{\lambda^\alpha}{\Gamma(\alpha)}\int_0^\infty x^{\alpha-1}e^{-\lambda x}dx = 1.$$

(by a change of variable $t := \lambda x$). Hence, g is a density function.

(ii) Suppose $X \sim n(0, \sigma^2)$ and $Y := X^2$, then the density function of X is

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

By Example 2.5, the density function of Y is given by

$$g(y) = \begin{cases} \frac{1}{2\sqrt{y}}(f(\sqrt{y}) - f(-\sqrt{y})) & : y > 0 \\ 0 & : y \leq 0. \end{cases}$$

A short calculation shows that

$$g(y) = \begin{cases} \frac{1}{\sigma\sqrt{2\pi}} e^{-y/2\sigma^2} & : y > 0 \\ 0 & : y \leq 0 \end{cases}$$

In other words, Y is a Gamma random variable with parameters $\alpha = 1/2$ and $\lambda = 1/2\sigma^2$.

(iii) Some values of Γ can be calculated. For instance,

$$\Gamma(1) = \int_0^\infty e^{-x} dx = 1.$$

(iv) One can prove that for any $\alpha > 0$,

$$\Gamma(\alpha + 1) = \alpha\Gamma(\alpha).$$

Therefore, it follows that $\Gamma(n) = (n - 1)!$ for all $n \in \mathbb{N}$.

(End of Day 22)

VI. Jointly Distributed Random Variables

1. Properties of Bivariate Distributions

Definition 1.1. Let X and Y be two random variables on a probability space (Ω, A, P). The joint distribution function is defined as $F : \mathbb{R}^2 \to \mathbb{R}$ by

$$F(x, y) := P(X \leq x \text{ and } Y \leq y)$$

Remark 1.2.

(i) We can directly use the joint distribution function to compute certain probabilities:

$$P(a < X \leq b, Y \leq d) = F(b, d) - F(a, d)$$
$$P(a < X \leq b, c < Y \leq d) = F(b, d) - F(b, c) - F(a, d) + F(a, c)$$

Definition 1.3.

(i) The marginal distribution functions are defined as

$$F_X(x) = P(X \leq x) \text{ and } F_Y(y) := P(Y \leq y).$$

(ii) If there is a function $f : \mathbb{R}^2 \to \mathbb{R}$ such that

$$F(x, y) = \int_{-\infty}^{y} \left(\int_{-\infty}^{x} f(u, v) dv \right) du$$

then f is called the joint density function. Again, as before, for continuous random variables, it is *not* true that

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

Remark 1.4.

(i) Observe that

$$P(a < X \leq b, c < Y \leq d) = \int_{a}^{b} \left(\int_{c}^{d} f(u, v) dv \right) du$$

More generally, for any subset $A \subset \mathbb{R}^2$ that is ‘nice’, we have

$$P((X, Y) \in A) = \int \int_A f(u, v) dv du$$

(ii) Taking $A = \mathbb{R}^2$, we get

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(u,v)dvdu = 1.$$

(iii) Given a joint density function $f : \mathbb{R}^2 \to \mathbb{R}$, we define

$$f_X(x) = \int_{-\infty}^{\infty} f(x,v)dv$$

Then f_X is a density function (as in Definition V.2.1) and

$$F_x(x) = \int_{-\infty}^{x} f_X(t)dt.$$

Hence, f_X is the density function for X and is called the marginal density. Similarly,

$$f_Y(y) = \int_{-\infty}^{\infty} f(u,y)du$$

is a density function for Y.

(iv) Under certain conditions (f should be twice continuously differentiable), we have

$$\frac{\partial^2 F}{\partial x \partial y} = f(x,y)$$

Definition 1.5. X and Y are said to be independent if whenever $a \leq b$ and $c \leq d$, we have

$$P(a < X \leq b, c < Y \leq d) = P(a < X \leq b)P(c < Y \leq d).$$

Equivalently, they are independent if and only if the joint distribution function is a product of marginal distribution functions:

$$F(x,y) = F_X(x)F_Y(y)$$

Remark 1.6.

(i) More generally, for any two sets $A, B \subset \mathbb{R}$, we have

$$P((X,Y) \in A \times B) = P(X \in A)P(Y \in B).$$

(ii) Moreover, if $f : \mathbb{R}^2 \to \mathbb{R}$ is a joint density function, then X and Y are independent if and only if

$$f(x,y) = f_X(x)f_Y(y).$$

(iii) Note that an abstract joint density function is a non-negative function $f : \mathbb{R}^2 \to \mathbb{R}$ such that

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)dydx = 1.$$

(iv) In particular, if f_1 and f_2 are two (one dimensional) density functions as in Definition V.2.1, then

$$f(x,y) := f_1(x)f_2(y)$$

defines a joint density function.

Example 1.7. If $X \sim n(0,1)$ and $Y \sim n(0,1)$ are independent, then the joint density function is given by

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

This is called the standard bivariate normal density.

Example 1.8. Suppose $f : \mathbb{R}^2 \to \mathbb{R}$ is given by

$$f(x,y) = ce^{-(x^2-xy+y^2)/2}$$

We need to find c so that it is a density function: Write

$$-(x^2 - xy + y^2) = -(y - x/2)^2 - 3x^2/4$$

So the marginal is given by

$$f_X(x) = ce^{-3x^2/8}\int_{-\infty}^{\infty} e^{-(y-x/2)^2/2}dy$$

A change of variable: $u := (y - x/2)$ gives

$$f_X(x) = ce^{-3x^2/8}\int_{-\infty}^{\infty} e^{-u^2/2}du = ce^{-3x^2/8}\sqrt{2\pi}.$$

Hence, $X \sim n(0,\sigma^2)$ where

$$c\sqrt{2\pi} = \frac{1}{\sigma\sqrt{2\pi}} \Rightarrow c = \frac{\sqrt{3}}{4\pi}$$

Hence,

$$f(x,y) = \frac{\sqrt{3}}{4\pi}e^{-(x^2-xy+y^2)/2}$$

is a density function. Note that

$$X \sim n(0,4/3).$$

Similarly, one can show that $Y \sim n(0,4/3)$. Since

$$f(x,y) \neq f_X(x)f_Y(y)$$

it follows that X and Y are not independent.

2. Distribution of Sums

Remark 2.1. Suppose $Z = X+Y$ is the sum of two random variables with joint density function f. Then, for any $z \in \mathbb{R}$, if

$$A_z = \{(x,y) : x + y \leq z\}$$

Then this is the half-plane under the line $y = z - x$. Hence,

$$\begin{aligned} F_Z(z) = P(A_z) &= \int\int_{A_z} f(x,y)dydx \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{z-x} f(x,y)dydx \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{z} f(x,v-x)dvdx \\ &= \int_{-\infty}^{z}\int_{-\infty}^{\infty} f(x,v-x)dxdv \end{aligned}$$

Therefore, the density function of Z is

$$f_{X+Y}(z) = \int_{-\infty}^{\infty} f(x, z-x)dx$$

Moreover, if X and Y are independent, then

$$f_{X+Y}(z) = \int_{-\infty}^{\infty} f_X(x)f_Y(z-x)dx$$

This is called the convolution of the two functions f_X and f_Y.

Example 2.2. Suppose X and Y are independent random variables each having an exponential distribution with parameter λ. Find the distribution function of $(X+Y)$.

Solution: The density of X is given by

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & : \text{ if } x \geq 0 \\ 0 & : \text{ if } x < 0. \end{cases}$$

and the same for Y as well. The density function for $(X+Y)$ is given by

$$f_{X+Y}(z) = \int_{-\infty}^{\infty} f_X(x)f_Y(z-x)dx$$

Since $f_X(x) = 0$ if $x < 0$ and $f_Y(z-x) = 0$ if $x > z$, it follows that

$$f_{X+Y}(z) = 0$$

if $z < 0$ and if $z \geq 0$, then

$$\begin{aligned} f_{X+Y}(z) &= \int_0^z f_X(x) f_Y(z-x)dx \\ &= \lambda^2 \int_0^z e^{-\lambda x} e^{-\lambda(z-x)} dx \\ &= \lambda^2 e^{-\lambda z} \end{aligned}$$

Hence, $X + Y \sim \Gamma(2, \lambda)$. □

Theorem 2.3. *Suppose $X \sim n(\mu_1, \sigma_1^2)$ and $Y \sim n(\mu_2, \sigma_2^2)$ are both independent random variables, then*

$$X + Y \sim n(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2).$$

Proof. We will assume that $\mu_1 = \mu_2 = 0$ and $\sigma_1 = \sigma_2 = 1$. The general proof is similar but more technical. Then,

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \text{ and } f_Y(y) = \frac{1}{\sqrt{2\pi}} e^{-y^2/2}$$

By the convolution formula,

$$f_{X+Y}(z) = \frac{1}{2\pi} \int_{\infty}^{\infty} e^{\frac{-1}{2}(x^2+(z-x)^2)} dx$$

Observe that

$$x^2 + (z-x)^2 = 2x^2 + z^2 - 2zx = (\sqrt{2}x - z/\sqrt{2})^2 + \frac{z^2}{2}$$

Hence, if $u := \sqrt{2}x - z/\sqrt{2}$, then $du = \sqrt{2}dx$ so

$$\begin{aligned} f_{X+Y}(z) &= \frac{e^{-z^2/4}}{2\pi} \int_{-\infty}^{\infty} e^{-u^2/2} \frac{1}{\sqrt{2}} du \\ &= \frac{1}{2\sqrt{2}\pi} e^{-z^2/4} \sqrt{2\pi} \\ &= \frac{1}{\sqrt{2\pi}\sqrt{2}} e^{-z^2/4} \end{aligned}$$

Hence,

$$(X + Y) \sim n(0, 2).$$

□

Theorem 2.4. *Suppose $X \sim \Gamma(\alpha_1, \lambda)$ and $Y \sim \Gamma(\alpha_2, \lambda)$ are independent random variables, then $X + Y \sim \Gamma(\alpha_1 + \alpha_2, \lambda)$.*

Proof. Omitted. See [HPS, Section 6.2]. □

3. Conditional Densities

Remark 3.1. Suppose X and Y are discrete random variables with joint density function f and $x, y \in \mathbb{R}$. If $P(X = x) > 0$, then

$$P(Y = y|X = x) = \frac{P(Y = y, X = x)}{P(X = x)} = \frac{f(x, y)}{f_X(x)}$$

The conditional density function may thus be defined as

$$f_{Y|X}(y|x) = \begin{cases} \frac{f(x,y)}{f_X(x)} & : \text{ if } 0 < f_X(x) < \infty \\ 0 & : \text{ otherwise.} \end{cases}$$

Then, $f_{Y|X}$ is a non-negative function and if $0 < f_X(x) < \infty$, then

$$\sum_{y \in \mathbb{R}} f_{Y|X}(y|x) = \sum_{y \in \mathbb{R}} \frac{f(x, y)}{f_X(x)} = \frac{f_X(x)}{f_X(x)} = 1.$$

For each such $x \in \mathbb{R}$, this is a density function in the sense of Lemma III.1.12.

