STAT 532: Foundations of Statistical Inference I **Fall 2020**

Lecture 1: Measure Space

Lecturer: Dennis Cox *Scribes: Shu-Hsien Cho*

Objectives of theoretical statistics

Statistics is about the mathematical modeling of observable phenomena, using stochastic models, and about analyzing data: estimating parameters of the model and testing hypotheses. Theoretical statistics relies heavily on probability theory, which in turn is based on measure theory. Thus, a student of advanced statistics needs to learn some measure theory. A proper introduction to measure theory is not provided here. Instead, definitions and concepts are given and the main theorems are stated without proof.

Measure theory is a rather difficult and dry subject, and many statisticians believe it is unnecessary to learn measure theory in order to understand statistics. To counter these views, we offer the following list of benefits from studying measure theory:

1. A good understanding of measure theory eliminates the artificial distinction between discrete and continuous random variables. Summations become an example of the abstract integral, so one need not dichotomize proofs into the discrete and continuous cases, but can cover both at once.
2. One can understand probability models which cannot be classified as either discrete or continuous. Such models do arise in practice, e.g. when censoring a continuous lifetime and in Generalized Random Effects Models such as the Beta-Binomial.
3. The measure theoretic statistics presented here provides a basis for understanding complex problems that arise in the statistical inference of stochastic processes and other areas of statistics.
4. Measure theory provides a unifying theme for much of statistics. As an example, consider the notion of likelihoods, which are rather mysterious in some ways, but at least from a formal point of view are measure theoretically quite simple. As with many mathematical theories, if one puts in the initial effort to understand the theory, one is rewarded with a deeper and clearer understanding of the subject.
5. Certain fundamental notions (such as conditional expectation) are arguably not completely understandable except from a measure theoretic point of view. Rather than spend more words on motivation, let us embark on the subject matter.

1.1 Measures

A measure space is a 3-element $(\Omega, \mathcal{F}, \mu)$, where Ω is a set (e.g. possible outcomes of random experiment, all "*sets*"), $\mathcal{F}$ is a collection of subsets of a set Ω, and μ is a *measure* / function from $\mathcal{F}$ to $[0, \infty)$. μ satisfies

1. $\mu(\emptyset) = 0$
2. A sequence of measurable disjoint sets $A_1, A_2, \cdots$ in $\mathcal{F}$, then $\mu\left(\bigcup\limits_{n=1}^{\infty} A_n\right) = \sum\limits_{n=1}^{\infty} \mu(A_n)$

Note. ω (element) $\in \Omega$, but $\omega \neq \{\omega\}$ (set). That is, $\mu(\{\omega\})$ may be meaningful, but $\mu(\omega)$ is nonsense.

1.1.1 σ-fields

Here we need some requisite properties for the class of sets on which a measure is well defined.

Definition 1.1.1. Let $\mathcal{F}$ be a collection of subsets of a set Ω, then $\mathcal{F}$ is called a σ-field iff it satisfies :

1. $\emptyset \in \mathcal{F}$
2. If $A \in \mathcal{F}$, the $A^c \in \mathcal{F}$
3. If $A_1, A_2, \cdots$ is a sequence of elements (that is, $\{A_1, A_2, \cdots\}$ is a countable subset) of $\mathcal{F}$, $\bigcup\limits_{n=1}^{\infty} A_n \in \mathcal{F}$

$(\Omega, \mathcal{F})$ is a measurable space, and the element of $\mathcal{F}$ is measurable set / measurable event. Given the measurable space, a measure is a function $\mu : \mathcal{F} \to \bar{\mathbb{R}} = \mathbb{R} \cup \{-\infty, \infty\}$ (extended real numbers) satisfying

1. nonnegative: $\forall A \in \mathcal{F}$, $0 \leq \mu(A) \leq \infty$.
2. $\mu(\emptyset) = 0$
3. (Countable additivity property) If $A_1, A_2, \cdots$ is a sequence of disjoint sets of $\mathcal{F}$, then $\mu\left(\bigcup\limits_n A_n\right) = \sum\limits_n \mu(A_n)$

Remark. Probability Spaces

1. $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space with $\mathbb{P}(\Omega) = 1$, $\mathbb{P}$ is *probability measure.*
2. Elements of $\mathcal{F}$ (Measurable sets) are *events.*
3. Ω is the *sample space.* (underlying space)
4. $\emptyset \in \mathcal{F}$, that is to say, $\Omega \in \mathcal{F}$
5. Given any set Ω, the most trivial (smallest) σ-field is $\mathcal{F} = \{\emptyset, \Omega\}$. The power set $\mathcal{P}(\Omega) = \{A : A \subset \Omega\}$ consisting all subsets of Ω is the largest σ-field on Ω. (2^Ω) It is easy to prove that the $\mathcal{F}$ is a σ-field.
   (a) $\emptyset \in \mathcal{F}$
   (b) $A \in \mathcal{F}$ implies $A^c = \Omega \setminus A \in \mathcal{F}$
   (c) $A_1, A_2, \cdots$ in $\mathcal{F}$ implies $\bigcup\limits_n A_n \in \mathcal{F}$

Given any $\mathcal{A} \subset \mathcal{P}(\Omega)$, $\sigma(\mathcal{A})$ is the smallest σ-field containing $\mathcal{A}$.

Example 1.1.1. (Jun Shao Equation 1.1, Robert Ash Example 1.2.2)

Let A be a nonempty subset of Ω ($A \subseteq \Omega$), the smallest σ-field containing A, $\sigma(\{A\}) = \{\emptyset, A, A^c, \Omega\}$

Definition 1.1.2. The smallest σ-field containing $\mathcal{C}$, a collection of subsets of Ω, is denoted by $\sigma(\mathcal{C})$ and is called the σ-field generated by $\mathcal{C}$. That is to say, if $\mathcal{F}$ is any σ-field containing $\mathcal{C}$, then $\sigma(\mathcal{C}) \subset \mathcal{F}$. (The details and proof are shown in the main text Prop. 1.1.1, it also shows that the exception of the countable additivity property that intersection may be an uncountable collection Γ of σ-fields.)

Definition 1.1.3. Borel σ-field $\mathcal{B}$ on $\mathbb{R}$ is the σ-field generated by the collection of all finite open intervals. Any "*practical*" subset of $\mathbb{R}$ is a Borel set. Actually, all open/closed sets, all intervals(semi-infinite interval or half open interval), and all finite subsets of $\mathbb{R}$ are Borel sets.

$$\begin{aligned}\mathcal{B} &= \sigma\left(\{(a,b) : -\infty < a < b < \infty\}\right)\\ &= \sigma\left(\{[a,b] : -\infty < a < b < \infty\}\right)\\ &= \sigma\left(\{[a,\infty) : a \in \mathbb{R}\}\right)\end{aligned}$$

So, there exists a unique measure m (Borel measure) on $(\mathbb{R}, \mathcal{B})$ satisfies $\forall a, b \in \mathbb{R}, a < b$, then $m((a,b)) = b-a$.

Example 1.1.2. Counting Measure of any $(\Omega, \mathcal{F}) : \mu(A) = \#(A)$ the number of elements in A. If A is an infinite set, then $\#(A) = \infty$. It is fairly easy to check that $(\Omega, \mathcal{F}, \#)$ is a measure space (check the three properties). Unless otherwise stated, we will use the power set for the σ–field when dealing with counting measure, i.e. $\mathcal{F} = \mathbb{P}(\Omega)$), the collection of all subsets of Ω. Note that most of the unions and intersections have been *countable.*

Example 1.1.3. Unit Point Mass Measure: Given a measurable space $(\Omega, \mathcal{F})$ and $x \in \Omega$.

$$\delta_x(A) = \begin{cases} 1, & \text{if } x \in A \\ 0, & \text{if } x \notin A \end{cases}$$

Note that counting measure on $\{x_1, x_2, \cdots\}$ can be written in terms of unit point masses as $\# = \sum\limits_i \delta_{x_i}$, and the sum of measures is a measure. Then we could check it is a probability measure easily. this measure is useful in empirical distribution (see below).

To compute some other values of $m(B), B \in \mathcal{B}$, we need Proposition below. We will also add another part to this result for increasing unions.

Proposition 1.1.1. *Basic Properties of Measures. Let $(\Omega, \mathcal{F}, \mu)$ be a measure space.*

1. *Monotonicity: $A \subset B$ implies $\mu(A) \leq \mu(B)$*
2. *Subadditivity: For any sequence $A_1, A_2, \cdots, \mu(\cup A_n) \leq \sum \mu(A_n)$*
3. *Continuity (increasing/decreasing intersections): If $A_1 \subset A_2 \subset A_3 \subset \cdots$ $(orA_1 \supset A_2 \supset A_3 \supset \cdots)$ and $\mu(A_i) \leq \infty$, then*

$$\mu\left(\lim_{n\to\infty} A_n\right) = \lim_{n\to\infty} \mu(A_n), \text{ where } \lim_{n\to\infty} A_n = \bigcup_{i=1}^{\infty} A_i (or \bigcap_{i=1}^{\infty} A_i)$$

Proof. 1. $A \subset B, B = A \cup (A^c \cap B)$, A and $(A^c \cap B)$ are disjoint. By countable additivity property (Definition 1.1.1-3), $\mu(B) = \mu(A) + \mu(A^c \cap B) \geq \mu(A)$ by Definition 1.1.1-1.

Some other detailed proofs could be found in K.L. Chung Section 2.2. □

Example 1.1.4. 1. $m : x \in \mathbb{R}, m(\{x\}) = m\left(\bigcap\limits_{n=1}^{\infty}(x - \frac{1}{n}, x + \frac{1}{n})\right) = \lim\limits_{n\to\infty} m\left[(x - \frac{1}{n}, x + \frac{1}{n})\right] = 0$ (Singleton set in Lebesgue measure has length 0.)

2. $m([a,b]) = m(\{a\} \cup (a,b) \cup \{b\}) = m(\{a\}) + m(\{b\}) + m((a,b)) = m((a,b)) = b - a$

Proposition 1.1.2. *1. $\mu_1, \mu_2, \cdots$ are a finite/infinite sequence of measures on $(\Omega, \mathcal{F})$, and $a_1, a_2, \cdots$ are nonnegative real numbers. Then $\mu = \sum\limits_i a_i \mu_i$ is also a measure on $(\Omega, \mathcal{F})$*

2. If each of the μ_i is a probability measure and $\sum a_i = 1$, then μ is also a probability measure. We also allow some $a_i = \infty$ here for $\infty \cdot 0 = 0$

How do we get a $\mathbb{R}^p$ measure? see Section 1.3 about the product measure theorem. With these measure knowledge, we could simplify some proof steps.

