Preliminaries for Statistics
Expectation, Independence & Conditioning

Zhu Xuelin

This note serves as a summary of foundational concepts in statistical theory. We begin with expectation, examining its construction and convergence. Next, we explore distributions, focusing on computing expectations under two types of distributions. We then discuss independence in the view of product measures, leveraging Fubini's theorem for deeper insights. Finally, we conclude with the art of conditioning—your key to unlocking the depths of probability theory.

Acknowledgement I would like to express my sincerest gratitude to all the professors who have introduced me to the probability and statistics at various levels. Their guidance has been invaluable throughout my learning journey.

- Professor Chen Anyue and Donghan Kim for elementary probability,
- Professor Kean Ming Tan and Tian Guoliang for elementary statistics,
- Professor Li Zhan and Jing Bingyi for real analysis and measure theory,
- Professor Tailen Hsing, Mark Rudelson, and Somabha Mukherjee for modern probability,
- Professor Moulinath Banerjee and Huang Dongming for modern statistics.

Contents

1 Expectation

In this section, our goal is to redefine the concept *expectation* like $\mathbb{E}X$ in a unifying way, regardless of its discrete or continuous distribution, and discuss when the expectation converges as $\mathbb{E}X_n \to \mathbb{E}X$.

1.1 Re-construction

For a complete and detailed construction, refer to classics like *Probability Theory and Examples* by Durrett[1], or *Real Analysis* by Folland[2] and Stein[3]. Here, we'll proceed as:

$$\text{simple random variables} \;\Rightarrow\; \text{Non-negative rvs} \;\Rightarrow\; \text{Measurable rvs.}$$

Simple random variables

Let's assume all random variables (rvs) discussed below are measurable from $(\Omega, \mathcal{F}, \mathbb{P})$ to $(\mathbb{R}, \mathcal{B}_{\mathbb{R}})$.

Definition 1 (Simple random variable)**.** X is said to be a simple rv if $X(\omega) = \sum_1^n a_i \mathbb{1}_{A_i}(\omega)$, where $a_i \in \mathbb{R}$ and $\mathbb{P}(A_i) < \infty$. And we define the integral of X to be

$$\mathbb{E}X := \sum_{i=1}^{n} a_i \mathbb{P}(A_i).$$

In the language of measure theory, $\mathbb{E}X$ is also written as $\int X d\mathbb{P}$.

A simple random variable can have multiple representations, such as $X = \sum_1^n a_i \mathbb{1}_{A_i} = \sum_1^m b_j \mathbb{1}_{B_j}$. However, if we further require that $a_i \neq a_j$ for $i \neq j$ and $\bigsqcup_{i=1}^n A_i = \Omega$, there exists a unique representation, which we call the *canonical form*. It can be shown that the defined integral does not depend on the choice of representation (for details, refer to Stein's book[3]).

We will check 6 properties of integral each time after definition. Here are the first three.

Lemma 2. *Let X and Y be simple rvs.*

(i) If $X \geq 0$ almost surely then $\mathbb{E}X \geq 0$.

(ii) For any $a \in \mathbb{R}$, $\mathbb{E}(aX) = a \cdot \mathbb{E}X$.

(iii) $\mathbb{E}(X + Y) = \mathbb{E}X + \mathbb{E}Y$.

Proof. (i) and (ii) are from definition. To prove (iii), suppose

$$X = \sum_{i=1}^{m} a_i \mathbb{1}_{A_i} \quad \text{and} \quad Y = \sum_{j=1}^{n} b_j \mathbb{1}_{B_j}.$$

To make the supports of the two functions the same, we let $A_0 = \cup_1^n B_i - \cup_1^m A_i$, and $B_0 = \cup_1^m A_i - \cup_1^n B_i$ and $a_0 = b_0 = 0$. Draw a graph to see what is doing here. Now

$$X + Y = \sum_{i=0}^{m} \sum_{j=0}^{n} (a_i + b_j) \, \mathbb{1}_{A_i \cap B_j},$$

where $A_i \cap B_j$ are pairwise disjoint, so

$$\mathbb{E}(X+Y) = \sum_{i=0}^{m}\sum_{j=0}^{n}(a_i+b_j)\,\mathbb{P}(A_i\cap B_j) = \sum_{i=0}^{m}\sum_{j=0}^{n} a_i\mathbb{P}(A_i\cap B_j) + \sum_{j=0}^{n}\sum_{i=0}^{m} b_j\mathbb{P}(A_i\cap B_j)$$
$$= \sum_{i=0}^{m} a_i\mathbb{P}(A_i) + \sum_{j=0}^{n} b_j\mathbb{P}(B_j) = \mathbb{E}X + \mathbb{E}Y.$$

This completes the first three properties. □

Another three properties (iv)-(vi) are also important. They can be checked once we have (i)-(iii), so we only need to prove them once here.

Lemma 3. *If (i) and (iii) hold, then we have*

(iv) If $X \le Y$ almost surely then $\mathbb{E}X \le \mathbb{E}Y$.

(v) If $X = Y$ almost surely then $\mathbb{E}X = \mathbb{E}Y$.

In addition, if (ii) holds when $a = -1$, we have

(vi) $|\mathbb{E}X| \le \mathbb{E}|X|$.

Proof. For (iv), noting that $(Y - X) \ge 0$ is simple, by (i) we have $\mathbb{E}(Y - X) \ge 0$. Further with $Y = X + (Y - X)$, we have $\mathbb{E}Y = \mathbb{E}X + \mathbb{E}(Y - X)$ by (iii), concluding (iv). And (v) follows from two applications of (iv) in $X \le Y$ and $Y \le X$.

To prove (vi), note that $X \le |X|$, so (iv) implies $\mathbb{E}X \le \mathbb{E}|X|$. Also, with (ii), $-X \le |X|$ implies $-\mathbb{E}X \le \mathbb{E}|X|$, concluding (vi). □

Non-negative random variables

The integral of non-negative rvs depends on the next theorem of simple rvs approximation.

Lemma 4. *Let $X \ge 0$ be a rv. There exists a sequence of simple rvs $(X_n)_{n=1}^{\infty}$ such that $X_n \uparrow X$ pointwisely.*

Proof. Take $X_n(\omega) = \sum_{k=1}^{n2^n} \frac{k-1}{2^n}\mathbb{1}\left\{\frac{k-1}{2^n} \le X(\omega) < \frac{k}{2^n}\right\} + n\mathbb{1}\{X > n\}$. □

Now we define the integral of non-negative rv to be the limit of simple rv integrals.

Definition 5. Let $X \ge 0$ be a measurable function. Define the integral of X to be

$$\mathbb{E}X := \lim_{n\to\infty} \mathbb{E}X_n,$$

where (X_n) is a sequence of simple rvs such that $X_n \uparrow X$.

We should be careful for few things to make this definition well-defined:

(i) Whether such sequence exists?

(ii) If it exists, does the limit of RHS exist?

(iii) And if there exists multiple sequences, does this definition give the same value?

The first question is answered by Lemma 4. We now prove the second and third ones.

Theorem 6. *Let $X_n \geq 0$ and $Y_n \geq 0$ be simple rvs with $X_n \uparrow X$ and $Y_n \uparrow X$. Then*

(i) both $\lim_{n\to\infty} \mathbb{E}X_n$ and $\lim_{n\to\infty} \mathbb{E}Y_n$ exist, and

(ii) their limits $\lim_{n\to\infty} \mathbb{E}X_n = \lim_{n\to\infty} \mathbb{E}Y_n$.

Proof. Since X_n are increasing simple rvs, we have $\mathbb{E}X_n$ increasing as a sequence of real numbers. So, the limit exists.

To show (ii), we first fix a proportion $0 < t < 1$ and fix an integer $m \geq 1$, and define for every $n \in \mathbb{N}$:

$$A_n := \{\omega \in \Omega : X_n(\omega) \geq t \cdot Y_m(\omega)\}.$$

Since t, m are fixed, we can check $A_n \uparrow \Omega$. Then by the properties of simple rvs integral:

$$\mathbb{E}X_n \geq \mathbb{E}(X_n \mathbb{1}_{A_n}) \geq \mathbb{E}(tY_m \mathbb{1}_{A_n}) = t \cdot \mathbb{E}(Y_m \mathbb{1}_{A_n}), \quad \forall n \in \mathbb{N}.$$

Taking limits on both side (we can do it since both sides are increasing sequences of numbers), we get

$$\lim_{n\to\infty} \mathbb{E}X_n \geq t \cdot \lim_{n\to\infty} \mathbb{E}(Y_m \mathbb{1}_{A_n}) = t \cdot \mathbb{E}\Big[\lim_{n\to\infty} (Y_m \mathbb{1}_{A_n})\Big] = t \cdot \mathbb{E}Y_m, \tag{1}$$

where the second equality is a claim(to be proved in the end). Since it holds for all $0 < t < 1$ and $m \in \mathbb{N}$, taking $t \to 1$ and $m \to \infty$, we get $\lim_{n\to\infty} \mathbb{E}X_n \geq \lim_{m\to\infty} \mathbb{E}Y_m$. Switching the role of X and Y, we get our goal.

It now remains to prove the claim (1) (we need to do this without using any convergence theorem since we have not proved them). Note that Y_m, by theorem condition, are simple rvs. WLOG, suppose $Y_m = \sum_1^k b_i \mathbb{1}_{B_i}$, with $b_i \in \mathbb{R}$ and $\mathbb{P}(B_i) < \infty$. Then $Y_m \mathbb{1}_{A_n} = \sum_1^k b_i \mathbb{1}_{B_i \cap A_n}$ is also simple, and

$$\mathbb{E}(Y_m \mathbb{1}_{A_n}) = \sum_{i=1}^k b_i \mathbb{P}(B_i \cap A_n).$$

With $A_n \uparrow \Omega$ and the continuity of measure, letting $n \to \infty$, we get

$$\lim_{n\to\infty} \mathbb{E}(Y_m \mathbb{1}_{A_n}) = \lim_{n\to\infty} \sum_{i=1}^k b_i \mathbb{P}(B_i \cap A_n) = \sum_{i=1}^k b_i \lim_{n\to\infty} \mathbb{P}(B_i \cap A_n) = \sum_{i=1}^k b_i \mathbb{P}(B_i) = \mathbb{E}Y_m.$$

This completes the proof. □

Remark. What inspires us to introduce such a number t in the proof? And what role does it play?

We have verified this definition is well defined. It is time to show the properties, and it suffices to only show (i)-(iii).

Lemma 7. *Let X and Y be non-negative rvs.*

(i) If $X \geq 0$ almost surely then $\mathbb{E}X \geq 0$.

(ii) For any $a \in \mathbb{R}$, $\mathbb{E}(aX) = a \cdot \mathbb{E}X$.

(iii) $\mathbb{E}(X + Y) = \mathbb{E}X + \mathbb{E}Y$.

Proof. For (i), since $X \geq 0$ a.s, we can choose simple rvs $X_n \geq 0$ s.t. $X_n \uparrow X$ a.s. Then by the properties of

simple rvs integral $\mathbb{E}X_n \geq 0$ for all $n \in \mathbb{N}$, and $\mathbb{E}X = \lim_{n\to\infty} \mathbb{E}X_n \geq 0$. (ii) can be shown using the same idea. For (iii), suppose simple rvs $X_n \uparrow X$ and $Y_n \uparrow Y$. Check that $\{X_n + Y_n\}$ is also simple and increasing to $\uparrow (X + Y)$. Therefore, by definitions and using the properties of simple rvs integrals,

$$\mathbb{E}(X+Y) = \lim_{n\to\infty} \mathbb{E}(X_n + Y_n) = \lim_{n\to\infty} (\mathbb{E}X_n + \mathbb{E}Y_n) = \lim_{n\to\infty} \mathbb{E}X_n + \lim_{n\to\infty} \mathbb{E}Y_n = \mathbb{E}X + \mathbb{E}Y.$$

Hence (iv)-(vi) follows. □

Note that in property (ii), we prove for $a > 0$ instead of $a \in \mathbb{R}$ here. The reason is in $a < 0$ case, we will have negative function, for which we have not defined an integral yet.

Before we move on, we stop here to prove the equivalence between the approximation limit definition and the supremum definition. Let $\mathbb{S}^+$ be the set of all non-negative simple rvs.

Proposition 8. *Let $X \geq 0$ be a measurable function. Then*

$$\mathbb{E}X = \sup\left\{\mathbb{E}Y : Y \in \mathbb{S}^+ \text{ and } Y \leq X\right\}.$$

Remark. Since they are equivalent, we shall use either one when it is more convenient. The limit definition is more convenient in proving linearity properties, however it needs prove to be well defined as in Theorem 6. The supremum definition is directly well defined by the good properties of supremum, however it is challenging to prove the linearity properties of the integral. It's kind of a trade-off.

Proof. We first show LHS $\leq$ RHS. Since LHS does not depend on the choice of approximation, suppose simple rvs $X_n \uparrow X$. Since each X_n is an element of the set on RHS, we have (sup is an upper bound)

$$\mathbb{E}X_n \leq \sup\left\{\mathbb{E}Y : Y \in \mathbb{S}^+ \text{ and } Y \leq X\right\}, \quad \forall n \in \mathbb{N}.$$

Letting $n \to \infty$, we get $\mathbb{E}X \leq \sup\{\mathbb{E}Y : Y \in \mathbb{S}^+ \text{ and } Y \leq X\}$

For the other direction, we need to show $\mathbb{E}X \geq \mathbb{E}Y$ for all $Y \in \mathbb{S}^+$ and $Y \leq X$. So we fix a simple rv Y with $0 \leq Y \leq X$, and the goal is to show $\mathbb{E}X \geq \mathbb{E}Y$. Note that now $(Y - X)$ is a non-negative simple rv, so there exists a sequence of simple rvs $0 \leq Z_n \uparrow (X - Y)$ with $\mathbb{E}(X - Y) = \lim_{n\to\infty} \mathbb{E}Z_n$. By properties of integral, we can add $\mathbb{E}Y$ to both sides and get

$$\mathbb{E}X = \mathbb{E}(X - Y) + \mathbb{E}Y = \lim_{n\to\infty} \mathbb{E}Z_n + \mathbb{E}Y \geq \mathbb{E}Y,$$

where the last inequality comes from the fact $Z_n \geq 0$. □

Any random variables

For a general rv X, we can define its positive part X^+ and negative part $X^- : \Omega \mapsto \mathbb{R}$ by

$$X^+\omega = X\omega \wedge 0 \text{ and } X^-\omega = -(X\omega \vee 0).$$

Note that $|X| = X^+ + X^-$ and $X = X^+ - X^-$.

Definition 9. Let X be a rv. We say X is integrable if $\mathbb{E}X^+ < \infty$ and $\mathbb{E}X^- < \infty$, and define its expectation by $\mathbb{E}X = \mathbb{E}X^+ - \mathbb{E}X^-$.

We sometimes also write $X \in L_1(\Omega, \mathcal{A}, \mathbb{P})$ to indicate $X \in \mathcal{A}$ is an integrable rv on Ω with respect to the measure $\mathbb{P}$. And it can be checked that $f \in L_1(\mu)$ iff $\int |f| d\mu < \infty$. And all the properties follow by the definition and simple algebra (property (ii) may need a little work, see Lemma 1.4.6 by Durrett). We will skip them here.

Remark. In fact, the construction is not limited to a measurable function X on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. It works for any real-valued function f on any measure space $(\Omega, \mathcal{F}, \mu)$. And we write $\mathbb{E}f$ or $\mathbb{E}X$ or $\int f$ when there is no discrepancy.

