M40005 term 1 revision

Teddy Wu

April 2025

1 Taylor Expansions and the Method of Differential identities

Two common Taylor expansion that may appear in exam are

$$\sum_{n=0}^{\infty} \frac{x^n}{n!} = e^x \quad \text{and} \quad \sum_{n=1}^{\infty} \frac{x^n}{n} = -\ln(1-x)$$

We now introduce the method of differential identities, which can help us obtain certain infinite series easily. We start by the geometric series

$$\sum_{n=0}^{\infty} x^n = \frac{1}{1-x}$$

Now, differentiate both side with respect to x:

$$\frac{\partial}{\partial x} \sum_{n=0}^{\infty} x^n = \frac{d}{dx} \frac{1}{1-x}$$

By assuming that we can interchange the order of differentiation and summation, we obtain

$$\sum_{n=0}^{\infty} n x^{n-1} = \frac{1}{(1-x)^2}$$

Hence

$$\sum_{n=0}^{\infty} n x^n = \frac{x}{(1-x)^2}$$

What makes this method useful is we can continue to obtain new series. Differentiate again yields

$$\sum_{n=1}^{\infty} n^2 x^{n-1} = \frac{(1-x)^2 + 2x(1-x)}{(1-x)^4}$$

$$\sum_{n=0}^{\infty} n^2 x^n = \frac{x+x^2}{(1-x)^3}$$

We can also integrate instead of taking derivative. Start with

$$\sum_{n=1}^{\infty} x^{n-1} = \frac{1}{1-x}$$

Integrate both sides with respect to x and interchange the order of integration and summation, we obtain

$$\sum_{n=1}^{\infty} \frac{x^n}{n} = -\ln(1-x)$$

and note that we obtained the Taylor series for $\ln(1-x)$.

Although this method is useful, it is not guaranteed to get marks in exam since interchanging differentiation and summation is not justified. Use Taylor series if you can and leave this method as a last resort.

As a fun extension, we will use this method to obtain the Taylor series for e^x. Let $f(x) = \sum_{n=0}^{\infty} \frac{x^n}{n!}$. Taking derivative of each sides we have

$$f'(x) = \frac{\partial}{\partial x} \sum_{n=0}^{\infty} \frac{x^n}{n!} = \sum_{n=0}^{\infty} \frac{nx^{n-1}}{n!} = \sum_{n=1}^{\infty} \frac{x^{n-1}}{(n-1)!} = \sum_{n=0}^{\infty} \frac{x^n}{n!} = f(x)$$

Solving the differential equation we obtain

$$\sum_{n=0}^{\infty} \frac{x^n}{n!} = Ae^x$$

Finally, plug in $x = 0$ yields $A = 1$, and we are done.

Below are two applications of infinite series:

The number of permutations of n objects is $n!$. We call the permutations that does not fix any object a derangement. A classical result in combinatorics states that the proportion of derangements of n objects tends to $\frac{1}{e}$ as n tends to infinity. (2021 Jan q1(b))

Let A_i denote that set of permutations that fix object i in its initial position. Then the set

$$A = A_1 \cup A_2 \cup ... \cup A_n$$

consists all permutation that fixes at least one object. By inclusion-exclusion principle:

$$|A_1 \cup A_2 \cup ... \cup A_n| = n|A_1| - \binom{n}{2} |A_1 \cap A_2| + \binom{n}{3} |A_1 \cap A_2 \cap A_3| - ... + (-1)^{n+1} |A_1 \cap A_2 \cap ... \cap A_n|$$

What is $|A_1 \cap ... \cap A_i|$? This set consists all permutations that fixes object $1, ..., i$, so we can permute the remaining $n - i$ objects, thus there are $(n-i)!$ such permutation.

$$|A_1 \cup A_2 \cup ... \cup A_n| = \binom{n}{1} \cdot (n-1)! - \binom{n}{2} \cdot (n-2)! + \binom{n}{3} \cdot (n-3)! - ... + \binom{n}{n} (-1)^{n+1} \cdot (n-n)!$$

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

$$= -n! \sum_{k=1}^{n} \frac{(-1)^k}{k!}$$

$$\approx -n! \sum_{k=1}^{\infty} \frac{(-1)^k}{k!}$$

$$= -n!(e^{-1} - 1)$$

Set A consists every permutation that is not a derangement. Thus, the number of derangement is just

$$n! + n!(e^{-1} - 1) = \frac{n!}{e}$$

We now consider (2020 final q2(d)): Imagine you toss a fair coin repeatedly. You denote by H the outcome Heads and by T the outcome Tails. How many times, on average, do you need to toss the coin to see the pattern HT for the first time?

Although there are clever solutions to this problem, suppose in exam you run out of ideas and have to brute force this problem. Let's define the random variable X be to the number of tosses needed until HT first appears, and we want to find $\mathrm{E}(X)$.

Clearly $P(X = 1) = 0$. $P(X = 2) = \frac{1}{4}$ because out of four possibilities there is only one in which HT appears. $P(X = 3) = \frac{2}{8}(HHT, THT)$ and $P(X = 4) = \frac{3}{16}(HHHT, THHT, TTHT)$. We can see the pattern that $P(X = k) = \frac{k-1}{2^k}$. Now

$$\mathrm{E}(X) = \sum_{k=0}^{\infty} kP(X = k) = \sum_{k=0}^{\infty} \frac{k^2 - k}{2^k}$$

$$= \sum_{k=0}^{\infty} k^2 \left(\frac{1}{2}\right)^k - \sum_{k=0}^{\infty} k \left(\frac{1}{2}\right)^k$$

We can thus apply the formula above to obtain

$$\mathrm{E}(X) = \frac{\frac{1}{2} + \frac{1}{4}}{(\frac{1}{2})^3} - \frac{\frac{1}{2}}{(\frac{1}{2})^2} = 6 - 2 = 4.$$

2 Axiomatic Definition of Probability

Suppose someone tosses a fair coin 2 times. There are four possible outcomes: HH, HT, TH, TT. Let $\Omega = \{HH, HT, TH, TT\}$. Ω is called the sample space of the event.