Definition 3.2. Let X and Y be two random variables with joint density function f. Then, the conditional density function is defined as

$$f_{Y|X}(y|x) = \begin{cases} \frac{f(x,y)}{f_X(x)} & : \text{ if } 0 < f_X(x) < \infty \\ 0 & : \text{ otherwise.} \end{cases}$$

Remark 3.3.

(i) Note that for any $a \leq b$ and $x \in \mathbb{R}$

$$P(a < Y \leq b|X = x) = \int_a^b f_{Y|X}(y|x)dy.$$

This can also be done using calculus (see the textbook).

(ii) Observe that for any $x, y \in \mathbb{R}$,

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

(iii) In particular, X and Y are independent if and only if

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

for any $x, y \in \mathbb{R}$.

(iv) You should think of the conditional density as the vertical cross-section of the joint density function surface at the point $X = x$, which is further scaled by $f_X(x)$ so that its area is 1.

Example 3.4. Consider X, Y with joint density function

$$f(x,y) = \frac{\sqrt{3}}{4\pi} e^{\frac{-x^2+xy-y^2}{2}}$$

We saw in Example 1.8 that X and Y are dependent normal variables with $X \sim n(0, 4/3)$ and $Y \sim n(0, 4/3)$. Note that for any $x \in \mathbb{R}$,

$$f_X(x) = \frac{\sqrt{3}}{2\sqrt{2\pi}} e^{-3x^2/8}$$

Hence, the conditional density is given by

$$\begin{aligned} f_{Y|X}(y|x) &= \frac{\frac{\sqrt{3}}{4\pi} e^{\frac{-x^2+xy-y^2}{2}}}{\frac{\sqrt{3}}{2\sqrt{2\pi}} e^{-3x^2/8}} \\ &= \frac{1}{\sqrt{2\pi}} e^{-(y-x/2)^2/2} \end{aligned}$$

Hence, for each $x \in \mathbb{R}$, the conditional density if Y given $X = x$ is the normal density $n(x/2, 1)$.

Theorem 3.5 (Bayes' Rule)**.** *For any $x, y \in \mathbb{R}$, we have*

$$f_{X|Y}(x|y) = \frac{f_X(x) f_{Y|X}(y|x)}{\int_{-\infty}^{\infty} f_X(x) f_{Y|X}(y|x) dx}$$

Proof. We know that

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

if $0 < f_Y(y) < \infty$. Now note that

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

by definition. Moreover,

$$f_Y(y) = \int_{-\infty}^{\infty} f(x,y) dx = \int_{-\infty}^{\infty} f_X(x) f_{Y|X}(y|x) dx.$$

□

(End of Day 23)

4. Properties of Multivariate Distributions

Definition 4.1. Suppose $X_1, X_2, \ldots, X_n$ are n random variables.

(i) The joint distribution function is given by $F : \mathbb{R}^n \to \mathbb{R}$ defined as

$$F(x_1, x_2, \ldots, x_n) = P(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n)$$

(ii) A function $f : \mathbb{R}^n \to \mathbb{R}$ is called a joint density function for $X_1, \ldots, X_n$ if it is non-negative and

$$F(x_1, x_2, \ldots, x_n) = \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} \cdots \int_{-\infty}^{x_n} f(u_1, u_2, \ldots, u_n) du_n du_{n-1} \ldots du_1.$$

Remark 4.2.

(i) Under certain mild conditions, we have

$$f(x_1, x_2, \ldots, x_n) = \frac{\partial^n F}{\partial x_1 \partial x_2 \ldots \partial x_n}(x_1, x_2, \ldots, x_n)$$

(ii) For any 'nice' subset $A \subset \mathbb{R}^n$, we have

$$P((X_1, X_2, \ldots, X_n) \in A) = \int_A f(u_1, u_2, \ldots, u_n) du_n du_{n-1} \ldots du_1.$$

(iii) In particular,

$$\int_{\mathbb{R}^n} f(u_1, u_2, \ldots, u_n) du_n du_{n-1} \ldots du_1 = 1.$$

and for any n-cell ('rectangle' in n dimensions), we have

$$P(a_1 < X_1 \leq b_1, \ldots, a_n < X_n \leq b_n) = \int_{a_1}^{b_1} \cdots \int_{a_n}^{b_n} f(u_1, u_2, \ldots, u_n) du_n du_{n-1} \ldots du_1.$$

(iv) The marginal distribution function for X_j is defined $F_{X_j} : \mathbb{R} \to \mathbb{R}$ defined as

$$F_{X_j}(x) = P(X_j \leq x).$$

(v) If $f : \mathbb{R}^n \to \mathbb{R}$ is the joint density function, then the marginal density functions are defined as

$$f_{X_1}(x_1) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, u_2, \ldots, u_n) du_n du_{n-1} \ldots du_2.$$

and others are defined similarly.

(vi) The variables $X_1, X_2, \dots, X_n$ are said to be independent if

$$P(a_1 < X_1 \leq b_1, \dots, a_n < X_n \leq b_n) = \prod_{i=1}^{n} P(a_i < X_i \leq b_i).$$

Equivalently, this happens if and only if

$$F(x_1, x_2, \dots, x_n) = \prod_{i=1}^{n} F_{X_i}(x_i).$$

and equivalently, if and only if

$$f(x_1, x_2, \dots, x_n) = \prod_{i=1}^{n} f_{X_i}(x_i).$$

Example 4.3. Suppose $X_1, X_2, \dots, X_n$ are independent random variables, each of which is exponentially distributed with density λ. Then, the joint density function is given by

$$f(x_1, x_2, \dots, x_n) = \begin{cases} \lambda^n e^{-\lambda(x_1+x_2+\ldots+x_n)} & : \text{ if } x_i > 0 \text{ for all } 1 \leq i \leq n \\ 0 & : \text{ otherwise.} \end{cases}$$

Proof. Exercise. Try it for $n = 2$ first! □

Theorem 4.4. *Suppose $X_1, X_2, \dots, X_n$ are independent random variables. Suppose*

$$Y = \varphi(X_1, X_2, \dots, X_m) \text{ and } Z = \psi(X_{m+1}, X_{m+2}, \dots, X_n)$$

for two functions φ and ψ. Then, Y and Z are independent.

Corollary 4.5. *Suppose $X_1, X_2, \dots, X_n$ are independent random variables with $X_i \sim n(\mu_i, \sigma_i^2)$. Then,*

$$X_1 + X_2 + \ldots + X_n \sim n(\mu, \sigma^2)$$

where $\mu = \sum_{i=1}^n \mu_i$ and $\sigma^2 = \sum_{i=1}^n \sigma_i^2$.

Proof. For $n = 2$, this was proved in Theorem 2.3. For $n > 2$, we use induction. Assume that the result is true for $(n-1)$ variables. Let $Y := X_2 + X_3 + \ldots + X_n$. By induction hypothesis,

$$Y \sim n(\nu, \lambda^2)$$

where $\nu = \sum_{i=2}^n \mu_i$ and $\lambda_2 = \sum_{i=2}^n \sigma_i^2$. Moreover, by Theorem 4.4, Y and X_1 are independent. Therefore, by the $n = 2$ case,

$$X_1 + Y \sim n(\mu, \sigma^2)$$

as desired. □

VII. Expectations and the Central Limit Theorem

1. Expectations of Continuous Random Variables

Definition 1.1. Let X be a continuous random variable with density function f. We say that X has finite expectation if

$$\int_{-\infty}^{\infty} |x| f(x)dx < \infty.$$

If this happens, we define the expectation of X to be

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

This is also called the mean of X.

Example 1.2. If $X \sim U(a, b)$, then

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

Hence,

$$EX = \int_a^b \frac{x}{(b-a)} dx = \frac{x^2}{2(b-a)} |_a^b = \frac{(b+a)}{2}.$$

Example 1.3. If $X \sim \Gamma(\alpha, \lambda)$, then

$$f(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}.$$

Hence,

$$\begin{aligned} EX &= \frac{\lambda^\alpha}{\Gamma(\alpha)} \int_{-\infty}^{\infty} x^\alpha e^{-\lambda x} dx \\ &= \frac{\lambda^\alpha}{\Gamma(\alpha)} \frac{\Gamma(\alpha+1)}{\lambda^{\alpha+1}} \\ &= \frac{\alpha}{\lambda} \end{aligned}$$

because $\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$.