In particular, we now take a detour to $(\mathbb{R}, \mathcal{B}_{\mathbb{R}}, \lambda)$, where λ is the Lebesgue measure, and end this integration construction with the relation between Lebesgue and Riemann integral. The proof can be found on Page 57 of *Real Analysis* by Folland or Stein.

Theorem 10. *Let f be a bounded real-valued function on $[a, b]$.*

(i) If f is Riemann integrable, then f is measurable, integrable and

$$\int_a^b f(x)dx = \int f \mathbb{1}_{[a,b]} d\lambda.$$

(ii) f is Riemann integrable iff $\lambda\{x \in [a, b] : f \text{ is discontinuous at } x\} = 0$

However, this theorem does not apply to improper integral, failing in the following example.

Example 1. Consider the function $f(x) = \sin x/x$ on $[0, \infty)$. The Lebesgue integral is not defined for it since

$$\int_{(n-1)\pi}^{n\pi} \frac{|\sin x|}{x} dx \geq \frac{1}{n\pi} \int_{(n-1)\pi}^{n\pi} |\sin x| dx = \frac{2}{n\pi},$$

and then for any $N \in \mathbb{N}$,

$$\int_0^{N\pi} \frac{|\sin x|}{x} dx \geq \frac{2}{\pi} \sum_{n=1}^{N} \frac{1}{n}.$$

Define $g_N = |f| \mathbb{1}_{[0,N]}$. Since $0 \leq g_N \to f$, by MCT,

$$\int |f| d\lambda = \lim_{N\to\infty} \int g_N(x)dx \geq \lim_{N\to\infty} \frac{2}{\pi} \sum_{n=1}^{N} \frac{1}{n} = \infty.$$

Therefore, the f is not Lebesgue integrable.

But using the basic calculus way, for any fixed $n \in \mathbb{N}$, we can do integration by parts and get

$$\int_0^n \frac{\sin x}{x} dx = \int_0^n \frac{1 - \cos x}{x^2} dx - \frac{1 - \cos n}{n}.$$

Then the improper Riemann integral of f is defined as

$$\int_0^\infty f(x)dx = \lim_{n\to\infty} \int_0^n f(x)dx = \lim_{n\to\infty} \left(\int_0^n \frac{1 - \cos x}{x^2} dx - \frac{1 - \cos n}{n} \right) = \frac{\pi}{2}.$$

It does not say that there is a Riemann but not Lebesgue integrable function. This example exists only by the definition of improper integral.

1.2 Integral Convergence Theorems

If $X_n \to X$ in some sense, we may guess $\mathbb{E}X_n \to \mathbb{E}X$ as a sequence of real numbers. However, this is not necessarily true. Three theorems allowing us to switch the order of a limit and integration. We start from the first one, the monotone convergence theorem.

Theorem 11 (MCT). *If $X_n \geq 0$ and $X_n \uparrow X$, then $\mathbb{E}X_n \uparrow \mathbb{E}X$.*

Proof. We will show two directions:

$$\lim_{n\to\infty} \mathbb{E}X_n \leq \mathbb{E}X \quad \text{and} \quad \lim_{n\to\infty} \mathbb{E}X_n \geq \mathbb{E}X.$$

The first one follows from $\mathbb{E}X_n \leq \mathbb{E}Xx$ for every $n \in \mathbb{N}$ by the properties of integral.

The second part is the same as proving the uniqueness of integral definition. If we show $\lim_{n\to\infty} \mathbb{E}X_n \geq \mathbb{E}Y$ for every $Y \in \mathbb{S}^+$ such that $Y \leq X$, then by Proposition 8, we are done. Now fix a Y satisfying that condition, and fix a proportion $0 < t < 1$. For each $n \in \mathbb{N}$, we define $A_n := \{X_n \geq tY\}$. And we can check $A_n \uparrow \Omega$ since $t < 1$. It follows that $\mathbb{E}X_n \geq \mathbb{E}(X_n \mathbb{1}_{A_n}) \geq t \cdot \mathbb{E}(Y\mathbb{1}_{A_n})$ for all $n \in \mathbb{N}$. Taking limits on both sides, and use claim (1) proved before, we get

$$\lim_{n\to\infty} \mathbb{E}X_n \geq t \cdot \lim_{n\to\infty} \mathbb{E}(Y\mathbb{1}_{A_n}) = t \cdot \mathbb{E}\left(\lim_{n\to\infty} Y\mathbb{1}_{A_n}\right) = t \cdot \mathbb{E}Y.$$

Letting $t \to 1$, we conclude MCT. □

Remark. The whole MCT works because a baby version of it, claim (1), holds. It does not rely on anything but just the definition.

If we do not have increasing property, the pointwise limit for an arbitrary sequence (X_n) may not exists. However, the $\limsup X_n\omega$ and $\liminf X_n\omega$ always exist for all $\omega \in \Omega$. And Fatou's lemma guarantees the order switch in the following way.

Theorem 12 (Fatou). *Let $X_n \geq 0$ for all $n \in \mathbb{N}$. Then $\mathbb{E}\left(\liminf_{n\to\infty} X_n\right) \leq \liminf_{n\to\infty} \mathbb{E}X_n$.*

Proof. Define $Y_n = \inf_{k\geq n} X_k$ for each $n \in \mathbb{N}$. Then we have $X_n \geq 0$ and

$$Y_n \uparrow \liminf_{n\to\infty} X_n \quad \overset{\text{MCT}}{\Longrightarrow} \quad \lim_{n\to\infty} \mathbb{E}Y_n = \mathbb{E}\left(\liminf_{n\to\infty} X_n\right).$$

We further notice that $Y_n = \inf_{k\geq n} X_k \leq X_n$ for all $n \in \mathbb{N}$. By the integral properties, it follows that $\mathbb{E}Y_n \leq \mathbb{E}X_n$ as a sequence of number for all $n \in \mathbb{N}$, and

$$\lim_{n\to\infty} \mathbb{E}Y_n = \liminf_{n\to\infty} \mathbb{E}Y_n \leq \liminf_{n\to\infty} \mathbb{E}X_n.$$

Combining the two quantities, we conclude Fatou's lemma. □

The last convergence theorem is the most general since it does not require increasing, and also reveals equality. It is the dominated convergence theorem.

Theorem 13 (DCT). *Let (X_n) be a sequence of rvs such that $|X_n| \leq Y$ for some $Y \in L_1(\mathbb{P})$. If $X_n \to X$ almost surely, then $X \in L_1(\mathbb{P})$ and $\lim_{n\to\infty} \mathbb{E}X_n$ exists with*

$$\lim_{n\to\infty} \mathbb{E}X_n = \mathbb{E}X.$$

Proof. With $|X_n| \leq Y$, we know $\mathbb{E}|X| \leq \mathbb{E}Y < \infty$, hence $X \in L_1(\mathbb{P})$.

Further with almost surely $Y - X_n \geq 0$ and $(Y - X_n) \to (Y - X) \geq 0$, by Fatou's lemma,

$$\mathbb{E}Y - \mathbb{E}X = \mathbb{E}(Y - X) = \mathbb{E}\left[\liminf_{n\to\infty}(Y - X_n)\right] \leq \liminf_{n\to\infty} \mathbb{E}(Y - X_n) = \mathbb{E}Y - \limsup_{n\to\infty} \mathbb{E}X_n,$$

where the first and last equalities follows from integral properties. Rearranging terms leads to

$$\limsup_{n\to\infty} \mathbb{E}X_n \le \mathbb{E}X.$$

Using $Y + X_n \ge 0$ and $(Y + X_n) \to (Y + X) \ge 0$, by Fatou's lemma again,

$$\mathbb{E}Y + \mathbb{E}X = \mathbb{E}(Y + X) = \mathbb{E}\Big[\liminf_{n\to\infty}(Y + X_n)\Big] \le \liminf_{n\to\infty} \mathbb{E}(Y + X_n) = \mathbb{E}Y + \liminf_{n\to\infty} \mathbb{E}X_n.$$

Combining two inequalities, we get $\limsup \mathbb{E}X_n \le \mathbb{E}X \le \liminf \mathbb{E}X_n$, forcing everything to be equal. □

1.3 Why DCT May Fail?

As we said, convergence theorems answer the question: when lim can be switched with integral. By DCT, the answer is affirmative when there exists a dominating function $Y \ge |X_n|$. And this dominating function, in essence, prevents two things:

- some area under the graph escapes to infinity as $n \to \infty$,
- some measure-zero set becomes unbounded.

Example 2. Consider the space $(\mathbb{R}, \mathcal{B}_\mathbb{R}, \lambda)$ with functions $f_n(x) = \mathbb{1}_{(n,n+1]}(x)$, whose area under the graph escapes to infinity. Since for any fixed $x \in \mathbb{R}$, there exists $N \in \mathbb{N}$ s.t. for all $n > N$, $f_n(x) = 0$, we have $\lim_{n\to\infty} f(x) \equiv 0$, and the strictly inequality

$$\int \lim_{n\to\infty} f_n d\lambda = 0 < 1 = \lim_{n\to\infty} \int f_n d\lambda.$$

Example 3. Consider the probability space $((0, 1], \mathcal{B}_{(0,1]}, \mathbb{P})$ with rvs $X_n = n\mathbb{1}_{(0,1/n]}$. Hence $X_n \to 0$, but

$$\mathbb{E}\Big[\lim_{n\to\infty} X_n\Big] = 0 < 1 = \lim_{n\to\infty} \mathbb{E}X_n.$$

This happens because at $\omega = 0$, the function blows up. And if we want to find another smaller rv Y that dominates all X_n, the choice must be $Y \equiv \infty$, which is not integrable.

Remark. Though the DCT fails in the two cases, Fatou's lemma still holds.

The DCT provides a one-way ticket from the convergence of rvs $X_n \to X$ themselves to the convergence of their expectations $\mathbb{E}X_n \to \mathbb{E}X$ as a real sequence. However, reflecting on Example 3, we observe something remarkably strong:

$$\begin{cases} X_n \to 0 \ \text{ everywhere}, \\ \mathbb{E}|X_n| = 1 \ \text{ for all } n \in \mathbb{N}, \end{cases} \qquad \text{BUT } \mathbb{E}X_n \not\to \mathbb{E}X!$$

This raises a natural question: is there any way to reverse the process? The answer is yes! The following concept provides the last word to end the integral convergence topic.

Definition 14 (Uniform integrability)**.** A sequence of functions $(X_n)_{n=1}^\infty \subset L_1(\mathbb{P})$ is called uniformly integrable if for any $\epsilon > 0$, there exists $M > 0$ such that

$$\mathbb{E}\big\{|X_n|\mathbb{1}_{\{|X_n|>M\}}\big\} \le \epsilon \ \text{ for all } n \in \mathbb{N}.$$

In other words, we are requiring the whole sequence has a similar tail behavior.

Remark. In the CLT theory, we may encounter another weaker condition which does a similar thing to control the tail behavior: *tightness*. It applies to all rv sequence (doesn't need the sequence to be in L_1), requiring $\mathbb{P}(|X_n| > M) < \epsilon$ for all $n \in \mathbb{N}$.

The uniform integrability prevents both two cases in §1.2.1 that fails the DCT to happen, as in the next characterization.

Proposition 15. *Let $(X_n)_{n=1}^{\infty} \subset L_1(\mathbb{P})$ be a sequence of rvs. (X_n) is uniformly integrable iff*

(i) for any $\epsilon > 0$, there exists $\delta > 0$ such that for all event $A \subset \Omega$ with $\mathbb{P}(A) < \delta$,

$$\mathbb{E}\big(|X_n|\mathbb{1}_A\big) < \epsilon \text{ for all } n \in \mathbb{N};$$

(ii) and there exists $K > 0$ such that $\mathbb{E}|X_n| < K$ for all $n \in \mathbb{N}$.

Proof. Suppose (X_n) is ui. Let $\epsilon > 0$, and $A \subset \Omega$ be an event. Then for all $n \in \mathbb{N}$, it follows that

$$\mathbb{E}|X_n|\mathbb{1}_A = \mathbb{E}\big[|X_n|\mathbb{1}_{A\cap\{|X_n|>M|\}}\big] + \mathbb{E}\big[|X_n|\mathbb{1}_{A\cap\{|X_n|\le M|\}}\big] \le \mathbb{E}\big[|X_n|\mathbb{1}_{\{|X_n|>M|\}}\big] + M\cdot\mathbb{P}(A)$$

using ui for the first term, and choosing A with $\mathbb{P}(A) < \epsilon/M$, we get

$$\le \epsilon + \epsilon = 2\epsilon,$$

completing (i). For (ii), take $\epsilon = 1$ and decompose $\mathbb{E}|X_n| = \mathbb{E}|X_n|\mathbb{1}_{\{|X_n|>M\}} + \mathbb{E}|X_n|\mathbb{1}_{\{|X_n|\le M\}} = 1 + M$.

Suppose (i) and (ii) together. Then by Markov's inequality and (ii), we have

$$\mathbb{P}(|X_n| > M) \le \frac{\mathbb{E}|X_n|}{M} \le \frac{K}{M}, \text{ for all } n \in \mathbb{N}.$$

Hence let $\epsilon > 0$ and $\delta > 0$ be the one given by (i). We can choose $M > K/\delta$ to get $\mathbb{P}(|X_n| > M) \le \delta$ for all $n \in \mathbb{N}$, which makes $\mathbb{E}|X_n|\mathbb{1}_{\{|X_n|>M\}} \le \epsilon$ by (i). □

Using (i), ui ensures that the mass of (X_n) does not concentrate excessively in a small region, much like the case of $X_n = n\mathbb{1}_{(0,1/n]}$. Meanwhile, (ii) prevents (X_n) from having mass escape to infinity, as in the example $f_n(x) = \mathbb{1}_{(n,n+1]}(x)$. With ui in place, we can establish an equivalent condition for $\mathbb{E}X_n \to \mathbb{E}X$.

Theorem 16. *Let $(X_n)_{n=1}^{\infty}$ be a sequence of $L_1(\mathbb{P})$ rvs such that $X_n \to X$ in probability. TFAE:*

(i) $(X_n)_{n=1}^{\infty}$ is uniformly integrable, *(ii) $\mathbb{E}|X_n - X| \to 0$,* *(iii) $\mathbb{E}|X_n| \to \mathbb{E}|X|$.*

This is a solid theorem, implying the completeness of the L_1 space. We omit the proof here, but it can be found in Resnick [4] or my MA625 notes.

Remark. Note that (ii) implies $\mathbb{E}X_n \to \mathbb{E}X$. Consequently, $X_n \xrightarrow{p} X$ with ui ensures $\mathbb{E}X_n \to \mathbb{E}X$, and vice versa, concluding our discussion on integral convergence. However, in practice, when switching the order of integration and limits, the dominating condition is typically checked instead of ui even if ui is weaker.

2 Distribution

During the reconstruction of expectation, we avoided any reference to the specific distributions of rvs, such as Poisson or normal. In fact, we did not even define what the distribution means for a rv. In this section,

we show how the modern version of expectation aligns with the intuitive understanding from elementary probability. To do this formally, we introduce the following tool.

Definition 17. Let $(\Omega, \mathcal{A})$ be a measurable space, and let μ and ν be two measures on it. We say that ν is absolutely continuous with respect to μ, denoted $\nu \ll \mu$, if $\mu(A) = 0 \Rightarrow \nu(A) = 0$ for all $A \in \mathcal{A}$.