1.1.2 Distribution Functions

Given any $F : \mathbb{R} \to \mathbb{R}$ satisfying following theorem, there is a unique Borel probability measure $\mathbb{P}$ on $(\mathbb{R}, \mathcal{B})$ can define (cumulative) distribution function $F(x) = \mathbb{P}((-\infty, x]), \forall x \in \mathbb{R}$.

Theorem 1.1.3. *The c.d.f. of a Borel probability measure has the following properties*

1. $F(-\infty) = \lim\limits_{x \to -\infty} F(x) = 0$

2. $F(\infty) = \lim\limits_{x \to \infty} F(x) = 1$

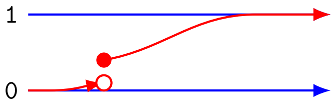
3. *F is **nondecreasing** i.e. $F(x) \le F(y)$ if $x < y$.*

4. *F is **right-continuous** $F(x+0) = \lim\limits_{z \downarrow x} F(z) = F(x)$, $z \downarrow x$ means $z > x$ and $z \to x$.*

5. *F's **left limit exists**: $F(x-0) = \lim\limits_{z \uparrow x} F(z) = F(x)$*

Example 1.1.5.

$$F(x) = \begin{cases} 0, & x < 0 \\ \dfrac{1}{2}(1+x), & \text{if } 0 \le x \le 1 \\ 1, & \text{if } x > 1 \end{cases}$$



Remark. **Quantile Function**: more detailed content of quantile function and its application could be found in Jun Shao 5.2 & 5.3 for Estimation in Nonparametric Models. Professor Cox mentioned the quantile function as an introduction after we discuss the inverse image in 1.2.1, but I still put this remark here. The quantile function in quite important in nonparametric statistics and empirical distribution (will be shown later).

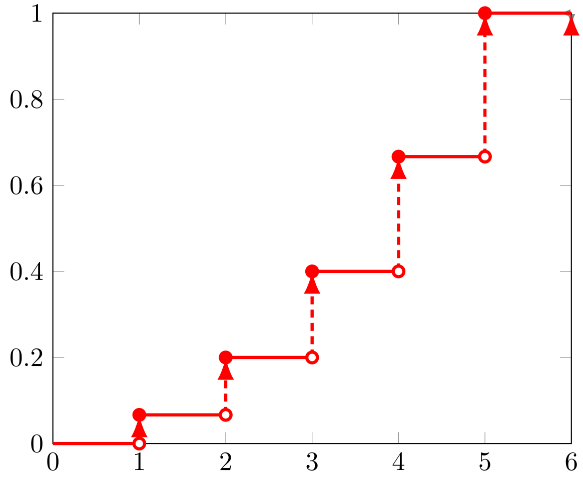
1. Quantile function is the inverse of a c.d.f.

2. For $\alpha \in (0,1)$ or $[0,1]$, $F^-(\alpha) = \inf\{x : F(x) \ge \alpha\}$, and $F^+(\alpha) = \sup\{x : F(x) \le \alpha\}$, if F is strictly increasing and continuous, iff $F^- = F^+ = F^{-1}$. (We can easily find that $\inf \emptyset = +\infty$ and $\sup \emptyset = -\infty$, so $F^-(\alpha) \le F^+(\alpha), \forall \alpha \in (0,1)$ in HW1 exercise 1.1.15)

3. If $0 < \alpha < 1$,then $\exists x$ s.t. $F(x) < \alpha$

4. $F^-(\alpha) = F^+(\alpha) = x$ if $\forall \epsilon > 0$ there exist $x_1 \in (x-\epsilon, x)$ and $x_2 \in (x, x+\epsilon)$ with $F(x_1) < F(x) < F(x_2)$. x is a *point of increase* for F

5. *Median* of a c.d.f. is defined as $\text{Med}(F) = \frac{1}{2}\left[F^-\left(\frac{1}{2}\right) + F^+\left(\frac{1}{2}\right)\right]$. Using order statistics and $k = \frac{n}{2}, x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)} \leq$, median is $\frac{x_{(k)} + x_{(k+1)}}{2}$



Example 1.1.6. Please see the figure above. $F^{-1}(0.2) = [2, 3), F^-(0.2) = 2, F^+(0.2) = 3$ because $\{x : F(x) \geq 0.2\} = [2, \infty)$ and $\{x : F(x) \leq 0.2\} = (-\infty, 3]$

Remark. **Empirical Distribution**: Here Professor Cox also discussed Empirical Distribution as an introduction after introducing inverse image operator. More details shown in Jun Shao 5.1 Distribution Estimators.

1. Data $(x_1, x_2, \cdots, x_n)$ are elements of some set $\Omega = \mathbb{R}$, we count the data only once in the set Ω, i.e. $x_i \neq x_j, \forall i \neq j$. But we allow replicates in data set. e.g. $\{1, 2, 2\} = \{1, 2\}$ but $(1, 2, 2) \neq (1, 2)$.

2. To include this data set, we use **unit point mass measure** here. We also need Proposition 1.1.2 to conduct an empirical distribution.

3. $\hat{P} = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}$.

4. Note that for any measurable set $A \in \Omega$, $\hat{P}(A)$ is the proportion of data points(observations) in A.

5. $\hat{P}_n(A) = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}(A) = \frac{1}{n}\#\{i : x_i \in A\}$

6. $x_i \in \mathbb{R}$, then $\hat{P}$ is Borel probability measure has c.d.f $\hat{F}(x) =$ (proportion of observations $\leq x) = \hat{P}((-\infty, x])$

1.2 Measurable Functions and Integration

A measure μ as a real valued function is defined on a class of subsets $\mathcal{F}$ of Ω. Every set $A \subset \Omega$ is associated with a unique real valued function called the indicator function of A. It is given by

$$I_A(x) = \begin{cases} 1, & \text{if } x \in A \\ 0, & \text{if } x \notin A \end{cases}$$

Thus, we may think of a measure as being defined on the class of indicator functions of sets $A \in \mathcal{F}$. Instead of writing $\mu(A)$, we could write "$\mu(I_A)$". In this section we define the abstract notion of integration which extends the definition of μ to a large class of real valued functions, i.e. we can define "$\mu(f)$," usually written $\int f d\mu$ (so that $\mu(A) = \int I_A d\mu$)

1.2.1 Measurable Functions

Since Ω can be quite arbitrary, it is often convenient to consider a function (mapping) f from Ω to a simpler space Λ (often $\Lambda = \mathbb{R}^k$). The **inverse image operator** is: any function $f : A \to B$, $f^{-1} : \mathbb{P}(B) \to \mathbb{P}(A)$. Where $C \subseteq B, f^{-1}(C) = \{a \in A : f(a) \in C\}$. **It maps a set to another set**

Proposition 1.2.1. *1.* $f^{-1}(A^c) = [f^{-1}(A)]^c$

2. For any $A \subset \Lambda$ *and* $A_1, A_2, \cdots$ *are subsets of* Λ, $f^{-1}\left(\bigcup_i A_i\right) = \bigcup_i f^{-1}(A_i)$ *and* $f^{-1}\left(\bigcap_i A_i\right) = \bigcap_i f^{-1}(A_i)$.

Proof could be found in Cox section 1.2.1.

Applying this to a c.d.f. $F : \mathbb{R} \to \mathbb{R}$

$$F^{-1}(\{\alpha\}) = \begin{cases} \emptyset, & \text{if no } x \text{ s.t. } F(x) = \alpha \\ \{x\}, & \text{if unique } x \text{ s.t. } F(x) = \alpha \\ [x_1, x_2), & \text{if } F(x) = \alpha \text{ and all } x \in [x_1, x_2), \text{but } F(x_2) \neq \alpha \\ [x_1, x_2], & \text{if } F(x) = \alpha \text{ and all } x \in [x_1, x_2), \text{and } F(x_2) = \alpha \end{cases}$$

If $F^{-1}(\{\alpha\}) = [x_1, x_2)$, then $F^-(\alpha) = x_1, F^+(\alpha) = x_2$ (that is to say, always $F^- \leq F^+$)

Definition 1.2.1. $(\Omega, \mathcal{F})$ and $(\Lambda, \mathcal{G})$ are measurable spaces. $f : \Omega \to \Lambda$ is measurable function iff $\forall A \in \mathcal{G}, f^{-1}(A) \in \mathcal{F}$ (or $f^{-1}(\mathcal{G}) \subset \mathcal{F}$, a sub-$\sigma$-field of $\mathcal{F}$). $\Lambda = \mathbb{R}$ and $\mathcal{G}$ is the Borel σ-field, f is Borel measurable (real value Borel function). All this kinds of function is more practical.

Definition 1.2.2. $f : (\Omega, \mathcal{F}) \to (\Lambda, \mathcal{G})$, the σ-field generated by f is $f^{-1}(\mathcal{G})$, denoted $\sigma(f)$

Take the indicator function above as an example, $I_A(x) = \delta_x(A)$ (Sure, it is similar to empirical distribution).

For $B \in \mathcal{B}$=Borel sets as an example.

$$I_A^{-1}(B) = \begin{cases} \emptyset, & \text{if } 0, 1 \notin B \\ A, & \text{if } , 1 \in B, 0 \notin B \\ A^c, & \text{if } 0 \in B, 1 \notin B \\ \Omega, & \text{if } 0, 1 \in B \end{cases}$$

$I_A^{-1}(B) \in \mathcal{F}$, I_A is a Borel function, $\sigma(I_A) = \sigma(\{A\})$ is a much smaller σ-field than $\mathcal{F}$ (a power set). We usually use this property to generate a random variable with appropriate σ-field with interested subsets.

Simple functions: $\phi(\omega) = \sum\limits_{i=1}^{n} a_i I_{A_i}(\omega)$ where A_i are measurable sets on Ω and a_i are real numbers. Let $A_1, \cdots, A_k$ be a partition of Ω, $\sigma(\phi) = \sigma(\{A_1, \cdots, A_n\})$. ϕ is also a Borel function, we need the following proposition.