Example 1.4. If $X \sim \text{Exp}(\lambda)$, then $X \sim \Gamma(1, \lambda)$, so

$$EX = \frac{1}{\lambda}$$

by the previous example.

Example 1.5. If X is a Cauchy distribution, then

$$f(x) = \frac{1}{\pi(1+x^2)}$$

so

$$\begin{aligned}
\frac{1}{\pi}\int_{-\infty}^{\infty} \frac{|x|}{(1+x^2)}dx &= \frac{2}{\pi}\int_{0}^{\infty} \frac{x}{(1+x^2)}dx \\
&= \frac{2}{\pi}\lim_{M\to\infty}\int_{0}^{M} \frac{x}{(1+x^2)}dx \\
&= \frac{1}{\pi}\lim_{M\to\infty} \ln(1+x^2)|_0^M \\
&= \frac{1}{\pi}\lim_{M\to\infty} \ln(1+M^2) = +\infty.
\end{aligned}$$

So X does not have finite expectation, so EX is undefined.

Theorem 1.6. *Let $X_1, X_2, \ldots, X_n$ be random variables with joint density $f : \mathbb{R}^n \to \mathbb{R}$. Let Z be the random variable defined as*

$$Z = \varphi(X_1, X_2, \ldots, X_n)$$

Then, Z has finite expectation if

$$\int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} |\varphi(x_1, x_2, \ldots, x_n)| f(x_1, x_2, \ldots, x_n) dx_1 dx_2 \ldots dx_n < \infty$$

and in that case,

$$EZ = \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} \varphi(x_1, x_2, \ldots, x_n) f(x_1, x_2, \ldots, x_n) dx_1 dx_2 \ldots dx_n$$

(End of Day 24)

2. Moments of Continuous Random Variables

Definition 2.1. Let X be a random variable with density function f and mean μ, and let $m \in \mathbb{N}$.

(i) If X^m has finite expectation, then the <u>m^{th} moment of X</u> is defined as

$$EX^m = \int_{-\infty}^{\infty} x^m f(x)dx.$$

(ii) The <u>m^{th} central moment of X</u> is defined as

$$E(X-\mu)^m = \int_{-\infty}^{\infty} (x-\mu)^m f(x)dx.$$

(iii) The <u>variance</u> of X is

$$\text{Var}(X) = E(X-\mu)^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx.$$

(iv) The <u>standard deviation</u> of X is $\sigma := \sqrt{\text{Var}(X)}$.

Remark 2.2.

(i) As in the discrete case, $\sigma = 0$ if and only if X is constant.

(ii) As in the discrete case,

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

Example 2.3. If $X \sim \Gamma(\alpha, \lambda)$, then we find the moments and variance of X.

Solution: For $m \geq 1$,

$$\begin{aligned} E(X^m) &= \frac{\lambda^\alpha}{\Gamma(\alpha)} \int_0^\infty x^{m+\alpha-1} e^{-\lambda x} dx \\ &= \frac{\lambda^\alpha}{\Gamma(\alpha)} \frac{\Gamma(\alpha+m)}{\lambda^{\alpha+m}} \\ &= \frac{\alpha(\alpha+1)\ldots(\alpha+m-1)}{\lambda^m} \end{aligned}$$

By the earler calculation, $E(X) = \alpha/\lambda$, so the variance is given by

$$\sigma^2 = E(X^2) - E(X)^2 = \frac{\alpha(\alpha+1)}{\lambda^2} - \frac{\alpha^2}{\lambda^2} = \frac{\alpha}{\lambda^2}.$$

□

Example 2.4. If $X \sim \text{Exp}(\lambda) = \Gamma(1, \lambda)$, then

$$\mu = \frac{1}{\lambda} \text{ and } \sigma^2 = \frac{1}{\lambda^2}.$$

Remark 2.5. Suppose X is a random variable with density function f that is *symmetric*, i.e.

$$f(-x) = f(x)$$

Then, X and $-X$ have the same density function. In particular, if $m \in \mathbb{N}$ is odd, then

$$E(X^m) = E((-X)^m) = -E(X^m) \Rightarrow E(X^m) = 0.$$

Example 2.6. Suppose $X \sim n(\mu, \sigma^2)$, then we calculate the mean and variance of X.

Solution: Note that $(X - \mu) \sim n(0, \sigma^2)$. This is a symmetric density function, so in particular,

$$E(X - \mu) = 0 \Rightarrow E(X) = \mu.$$

Now if $Y = (X - \mu)^2$, then by Remark V.3.6,

$$Y \sim \Gamma(1/2, 1/2\sigma^2).$$

Therefore,

$$\text{Var}(X) = EY = \frac{1/2}{1/2\sigma^2} = \sigma^2.$$

□

Definition 2.7. Suppose X and Y have joint density function $f : \mathbb{R}^2 \to \mathbb{R}$, means μ_X and μ_Y and finite second moments. Then the covariance of X and Y is defined as

$$\text{Cov}(X, Y) = E(X - \mu_X)(Y - \mu_Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x - \mu_X)(y - \mu_Y)f(x, y)dxdy$$

Remark 2.8.

(i) As before,

$$\text{Cov}(X, Y) = E(XY - \mu_X Y - \mu_Y X + \mu_X\mu_Y) = E(XY) - E(X)E(Y).$$

(ii) If X and Y are independent, then $f(x, y) = f_X(x)f_Y(y)$ for all $x, y \in \mathbb{R}^2$. Therefore,

$$\begin{aligned} E(XY) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf_X(x)f_Y(y)dydx \\ &= \int_{-\infty}^{\infty} xf_X(x)\left(\int_{-\infty}^{\infty} yf_X(y)dy\right)dx \\ &= \mu_Y\mu_X \end{aligned}$$

Hence, $\text{Cov}(X, Y) = 0$.

Example 2.9. Consider the function $f : \mathbb{R}^2 \to \mathbb{R}$ given by

$$f(x,y) = \frac{\sqrt{3}}{4\pi} e^{-[x^2-xy+y^2]/2}$$

Then, by Example VI.1.8, $X \sim n(0, 4/3)$ and $Y \sim n(0, 4/3)$. We now calculate $\mathrm{Cov}(X,Y)$.

Solution: We know from Example 2.6 that $\mu_X = \mu_Y = 0$. Now,

$$\begin{aligned}
E(XY) &= \frac{\sqrt{3}}{4\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xye^{-[x^2-xy+y^2]/2} dydx \\
&= \frac{\sqrt{3}}{4\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xye^{-3x^2/8} e^{-[(y-x/2)^2/2} dydx \\
&= \frac{\sqrt{3}}{4\pi} \int_{-\infty}^{\infty} xe^{-3x^2/8} \left(\int_{-\infty}^{\infty} ye^{-[(y-x/2)^2/2]} dy \right) dx
\end{aligned}$$

Now,

$$\begin{aligned}
\int_{-\infty}^{\infty} ye^{-[(y-x/2)^2/2]} dy &= \int_{-\infty}^{\infty} \left(u + \frac{x}{2} \right) e^{-u^2/2} du \\
&= \int_{-\infty}^{\infty} ue^{-u^2/2} du + \frac{x}{2} \int_{-\infty}^{\infty} e^{-u^2/2} du \\
&= \frac{x}{2} \sqrt{2\pi}.
\end{aligned}$$

Therefore,

$$E(XY) = \frac{\sqrt{6\pi}}{8\pi} \int_{-\infty}^{\infty} x^2 e^{-3x^2/8} dx$$

Now if $Z \sim n(0, 4/3)$, then $Z^2 \sim \Gamma(1/2, 3/8)$ by Remark V.3.6. Therefore, by

$$E(Z^2) = \frac{(1/2)(1/2+1)}{(3/8)^2} = \frac{3/4}{9/64} = \frac{3 \times 64}{4 \times 9} = \frac{16}{3}.$$

Hence, a short calculation shows that

$$E(XY) = \frac{1}{2} = \mathrm{Cov}(X,Y).$$

□

3. The Central Limit Theorem

Definition 3.1. $X_1, X_2, \ldots, X_n$ are said to be i.i.d. if they are independent and identically distributed.

We will always assume that they have finite mean μ and variance σ^2.

Remark 3.2. Define

$$S_n := X_1 + X_2 + \ldots + X_n$$

Then,

$$E(S_n) = n\mu \text{ and } \mathrm{Var}(S_n) = n\sigma^2.$$

Therefore, if

$$S_n^* := \frac{S_n - n\mu}{\sigma\sqrt{n}}$$

then S_n^* has mean 0 and variance 1.

Question: What is the distribution function of this random variable S_n^*?

Example 3.3.

(i) If $X_1 \sim n(\mu, \sigma^2)$, then $S_n \sim n(n\mu, n\sigma^2)$. Therefore,

$$S_n^* \sim n(0,1).$$

(ii) If $X_1 \sim \mathrm{Po}(\lambda)$, then $\mu = \sigma^2 = \lambda$, so

$$S_n^* = \frac{S_n - n\lambda}{\sqrt{n\lambda}}$$

(iii) If $X_1 \sim \mathrm{Bern}(p)$ for some $0 < p < 1$, then $S_n \sim B(n,p)$. Hence,

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

Here, $\mu = p$ and $\sigma^2 = p(1-p)$, so

$$S_n^* = \frac{S_n - np}{\sqrt{np(1-p)}}$$

De Moivre (ca 1700) and Laplace (ca 1800) proved that in this case,

$$P(S_n^* \leq x) \approx \Phi(x)$$

where Φ is the standard normal distribution. This was generalized to its current form by Lindemann in 1922.