If we suppose on $(\Omega, \mathcal{A}, \mu)$, we have a non-negative function f. Then we can define a new function ν by

$$\nu(A) := \int_A f d\mu, \quad \forall A \in \mathcal{A}.$$

It can be checked that $\nu(\cdot)$ is also a measure on $(\Omega, \mathcal{A})$ with the property

$$\mu(A) = 0 \quad \Rightarrow \quad \nu(A) = 0.$$

So defining in this way, we have $\nu \ll \mu$. The Radon-Nikodym theorem provides its converse.

Theorem 18 (Radon-Nikodym)**.** *Let $(\Omega, \mathcal{A})$ be a measurable space, and $\nu \ll \mu$ be two σ-finite measures on it. Then there exists a μ-almost everywhere measurable function $f \geq 0$ such that*

$$\nu(A) = \int_A f d\mu, \quad \forall A \in \mathcal{A}.$$

And we write $f := \frac{d\nu}{d\mu}$, and call it the Radon-Nikodym derivative.

The proof of this theorem can be done in various ways, such as using the Riesz representation, as outlined in Resnick [4]. We omit the proof here. In fact, we have used this theorem many times without explicitly recognizing it. It is implicitly present behind the change of variable rule in integration, as we can derive a simple proposition beyond Theorem 18:

Proposition 19. *Let $(\Omega, \mathcal{A})$ be a measurable space with two measures ν and μ such that $\nu \ll \mu$. Let g be a $L_1(\nu)$ function. Then*

$$\int g d\nu = \int g \frac{d\nu}{d\mu} d\mu, \quad \textit{where } \frac{d\nu}{d\mu} \textit{ is the RN derivative of } \nu \textit{ wrt } \mu.$$

Proof. The idea of proof is starting from indicator function to all L_1 functions. □

To see the usage of the RN derivative, we can consider this example:

$$\int_0^1 \sin(x^2) d(x^2) = \int_0^1 \sin(x^2)\,(2x)\,dx, \quad \text{where } 2x = \frac{d\mu_{x^2}}{d\mu_x}.$$

So any change of variable in Calculus 1 can be viewed as a RN derivative.

2.1 Computing Expectation by Distribution

Suppose we have a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ with a measurable function $X : (\Omega, \mathcal{A}) \mapsto (\mathbb{S}, \mathcal{S})$. Different choices of $(\mathbb{S}, \mathcal{S})$ make X have different names:

- if $(\mathbb{S}, \mathcal{S}) = (\mathbb{R}, \mathcal{B}_{\mathbb{R}})$, X is called a random variable;

- if $(\mathbb{S}, \mathcal{S}) = (\mathbb{R}^p, \mathcal{B}_{\mathbb{R}^p})$, X is called a random vector;
- if $(\mathbb{S}, \mathcal{S}) = (\mathbb{H}, \mathcal{B}_{\mathbb{H}})$, a topological Hilbert space, X is called a random function;
- in general, if the fundamental space is a probability space, we can call X a random element.

The set function on $(\mathbb{S}, \mathcal{S})$, defined by $\mu_X(S) := \mathbb{P}(X \in S)$ for all $S \in \mathcal{S}$, is called the **distribution** of X, sometimes is also called the induced measure. And we can easily check, no matter what $(\mathbb{S}, \mathcal{S})$ is, the distribution is always a measure.

Proposition 20. *Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space, and let $X : (\Omega, \mathcal{A}) \mapsto (\mathbb{S}, \mathcal{S})$ be a measurable function. Then μ_X is a probability on $(\mathbb{S}, \mathcal{S})$.*

Proof. Everything follows from the properties of an inverse mapping. □

Now, suppose there is another measurable mapping $g : (\mathbb{S}, S) \mapsto (\mathbb{R}, \mathcal{B}_{\mathbb{R}})$. For simplicity, we denote $g(X(\omega)) =: Y(\omega)$, then Y is a measurable function from $(\Omega, \mathcal{A})$ to $(\mathbb{R}, \mathcal{B}_{\mathbb{R}})$, and also induces a measure on $(\mathbb{R}, \mathcal{B}_{\mathbb{R}})$ by $\mu_Y(B) := \mathbb{P}(Y \in B)$ for all $B \in \mathcal{B}_{\mathbb{R}}$.

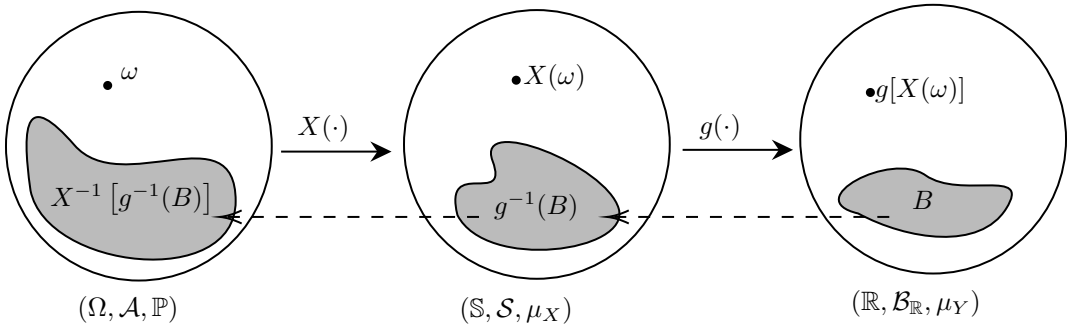

Figure 1: Measurable transformations

The question is: how do we compute the expectation $\mathbb{E}[Y]$ or $\mathbb{E}[g(X)]$? Here lies the most important tool for statisticians: it allows us to compute expectations without directly knowing the underlying probability space.

Theorem 21 (Change of variable formula)**.** *Let X be a random element from $(\Omega, \mathcal{A}, \mathbb{P})$ to $(\mathbb{S}, S)$, and g be a measurable function from $(\mathbb{S}, S)$ to $(\mathbb{R}, \mathcal{B}_{\mathbb{R}})$ with $g \geq 0$ or $\mathbb{E}|g(X)| < \infty$. Then*

$$\mathbb{E}g(X) = \int_\Omega g(X(\omega))\mathbb{P}(d\omega) = \int_{\mathbb{S}} g(x)\mu_X(dx) = \int_{\mathbb{R}} y\ \mu_Y(dy),$$

where μ_Y is the measure induced by $Y := g(X)$.

The proof can be found in Durrett[1], and this theorem states three ways to compute the expectation:

- from the underlying $(\Omega, \mathcal{A})$ with probability $\mathbb{P}$,
- from the intermediate $(\mathbb{S}, \mathbb{S})$ with the distribution μ_X,
- from the upper $(\mathbb{R}, \mathcal{B}_{\mathbb{R}})$ with the distribution μ_Y.

In most cases, we assume that the X_i's are iid from a specified distribution, as is typical in parametric models, even if the parameters are unknown. In practice, no one explicitly refers to the underlying probability space, which is why this is often called the *theorem of the unconscious statistician*(LOTUS). Moreover, when the distribution of X is given and we wish to compute the expected value of $Y = g(X)$, we can bypass computing the distribution of Y entirely.

Remark. The change of variable formula serves as a general solution for computing any expectation in the form of $g(X)$, with one caveat: the final result $g(X)$ must be a real-valued function, even though the intermediate random element X may take other forms, such as a vector or a function.

As integration theory developed, mathematicians first introduced the Lebesgue integral to handle functions of the form $f : (\mathbb{R}^d, \mathcal{B}_{\mathbb{R}^d}, \lambda) \to (\mathbb{R}^1, \mathcal{B}_{\mathbb{R}^1})$. This was later generalized to functions with domains in abstract measure spaces, $f : (\mathbb{S}, \mathcal{S}, \mu) \to (\mathbb{R}^1, \mathcal{B}_{\mathbb{R}^1})$. However, extending the range of integration beyond $(\mathbb{R}^1, \mathcal{B}_{\mathbb{R}^1})$ to more abstract spaces came much later, around the 1980s, with the introduction of the *Bochner integral*. In fact, the characteristic function $e^{\mathbb{i}tX}$ for a fixed $t \in \mathbb{R}$ can be viewed as a mapping $(\Omega, \mathcal{A}, \mathbb{P}) \to (\mathbb{C}, \mathcal{B}_{\mathbb{C}})$. Fortunately, it is straightforward to handle without resorting to the Bochner integral.

In the remainder of this section, we will derive specific formulas for continuous and discrete random variables and confirm their consistency with elementary probability theory.

2.2 Continuous Random Variables

We say a random variable X is (absolutely) **continuous** if its distribution $\mu_X \ll \lambda$, where λ denotes the Lebesgue measure. By Radon-Nikodym theorem, $\mu_X \ll \lambda$ implies there exists a non-negative function f_X s.t.

$$\mu_X(B) = \int_B f_X d\lambda, \quad \forall B \in \mathcal{B}_{\mathbb{R}}.$$

By change of variable and Proposition 19 , we have for any function g s.t. $\mathbb{E}|g(X)| < \infty$,

$$\mathbb{E}g(X) = \int_{\mathbb{R}} g(x)\mu_X(dx) = \int g(x)f_X(x)dx.$$

2.3 Discrete Random Variables

The idea is the same but we need to deal with the integral on a discrete space first. And we shall see it is actually a summation.

Integration under counting measures

A measure μ is said to be a **counting measure** μ if $\mu = 1$ on a countable subset $C = \{c_1, c_2, \dots\}$ and $\mu = 0$ everywhere else (ex. check this is a measure on $(\mathbb{R}, \mathcal{B}_{\mathbb{R}})$). First, we assume $f \geq 0$. Define for all $k \in \mathbb{N}$:

$$g_k(x) = \sum_{i=1}^{k} f(c_i)\mathbb{1}_{c_i}(x) \quad \Rightarrow \quad \lim_{k\to\infty} g_k \uparrow = \sum_{i=1}^{\infty} f(c_i)\mathbb{1}_{c_i}(x) + 0 \cdot \mathbb{1}_{C^c}(x) =: \tilde{f}(x).$$

Then by the counting measure and MCT, we have

$$\lim_{k\to\infty} \sum_{i=1}^{k} f(c_i) = \lim_{k\to\infty} \int g_k d\mu = \int \lim_{k\to\infty} g_k d\mu = \int \tilde{f} d\mu.$$

Note that f can be represented as

$$f(x) = \sum_{i=1}^{\infty} f(c_i)\mathbb{1}_{c_i}(x) + f(x) \cdot \mathbb{1}_{C^c}(x),$$

which implies $f = \tilde{f}$ μ-almost everywhere, further meaning that, by the integral properties,

$$\int f d\mu = \int \tilde{f} d\mu = \lim_{k\to\infty} \sum_{i=1}^{k} f(c_i) = \sum_{i=1}^{\infty} f(c_i) \text{ and } \int_B f d\mu = \int_B \tilde{f} d\mu = \sum_{i:c_i \in B}^{\infty} f(c_i)$$

for any $B \subset \mathbb{R}$. This conclude the integral for the counting measures, which in essense is the summation as we promised. And this integral in the summation form will lead us to the expectation for discrete rvs.

Expectation for discrete random variables

We say a random variable is **discrete** if there exists a countable subset $C = \{c_1, c_2, \dots\}$ of $\mathbb{R}$ such that $\mathbb{P}(X \in C) = 1$. We further define the counting measure on C by

$$\mu(B) = \#(B \cap C), \quad \forall B \in \mathcal{B}_{\mathbb{R}}.$$

Then we claim $\mu_X \ll \mu$, i.e. the distribution of X is dominated by the counting measure.

Proof. Suppose $N \subset \mathbb{R}$ with $\mu(N) = \#(N \cap C) = 0$. By a counting measure, this means $N \cap C = \emptyset$, and hence $N \subset C^c$. Then $\mu_X(N) = \mathbb{P}(X \in N) \leq \mathbb{P}(X \in C^c) = 0$. □

Remark. This proposition indicates that if X is discrete, then the induced measure space $(\mathbb{R}, \mathcal{B}_{\mathbb{R}}, \mu_X)$ is dominated by a counting measure instead of the Lebesgue measure.

So, by Radon-Nikodym theorem, there exists a μ-almost everywhere defined function $p(x) \geq 0$ such that

$$\mu_X(B) = \int_B p d\mu = \int p \mathbb{1}_B d\mu = \sum_{x_i \in B} p(x_i),$$

where the last equality comes from the counting measure. This is consistent with the definition of a mass function. But here, we allow p to take any values on C^c, since $\mu(C^c) = 0$. The last thing we check is its expectation. For any function g s.t. $\mathbb{E}|g(X)| < \infty$, change of variable formula, Proposition 19 and counting measure integral give

$$\mathbb{E}g(X) = \int_{\mathbb{R}} g(x)\mu_X(dx) = \int_{\mathbb{R}} g(x)p(x)d\mu(x) = \sum_{i=1}^{\infty} g(x_i)p(x_i).$$

2.4 Random Vector

When the range space $(\mathbf{S}, \mathcal{S})$ is $(\mathbb{R}^n, \mathcal{B}_{\mathbb{R}^n})$, in other words $\mathbf{X} : (\Omega, \mathcal{A}, \mathbb{P}) \mapsto (\mathbb{R}^n, \mathcal{B}_{\mathbb{R}^n})$, the distribution of $\mathbf{X}$, by Proposition 20, is still a measure on $(\mathbb{R}^n, \mathcal{B}_{\mathbb{R}^n})$ defined as

$$\mu_{X_1,\dots,X_n}(B) = \mathbb{P}\left\{(X_1, \dots, X_n) \in B\right\}, \quad \forall B \in \mathcal{B}_{\mathbb{R}^n}.$$

Remark. On the metric/topological space $(\mathbb{R}^n, d)$ where d is the usual metric, the Borel σ-algabra on it is defined to be $\mathcal{B}_{\mathbb{R}^n} = \sigma\{U \subset \mathbb{R}^n : U \text{ is open}\}$, the smallest σ-algebra containing all open sets. But with the good structure of $\mathbb{R}^n$, it can be shown that $\mathcal{B}_{\mathbb{R}^n} = \sigma\{E_1 \times \cdots \times E_n : E_i \in \mathcal{B}_{\mathbb{R}}\}$ (in Folland [2] or my ST6103 HW4). This property will be useful (such as to show $\mathbf{X}$ is $\mathcal{A}/\mathcal{B}_{\mathbb{R}^n}$ measurable iff each X_i is $\mathcal{A}/\mathcal{B}_{\mathbb{R}^1}$) and reflected when we define independence, as it allows us to decompose events in $\mathbb{R}^n$ into components from $\mathbb{R}$.

Sometimes, $\mu_{X_1,\dots,X_n}$ is also called the joint distribution. And in this case, for $g : (\mathbb{R}^n, \mathcal{B}_{\mathbb{R}^n}) \mapsto (\mathbb{R}, \mathcal{B}_{\mathbb{R}})$, the change of variable formula says

$$\mathbb{E}g(\mathbf{X}) = \int_{\Omega} g(\mathbf{X}(\omega))\mathbb{P}(d\omega) = \int_{\mathbb{R}^n} g(\mathbf{x})\mu_{X_1,\dots,X_n}(d\mathbf{x}).$$

Let's now consider $\mathbf{X} = (X, Y)$ and $g(X, Y) = XY$. Then the change of variable formula says

$$\mathbb{E}(XY) = \int_\Omega X(\omega)Y(\omega)\mathbb{P}(d\omega) = \int_{\mathbb{R}^2} xy\ \mu_{X,Y}(dxdy).$$

One may wonder, can we split the integral further as

$$\mathbb{E}(XY) = \int_{\mathbb{R}^2} xy\ \mu_{X,Y}(dxdy) = \int_{\mathbb{R}}\int_{\mathbb{R}} xy\ \mu_X(dx)\mu_Y(dy) = \int_{\mathbb{R}} y\left[\int_{\mathbb{R}} x\ \mu_X(dx)\right]\mu_Y(dy) = \mathbb{E}X \cdot \mathbb{E}Y?$$

The answer is NO in general, and the issue lies in the second equality, where the integral over $\mathbb{R}^2$ is split into two iterative integrals over $\mathbb{R}^1$. This step is valid only when the measure on $\mathbb{R}^2$ is a product measure of the two measures on $\mathbb{R}^1$, as guaranteed by Fubini's theorem. However, in our case $\mu_{X,Y}$ on $\mathbb{R}^2$ is not necessarily the product of μ_X and μ_Y on $\mathbb{R}^1$.