Proposition 1.2.2. *(Jun Shao Proposition 1.4)$(\Omega, \mathcal{F})$ is a measurable space.*

1. *f is Borel iff $f^{-1}((a, \infty)) \to \mathcal{F}, \forall a \in \mathbb{R}$*
2. *If f, g are Borel, fg and $af + bg, a, b \in \mathbb{R}$ are also Borel. f/g is Borel provided $g(\omega) \neq 0$*
3. *Suppose $f_1, f_2, \cdots$ are Borel. Let $L = \{\omega \in \Omega : \lim\limits_{n\to\infty} f_n(\omega) \text{ exists}\}$, then L is a measurable set in Ω and*
$$h(\omega) = \begin{cases} \lim\limits_{n\to\infty} f_n(\omega), & \forall \omega \in L \text{ is Borel.} \\ f_1(\omega), & \forall \omega \notin L \end{cases}$$
4. *f is measurable $(\Omega, \mathcal{F}) \to (\Lambda, \mathcal{G})$ and g is measurable $(\Lambda, \mathcal{G}) \to (\Delta, \mathcal{H})$, the composite function $g \circ f$ is measurable $(\Omega, \mathcal{F}) \to (\Delta, \mathcal{H})$*
5. *Ω is a Borel set in $\mathbb{R}^p$, if f is a continuous function from Ω to $\mathbb{R}^q$, then f is measurable*

The proof could be found in Billingsley Chapter 10 and 13, and Cox Proposition 1.2.1-1.2.3. Based on the above proposition, it is hard to find a non-Borel function.

Let f be a nonnegative Borel function on $(\Omega, \mathcal{F})$, there exists a sequence of simple functions $\{\phi_n\}$ satisfying $0 \leq \phi_1 \leq \phi_2 \cdots \leq f$ and $\lim\limits_{n\to\infty} \phi_n = f$. This is useful for technical proofs.

1.2.2 Induced Measure

We need Proposition 1.2.2-4 for us to construct an induced measure, which is very important in statistics. Just as we said before, the borel set Ω might contain too much useless information for us, and we try to generate a random variable to obtain some interested subsets. Here we use a measure μ (maps a set to a real number) and a measurable function f (maps one set to another set) on $\mathcal{G}$.

Let $(\Omega, \mathcal{F}, \mu)$ be a measure space, $(\Lambda, \mathcal{G})$ a measurable space, and $f : (\Omega, \mathcal{F}) \to (\Lambda, \mathcal{G})$ a measurable function. Define a function $\mu \circ f^{-1}$ on $\mathcal{G}$ by $(\mu \circ f^{-1})(C) = \mu(f^{-1}(C)), C \in \mathcal{G}$. Keep in mind that the measurable function f pulls a σ-fields backwards (i.e. $\sigma(f) \in \mathcal{F}$) but $\mu \circ f^{-1}$ is a measure on the range space $(\Lambda, \mathcal{G})$ How to verify the $\mu \circ f^{-1}$ is a measure? We need to prove Definition 1.1.1 for measure.

Proof.

1. $0 \leq \mu \circ f^{-1}(C) \leq \infty$ is trivial.
2. $\Omega = f^{-1}(\Lambda) = f^{-1}(\Lambda \cup \emptyset) = (f^{-1}(\Lambda) \cup f^{-1}(\emptyset)) = \Omega \cup f^{-1}(\emptyset)$, we can easily find that $f^{-1}(\emptyset) = \emptyset$, $\mu(f^{-1}(\emptyset)) = \mu(\emptyset) = 0$
3. $\mu \circ f^{-1}(C) = \mu \circ f^{-1}(\bigcup\limits_{i=1}^{\infty} C_i) = \mu \circ (\bigcup\limits_{i=1}^{\infty} f^{-1}(C_i)) = \sum\limits_{i=1}^{\infty} (\mu \circ f^{-1})(C_i)$ by Proposition 1.2.1-2.

□

In probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a measurable function $X : \Omega \to \mathbb{R}$ is a real valued random variable, then the induced measure $\mathbb{P} \circ X^{-1}$ is the *distribution of* X*, denoted* P_X. Please keep in mind that the distribution of X still depends on the underlying probability measure $\mathbb{P}$ we use here.

Example 1.2.1.

1. Take $B \subset \mathbb{R}$ a Borel set, an event $[X \in B] = \{\omega \in \Omega : X(\omega) \in B\} = X^{-1}(B)$ and $\mathbb{P}\{\omega \in \Omega : X(\omega) \in B\} = \mathbb{P}[X \in B] = \mathbb{P} \circ X^{-1}(B) = P_X(B)$. P_X is distribution of X.
2. $(\Omega, \mathcal{F}, \mu)$ an arbitrary measure space. $A \in \mathcal{F}$, what is $\mu \circ I_A^{-1}$? Take $B \subset \mathbb{R}$ a Borel set
$$\mu \circ I_A^{-1}(B) = \begin{cases} 0, & \text{if } \{0,1\} \cap B = \emptyset \\ \mu(A), & \text{if } \{0,1\} \cap B = \{1\} \\ \mu(A^c), & \text{if } \{0,1\} \cap B = \{0\} \\ \mu(\Omega), & \text{if } \{0,1\} \cap B = \{0,1\} \end{cases} = \mu(A)\delta_1(B) + \mu(A^c)\delta_0(B)$$
3. m is Lebesgue measure and $f : \mathbb{R} \to \mathbb{R}$. Assume f is strictly increasing and continuous differentiable and $\forall x \in \mathbb{R}$, $Df(x) \neq 0$. $f(-\infty) = -\infty, f(\infty) = \infty$. Then f is 1-1 and onto, f^{-1} exists. $m \circ f^{-1}([a, b]) = \int_a^b D(f^{-1})(x)dx$. This gives $m \circ f^{-1}$ for intervals. This involves formulating Jacobians. Note that with f having some propositions, $f^{-1}([a, b]) = [f^{-1}(a), f^{-1}(b)]$, $m \circ f^{-1}([a, b]) = m[f^{-1}(a), f^{-1}(b)] = f^{-1}(b) - f^{-1}(a)$. So this results follows by the fundamental theorem of calculus.

Remark. Underlying Probability Spaces

1. $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space with $\mathbb{P}(\Omega) = 1$, $\mathbb{P}$ is *probability measure* on real numbers.
2. Elements of $\mathcal{F}$ (Measurable sets) are *events*. $\mathcal{F}$ is the collection of events.
3. Ω is the *sample space.* (underlying space)
4. $\emptyset \in \Omega$, that is to say, $\Omega \in \mathcal{F}$
5. Given any set Ω, the most trivial (smallest) σ-field is $\mathcal{F} = \{\emptyset, \Omega\}$. The power set $\mathbb{P}(\Omega) = \{A : A \subset \Omega\}$ consisting all subsets of Ω is the largest σ-field on Ω. (2^Ω) It is easy to prove that the $\mathcal{F}$ is a σ-field.
   (a) $\emptyset \in \mathcal{F}$
   (b) $A \in \mathcal{F}$ implies $A^c = \Omega \setminus A \in \mathcal{F}$
   (c) $A_1, A_2, \cdots$ in $\mathcal{F}$ implies $\bigcup\limits_n A_n \in \mathcal{F}$

Given any $\mathcal{A} \subset \mathbb{P}(\Omega)$, $\sigma(\mathcal{A})$ is the smallest σ-field containing $\mathcal{A}$.

1.2.3 The Definition of an Integral

See Jun Shao 1.2, Robert Ash 1.5, and Patrick Billingsley Ch15. Expected values of simple random variables and Riemann integrals of continuous functions can be brought together with other related concepts under a general theory of integration. Here we consider a measure $(\Omega, \mathcal{F}, \mu)$ and $f : \Omega \to \bar{\mathbb{R}}$ for extended borel measurable functions. Note that $\infty - \infty$ is undefined, and it's a main issue with when $\int f d\mu$ is undefined. We will define $\int f d\mu$ in three steps.

1. nonnegative simple function: $f = \phi(\omega) = \sum\limits_{i=1}^n a_i I_{A_i}(\omega)$, a canonical form of simple functions. $\int \phi d\mu = \sum\limits_{i=1}^n a_i \mu(A_i)$ for all $a_i > 0$. Note that there is no problem with $\infty - \infty$ in the summation. This result will be ∞ for some i, $a_i > 0$ and $\mu(A_i) = \infty$. ($\int I_A d\mu = \mu(A)$, so $\int \phi d\mu = \sum a_i \int I_{B_i} d\mu$ by linearity property.)

2. nonnegative general function: $\forall \omega \in \Omega, f(\omega) \geq 0$ or $(f \geq 0)$: $\int f d\mu = \sup\{\int \phi d\mu : \phi$ is a simple function with $0 \leq \phi \leq f\}$. (It is equivalent to the collection of nonnegative simple functions). In words, $\int f d\mu$ is the supremum (least upper bound) of all integrals of nonnegative simple functions which are below f. $\int \phi d\mu \leq \int f d\mu$. Note that the set of the simple functions is nonempty since it contains $I_\emptyset = 0$. Also, $\int f d\mu = \infty$ is possible.

3. general function: For general f, $f = f_+($ the positive part of $f) = f_-($ the negative part of $f)$, $f_+, f_- \geq 0$. $f_+(\omega) = \max\{f(\omega), 0\}$, $f_-(\omega) = -(f(\omega))_+ = -\min\{f(\omega), 0\} = \max\{-f(\omega), 0\}$. Note that f_+, f_- are Borel functions, $f(\omega) = f_+(\omega) - f_-(\omega), |f(\omega)| = f_+(\omega) - f_-(\omega)$. Then, $\int f d\mu = \int f_+ d\mu - \int f_- d\mu$. It exists/is defined iff at least one of $\int f_+ d\mu$ and $\int f_- d\mu$ is finite (By step 2), and *integrable* iff both $\int f_+ d\mu$ and $\int f_- d\mu$ are finite.

Finally, we define the integral of f over the set $A \in \mathcal{F}$ as $\int_A f d\mu = \int I_A f d\mu$ (Just follow the Example 1.2.2-3 to prove fI_A is Borel measurable.)

Example 1.2.2. 1. $\Omega = \mathbb{R}, \mu = m$ Lebesgue measure, $\phi = \frac{1}{2} I_{[0,1]} + \frac{3}{4} I_{(1,2]}$, $\int \phi dm = \frac{1}{2} m([0,1]) + \frac{3}{4} m((1,2]) = \frac{1}{2} + \frac{3}{4} = \frac{5}{4} = \int_0^2 \phi(x) dx = \int_{-\infty}^{\infty} \phi(x) dx$ (Note: step function is a type of simple function. Then we can see that the Lebesgue integral is more powerful than Riemann integral with simple integral and sum of rectangles under the curve, that is, a step function.)