Now we want to define probability formally. For instance, one may ask the probability that we tossed HH ($\frac{1}{4}$), or the probability that we tossed HH or TT ($\frac{1}{2}$). To begin we must first understand what is an event. An event can be thought of a subset of Ω, e.g. $\{HH\}$ or $\{HH, TT\}$ corresponding to the previous example. It's clear that our probability function should map an event to a number. But which events should we consider? First let $\mathcal{F}$ be the collection of all subsets of Ω (events) that we assign a probability on it. $\mathcal{F}$ is called the event space.

Suppose $A \in \mathcal{F}$, so we assigned a probability to the subset (event) A. Then the complement of A, A^c correspond to the event of A not happening, to which we should assign a probability. Thus A^c should also be in $\mathcal{F}$, and we say $\mathcal{F}$ is closed under complement.

Suppose $A, B \in \mathcal{F}$, then the union $A \cup B$ correspond to the event that A or B happening, to which we should also assign a probability: $A \cup B \in \mathcal{F}$, and we say $\mathcal{F}$ is closed under union.

Let $\emptyset \in \mathcal{F}$ for consistency. Such $\mathcal{F}$ is (almost) called a σ-algebra. σ-algebra requires additionally that, $\mathcal{F}$ is closed under countable union.

Our probability function is thus $P : \mathcal{F} \to \mathbb{R}$. Since probability should be positive, we require $P(A) \geq 0$ for all $A \in \mathcal{F}$. We also want $P(\Omega) = 1$ since the probability of everything happening should be 1. Finally, we want

$$P\left(\bigcup_{i \in \mathcal{I}} A_i\right) = \sum_{i \in \mathcal{I}} P(A_i)$$

for any countable sequence of disjoint events $(A_i)_{i \in \mathcal{I}}$. (Note that $\mathcal{F}$ being a σ-algebra ensures that $\bigcup_{i \in \mathcal{I}} A_i$ is in $\mathcal{F}$). Such P is called a probability measure. The reason we want this property is that it makes sense to let $P(A \text{ or } B) = P(A) + P(B)$ if A and B are disjoint events.

We say $(\Omega, \mathcal{F}, P)$ is a probability space.

The formal definitions are summarized below:

σ-algebra: A collection of subsets of Ω denoted by $\mathcal{F}$ is an σ-algebra on Ω if

(i) $\emptyset \in \mathcal{F}$;

(ii) $\mathcal{F}$ is closed under complement;

(iii) $\mathcal{F}$ is closed under countable union.

Probability measure: Let $\mathcal{F}$ be a σ-algebra on Ω. A mapping $P : \mathcal{F} \to \mathbb{R}$ is called a probability measure on $(\Omega, \mathcal{F})$ if

a(i) $P(A)$ for all events $A \in \mathcal{F}$;

a(ii) $P(\Omega) = 1$;

a(iii) $P\left(\bigcup_{i \in \mathcal{I}} A_i\right) = \sum_{i \in \mathcal{I}} P(A_i)$.

Probability space: Let $\mathcal{F}$ be a σ-algebra on Ω and P be a probability measure on $(\Omega, \mathcal{F})$. We say $(\Omega, \mathcal{F}, P)$ is a probability space. Ω is called the sample space and $\mathcal{F}$ is called the event space.

Properties of probability measure:

b(i) $P(A^c) = 1 - P(A)$;

b(ii) If $A \subseteq B$ then $P(A) \leq P(B)$;

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

Proof: To prove these results we will frequently make use of a(iii) in probability measure.

b(i) Since Ω can be written as disjoint union $A \cup A^c$ we have

$$P(\Omega) \overset{\text{a(iii)}}{=} P(A) + P(A^c)$$

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

Particularly, $P(\Omega) = 1 - P(\emptyset) = 1$.

b(ii) Before we start, note that a common way to create disjoint union is to write

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

If $A \subseteq B$, write B as disjoint union $(B \cap A) \cup (B \cap A^c) = A \cup (B \cap A^c)$

$$P(B) \overset{\text{a(iii)}}{=} P(A) + P(B \cup A^c)$$

$$P(B) - P(A) = P(B \cup A^c) \geq 0$$

b(iii) Write A and B as disjoint union

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

$$P(A) \overset{\text{a(iii)}}{=} P(A \cap B) + P(A \cap B^c), P(B) \overset{\text{a(iii)}}{=} P(B \cap A) + P(B \cap A^c)$$

Now write $A \cup B$ as disjoint union

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

$$P(A \cup B) \overset{\text{a(iii)}}{=} P(A \cap B) + P(A^c \cap B) + P(A \cap B^c)$$

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

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

In most cases the event space $\mathcal{F} = \mathcal{P}(A)$, the power set of A. Returning to the coin tossing example,

$$\Omega = \{HH, HT, TH, TT\}$$

$$\mathcal{F} = \{\emptyset, \{HH\}, \{TT\}, \{HT\}, \{TH\}, \{HH, HT\}, ..., \Omega\}$$

To define a probability measure that is consistent with our common sense, let $P : \mathcal{F} \to \mathbb{R}$ defined by

$$P(A) = \frac{\text{card}(A)}{\text{card}(\Omega)}$$

For instance, $P(\{HH\}) = \frac{1}{4}$ and $P(\{HH, HT\}) = \frac{2}{4} = \frac{1}{2}$. Such P is called naive probability measure. (2021 final q1(b)(iii), 2022 final q1(a)(iii))