What does it mean for $\mu_{X,Y}$ to be the product of μ_X, μ_Y? And when is it? This leads naturally leads us to the next section: independence. We shall see $\mu_{X,Y}$ is the product of μ_X, μ_Y when (X, Y) is independent.

3 Independence

The concept of independence marks a key divergence between probability theory and measure theory. On one hand, it abstracts the idea of independent events observed in the real world. On the other hand, it prepares the ground for later topics such as the laws of large numbers. So what is it?

Definition 22 (Independence)**.** Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space. We say

(i) two events A and B are independent if $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$;

(ii) two classes of sets $\mathcal{F}$ and $\mathcal{G}$ are independent if the events A and B are independent for all $A \in \mathcal{F}$ and $B \in \mathcal{G}$;

(iii) two random elements X and Y are independent if $\sigma(X)$ and $\sigma(Y)$ are independent.

Moreover, let I be an index set. We say

(iv) the classes of sets $(\mathcal{F}_i)_{i\in I}$ are independent if for any finite set $F \subset I$ and any $A_i \in \mathcal{F}_i$ with $i \in F$,

$$\mathbb{P}\left(\cap_{i\in F} A_i\right) = \prod_{i\in F} \mathbb{P}(A_i);$$

(v) the random elements $(X_i)_{i\in I}$ are independent if $\{\sigma\left(X_i\right)\}_{i\in I}$ are independent.

The (iv) and (v) allow us to make the independent argument for infinite sequence of rvs, and we can check (iv) and (v) are consistent with (ii) and (iii) when $I = \{1, 2\}$, hence the definition is well-defined.

Furthermore, we allow the random elements to have different range space, such as $X : (\Omega, \mathcal{A}) \mapsto (\mathbb{S}_1, \mathcal{S}_1)$ and $Y : (\Omega, \mathcal{A}) \mapsto (\mathbb{S}_2, \mathcal{S}_2)$. With the definition $\sigma(X) = \sigma\{X^{-1}(S_1) : S_1 \in \mathcal{S}_1\}$, the independence can always be defined even when $\mathbb{S}_1 \neq \mathbb{S}_2$.

Example 4. Let $X : (\Omega, \mathcal{A}) \mapsto (\mathbb{R}, \mathcal{B}_\mathbb{R})$ be a rv and $(Y, Z) : (\Omega, \mathcal{A}) \mapsto \left(\mathbb{R}^2, \mathcal{B}_{\mathbb{R}^2}\right)$ be a random vector. By saying X and (Y, Z) are independent, it is required to hold that for all $B_1 \in \mathcal{B}_\mathbb{R}$ and $B_2 \in \mathcal{B}_{\mathbb{R}^2}$,

$$\mathbb{P}[X \in B_1,\ (Y, Z) \in B_2] = \mathbb{P}(X \in B_1) \cdot \mathbb{P}[(Y, Z) \in B_2].$$

Independence can also be defined for a random variable $X : (\Omega, \mathcal{A}) \to (\mathbb{R}, \mathcal{B}_\mathbb{R})$ and a random function $Y : (\Omega, \mathcal{A}) \to (\mathbb{H}, \mathcal{B}_\mathbb{H})$. As you might notice, regardless of the range space, we often prefer the Borel σ-

algebra as long as there is a topology. Why? Because checking independence across the entire σ-algebra is cumbersome, while the Borel σ-algebra offers a much simpler approach.

Lemma 23. *Let $(\mathcal{C}_i)_{i\in I}$ be independent π-classes. Then $\{\sigma(\mathcal{C}_i)\}_{i\in I}$ are independent.*

Proof. It can be done by Dynkin's π-λ theorem, or the approximation by σ-δ sets in my MA625 course, or the Carathéodory extension in my ST6103 course. □

Recalling the fact that $\mathcal{B}_{\mathbb{R}} = \sigma\{(-\infty, x] : x \in \mathbb{R}\}$, Lemma 23 matched the definition of independence in elementary probability: the random variables $X_1, \ldots, X_n$ are independent if, for any $x_1, \ldots, x_n \in \mathbb{R}$,

$$\mathbb{P}(X_1 \le x_1, \ldots, X_n \le x_n) = \mathbb{P}(X_1 \le x_1) \times \cdots \times \mathbb{P}(X_n \le x_n).$$

Taking partial derivatives, when the densities exist, gives the density version of this condition. Furthermore, there is another way to check independence without looking at the distribution functions.

Lemma 24 (Grouping)**.** *Let $(\mathcal{F}_{ij} : i = 1, \ldots, n;\ j = 1, \ldots, m_i)$ be independent σ-algebras and let $\mathcal{G}_i = \sigma\left(\cup_j \mathcal{F}_{i,j}\right)$. Then $\mathcal{G}_1, \ldots, \mathcal{G}_n$ are independent.*

Proof. Define $\mathcal{A}_i = \left\{\cap_{j=1}^{m_i} A_{ij} : A_{ij} \in \mathcal{F}_{ij}\right\}$ for all $i = 1, \ldots, n$. Then each $\mathcal{A}_i$ is a π-class containing $\cup_{j=1}^{m_i} \mathcal{F}_{ij}$. Hence, by Lemma 23, we know $\{\sigma(\mathcal{A}_i)\}_{i=1}^n$ are independent, and so are $\mathcal{G}_1, \ldots, \mathcal{G}_n$. □

This lemma explains why functions of independent rvs are independent, such as the following concrete example.

Example 5. Let $X_1, \ldots, X_4$ be ind standard normal rvs, and let $Y_1 = X_1/X_2$ and $Y_2 = \exp\{X_3 + X_4\}$. If we consider $\mathcal{F}_{11} = \sigma(X_1)$, $\mathcal{F}_{12} = \sigma(X_2)$ and $\mathcal{F}_{21} = \sigma(X_3)$, $\mathcal{F}_{22} = \sigma(X_4)$, then $\sigma(X_1, X_2)$ and $\sigma(X_3, X_4)$ are independent, hence so are Y_1 and Y_2.

3.1 Product Space

Let's revisit the earlier question. Given $\mathbf{X} = (X, Y)$ as a random vector, can we compute

$$\mathbb{E}(XY) = \int_{\mathbb{R}^2} xy\ \mu_{X,Y}(dxdy) = \int_{\mathbb{R}} \left[\int_{\mathbb{R}} xy\ \mu_X(dx)\right] \mu_Y(dy) = \int_{\mathbb{R}} \left[\int_{\mathbb{R}} xy\ \mu_Y(dy)\right] \mu_X(dx) = \mathbb{E}X \cdot \mathbb{E}Y?$$

At first glance, it seems plausible. In Lebesgue integration theory, for a function $g : \mathbb{R}^2 \to \mathbb{R}$ that is either integrable or non-negative, we know:

$$\iint_{\mathbb{R}^2} g(x,y) d\lambda_2(x,y) = \int_{\mathbb{R}} \left[\int_{\mathbb{R}} g(x,y) d\lambda_1(x)\right] d\lambda_1(y) = \int_{\mathbb{R}} \left[\int_{\mathbb{R}} g(x,y) d\lambda_1(y)\right] d\lambda_1(x),$$

where λ_2 is the Lebesgue measure on $\left(\mathbb{R}^2, \mathcal{B}_{\mathbb{R}^2}\right)$, and λ_1 is the Lebesgue measure on $(\mathbb{R}, \mathcal{B}_{\mathbb{R}})$.

However, a hidden condition for this result is that the space on the left must be a product space, i.e. $\left(\mathbb{R}^2, \mathcal{B}_{\mathbb{R}^2}\right) = (\mathbb{R}, \mathcal{B}_{\mathbb{R}}) \otimes (\mathbb{R}, \mathcal{B}_{\mathbb{R}})$, and the measure on the left must be a product measure, i.e. $\lambda_2 = \lambda_1 \times \lambda_1$. Only in this *product* sense can we split the integral, and that's where independence helps, and also what we will explore it formally in this section. Let's begin.

Definition 25 (Product space)**.** Let $(\Omega_1, \mathcal{A}_1, \mu_1)$ and $(\Omega_2, \mathcal{A}_2, \mu_2)$ be two measure spaces. We call

$$\mathcal{A}_1 \otimes \mathcal{A}_2 = \sigma\left\{A_1 \times A_2 : A_1 \in \mathcal{A}_1 \text{ and } A_2 \in \mathcal{A}_2\right\} \text{ the product } \sigma\text{-algebra},$$

and call the product measure by $\mu_1 \times \mu_2(A_1 \times A_2) = \mu_1(A_1) \times \mu_2(A_2)$ for all $A_1 \in \mathcal{A}_1$ and $A_2 \in \mathcal{A}_2$.

Remark. When $\mathcal{A}_1 = \mathcal{A}_2 = \mathcal{A}$, we also write $\mathcal{A}^2$ for $\mathcal{A}_1 \otimes \mathcal{A}_2$. And this definition naturally extends to multiple or even countable product space. But for uncountable case, the thing is a little different, and we will not cover it.

The newly defined $(\Omega_1 \times \Omega_2, \mathcal{A}_1 \otimes \mathcal{A}_2)$ is clearly a measurable space. However, the function $\mu_1 \times \mu_2$ is initially defined only on a restricted subset of $\mathcal{A}_1 \otimes \mathcal{A}_2$, namely $\{A_1 \times A_2 : A_1 \in \mathcal{A}_1,\ A_2 \in \mathcal{A}_2\}$. Why do we call $\mu_1 \times \mu_2$ a measure? Because $\{A_1 \times A_2 : A_1 \in \mathcal{A}_1,\ A_2 \in \mathcal{A}_2\}$ forms a semi-algebra, and Carathéodory extends it to $\mathcal{A}_1 \otimes \mathcal{A}_2$ (uniquely when two measures are σ-finite).

Example 6 (Eucleadean space)**.** Take $(\Omega_1, \mathcal{A}_1, \mu_1) = (\Omega_2, \mathcal{A}_2, \mu_2) = (\mathbb{R}, \mathcal{B}_\mathbb{R}, \lambda)$. Then the product σ-algebra is $\mathcal{B}_\mathbb{R} \otimes \mathcal{B}_\mathbb{R} = \sigma\{A_1 \times A_2 : A_1, A_2 \in \mathcal{B}_\mathbb{R}\}$ with product measure $\lambda \times \lambda(A_1 \times A_2) = \lambda(A_1) \times \lambda(A_2)$. And the triplet $\left(\mathbb{R}^2, \mathcal{B}_\mathbb{R} \otimes \mathcal{B}_\mathbb{R}, \lambda \times \lambda\right)$ forms a measure space.

But a natural question is whether $\left(\mathbb{R}^2, \mathcal{B}_\mathbb{R} \otimes \mathcal{B}_\mathbb{R}, \lambda \times \lambda\right) = (\mathbb{R}^2, \mathcal{B}_{\mathbb{R}^2}, \lambda_2)$, where the latter is what we learned in real analysis. We already encountered the affirmative answer when discussing the random vector. The proof can be found in *Real Analysis*[2] or my ST6103 HW4. And a formal statement for it is $\sigma\{B \subset \mathbb{R}^n : B \text{ is open}\} = \sigma\{B_1 \times \cdots \times B_n : B_i \in \mathcal{B}_\mathbb{R}\}$.

As we promised, the product measure space has a good property to split the integral for integrable functions, known as Fubini's theorem. We now state the it together with the Tonelli's theorem, which is for the non-negative measurable functions.

Theorem 26 (Fubini)**.** *Let $(\Omega_1, \mathcal{A}_1, \mu_1)$, $(\Omega_2, \mathcal{A}_2, \mu_2)$ be two σ-finite measure spaces, and let $f \geq 0$ be $\mathcal{A}_1 \otimes \mathcal{A}_2$ measurable [and $f \in L^1(\mu_1 \times \mu_2)$]. Then*

(i) for μ_1-almost everywhere $\omega_1 \in \Omega_1$, the slicing f_{ω_1} is $\mathcal{A}_2$ is measurable [and in $L_1(\mu_2)$],

(ii) the μ_1-almost everywhere defined function $g(\omega_1) = \int f_{\omega_1} d\mu_2$ is $\mathcal{A}_1$ is measurable [and in $L_1(\mu_1)$],

(iii) moreover, the following integrals are equal:

$$\int f d(\mu_1 \times \mu_2) = \int \left[\int f(\omega_1, \omega_2) d\mu_2(\omega_2)\right] d\mu_1(\omega_1) = \int \left[\int f(\omega_1, \omega_2) d\mu_1(\omega_1)\right] d\mu_1(\omega_2).$$

The proof can be found anywhere, such as in Folland [2] or Resnick [4], and the integrability condition is necessary; otherwise, we may have counterexamples where the iterative integrals are not equal.

Example 7. The following examples illustrate how integrability conditions affect the interchange of integration or summation orders.

(7.1) Consider $\left(\mathbb{R}^2, \mathcal{B}_{\mathbb{R}^2}, \lambda\right)$

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

Then $\iint f\,dx\,dy = 1$ but $\iint f\,dy\,dx = -1$. The discrepancy arises because $\iint |f|\,dx\,dy = \infty$, violating the integrability condition.

(7.2) Consider $\left(\mathbb{Z}^2, \mathcal{P}(\mathbb{Z}^2), \#\right)$ with $f(m,n)$ as:

	0	0	0	1	$\dots$
$\uparrow$	0	0	0	1	-1
n	0	1	-1	0	$\dots$
	1	-1	0	0	$\dots$
		m	$\rightarrow$		

Here $\iint f\#(dn)\#(dm) = \sum_m \sum_n f(m,n) = 1$ but $\iint f\#(dm)\#(dn) = \sum_n \sum_m f(m,n) = 0$.

3.2 Expectation under Independence

After our detour into product measures, we can now answer the original question. To have $\mathbb{E}(XY) = \mathbb{E}X \cdot \mathbb{E}Y$, it must hold that

$$\int_{\mathbb{R}^2} xy \; \mu_{X,Y}(dx\, dy) = \int_{\mathbb{R}} \left[\int_{\mathbb{R}} xy \; \mu_X(dx) \right] \mu_Y(dy).$$

From our preparation of product measures, it sufficies to require $\mu_X \times \mu_Y = \mu_{X,Y}$ on $\left(\mathbb{R}^2, \mathcal{B}_{\mathbb{R}^2}\right)$, which is exactly what independence gives us.