2. We could define Riemann integral $\int f(x) dx = \sup\{\int \psi(x) dx : \psi$ is a step function. $0 < \psi < f\}$. It's a subset of simple function, could replace ψ with ϕ. If Riemann integral $\int_{\mathbb{R}} f(x) dx$ exist, it equals Lebesgue integral $\int f dm, f > 0$.

Remark. 1. (Ash P.37) $\{\omega : f_+(\omega) \in A\} = \{\omega : f(\omega) \geq 0, f(\omega) \in A\} \cup \{\omega : f(\omega) < 0, 0 \in A\}$. The first set is $f^{-1}[0, \infty] \cap f^{-1}(A) \in \mathcal{F}$. The first set is $f^{-1}[-\infty, 0)$ if $0 \in A$ and $\emptyset$ if $0 \notin A$. Therefore, $(f_+)^{-1}(A) \in \mathcal{F}$ for each $A \in \mathcal{B}(\hat{\mathbb{R}})$, and similarly for f_-, are both Borel measurable.

2. We also note that it is common to write $d\mu(\omega)$ as $\mu(d\omega)$ as in $\int f(\omega) d\mu(\omega) = \int f(\omega)\mu(d\omega)$. To explain this notation, if $\sum a_i I_{A_i}$ is a simple function approximation to $f(x)$ so that $\int \sum a_i I_{A_i} = \sum a_i \mu(A_i) \doteq \int f d\mu$, then the values a_i will be approximately $f(\omega)$ for some $\omega_i \in A_i$. and the sets A_i will have small measure. If we write $d\omega_i$ to represent the "differential" set A_i, then we obtain notationally $\sum f(\omega_i)\mu(d\omega_i) \doteq \int f d\mu$. The notation $\mu(d\omega)$ is meant to remind us of the measure of these differential sets, which are multiplied by $f(\omega)$ and summed. We will sometimes use this notation when it helps to aid understanding.

3. $\Omega = \{a_1, a_2, \cdots\}$: discrete set (finite or infinite). Take $\mathcal{F} = \mathcal{P}(\Omega)$ as the σ-field, and $\mu = \#$, counting measure. For any $f : \Omega \to \mathbb{R}$ is measurable, $\int f d\# = \sum\limits_i f(a_i)$. It's a classical example of measure theory includes summation and Riemann integral.

4. **unit point mass measure**:

   (a) on a measurable space $(\Omega, \mathcal{F})$, if $\phi = \sum\limits_i c_i I_{A_i}$ is a simple function, then $\int \phi d\delta_x = \sum\limits_i c_i \delta_x(A_i) = \sum\limits_i c_i I_{A_i}(x) = \phi(x) \leq f(x)$. And taking $\phi = f(x) I_{\{x\}}$, we get $\int \phi d\delta_x = \phi(x) = f(x)$.

   (b) If $f \geq 0$, $\sup\{\int \phi d\delta_x = \phi(x) : 0 \leq \phi \leq f\}$. We could get $0 \leq f(x) I_{\{x\}} \leq f$. Assuming $f(x) < \infty, \{x\} \in \mathcal{F}$. $(f(x) = \infty$, then use sequence of simple function $\phi_n = n I_{\{x\}}, \int \phi d\delta_x = \phi(x) = n \to \infty, 0 \leq \phi \leq f$. That is, $\int f d\delta_x = f(x)$.)

   (c) $f = f_+ = f_-, \int f d\delta_x = \int f_+ d\delta_x = \int f_- d\delta_x = f_+(x) - f_-(x) = f(x)$ That is, $\int f d\delta_x = f(x)$.

For a linear combination of unit point mass measure, if $\mu = \sum\limits_i a_i \delta_{x_i} then \int f d\mu = \sum_i a_i f(x_i)$. Let $(x_1, \cdots, x_n)$ be a dataset, and g is a real valued function defined on the space Ω of possible observations. $\int_\Omega g(x) d\hat{P}_n(x) = \frac{1}{n} \sum\limits_{i=1}^{n} g(x_i)$. That is, if P is the true probability model for the experiment, $E[g(X)] = \int g(x)dP(x)$ is used to the sample average $\int g(x)d\hat{P}_n(x)$, and $\int g(x)d\hat{P}_n(x) \to E[g(x)]$ as $n \to \infty$.

5. Show dummy variables: $\int f(\cdot, \theta)dm$, x integrated out here, and it would be as a function of θ.

Riemann Integral: Step function $\psi(x) = \sum\limits_{i=1}^{n} c_i I_{[a_{i-1}, a_i)}(x), a_0 < a_1 < \cdots < a_n$. it's a kind of simple function where A_i are required to be intervals. $[a_0, a_1), \cdots, [a_{n-1}, a_n)$ form a partition of $[a_0, a_n)$ consisting of finitely many intervals. $\Pi = \max\{a_i - a_{i-1} : 1 \leq \ i \leq n\}$ Given Π, the partition and $f : [a, b) \to \mathbb{R}$:

1. upper Riemann integral: $\bar{\int}_a^b f(x)dx = \inf\limits_\Pi \mathcal{U}(f, \Pi)$, $\mathcal{U}(f, \Pi) = \sum\limits_{i=1}^{n} \left(\sup\limits_{[a_{i-1}, a_i]} f \right) (a_{i-1}, a_i) = \int \bar{\psi}_{f,\Pi}(x)dx$, where step function $\bar{\psi}_{f,\Pi}(x) = \sum\limits_{i=1}^{n} \left(\sup\limits_{[a_{i-1}, a_i]} f \right) I_{[a_{i-1}, a_i]}(x)$

2. lower Riemann integral: $\underline{\int}_a^b f(x)dx = \sup\limits_\Pi \mathcal{L}(f, \Pi)$, $\mathcal{L}(f, \Pi) = \sum\limits_{i=1}^{n} \left(\inf\limits_{[a_{i-1}, a_i]} f \right) (a_{i-1}, a_i) = \int \underline{\psi}_{f,\Pi}(x)dx$, where step function $\underline{\psi}_{f,\Pi}(x) = \sum\limits_{i=1}^{n} \left(\inf\limits_{[a_{i-1}, a_i]} f \right) I_{[a_{i-1}, a_i]}(x)$

The Riemann integral exists when $\bar{\int}_a^b f(x)dx = \inf\limits_\Pi \mathcal{U}(f, \Pi) = \underline{\int}_a^b f(x)dx = \sup\limits_\Pi \mathcal{L}(f, \Pi) = \mathcal{R} \int_a^b f(x)dx$. We here try to prove that Lebesgue integral is between lower and upper Riemann integral. $\underline{\int}_a^b f(x)dx = \sup\limits_{\underline{\psi}_{f,\Pi}} \int \underline{\psi}_{f,\Pi}(x)dx$. Step function, as a subclass of simple function, satisfies the simple function property, $0 \leq \psi \leq f$ on $[a, b)$. Since the supremum of a subset is smaller than a superset, $\underline{\int}_a^b f(x)dx \leq \int_{[a,b)} f(x)dm(x)$. Each of the step functions $\bar{\phi}_{f,\Pi}$ that goes into the definition of the upper Riemann integral satisfies $f \leq \bar{\psi}_{f,\Pi}$, so $\int_{[a,b)} f(x)dm(x) \leq \int_{[a,b)]} \bar{\psi}_{f,\Pi} dm(x)$ by proposition below. Then taking infimum over all such step functions gives $\int_{[a,b)} f(x)dm(x) \leq \bar{\int}_{[a,b)]} f(x)dx$.

Example 1.2.3. Not all Lebesgue integrable functions are Riemann integrable. Let $f(x) = I_A(x)$ be the indicator of $A = \{x \in [0, 1) : x \in \mathbb{Q}\}$, then $m(A) = 0$. $[a_{i-1}, a_i) \in \mathbb{R}^+$, $\sup\limits_{[a_{i-1}, a_i)} = 1$ and $\inf\limits_{[a_{i-1}, a_i)} = 0$. Even the upper step on $[0, 1)$ =1 and lower step on $[0, 1)$ =0 is allowable in the partition Π of $[0, 1)$, the upper Riemann integral is 1 and lower Riemann integral =0, and the Riemann integral doesn't exist.

Example 1.2.4. Improper integral: A integral is improper if it is over an infinite interval or ig the function is not bounded, e.g. $\mathcal{R} \int_0^\infty f(x)dx = \lim\limits_{b \to \infty} \mathcal{R} \int_0^b f(x)dx$. An improper Riemann integral may exist, but its Lebesgue integral may *fail to exist.*

$$f(x) = \begin{cases} \frac{1}{n}, & \text{if } 2n - 1 \leq x < 2n \\ -\frac{1}{n}, & \text{if } 2n \leq x < 2n + 1 \\ 0, & \text{if } x < 1 \end{cases}$$

where $\{n : n \in \mathbb{N}\}$. Then for $b > 1$,

$$\int_0^b f(x)dx = \begin{cases} \dfrac{b-(2n-1)}{n}, & \text{if } 2n-1 \leq b < 2n \\ \dfrac{1-(b-2n)}{n}, & \text{if } 2n \leq b < 2n+1 \end{cases}$$

Note that $|\int_0^b f(x)dx| \leq \dfrac{1}{n} \to 0$ as $b \to \infty$. $\mathcal{R}\int_0^\infty f(x)dx = \lim_{b\to\infty} \mathcal{R}\int_0^b f(x)dx = 0$. However, the Lebesgue integral $\int f(x)dm(x) = \int f_+(x)dm(x) + \int f_-(x)dm(x)$ does not exist because $\int f_+(x)dm(x) = \int f_-(x)dm(x) = \sum_{n=1}^{\infty} \dfrac{1}{n} = \infty$, which violates the definition that Lebesgue integral exists/is defined iff at least one of $\int f_+ d\mu$ and $\int f_- d\mu$ is finite (By step 2), and *integrable* iff both $\int f_+ d\mu$ and $\int f_- d\mu$ are finite. That is, we need the improper Riemann integral should be absolutely convergent.

Example 1.2.5. μ is counting measure on $\mathbb{N}$, $\mu(A) = \#$ elements in A, $f : \mathbb{N} \to \mathbb{R}$. $\int f d\mu = \sum_{k=0}^{\infty} f(k) = \lim_{N\to\infty} \sum_{k=0}^{N} f(k)$

1.2.4 Properties of the integral

Please also follow the reference Jun Shao 1.2, Robert Ash 1.5, and Patrick Billingsley CH15.