Note that such definition only works when Ω is finite. When Ω is infinite a possible definition for

probability measure is

$$P(A) = \frac{\text{area of } A}{\text{area of } \Omega}$$

(2021 final q1(a))

3 Conditional Probability and Independence

Conditional probability: The conditional probability of A given B is defined as

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

Thus $P(A \cap B) = P(A|B)P(B)$. But note that $P(B \cap A) = P(B|A)P(A)$. From this we can derive Bayes' rule:

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

Law of total probability: Let $(B_i)_{i \in \mathcal{I}}$ be a partition of Ω. Then

$$P(A) = \sum_{i \in \mathcal{I}} P(A \cap B_i) = \sum_{i \in \mathcal{I}} P(A|B_i)P(B_i)$$

Proof:

$$P(A) = P(A \cap \Omega) = P\left(A \cap \bigcup_{i \in \mathcal{I}} B_i\right) = P\left(\bigcup_{i \in \mathcal{I}} (A \cap B_i)\right) = \sum_{i \in \mathcal{I}} P(A \cap B_i) = \sum_{i \in \mathcal{I}} P(A|B_i)P(B_i)$$

We also have the law of total probability with condition:

$$P(A|E) = \sum_{i \in \mathcal{I}} P(A|B_i \cap E)P(B_i|E)$$

Proof:

$$P(A|E) = \frac{P(A \cap E)}{P(E)} \overset{\text{law of total prob.}}{=} \frac{\sum_{i \in \mathcal{I}} P((A \cap E) \cap B_i)}{P(E)}$$

$$= \frac{\sum_{i \in \mathcal{I}} P(A \cap (B_i \cap E))}{P(E)}$$

$$= \frac{\sum_{i \in \mathcal{I}} P(A|B_i \cap E)P(B_i \cap E)}{P(E)}$$

$$= \frac{\sum_{i \in \mathcal{I}} P(A|B_i \cap E)P(B_i|E)P(E)}{P(E)}$$

$$= \sum_{i \in \mathcal{I}} P(A|B_i \cap E)P(B_i|E)$$

Independence: Two events $A, B \in \mathcal{F}$ are said to be independent if

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

If A and B are independent, then the same is true for A^c and B, A and B^c, A^c and B^c:

Proof: Since B can be written as disjoint union $(A \cap B) \cup (A^c \cap B)$ we have

$$P(B) \overset{\text{a(iii)}}{=} P(A \cap B) + P(A^c \cap B)$$

$$P(A^c \cap B) = P(B) - P(A \cap B) = P(B) - P(A)P(B) = (1 - P(A))P(B) = P(A^c)P(B)$$

So A^c and B are independent. Similarly A and B^c are independent.

As for A^c and B^c write A^c in disjoint union

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

$$P(A^c) \overset{\text{a(iii)}}{=} P(A^c \cap B) + P(A^c \cap B^c)$$

$$P(A^c \cap B^c) = P(A^c) - P(A^c \cap B) = P(A^c) - P(A^c)P(B) = P(A^c)(1 - P(B)) = P(A^c)P(B^c)$$

An alternate way is to note that

$$P(A^c \cap B^c) \overset{\text{De Morgan's law}}{=} P((A \cup B)^c) \overset{\text{b(i)}}{=} 1 - P(A \cup B) \overset{\text{b(iii)}}{=} 1 - (P(A) + P(B) - P(A \cap B))$$

$$= 1 - P(A) - P(B) + P(A)P(B) = (1 - P(A))(1 - P(B)) = P(A^c)P(B^c)$$

(2021 final q1(c))

4 Discrete Random Variables

Intuitively, a discrete random variable assigns a number to every element in Ω, that is, every possible outcome. Recall the coin tossing example: a possible discrete random variable is the number of heads appearing in two tosses. Thus we can define a mapping X that maps HH to 2, HT and TH to 1, and TT to 0.

After defining a random variable, it is very natural to ask the probability that X takes certain value. For instance: What is $P(X = 1)$? We can define the probability mass function $P_X(x) = P(X = x)$.

$P(X = 1)$ is just the probability of the event HT or TH happening, thus $P(X = 1) = P(\{HT, TH\}) = \frac{1}{2}$. Note that $\{HT, TH\} = X^{-1}(1)$.

Below are the formal definitions:

Discrete random variable: A discrete random variable on the probability space $(\Omega, \mathcal{F}, P)$ is defined by a mapping $X : \Omega \to \mathbb{R}$ s.t.

(i) The image of Ω under X ($\text{Im}(X) = \{X(w) : w \in \Omega\}$) is a countable subset of $\mathbb{R}$;

(ii) $X^{-1}(x) = \{w \in \Omega : X(w) = x\} \in \mathcal{F}$

Probability mass function: The probability mass function of the discrete random variable X is

defined as the function $p_X : \mathbb{R} \to [0, 1]$ given by

$$p_X(x) = P(\{w \in \Omega : X(w) = x\}) = P(X^{-1}(x))$$

and is typically written as $P(X = x)$.

Comments on the definitions: (i) requires the image to be countable, since we are considering discrete random variable. (ii) requires $X^{-1}(x) \in \mathcal{F}$, so that in the def of p.m.f. $P(X^{-1}(x))$ makes sense. What p.m.f. is actually doing is that it picks the event $X^{-1}(x)$, consisting all outcomes that are mapped to x by the discrete random variable X, and then calculate the probability that this event happening.

Properties of p.m.f:

c(i) $p_X(x) = 0$ if $x \notin \text{Im}(X)$;

c(ii) $\sum_{x \in \text{Im}(X)} p_X(x) = 1$. That is, all probabilities sum up to 1.

Proof: (i) If $x \notin \text{Im}(X)$ then clearly $X^{-1}(x) = \emptyset$. Now $p_X(x) = P(X^{-1}(x)) = P(\emptyset) = 0$.