(End of Day 25)

Theorem 3.4. *Let $X_1, X_2, \ldots, X_n$ be i.i.d random variables with mean μ and finite non-zero variance σ^2. If $S_n = X_1 + X_2 + \ldots + X_n$, then for any $x \in \mathbb{R}$,*

$$\lim_{n\to\infty} P\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \leq x\right) = \Phi(x)$$

a. Normal Approximation

Remark 3.5. For n large, the Central Limit Theorem says that

$$P\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \leq x\right) \approx \Phi(x)$$

Rewriting this, we get

$$P(S_n \leq x) \approx \Phi\left(\frac{x - n\mu}{\sigma\sqrt{n}}\right)$$

This is called a normal approximation formula.

Example 3.6. Suppose the length of life of a light bulb is exponentially distributed by $\mu = 10$ days. As soon as one bulb burns out, another is installed in its place. Find the probability that more than 50 light bulbs will be needed in year.

Solution: Let X_n denote the length of life of the n^{th} bulb. We assume that $X_1, X_2, \ldots$ are all independent with $X_i \sim \text{Exp}(\lambda)$ where $\lambda = 1/10$ (so that $\mu = 10$ and $\sigma^2 = 100$). Then, if

$$S_n := X_1 + X_2 + \ldots + X_n$$

then we wish to calculate $P(S_{50} < 365)$. Now, by the normal approximation formula,

$$\begin{aligned} P(S_{50} < 365) &\approx \Phi\left(\frac{365 - 50\mu}{\sigma\sqrt{50}}\right) \\ &= \Phi\left(\frac{365 - 500}{10\sqrt{50}}\right) \\ &= \Phi(-1.91) = 0.028. \end{aligned}$$

Therefore, it is very unlikely we will need more than 50 lightbulbs. □

Example 3.7. An instructor has 60 exams to grade in sequence. The time required to grade each exam is independent and identically distributed with mean 15 minutes and standard deviation 2 minutes. Approximate the probability that she will grade all the exams

(i) in the first 14 hours?

(ii) in the first 16 hours?

Solution: Let X_i be the time taken to grade the i^{th} exam, then we are interested in

$$S_{60} = \sum_{i=1}^{60} X_i.$$

(i) For part (i), we are interested in

$$
\begin{aligned}
P(S_{60} \leq 840) &\approx \Phi\left(\frac{840 - (60 \times 15)}{2\sqrt{60}}\right) \\
&= \Phi\left(-\frac{30}{\sqrt{60}}\right) \\
&= \Phi(-5.47)
\end{aligned}
$$

Note that Φ is an increasing function, and $\Phi(-3.9) \approx 0$. Hence,

$$P(S_{60} \leq 840) \approx 0.$$

Hence, she will almost certainly not finish in 14 hours.

(ii) However, for part (ii), we are interested in

$$P(S_{60} \leq 960) \approx \Phi(+5.47) \approx 1$$

because $\Phi(3.9) \approx 1$.

Hence, she will almost certainly be done in 16 hours. □

Note: In practice, X_1 will be much larger than X_{60} because it gets easier to grade as you go along. Therefore, X_1 is most certainly not a normal distribution!

b. Additional Material

Here are some additional material to help understand the Central Limit Theorem:

(i) Youtube video: https://www.youtube.com/watch?v=jvoxEYmQHNM

(ii) Geogebra applet: https://www.geogebra.org/m/xqvcg8sm

(iii) Geogebra applet: https://www.geogebra.org/m/n4SujFmy

c. Random Sampling

Remark 3.8. Consider a probability space $(\Omega, \mathcal{A}, P)$ and a random variable $X : \Omega \to \mathbb{R}$. We think of values $x := X(w)$ taken by X as outcomes of an experiment. Let F be the distribution function of X. In practice, we often do not know what F is.

(i) To understand F better, a statistician would obtain n independent observations on X, i.e. She would obtain n values $x_1, x_2, \ldots, x_n$ assumed by X. Each x_i is regarded as a value assumed by a random variable $X_i, 1 \leq i \leq n$ where $X_1, X_2, \ldots, X_n$ are independent RVs with common distribution F. The observed values $(x_1, x_2, \ldots, x_n)$ are then taken to be values of $(X_1, X_2, \ldots, X_n)$.

(ii) The set $\{X_1, X_2, \ldots, X_n\}$ is called a sample of size n taken from a population distribution F.

(iii) The set $\{x_1, x_2, \ldots, x_n\}$ is called a realization of the sample.

(iv) A simple random sample is one choice of tuple $(x_1, x_2, \ldots, x_n)$ which is chosen without replacement. Note: In this case, the variables $X_1, X_2, \ldots, X_n$ are not independent, but for large populations and small sample size, it is not very different from sampling with replacement (in which case they are independent).

(v) In practice, one often observes not the values $\{x_1, x_2, \ldots, x_n\}$ but a single value

$$\varphi(x_1, x_2, \ldots, x_n).$$

The random variable $Z := \varphi(X_1, X_2, \ldots, X_n)$ is called a (sample) statistic, provided it does not depend on any unknown parameters.

(vi) For example,

$$\overline{X} := \frac{X_1 + X_2 + \ldots + X_n}{n}$$

is called the sample mean, and

$$S^2 = \sum_{i=1}^{n} \frac{(X_i - \overline{X})^2}{n-1} = \frac{\sum_{i=1}^{n} X_i^2 - n\overline{X}^2}{n-1}$$

is called the sample variance. S is called the sample standard deviation.

Example 3.9. Suppose $X \sim \text{Bern}(p)$ for some $0 < p < 1$, which is possibly unknown. If five independent observations are $\{0, 1, 1, 0, 0\}$, then this is a realization of the sample $\{X_1, X_2, \ldots, X_5\}$. The sample mean is

$$\overline{x} = \frac{2}{5}$$

which is the value assumed by the RV $\overline{X}$, and the sample variance is

$$s^2 = \frac{\sum_{i=1}^{5} x_i^2 - 5\overline{x}^2}{4} = 0.3$$

which is the value assumed by the RV S^2.

We may now rephrase the Central Limit theorem as follows.

Theorem 3.10. *For sufficiently large n, the distribution of the sample mean $\overline{X}$ approximates a normal distribution (regardless of the original distribution F).*

Remark 3.11. Fix $c > 0$ and consider

$$\begin{aligned} P\left(\left|\frac{S_n}{n} - \mu\right| \geq c\right) &= P(S_n \leq n\mu - nc) + P(S_n \geq n\mu + nc) \\ &\approx \Phi\left(\frac{-nc}{\sigma\sqrt{n}}\right) + 1 - \Phi\left(\frac{nc}{\sigma\sqrt{n}}\right) \\ &= 2\left[1 - \Phi\left(\frac{c\sqrt{n}}{\sigma}\right)\right] \end{aligned}$$

Hence, if $\delta := c\sqrt{n}/\sigma$, then

$$P\left(\left|\frac{S_n}{n} - \mu\right| \geq c\right) \approx 2(1 - \Phi(\delta))$$

This is also called a <u>normal approximation formula</u>.

(End of Day 26)

Example 3.12. An astronomer measures the distance d to a star in light years. He believes that each measurement is independent and identically distributed with common variance of 4 light years. How many measurements does he need to make to be 95% sure that his estimated distance (the sample mean) is accurate to within ±0.1 light years?

Solution: We wish to find n to ensure that

$$P\left(\left|\frac{S_n}{n} - d\right| \geq 0.1\right) < 0.05$$

We use the normal approximation formula to find n so that

$$2(1 - \Phi(\delta)) = 0.05$$

where

$$\delta = c\sqrt{n}/\sigma = \frac{0.1\sqrt{n}}{2} = \sqrt{n}20$$

In other words,

$$\Phi\left(\frac{\sqrt{n}}{20}\right) = 1 - 0.025 = 0.975$$

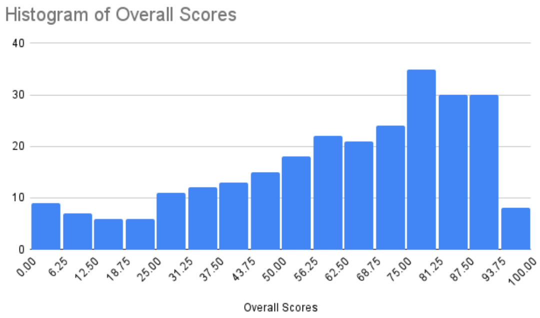
From the Standard Normal table, we see that

$$\sqrt{n}20 = 1.96 \Rightarrow n \approx 307.35$$

Hence, he should make 308 observations. □

d. Understanding the CLT

Remark 3.13. Suppose you have 10,000 people, each weighing between 30-100 kilograms.

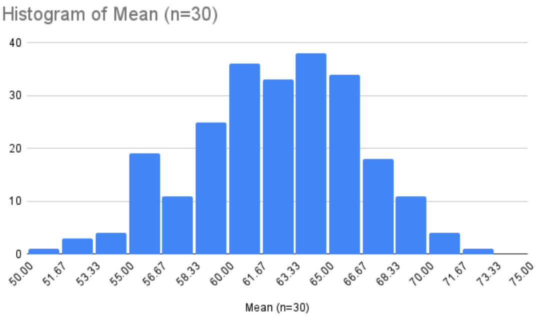


If you had all their weights, you could take an average to find the *population mean*.

(i) If you don't have all that data (or you cannot process so much), you could find the population mean by the following steps:

- Choose a group of 30 and take the average weight. This is the *sample mean*.
- Repeat this process 50 times. Each time you choose a group of 30 people and take an average, you get a different sample mean.
- Take the average of all these 50 values. This number will be close to the population mean.

(ii) You may also tabulate all the 50 sample means and draw a histogram. This histogram will resemble a bell curve.



- The mean of this bell curve is close to the population mean.
- The standard deviation (multiplied by $\sqrt{30}$) of this bell curve is close to the population standard deviation.