Proposition 1.2.3. ***(Basic properties of the integral)****: Let $(\Omega, \mathcal{F}, \mu)$ be a measurable space and f, g are extended Borel functions on Ω. See Billingsley Theorem 15.1 and Cox Proposition 1.2.5*

1. *If $f = \sum_i x_i I_{A_i}$ is a nonnegative simple function, $\{A_i\}$ being a finite decomposition of Ω into F-sets, then $\int f d\mu = \sum_i x_i \mu(A_i)$*

   Proof. (In Billingsley 15.1) $\{B_j\}$ a finite decomposition of Ω and let β_j be the infimum of f over B_j. If $A_i \cap B_j \neq \emptyset$, then $\beta_j \leq x_i$. Therefore, $\sum_j \beta_j \mu(B_j) = \sum_{ij} \beta_j \mu((A_i \cap B_j)) \leq \sum_{ij} x_i \mu(A_i \cap B_j) = \sum_i x_j \mu(A_i)$ □

2. ***(Monotonicity)****: $\forall \omega, 0 \leq f(\omega) \leq g(\omega)$, then $\int f d\mu \leq \int g d\mu$.*

   Proof. For $0 \leq f \leq g$ a.e., this would follows Proposition 1.2.4-4. And for general integrable $f \leq g$ a.e., $f_+ \leq g_+$ and $f_- \geq g_-$ a.e. By step 3 in 1.2.3 The definition of integral, $\int f d\mu = \int f_+ d\mu - \int f_- d\mu$, prove it. □

3. *$\forall \omega, 0 \leq f_n(\omega) \uparrow f(\omega)$, then $0 \leq \int f_n d\mu \uparrow \int f d\mu$*

   Proof. See Billingsley. We need to show $\int f d\mu \leq \lim_n \int f d\mu$, equivalent to $\lim_n \int f d\mu \geq S = \sum_{i=1}^{m} v_i \mu(A_i)$ and $v_i = \{\inf_{\omega \in A_i} f(\omega)\}$. First, suppose that S is finite and all the v_i and $\mu(A_i)$ are positive and finite. $\forall 0 < \epsilon < v_i, A_{in} = [\omega \in A_i : f_n(\omega) > v_i - \epsilon]$. $f_n \uparrow f, A_{in} \uparrow A_i$. $\int f_n d\mu \leq \sum_{i=1}^{m} (v_i - \epsilon)\mu(A_{in}) \to \sum_{i=1}^{m} (v_i - \epsilon)\mu(A_i) = S - \epsilon \sum_{i=1}^{m} \mu(A_i)$

Next, suppose only S is finite. Each product of $v_i\mu(A_i)$ is then finite. $i \le m_0$ is positive and $i > m_0$ is 0. (If $m_0 < m, S = 0$, and trivial.) Now $v_i, \mu(A_i)$ are positive and finite for $i \le m_0$ then replace m by m_0 and follow the same procedure to proof.

For an a.e. version, $0 \le f_n \uparrow f$ on a set A with $\mu(A^c) = 0$, then $0 \le f_nI_A \uparrow fI_A$ and $\int f_n d\mu = \int f_nI_Ad\mu \uparrow \int fI_Ad\mu = \int fd\mu$ □

4. ***(Linearity)**: for $a, b \in \mathbb{R}$, $\int afd\mu = a\int fd\mu$. Also, $\int(af + bg)d\mu = a\int fd\mu + b\int gd\mu$*

   Proof. (In Billingsley 15.1) Suppose at first that $f = \sum\limits_i x_iI_{A_i}, g = \sum\limits_j y_jI_{B_j}, af + bg = \sum\limits_{ij}(ax_i + by_j)I_{A_i\cap B_j}$. $\int(af + bg)d\mu = \sum\limits_{ij}(ax_i + by_j)\mu A_i \cap B_j = a\sum\limits_i x_i\mu(A_i) + b\sum\limits_j y_j\mu(B_j) = a\int fd\mu + b\int gd\mu$. Here we check $a = b = 1$ is integrable. $(f + g)_+ - (f + g)_- = f_g = f_+ - f_- + g_+ - g_-$ and $(f+g)_+ + f_- + g_- = (f+g)_- + f_+ + g_+$. All of them are nonnegative. Also, $\int(f+g)_+d\mu + \int f_-d\mu + \int g_-d\mu = \int(f+g)_-d\mu + \int f_+d\mu + \int g_+d\mu$. Finally, $\int(f+g)_+d\mu - \int(f+g)_-d\mu = \int f_+d\mu - \int f_-d\mu + \int g_+d\mu - \int g_-d\mu = \int fd\mu + \int gd\mu$ □

5. *$|\int fd\mu| \le \int|f|d\mu$ if $\int fd\mu$ exists.*

   Proof. Follows the previous properties, it can be easily proved. □

Remark. If A is an event with $\mu(A) = 0$ and the statement $S(\omega)$ (f is continuous at ω) holds for all ω in the complement A^c, then the statement is said to hold a.e.μ (almost everywhere). If μ is a probability measure, a.e. → a.s. (almost surely). e.g. 2 functions f, g. If $f = g$ a.e. (implies $\int fd\mu = \int gd\mu$), $\mu(\{\omega : f(\omega) \ne g(\omega)\}) = 0$.

Proposition 1.2.4. ***Almost everywhere**: If f, g are extended Borel function on $(\Omega, \mathcal{F}, \mu)$*

1. *If $f = 0$ a.e., then $\int fd\mu$=0*

   Proof. Suppose $f = 0$ a.e., if A_i meet $[\omega : f(\omega) = 0]$, then $\inf\limits_{\omega\in A_i} f(\omega) = 0$. Otherwise, $\mu(A_i) = 0$. Hence, $\sum\limits_i \left[\inf\limits_{\omega\in A_i} f(\omega)\right]\mu(A_i) = 0$ □

2. *If $[\omega : f(\omega) > 0]$, $\int fd\mu > 0$*

   Proof. If $A_\epsilon \in [\omega : f(\omega) \ge \epsilon], A_\epsilon \uparrow [\omega : f(\omega) > 0]$ as $\epsilon \downarrow 0$. There exists a $\epsilon > 0$ for which $\mu(A_\epsilon) > 0$. Decomposing Ω into A_ϵ and its complements shows that $\int fd\mu \ge \epsilon\mu(A_\epsilon) > 0$ □

3. *If $\int fd\mu < \infty$ then $f < \infty$ a.e.*

   Proof. If $\mu[f = \infty] > 0$, decompose Ω into $[f = \infty]$ and its complement: $\int fd\mu \ge \infty\cdot\mu[f = \infty] > 0$. □

4. *$f \le g$ a.e., then $\int fd\mu \le \int gd\mu$, provided the integral exist.*

   Proof. Let $G = [f \le g]$, for any finite decomposition $[A_1, \cdots, A_m]$ of Ω, $\sum\left[\inf\limits_{A_i} f\right]\mu(A_i) = \sum\left[\inf\limits_{A_i} f\right]\mu(A_i\cap G) \le \sum\left[\inf\limits_{A_i\cap G} f\right]\mu(A_i \cap G) \le \sum\left[\inf\limits_{A_i\cap G} g\right]\mu(A_i \cap G) \le \int gd\mu$. It also shows that if $f = g$ a.e. (implies $\int fd\mu = \int gd\mu$). □

5. *$f \geq 0, \mu$ a.e. and $\int f d\mu = 0$, then $f = 0, \mu$-a.e.*

   Proof. $A = \{f > 0\}, A_n = \{f > \dfrac{1}{n}\}, n = 1, 2, \cdots$. Then $A_n \subset A$ for any n. By Proposition 1.1.1-3 and Definition 1.1.1-3, $\lim\limits_{n\to\infty} A_n = \cup A_n = A$, $\lim\limits_{n\to\infty} \mu(A_n) = \mu(A)$. And by 1. and Proposition 1.2.3, $\dfrac{1}{n}\mu(A_n) = \int \dfrac{1}{n} I_{A_n} d\mu \leq \int f I_{A_n} d\mu \leq \int f d\mu = 0$. □

Remark. 1. Some direct consequences of Proposition 1.2.4-1 are $|\int f d\mu| \leq \int |f| d\mu$; if $f \geq 0$ a.e., then $\int f d\mu \geq 0$. and if $f = g$ a.e. (implies $\int f d\mu = \int g d\mu$), $\mu(\{\omega : f(\omega) \neq g(\omega)\}) = 0$.

2. $\int \lim\limits_{n\to\infty} f_n d\mu = \lim\limits_{n\to\infty} \int f_n d\mu$ where $\{f_n : n = 1, 2, \cdots\}$ is a sequence of Borel functions. We only require $\lim\limits_{n\to\infty} f_n$ (also a Borel function by proposition 1.2.2-3) exists a.e.

   Example 1.2.6. $\{f_n : n = 1, 2, \cdots\}$ is a sequence of Borel functions on Lebesgue measure $(\mathbb{R}, \mathcal{B})$. Let $f_n(x) = nI_{[0,\frac{1}{n}]}(x)$. Then $\lim\limits_{n\to\infty} f_n(x) = 0, \forall x \setminus \{0\}$ ($\lim\limits_{n\to\infty} f_n = \infty$ at $x = 0$). Since $m(\{0\}) = 0$, we may say $f_n \to 0$ for m-a.e., $\lim\limits_{n\to\infty} \int f_n dm = 1$, $\int \lim\limits_{n\to\infty} f_n dm = 0$. That is, the interchange is not feasible here.