(ii) We first note that $X^{-1}(x)$ and $X^{-1}(y)$ are disjoint if $x \neq y$. Otherwise there is a element that is mapped to both x and y by X, which is impossible. Now

$$\sum_{x \in \text{Im}(X)} p_X(x) = \sum_{x \in \text{Im}(X)} P(X^{-1}(x)) \overset{\text{disjoint}}{=} P\left(\bigcup_{x \in \text{Im}(X)} X^{-1}(x) \right)$$

$$= P\left(\bigcup_{x \in \text{Im}(X)} \{w \in \Omega : X(w) = x\} \right)$$

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

5 Continuous Random Variables

For an example of a continuous random variable, consider the event of randomly picking a real number in $[0, 1]$. Clearly, $P(X = x) = 0$ for any $x \in [0, 1]$, so discrete p.m.f. fails to give us any information about this event. However, we can instead consider $P(-\infty \leq X \leq x)$. By our common sense, this is just x if $x \in [0, 1]$, which gives us some information about this event.

Analogously, $P(-\infty \leq X \leq x) = P(X^{-1}((-\infty, x]))$, hence we want $X^{-1}((-\infty, x]) \in \mathcal{F}$ so that we can assign a probability to it using a probability measure. This inspires the definition of a random variable, which works for both discrete and continuous case:

Random variable: A random variable on the probability space $(\Omega, \mathcal{F}, P)$ is a mapping $X : \Omega \to \mathbb{R}$ satisfies

$$X^{-1}((-\infty, x]) \in \mathcal{F} \text{ for all } x \in \mathbb{R}$$

Cumulative distribution function: Suppose X is a random variable on $(\Omega, \mathcal{F}, P)$. The cumulative

distribution function of X is defined as the mapping $F_X : \Omega \to [0,1]$ given by

$$F_X(x) = P(X^{-1}((-\infty, x])) = P(\{w \in \Omega : X(w) \leq x\})$$

typically written as $P(X \leq x)$.

Returning to the previous example, we can define the sample space $\Omega = [0,1]$, the event space $\mathcal{F} = \mathcal{B}([0,1])$ (consists all sets formed by open intervals using countable union, intersection and complement), probability measure $P =$ Lebesgue measure (the 'length' of a set), and $X(w) = w$.

Properties of c.d.f:

(i) F_X is monotonic increasing;

(ii) F_X is right continuous;

(iii) $\lim_{x \to -\infty} F_X(x) = 0$, $\lim_{x \to \infty} F_X(x) = 1$.

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

Continuous random variable: A random variable X is called continuous if its c.d.f. can be written as

$$F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f_X(u)du, \text{ for all } x \in \mathbb{R}$$

where the function $f_X : \mathbb{R} \to \mathbb{R}$ satisfies

d(i) $f_X(u) \geq 0$ for all $u \in \mathbb{R}$

d(ii) $\int_{-\infty}^{\infty} f_X(u)du = 1$

Point probabilities are 0 for continuous random variable: For a continuous random variable X with density f_X, we have

$$P(X = x) = 0 \text{ for all } x \in \mathbb{R}$$

From this we can infer that

$$P(a \leq x \leq b) = P(a < x \leq b) = F_X(b) - F_X(a) = \int_a^b f_X(u)du$$

We call f_X the probability density function.

6 Transformation of Random Variables

A very common exam question is to find the p.d.f. of a random variable X after a transformation g. The general method is to:

(i) Find the cumulative distribution function $F_X(x)$;

(ii) Write $F_Y(y) = P(Y \leq y) = P(g(X) \leq y)$ to solve for the cumulative distribution function for Y;

(iii) Differentiate $F_Y(y)$ to obtain $f_Y(y)$.

This is best illustrated with an example (2024 Jan q2(a)): Let

$$f_X(x) = \begin{cases} 2x, & \text{for } x \in [0,1], \\ 0, & \text{otherwise.} \end{cases}$$

Find the p.d.f. of $Y = X^2$.

We first find $F_X(x)$. Since

$$\int_{-\infty}^{x} f_X(u)du = \int_0^x 2udu = x^2$$

$$F_X(x) = \begin{cases} 0, & \text{for } x < 0 \\ x^2, & \text{for } 0 \le x < 1, \\ 1, & \text{for } 1 \le x. \end{cases}$$

Now

$$F_Y(y) = P(Y \le y) = P(X^2 \le y) = P(-\sqrt{y} \le X \le \sqrt{y})$$

$$= P(0 \le X \le \sqrt{y})$$

since X only takes nonnegative values.

$$= F_X(\sqrt{y}) - F_X(0)$$

$$= \begin{cases} 0, & \text{for } \sqrt{y} < 0 \\ y, & \text{for } 0 \le \sqrt{y} < 1, \\ 1, & \text{for } 1 \le \sqrt{y} \end{cases}$$

$$= \begin{cases} 0, \text{ for } y < 0, \\ (\text{for } y < 0 \text{ the p.d.f. is not defined, set it to } 0) \\ y, \text{ for } 0 \le y < 1, \\ 1, \text{ for } 1 \le y \end{cases}$$

Differentiate yields

$$f_Y(y) = \begin{cases} 1, & \text{for } 0 \le y < 1 \\ 0, & \text{otherwise} \end{cases}$$

7 Expectation and Variance

Expectation gives us information about the center of the distribution. We can think of it as a weighted average of all possible values taken by the distribution.

Expectation of discrete random variable: Let X denote a discrete random variable, then the

expectation of X is defined as

$$\mathrm{E}(X) = \sum_{x \in \mathrm{Im}(X)} xP(X = x)$$

typically abbreviated as

$$\sum_x x p_X(x)$$

Expectation of continuous random variable: Let X denote a continuous random variable, then the expectation of X is defined as

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

Note: $f_X(x)dx \approx P(x < X \leq x + dx)$ so it can be view as a continuous analogy to $p_X(x)$. Thus the formula is just a continuous analogy to the discrete case.

