Statistical Theory
Preliminaries in Integration and Conditioning

By Zhu Xuelin
Email: xuelin@u.nus.edu

September 28, 2024

This is a summary of integration construction, Radon-Nikodym derivatives, change of variable formula, and informal but useful conditional distribution.

Contents

1 Integration

The completely analysis way of construction can be found in *Probability Theory and Examples* by Durrett[1], *Real Analysis*[2] by Folland or Stein[3]:

$$\text{Simple functions} \Rightarrow \text{Bounded functions} \Rightarrow \text{Non-negative functions} \Rightarrow \text{Measurable functions.}$$

We do not focus on the first way, but on a probabilistic way, skipping the second step:

$$\text{Simple functions} \Rightarrow \text{Non-negative functions} \Rightarrow \text{Measurable functions.}$$

The difference between the two ways is the usage of **supremum** and **infimum**. To end it, we will show they are equivalent by MCT in the end.

1.1 Simple Functions

To begin with, let's assume all functions discussed below are measurable from $(\Omega, \mathscr{A}, \mu)$ to $(\mathbb{R}, \mathcal{B}_{\mathbb{R}})$.

Definition 1.1 (Simple functions)**.** φ is said to be a simple function if $\varphi(\omega) = \sum_1^n a_i \mathbb{1}_{A_i}(\omega)$, where $a_i \in \mathbb{R}$ and $\mu(A_i) < \infty$. And we define the integral of φ to be

$$\int \varphi \, d\mu := \sum_{i=1}^{n} a_i \mu(A_i).$$

A simple function may take different forms, i.e. $\varphi = \sum_1^n a_i \mathbb{1}_{A_i} = \sum_1^n b_j \mathbb{1}_{B_j}$. If we further require $a_i \neq a_j$ and $\sqcup_1^n A_i = \Omega$, there exists only one representation, and we call it the canonical form. It can be shown that the defined integral does not depend on the representation (Details refer Stein's book).

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

Lemma 1.1. *Let φ and ψ be simple functions.*

(i) If $\varphi \geq 0$ a.e. then $\int \varphi \, d\mu \geq 0$.

(ii) For any $a \in \mathbb{R}$, $\int a\varphi \, d\mu = a \int \varphi \, d\mu$.

(iii) $\int \varphi + \psi \, d\mu = \int \varphi \, d\mu + \int \psi \, d\mu$.

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

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

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

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

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

$$\begin{aligned}\int \varphi + \psi \, d\mu &= \sum_{i=0}^{m}\sum_{j=0}^{n} (a_i + b_j)\, \mu\left(A_i \cap B_j\right) \\ &= \sum_{i=0}^{m}\sum_{j=0}^{n} a_i \mu\left(A_i \cap B_j\right) + \sum_{j=0}^{n}\sum_{i=0}^{m} b_j \mu\left(A_i \cap B_j\right) \\ &= \sum_{i=0}^{m} a_i \mu\left(A_i\right) + \sum_{j=0}^{n} b_j \mu\left(B_j\right) = \int \varphi \, d\mu + \int \psi \, d\mu.\end{aligned}$$

□

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

(iv) If $\varphi \le \psi$ a.e. then $\int \varphi \, d\mu \le \int \psi \, d\mu$.

(v) If $\varphi = \psi$ a.e. then $\int \varphi \, d\mu = \int \psi \, d\mu$.

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

(vi) $\left|\int \varphi \, d\mu\right| \le \int |\varphi| \, d\mu$.

Proof. For (iv), noting that $(\psi - \varphi) \ge 0$ is simple, by (i) we have $\int(\psi - \varphi) \, d\mu \ge 0$. Further with $\psi = \varphi + (\psi - \varphi)$, we have $\int \psi \, d\mu = \int \varphi \, d\mu + \int(\psi - \varphi) \, d\mu$ by (iii), concluding (iv). And (v) follows from two applications of (iv) in $\varphi \le \psi$ and $\psi \le \varphi$.

To prove (vi), note that $\varphi \le |\varphi|$, so (iv) implies $\int \varphi \, d\mu \le \int |\varphi| \, d\mu$. Also, with (ii), $-\varphi \le |\varphi| \Rightarrow -\int \varphi \, d\mu \le \int |\varphi| \, d\mu$, concluding (vi). □

1.2 Non-negative Functions

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

Theorem 1.3. *Suppose $f \ge 0$ is a measurable function. There exists a sequence of simple functions $\{\varphi_n\}$ s.t. $\varphi_n \uparrow f$ pointwisely.*

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

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

Definition 1.2. Let $f \ge 0$ be a measurable function. Define the integral of f to be

$$\int f \, d\mu := \lim_{n\to\infty} \int \varphi_n \, d\mu,$$

where $\{\varphi_n\}$ is a sequence of simple functions s.t. $\varphi_n \uparrow f$.

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 Theorem 2.1. We now prove the second and third ones.

Theorem 1.4. *Let $\varphi_n \geq 0$ and $\psi_n \geq 0$ be simple functions with $\varphi_n \uparrow f$ and $\psi_n \uparrow f$. Then*

(i) $\lim_{n\to\infty} \int \varphi_n \, d\mu$ *and* $\lim_{n\to\infty} \int \psi_n \, d\mu$ *exists.*

(ii) $\lim_{n\to\infty} \int \varphi_n \, d\mu = \lim_{n\to\infty} \int \psi_n \, d\mu$

Proof. Since φ_n are increasing simple functions, we have $\int \varphi_n \, d\mu$ increasing as a sequence of numbers, say c_n. 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 $n \in \mathbb{N}$:

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

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

$$\int \varphi_n \, d\mu \geq \int \varphi_n \mathbb{1}_{A_n} d\mu \geq \int t\psi_m \mathbb{1}_{A_n} d\mu = t \int \psi_m \mathbb{1}_{A_n} d\mu, \quad \forall n \in \mathbb{N}.$$

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

$$\lim_{n\to\infty} \int \varphi_n d\mu \geq t \cdot \lim_{n\to\infty} \int \psi_m \mathbb{1}_{A_n} d\mu = t \cdot \int \lim_{n\to\infty} (\psi_m \mathbb{1}_{A_n}) d\mu = t \cdot \int \psi_m d\mu, \tag{1}$$

where the second equality comes from the claim (we admit it here and prove in the end)

$$\lim_{n\to\infty} \int \psi_m \mathbb{1}_{A_n} d\mu = \int \lim_{n\to\infty} (\psi_m \mathbb{1}_{A_n}) d\mu. \tag{2}$$

Since (1) holds for all $0 < t < 1$ and $m \in \mathbb{N}$, taking $t \to 1$ and $m \to \infty$, we get

$$\lim_{n\to\infty} \int \varphi_n d\mu \geq \lim_{m\to\infty} \int \psi_m d\mu.$$

Switching the role of φ and ψ, we get our goal.