For part (i), you do not need the CLT. The Weak Law of Large Numbers assures you of it already. For part (ii), however, you do need the CLT.

Remark 3.14. What the CLT does not say:

(i) It does not say anything about any *individual's* weight. That is still a random phenomenon.

(ii) It does not answer questions like *What fraction of people weigh more than 60kg?*. i.e. *What is* $P(X > 60)$*?* where X is the random variable that assigns each person their weight. To answer this, we need the probability distribution for X, which we do not have!

(iii) The number 30 chosen above may not be good enough, depending on your population. The vague term 'close to' used above is also not precise unless you know something about the distribution of X.

VIII. Moment Generating Functions and Characteristic Functions

1. Moment Generating Function

Remark 1.1. If X is a random variable and $t \in \mathbb{R}$, then $Y := e^{tX}$ is a random variable given by

$$Y(w) := e^{tX(w)} = \sum_{n=0}^{\infty} \frac{t^n X(w)^n}{n!}.$$

Note that this series always converges, so Y is well-defined. It may or may not have finite expectation though.

Definition 1.2. The moment generating function of a random variable X is defined as

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

which is defined for all $t \in \mathbb{R}$ such that e^{tX} has finite expectation.

Example 1.3. If $X \sim \text{Bern}(p)$, then

$$\begin{aligned} M_X(t) &= E(e^{tX}) \\ &= e^{t\cdot 0} P(X = 0) + e^{t\cdot 1} P(X = 1) \\ &= (1 - p) + e^t p \end{aligned}$$

Remark 1.4. Suppose X is a random variable taking only positive integer values, then the probability generating function (see Definition III.5.2) is given by

$$\Phi_X(t) = \sum_{n=0}^{\infty} t^n P(X = n)$$

Hence,

$$M_X(t) = E(e^{tX}) = \sum_{n=0}^{\infty} e^{tn} P(X = n) = \Phi_X(e^t).$$

Example 1.5. If $X \sim B(n, p)$, then we know that the probability generating function is given by

$$\Phi_X(t) = (pt + 1 - p)^n$$

Hence,

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

Example 1.6. If $X \sim \text{Po}(\lambda)$, then we had computed in Example III.5.5 that

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

Therefore,

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

Example 1.7. Suppose $X \sim n(\mu, \sigma^2)$, then

$$\begin{aligned} M_X(t) = E(e^{tX}) &= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{tx} e^{-(x-\mu)^2/2\sigma^2} dx \\ &= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{t(y+\mu)} e^{-y^2/2\sigma^2} dy \\ &= \frac{e^{\mu t}}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{ty-y^2/2\sigma^2} dy \end{aligned}$$

Now note that

$$ty - \frac{y^2}{2\sigma^2} = -\frac{(y-\sigma^2 t)^2}{2\sigma^2} + \frac{\sigma^2 t^2}{2}$$

Hence,

$$\begin{aligned} M_X(t) &= \frac{e^{\mu t} e^{\sigma^2 t^2/2}}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(y-\sigma^2 t)^2/2\sigma^2} dy \\ &= e^{\mu t + \sigma^2 t^2/2} \end{aligned}$$

Example 1.8. If $X \sim \Gamma(\alpha, \lambda)$, then

$$\begin{aligned} M_X(t) &= \frac{\lambda^\alpha}{\Gamma(\alpha)} \int_0^\infty e^{tx} x^{\alpha-1} e^{-\lambda x} dx \\ &= \frac{\lambda^\alpha}{\Gamma(\alpha)} \int_0^\infty x^{\alpha-1} e^{(-(\lambda-t)x} dx \\ &= \frac{\lambda^\alpha}{\Gamma(\alpha)} \frac{\Gamma(\alpha)}{(\lambda-t)^\alpha} \end{aligned}$$

provided $-\infty < t < \lambda$. If $t \geq \lambda$, then this integral diverges. Hence,

$$M_X(t) = \left(\frac{\lambda}{\lambda - t}\right)^\alpha, -\infty < t < \lambda.$$

Lemma 1.9.

(i) If $c \in \mathbb{R}$ is a scalar and X is a random variable, then

$$M_{cX}(t) = M_X(ct).$$

(ii) If X and Y are independent, then

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

(iii) If $X_1, X_2, \ldots, X_n$ are i.i.d and $S_n = X_1 + X_2 + \ldots + X_n$, then

$$M_{S_n} = (M_{X_1}(t))^n.$$

Proof.

(i) Follows from the definition.

(ii) Here, e^{tX} and e^{tY} are also independent, so

$$M_{X+Y}(t) = E(e^{t(X+Y)}) = E(e^{tX}e^{tY}) = E(e^{tX})E(e^{tY}) = M_X(t)M_Y(t).$$

(iii) Follows from part (ii).

□

Theorem 1.10. *Suppose X is a random variable such that $M_X(t)$ is finite for all $t \in (-\delta, \delta)$ for some $\delta > 0$, then for each $n \in \mathbb{N}$,*

$$E(X^n) = \frac{d^n}{dt^n} M_X(t)|_{t=0}$$

Proof. Consider the series

$$\begin{aligned} M_X(t) = E(e^{tX}) &= E\left(\sum_{n=0}^{\infty} \frac{t^n X^n}{n!}\right) \\ &= \sum_{n=0}^{\infty} \frac{t^n}{n!} E\left(X^n\right) \\ &= 1 + tE(X) + \frac{t^2}{2!}E(X^2) + \ldots \end{aligned}$$

Differentiating term by term and plugging in $t = 0$ gives the desired result. All this works because the series converges absolutely by assumption. □

Example 1.11. If $X \sim n(0, \sigma^2)$, then

$$M_X(t) = e^{\sigma^2 t^2/2} = \sum_{n=0}^{\infty} \frac{\sigma^{2n} t^{2n}}{2^n n!}$$

Hence,

$$E(X^m) = 0$$

whenever m is odd. Moreover, if $m = 2n$ is even,

$$E(X^m) = \frac{\sigma^{2n}(2n)!}{2^n n!}.$$

In particular, $E(X^2) = \sigma^2$ (as we already know from Example VII.2.6).

2. Characteristic functions

Remark 2.1.

(i) A complex number is one of the form $z = x + iy$ where $x, y \in \mathbb{R}$ and $i^2 = -1$. Write $\mathbb{C}$ for the set of all complex numbers. For $z \in \mathbb{C}$, define

$$|z| = (x^2 + y^2)^{1/2}$$

and the distance between two complex numbers if $|z_1 - z_2$.

(ii) For $z \in \mathbb{C}$, define

$$e^z = \sum_{n=0}^{\infty} \frac{z^n}{n!}.$$

This series converges absolutely for each $z \in \mathbb{C}$, so defines a function $\exp : \mathbb{C} \to \mathbb{C}$. Note that

$$e^{z_1+z_2} = e^{z_1} e^{z_2}$$

Moreover, for any $t \in \mathbb{R}$,

$$e^{it} = \cos(t) + i\sin(t).$$

Then, $|e^{it}| = 1$ for all $t \in \mathbb{R}$.

(iii) Since $\cos(-t) = \cos(t)$ and $\sin(-t) = -\sin(t)$, we have

$$e^{-it} = \cos(t) - i\sin(t).$$

Thus,

$$\cos(t) = \frac{e^{it} + e^{-it}}{2}, \text{ and } \sin(t) = \frac{e^{it} - e^{-it}}{2i}$$

(iv) If $h(t) = f(t) + ig(t)$ is a complex valued function, then

$$h'(t) = f'(t) + ig'(t)$$

provided f and g are differentiable. Similarly,

$$\int h(t)dt = \int f(t)dt + i \int g(t)dt$$

provided the integrals exist.

(v) In particular,

$$\frac{d}{dt} e^{ct} = ce^{ct}$$

And

$$\int_a^b e^{ct} dt = \frac{e^{bc} - e^{ac}}{c}.$$

(vi) Let $(\Omega, \mathcal{A}, P)$ be a probability space. A complex valued function $Z : \Omega \to \mathbb{C}$ can be expressed in the form

$$Z(w) = X(w) + iY(w)$$

where X and Y are real-valued. We say that Z is a (complex) random variable if both X and Y are random variables.

(vii) We say that Z has finite expectation if X and Y both have finite expectation, which is equivalent to requiring that

$$E(|Z|) < \infty.$$

In that case, we define the expectation of Z as

$$EZ := E(X) + iE(Y).$$

(viii) Note that expectation is linear in this case as well:

$$E(a_1 Z_1 + a_2 Z_2) = a_1 E(Z_1) + a_2 E(Z_2)$$

whenever $a_1, a_2 \in \mathbb{C}$ and Z_1, Z_2 are complex random variables with finite expectation.

(ix) Given a real-valued random variable X, consider a new random variable $Y : \Omega \to \mathbb{C}$ by

$$Y(w) := e^{itX(w)} = \sum_{n=0}^{\infty} \frac{(itX(w))^n}{n!}$$

Definition 2.2. The characteristic function of X is defined as $\varphi_X : \mathbb{R} \to \mathbb{C}$ by

$$\varphi_X(t) := E(e^{itX})$$

(End of Day 27)

Remark 2.3.

(i) Whenever $M_X(t)$ is defined, we have

$$\varphi_X(t) = M_X(it).$$

However, φ_X is defined even when M_X may not be. Indeed, $|e^{itX(w)}| = 1$ for all $w \in \Omega$, so that

$$|\varphi_X(t)| \leq E(1) = 1$$

for all $t \in \mathbb{R}$.

(ii) If X has density function $f_X : \mathbb{R} \to \mathbb{R}$, then

$$\varphi_X(t) = \int_{-\infty}^{\infty} e^{itx} f_X(x) dt$$

This is a sum if X is a discrete random variable.