Theorem 27. *Let X, Y be independent rvs. Then $\mu_{X,Y} = \mu_X \times \mu_Y$ on $\left(\mathbb{R}^2, \mathcal{B}_{\mathbb{R}^2}\right)$.*

Proof. Dynkin's π-λ theorem is a good way; however, using the uniqueness of Carathéodory is even faster. Noticing that $\mu_{X,Y} = \mu_X \times \mu_Y$ agree on $\{B_1 \times B_2 : B_1, B_2 \in \mathcal{B}_{\mathbb{R}}\}$, which is a generator of $\mathcal{B}_{\mathbb{R}^2}$, we are done. □

By independence, we can make a further leap in the change of variable formula to conclude:

$$\mathbb{E}g(X,Y) = \iint_{\mathbb{R}^2} g(x,y) d\mu_{X,Y}(dx, dy) = \int_{\mathbb{R}} \int_{\mathbb{R}} g(x,y) d\mu_X(x) d\mu_Y(y)$$

provided that $\mathbb{E}|g(X,Y)| < \infty$. And the similar result holds for the distribution on $(\mathbb{R}^n, \mathcal{B}_{\mathbb{R}^n})$ when we have n independent rvs.

3.3 Existence of Random Process

By the change of variable formula, we can now compute expectations given a specified joint distribution, such as $(X,Y) \sim \mathbf{N}(\mathbf{0}, \mathbf{1}_{2\times 2})$, or with an independent structure, such as $X_1, \ldots, X_n$ being iid standard exponential random variables. However, as a mathematical concept, does such a sequence exist? We will address this question in steps.

Specifying Random Variable given Distribution

Consider the probability space $(\Omega, \mathcal{A}, \mathbb{P}) = \left((0,1], \mathcal{B}_{(0,1]}, \lambda\right)$. A uniformly distributed random variable exists on this space by defining $X : (0,1] \mapsto (0,1]$ as $X(\omega) = \omega$. By saying X is uniformly distributed on $(0,1]$, we mean that for any $(a,b] \subset (0,1]$, the probability satisfies $\mathbb{P}(X \in (a,b]) = b - a$.

Given the existence of a uniform rv, for any given cdf F, a rv with this distribution exists as follows: define $X = F^{-1}(U)$, where U is a uniform rv. Recall that F is an increasing and right-continuous function with $F(-\infty) = 0$ and $F(+\infty) = 1$, which introduces a few technical points worth discussing:

(i) Since F is not necessarily one-to-one, what does the inverse function $F^{-1}(u)$ mean?

(ii) How do we define F^{-1} to ensure $F^{-1}(U)$ is a random variable (i.e., a measurable function)?

(iii) Doesn't $F^{-1}(U)$ being a rv with distribution F need any extra conditions?

To address the issue of F not being strictly increasing or continuous, we define $F^{-1}(u) = \inf\{t \in \mathbb{R} : F(t) \geq u\}$ for $u \in [0,1]$. Observe that F^{-1} is increasing by definition, ensuring that $F^{-1}(U)$ is always a random variable. Furthermore, we can show, without requiring any additional conditions, that $F(x) \geq u$ iff $x \geq F^{-1}(u)$ for any $x, u \in \mathbb{R}$. This implies $\{F^{-1}(U) \leq x\} = \{U \leq F(x)\}$, meaning $F^{-1}(U) \sim F$.

Additionally, we can show that $F(X)$ is a uniform random variable **under the extra condition** that F is continuous. However, this is more technical, as proving $F(X)$ is measurable when F is only right-continuous is non-trivial. A complete argument can be found in my ST6103 HW2.

Independent Random Process

Once we can already generate one rv with specified distrbution F in the last section, we may consider to get independent ones in the product space.

Lemma 28. *Let $\mu_1, \dots, \mu_n$ be probaiblity measures on $(\mathbb{R}, \mathcal{B}_\mathbb{R})$. Then there exists a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ and independent rvs $X_1, \dots, X_n$ on it, such that $\mu_{X_i} = \mu_i$ for all $i = 1, \dots, n$.*

Proof. Consider $(\Omega, \mathcal{A}, \mathbb{P}) = (\mathbb{R}^n, \mathcal{B}_{\mathbb{R}^n}, \mu_1 \times \cdots \times \mu_n)$. For any $\omega = (\omega_1, \dots, \omega_n) \in \Omega$, let $X_i(\omega) = \omega_i$. Then

$$\mathbb{P}(X_1 \in B_1, \dots, X_n \in B_n) = \mu_1 \times \cdots \times \mu_n(B_1 \times \cdots \times B_n) = \mu_1(B_1) \times \cdots \times \mu_n(B_n),$$

for all $B_1, \dots, B_n \in \mathcal{B}_\mathbb{R}$, showing both independence and marginal distribution. □

Extending this result to an infinite sequence of random variables is non-trivial. The technical challenge lies in constructing the infinite product $(\mathcal{B}_\mathbb{R})^\infty$, while ensuring two key properties: a generator for $(\mathcal{B}_\mathbb{R})^\infty$ can be defined, and a probability measure can be defined on this generator. This general problem was resolved by Kolmogorov. Moreover, Kolmogorov relaxed the independence condition: he showed that it is possible to construct a random process with any joint distribution, provided the joint distribution satisfies a key property—namely, that it is *consistent*.

Theorem 29 (Kolmogorov's extension)**.** *Let μ_n be probability measures on $(\mathbb{R}^n, \mathcal{B}_{\mathbb{R}^n})$ for each $n \in \mathbb{N}$, satisfying $\mu_{n+1}(B \times \mathbb{R}) = \mu_n(B)$ for all $B \in \mathcal{B}_{\mathbb{R}^n}$. Then there is a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ with a random process $(X_n)_{n=1}^\infty$ such that $(X_1, \dots, X_n)$ has distribution μ_n for any $n \in \mathbb{N}$.*

This theorem can be applied to construct processes such as Brownian motion on a rational trajectory and the normal AR(1) process (see my ST6103 HW4). Additionally, a more general version of the theorem, which does not require a countable collection, exists (see my MA625 notes). This extended version allows for the construction of standard Brownian motion on the entire $\mathbb{R}^+$ trajectory and general Gaussian processes.

4 Conditioning

Another distinguishing feature between probability and measure theory is the concept of conditioning, which encapsulates the idea of *given some known information*. Unlike the usual approach of first defining a distribution and then computing the expectation, we need to define the conditional expectation first, followed by the conditional distribution. We will see why in the end.

Since the basic construction and properties of conditional expectation can be found everywhere, we will omit them here (or see my MA625 notes). Instead, our focus will include:

- The relationship between conditional expectation in elementary probability and its modern counterpart.
- Some interesting and perhaps surprising results with illustrative examples.
- Introducing conditional independence as a natural by-product.
- Finally, exploring conditional distribution with examples.

4.1 Conditional Expectation

On a probability space $L_1(\Omega, \mathcal{A}, \mathbb{P})$, let $X \in L^1(\mathbb{P})$ be a random variable and $\mathcal{F} \subset \mathcal{A}$ be a sub σ-algebra. The **conditional expectation** of X given $\mathcal{F}$, denoted by $\mathbb{E}[X|\mathcal{F}]$, is an $\mathcal{F}$-measurable random variable satisfying

the partial averaging property: $\mathbb{E}[\mathbb{1}_A X] = \mathbb{E}\big[\mathbb{1}_A \mathbb{E}[X|\mathcal{F}]\big]$ for all $A \in \mathcal{F}$. Conditioning on another random variable Y is simply defined as $\mathbb{E}[X|\sigma(Y)]$, which can also be written as $\mathbb{E}[X|Y]$.

In statistics, we often deal with a family of distributions, say $\mathcal{P}$, as inference candidates. Under different distributions, the conditional expectation may vary, so it is standard practice to include a subscript to denote the governing measure. For instance, let $P, Q \in \mathcal{P}$ be two distributions. Then both $\mathbb{E}_P[X|Y]$ and $\mathbb{E}_Q[X|Y]$ are $\sigma(Y)$-measurable random variables, and for all $A \in \sigma(Y)$,

$$\int_A X dP = \int_A \mathbb{E}_P[X|Y] dP \text{ and } \int_A X dQ = \int_A \mathbb{E}_Q[X|Y] dQ,$$

but $\mathbb{E}_P[X|Y] \neq \mathbb{E}_Q[X|Y]$ necessarily. More techniques for integrals under different measures will be introduced in Shao Jun[5] when we discuss the sufficiency and factorization theorem.

$\mathbb{E}[X|Y]$ equals?

Returning to the case where we have a single probability measure $\mathbb{P}$ with two random variables X and Y, the conditional expectation $\mathbb{E}[X|Y]$ is $\sigma(Y)$-measurable. This measurability allows us to say something more.

Proposition 30. *Let $X : (\Omega, \mathcal{A}) \mapsto (\mathbb{S}_1, \mathcal{S}_1)$ and $Y : (\Omega, \mathcal{A}) \mapsto (\mathbb{S}_2, \mathcal{S}_2)$ be two random elements. Then $X \in \sigma(Y)$ if and only if there exists a measurable function $h : (\mathbb{S}_2, \mathcal{S}_2) \mapsto (\mathbb{S}_1, \mathcal{S}_1)$ such that $X = h(Y)$.*

This proposition in particular says $\mathbb{E}[X|Y] = h(Y)$ for some Borel function h, and we further denote $\mathbb{E}[X|Y = y] := h(y)$. While the proof of this proposition (see my ST621 HW3) establishes the existence of such a function h, it is non-constructive and does not explicitly describe what h is. But recall our knowledge in elementary probability, we may have a guess of what it is!

Depending on whether the joint distribution is a mass p or a density f, in elementary probability, we define the conditional mass or density by

$$p_{X|Y=y}(x) = \frac{p(x,y)}{\sum_x p(x,y)} \text{ or } f_{X|Y=y}(x) = \frac{f(x,y)}{\int_{\mathbb{R}} f(x,y)dx}$$

for all $y \in \mathbb{R}$ such that $\sum_x p(x,y) \neq 0$ or $\int_{\mathbb{R}} f(x,y)dx \neq 0$, leaving other y undefined. Then we can compute the elementary conditional expectation X given $Y = y$ by

$$\begin{aligned} E(X|Y=y) &= \sum_x x p_{X|Y=y}(x) = \frac{\sum_x x p(x,y)}{\sum_x p(x,y)} \text{ or,} \\ &= \int_{\mathbb{R}} x f_{X|Y=y}(x)dx = \frac{\int_{\mathbb{R}} x f(x,y)\, dx}{\int_{\mathbb{R}} f(x,y)dx} \end{aligned}$$

for all $y \in \mathbb{R}$ such that $\sum_x p(x,y) \neq 0$ or $\int_{\mathbb{R}} f(x,y)dx \neq 0$, leaving other y undefined. In fact, the discrete case can be viewed as the integral of a density under a counting measure as discussed in §2.3.

Remark. This $E(X|Y = y)$ is defined in a completely different way than $\mathbb{E}[X|Y = y]$ in Proposition 30. Though we will shortly show they are equal as a function of y, don't mix them up!

Note that in both cases, after integrating x out, the conditional expectation $E(X|Y = y)$ is some function of y, say $h(y)$. And plugging the rv Y instead of a fixed $y \in \mathbb{R}$, we get $h(Y)$ as the definition of conditional expectation in elementary probability. In fact, we can see an explicit form of conditional expectation as

$$E(X|Y) := h(Y) = \frac{\int_{\mathbb{R}} x f(x,Y)\, dx}{\int_{\mathbb{R}} f(x,Y)dx}.$$

A natural question is: $\mathbb{E}[X|Y] = E[X|Y]$? Or equivalently is this h the same as the one in Proposition 30? As we can guess, the answer is yes, because this function h here is not coincidental.

Proposition 31. *Let $X \in \mathbb{R}^n$ and $Y \in \mathbb{R}^m$ be two random vectors with a joint density $f(x,y)$ wrt $\nu \times \lambda$, where ν and λ are σ-finite measures on $(\mathbb{R}^n, \mathcal{B}_{\mathbb{R}^n})$ and $(\mathbb{R}^m, \mathcal{B}_{\mathbb{R}^m})$ respectively. Let $g : \mathbb{R}^{n+m} \mapsto \mathbb{R}$ be a Borel function such that $\mathbb{E}|g(X,Y)| < \infty$. Then almost surely,*

$$\mathbb{E}[g(X,Y)|Y] = \frac{\int g(x,Y)f(x,Y)d\nu(x)}{\int f(x,Y)d\nu(x)}.$$

Proof. By Fubini's theorem 26 part (i) and (ii), the integral of the slicing $\int g(x,\cdot)f(x,\cdot)d\nu(x) =: \varphi(\cdot)$ is $\mathcal{B}_{\mathbb{R}^m}$-measurable. Hence $\varphi(Y)$ is $\sigma(Y)$-measurable, which is the numerator on the RHS. Similarly is the denominator, thereby the $\mathbb{E}[g(X,Y)|Y]$. The partial average property is due to Fubini part (iii). □

This proposition is insightful. First, it verifies again $\mathbb{E}[g(X,Y)|Y]$ is indeed a function of Y. Secondly, since we would assume a density in almost every case in statistics, it gives out the explicit form of conditional expectation: when X (or ν) is continuous, we integrate; when discrete, we sum. In fact, though the rigorous proof was just introduced now, we have used this result when we deal with mixed type joint distribution, especially in the latent variable model.

Example 8 (EM for Gaussian mixture model)**.** Assume the latent class $Z \sim \text{Ber}(p)$, and $X|Z=0 \sim \text{N}(\mu_0, 1)$ while $X|Z=1 \sim \text{N}(\mu_1, 1)$. Though we have not defined the rigorous conditional distribution (will do later), intuitively the complete data $Y = (X, Z)$ has

$$\begin{aligned} f_Y(x,z) = f_{X|Z}(x|z) \cdot f_Z(z) &= \begin{cases} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{(x-\mu_0)^2}{2}\right\} \cdot (1-p) & \text{if } z=0, \\ \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{(x-\mu_1)^2}{2}\right\} \cdot p & \text{if } z=1. \end{cases} \\ &= \left\{\frac{1-p}{\sqrt{2\pi}} \exp\left[-\frac{(x-\mu_0)^2}{2}\right]\right\}^{\mathbb{1}_{\{z=0\}}} \times \left\{\frac{p}{\sqrt{2\pi}} \exp\left[-\frac{(x-\mu_1)^2}{2}\right]\right\}^{\mathbb{1}_{\{z=1\}}}. \end{aligned} \tag{2}$$

Hence the complete data log-likelihood is (the reason to compute this quantity will be discussed in EM)

$$\ell_Y(\theta) = -\log\sqrt{2\pi} + \mathbb{1}_{\{z=0\}}\left[\log(1-p) - \frac{(x-\mu_0)^2}{2}\right] + \mathbb{1}_{\{z=1\}}\left[\log(p) - \frac{(x-\mu_1)^2}{2}\right],$$

where $\theta = (\mu_0, \mu_1, p)$ denotes all parameters. Hence the function to be optimized $Q(\theta|\theta_t)$ is

$$\mathbb{E}_{\theta_t}[\ell_Y(\theta)|X] = -\log\sqrt{2\pi} + \left[\log(1-p) - \frac{(x-\mu_0)^2}{2}\right] \cdot \mathbb{P}_{\theta_t}(Z=0|X) + \left[\log p - \frac{(x-\mu_1)^2}{2}\right] \cdot \mathbb{P}_{\theta_t}(Z=1|X)$$

where θ_t is the computed parameters at step t.