It now remains to prove the claim (2) (we need to do this without using any convergence theorem since we have not proved them). Note that ψ_m, by theorem condition, is a simple function. WLOG, suppose $\psi_m = \sum_1^k b_i \mathbb{1}_{B_i}$, with $b_i \in \mathbb{R}$ and $\mu(B_i) < \infty$. Then $\psi_m \mathbb{1}_{A_n} = \sum_1^k b_i \mathbb{1}_{B_i \cap A_n}$ is also simple, and

$$\int \psi_m \mathbb{1}_{A_n} d\mu = \sum_{i=1}^k b_i \mu(B_i \cap A_n).$$

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

$$\lim_{n\to\infty} \int \psi_m \mathbb{1}_{A_n} d\mu = \lim_{n\to\infty} \sum_{i=1}^{k} b_i \mu(B_i \cap A_n) = \sum_{i=1}^{k} b_i \lim_{n\to\infty} \mu(B_i \cap A_n) = \sum_{i=1}^{k} b_i \mu(B_i) = \int \psi_m d\mu.$$

□

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

Lemma 1.5. *Let f and g be non-negative measurable functions.*

(i) If $f \geq 0$ a.e. then $\int f \, d\mu \geq 0$.

(ii) For any $a > 0$, $\int af \, d\mu = a \int f \, d\mu$.

(iii) $\int f + g \, d\mu = \int f \, d\mu + \int g \, d\mu$.

Proof. For (i), since $f \geq 0$ a.e. by Theorem 2.1, we can choose simple functions $\varphi_n \geq 0$ s.t. $\varphi \uparrow f$ a.e. Then by the properties of simple functions integral $\int \varphi_n \geq 0$ for all $n \in \mathbb{N}$, and $\int f d\mu = \lim_{n\to\infty} \int \varphi_n d\mu \geq 0$. (ii) can be shown using the same idea.

For (iii), suppose simple functions $\varphi_n \uparrow f$ and $\psi_n \uparrow g$. Check that $\{\varphi_n + \psi_n\} \uparrow (f + g)$ and are also simple. Therefore, by definitions and using the properties of simple function integrals,

$$\int f + g \, d\mu = \lim_{n\to\infty} \int (\varphi_n + \psi_n) d\mu = \lim_{n\to\infty} \left(\int \varphi_n d\mu + \int \psi_n d\mu \right) = \int f d\mu + \int g d\mu.$$

□

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.

Proposition 1.6. *Let $f \geq 0$ be a measurable function. Then*

$$\int f d\mu = \sup \left\{ \int \varphi d\mu : \varphi \text{ is simple and } 0 \leq \psi \leq f \right\}.$$

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 1.4. 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 functions $\varphi_n \uparrow f$. Since each φ_n is an element of the set on RHS, we have (sup is an upper bound)

$$\int \varphi_n d\mu \leq \sup \left\{ \int \varphi d\mu : \varphi \text{ is simple and } 0 \leq \varphi \leq f \right\}, \quad \forall n \in \mathbb{N}.$$

Letting $n \to \infty$, we get

$$\int f d\mu = \lim_{n\to\infty} \int \varphi_n d\mu \le \sup\left\{ \int \varphi d\mu : \varphi \text{ is simple and } 0 \le \varphi \le f \right\}.$$

The other direction is tricky. If we show:

$$\int f d\mu \ge \int \varphi d\mu \text{ for every } \varphi \text{ which is simple and } 0 \le \varphi \le f,$$

then $\int f d\mu$ is no less than every elements in that set, meaning

$$\int f d\mu \ge \sup\left\{ \int \varphi d\mu : \varphi \text{ is simple and } 0 \le \varphi \le f \right\}.$$

So we can fix a simple function φ with $0 \le \varphi \le f$, and the goal is to show $\int f d\mu \ge \int \varphi d\mu$.

Note that now φ is fixed and $\varphi \le f$. So there exists, by approximation theorem 2.1, a sequence of simple functions $0 \le \varphi_n \uparrow (f - \varphi)$ and hence $\int (f-\varphi) d\mu = \lim_{n\to\infty} \int \varphi_n d\mu$. By properties of integral, we can add $\int \varphi d\mu$ to both sides and get

$$\int f d\mu = \int (f-\varphi) d\mu + \int \varphi d\mu = \lim_{n\to\infty} \int \varphi_n d\mu + \int \varphi d\mu \ge \int \varphi d\mu,$$

where the last inequality comes from the fact $\varphi_n \ge 0$. □

1.3 General Measurable Functions

For a general measurable function f, we can define its positive and negative parts by

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

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

Definition 1.3. Let f be measurable. We say f is integrable if $\int f^+ \, d\mu < \infty$ and $\int f^- \, d\mu < \infty$. And define its integral by $\int f \, d\mu := \int f^+ \, d\mu - \int f^- \, d\mu$

We also write $f \in L_1(\Omega, \mathcal{A}, \mu)$ to indicate f is an integrable function on Ω with respect to the measure μ. 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.

We end the 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 1.7. *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. 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.

1.4 Convergence Theorems

There are three theorems allowing us to switch the order of a limit and integration. We start from the first one, the Monotone Convergence Theorem.

Theorem 1.8 (MCT). *If $f_n \geq 0$ and $f_n \uparrow f$, then $\int f_n d\mu \uparrow \int f d\mu$.*

Proof. We will show two directions:

$$\lim_{n\to\infty} \int f_n d\mu \leq \int f d\mu \quad \text{and} \quad \lim_{n\to\infty} \int f_n d\mu \geq \int f d\mu.$$

The first one follows from letting $n \to \infty$ in $\int f_n d\mu \leq \int f d\mu$, which holds 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 $\lim_{n\to\infty} \int f_n d\mu \geq \int \varphi d\mu$ for every simple function φ s.t. $0 \leq \varphi \leq f$, then by Proposition 2.5, we are done. Now fix a φ satisfying that condition, and fix a proportion $0 < t < 1$. For each $n \in \mathbb{N}$, we define $A_n := \{f_n \geq t\varphi\}$. And we can

check $A_n \uparrow \Omega$ since $t < 1$. It follows that

$$\int f_n d\mu \geq \int f_n \mathbb{1}_{A_n} d\mu \geq t \int \varphi \mathbb{1}_{A_n} d\mu, \quad \forall n \in \mathbb{N}.$$

Taking limits on both sides, and use claim (2) proved before, we get

$$\lim_{n\to\infty} \int f_n d\mu \geq t \cdot \lim_{n\to\infty} \int \varphi \mathbb{1}_{A_n} d\mu = t \cdot \int \varphi \lim_{n\to\infty} \mathbb{1}_{A_n} d\mu = t \cdot \int \varphi d\mu.$$

Letting $t \to 1$, we conclude MCT. □

If we do not have increasing property, the pointwise limit for an arbitrary sequence $\{f_n\}$ may not exists. However, the $\limsup_n f_n(\omega)$ and $\liminf_n f_n(\omega)$ always exist for all $\omega \in \Omega$. And Fatou's Lemma guarantees the order switch in the following way.