(iii) Moreover,

$$\varphi_X(0) = E(e^{it0}) = 1$$

(iv) If $c \in \mathbb{R}$ then

$$\varphi_{cX}(t) = E(e^{itcX}) = \varphi_X(ct).$$

(v) If X and Y are independent random variables, then

$$\begin{aligned}\varphi_{X+Y}(t) = E(e^{it(X+Y)}) &= E(e^{itX}e^{itY}) \\ &= E(e^{itX})E^{itY}) \\ &= \varphi_X(t)\varphi_Y(t)\end{aligned}$$

(vi) If $X = a$ is a constant random variable, then

$$\varphi_X(t) = E(e^{ita}) = e^{ita}$$

for all $t \in \mathbb{R}$.

(vii) Hence if X is a random variable and $a, b \in \mathbb{R}$, then

$$\varphi_{a+bX}(t) = e^{ita}\varphi_X(bt)$$

for all $t \in \mathbb{R}$.

Example 2.4. If $X \sim B(n, p)$, then

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

Hence,

$$\varphi_X(t) = (pe^{it} + 1 - p)^n$$

Example 2.5. Similarly, if $X \sim \text{Po}(\lambda)$, then

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

Example 2.6. If $X \sim U(a, b)$, then

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

Hence,

$$\begin{aligned}\varphi_X(t) &= \int_{-\infty}^{\infty} e^{itx} f_X(x)dt \\ &= \frac{1}{(b-a)}\int_a^b e^{itx}dt \\ &= \frac{e^{ibt}-e^{iat}}{it(b-a)}\end{aligned}$$

Example 2.7. If $X \sim n(\mu,\sigma^2)$, then we had seen that

$$M_X(t) = e^{\mu t+\sigma^2/t^2/2}.$$

Hence,

$$\varphi_X(t) = e^{it\mu-\sigma^2 t^2/2}.$$

Example 2.8. If $X \sim \mathrm{Exp}(\lambda)$, then

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & : \text{ if } x \geq 0 \\ 0 & : \text{ if } x < 0. \end{cases}$$

Hence,

$$\begin{aligned}\varphi_X(t) &= \lambda\int_0^{\infty} xe^{itx-\lambda x}dx \\ &= \frac{\lambda}{(\lambda-it)}e^{itx}|_0^{\infty} \\ &= \frac{\lambda}{(\lambda-it)}\end{aligned}$$

Note that unlike $M_X(t)$ in Example 1.8, this converges everywhere on $\mathbb{R}$.

Theorem 2.9. *If $X_1, X_2, \ldots, X_n$ are i.i.d. random variables and $S_n = X_1+X_2+\ldots+X_n$, then*

$$\varphi_{S_n}(t) = (\varphi_{X_1}(t))^n$$

Proof. Exactly as in Lemma 1.9. □

Remark 2.10. Note that

$$\varphi_X(t) = \sum_{n=0}^{\infty} \frac{i^n E(X^n)}{n!}t^n$$

This series converges absolutely, so by differentiating, we get

$$\varphi_X'(0) = iE(X)$$

More generally,

$$\varphi_X^{(n)}(0) = i^n E(X^n)$$

This can be used to compute the higher moments of X exactly as in Theorem 1.10.

3. Inversion Formulas and the Continuity Theorem

Remark 3.1.

(i) Recall that if X is a non-negative integer valued discrete random variable, then we defined the probability generating function as

$$\varphi_X(t) = \sum_{n=0}^{\infty} t^n f_X(n).$$

This can then be used to calculate the individual probabilities $f_X(n) = P(X = n)$ by repeated differentiation (see Remark III.5.3).

(ii) Now suppose X is any integer-valued random variable, then the characteristic function

$$\varphi_X(t) = \sum_{n=-\infty}^{\infty} e^{itn} f_X(n)$$

Then, we wish to use it to recover the values $f_X(n)$.

Lemma 3.2. *For any integers $j, k \in \mathbb{Z}$,*

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} e^{i(j-k)t} dt = \begin{cases} 1 & : \text{ if } j = k \\ 0 & : \text{ if } j \neq k \end{cases}$$

Proof. If $j = k$, then

$$e^{i(j-k)t} = 1$$

for all $t \in [-\pi, \pi]$. So,

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} e^{i(j-k)t} dt = 1.$$

If $j \neq k$, then

$$\begin{aligned}\int_{-\pi}^{\pi} e^{i(j-k)t} dt &= \frac{e^{i(j-k)t}}{(i(j-k)}|_{-\pi}^{\pi} \\ &= \frac{e^{i(j-k)\pi} - e^{-i(j-k)\pi}}{i(j-k)} \\ &= 2i\frac{\sin((j-k)\pi)}{i(j-k)} \\ &= 0\end{aligned}$$

because $\sin(m\pi) = 0$ for all $m \in \mathbb{Z}$. Hence the result. □

Theorem 3.3. *Let X be an integer-valued random variable. Then, for each $k \in \mathbb{Z}$,*

$$f_X(k) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \varphi_X(t) e^{-ikt} dt$$

This is called the inversion formula.

Proof. By definition,

$$\varphi_X(t) = \sum_{n=-\infty}^{\infty} e^{itn} f_X(n).$$

Therefore,

$$\begin{aligned}\frac{1}{2\pi}\int_{-\pi}^{\pi} \varphi_X(t)e^{-ikt}dt &= \sum_{n=-\infty}^{\infty} \frac{1}{\pi}\int_{-\pi}^{\pi} e^{-ikt}e^{itn} f_X(n)dt \\ &= \sum_{n=-\infty}^{\infty} f_X(n)\frac{1}{\pi}\int_{-\pi}^{\pi} e^{it(n-k)}dt \\ &= f_X(k)\end{aligned}$$

by Lemma 3.2. □

The following is an analogue for continuous random variables.

Definition 3.4. We say that a function $g : \mathbb{R} \to \mathbb{R}$ is integrable if

$$\int_{-\infty}^{\infty} |g(t)|dt < \infty.$$

Theorem 3.5 (Inversion Formula)**.** *Suppose that X is a continuous random variable such that φ_X is integrable, then for each $x \in \mathbb{R}$, we have*

$$f_X(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-ixt}\varphi_X(t)dt.$$

Example 3.6. Suppose X is a random variable such that

$$\varphi_X(t) = e^{-\sigma^2 t^2/2}$$

This function is integrable so the density function is given by

$$\begin{aligned}f_X(x) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-ixt}e^{-\sigma^2t^2/2}dt \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-ixt-\sigma^2t^2/2}dt\end{aligned}$$

Note that

$$ixt + \frac{\sigma^2t^2}{2} = \frac{2ixt+\sigma^2t^2}{2} = \frac{(\sigma t + ix/\sigma)^2}{2} + \frac{x^2}{2\sigma^2}$$

Hence,

$$\begin{aligned} f_X(x) &= \frac{e^{-x^2/2\sigma^2}}{2\pi}\int_{-\infty}^{\infty} e^{-(\sigma t+ix/\sigma)^2/2}dt \\ &= \frac{e^{-x^2/2\sigma^2}}{2\pi\sigma}\int_{-\infty}^{\infty} e^{-y^2/2}dy \\ &= \frac{1}{\sigma\sqrt{2\pi}}e^{-x^2/2} \end{aligned}$$

Hence, $X \sim n(0, \sigma^2)$.

Remark 3.7. Suppose X and Y are two random variables such that

$$\varphi_X(t) = \varphi_Y(t)$$

for all $t \in \mathbb{R}$, and suppose we know that φ_X is integrable. Then, by Theorem 3.5, we have

$$f_X(x) = f_Y(x)$$

for all $x \in \mathbb{R}$. Hence, $X \sim Y$. The next theorem tells us that this result is true even if we do not assume that φ_X is integrable.

Theorem 3.8 (Uniqueness Theorem)**.** *If X and Y are two random variables with the same characteristic functions, then X and Y have the same distribution function.*

(End of Day 28)

Example 3.9. Suppose $X \sim n(\mu_1, \sigma_1^2)$ and $Y \sim n(\mu_2, \sigma_2^2)$ are two independent normal random variables. Then,

$$\varphi_X(t) = e^{i\mu_1 t - \sigma_1^2 t^2/2} \text{ and } \varphi_Y(t) = e^{i\mu_2 t - \sigma_2^2 t^2/2}.$$

Since they are independent, we know that

$$\varphi_{X+Y}(t) = \varphi_X(t)\varphi_Y(t) = e^{i(\mu_1+\mu_2)t - (\sigma_1^2+\sigma_2^2)t^2/2}.$$

By the uniqueness theorem, it follows that

$$X + Y \sim n(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2).$$

This was the conclusion of Theorem VI.2.3.

Theorem 3.10 (Continuity Theorem)**.** *Let (X_n) be a sequence of random variables and X be a random variable such that*

$$\lim_{n\to\infty} \varphi_{X_n}(t) = \varphi_X(t)$$

for each $t \in \mathbb{R}$. Then,

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

for each $x \in \mathbb{R}$.