Law of the Unconscious Statistician: Let X be a discrete random variable and $g : \mathbb{R} \to \mathbb{R}$, then

$$\mathrm{E}(g(X)) = \sum_{x \in \mathrm{Im}(X)} g(x)P(X = x)$$

Let X be a continuous random variable with density f_X, consider a function $g : \mathbb{R} \to \mathbb{R}$ s.t. $g(X)$ is also a continuous random variable, then

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

Variance: Let X be a discrete/continuous random variable, then its variance is defined as

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

Linearity of Expectation: We have

$$\mathrm{E}(aX + b) = a\,\mathrm{E}(X) + b$$

Proof: For discrete case

$$\mathrm{E}(aX + b) \overset{\text{LOTUS}}{=} \sum_x (ax + b)P(X = x)$$

$$= a \sum_x xP(X = a) + b \sum_x P(X = x)$$

$$\overset{\text{c(ii)}}{=} a\,\mathrm{E}(X) + b$$

For continuous case

$$\mathrm{E}(aX + b) \overset{\text{LOTUS}}{=} \int_{-\infty}^{\infty} (ax + b) f_X(x) dx$$

$$= a \int_{-\infty}^{\infty} x f_X(x) dx + b \int_{-\infty}^{\infty} f_X(x) dx$$

$$\overset{\text{d(ii)}}{=} a\,\mathrm{E}(X) + b$$

Note that this is also a direct consequence of linearity of expectation for joint variables.

Basic properties of Variance:

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

(ii) $\text{Var}(aX + b) = a^2 \text{Var}(x)$

Proof: (i) For convenience, let $\mu = \text{E}(X)$.

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

Using linearity of expectation:

$$= \text{E}(X^2) - 2\mu \,\text{E}(X) + \mu^2$$

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

(ii)

$$\text{Var}(aX + b) = \text{E}((aX + b)^2) - (\text{E}(aX + b))^2$$

$$= \text{E}(a^2X^2 + 2abX + b^2) - (a \,\text{E}(X) + b)^2$$

$$\overset{\text{linearity}}{=} a^2 \,\text{E}(X^2) + 2ab \,\text{E}(X) + b^2 - a^2 \,\text{E}(X)^2 - 2ab \,\text{E}(X) - b^2$$

$$= a^2 \,\text{E}(X^2) - a^2 \,\text{E}(X)^2 = a^2 \,\text{Var}(X)$$

8 Multivariate Distributions

Consider two random variables on the same probability space $(\Omega, \mathcal{F}, P)$. To study their relation we write them as a random vector (X, Y) taking values in $\mathbb{R}^2$.

Joint Distribution Function: The joint distribution function of the random vector (X, Y) is defined as the mapping $F_{X,Y} : \mathbb{R}^2 \to [0, 1]$ given by

$$F_{X,Y}(x, y) = P(\{w \in \Omega : X(w) \leq x, Y(w) \leq y\})$$

typically written as

$$P(X \leq x, Y \leq y)$$

Independence: The random variables X and Y are independent if

$$P(X \leq x, Y \leq y) = P(X \leq x)P(Y \leq y), \text{ for all } x, y \in \mathbb{R}$$

which is equivalent to

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

A useful consequence is (2020 Jan q1(e))

$$P(a < X \leq b, c < Y \leq d) = P(a < X \leq b)P(c < Y \leq d)$$

Proof: We can write $\{w \in \Omega : X(w) \leq b, Y(w) \leq d\}$ as disjoint union

$$\{w \in \Omega : a < X(w) \leq b, c < Y(w) \leq d\} \cup \{w \in \Omega : X(w) \leq a, c < Y(w) \leq d\}$$

$$\cup \{w \in \Omega : X(w) \leq b, Y(w) \leq c\}$$

By a(iii) we have

$$P(\{w \in \Omega : X(w) \leq b, Y(w) \leq d\})$$

$$= P(\{w \in \Omega : a < X(w) \leq b, c < Y(w) \leq d\}) + P(\{w \in \Omega : X(w) \leq a, c < Y(w) \leq d\})$$

$$+P(\{w \in \Omega : X(w) \leq b, Y(w) \leq c\})$$

Adding both sides by $P(\{w \in \Omega : X(w) \leq a, Y(w) \leq c\})$ we obtain

$$P(\{w \in \Omega : X(w) \leq b, Y(w) \leq d\}) + P(\{w \in \Omega : X(w) \leq a, Y(w) \leq c\})$$

$$= P(\{w \in \Omega : a < X(w) \leq b, c < Y(w) \leq d\})$$

$$+P(\{w \in \Omega : X(w) \leq a, c < Y(w) \leq d\}) + P(\{w \in \Omega : X(w) \leq a, Y(w) \leq c\})$$

$$+P(\{w \in \Omega : X(w) \leq b, Y(w) \leq c\})$$

$$= P(\{w \in \Omega : a < X(w) \leq b, c < Y(w) \leq d\})$$

$$+P(\{w \in \Omega : X(w) \leq a, Y(w) \leq d\})$$

$$+P(\{w \in \Omega : X(w) \leq b, Y(w) \leq c\})$$

Thus

$$F_{X,Y}(b,d) + F_{X,Y}(a,c) = P(a < X \leq b, c < Y \leq d) + F_{X,Y}(a,d) + F_{X,Y}(b,c)$$

Now

$$P(a < X \leq b, c < Y \leq d) = F_{X,Y}(b,d) + F_{X,Y}(a,c) - F_{X,Y}(a,d) - F_{X,Y}(b,c)$$

By independence

$$= F_X(b)F_Y(d) + F_X(a)F_Y(c) - F_X(a)F_Y(d) - F_X(b)F_Y(c)$$

$$= (F_X(b) - F_X(a))(F_Y((d)) - F_Y(c))$$

$$= P(a < X \leq b)P(c < Y \leq d)$$

Joint distributions can also be discrete or continuous. We define them as follows:

Joint probability mass function: Let X, Y denote discrete random variables $(\Omega, \mathcal{F}, P)$. Their joint probability mass function is the function $p_{X,Y} : \mathbb{R}^2 \to [0,1]$ given by

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

typically written as $P(X = x, Y = y)$

We have that $p_{X,Y}(x,y) \geq 0$ and $\sum_x \sum_y p_{X,Y}(x,y) = 1$.