Theorem 1.9 (Fatou)**.** *Let $f_n \geq 0$ for all $n \in \mathbb{N}$. Then $\int \left(\liminf_{n\to\infty} f_n\right) d\mu \leq \liminf_{n\to\infty} \int f_n d\mu$.*

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

$$g_n \uparrow \liminf_{n\to\infty} f_n \quad \overset{\text{MCT}}{\Longrightarrow} \quad \lim_{n\to\infty} \int g_n d\mu = \int \liminf_{n\to\infty} f_n.$$

We further notice that $g_n = \inf_{k\geq n} f_k \leq f_n$ for all $n \in \mathbb{N}$. By the integral properties, it follows that $\int g_n d\mu \leq \int f_n d\mu$ as a sequence of number for all $n \in \mathbb{N}$, and

$$\lim_{n\to\infty} \int g_n d\mu = \liminf_{n\to\infty} \int g_n d\mu \leq \liminf_{n\to\infty} \int f_n d\mu.$$

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 1.10 (DCT)**.** *Let f_n be a sequence of functions s.t. $|f_n| \leq g$ for some $g \in L_1(\mu)$. If $f_n \to f$ a.e., then $f \in L_1(\mu)$ and $\lim_{n\to\infty} \int f_n d\mu$ exists with*

$$\lim_{n\to\infty} \int f_n d\mu = \int f d\mu.$$

Proof. With $|f_n| \leq g$, we know

$$|f| = \lim_{n\to\infty} |f_n| \leq g \quad \Rightarrow \quad \int |f| d\mu \leq \int g d\mu < \infty \quad \Rightarrow \quad f \in L_1(\mu).$$

Further with $g - f_n \geq 0$ and $(g - f_n) \to (g - f) \geq 0$, by Fatou's lemma,

$$\int g d\mu - \int f d\mu = \int (g - f) d\mu = \int \liminf_{n\to\infty} (g - f_n) d\mu \leq \liminf_{n\to\infty} \int (g - f_n) d\mu = \int g d\mu - \limsup_{n\to\infty} f_n d\mu,$$

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

$$\limsup_{n\to\infty} f_n d\mu \le \int f d\mu.$$

Using $g + f_n \ge 0$ and $(g + f_n) \to (g + f) \ge 0$, by Fatou's lemma,

$$\int g d\mu + \int f d\mu = \int (g + f) d\mu = \int \liminf_{n\to\infty} (g + f_n) d\mu \le \liminf_{n\to\infty} \int (g + f_n) d\mu = \int g d\mu + \liminf_{n\to\infty} f_n d\mu.$$

Combining two inequalities, we get

$$\limsup_{n\to\infty} f_n d\mu \le \int f d\mu \le \liminf_{n\to\infty} f_n d\mu.$$

So, all inequalities above should be equality. □

1.4.1 Why the DCT may Fail?

In essence, convergence theorems answer the question: when lim can be switched with integral. By DCT, the answer is affirmative when there exists a dominating function $g \ge |f_n|$. 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. Consider the function $f_n(x) = \mathbb{1}_{(n,n+1]}(x)$, 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. Consider the function $f_n = n\mathbb{1}_{(0,1/n]}$ with $\lim_{n\to\infty} f \equiv 0$, but

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

This happens because at $x = 0$, the function blows up. And if we want to find a smallest function g that dominates f_n, the choice must be $g = \mathbb{1}_{(1,\infty)}$, which is not integrable.

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

2 Radon-Nikodym Derivatives

This concept is useful when we reach the topic about the **general density** of a distribution or **conditional expectation**. Let's start with the definition of absolutely continuity.

Definition 2.1. Let $(\Omega, \mathcal{A})$ be a measurable space, and μ, ν be two measures on it. ν is said to be absolutely continuous (a.c.) with respect to μ, written $\nu \ll \mu$, 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 gives the converse proposition.

Theorem 2.1 (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.

We have used this theorem many times without knowing it. It is hidden behind the change of variable when we do integration, since we can prove a simple proposition beyond Theorem 2.2:

Proposition 2.2. *For all function $g \in L_1(\Omega, \mathcal{A}, \nu)$, we have*

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

The idea of proof is starting from indicator function to all L_1 functions. To be specific, we can consider this example:

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

3 Computing Expectation

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

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

No matter what $(\mathcal{S}, S)$ is, we can define a measure on $(\mathcal{S}, S)$ by $\mathbb{P}_X(S) := \mathbb{P}(X \in S)$ for all $A \in \mathcal{S}$, called induced measure by X, or distribution of X (ex. check this is a measure).

Now, suppose there is another measurable mapping $g : (\mathcal{S}, S) \mapsto (\mathbb{R}, \mathcal{B}_{\mathbb{R}})$. 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 $\mathbb{P}_Y(B) := \mathbb{P}(Y \in B)$ for all $B \in \mathcal{B}_{\mathbb{R}}$.

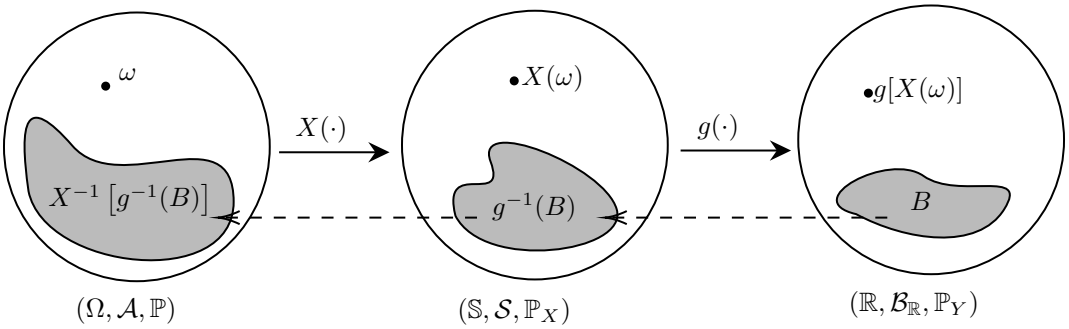

Figure 1: Measurable transformations

Here is the most important tool for statisticians to compute expectation, even without touching the underlying probability space.

Theorem 3.1 (Change of variable formula)**.** *Let X be a random element from $(\Omega, \mathcal{A}, \mathbb{P})$ to $(\mathcal{S}, S)$, and f be a measurable function from $(\mathcal{S}, S)$ to $(\mathbb{R}, \mathcal{B}_{\mathbb{R}})$ with $f \geq 0$ or $\mathbb{E}|f(X)| < \infty$. Then*

$$\mathbb{E}f(X) = \int_{\Omega} f(X(\omega))\mathbb{P}(d\omega) = \int_{\mathbb{S}} f(x)\mathbb{P}_X(dx) = \int_{\mathbb{R}} y\, \mathbb{P}_Y(dy),$$

where $\mathbb{P}_Y$ is the measure induced by $Y := f(X)$.

The proof can be found in *Probability Theory and Examples* by Durrett or Jing's manuscript. And this theorem states three ways to compute the expected value:

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

Most of the time, we are assuming X_i's are iid from a certain distribution, which is given in parametric models, though parameters may be unknown, and no one ever touches the underlying probability space. So, this is also called the theorem of unconscious statistician. And when we are giving the distribution of X and want to compute the expected value of $Y = f(X)$, we need not compute the distribution of Y anymore.

In the rest of this section, we will derive the specific formula for continuous and discrete r.v.s, and see the consistency with baby probability theory.