Though (X, Z) has a mixed distribution, the joint density (2) on $\left(\mathbb{R}^2, \mathcal{B}_{\mathbb{R}^2}\right)$ is under $\lambda \times \#$ measure. By Proposition 31 and counting measure, we can compute

$$\mathbb{P}(Z=1|X) = \frac{\int \mathbb{1}_{\{z=1\}} f_Y(X,z)d\#(z)}{\int f_Y(X,z)d\#(z)} = \frac{pf_Y(X,1)}{(1-p)f_Y(X,0) + pf_Y(X,1)} = \frac{pe^{-\frac{(x-\mu_1)^2}{2}}}{(1-p)e^{-\frac{(x-\mu_0)^2}{2}} + pe^{-\frac{(x-\mu_1)^2}{2}}}.$$

In EM, don't forget to plug in the version of parameter θ_t, and proceed with the rest.

Interesting Properties

Now let's dive into two interesting while misleading results:

- $\mathbb{E}[g(X,Y)|Y=y]=\mathbb{E}[g(X,y)]$, and
- $\mathbb{E}[X|(Y,Z)]=\mathbb{E}[X|Y]$ when X and Z are independent.

The above two identities may appear sensible and correct at first glance, but **(important)** the above two identities are both wrong, or partly wrong! Introducing the modern version of conditional expectation allows us to show them.

What's wrong with $\mathbb{E}[g(X,Y)|Y=y]=\mathbb{E}[g(X,y)]$? Let's see the following example.

Example 9. Let (X,Y) be a bivariate normal with $\mu_X=\mu_Y=1$ and $\sigma_X^2=\sigma_Y^2=\rho=1$. For the function $g(x,y)=xy$ and a given $y\in\mathbb{R}$, it is easy to compute $\mathbb{E}[g(X,y)]=\mathbb{E}[Xy]=y\mathbb{E}X=0$. But on the other hand, we can compute $\mathbb{E}[g(X,Y)|Y=y]=y^2$ in two ways.

First, note that $\rho=1$ in bivariate normal means $X=Y$, hence $g(X,Y)=XY=Y^2$. Then we can get $\mathbb{E}[g(X,Y)|Y=y]=\mathbb{E}[Y^2|Y=y]=y^2$. Or we can first use $\mathbb{E}[XY|Y]=Y\mathbb{E}[X|Y]$, and hence $\mathbb{E}[XY|Y=y]=y\mathbb{E}[X|Y=y]$ as functions of y. Then compute by conditional distribution of multivariate normal that

$$\mathbb{E}[g(X,Y)|Y=y]=\mathbb{E}[XY|Y=y]=y\mathbb{E}[X|Y=y]=y\cdot\left[\mu_X+\rho\frac{\sigma_X}{\sigma_Y}(y-\mu_y)\right]=y^2.$$

But we can see $\mathbb{E}[g(X,y)]=0\neq y^2=\mathbb{E}[g(X,Y)|Y=y]$ necessarily. So this equality does not hold when there exists a correlation between X and Y like $\mathrm{Cov}(X,Y)=1$ in this example. A natural guess of independence leads us to the remedy of this result.

Proposition 32. *Let $X\in\mathbb{R}^m, Y\in\mathbb{R}^n$ be two independent random vectors and $g:\mathbb{R}^{m+n}\mapsto\mathbb{R}$ be Borel such that $\mathbb{E}|g(X,Y)|<\infty$. Then $\mathbb{E}[g(X,Y)|Y=y]=\mathbb{E}[g(X,y)]$ holds for μ_Y-almost everywhere $y\in\mathbb{R}^n$.*

Before proceeding with the proof, let's recall the meaning of both sides. By Proposition 30, the LHS $\mathbb{E}[g(X,Y)|Y]$ is some function of Y, say $h(Y)$. Then we define $\mathbb{E}[g(X,Y)|Y=y]=h(y)$. For the RHS, we fix a specific value $y\in\mathbb{R}^n$, making $g(\cdot,y):\mathbb{R}^m\to\mathbb{R}$ a Borel function. Thus, $g(X,y)$ becomes a random variable by Fubini 26 part (i), and we can compute its expectation as in the RHS. This explanation clarifies why the equality doesn't hold trivially, as the LHS and RHS are defined in entirely different ways.

Proof. Let $g(x,y)=\mathbb{1}_{B_1\times B_2}(x,y)$ be simple where $B_1\in\mathcal{B}_{\mathbb{R}^m}$ and $B_2\in\mathcal{B}_{\mathbb{R}^n}$. Then it follows that

$$\mathbb{E}[g(X,Y)|Y]=\mathbb{E}[\mathbb{1}_{B_1}(X)\mathbb{1}_{B_2}(Y)|Y]=\mathbb{1}_{B_2}(Y)\cdot\mathbb{P}(X\in B_1|Y)=\mathbb{1}_{B_2}(Y)\cdot\mathbb{P}(X\in B_1),$$

where the last equality depends on the independence. Hence $\mathbb{E}[g(X,Y)|Y=y]=\mathbb{1}_{B_2}(y)\cdot\mathbb{P}(X\in B_1)$ by definition. While the RHS $=\mathbb{E}[g(X,y)]=\mathbb{E}[\mathbb{1}_{B_1}(X)\cdot\mathbb{1}_{B_2}(y)]=\mathbb{1}_{B_2}(y)\cdot\mathbb{P}(X\in B_1)$. Hence LHS $=$ RHS when g is simple (here used $\mathcal{B}_{\mathbb{R}^{m+n}}=\mathcal{B}_{\mathbb{R}^m}\otimes\mathcal{B}_{\mathbb{R}^n}$). Then use MCT to complete the non-negative and general functions. □

Though the independence looks useless in the statement itself, when we read the proof, we immediately recognize it is necessary. Now, what's wrong with $\mathbb{E}[X|(Y,Z)]=\mathbb{E}[X|Y]$ when X and Z are independent?

Example 10. Let $U_1,\dots,U_4$ be iid standard normal rvs, set $X=U_1+U_2$, $Y=U_2+U_3$ and $Z=U_3+U_4$. Then by grouping of independent rvs, we know X,Z are independent. But using multivariate normal properties, we also know

$$\begin{pmatrix}X\\Y\\Z\end{pmatrix}\sim\mathbf{N}_3\left(\mathbf{0},\begin{bmatrix}2&1&0\\1&2&1\\0&1&2\end{bmatrix}\right)\quad\Rightarrow\quad\mathbb{E}(X|Y,Z)=\frac{2}{3}Y-\frac{1}{3}Z.$$

Hence, even though X and Z are independent, the quantity $\mathbb{E}(X|Y,Z)$ still depends on Z! The intuitive explanation is that Z can influence Y, which in turn affects X. But what if Z cannot influence Y, or if Z and Y are also independent? Exactly, this leads us to the following proposition.

Proposition 33. *Let $X \in L^1(\mathbb{P})$ be a rv and let $Y \in \mathbb{R}^m, Z \in \mathbb{R}^n$ be two random vectors. If (X,Y) and Z are independent, then $\mathbb{E}[X|(Y,Z)] = \mathbb{E}[X|Y]$ almost surely.*

Recall Definition 22 of independence for random elements, we say $(X,Y) \in \mathbb{R}^{m+1}$ and $Z \in \mathbb{R}^n$ are independent, if $\mathbb{P}\big\{(X,Y) \in B_1,\ Z \in B_2\big\} = \mathbb{P}\big\{(X,Y) \in B_1\big\}\mathbb{P}(Z \in B_2)$ for all $B_1 \in \mathcal{B}_{\mathbb{R}^{m+1}}$ and $B_2 \in \mathcal{B}_{\mathbb{R}^n}$.

Proof. The idea is to verify $\mathbb{E}[X|Y]$ is a version of $\mathbb{E}[X|(Y,Z)]$, and since the measurability is trivial by $\sigma(Y) \subset \sigma(Y,Z)$, it remains to show $\mathbb{E}[X|Y]$ satisfies the partial average property on $\sigma(Y,Z)$.

Let $B = B_1 \times B_2$ where $B_1 \in \mathcal{B}_{\mathbb{R}^m}$ and $B_2 \in \mathcal{B}_{\mathbb{R}^n}$. By the simple choice of B, we have

$$\int_{\{(Y,Z)\in B\}} \mathbb{E}[X|Y]d\mathbb{P} = \int_{\{(Y,Z)\in(B_1,B_2)\}} \mathbb{E}[X|Y]d\mathbb{P} = \int_\Omega \mathbb{E}[X|Y]\mathbb{1}_{\{Y\in B_1\}}\mathbb{1}_{\{Z\in B_2\}}d\mathbb{P} =$$

where the three terms are functions of (Y,Z), not X. By change of variable and independence,

$$= \int_{\mathbb{R}^2} \mathbb{E}[X|Y=y]\mathbb{1}_{\{y\in B_1\}}\mathbb{1}_{\{z\in B_2\}}d\mu_{Y,Z}(y,z) = \int_{\mathbb{R}} \mathbb{1}_{\{z\in B_2\}}d\mu_Z(z)\int_{\mathbb{R}} \mathbb{E}[X|Y=y]\mathbb{1}_{\{y\in B_1\}}d\mu_Y(y),$$

by change of variable for the second term and the average property of $\mathbb{E}[X|Y]$,

$$= \int_{\mathbb{R}} \mathbb{1}_{\{z\in B_2\}}d\mu_Z(z)\int_\Omega \mathbb{E}[X|Y]\mathbb{1}_{\{Y\in B_1\}}d\mathbb{P} = \int_{\mathbb{R}} \mathbb{1}_{\{z\in B_2\}}d\mu_Z(z)\int_\Omega X\mathbb{1}_{\{Y\in B_1\}}d\mathbb{P},$$

by change of variable back for the second term and (X,Y) independent of Z,

$$= \int_{\mathbb{R}} \mathbb{1}_{\{z\in B_2\}}d\mu_Z \int_{\mathbb{R}^2} x\mathbb{1}_{\{y\in B_1\}}d\mu_{X,Y}(x,y) = \int_{\mathbb{R}^3} x\mathbb{1}_{\{y\in B_1\}}\mathbb{1}_{\{z\in B_2\}}d\mu_{X,Y,Z}(x,y,z),$$

finally one more change of variable for (X,Y,Z),

$$= \int_\Omega X\mathbb{1}_{\{Y\in B_1, Z\in B_2\}}d\mathbb{P} = \int_{\{(Y,Z)\in B\}} Xd\mathbb{P},$$

completing partial average property for the simple case. For a general set $B \in \mathcal{B}_{\mathbb{R}^{m+n}}$, use the fact $\mathcal{B}_{\mathbb{R}^{m+n}} = \mathcal{B}_{\mathbb{R}^m} \otimes \mathcal{B}_{\mathbb{R}^n}$, good set principle ($\pi$-$\lambda$ theorem), and the MCT of conditional expectation. □

As the proof suggests, the result holds for all Borel transformation of X, say $h(X)$. So, $\mathbb{P}(A|Y,Z) = \mathbb{P}(A|Y)$ almost surely (both terms are rvs now) for all $A \in \sigma(X)$ when (X,Y) and Z are independent. This, as a by-product, suggests the notion of conditionally independent.

Definition 34 (Conditional independence)**.** Let X,Y,Z be three random elements. We say X and Z are conditionally independent given Y if $\mathbb{P}(A|Y,Z) = \mathbb{P}(A|Y)$ almost surely for all $A \in \sigma(X)$.

This definition is used sometimes when there is a latent categorical label behind the observed data (see my ST601 notes). From the construction, the $(X,Y) \perp\!\!\!\perp Z$ implies $X \perp\!\!\!\perp Z|Y$ (for short let me use $\perp\!\!\!\perp$ for independence, and | for given).

This ends our discussion on conditional expectation.

4.2 Regular Conditional Distribution

As the EM example 8 suggests, we are often interested in conditional structures, such as $X|Y = y \sim \mathrm{N}(\mu_y, 0)$ and $Y \sim \mathrm{Ber}(p)$. This raises a natural question: why not structure this section by first defining the conditional distribution of $X|Y = y$ and then use it to compute the conditional expectation $\mathbb{E}[X|Y = y]$? This approach aligns with elementary probability, yet in the modern formulation, we seem to reverse the order.

More generally, given two arbitrary random elements $X : (\Omega, \mathcal{A}) \to (\mathbb{S}_1, \mathcal{S}_1)$ and $Y : (\Omega, \mathcal{A}) \to (\mathbb{S}_2, \mathcal{S}_2)$, the joint distribution $\mu_{X,Y}$ always exists on the measurable space $(\mathbb{S}_1 \times \mathbb{S}_2, \mathcal{S}_1 \otimes \mathcal{S}_2)$. However, the conditional distribution of $X|Y = y$, as described earlier, does not always exist—at least, for now we don't know.

If we can establish its existence and identify the relationship between the conditional distribution and the joint distribution, we can confidently use the conditional structure to specify models without ambiguity. This is precisely the goal of this section.

Construction

Suggested by the conditional expectation, if we fix an $B \in \mathcal{B}_\mathbb{R}$, we may want define the conditional probability of A given $\mathcal{F}$ by $\mathbb{P}(B|\mathcal{F}) = \mathbb{E}[\mathbb{1}_B|\mathcal{F}]$. This notation looks well-defined and satisfies good defining properties of probability:

- $0 \le \mathbb{P}(B|\mathcal{F}) \le 1$ since $0 \le \mathbb{1}_A \le 1$ with $\mathbb{P}(\emptyset|\mathcal{F}) = 0$ and $\mathbb{P}(\Omega|\mathcal{F}) = 1$;
- let $(B_n)_{n=1}^\infty$ be disjoint sets, then $\mathbb{P}(\cup_{n=1}^\infty B_n|\mathcal{F}) = \mathbb{E}\left[\sum_{n=1}^\infty \mathbb{1}_{B_n}|\mathcal{F}\right] = \sum_{n=1}^\infty \mathbb{P}[B_n|\mathcal{F}]$ by MCT.

However there is a potential problem. Remember $\mathbb{P}(B|\mathcal{F})$ is an almost surely defined rv, and let's denote by N_B the set where $\mathbb{P}(B|\mathcal{F})$ is not defined. As a probability, we need $\mathbb{P}(\cdot|\mathcal{F})$ to be defined for all $B \in \mathcal{B}_\mathbb{R}$, but

$$\{\omega \in \Omega : \text{where } \mathbb{P}(\cdot \mid \mathcal{F})_{(\omega)} \text{ is not defined for some } B \in \mathcal{B}_\mathbb{R}\} = \bigcup_{B \in \mathcal{B}_\mathbb{R}} N_B,$$

and even though each $\mathbb{P}(N_B) = 0$, the uncountable union may not has 0 probability. Then our definition makes sense only on a set with probability

$$\mathbb{P}\Big\{\omega \in \Omega : \mathbb{P}(\cdot \mid \mathcal{F})_{(\omega)} \text{ is defined for all } B \in \mathcal{B}_\mathbb{R}\Big\} \neq 1 \text{ necessarily.}$$

So, if X is a rv and $\mathcal{F}$ is a sub σ-algebra, what do we want to solve? We want a function $g : \mathcal{B}_\mathbb{R} \times \Omega \mapsto [0, 1]$ such that when we fix $B \in \mathcal{B}_\mathbb{R}$, the $g(B, \cdot)$ is a version of $\mathbb{P}(X \in B|\mathcal{F})$, and at the same time, for ω in a probability 1 set, we need $g(\cdot, \omega)$ represents the conditional probability on $(\mathbb{R}, \mathcal{B}_\mathbb{R})$. And luckily, when the X is a rv or a random vector, the existence of such a function is guaranteed.