The marginal density of X and Y are given by

$$p_X(x) = \sum_y p_{X,Y}(x,y), \qquad p_Y(y) = \sum_x p_{X,Y}(x,y)$$

Note: To check that they are actually the p.m.f. of X and Y we can use the law of total probability:

$$P(X = x) = P(\{w \in \Omega : X(w) = x\}) = \sum_{y \in \mathrm{Im}(Y)} P(\{w \in \Omega : X(w) = x\} \cup \{Y(w) = y\})$$

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

Hence we can also write

$$P(X = x) = \sum_y P(X = x, Y = y), \qquad P(Y = y) = \sum_x P(X = x, Y = y)$$

Independence of discrete random variables: Let X, Y denote discrete random variables $(\Omega, \mathcal{F}, P)$. X and Y are said to be independent if

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

Note: this can actually be derived from the def of independence for general random variables. We need the following result

$$P(X = x) = F_X(x) - F_X(x^-), P(Y = y) = F_Y(y) - F_Y(y^-)$$

and

$$P(X = x, Y = y) = F_{X,Y}(x,y) - F_{X,Y}(x^-, y) - F_{X,Y}(x, y^-) + F_{X,Y}(x^-, y^-)$$

by independence

$$P(X = x, Y = y) = F_X(x)F_Y(y) - F_X(x^-)F_Y(y) - F_X(x)F_Y(y^-) + F_X(x^-)F_Y(y^-)$$

$$= (F_X(x) - F_X(x^-))(F_Y(y) - F_Y(y^-))$$

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

Continuous random vector: We call the random vector (X, Y) on $(\Omega, \mathcal{F}, P)$ continuous if

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

for a function $f_{X,Y} : \mathbb{R}^2 \to \mathbb{R}^2$ satisfying

(i) $f_{X,Y}(u,v) \geq 0$ for all $u, v \in \mathbb{R}$;

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

To find the marginal densities note that

$$F_X(x) = P(X \leq x, Y \leq \infty) = \lim_{y\to\infty} F_{X,Y}(x,y)$$

$$= \int_{u=-\infty}^{x} \int_{v=-\infty}^{\infty} f_{X,Y}(u,v)dvdu$$

Hence

$$f_X(x) = \frac{d}{dx}F_X(x) = \frac{d}{dx}\int_{u=-\infty}^{x} \int_{v=-\infty}^{\infty} f_{X,Y}(u,v)dvdu$$

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

Similarly, $f_Y(y) = \int_{u=-\infty}^{\infty} f_{X,Y}(u,y)du$. Intuitively, by summing/integrating over a variable we eliminate the effect of that variable so we obtain the pmf/pdf of the other variable.

Independence of continuous random variables: If X, Y are independent then $F_{X,Y}(x,y) = F_X(x)F_Y(y)$. Differentiating yields

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

A useful result about independence states that if X, Y are independent, then $f(X), g(Y)$ are also independent.

2D LOTUS: Let X, Y denote discrete random variables $(\Omega, \mathcal{F}, P)$ and let $g : \mathbb{R}^2 \to \mathbb{R}$. Then $g(X,Y)$ is also a discrete random variable on $(\Omega, \mathcal{F}, P)$ and its expectation is given by

$$\mathrm{E}(g(X,Y)) = \sum_{x\in \mathrm{Im}(X)} \sum_{y\in \mathrm{Im}(Y)} g(x,y)P(X=x, Y=y)$$

Let X, Y be jointly continuous random variables on $(\Omega, \mathcal{F}, P)$ with density $f_{X,Y}$ and let $h : \mathbb{R}^2 \to \mathbb{R}$. Then

$$\mathrm{E}(h(X,Y)) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(x,y)f_{X,Y}(x,y)dxdy$$

Finding joint probabilities: For set $A \subseteq \mathbb{R}^2$ we have that

$$P((X,Y) \in A) = \sum \sum_{(x,y)\in A} P(X=x, Y=y)$$

and for continuous case

$$P((X,Y) \in A) = \int \int_{(x,y)\in A} f_{X,Y}(x,y)dxdy$$

A frequently asked question will be like 2024 Jan q2(d): Find

$$P(\min\{X_1, X_2, ..., X_n\} < x)$$

if $X_1, ..., X_n$ are independent.