3.1 Absolutely Continuous Random Variables

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

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

By Proposition 2.2, we have for any function g s.t. $\mathbb{E}|g(X)| < \infty$,

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

3.2 Discrete Random Variables

3.2.1 Integration with respect to Counting Measures

To deal with it, we first need to derive the integration formula w.r.t. a **counting measure** μ, i.e. $\mu = 1$ on a countable subset $C = \{c_1, c_2, \dots\}$ and $\mu = 0$ everywhere else. First, we assume $f \geq 0$. Define $\forall k \in \mathbb{N}$,

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

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,

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

And similarly, for any $B \subset \mathbb{R}$, we have

$$\int_B f d\mu = \int_B \tilde{f} d\mu = \sum_{i:c_i \in B}^{\infty} f(c_i)$$

3.2.2 Expected Values of Discrete Random Variables

Now, 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 (ex. check this is a measure)

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

Then we claim $\mathbb{P}_X \ll \mu$, 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 $\mathbb{P}_X(N) = \mathbb{P}(X \in N) \leq \mathbb{P}(X \in C^c) = 0$. □

Remark. This proposition indicates that if X is discrete, then the induced measure space $(\mathbb{R}, \mathcal{B}_{\mathbb{R}}, \mathbb{P}_X)$ 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

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

where the last equality comes from Section 3.2.1. 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 f s.t. $\mathbb{E}|f(X)| < \infty$, change of variable formula, Proposition 2.3 and counting measure integral give

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

4 Conditional Distributions

4.1 Recaps of Elementary Definition

4.1.1 Discrete Random Variables

Suppose $(X, Y) \in \mathbb{R}^2$. In baby probability, we first defined the conditional distributions when X is discrete (Y can be any type), by setting the conditional mass function

$$\mathbb{P}_{Y|X=x}(B) = \mathbb{P}(Y \in B \mid X = x) = \frac{\mathbb{P}(X = x, Y \in B)}{\mathbb{P}(X = x)}, \quad \forall x \in \mathbb{R} \text{ such that } \mathbb{P}_X(x) > 0.$$

And leave the points with $\mathbb{P}_X(x) = 0$ undefined. One can check that $\mathbb{P}_{Y|X=x}$ is a probability measure for every x which is defined as exercise.

4.1.2 Continuous Random Variables

When X, Y are absolutely continuous, the above definition becomes pathological since $\mathbb{P}_X(X = x) = 0$ for all $x \in \mathbb{R}$. But we can fix it using the idea of limit. Define a new function on the semi-ring $\mathcal{S} = \{(a, b] : -\infty < a < b \leq \infty\}$ by

$$g_x((a, b]) = \lim_{h \to 0^+} \mathbb{P}\big[Y \in (a, b] \mid X \in (x - h, x + h)\big] = \lim_{h \to 0^+} \frac{\mathbb{P}\big[Y \in (a, b],\ X \in (x - h, x + h)\big]}{\mathbb{P}\big[X \in (x - h, x + h)\big]}.$$

Intuitively, the function g_x is computing the limiting value for the proportion in red box of Figure 2 as the bandwidth $h \to 0$.

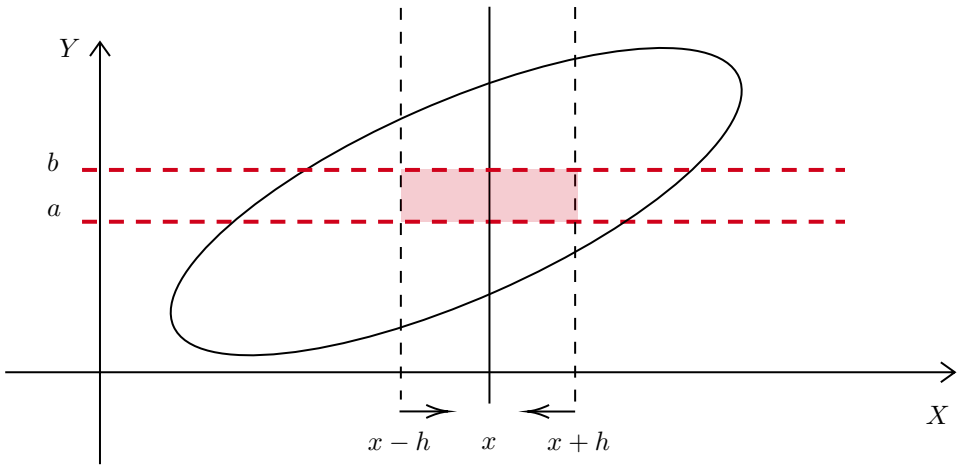

Figure 2: Illustration of conditional distribution

Thanks to the existence of the density of X, this function can be simplified explicitly.

$$g_x((a, b]) = \lim_{h \to 0^+} \frac{\mathbb{P}\big[Y \in (a, b],\ X \in (x - h, x + h)\big]}{\mathbb{P}\big[X \in (x - h, x + h)\big]} = \lim_{h \to 0^+} \frac{\int_{x-h}^{x+h} \int_a^b f(u, v) dv du}{\int_{x-h}^{x+h} f_X(u) du} = \frac{\int_a^b f(x, v) dv}{f_X(x)},$$

where the last equality follows from Lebesgue differentiation theorem. One can check that this function g_x is a probability on the semi-ring $\mathcal{S}$ for all x s.t. $f_X(x) > 0$. So, it uniquely determines a probability on $(\mathbb{R}, \mathcal{B}_{\mathbb{R}})$, which is called the **conditional distribution** of $Y \mid X = x$.

We further note that the measure $g_x \ll \lambda$, meaning that it has a unique density, normally written $f_{Y|X=x}(y)$, such that,

$$g_x(B) = \int_B f_{Y|X=x}(y)dy, \quad \forall B \in \mathcal{B}_{\mathbb{R}}.$$

And we call it the conditional density of $Y \mid X$. Indeed, this density can be computed explicitly by taking $B = (y-r, y+r]$ and Lebesgue differentiation theorem,

$$f_{Y|X=x}(y) = \lim_{r\to 0} g_x\big\{(y-r, y+r]\big\} = \lim_{r\to 0} \frac{\int_{y-r}^{y+r} f(x,v)dv}{f_X(x)} = \frac{f(x,y)}{f_X(x)}.$$

4.2 Formal Definition for Conditional Distribution

Now we want to define conditional distribution in a more general way. To do this, we need a new concept:

Definition 4.1 (Transition function)**.** Let $(\Omega_1, \mathcal{B}_1)$ and $(\Omega_2, \mathcal{B}_2)$ be two measure spaces, and $(\Omega_1 \times \Omega_2, \mathcal{B}_1 \times \mathcal{B}_2)$ be the product space. A function

$$K(\omega_1, B_2) : \Omega_1 \times \mathcal{B}_2 \mapsto [0,1]$$

is called a transition function if

(i) for each $\omega_1 \in \Omega_1$, $K(\omega_1, \cdot)$ is a probability on $(\Omega_2, \mathcal{B}_2)$,

(ii) for each $B_2 \in \mathcal{B}_2$, $K(\cdot, B_2)$ is $\mathcal{B}_1/\mathcal{B}_{[0,1]}$ measurable.

Intuitively, $K(\omega_1, \cdot)$ can be viewed as a distribution on $(\Omega_2, \mathcal{B}_2)$, which may change as ω_1 varies. We will borrow this idea to define conditional distribution.

From the probability view, if the random variables X, Y are given, we can derive their induced joint distribution $\mathbb{P}_{X,Y}$ on $(\mathbb{R}^2, \mathcal{B}_{\mathbb{R}^2})$ and the marginal $\mathbb{P}_X$ on $(\mathbb{R}, \mathcal{B}_{\mathbb{R}})$. Now we define the conditional distribution of $Y \mid X$.