3. And, we need Fatou Lemma here, and for Monotone Convergence Theorem and Dominated Convergence Theorem.

Theorem 1.2.5. *1.* ***Fatou's Lemma****: For nonnegative f_n, $\int \liminf\limits_n f_n d\mu \leq \liminf\limits_n \int f_n d\mu$*

   Proof. If $g_n = \inf\limits_{k\geq n} f_k$, then $0 \leq g_n \uparrow g = \liminf\limits_n f_n$ and by Proposition 1.2.3 and 1.2.4, $\int f_n d\mu \geq \int g_n d\mu \to \int g d\mu$ □

2. ***Monotone Convergence Theorem****: $0 \leq f_1 \leq f_2 \leq \cdots \leq f_n \leq \cdots, (0 \leq f_n \uparrow f), f(\omega) = \lim\limits_{n\to\infty} f_n(\omega)$ a.e., then*

$$\lim \int f_n d\mu = \int \lim f_n d\mu.$$

   Proof. By Proposition 1.2.4-4, there exists

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

   Let ϕ be a simple function, $0 \leq \phi \leq f$ and let $A_\phi = \{\phi > 0\}$. Suppose that $\mu(A_\phi) = \infty$, then $\int f d\mu = \infty$. Let $a = \frac{1}{2} \min\limits_{\omega\in A_\phi}$ and $A_n\{f_n > a\}$. Then $a > 0, A_1 \subset A_2 \subset \cdots$, and $A_\phi \subset \cup A_n$. By Proposition 1.1.1-1,

$$\mu(A_n) \to \mu(\cup A_n) \geq \mu(A_\phi) = \infty \text{ and } \int f_n d\mu \geq \int_{A_n} f_n d\mu \geq a\mu(A_n) \to \infty.$$

   Suppose now $\mu(A_\phi) < \infty$, by Egoroff's Theorem, for any $\epsilon > 0$, there is $B \subset A_\phi$ with $\mu(B) < \epsilon$ s.t. $f_n \to f$ uniformly on $A_\phi \cap B^c$. Hence,

$$\int f_n d\mu \geq \int_{A_\phi\cap B^c} f_n d\mu \to \int_{A_\phi\cap B^c} f d\mu \geq \int_{A_\phi\cap B^c} \phi d\mu = \inf \phi d\mu - \int_B \phi d\mu \geq \int \phi d\mu - \epsilon \max_\omega \phi(\omega).$$

Since ϵ is arbitrary, $\lim\limits_{n\to\infty}\int f_n d\mu \geq \int \phi d\mu$. Since ϕ is also arbitrary, by Definition of Integral 2,

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

□

3. ***Dominated Convergence Theorem*** *Assume $f_n \to f$ a.e., and $\exists g$ dominating function s.t. $\forall n, |f_n| \leq g$ a.e.& $\int g d\mu < \infty$ (integrable), then $\int \lim\limits_{n\to\infty} f_n d\mu = \lim\limits_{n\to\infty}\int f_n d\mu$*

   Proof. Applying Fatou's Lemma to functions $g + f_n$ and $g - f_n$. We obtain that

$$\begin{aligned}\int g d\mu + \int \liminf_n f_n d\mu &= \int \liminf_n (g + f_n) d\mu \\ &\leq \liminf_n \int (g + f_n) d\mu = \int g d\mu + \liminf_n \int f_n d\mu\end{aligned}$$

   and

$$\begin{aligned}\int g d\mu - \int \limsup_n f_n d\mu &= \int \limsup_n (g - f_n) d\mu \\ &\leq \limsup_n \int (g - f_n) d\mu = \int g d\mu - \limsup_n \int f_n d\mu.\end{aligned}$$

   Therefore,

$$\int \liminf_n f_n d\mu \leq lim\inf_n \int f_n d\mu \leq \limsup_n \int f_n d\mu \leq \int \limsup_n f_n d\mu.$$

   Now use asusmptions $f_n \to f$ a.e., f is dominated by g. Since g is integrable, these results imply that

$$\int f d\mu \leq \liminf_n \int f_n d\mu \leq \limsup_n \int f_n d\mu \leq \int f d\mu.$$

   That is, Fatou's Lemma implies DCT. □

Remark. Note that for each ω, $\lim\limits_{n\to\infty}$ exists, but may be $+\infty$. In DCT, there is no unique g dominating function, we could choose a convenient one.

Example 1.2.7. Back to $\int f d\delta_x$ (unit point mass measure), recall that we assume $\{x\} \in \mathcal{F}$ e.f. $\mathcal{F} = \{\phi, \Omega\}$. We don't need this assumptions by the following explanation. For any $f \geq 0$ have simple functions ϕ_n s.t. $0 \leq \phi_n \uparrow f$. We have $\int \phi_n d\mu \uparrow \int f d\mu$. Thus, we can "calculate" $\int f d\mu$ using these simple functions instead of $\sup\{\int \phi d\mu : 0 \leq \phi \leq f\}$. In particular, $\int \phi d\delta_x = \phi(x)$, $0 \leq \phi \uparrow f \implies \int \phi_n d\delta_x = \phi_n(x) \to \int f d\delta_x$

Proposition 1.2.6. *Simple function approximation: Let $f : (\Omega, \mathcal{F}) \to (\bar{\mathbb{R}}, \bar{\mathcal{B}})$, $f \geq 0$, then $\exists \phi_n$, a sequence of simple functions s.t. $0 \leq \phi_n \uparrow f$ and $\forall n, |\phi_n| \leq |f|$. If $f \geq 0$ then we may take $\phi_n \geq 0$ for all n. Further, if μ is a measure of $(\Omega, \mathcal{F})$ and $\int f d\mu$ is defined, then $\int \phi d\mu \to \int f d\mu$.*

Theorem 1.2.7. ***Measures Defined by Densities****: let $f : (\Omega, \mathcal{F}, \mu) \to (\bar{\mathbb{R}}, \bar{\mathcal{B}})$ be nonnegative, and put $\nu(A) = \int_A f d\mu, A \in \mathcal{F}$. Show that ν is a measure on $(\Omega, \mathcal{F})$*

1. *$0 \leq \nu(A) \leq \infty$? Yes, by monotonicity*
2. *$\nu(\phi) = 0$? Yes, $I_\phi \equiv 0, I_\phi f \equiv 0$*

3. *Take $\omega \in \Omega$, $I_U(\omega) = 1$ iff $\exists n$ s.t. $\omega \in A_n$. But there only one such n so that $\sum\limits_n I_{A_n}(\omega) = 1$, $I_{\bigcup\limits_n A_n}(\omega) = 0$ iff all $I_{A_n}(\omega) = 0$. Note that $I_{\bigcup\limits_n A_n} = \sum\limits_n I_{A_n}$ by disjointness. Thus, $\nu(\bigcup\limits_n A_n) = \int \left(\sum\limits_{n=1}^{\infty} I_{A_n} f d\mu\right)$*

In the context of this theorem, the function f is called the density of ν with respect to (w.r.t.) μ. Statistical models have possible distributions for Y, a random vector with possible distributions. All having densities w.r.t. some μ, $f(y, \theta)$ where θ is unknown parameter, $\int f(y, \theta) d\mu(y) = 1$. Most of the probability measures we use in practice will be constructed through densities, either w.r.t. Lebesgue measure (so-called continuous distributions) or w.r.t. counting measure (discrete distributions). We will later provide necessary and sufficient conditions for when one measure has a density w.r.t another measure (the Radon-Nikodym theorem). This result has many ramifications in probability and statistics.

Theorem 1.2.8. ***Change of variables**: Suppose $f : (\Omega, \mathcal{F}, \mu) \to (\Lambda, \mathcal{G})$ and $g : (\Lambda, \mathcal{G}) \to (\bar{\mathbb{R}}, \bar{\mathcal{B}})$. Then*

$$\int_\Omega (g \circ f)(\omega) d\mu(\omega) = \int_\Lambda g(\lambda) d(\mu \circ f^{-1})(\lambda)$$

, if either integral is defined, then so is the other and the two are equal.

Proof. First assume g is a nonnegative simple function, say

$$g(\lambda) = \sum_{i=1}^{n} a_i I_{A_i}(\lambda)$$

where $a_i \geq 0, \forall i$. Then $g \circ f \geq 0$ so both integral exist. Now

$$\int_\Omega (g \circ f)(\omega) d\mu(\omega) = \int \sum a_i I_{A_i}(f(\omega)) d\mu(\omega) = \sum a_i \int I_{A_i}(f(\omega)) d\mu(\omega)$$

by linearity property. Note that $I_A(f(\omega)) = 1$ iff $f(\omega) \in A$ iff $\omega \in f^{-1}(A)$ iff $I_{f^{-1}(A)}(\omega) = 1$, so $I_A \circ f = I_{f^{-1}(A)}$. Using this,

$$\begin{aligned}\int_\Omega (g \circ f)(\omega) d\mu(\omega) &= \sum a_i \int I_{A_i}(f(\omega)) d\mu(\omega) \\ &= \sum a_i \int I_{f^{-1}(A)}(\omega) d\mu(\omega) = \sum a_i \mu(f^{-1}(A_i)) \\ &= \sum a_i (\mu \circ f^{-1})(A_i) = \sum a_i \int I_{A_i} d(\mu \circ f^{-1}) \\ &= \int g d(\mu \circ f^{-1}).\end{aligned}$$

Now suppose that $g \geq 0$. Then both integrals are still defined. Let ϕ_n be simple functions with $0 \leq \phi_n \uparrow g$ by Proposition 1.2.6. Then

$$\int \phi_n d(\mu \circ f^{-1}) \to \int g d(\mu \circ f^{-1})$$

by MCT. $I_A \circ f = I_{f^{-1}(A)}$ shows that $\phi_n \circ f$ are nonnegative simple functions on Ω, $0 \leq \phi_n \circ f \uparrow g \circ f$, so

$$\int (\phi_n \circ f) d\mu \to \int (g \circ f) d\mu$$

by MCT. Since $\int(\phi_n \circ f)d\mu = \int \phi_n d(\mu \circ f^{-1})$ by the first part of proof, we have $\int(g \circ f)d\mu = \int gd(\mu \circ f^{-1})$

Now let g be a general extended Borel function on Λ and consider g_+, g_-. $(g\circ f)_+ = g_+\circ f$ and $(g\circ f)_- = g_-\circ f$, by previous part of the proof and the nonnegative function,

$$\int (g \circ f)_+ d\mu = \int g_+ d(\mu \circ f^{-1}), \int (g \circ f)_- d\mu = \int g_- d(\mu \circ f^{-1}).$$

Hence, if say $\int(g \circ f)_- d\mu < \infty$, so that the $\int_\Omega (g \circ f)(\omega)d\mu(\omega)$ is defined, then $\int g_- d(\mu \circ f^{-1}) < \infty$ and $\int_\Lambda g(\lambda)d(\mu \circ f^{-1}(\lambda)$ is defined, and

$$\int g \circ f d\mu = \int (g \circ f)_+ d\mu - \int (g \circ f)_- d\mu = \int g_+ d(\mu \circ f^{-1}) - \int g_- d(\mu \circ f^{-1}) = \int gd(\mu \circ f^{-1}).$$

A similar argument applies if $\int(g \circ f)_+ d\mu < \infty$, which is the other way $\int_\Omega (g \circ f)(\omega)d\mu(\omega)$ can exist. $\int_\Lambda g(\lambda)d(\mu \circ f^{-1}(\lambda)$ exists just in case one of $g_+ d(\mu \circ f^{-1}) < \infty$ or $\int g_- d(\mu \circ f^{-1}) < \infty$, and the proof goes through without difficulty again. □

Remark. Start with simple functions, use Proposition 1.2.6 and Theorem 1.2.5 to extend to nonnegative functions, and finally to general functions using the decomposition into positive and negative parts.