The key is to find the complement

$$P(\min\{X_1, X_2, ..., X_n\} \geq x) = P(X_1 \geq x, X_2 \geq x..., X_n \geq x)$$

$$= P(X_1 \geq x)P(X_2 \geq x)...P(X_n \geq x) \text{ (by independence)}$$

We can apply a similar method to find

$$P(\max\{X_1, X_2, ..., X_n\} > x)$$

Linearity of expectation: Let X, Y denote jointly discrete/continuous random variables on $(\Omega, \mathcal{F}, P)$, then

$$\mathrm{E}(aX + bY) = a\,\mathrm{E}(X) + b\,\mathrm{E}(Y)$$

Proof: For discrete case

$$\mathrm{E}(aX + bY) \overset{\text{LOTUS}}{=} \sum_x \sum_y (ax + by)P(X = x, Y = y)$$

$$= a\sum_x x \sum_y P(X = x, Y = y) + b\sum_y y \sum_x P(X = x, Y = y)$$

$$= a\sum_x xP(X = x) + b\sum_y yP(Y = y)$$

$$= a\,\mathrm{E}(X) + b\,\mathrm{E}(Y)$$

For continuous case

$$\mathrm{E}(aX + bY) \overset{\text{LOTUS}}{=} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (ax + by)f_{X,Y}(x, y)dxdy$$

$$= a\int_{-\infty}^{\infty} x \int_{-\infty}^{\infty} f_{X,Y}(x, y)dydx + b\int_{-\infty}^{\infty} y \int_{-\infty}^{\infty} f_{X,Y}(x, y)dxdy$$

$$= a\int_{-\infty}^{\infty} xf_X(x)dx + b\int_{-\infty}^{\infty} yf_Y(y)dy$$

$$= a\,\mathrm{E}(X) + b\,\mathrm{E}(Y)$$

Covariance: The covariance of two random variables is defined as

$$\mathrm{Cov}(X, Y) = \mathrm{E}((X - \mu_X)(Y - \mu_Y))$$

We have that

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

Proof:

$$\mathrm{Cov}(X, Y) = \mathrm{E}((X - \mu_X)(Y - \mu_Y)) = \mathrm{E}(XY - \mu_X Y - \mu_Y X + \mu_X\mu_Y)$$

$$= \mathrm{E}(XY) - \mu_X\,\mathrm{E}(Y) - \mu_Y\,\mathrm{E}(X) + \mu_X\mu_Y$$

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

Expectation of product: Let X, Y denote independent and jointly discrete/continuous random

variables. Then

$$\mathrm{E}(XY) = \mathrm{E}(X)\,\mathrm{E}(Y)$$

Proof: For discrete case

$$\mathrm{E}(XY) \overset{\text{LOTUS}}{=} \sum_x \sum_y xyP(X=x, Y=y)$$

$$\overset{\text{independence}}{=} \sum_x \sum_y xyP(X=x)P(Y=y)$$

$$= \sum_x xP(X=x) \sum_y yP(Y=y)$$

$$= \mathrm{E}(X)\,\mathrm{E}(Y)$$

The proof for continuous case is similar.

Hence, $\mathrm{Cov}(X,Y) = 0$ if X and Y are independent. However the converse is not true.

'Linearity' of variance: Let X, Y denote and jointly discrete/continuous random variables. Then

$$\mathrm{Var}(aX+bY) = a^2\,\mathrm{Var}(X) + b^2\,\mathrm{Var}(Y) + 2ab\,\mathrm{Cov}(X,Y)$$

Proof:

$$\mathrm{Var}(aX+bY) = \mathrm{E}((aX+bY-a\mu_X-b\mu_Y)^2)$$

$$= \mathrm{E}((aX-a\mu_X)^2 + (bY-b\mu_Y)^2 + 2ab(X-\mu_X)(Y-\mu_Y))$$

$$= a^2\,\mathrm{E}((X-\mu_X)^2) + b^2\,\mathrm{E}(Y-\mu_Y)^2) + 2ab\,\mathrm{E}((X-\mu_X)(Y-\mu_Y))$$

$$= a^2\,\mathrm{E}(X) + b^2\,\mathrm{E}(Y) + 2ab\,\mathrm{Cov}(X,Y)$$

9 Conditional Distributions

Suppose we have a joint distribution of X and Y, and say we know something about X, e.g. some event B occurs. What do we now know about Y? This is answered by the conditional distribution.

Contditional distribution and expectation, discrete case: Let Y denote a discrete random variable on the probability space $(\Omega, \mathcal{F}, P)$. Consider an event $B \in \mathcal{F}$ such that $P(B) > 0$. The conditional distribution of Y given B is defined as

$$P(Y=y|B) = \frac{P(\{Y=y\} \cap B)}{P(B)}, \text{ for } y \in \mathbb{R}$$

The conditional expectation of Y given B is defined as

$$\mathrm{E}(Y|B) = \sum_{y \in \mathrm{Im}(Y)} yP(Y=y|B)$$

Law of total expectation, discrete case: Consider a partition $\{B_i : i \in \mathcal{I}\}$ of Ω with $P(B_i) > 0$.

Let Y denote a discrete random variable, then

$$\mathrm{E}(Y) = \sum_{i \in \mathcal{I}} \mathrm{E}(Y|B_i)P(B_i)$$

Proof:

$$\mathrm{E}(Y) = \sum_{y \in \mathrm{Im}(Y)} yP(Y=y) \overset{\text{law of total prob.}}{=} \sum_{y \in \mathrm{Im}(Y)} y \sum_{i \in \mathcal{I}} P(Y=y|B_i)P(B_i)$$

$$= \sum_{i \in \mathcal{I}} P(B_i) \sum_{y \in \mathrm{Im}(Y)} yP(Y=y|B_i)$$

$$= \sum_{i \in \mathcal{I}} P(B_i)E(Y|B_i)$$

Conditioning by $X = x$: Suppose we take the event $B = \{X = x\}$ as the condition for Y:

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

Thus we define the conditional probability mass function as

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

Note that, if X and Y are independent then

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

$$P(Y=y|X=x) = P(Y=y)$$

This makes sense because if X and Y are independent, then knowing $X = x$ will not give us any information about Y.