Definition 4.2 (Conditional distribution)**.** Let $X : \Omega \mapsto \mathbb{R}^n$ and $Y : \Omega :\mapsto \mathbb{R}^m$ be two random vectors on the probability space $(\Omega, \mathcal{A}, \mathbb{P})$. A function $G(x, B) : \mathbb{R}^n \times \mathcal{B}_{\mathbb{R}^m} \mapsto [0,1]$ is called the conditional distribution for Y given X, written

$$Y \mid X = x \sim G(x, \cdot),$$

if G is a transition function, i.e.

(i) for each $x \in \mathbb{R}$, $G(x, \cdot)$ is a probability on $(\mathbb{R}^m, \mathcal{B}_{\mathbb{R}^m})$,

(ii) for each $B \in \mathcal{B}_{\mathbb{R}^m}$, $G(\cdot, B)$ is $\mathcal{B}_{\mathbb{R}^n}/\mathcal{B}_{[0,1]}$ measurable,

and furthermore,

(iii) for any Borel sets $B_x \in \mathcal{B}_{\mathbb{R}^n}$ and $B_y \in \mathcal{B}_{\mathbb{R}^m}$,

$$\mathbb{P}(X \in B_x, Y \in B_y) = \int_{B_x} G(x, B_y)\mathbb{P}_X(dx),$$

where $\mathbb{P}_X(\cdot)$ is the distribution of X on $(\mathbb{R}^n, \mathcal{B}_{\mathbb{R}^m})$.

Remark. Here are two things we need to be careful:

(i) Since this is a definition, we need to ask: (i) does this function G exist? (ii) if so, is it unique? Fortunately, the existence and uniqueness are both guaranteed in more advanced theory, cf. Theorem 33.3 by Billingsley (2005). We only need this definition and know the conditional distribution exists. That's enough.

(ii) Since $\mathbb{P}_{X,Y}$ is a measure on $(\mathbb{R}^{m+n}, \mathcal{B}_{\mathbb{R}^{m+n}})$, it suffices to specify $\{(-\infty, x] \times (-\infty, y], x \in \mathbb{R}^n, y \in \mathbb{R}^m\}$, which is more convenient for computation.

(iii) For convenience, one can understand the function G as (informal notation)

$$G(x, B) = \mathbb{P}(Y \in B | X = x) = \mathbb{P}_{Y|X=x}(B).$$

And actually, this notation makes sense since if we consider for a fixed $B \in \mathcal{B}_{\mathbb{R}^m}$, the formal definition of conditional expectation says $\mathbb{E}\big[\mathbb{1}_{Y^{-1}(B)}\big|X\big]$ is $\sigma(X)$-measurable. Hence there exists a Borel function, say

$$h : (\mathbb{R}^n, \mathcal{B}_{\mathbb{R}^n}) \mapsto (\mathbb{R}, \mathcal{B}_{\mathbb{R}}) \text{ such that } \mathbb{P}(Y \in B|X) = \mathbb{E}\big[\mathbb{1}_{Y^{-1}(B)}\big|X\big] = h(X).$$

Then we can define $\mathbb{P}(Y \in B|X = x) = h(x)$. Note that this function h could only be defined on the range of X. Using this notation, everything is well-defined, and in fact $G(x, B) = \mathbb{P}(Y \in B|X = x)$. For this alternative defining way, confer Shao Jun [4].

4.3 Examples of Conditional Distribution

Let's see an example.

Example. Suppose (X, Y) are two random variables with the marginal distribution of X and conditional of $Y \mid X = x$ to be

$$F_X(x) = \begin{cases} 0, & x < 0 \\ x, & 0 \le x \le 1 \\ 1, & x > 1 \end{cases} \quad \text{and} \quad \begin{cases} G\Big(x, \emptyset\Big) = 0, \\ G\Big(x, \{1\}\Big) = x, \\ G\Big(x, \{0\}\Big) = 1 - x, \\ G\Big(x, \{0, 1\}\Big) = 1. \end{cases}$$

Find the traditional meaning of the conditional $Y \mid X = x$.

Solution. For any $0 < x < 1$, we can easily check $G(x, \cdot)$ is a Borel measure on $(\mathbb{R}, \mathcal{B}_{\mathbb{R}})$ which is dominated by the counting measure on $\{0, 1\}$. Our goal is to find what does it really means. By Definition 4.2 (iii), for a fixed $x \in [0, 1]$, if we take $B_y = \{1\}$ and $B_x = (x - h, x + h) \subset [0, 1]$, we will see

$$\mathbb{P}\left(x - h < X < x + h, Y = 1\right) = \int_{x-h}^{x+h} G\left(\tilde{x}, \{1\}\right) \mathbb{P}_X(d\tilde{x}) = \int_{x-h}^{x+h} \tilde{x} \mathbb{P}_X(d\tilde{x}) = 2hx.$$

And similarly, if we take $B_y = \{0\}$ and $B_x = (x-h, x+h) \subset [0,1]$, we will see

$$\mathbb{P}\left(x-h < X < x+h, Y = 0\right) = \int_{x-h}^{x+h} G\left(\tilde{x}, \{0\}\right) \mathbb{P}_X(d\tilde{x}) = \int_{x-h}^{x+h} 1 - \tilde{x}\, \mathbb{P}_X(d\tilde{x}) = 2h - 2hx.$$

Now by basic probability, we can compute by plugging in

$$\mathbb{P}(Y = 1 \mid x-h < X < x+h) = \frac{\mathbb{P}(x-h < X < x+h, Y = 1)}{\mathbb{P}(x-h < X < x+h)} = \frac{2hx}{(2hx) + (2h - 2hx)} = x.$$

Taking the limit, we get

$$\lim_{h \to 0} \mathbb{P}(Y = 1 \mid x-h < X < x+h) = x,$$

which may be hard to handle if we are limited to the traditional definition, since Y is discrete and X is continuous. This is the advantage to bring in the definition of G.

4.3.1 From $Y \mid X$ and X to (X, Y)

For now, let's turn to another topic: if we are only given the marginal distribution of X and conditional distribution of $Y \mid X$, can we get the joint distribution? The answer is affirmative. Just like $(X, Y) = (Y \mid X) \times X$ in baby probability, if we know the marginal distribution of X and the conditional distribution of $Y \mid X$, then joint distribution is uniquely determined from a theoretical view.

Theorem 4.1. *Let $\mathbb{P}_X$ be the distribution of X on $(\mathbb{R}^n, \mathcal{B}_{\mathbb{R}^n})$, and suppose the conditional distribution of $Y \mid X = x$ on $(\mathbb{R}^m, \mathcal{B}_{\mathbb{R}^m})$ is*

$$G(x, B) : \mathbb{R}^n \times \mathcal{B}_{\mathbb{R}^m} \mapsto [0,1].$$

Then G and $\mathbb{P}_X$ uniquely determine a probability on $(\mathbb{R}^n \times \mathbb{R}^m, \mathcal{B}_{\mathbb{R}^n} \times \mathcal{B}_{\mathbb{R}^m})$ by

$$\mu(B_x \times B_y) = \int_{B_x} G(x, B_y)\, \mathbb{P}_X(dx),$$

for all $B_x \in \mathcal{B}_{\mathbb{R}^n}$ and $B_y \in \mathcal{B}_{\mathbb{R}^m}$.