Example 1.2.8. We briefly indicate the importance of Theorem 1.2.8. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and X a r.v. defined thereon. If $E[X] = Xd\mathbb{P}$ exists, then $\int fE[X]$ by $\int_R xdP_X(x)$ where $P_X = \mathbb{P} \circ X^{-1}$ is the distribution of X. Thus, we compute an integral over the real line rather than an integral over the original probability space. If $X : \Omega \to \mathbb{R}, Y : \Omega \to \mathbb{R}, g : \mathbb{R} \to \mathbb{R}$, then $E[g(X)]$ is typically computed as $\int_\mathbb{R} g(x)dP_X(x)$ rather then $\int_\mathbb{R} ydP_{g(X)}(y)$, i.e. one integrates w.r.t. the distribution of the original r.v. X rather than w.r.t. the distribution of $g(X)$. Theorem 1.2.8 is used so often by statisticians without giving it any thought that it is sometimes referred to as "the law of the unconscious statistician." It should be noted that calculation of $\mu \circ f^{-1}$ may be complicated, e.g. involving Jacobians, a subject treated in Chapter 2, Section 2.4.

Theorem 1.2.9. *interchange of differentiation and integration*: *Let $(\Omega, \mathcal{F}, \mu)$ and suppose $g(\omega, \theta)$ is a real valued function on the cartesian product space $\Omega \times (a, b)$ where (a, b) is a finite open interval in $\mathbb{R}$. Assume g satisfies:*

1. *For each fixed $\theta \in (a, b)$, the function $f_\theta(\omega) = g(\omega, \theta)$ is a Borel function of ω and $\int |g(\omega, \theta)|d\omega < \infty$*
2. *a null set N s.t. $\forall \omega \notin B$, $\dfrac{\partial g(\omega, \theta)}{\partial \theta}$ exist for all $\theta \in (a, b)$*
3. *an integrable function $G : \Omega \to \bar{\mathbb{R}}$ s.t. $\forall \omega \notin N$ and all $\theta \in (a, b)$, $\left|\dfrac{\partial g}{\partial \theta}(\omega, \theta)\right| \le G(\omega)$*

Then for each fixed $\theta \in (a, b)$, $\dfrac{\partial g}{\partial \theta}(\omega, \theta)$ is integrable w.r.t μ and $\dfrac{d}{d\theta} \int_\Omega g(\omega, \theta)d\mu(\omega) = \int_\Omega \dfrac{\partial g}{\partial \theta}(\omega, \theta)d\mu(\omega)$

Proof. Let $H(\theta) = \int g(\omega, \theta)d\mu(\omega)$. Suppose $\omega \notin N$, then by the mean value theorem, if $\theta \in (a, b)$ and $\theta + \delta \in (a, b)$, then $\dfrac{g(\omega, \theta + \delta) - g(\omega, \theta)}{\delta} = \dfrac{\partial g}{\partial \theta}(\omega, \theta + \alpha\delta)$ for some $\alpha \in [0, 1]$, and in particular $\forall \omega \notin N$, $\left|\dfrac{g(\omega, \theta + \delta) - g(\omega, \theta)}{\delta}\right| \le G(\omega)$. Now let η_n be any sequence in $\mathbb{R}$ converging to 0. Then by Proposition 1.2.4, $\dfrac{H(\theta + \eta_n) - H(\theta)}{\eta_n} = \int \dfrac{g(\omega, \theta + \eta_n) - g(\omega, \theta)}{\eta_n} d\mu(\omega)$. Thus, we have for each fixed $\theta \in (a, b)$, the sequence of functions $f_n(\omega) = \dfrac{g(\omega, \theta + \eta_n) - g(\omega, \theta)}{\eta_n}$ converges μ-a.e. to $\dfrac{\partial g(\omega, \theta)}{\partial \theta}$. And by $\left| \dfrac{g(\omega, \theta + \delta) - g(\omega, \theta)}{\delta} \right| \leq G(\omega)$, $|f_n| \leq G$, μ-a.e., and G is μ-integrable by assumption. Hence, by DCT, $\int f_n d\mu \to \int \left[\dfrac{\partial g(\omega, \theta)}{\partial \theta} \right] d\mu$, i.e. $\lim_{n \to \infty} \dfrac{H(\theta + \eta_n) - H(\theta)}{\eta_n} = \int_\Omega \dfrac{\partial g}{\partial \theta}(\omega, \theta) d\mu(\omega)$. Since the sequence $\eta_n \to 0$ was arbitrary, it follows that $\lim_{\delta \to 0} \dfrac{H(\theta + \eta_n) - H(\theta)}{\delta} = \int_\Omega \dfrac{\partial g}{\partial \theta}(\omega, \theta) d\mu(\omega)$. The equation just before this equation states that $H(\theta)$ is differentiable and the derivative is the right hand side. □

1.3 Measures on Product Spaces

The Cartesian product of sets Γ_i is defined as the set of all $(a_1, a_2, \cdots)$, $a_i \in \Gamma_i$, and is denoted by $\prod\limits_{i\in\{1,2,\cdots\}} \Gamma_i = \Gamma_1 \times \Gamma_2 \times \cdots$. Given measurable spaces $(\Omega_i, \mathcal{F}_i), i = 1, 2, \cdots$. Since $\prod\limits_{i\in\{1,2,\cdots\}} \mathcal{F}_i$ is not necessarily a σ-field, $\sigma\left(\prod\limits_{i\in\{1,2,\cdots\}} \mathcal{F}_i\right)$ is called the *product σ-field* on the *product space* $\prod\limits_{i\in\{1,2,\cdots\}} \Omega_i$ and $\left(\prod\limits_{i\in\{1,2,\cdots\}} \Omega_i, \prod\limits_{i\in\{1,2,\cdots\}} \mathcal{F}_i\right)$ is denoted by $\prod\limits_{i\in\{1,2,\cdots\}} (\Omega_i, \mathcal{F}_i)$

A measure space $(\Lambda, \mathcal{G}, \mu)$ with $\Lambda \in \mathcal{B}_n$ and $\mathcal{G} = \{B \cap \Lambda : B \in \mathcal{B}_n\}$ is called a **Euclidean Space**, which is usually used in most of the statistical application.

1.3.1 Definitions and Results

Definition 1.3.1. A measure space $(\Omega, \mathcal{F}, \mu)$ is called a σ-finite iff there is an infinite sequence $A_1, A_2, \cdots$ in $\mathcal{F}$ s.t.

1. $\mu(A_i) < \infty$
2. $\bigcup\limits_{i=1}^{\infty} A_i = \Omega$

Any finite measure (e.g. probability measure) is σ-finite, since $\mathbb{R} = \cup A_n$. The counting measure is σ-finite iff Ω is countable.

Example 1.3.1. $f \equiv \infty$, $\nu(A) = \int A d\mu$. This fives non-σ-finite measure where

$$\nu(A) = \begin{cases} \infty, & \text{if } \mu(A) > 0 \\ 0, & \text{if } \mu(A) = 0 \end{cases}$$

Theorem 1.3.1. ***Product Measure Theorem****: Let $(\Omega_i, \mathcal{F}_i, \mu_i)$ be measure spaces with σ-finite measures (that is, they are σ-finite measure spaces). There exists a unique σ-finite measure on the product σ-field $\sigma\left(\prod\limits_{i\in\{1,2,\cdots\}} \mathcal{F}_i\right)$, called the product measure and denoted by $\mu_1 \times \mu_2 \times \cdots$ s.t. $\mu_1 \times \cdots \times \mu_k(A_1 \times \cdots A_k) = \mu_1(A_1) \times \cdots \mu_k(A_k)$, for all $A_i \subseteq \Omega_i(A_i \in \mathcal{F}_i)$*

Example 1.3.2. The usual length of an interval $[a, b] \subset \mathbb{R}$ is the same as the Lebesgue measure of $[a, b]$. Consider a measurable rectangle $[a_1, b_1] \times [a_2, b_2] \subset \mathbb{R}^2$, the usual area of $[a_1, b_1] \times [a_2, b_2]$ is $(b_1 - a_1)(b_2 - a_2) = m([a_1, b_1])m([a_2, b_2])$, i.e. the product of the Lebesgue measures of two intervals $[a_1, b_1]$ and $[a_2, b_2]$. Note that $[a_1, b_1] \times [a_2, b_2]$ is a measurable set by the definition of the product σ-field. We need the above definition and theorem to verify the above calculation.

In $\mathbb{R}^2$, there is a unique measure, the product measure $m \times m = m^2$, $m^2([a_1, b_1] \times [a_2, b_2]) = (b_1 - a_1)(b_2 - a_2) = m([a_1, b_1])m([a_2, b_2])$ on 2-dimensional Lebesgue measure (there also exist n-dimensional Lebesgue measure for $n \in \mathbb{N}$ which is similarly defined.)

Example 1.3.3. μ_i are counting measures on $\mathbb{N}$, $\mu_1 \times \mu_2$ is counting measure on $\mathbb{N} \times \mathbb{N}$. Take $A_1, A_2 \subseteq \mathbb{N}$, the number of elements in $A_1 \times A_2$ is the number of elements in $A_1\times$ the number of elements in A_2

Note. 1. Trivial isomorphism: $\Omega_1 \times \Omega_2 \cong \Omega_2 \times \Omega_1$, there exists one to one mapping $(\omega_1, \omega_2) \mapsto (\omega_2, \omega_1)$.