Theorem 35. *Let $X \in \mathbb{R}^n$ be a random vector and $\mathcal{F}$ be a sub σ-algebra on $(\Omega, \mathcal{A}, \mathbb{P})$. Then there exists a function $g : \mathcal{B}_{\mathbb{R}^n} \times \Omega \mapsto [0, 1]$ such that*

(i) fixing any $B \in \mathcal{B}_{\mathbb{R}^n}$, we have $g(B, \cdot)$ is a version of $\mathbb{P}(X \in B|\mathcal{F})$ and

(ii) fixing $\mathbb{P}$-almost any $\omega \in \Omega$, we have $g(\cdot, \omega)$ is a probability on $(\mathbb{R}^n, \mathcal{B}_{\mathbb{R}^n})$.

This function g is called the **regular conditional distribution** (rcd) of X given $\mathcal{F}$. Its existence is guaranteed in part due to the separability of $\mathbb{R}^n$, which allows us to appropriately choose $B \in \mathcal{B}_{\mathbb{R}^n}$ to cover all Borel sets effectively. For a detailed proof, refer to Durrett [1] or Billingsly [6].

As is often the case, we are particularly interested in situations where $\mathcal{F} = \sigma(Y)$, where $Y : (\Omega, \mathcal{A}) \to (\mathbb{S}, \mathcal{S})$ is a random element. In this context, the existence of a rcd in the general form allows us to derive a powerful result that resolves all the uncertainties we previously faced regarding conditional distributions.

Theorem 36. *Let $X \in \mathbb{R}^n$ be a random vector and $Y : (\Omega, \mathcal{A}) \mapsto (\mathbb{S}, \mathcal{S})$ be a random element. Then there exists a function $P_{X|Y}(B \mid y) : \mathbb{S} \times \mathcal{B}_{\mathbb{R}^n} \mapsto [0,1]$ such that*

(i) fixing any $B \in \mathcal{B}_{\mathbb{R}^n}$, we have $P_{X|Y}(B|y) = \mathbb{E}[X \in B|Y = y]$ for μ_Y-almost all $y \in \mathbb{S}$,

(ii) fixing μ_Y-almost any $y \in \mathbb{S}$, we have $P_{X|Y}(\cdot|y)$ is a probability on $(\mathbb{R}^n, \mathcal{B}_{\mathbb{R}^n})$ such that

$$\mathbb{E}[g(X,Y)|Y = y] = \int_{\mathbb{R}^n} g(x,y) P_{X|Y}(dx|y) \text{ for } \mu_Y\text{-almost all } y \in \mathbb{S}.$$

Proof. For any fixed $B \in \mathcal{B}_{\mathbb{R}^n}$, by Theorem 35, the rcd $g(\omega, B) = \mathbb{P}(X \in B|Y)$. Meanwhile, Proposition 30 says $\mathbb{P}(X \in B|Y)$ is some Borel function of Y. Hence, we can write $g(B, \omega) = h(B, Y\omega)$ where $h : \mathcal{B}_{\mathbb{R}^n} \times \mathbb{S} \mapsto [0,1]$ is some function depending on both B and ω, but on ω only through $Y\omega$.

Theorem 35 part (i) says for all $B \in \mathcal{B}_{\mathbb{R}^n}$, we have $h(B, Y\omega) = \mathbb{P}(X \in B|Y)$ for almost surely $\omega \in \Omega$. Hence by the meaning of given $Y = y$, we have $h(B, y) = \mathbb{P}(X \in B|Y = y)$ for μ_Y-almost all $y \in \mathbb{S}$. And $h(B, y)$ is a probability on $(\mathbb{R}^n, \mathcal{B}_{\mathbb{R}^n})$ for any $y \in \mathbb{S}$ directly follows from Theorem 35 part (ii).

To get the equality, let first take $g(x, y) = \mathbb{1}_{B_1 \times B_2}(x, y)$ where $B_1 \in \mathcal{B}_{\mathbb{R}^m}$ and $B_2 \in \mathcal{S}$. Then LHS

$$\mathbb{E}[g(X,Y)|Y = y] = \mathbb{E}\big[\mathbb{1}_{\{X \in B_1\}} \mathbb{1}_{\{Y \in B_2\}} \big| Y = y\big],$$

since $\mathbb{E}\big[\mathbb{1}_{\{X \in B_1\}} \mathbb{1}_{\{Y \in B_2\}} \big| Y\big] = \mathbb{1}_{\{Y \in B_2\}} \mathbb{E}\big[\mathbb{1}_{\{X \in B_1\}} \big| Y\big]$, by the meaning of given $Y = y$, we get

$$= \mathbb{1}_{\{y \in B_2\}} \mathbb{E}\big[\mathbb{1}_{\{X \in B_1\}} \big| Y = y\big] = \mathbb{1}_{\{y \in B_2\}} \mathbb{P}(X \in B_1 | Y = y) = \mathbb{1}_{\{y \in B_2\}} h(B_1, y),$$

note that y is some fixed number, hence $h(\cdot, y)$ is a probability on $(\mathbb{R}^n, \mathcal{B}_{\mathbb{R}^n})$. So we can write

$$= \mathbb{1}_{\{y \in B_2\}} \int_{\mathbb{R}^n} \mathbb{1}_{\{x \in B_2\}} h(dx, y) = \int_{\mathbb{R}^n} \mathbb{1}_{\{x \in B_2\}} \mathbb{1}_{\{y \in B_2\}} h(dx, y) = \int_{\mathbb{R}^n} g(x,y) h(dx, y).$$

Hence this function h serves as $P_{X|Y}$ and the equality holds when g is simple. Then we extend this result to general functions g by good set principle. □

Recall that we encountered a similar formula in Proposition 31:

$$\mathbb{E}[g(X,Y)|Y] = \frac{\int g(x,Y) f(x,Y) d\nu(x)}{\int f(x,Y) d\nu(x)},$$

when (X, Y) has a joint density $f(x, y)$ with respect to the product measure $\nu \times \lambda$. Notably, the proof of this result there did not involve any reference to the rcd. Now, combining these two formula, we can identify the Radon-Nikodym derivative of the conditional distribution $X|Y = y$ as follows:

$$\frac{dP_{X|Y}(\cdot \mid y)}{d\nu}(x) = \frac{f(x,y)}{\int f(x,y) d\nu(x)} \text{ aligns with } f_{X|Y}(x|y) = \frac{f(x,y)}{\int f(x,y) dx} \text{ in elementary probability!}$$

Furthermore, Theorem 36 part (ii) also verifies the conditioning way to compute the marginal $\mathbb{E}g(X,Y)$, as what we did in elementary probability, since

$$\mathbb{E}g(X,Y) = \mathbb{E}\big[\mathbb{E}[g(X,Y)|Y]\big] = \int_{\mathbb{S}} \mathbb{E}[g(X,Y)|Y = y] d\mu_Y(y) = \int_{\mathbb{S}} \left[\int_{\mathbb{R}^n} g(x,y) dP_{X|Y}(x|y)\right] d\mu_Y(y). \tag{3}$$

In fact, when there is a dependence structure, the model will directly be given in a conditional way. And this formula is how we compute expectation. And as as a by-product, we know that one conditioning $P_{X|Y}$ and one marginal μ_Y could determine the joint distribution $\mu_{X,Y}$ since taking $g(x, y) = \mathbb{1}_{B_1 \times B_2}(x, y)$ and

from (3)

$$\mathbb{P}(X \in B_1,\ Y \in B_2) = \int_{B_2} P_{X|Y}(B_1|y)d\mu_Y(y).$$

Remark. Why in (3) the integrand after the second equality is $\mathbb{E}[g(X,Y)|Y=y]$ but not $\mathbb{E}[g(X,y)|Y=y]$? Treat $g(X,Y)$ as Z. Then $\mathbb{E}(Z|Y)=h(Y)$ for some Borel h and $\mathbb{E}(Z|Y=y)=h(y)$ by the meaning of given $Y=y$. By change of variable only for Y, we have

$$\mathbb{E}\big[\mathbb{E}[g(X,Y)|Y]\big] = \mathbb{E}h(Y) = \int_{\mathbb{S}} h(y)d\mu_Y(y) = \int_{\mathbb{S}} \mathbb{E}(Z|Y=y)d\mu_Y(y) = \int_{\mathbb{S}} \mathbb{E}[g(X,Y)|Y=y]d\mu_Y(y).$$

We have actually posed this question in the last section: when does $\mathbb{E}[g(X,Y) \mid Y=y] = \mathbb{E}[g(X,y)]$ hold true? There we know the independence is necessary. And we can prove it in another way now. By Theorem 36, it holds without condition that

$$\mathbb{E}[g(X,Y)|Y=y] = \int_{\mathbb{R}^n} g(x,y)dP_{X|Y}(x|y),$$

when we further have independence, the probability $\mathbb{P}_{X|Y}(\cdot \mid y)$ should be the same across all $y \in \mathbb{S}$. Hence

$$= \int_{\mathbb{R}^n} g(x,y)d\mu_X(x) = \mathbb{E}[g(X,y)].$$

This should end our all confusions about conditional distribution, such as the one in the EM example 8 where we have a counting measure product a Lebesgue measure.

4.3 Computation by Conditioning

With the help of the conditioning techniques, especially Proposition 32 and the rcd property (3), we can go over some interesting examples. Let's start with an appetizer, and followed by the order statistics from the uniform distribution, and we end this topic with the random sum examples.

Question 37. Suppose (X,Y) are two random variables with the marginal distribution of X and conditional of $Y|X=x$ to be

$$F_X(x) = \begin{cases} 0, & x<0 \\ x, & 0 \le x \le 1 \\ 1, & x>1 \end{cases} \quad \text{and} \quad \begin{cases} P_{Y|X}\big(\emptyset|x\big) = 0, \\ P_{Y|X}\big(\{1\}|x\big) = x, \\ P_{Y|X}\big(\{0\}|x\big) = 1-x, \\ P_{Y|X}\big(\{0,1\}|x\big) = 1. \end{cases}$$

(i) How should we understand the whole model?

(ii) Find the joint distribution of (X,Y).

(iii) Find the marginal distribution of Y.

(iv) Find the rcd of $X|Y$.

Solution. (i) It obvious that $X \sim \mathrm{U}[0,1]$. Then given $X=x$, we can see $\mathbb{P}_{Y|X}(\cdot \mid x)$ is a counting measure on $\{0,1\}$. Hence it can be understood as $Y|X=x \sim \mathrm{Ber}(x)$.

(ii) How to find the joint distribution of (X,Y)? If this can be done, then we know everything about this model. The idea is to use the by-product of rcd $\mathbb{P}(Y \in B_1,\ X \in B_2) = \int_{B_2} P_{Y|X}(B_1|x)d\mu_X(x)$.

- When $B_2 \cap [0,1] = \emptyset$, we know immediately $\mathbb{P}(X \in B_1, Y \in B_2) = 0$.
- Otherwise, take $B_1 = (-\infty, \tilde{x}]$ where $0 < \tilde{x} < 1$ and $B_2 = \{1\}$. We get

$$\mathbb{P}(X \le \tilde{x}, Y = 1) = \int_{-\infty}^{\tilde{x}} P_{Y|X}(\{1\}|x)\mu_X(dx) = \int_0^{\tilde{x}} x dx = \frac{\tilde{x}^2}{2}.$$

  When $B_2 = \{0\}$, we get

$$\mathbb{P}(X \le \tilde{x}, Y = 0) = \int_{-\infty}^{\tilde{x}} P_{Y|X}(\{0\}|x)\mu_X(dx) = \int_0^{\tilde{x}} 1 - x \, dx = \tilde{x} - \frac{\tilde{x}^2}{2}.$$

  Note that $\{(-\infty, \tilde{x}] \times \{y\} : \tilde{x} \in (0,1) \text{ and } y \in \{0,1\}\}$ is enough to specify the whole $\mu_{X,Y}$

(iii) Then let's focus on the marginal distribution of Y. This is easy since we can take $\tilde{x} = 1$ or > 1 and get

$$\mathbb{P}(Y = 1) = \mathbb{P}(X \le \infty, Y = 1) = 1/2.$$

So, the marginal distribution of Y is Ber(1/2).

(iv) Finally, how to get the other way round conditional distribution $X|Y$? Specifically, we already know $\mu_{X,Y}$ and μ_Y in the previous steps, and we want to find $P_{X|Y}$ such that for all $B_1, B_2 \in \mathcal{B}_{\mathbb{R}}$,

$$\mathbb{P}(X \in B_1, Y \in B_2) = \int_{B_2} P_{X|Y}(B_1|y) d\mu_Y(y).$$

Smart choices of B_1 and B_2 yields the results. Since Y is discrete, we still discuss case by case. Starting with $Y = 1$, by the already calculated result, definition of $P_{X|Y}$ and the counting measure, we have for $0 < \tilde{x} < 1$,

$$\frac{\tilde{x}^2}{2} = \mathbb{P}(Y = 1, X \le \tilde{x}) = \int_{\{1\}} P_{X|Y}\left((0,\tilde{x}]\Big| y\right)\mu_Y(dy) = \frac{1}{2} P_{X|Y}\left((0,\tilde{x}]\Big| 1\right)$$

which leads to for all $0 < \tilde{x} < 1$,

$$P_{X|Y}\left((0,\tilde{x}]\Big| 1\right) = \tilde{x}^2 \quad \Rightarrow \quad X \mid Y = 1 \text{ has the cdf } F_{X|Y=1}(x) = \frac{x^2}{2}\mathbb{1}(0 \le x \le 1).$$

And a similar argument leads to $P_{X|Y}\left((0,\tilde{x}]\big| 0\right) = 2\tilde{x} - \tilde{x}^2$.

The following example is the conditional distribution of iid uniform rvs given the largest order statistics. The interesting thing is that we will see a jump in the distribution the end.

Question 38. Let $X_1, \ldots, X_n$ be iid copy from U(0, 1), and $X_{(1)} < \cdots < X_{(n)}$ be the order statistics. Find the conditional distribution of $X_1 \mid X_{(n)}$.

Solution. We first find the marginal of $X_{(n)}$. For all $\tilde{y} \in [0,1]$, we have

$$F_{X_{(n)}}(\tilde{y}) = \tilde{y}^n \text{ and } f_{X_{(n)}}(\tilde{y}) = n\tilde{y}^{n-1}.$$

The rcd of $X_1|X_{(n)}$ (for simplicity, we omit the subscript) $P : \mathcal{B}_{\mathbb{R}} \times \mathbb{R} \mapsto [0,1]$ satisfies for all $B_1, B_2 \in \mathcal{B}_{\mathbb{R}}$,

$$\mathbb{P}\left\{X_1 \in B_1, X_{(n)} \in B_2\right\} = \int_{B_2} P(B_1|y) d\mu_{X_{(n)}}(y).$$

To find this P, take $B_1 = (-\infty, \tilde{x}]$ and $B_2 = (-\infty, \tilde{y}]$. It must hold that (with abuse of notation)

$$\mathbb{P}(X_1 \le \tilde{x}, X_{(n)} \le \tilde{y}) = \int_0^{\tilde{y}} P\big(X_1 \le \tilde{x} | X_{(n)} = y\big) d\mu_{X_{(n)}}(y). \tag{4}$$

For any fixed $\tilde{y} \in [0,1]$, the $P\big(X_1 \le \tilde{x} | X_{(n)} = y\big)$ is enough to specify the distribution of $X_1 | X_{(n)} = \tilde{y}$. And the remaining job is to find it by smart choice of $\tilde{x}, \tilde{y}$ and plugging in (4)

Case 1: If $0 < \tilde{x} < \tilde{y} < 1$ where $\tilde{x}$ and $\tilde{y}$ are fixed.