Proof. Assume that all the characteristic functions are (uniformly) integrable. Then, by a convergence theorem from analysis, we can say that, for each $x \in \mathbb{R}$,

$$\lim_{n\to\infty} \int_{-\infty}^{\infty} e^{-itx} \varphi_{X_n}(t)dt = \int_{-\infty}^{\infty} e^{-itx} \varphi_X(t)dt.$$

By the inversion formula, we conclude that

$$\lim_{n\to\infty} f_{X_n}(x) = f_X(x).$$

Again, by the same theorem, we see that

$$\lim_{n\to\infty} F_{X_n}(x) = \lim_{n\to\infty} \int_{-\infty}^{x} f_{X_n}(y)dy = \int_{-\infty}^{x} f_X(y)dy = F_X(x).$$

Now if the characteristic functions are not integrable, one must apply this theorem to the function $\varphi_X(t)e^{-c^2t^2/2}$ (see the book for the details). □

4. The Weak Law of Large Numbers and the Central Limit Theorem

Remark 4.1. (i) Let $z \in \mathbb{C}$ be a complex number such that $|z-1| < 1$, then we can define

$$\log(z) = (z-1) - \frac{(z-1)^2}{2} + \frac{(z-1)^3}{3} + \ldots = \sum_{n=1}^{\infty} (-1)^{n-1} \frac{(z-1)^n}{n}$$

Then, one sees that

- $e^{\log(z)} = z$ for all $z \in \mathbb{C}$ with $|z-1| < 1$.
- $\log(1) = 0$.
- If $h : (a,b) \to \mathbb{C}$ is a complex valued function on (a,b) such that $|h(t)-1| < 1$ for all $t \in (a,b)$, then
$$\frac{d}{dt}\log(h(t)) = \frac{h'(t)}{h(t)}$$

(ii) Now let X be a random variable with characteristic function φ_X. Then,

$$\varphi_X(0) = 1$$

and φ_X is continuous. Hence, there is an interval $(-\delta, \delta)$ such that

$$\log(\varphi_X(t))$$

is well-defined on that interval. Fix one such interval for now.

(iii) Suppose X has finite mean μ. Then,

$$\begin{aligned}
\varphi_X'(0) &= i\mu \\
\Rightarrow \lim_{t\to 0}\frac{\log(\varphi_X(t))}{t} &= \lim_{t\to 0}\frac{\log(\varphi_X(t)) - \log(\varphi_X(0))}{t-0} \\
&= \frac{d}{dt}\log(\varphi_X(t))|_{t=0} \\
&= \frac{\varphi_X'(0)}{\varphi_X(0)} \\
&= i\mu
\end{aligned}$$

Hence,

$$\lim_{t\to 0}\frac{\log(\varphi_X(t)) - i\mu t}{t} = 0.$$

(iv) Now suppose X has finite variance σ^2. Then,

$$\varphi_X''(0) = -E(X^2) = -(\mu^2+\sigma^2)$$

Now by L'Hospital's rule,

$$\begin{aligned}
\lim_{t\to 0}\frac{\log(\varphi_X(t)) - i\mu t}{t^2} &= \lim_{t\to 0}\frac{\frac{\varphi_X'(t)}{\varphi_X(t)} - i\mu}{2t} \\
&= \lim_{t\to 0}\frac{\varphi_X'(t) - i\mu\varphi_X(t)}{2t\varphi_X(t)} \\
&= \lim_{t\to 0}\frac{\varphi_X'(t) - i\mu\varphi_X(t)}{2t} \\
&= \lim_{t\to 0}\frac{\varphi_X''(t) - i\mu\varphi_X'(t)}{2} \\
&= \frac{\varphi_X''(0) - i\mu\varphi_X'(0)}{2} \\
&= \frac{-(\mu^2+\sigma^2)+\mu^2}{2} \\
&= -\frac{\sigma^2}{2}.
\end{aligned}$$

(v) Let $Z : \Omega \to \mathbb{R}$ denote the random variable

$$Z(w) = 0 \text{ for all } w \in \Omega.$$

Then, $P(Z=0)=1$, so

$$\varphi_Z(t) = e^{i\cdot 0}P(Z=0) = 1.$$

Moreover, the distribution function of Z is

$$F_Z(x) = \begin{cases} 0 & : \text{ if } x < 0 \\ 1 & : \text{ if } x \geq 0 \end{cases}$$

Theorem 4.2 (Weak Law of Large Numbers)**.** *Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables having finite mean μ. Set*

$$S_n := X_1 + X_2 + \ldots + X_n$$

Then, for any $\epsilon > 0$,

$$\lim_{n\to\infty} P\left(\left|\frac{S_n}{n} - \mu\right| > \epsilon\right) = 0.$$

Proof. By Theorem 2.9,

$$\varphi_{S_n}(t) = (\varphi_{X_1}(t))^n$$

Hence, if $Z_n := \frac{S_n}{n} - \mu$, then by Remark 2.3,

$$\varphi_{Z_n}(t) = e^{-i\mu t}(\varphi_{X_1}(t/n))^n$$

Fix $t \in \mathbb{R}$ and choose $n \in \mathbb{N}$ so that t/n is close enough to zero so that

$$|\varphi_X(t/n) - 1| = |\varphi_X(t/n) - \varphi_X(0)| < 1.$$

Then, $\log(\varphi_X(t/n))$ is well-defined. Now note that

$$\varphi_{Z_n}(t) = \exp[n \log(\varphi_X(t/n) - i\mu(t/n))]$$

Now note that

$$\begin{aligned}\lim_{n\to\infty} n\log(\varphi_X(t/n) - i\mu(t/n)) &= t \lim_{n\to\infty} \frac{\varphi_X(t/n) - i\mu(t/n)}{t/n} \\ &= 0\end{aligned}$$

by part (iii) of Remark 4.1. Hence,

$$\lim_{n\to\infty} \varphi_{Z_n}(t) = \exp(0) = 1.$$

This is true for any $t \in \mathbb{R}$, so if Z denotes the random variable

$$Z(w) = 0 \text{ for all } w \in \Omega$$

then

$$\lim_{n\to\infty} \varphi_{Z_n}(t) = \varphi_Z(t)$$

by part (v) of Remark 4.1. Therefore, by the Continuity Theorem (Theorem 3.10),

$$\lim_{n\to\infty} F_{Z_n}(x) = F_Z(x).$$

Hence, for any $\epsilon > 0$, we have

$$\begin{aligned}
\lim_{n\to\infty} P(Z_n < -\epsilon) &= \lim_{n\to\infty} F_{Z_n}(-\epsilon) \\
&= F_Z(-\epsilon) = 0 \\
\lim_{n\to\infty} P(Z_n > \epsilon) &= 1 - \lim_{n\to\infty} P(Z_n < \epsilon) \\
&= 1 - \lim_{n\to\infty} F_{Z_n}(\epsilon) \\
&= 1 - F_Z(\epsilon) \\
&= 0
\end{aligned}$$

Hence,

$$\lim_{n\to\infty} P(|Z_n| > \epsilon) = 0.$$

This proves the theorem. □

(End of Day 29)

Theorem 4.3 (Central Limit Theorem)**.** *Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables with finite mean μ and finite non-zero variance σ^2. Then for any $x \in \mathbb{R}$,*

$$\lim_{n\to\infty} P\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \leq x\right) = \Phi(x)$$

Proof. Define

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

Then fix $t \in \mathbb{R}$. By Remark 2.3,

$$\begin{aligned}
\varphi_{S_n^*}(t) &= e^{-in\mu t/\sigma\sqrt{n}} \varphi_{S_n}(t/\sigma\sqrt{n}) \\
&= e^{-in\mu t/\sigma\sqrt{n}} (\varphi_{X_1}(t/\sigma\sqrt{n}))^n \\
&= \exp\left[\frac{-in\mu t}{\sigma\sqrt{n}} + n\log(\varphi_{X_1}(t/\sigma\sqrt{n}))\right]
\end{aligned}$$

Consider

$$\begin{aligned}
\lim_{n\to\infty} \frac{-in\mu t}{\sigma\sqrt{n}} + n\log(\varphi_{X_1}(t/\sigma\sqrt{n})) &= \frac{t^2}{\sigma^2} \lim_{n\to\infty} \frac{\log(\varphi_{X_1}(t/\sigma\sqrt{n}) - i\mu(t/\sigma\sqrt{n})}{(t/\sigma\sqrt{n})^2} \\
&= \frac{t^2}{\sigma^2}\frac{-\sigma^2}{2} \\
&= \frac{-t^2}{2}
\end{aligned}$$

whenever $t \neq 0$ by part (iv) of Remark 4.1. Moreover,

$$\lim_{n\to\infty} \frac{-in\mu t}{\sigma\sqrt{n}} + n\log(\varphi_{X_1}(t/\sigma\sqrt{n})) = \frac{-t^2}{2}$$

even when $t = 0$. Hence, we conclude that for any $t \in \mathbb{R}$,

$$\lim_{n\to\infty} \varphi_{S_n^*}(t) = e^{-t^2/2}$$

This is the characteristic function of $Z \sim n(0,1)$. So by the Continuity Theorem Theorem 3.10, we have that for any $x \in \mathbb{R}$,

$$\lim_{n\to\infty} F_{S_n^*}(x) = \Phi(x).$$

Hence the result. □

Example 4.4. Suppose a fair coin is tossed 100 times. Use a normal approximation to estimate the probability that there will be more than 60 heads.

Solution: Here, $X = \text{Bern}(1/2)$ represents a single coin toss, so we are interested in $S_{100} = \sum_{i=1}^{100} X_i$ where the X_i are i.i.d. random variables with $X_i \sim \text{Bern}(1/2)$. We wish to determine

$$P(S_{100} > 60) = 1 - P(S_{100} \leq 60)$$

We know that

$$\begin{aligned} E(S_{100}) &= 100E(X) = 50 \\ \text{Var}(S_{100}) &= 100\text{Var}(X) = 100(1/2)(1-1/2) = 25. \end{aligned}$$

So we use the normal approximation formula to estimate

$$\begin{aligned} P(S_{100} > 60) &\approx 1 - \Phi\left(\frac{60-50}{\sqrt{25}}\right) \\ &= 1 - \Phi(2) \\ &= 0.0228. \end{aligned}$$

□

Example 4.5. A drug is supposed to be 75% effective. It is tested on 100 people. Using a normal approximation, estimate the probability that at least 70 people will be cured.