Proof. For simplicity, we call RECT $:= \{B_x \times B_y : B_x \in \mathcal{B}_{\mathbb{R}^n}, B_y \in \mathcal{B}_{\mathbb{R}^m}\}$. Since the formula is defined for every element in RECT, which is a semi-algebra (check as ex.), once we claim μ is σ-additive on this semi-algebra, by Carathéodory extension theorem, it is a probability on $\mathcal{B}_{\mathbb{R}^n} \times \mathcal{B}_{\mathbb{R}^m}$.

Let $\left\{B_1^{(n)} \times B_2^{(n)}\right\}_{n=1}^{\infty}$ be a sequence of disjoint elements in RECT and further assume their union is in RECT. The goal is to show

$$\mu\left\{\sqcup_{n=1}^{\infty}\left[B_1^{(n)} \times B_2^{(n)}\right]\right\} = \sum_{n=1}^{\infty} \mu\left(B_1^{(n)} \times B_2^{(n)}\right).$$

Note if $\sqcup_n \left[B_1^{(n)} \times B_2^{(n)}\right] = B_1 \times B_2$, then

$$\mathbb{1}_{B_1}(x)\, \mathbb{1}_{B_2}(y) = \mathbb{1}_{B_1 \times B_2}(x, y) = \sum_n \mathbb{1}_{B_1^{(n)} \times B_2^{(n)}}(x, y) = \sum_n \mathbb{1}_{B_1^{(n)}}(x)\, \mathbb{1}_{B_2^{(n)}}(y),$$

which further leads to (use MCT, not Fubini),

$$\begin{aligned}
\mu\left(B_1 \times B_2\right) &= \int_x \mathbb{1}_{B_1}(x)\, G\left(x, B_2\right) \mathbb{P}_X\left(dx\right) \\
&= \int_x \left[\int_y \mathbb{1}_{B_1}(x)\, \mathbb{1}_{B_2}(y)\, G\left(x, dy\right)\right] \mathbb{P}_X\left(dx\right) \\
&= \int_x \left[\int_y \sum_n \mathbb{1}_{B_1^{(n)}}(x)\, \mathbb{1}_{B_2^{(n)}}(y)\, G\left(x, dy\right)\right] \mathbb{P}_X\left(dx\right) \\
&= \int_x \sum_n \left[\int_y \mathbb{1}_{B_1^{(n)}}(x)\, \mathbb{1}_{B_2^{(n)}}(y)\, G\left(x, dy\right)\right] \mathbb{P}_X\left(dx\right) \\
&= \sum_n \int_x \mathbb{1}_{B_1^{(n)}}(x) \left[\int_y \mathbb{1}_{B_2^{(n)}}(y)\, G\left(x, dy\right)\right] \mathbb{P}_X\left(dx\right) \\
&= \sum_n \int_x \mathbb{1}_{B_1^{(n)}}(x)\, G\left(x, B_2^{(n)}\right) \mathbb{P}_X\left(dx\right) \\
&= \sum_n \int_{B_1^{(n)}} G\left(x, B_2^{(n)}\right) \mathbb{P}_X\left(dx\right) \\
&= \sum_n \mu\left(B_1^{(n)} \times B_2^{(n)}\right).
\end{aligned}$$

□

This theorem says if the conditional distribution of $Y \mid X$, written G, and marginal distribution of X, written $\mathbb{P}_X$, are given, then there exists a unique probability μ on $(\mathbb{R}^n \times \mathbb{R}^m, \mathcal{B}_{\mathbb{R}^n} \times \mathcal{B}_{\mathbb{R}^m})$ such that

$$\mu\left(B_x \times B_y\right) = \int_{B_x} G\left(x, B_y\right) \mathbb{P}_X\left(dx\right),$$

for all $B_x \in \mathcal{B}_{\mathbb{R}^n}$ and $B_y \in \mathcal{B}_{\mathbb{R}^m}$. But recall Definition 4.2 (iii), it must be satisfied that

$$\mathbb{P}(X \in B_x, Y \in B_y) = \int_{B_x} G(x, B_y) \mathbb{P}_X(dx),$$

meaning that the measure μ is exactly the joint distribution of (X, Y) on $(\mathbb{R}^n \times \mathbb{R}^m, \mathcal{B}_{\mathbb{R}^n} \times \mathcal{B}_{\mathbb{R}^m})$.

Let's continue our example.

Example. Let X, Y be two random variables. Suppose $X \sim U[0,1]$ and $Y \mid X = x \sim \text{Ber}(x)$. Find the joint distribution of (X, Y) and the marginal distribution of Y.

Solution (Traditional)**.** We can do this without touching the modern definition, which is left as an exercise.

Solution (Modern)**.** Recall the last example. The assumption is just that distributions of X, written $F_X(\cdot)$, and $Y \mid X = x$, written $G(x, \cdot)$, on $(\mathbb{R}, \mathcal{B}_{\mathbb{R}})$ are

$$F_X(x) = \begin{cases} 0 & x < 0 \\ x & 0 \le x \le 1 \\ 1 & x > 1 \end{cases} \quad \text{and} \quad \begin{cases} G\Big(x, \emptyset\Big) = 0 \\ G\Big(x, \{1\}\Big) = x \\ G\Big(x, \{0\}\Big) = 1 - x \\ G\Big(x, \{0,1\}\Big) = 1 \end{cases}.$$

By the last theorem, we must have for any Borel sets B_x and B_y,

$$\mathbb{P}(X \in B_x, Y \in B_y) = \int_{B_x} G(x, B_y)\mathbb{P}_X(dx).$$

When $B_x \cap [0,1] = \emptyset$, we know immediately $\mathbb{P}(X \in B_x, Y \in B_y) = 0$ since we are integrating under a measure zero set B_x.

Taking $B_1 = (-\infty, \tilde{x}]$ where $0 < \tilde{x} < 1$ and $B_2 = \{1\}$, we get

$$\mathbb{P}(X \leq \tilde{x}, Y = 1) = \int_{-\infty}^{\tilde{x}} G\Big(x, \{1\}\Big)\mathbb{P}_X(dx) = \int_{-\infty}^{\tilde{x}} x dx = \frac{\tilde{x}^2}{2}.$$

And take $B_x = (-\infty, \tilde{x}]$ where $0 < x < 1$ and $B_y = \{0\}$, we get

$$\mathbb{P}(X \leq \tilde{x}, Y = 0) = \int_{-\infty}^{\tilde{x}} G\Big(x, \{0\}\Big)\mathbb{P}_X(dx) = \int_{-\infty}^{\tilde{x}} 1 - x \, dx = \tilde{x} - \frac{\tilde{x}^2}{2}.$$

And to sum up, for any $B_x = (-\infty, \tilde{x}]$ where $0 < x < 1$ and $B_y \in \mathcal{B}_{\mathbb{R}}$, we get

$$\mathbb{P}(X \leq \tilde{x}, Y \in B_y) = \int_{-\infty}^{\tilde{x}} G(x, B_y)\mathbb{P}_X(dx).$$

Also note that $G(x, \cdot)$ is a discrete measure on $\mathbb{R}$ for any $0 < x < 1$. We can further conclude

$$\mathbb{P}(X \leq \tilde{x}, Y \in B_y) = \int_{-\infty}^{\tilde{x}} G(x, B_y)\mathbb{P}_X(dx) = \int_{-\infty}^{\tilde{x}} \sum_{y \in B_y} G\Big(x, \{y\}\Big)\mathbb{P}_X(dx).$$