The LHS of (4) can be directly computed by independence:

$$\mathbb{P}\left(X_1 \le \tilde{x}, X_{(n)} \le \tilde{y}\right) = \mathbb{P}(X_1 \le \tilde{x}, X_1 \le \tilde{y}, \dots X_n \le \tilde{y}) = \mathbb{P}(X_1 \le \tilde{x}) \cdots \mathbb{P}(X_n \le \tilde{y}) = \tilde{x}\tilde{y}^{n-1}.$$

And with the marginal of $X_{(n)}$, the RHS of (4) can be computed by:

$$\int_0^{\tilde{y}} P\left(X_1 \le \tilde{x} | X_{(n)} = \tilde{y}\right) d\mu_{X_{(n)}}(y) = \int_0^{\tilde{y}} P\left(X_1 \le \tilde{x} | X_{(n)} = \tilde{y}\right) \cdot n y^{n-1} dy.$$

Equating them and differentiating both sides, we get for all $0 < \tilde{x} < \tilde{y} < 1$,

$$P\left(X_1 \le \tilde{x} | X_{(n)} = \tilde{y}\right) = \left(1 - \frac{1}{n}\right) \frac{\tilde{x}}{\tilde{y}}.$$

Case 2: If $0 < \tilde{y} \le \tilde{x} < 1$ where $\tilde{x}$ and $\tilde{y}$ are fixed.

Due to now $\{X_1 \le \tilde{x}, X_{(n)} \le \tilde{y}\} \subset \{X_{(n)} \le \tilde{y}\}$, the LHS of (4) can be directly computed by independence:

$$\mathbb{P}(X_{(n)} \le \tilde{y}, X_1 \le \tilde{x}) = \mathbb{P}(X_1 \le \tilde{y}, \dots X_n \le \tilde{y}) = \mathbb{P}(X_1 \le \tilde{y}) \cdots \mathbb{P}(X_n \le \tilde{y}) = \tilde{y}^n.$$

And by the marginal of $X_{(n)}$, the RHS of (4) is:

$$\int_0^{\tilde{y}} P\left(X_1 \le \tilde{x} | X_{(n)} = \tilde{y}\right) \mu_{X_{(n)}}(dy) = \int_0^{\tilde{y}} P\left(X_1 \le \tilde{x} | X_{(n)} = \tilde{y}\right) \cdot n y^{n-1} dy.$$

Equating them and differentiating both sides, we get $P\left(X_1 \le \tilde{x} | X_{(n)} = \tilde{y}\right) = 1$ for all $0 < \tilde{y} \le \tilde{x} < 1$,

Case 3: If $\tilde{y} < 0$ or $\tilde{y} > 1$

Recall our goal is to find for any Borel sets $B_1 \in \mathcal{B}_{\mathbb{R}}$ and $B_2 \in \mathcal{B}_{\mathbb{R}}$,

$$\mathbb{P}(X_{(n)} \in B_1, X_1 \in B_2) = \int_{B_1} P(X_1 \in B_2 | X_{(n)} = \tilde{y}) d\mu_{X_{(n)}}(y).$$

When $B_1 \cap [0,1] = \emptyset$, the B_1 is out of the support of $X_{(n)}$, and hence the integral on the RHS is always 0 since $\mathbb{P}(X_{(n)} \in B_1) = 0$. So, we can define $P\left(X_1 \in B_2 | X_{(n)} = y\right)$ to be any value since it does not matter. And we will leave it here.

Remark. To see the meaning in the traditional way, let's fix $0 < \tilde{y} < 1$. Our answer says

$$\lim_{h \to \infty} \mathbb{P}\Big\{0 < X_1 \le \tilde{x} \mid X_{(n)} \in (\tilde{y} - h, \tilde{y} + h)\Big\} = P\left(X_1 \le \tilde{x} | X_{(n)} = \tilde{y}\right) = 1 = \begin{cases} \left(1 - \frac{1}{n}\right) \frac{\tilde{x}}{\tilde{y}} & \text{if } \tilde{x} < \tilde{y}, \\ 1 & \text{if } \tilde{x} \ge \tilde{y}. \end{cases}$$

Figure 3 is a graph for fixed $\tilde{y}$ and varying $\tilde{x} \in [0, \tilde{y}]$. When $\tilde{y}$ is fixed, we can compute

$$P\left(X_1 = \tilde{y} | X_{(n)} = \tilde{y}\right) = P\left(X_1 \le \tilde{y} | X_{(n)} = \tilde{y}\right) - \lim_{\tilde{x} \uparrow \tilde{y}} P\left(X_1 \le \tilde{x} | X_{(n)} = \tilde{y}\right) = 1 - \lim_{\tilde{x} \uparrow \tilde{y}} \left(1 - \frac{1}{n}\right) \frac{\tilde{x}}{\tilde{y}} = \frac{1}{n},$$

meaning that the conditional distribution has a jump with mass $1/n$. This is consistent since if we have n observations, X_1 should have probability $1/n$ to be the maximum.

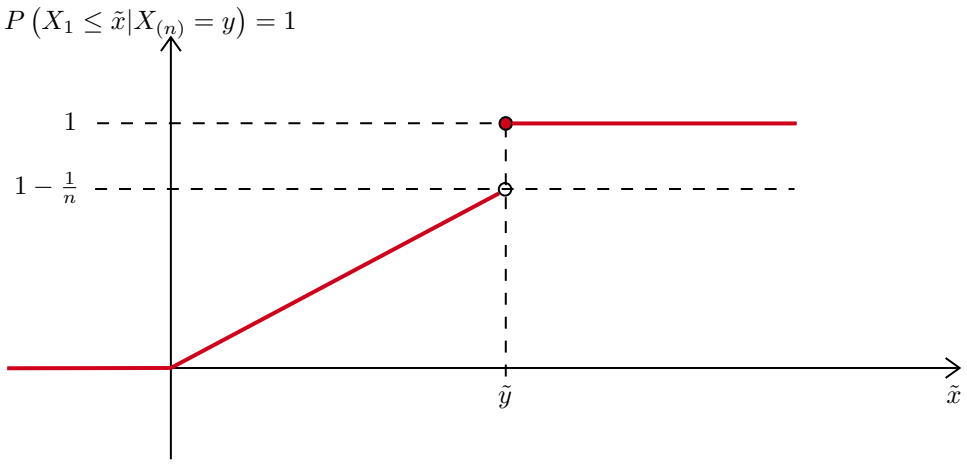

Figure 3: Conditional distribution of $X_1 \mid X_{(n)} = \tilde{y}$

Question 39. A Poisson process $(N_t)_{t\ge 0}$ with rate $\lambda > 0$ is defined, for all $t \ge 0$, via

$$N_t = \max\{n \ge 0 : T_n \le t\} = \sum_{n=1}^{\infty} \mathbb{1}(T_n \le t),$$

where $T_n = \tau_1 + \cdots + \tau_n$ with $\tau_i \overset{\text{iid}}{\sim} \text{Exp}(\lambda)$. For a fixed $t > 0$, find the distribution of N_t.

Solution. Fix $t > 0$. By the definition, we know N_t is an integer-valued rv, and the task is to find its mass function. By one important observation that

$$\{N_t = n\} = \left\{\sum_{i=1}^{\infty} \mathbb{1}(T_i \le t) = n\right\} = \Big\{T_n \le t, T_{n+1} > t\Big\},$$

it follows that $\mathbb{P}(N_t = n) = \mathbb{P}(T_n \le t, T_{n+1} > t)$, then

$$\mathbb{P}(N_t = n) = \mathbb{P}\Big\{T_n \le t,\ T_{n+1} > t\Big\} = \int_0^t P\left(T_{n+1} > t \mid T_n = s\right) d\mu_{T_n}(s)$$

and elementary conditioning helps us compute this conditional probability

$$P\left(T_{n+1} > t \mid T_n = s\right) = \int_t^{\infty} \frac{f_{T_{n+1},T_n}(u,s)}{f_{T_n}(s)} du.$$

Since $T_{n+1} = T_n + \tau_{n+1}$, by substitution formula and independence, we can get

$$f_{T_{n+1},T_n}(u,s) = f_{T_n}(s) \cdot f_{\tau_{n+1}}(u-s).$$

Hence we get

$$\mathbb{P}(N_t = n) = \int_0^t \int_t^{\infty} f_{\tau_{n+1}}(u-s) du\, d\mu_{T_n}(s) = \int_0^t \mathbb{P}(\tau_{n+1} > t - s) d\mu_{T_n}(s).$$

Plugging $T_n \sim \text{Gamma}(n, \lambda)$ and $\tau_{n+1} \sim \text{Exp}(\lambda)$, we get

$$= \int_0^t e^{-\lambda(t-s)} \frac{\lambda^n}{(n-1)!} s^{n-1} e^{-\lambda s} ds = e^{-\lambda t} \frac{(\lambda t)^n}{n!} = \mathbb{P}\Big\{\text{Pois}(\lambda t) = n\Big\}.$$

Let's end this topic with a hard question. It can be solved easily by computer simulation, but we will see to theoretically solve it, it takes some work.

Question 40. Let $(X_n)_{n=1}^{\infty}$ be iid U(0, 1) rvs, and set $S_n = X_1 + \cdots + X_n$ for all $n \in \mathbb{N}$. Moreover, let N be the first time S_n exceeds 1, or formally $N = \inf\{n \in \mathbb{N} : S_n \geq 1\}$. Find $\mathbb{E}N$.

Solution. To find $\mathbb{E}N$, we have no choice but to compute $\mathbb{P}(N = k)$ for all $k \in \mathbb{N}$. Notice $\mathbb{P}(N = 1) = \mathbb{P}(X_1 \geq 1) = 0$, so we can start with $k = 2$:

$$\mathbb{P}(N = 2) = \mathbb{P}(S_2 \geq 1, S_1 < 1) = \mathbb{P}(X_1 + X_2 \geq 1),$$

conditioning on $X_1 = x$ and using rcd property (3),

$$= \int_0^1 \mathbb{P}(X_1 + X_2 \geq 1 | X_1 = x) f_{X_1}(x) dx$$

by independence and Proposition 32, we can replace the random X_1 by x in the conditional probability,

$$= \int_0^1 \mathbb{P}(X_2 \geq 1 - x) f_{X_1}(x) dx = \int_0^1 x dx = \frac{1}{2}.$$

For $k = 3$, we can similarly compute

$$\mathbb{P}(N = 3) = \mathbb{P}(S_3 \geq 1, S_2 < 1) = \mathbb{P}(X_1 + X_2 + X_3 \geq 1, X_1 + X_2 < 1)$$

conditioning on $X_1 + X_2 =: Y = y \in [0, 2]$ and using rcd property (3),

$$= \int_0^2 \mathbb{P}(X_3 + Y \geq 1,\ Y < 1 | Y = y) f_Y(y) dy$$

by independence of Y and X_2, using Proposition 32

$$= \int_0^1 \mathbb{P}(X_3 \geq 1 - y,\ y < 1) f_Y(y) dy + \int_1^2 \mathbb{P}(X_3 \geq 1 - y,\ y < 1) f_Y(y) dy,$$

for the second term, the range $y \in [1, 2]$ and condition $y < 1$ forces the probability to be 0, hence we drop it,

$$= \int_0^1 \mathbb{P}(X_3 \geq 1 - y) f_Y(y) dy = \int_0^1 y f_Y(y) dy. \tag{5}$$

Now, we compute the density for $Y = X_1 + X_2$. Actually, we only need $f_Y(y)$ on $[0, 1]$ because we already saw in (5), we don't need $f_Y(y)$ on $[1, 2]$ (the result is more complicated for $y \in [1, 2]$, but won't be used in the future. That's why I ignored it here). Hence for $y \in [0, 1]$, using rcd (3) and Proposition 32,

$$\mathbb{P}(Y \leq y) = \mathbb{P}(X_1 + X_2 \leq y) = \int_0^1 \mathbb{P}(X_1 + X_2 \leq y | X_1 = x) f_{X_1}(x) dx = \int_0^1 \mathbb{P}(X_2 \leq y - x) dx$$

be careful that when $x \in [y, 1]$, the $\mathbb{P}(X_2 \le y - x)$ is 0. Hence only the $x \in [0, y]$ is integrating,

$$= \int_0^y (y - x)dx = \frac{y^2}{2}. \tag{6}$$

Hence $f_Y(y) = y$ for all $y \in [0, 1]$, and in (5),

$$\mathbb{P}(N = 3) = \int_0^1 y \cdot y dy = \frac{1}{1! \times 3}.$$

For $k = 4$, we can similarly compute

$$\mathbb{P}(N = 4) = \mathbb{P}(S_4 \ge 1, S_3 < 1) = \mathbb{P}(X_1 + X_2 + X_3 + X_4 \ge 1,\ X_1 + X_2 + X_3 < 1)$$

conditioning on $X_1 + X_2 + X_3 =: Y = y \in [0, 3]$ and using rcd property (3),

$$\begin{aligned}
&= \int_0^3 \mathbb{P}(X_4 + Y \ge 1,\ Y < 1 | Y = y) f_Y(y) dy \\
&= \int_0^1 \mathbb{P}(X_4 \ge 1 - y,\ y < 1) f_Y(y) dy + \int_1^3 \mathbb{P}(X_4 \ge 1 - y,\ y < 1) f_Y(y) dy \\
&= \int_0^1 \mathbb{P}(X_4 \ge 1 - y) f_Y(y) dy = \int_0^1 y f_Y(y) dy.
\end{aligned}$$

For density for $Y = X_1 + X_2 + X_3$ on $[0, 1]$, we use the exactly the same argument:

$$\mathbb{P}(Y \le y) = \mathbb{P}(X_1 + X_2 + X_3 \le y) = \int_0^1 \mathbb{P}(X_1 + X_2 \le y - x) f_{X_3}(x) dx = \int_0^y \mathbb{P}(X_1 + X_2 \le y - x) dx$$

noticing that the cdf of $X_1 + X_2$ is computed in (6), hence

$$= \int_0^y \frac{(y - x)^2}{2} dx = \frac{y^3}{3!}.$$

Hence $f_Y(y) = y^2/2!$ for all $y \in [0, 1]$, and

$$\mathbb{P}(N = 4) = \int_0^1 y \cdot \frac{y^2}{2!} dy = \frac{1}{2! \times 4}$$

By induction, we can get $\mathbb{P}(N = k) = \frac{1}{k \times (k-2)!}$. Hence finally,

$$\mathbb{E}N = \sum_{k=0}^{\infty} k \cdot \mathbb{P}(N = k) = 2 \cdot \frac{1}{2 \times 0!} + 3 \cdot \frac{1}{3 \times 1!} + \cdots = \sum_{k=0}^{\infty} \frac{1}{k!} = e.$$

References

[1] Rick Durrett. *Probability: theory and examples*, volume 49. Cambridge university press, 2019.

[2] Gerald B Folland. *Real analysis: modern techniques and their applications*, volume 40. John Wiley & Sons, 1999.

[3] Elias M Stein and Rami Shakarchi. *Real analysis: measure theory, integration, and Hilbert spaces*. Princeton University Press, 2009.

[4] Sidney I Resnick. *A probability path*. Springer Science & Business Media, 2013.

[5] Jun Shao. *Mathematical statistics*. Springer Science & Business Media, 2008.

[6] Gavin Brown. P. billingsley, probability and measure (wiley, 1979), pp. 532,£ 28· 95. *Proceedings of the Edinburgh Mathematical Society*, 26(3):398–399, 1983.