Now the conditional expectation becomes

$$\mathrm{E}(Y|X=x) = \sum_{y \in \mathrm{Im}(Y)} yP(Y=y|X=x) = \sum_{y \in \mathrm{Im}(Y)} yp_{Y|X}(y|x)$$

Also, the LOTUS for conditional expectation says that

$$\mathrm{E}(g(Y)|X=x) = \sum_{y \in \mathrm{Im}(Y)} g(y)p_{Y|X}(y|x)$$

Conditional distribution and conditional density: For two jointly continuous random variable X and Y, we define the conditional density of Y given $X = x$ as

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

Similarly, if X and Y are independent then

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

A nice way of interpreting the conditional distribution is the following:

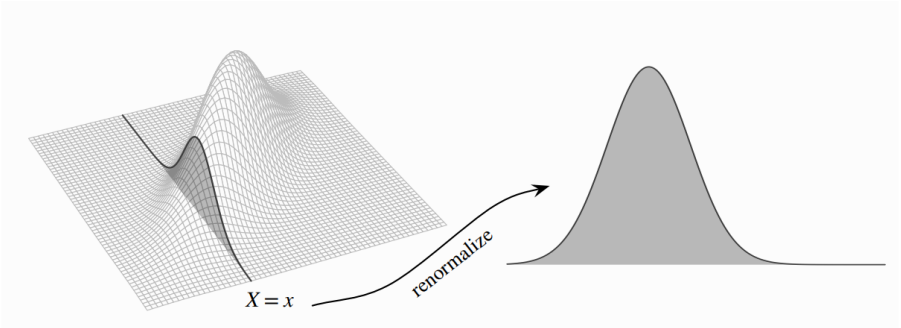

Figure 1: Conditional PDF of Y given $X = x$.

We take a vertical slice of the joint PDF corresponding to the observed value of X; since the total area under this slice is $f_X(x)$, we then divide by $f_X(x)$ to ensure that the conditional PDF will have an area of 1. Therefore conditional PDFs satisfy the properties of a valid PDF.

The corresponding conditional distribution function of Y given $X = x$ is

$$F_{Y|X}(y|x) = \frac{\int_{v=-\infty}^{y} f_{X,Y}(x,v)dv}{f_X(x)}$$

for all $x \in \mathbb{R}$ for which $f_X(x) > 0$.

(2021 final q2(c)): Show that

$$P(Y \leq y) = \int_{x:f_X(x)>0} F_{Y|X}(y|x) f_X(x) dx$$

Proof:

$$\begin{aligned}
\int_{x:f_X(x)>0} F_{Y|X}(y|x) f_X(x) dx &= \int_{x:f_X(x)>0} \frac{\int_{v=-\infty}^{y} f_{X,Y}(x,v)dv}{f_X(x)} f_X(x) dx \\
&= \int_{x:f_X(x)>0} \int_{v=-\infty}^{y} f_{X,Y}(x,v) dv dx \\
&= \int_{v=-\infty}^{y} \int_{x:f_X(x)>0} f_{X,Y}(x,v) dv dx \\
&= \int_{v=-\infty}^{y} f_Y(v) dv \\
&= P(Y \leq y)
\end{aligned}$$

Conditional expectation, continuous case: For two jointly continuous random variables X, Y, we define the conditional expectation of Y given $X = x$ as

$$\mathrm{E}(Y|X = x) = \int_{-\infty}^{\infty} y f_{Y|X}(y|x) dy$$

Law of total expectation, continuous case: For jointly continuous random variables X, Y we have

$$\mathrm{E}(Y) = \int_{x:f_X(x)>0} \mathrm{E}(Y|X=x) f_X(x) dx$$

Proof:

$$\begin{aligned}
\int_{x:f_X(x)>0} \mathrm{E}(Y|X=x) f_X(x) dx &= \int_{x:f_X(x)>0} \int_{y=-\infty}^{\infty} y f_{Y|X}(y|x) dy \cdot f_X(x) dx \\
&= \int_{x:f_X(x)>0} \int_{y=-\infty}^{\infty} y f_{X,Y}(x,y) dy dx \\
&= \int_{y=-\infty}^{\infty} y \int_{x:f_X(x)>0} f_{X,Y}(x,y) dx dy \\
&= \int_{y=-\infty}^{\infty} y f_Y(y) dy \\
&= \mathrm{E}(Y)
\end{aligned}$$

10 Generating Functions

Probability generating function: For a discrete random variable X with $\mathrm{Im}(X) \subseteq \mathbb{N} \cup \{0\}$, the probability generating function $G_X(s)$ is given by

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

Clearly

$$G_X(0) = P(X=0) \text{ and } G_X(1) = 1$$

(Note that $0^0 = 1$ for consistency.)

Probability generating functions uniquely determine a random variable. Specifically

$$\frac{d^n}{ds^n} G_X(s)\bigg|_{s=0} = n! P(X=n)$$

Perhaps the most useful property of p.g.f. is

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

when X, Y are independent. This allows us to find the distribution of $X+Y$. A similar result is

$$G_{aX+b}(s) = s^b G_X(s^a)$$

We can use p.g.f. to find moments of X using the formula

$$\frac{d^n}{ds^n} G_X(s)\bigg|_{s=1} = \mathrm{E}(X(X-1)...(X-n+1))$$

Hence

$$\frac{d}{ds}G_X(s)\Big|_{s=1} = \mathrm{E}(X)$$

and

$$\frac{d^2}{ds^2}G_X(s)\Big|_{s=1} = \mathrm{E}(X(X-1))$$

Thus $\mathrm{E}(X^2) = G''_X(1) + G'_X(1)$ and we also obtained a formula for variance

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

Moment generating functions: Let X be a random variable. Its moment generating function is defined as

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

Relation between p.g.f. and m.g.f. We have that

$$M_X(t) = G_X(e^t)$$

Properties of m.g.f.:

$$\frac{d^n}{dt^n}M_X(t)\Big|_{s=0} = \mathrm{E}(X^n)$$

$$M_{aX+b}(t) = e^{bt}M_X(at)$$

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

when X, Y are independent.