Solution: Again, if X is the random variable so that $X(\omega) = 1$ if ω is cured and $X(\omega) = 0$ if not, then $X \sim \text{Bern}(0.75)$ by hypothesis. We let $X_1, X_2, \ldots, X_{100}$ denote the 100 experiments with $X_i \sim X$ and set

$$S_{100} = \sum_{i=1}^{100} X_i$$

We wish to determine

$$P(S_{100} \geq 70).$$

We know that

$$\begin{aligned} E(S_{100}) &= 100E(X) = 75 \\ \text{Var}(S_{100}) &= 100\text{Var}(X) = 100 \times 0.75 \times 0.25 = 18.75. \end{aligned}$$

Hence,

$$\begin{aligned} P(S_{100} \geq 70) &= 1 - P(S_{100} \leq 69) \\ &= 1 - \Phi\left(\frac{69 - 75}{\sqrt{18.75}}\right) \\ &\approx 0.87. \end{aligned}$$

Hence, there is an 87% chance that at least 70 people will be cured. □

5. Review and Important Formulae

(i) Probability Measure:

- (Sum Rule) If $A_1, A_2, \dots$ are mutually disjoint, then $P(\bigsqcup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} P(A_n)$.
- $P(A^c) = 1 - P(A)$.
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.
- $P(B|A) = \frac{P(A \cap B)}{P(A)}$.
- Bayes' Rule: $P(A_j|B) = \frac{P(A_j)P(B|A_j)}{\sum_{i=1}^{n} P(A_i)P(B|A_i)}$
- Independent Events: $P(A \cap B) = P(A)P(B)$.

(ii) Density function:

- Discrete RV: $f_X(x) = P(X = x)$.
- Continuous RV: $f_X(x) = F'_X(x)$.

(iii) Distribution function: $F_X(x) = P(X \leq x)$.

(iv) Probability of events related to a RV:

- Discrete RV: $P(X \in B) = \sum_{x \in B} P(X = x) = \sum_{x \in B} f_X(x)$.
- Continuous RV: $P(X \in B) = \int_B f_X(x)dx$
- Discrete RV: $P(X \leq x) = \sum_{t \leq x} P(X = t)$
- Continuous RV: $P(X \leq x) = \int_{-\infty}^{x} f_X(t)dt$.

(v) Jointly distributed RVs:

- $P((X,Y)\in B)) = \int\int_B f(x,y)dydx$
- Marginal of X: $f_X(x) = \int_{-\infty}^{\infty} f(x,y)dy$.
- X and Y independent $\Leftrightarrow f(x,y) = f_X(x)f_Y(y)$.

(vi) Important Distributions:

- Discrete:

Name	Density Function	Set of values	$E(X)$	Var(X)
Bern(p)	$p^x(1-p)^{1-x}$	$\{0,1\}$	p	$p(1-p)$
$B(n,p)$	$\binom{n}{x}p^x(1-p)^{n-x}$	$\{0,1,2,\ldots,n\}$	np	$np(1-p)$
Po(λ)	$e^{-\lambda}\frac{\lambda^x}{x!}$	$\{0,1,\ldots\}$	λ	λ
Geom(p)	$p(1-p)^{x-1}$	$\{1,2,\ldots\}$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
Hyp(n,r,N)	$\frac{\binom{r}{x}\binom{N-r}{n-x}}{\binom{N}{n}}$	$\{0,1,\ldots,n\}$	np	$n\frac{r}{N}(1-\frac{r}{N})\frac{N-n}{N-1}$

- Continuous:

Name	Density Function	Set of values	$E(X)$	Var(X)
$U(a,b)$	$\frac{1}{(b-a)}$	(a,b)	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Exp(λ)	$\lambda e^{-\lambda x}$	$[0,\infty)$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Gamma(α,λ)	$\frac{\lambda^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\lambda x}$	$[0,\infty)$	$\frac{\alpha}{\lambda}$	$\frac{\alpha}{\lambda^2}$
$n(\mu,\sigma^2)$	$\frac{1}{\sigma\sqrt{2\pi}}e^{-(x-\mu)^2/2\sigma^2}$	$\mathbb{R}$	μ	σ^2

(vii) Expectation:

- Discrete RV: $E(X) = \sum_x xP(X=x)$
- Continuous RV: $E(X) = \int_{-\infty}^{\infty} xf(x)dx$.
- Function of RV: $E(\varphi(X)) = \int_{-\infty}^{\infty} \varphi(x)f(x)dx$.
- Linearity: $E(aX+bY) = aE(X)+bE(Y)$.
- If X and Y are independent, then $E(XY) = E(X)E(Y)$.
- Markov Inequality: If X is non-negative, then

$$P(X\geq t)\leq \frac{E(X)}{t}.$$

- Chebyshev's inequality:

$$P(|X-\mu|\geq t)\leq \frac{\sigma^2}{t^2}.$$

(End of Day 30)

(viii) Variance and Covariance:

- $\text{Var}(X) = E(X^2) - E(X)^2$.
- $\text{Var}(a + bX) = b^2\text{Var}(X)$
- $\text{Cov}(X, Y) = E(XY) - E(X)E(Y)$.
- If X and Y are independent, then $\text{Cov}(X, Y) = 0$.
- $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + \text{Cov}(X, Y)$.

(ix) Probability Generating Functions: If RV takes values in $\{0, 1, 2 \ldots, \}$.

- $\Phi_X(t) = \sum_{x=0}^{\infty} f_X(x)t^x$ (for $|z| < 1$).
- $f_X(n) = \frac{1}{n!}\Phi_X^{(n)}(0)$.
- $E(X) = \Phi'_X(1)$
- $\text{Var}(X) = \Phi''_X(1) + \Phi'_X(1) - (\Phi'_X(1))^2$.

(x) Moment Generating Functions: If e^{tX} has finite expectation.

- $M_X(t) = E(e^{tX})$.
- $E(X^n) = M_X^{(n)}(0)$.
- If X and Y are independent, then $M_{X+Y}(t) = M_X(t)M_Y(t)$.

(xi) Characteristic Function: Defined for all $t \in \mathbb{R}$.

- $\varphi_X(t) = E(e^{itX})$.
- $\varphi_{a+bX}(t) = e^{ita}\varphi_X(bt)$.
- If X and Y are independent, then $\varphi_{X+Y}(t) = \varphi_X(t)\varphi_Y(t)$.
- Inversion Formula: If φ_X is integrable, then $f_X(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-ixt}\varphi_X(t)dt$.
- Uniqueness Theorem: If $\varphi_X = \varphi_Y$, then $X \sim Y$.
- Continuity Theorem: If $\varphi_{X_n}(t) \to \varphi_X(t)$ for all $t \in \mathbb{R}$, then $F_{X_n}(x) \to F_X(x)$ for all $x \in \mathbb{R}$.

Distribution	PGF	MGF	Char Fn
Bern(p)	$(tp + 1 - p)$	$\Phi_X(e^t)$	$M_X(it)$
$B(n, p)$	$(tp + 1 - p)^n$	”	”
Po(λ)	$e^{\lambda(t-1)}$	”	”
Geom(p)	$\frac{p}{1-t(1-p)}$ (for $\lvert t\rvert < \frac{1}{1-p}$)	”	”
$U(a, b)$	NA	$\frac{e^{bt}-e^{at}}{t(b-a)}$	$M_X(it)$
$\Gamma(\alpha, \lambda)$	NA	$\left(\frac{\lambda}{\lambda-t}\right)^{\alpha}$ (for $t < \lambda$)	” (for all t)
$n(\mu, \sigma^2)$	NA	$e^{t\mu+\sigma^2t^2/2}$	”

(xii) Weak Law of Large Numbers:

$$\lim_{n\to\infty} P\left(\left|\frac{S_n}{n} - \mu\right| > \epsilon\right) = 0.$$

(xiii) Central Limit Theorem:

$$\lim_{n\to\infty} P\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \leq x\right) = \Phi(x)$$

where Φ is the distribution function for the standard normal.

(xiv) Normal Approximation Formulae:

$$P(S_n \leq x) \approx \Phi\left(\frac{x - n\mu}{\sigma\sqrt{n}}\right)$$

$$P\left(\left|\frac{S_n}{n} - \mu\right| \geq c\right) \approx 2(1 - \Phi(c\sqrt{n}/\sigma))$$

(You will be given the standard normal table on an exam)

(End of Day 31)

IX. Instructor Notes

(i) The course is good and the textbook [HPS] is excellent. I followed it almost verbatim and completed eight chapters (more or less). It may have been possible to do more (I lost some classes to holidays), but the syllabus is definitely larger than necessary.

(ii) I avoided most proofs, but I did some for the interested students. I feel that the balance was correct.

(iii) The class response was mixed. Some students were clearly very interested and worked hard. However, one-third of the class was repeating the course. This meant that many students were there just to mark their attendance and leave, which left the others feeling distracted and disheartened. Indeed, it was extremely annoying for me as well.

(iv) Another issue is that I was teaching two lectures at 6:00PM, which meant that students were very tired by then. This significantly impacted their enthusiasm.

(v) Overall, the course was fun though and can largely be repeated as is.

Bibliography

[HPS] Hoel, Port, Stone, *Introduction to Probability Theory*, Houghton Mifflin Co. (1971)

[Kroese] D.P. Kroese, *A Short Introduction to Probability*, https://people.smp.uq.edu.au/DirkKroese/asitp.pdf

[Rohatgi-Saleh] Rohatgi, Saleh, *An Introduction to Probability and Statistics (2nd Edition)*, Wiley (2001)

[Ross] Ross, *A First Course in Probability (Ninth Edition)*, Pearson (2014)