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

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

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

4.3.2 From $Y \mid X$ and X to $X \mid Y$

We have gone from $Y \mid X$ and X to (X, Y) now. Formally, we already know the distribution for any Borel sets $B_x \in \mathcal{B}_{\mathbb{R}^n}$ and $B_y \in \mathcal{B}_{\mathbb{R}^m}$,

$$\mathbb{P}(X \in B_x, Y \in B_y) = \int_{B_x} G(x, B_y)\mathbb{P}_X(dx).$$

One natural next step is to derive the marginal of Y and the conditional $X \mid Y$, just as the Bayes formula. And formally, our goal is to find $\mathbb{P}_Y(\cdot)$ and $\tilde{G}(y, \cdot)$ such that

$$\mathbb{P}(Y \in B_y, X \in B_x) = \int_{B_y} \tilde{G}(y, B_x)\mathbb{P}_Y(dy).$$

Note that the LHS is already known by the last step, and $\mathbb{P}_Y$ can be obtained by taking $B_x = \mathbb{R}^n$ normally. And usually by some special choices of B_y, the conditional $\tilde{G}(y, B_x)$ can be found. Let's illustrate by examples.

Example. Let X, Y be two random variables. Suppose $X \sim U[0,1]$ and $Y \mid X = x \sim \text{Ber}(x)$. We already found

$$\mathbb{P}(X \le \tilde{x}, Y = 1) = \frac{\tilde{x}^2}{2} \text{ and } \mathbb{P}(X \le \tilde{x}, Y = 0) = \tilde{x} - \frac{\tilde{x}^2}{2}.$$

And we already found the marginal of Y is $\mathbb{P}_Y(\{0\}) = \mathbb{P}_Y(\{1\}) = 1/2$. Find $X \mid Y$.

Solution. Since Y is discrete, we still discuss case by case. Starting with $Y = 1$, by the already calculated result, definition of $\tilde{G}$ and integration wrt counting measure, we have for $0 < \tilde{x} < 1$,

$$\frac{\tilde{x}^2}{2} = \mathbb{P}(Y = 1, X \le \tilde{x}) = \int_{\{1\}} \tilde{G}\Big(y, (0, \tilde{x}]\Big)\mathbb{P}_Y(dy) = \frac{1}{2}\tilde{G}\Big(1, (0, \tilde{x}]\Big),$$

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

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

And a similar argument leads to

$$\tilde{G}\Big(0, (0, \tilde{x}]\Big) = 2\tilde{x} - \tilde{x}^2 = \mathbb{P}(X \le \tilde{x} \mid Y = 0).$$

Let's see another example.

Example. 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. Since $X_n := \vee_{i=1}^n X_i$ is a function of $(X_1, \ldots, X_n)$, we know $X_1, X_{(n)}$ are both random variables on a same probability space $(\Omega, \mathcal{A}, \mathbb{P})$, and this question is well-defined.

To do it, we must first find the marginal of $X_{(n)}$ since we will need its marginal distribution. This is easy and we will get for all $\tilde{y} \in [0,1]$,

$$F_{X_{(n)}}(\tilde{y}) = \tilde{y}^n \text{ and } f_{X_{(n)}}(\tilde{y}) = n\tilde{y}^{n-1}. \tag{3}$$

By the existence of conditional distribution, there is a function $G(y, B) : \mathbb{R} \times \mathcal{B}_{\mathbb{R}} \mapsto [0,1]$ such that

(i) for each $y \in \mathbb{R}$, $G(y, \cdot)$ is a probability on $(\mathbb{R}, \mathcal{B}_{\mathbb{R}})$,
(ii) for each $B \in \mathcal{B}_{\mathbb{R}}$, $G(\cdot, B)$ is $\mathcal{B}_{\mathbb{R}}/\mathcal{B}_{[0,1]}$ measurable,
(iii) for any Borel sets $B_1 \in \mathcal{B}_{\mathbb{R}}$ and $B_2 \in \mathcal{B}_{\mathbb{R}}$,

$$\mathbb{P}(X_{(n)} \in B_1, X_1 \in B_2) = \int_{B_1} G(y, B_2)\mathbb{P}_{X_{(n)}}(dy),$$

where $\mathbb{P}_{X_{(n)}}(\cdot)$ is the distribution of $X_{(n)}$ on $(\mathbb{R}, \mathcal{B}_{\mathbb{R}})$. The goal is to find G, and to make our life easier, we are going to take $B_1 = (-\infty, \tilde{y}]$ and $B_2 = (-\infty, \tilde{x}]$, i.e.

$$\mathbb{P}(X_{(n)} \le \tilde{y}, X_1 \le \tilde{x}) = \int_0^{\tilde{y}} G\Big(y, (0, \tilde{x}]\Big)\mathbb{P}_{X_{(n)}}(dy), \tag{4}$$

since cdfs are measures are 1-1. In fact, special choices of $\tilde{x}$ and $\tilde{y}$ will solve this problem.

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:

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

And the RHS of (4) can be computed by the marginal (3):

$$\int_0^{\tilde{y}} G\Big(y, (0, \tilde{x}]\Big) \mathbb{P}_{X_{(n)}}(dy) = \int_0^{\tilde{y}} G\Big(y, (0, \tilde{x}]\Big) \times n y^{n-1} dy.$$

Equating them, it must hold that for all $0 < \tilde{x} < \tilde{y} < 1$ that

$$\int_0^{\tilde{y}} G\Big(y, (0, \tilde{x}]\Big) \times n y^{n-1} dy = \tilde{x}\tilde{y}^{n-1}.$$

Differentiating both sides, by Lebesgue differentiation theorem, we get for all $0 < \tilde{x} < \tilde{y} < 1$,

$$G\Big(\tilde{y}, (0, \tilde{x}]\Big) \times n\tilde{y}^{n-1} = (n-1)\tilde{x}\tilde{y}^{n-2} \quad \iff \quad G\Big(\tilde{y}, (0, \tilde{x}]\Big) = \left(1 - \frac{1}{n}\right) \frac{\tilde{x}}{\tilde{y}}.$$

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 the RHS of (4) can be computed by the marginal (3):

$$\int_0^{\tilde{y}} G\Big(y, (0, \tilde{x}]\Big) \mathbb{P}_{X_{(n)}}(dy) = \int_0^{\tilde{y}} G\Big(y, (0, \tilde{x}]\Big) \times n y^{n-1} dy.$$

Equating them, it must hold that for all $0 < \tilde{y} \le \tilde{x} < 1$ that

$$\int_0^{\tilde{y}} G\Big(y, (0, \tilde{x}]\Big) \times n y^{n-1} dy = \tilde{y}^n.$$

Differentiating both sides, by Lebesgue differentiation theorem, we get for all $0 < \tilde{y} \le \tilde{x} < 1$,

$$G\Big(\tilde{y}, (0, \tilde{x}]\Big) \times n\tilde{y}^{n-1} = n\tilde{y}^{n-1} \quad \iff \quad G\Big(\tilde{y}, (0, \tilde{x}]\Big) = 1.$$

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 $B_2 \in \mathcal{B}_{\mathbb{R}}$,

$$\mathbb{P}(X_{(n)} \in B_1, X_1 \in B_2) = \int_{B_1} G(y, B_2) \mathbb{P}_{X_{(n)}}(dy).$$

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 $\mathbb{P}_{X_{(n)}}(B_1) = 0$. So, we can actually define $G(y, B_2)$ 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

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

Here is a graph for fixed $\tilde{y}$ and varying $\tilde{x} \in [0, \tilde{y}]$.

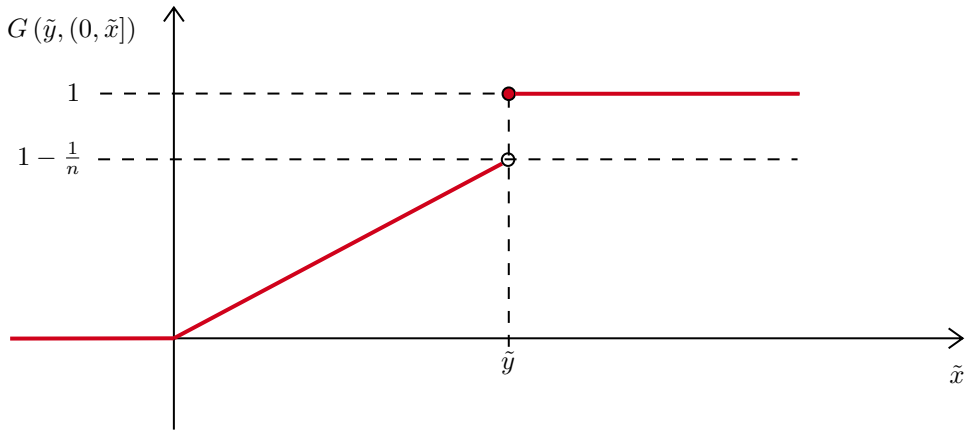

Figure 3: Conditional distribution of $X_1 \mid X_{(n)} = \tilde{y}$

Since when $\tilde{y}$ is fixed, the function $G(\tilde{y}, \cdot)$ is a measure. And we can compute (cf. Jing's manuscript §2.9)

$$\begin{aligned} G\Big(\tilde{y}, \{\tilde{y}\}\Big) &= G\Big(\tilde{y}, (-\infty, \tilde{y}]\Big) - G\Big(\tilde{y}, (-\infty, \tilde{y})\Big) \\ &= G\Big(\tilde{y}, (-\infty, \tilde{y}]\Big) - \lim_{k\to\infty} G\Big\{\tilde{y}, \Big(-\infty, \tilde{y} - \frac{1}{k}\Big]\Big\} \\ &= 1 - \lim_{k\to\infty} \left\{\left(1-\frac{1}{n}\right)\frac{\tilde{y}-1/k}{\tilde{y}}\right\} = \frac{1}{n}, \end{aligned}$$

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.

4.4 Computing Conditional Distribution

The definition is not constructive, which means it provides no help to really find the distribution. But intuitively, the distribution should be a quotient form since

$$p_{Y|X}(y|x) = \frac{p_{X,Y}(x,y)}{p_X(x)} \text{ is the conditional density for discrete case wrt the counting measure,}$$
$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)} \text{ is the conditional density for continuous case wrt the Lebesgue measure,}$$

and both forms can be viewed as densities. And here is the result.

Theorem 4.2. *Let $X \in \mathbb{R}^m$ and $Y \in \mathbb{R}^n$ be random vectors, and suppose*

(i) (X, Y) has a joint density $p : \mathbb{R}^m \times \mathbb{R}^n \mapsto \mathbb{R}$ wrt a product measure $\mu \times \nu$ on $(\mathbb{R}^{m+n}, \mathcal{B}_{\mathbb{R}^{m+n}})$,

(ii) X has a marginal density $p_X(x) = \int p(x,y)d\nu(y)$.

Let $E = \{x \in \mathbb{R}^m : p_X(x) > 0\}$. Define

$$p_{Y|X}(y|x) = \begin{cases} \frac{p(x,y)}{p_X(x)} & \text{for all } x \in E, \\ p_0(y) & \text{for all } x \notin E, \end{cases}$$

where p_0 is the density for an arbitrary fixed probability distribution P_0. Then for any $x \in \mathbb{R}^m$, the defined $p_{Y|X}(\cdot|x)$ is a density on $(\mathbb{R}^n, \mathcal{B}_{\mathbb{R}^n})$ wrt ν, and is the density of $Y \mid X = x$.

Computable for all cases.

Though the dominating measures may differ, this concept and computation works all the time. Specifically, when X is discrete, we divide the mass p_X; when X is continuous, we divide the density f_X.

Negligibility outside E.

The set $E = \{x \in \mathbb{R}^m : p_X(x) > 0\}$ serves as the support of X. When conditioning outside E, the conditional distribution does not matter, and can actually be left undefined.

Proof. Confer §6.2 by Keener [5].

4.4.1 Example: Poisson Process

We use the Poisson process as an example, which is defined by

Definition 4.3 (Poisson process)**.** A Poisson Process $(N_t)_{t\geq 0}$, with intensity (rate) $\lambda > 0$, is a random process defined, for all $t \geq 0$, via

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

where $T_n = \tau_1 + \cdots + \tau_n$, with independent rvs $\tau_i \overset{\text{iid}}{\sim} \text{Exp}(\lambda)$.

One of the most important feature, for which the Poisson process got its name, is the following:

Theorem 4.3. *If $(N_t)_{t\geq 0}$ is a Poisson process with intensity λ, then $N_t \sim \text{Pois}(\lambda t)$ for any fixed $t > 0$.*

We need to use the conditioning distribution for help, which is why it is here.

Proof. 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

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

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

$$\mathbb{P}(N_t = n) = \mathbb{P}\Big\{T_n \leq t, T_{n+1} > t\Big\}$$
$$\left(= \int_0^t P\left(T_n \in ds, T_{n+1} > t\right) = \int_0^t P\left(T_{n+1} > t \mid T_n = s\right) P_{T_n}(ds) = \int_0^t P\left(\tau_{n+1} > t - s\right) f_{T_n}(s) ds\right),$$

But how to make this formal? Denote the conditional distribution of $T_{n=1}|T_n = s$ by $G(s,\cdot)$, the definition guarantees that

$$= \int_0^t G\Big(s, [t,\infty)\Big)\, \mathbb{P}_{T_n}(ds)$$

where the computation theorem tells us the conditional density is in fact

$$G\Big(s, [t,\infty)\Big) = \int_t^\infty \frac{f_{T_{n+1},T_n}(u,s)}{f_{T_n}(s)} du.$$

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

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

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

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

Remark. In fact, we can compute without conditioning when arriving at $\mathbb{P}(N_t = n) = \mathbb{P}\Big\{T_n \le t, T_{n+1} > t\Big\}$, which is even simpler. The reason for conditioning is to utilize the new method here.

□

There is a unifying and rigorous definition of conditional distribution, just as the conditional expectation. This is called the **regular conditional distribution**, which is beyond the scope of this course. See theorem 33.3 in Billingsley[6] , or the chapter 5 by Rao[7]. Chapter 6 and Definition 6.2 in by Keener[5] will be enough for this course.

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

[5] Robert W Keener. *Theoretical statistics: Topics for a core course*. Springer, 2010.

[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.

[7] Malempati Madhusudana Rao. *Conditional measures and applications*, volume 271. CRC Press, 2005.