2. Higher dimensional Lebesgue measure m^n, we have n-dimensional *volume*

1.3.2 Integration with Product Measures

Theorem 1.3.2. ***Fubini's theorem (Fubini-Tonelli's theorem):*** *Let μ_i be a σ-finite measure on $(\Omega_i, \mathcal{F}_i), i = 1, 2,$ and let f be a Borel function on $\prod_{i-1}^{2}(\Omega_i, \mathcal{F}_i)$. Suppose that f is either nonnegative or integrable w.r.t. $\mu_1 \times \mu_2$. ($\int f d(\mu_1 \times \mu_2)$ exists) Then $g(\omega_2) = \int_{\Omega_1} f(\omega_1, \omega_2) d\mu_1(\omega_1)$ exists a.e. μ_2 and defines a Borel function on Ω_2 whose integral w.r.t. μ_2 exists, and*

$$\int_\Omega f(\omega) d\mu(\omega) = \int_{\Omega_1 \times \Omega_2} f(\omega_1, \omega_2) d(\mu_1 \times \mu_2)(\omega_1, \omega_2) = \int_{\Omega_2} \left[\int_{\Omega_1} f(\omega_1, \omega_2) d\mu_1(\omega_1) \right] d\mu_2(\omega_2) \tag{1.1}$$

This result can be extended to the integral w.r.t. the product measure on $\prod_{i=1}^{k}(\Omega_i, \mathcal{F}_i)$ for $k \in \mathbb{N}$

Example 1.3.4. Let $\Omega_1 = \Omega_2 = \{0, 1, 2, \cdots\}$ and $\mu_1 = \mu_2 = \#$. One can check that $\# \times \#$ on $\mathbb{N} \times \mathbb{N}$ is $\#$ on $\mathbb{N}^2$. A function $f(i, j) \geq 0$ or $\int |f| d(\mu_1 \times \mu_2) < \infty$ defines a double sequence, then

$$\int f d(\mu_1 \times \mu_2) = \int f d\# = \int \left[\int f(i, j) d\#(i) \right] d\#(j) = \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} f(i, j) = \sum_{j=0}^{\infty} \sum_{i=0}^{\infty} f(i, j)$$

Thus we have shown a well known fact from advanced calculus: if a double series is absolutely summable (i.e. $\sum_{i=0}^{\infty} \sum_{j=0}^{\infty} |f(i, j)| < \infty$ holds), then it can be summed in either order. In fact, by Fubini's theorem, it suffices for either the sum of the positive terms to be finite or the sum of the negative terms to be finite. That some condition is required for interchanging the order of the summations

1.3.3 Random Vectors and Stochastic Independence

A function $\underline{X}$: $(\Omega, \mathcal{F}, \mathbb{P}) \to \mathbb{R}^n$ is a n dimensional random vector. $\mathbb{P} \circ \underline{X}^{-1}$ on $\mathbb{R}^n$ is called the *distribution* or *law* if $\underline{X}$ and is denoted $P_{\underline{X}}$ or Law$[\underline{X}]$. We will write a vector as a column vector or as an ordered n–tuple, i.e.

$$(x_1, x_2, \cdots, x_n) = \begin{bmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{bmatrix}$$

We need to use the r.h.s. of this last equation wherein $\underline{x}$ is represented as an $n \times 1$ matrix whenever we do matrix operations. The component functions $(X_1, X_2, \cdots, X_n)$ of a random n–vector $\underline{X}$ are random variables, and their distributions on $\mathbb{R}^1$ are referred to as *marginal distributions.* The distribution of $\underline{X}$ on $\mathbb{R}^n$ is sometimes referred to as the *joint distribution* of $(X_1, X_2, \cdots, X_n)$.

The random variables $(X_1, X_2, \cdots, X_n)$ are said to be (jointly) independent iff for all $B_1, B_2, \cdots, B_n \in \mathcal{B}$, $\mathbb{P}\{X_1 \in B_1, \cdots, X_n \in B_n\} = \prod_{i=1}^{n} \mathbb{P}\{X_i \in B_i\}$. This definition extends to arbitrary random elements $(X_1, X_2, \cdots, X_n)$. This last displayed equation is equivalent to $\prod_{i=1}^{n} \mathbb{P}\{X_i \in B_i\} = (\mathbb{P} \circ \underline{X}^{-1})(\prod_{i=1}^{n} B_i) = \prod_{i=1}^{n} (\mathbb{P} \circ \underline{X}^{-1})(B_i)$

Proposition 1.3.3. *Let $\underline{X} = (X_1, X_2, \cdots, X_n)$ be a random vector. Then $(X_1, X_2, \cdots, X_n)$ are independent iff $Law[\underline{X}] = \prod_{i=1}^{n} Law[X_i]$*

Proof. Suppose $(X_1, X_2, \cdots, X_n)$ are jointly independent, so $\mathbb{P}\{X_1 \in B_1, \cdots, X_n \in B_n\} = \prod_{i=1}^{n} \mathbb{P}\{X_i \in B_i\}$ holds for all $B_1, B_2, \cdots, B_n \in \mathcal{B}$. Note that the l.h.s. of $\mathbb{P}\{X_1 \in B_1, \cdots, X_n \in B_n\}$ is the joint distribution $P =$Law$[\underline{X}]$ evaluated at the rectangle set $B_1 \times B_2 \times \cdots \times B_n$, and this equals the product of the corresponding measures of the factor sets. Since this holds for arbitrary rectangle sets, it follows that $\mathbb{P} \circ \underline{X}^{-1} = \prod(P \circ X_i^{-1})$ by uniqueness in the Product Measure Theorem. Conversely, if $\mathbb{P} \circ \underline{X}^{-1} = \prod(\mathbb{P} \circ X_i^{-1})$, then $\mathbb{P}\{X_1 \in B_1, \cdots, X_n \in B_n\} = \prod_{i=1}^{n} \mathbb{P}\{X_i \in B_i\}$ holds for all $B_1, B_2, \cdots, B_n$ by the definition of the product measure, and hence $X_1, X_2, \cdots, X_n$ are jointly independent. □

We say $X_1, X_2, \cdots, X_n$ are pairwise independent iff for all $i \neq j$, the pair X_i and X_j are independent. Joint independence implies pairwise independence, but the converse is false. The following result gives some useful consequences of independence.

Theorem 1.3.4. *Let X and Y be random elements defined on a common probability space.*

1. *For all appropriate sets A, B, $\mathbb{P}[X \in A \& Y \in B] = \mathbb{P}[X \in A]\mathbb{P}[Y \in B]$, $\mathbb{P}[(X,Y) \in A \times B] = P_{XY}(A \times B) = P_X(A)P_Y(B)$ by property of product measure.*
2. *If X and Y are independent, then so are $g(X)$ and $h(Y)$ where g and h are appropriately measurable functions.*
3. *If g and h in (1) are real-valued, then $E[g(X)h(Y)] = E[g(X)]E[h(Y)]$*

Proof. 1. Skip.

2. X, Y are independent if and only if $X^{-1}(A)$ and $Y^{-1}(B)$ are independent for all Borel measurable set A, B. Consider $(g \circ X)^{-1}(A) = X^{-1}(g^{-1}(A))$ (which is equivalent to $[X \in g^{-1}(A)]$) and $(h \circ Y)^{-1}(A) = Y^{-1}(h^{-1}(B))$ (which is equivalent to $[Y \in h^{-1}(B)]$). We could know that $P[g(X) = A \& h(Y) = B] = P[X \in g^{-1}(A) \& Y \in h^{-1}(B)] = P[X \in g^{-1}(A)]P[Y \in h^{-1}(B)] = P[g(X) = A]P[h(Y) = B]$
3. Let $P = $ Law$[g(X)]$ and $Q = $ Law$[h(Y)]$. By (1), Law$[g(X), h(Y)] = P \times Q$. By Change of Variable theorem (Theorem 1.2.8)

$$\begin{aligned} E[g(X)h(Y)] &= \int_{\mathbb{R}^2} g \cdot h d(P \times Q)(g,h) = \int_{\mathbb{R}} \left[\int_{\mathbb{R}} g \cdot h dP(g) \right] dQ(h) \qquad \text{(by Fubini's theorem)} \\ &= \int_{\mathbb{R}} \left[\int_{\mathbb{R}} g dP(g) \right] h dQ(h) \qquad \text{(by Proposition 1.2.3-(4), } h \text{ can be factored out of the integral of } P(g)) \\ &= \left[\int_{\mathbb{R}} g dP(g) \right] \cdot \left[\int_{\mathbb{R}} h dQ(h) \right] \qquad \text{(by Proposition 1.2.3-(4), } \int g dP(g) \text{ is a constant)} \\ &= E[g(X)] \cdot E[h(Y)] \end{aligned}$$

□

Definition 1.3.2. Events A, B (subsets of underlying sample space) are independent iff I_A, I_B are independent.

If A, B are independent events, $\mathbb{P}[A \cap B] = E[I_{A\cap B}] = E[I_A I_B]$. $I_A(\omega)I_B(\omega) = 1$ iff ($\omega \in A \& \omega \in B$) iff ($\omega \in A \cap B$) iff $I_{A\cap B}(\omega) = 1$. $A \& B$ are independent events iff I_A, I_B are independent random variables (Bernoulli random variables) iff $P_{I_A} \times P_{I_B} = P_{I_A I_B}$.

$$\begin{aligned}
P_{I_A} &= \mathbb{P}(A^c)\delta_0 + \mathbb{P}(A)\delta_1 \\
P_{I_A I_B} &= \mathbb{P}(A \cap B)\delta_{(1,1)} + \mathbb{P}(A \cap B^c)\delta_{(1,0)} + \mathbb{P}(A^c \cap B)\delta_{(0,1)} + \mathbb{P}(A^c \cap B^c)\delta_{(0,0)} \\
P_{I_A} \times P_{I_B} &= (\mathbb{P}(A^c)\delta_0 + \mathbb{P}(A)\delta_1) \times (\mathbb{P}(B^c)\delta_0 + \mathbb{P}(B)\delta_1) \\
&= \mathbb{P}(A^c)(B^c)(\delta_0 \times \delta_0) + \mathbb{P}(A)(B^c)(\delta_1 \times \delta_0) + \mathbb{P}(A^c)(B)(\delta_0 \times \delta_1) + \mathbb{P}(A)(B)(\delta_1 \times \delta_1)
\end{aligned}$$

That is, coefficient of $\delta_{(1,1)}$ match. i.e. $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$, i.e. independent of I_A, I_B → independent of events A, B. This also implies [Independent of A, B^c] & [Independent of A^c, B] & [Independent of A^c, B^c]. Also implies for four coefficient of $\delta_{(0,0)}$,$\delta_{(0,1)}$, $\delta_{(1,0)}$, $\delta_{(1,1)}$

Remark. 1. $\delta_x \times \delta_y = \delta_{(x,y)}$

Proof. Defining product measure

$$(\delta_x \times \delta_y)(A \times B) = \delta_x(A)\delta_y(B) = \begin{cases} 1, & \text{if } x \in A \& y \in B \\ 0, & \text{else} \end{cases}$$

$$\delta_{(x,y)}(A \times B) = \begin{cases} 1, & \text{if } (x, y) \in A \times B \\ 0, & \text{else} \end{cases}$$

$(x, y) \in A \times B$ iff $x \in A$ and $y \in B$

□

2. Another fact: $\nu \times (\mu_1 + \mu_2) = (\nu \times \mu_1) + (\nu \times \mu_2)$

$$\begin{aligned}
[\nu \times (\mu_1 + \mu_2)](A \times B) &= \nu(A)[(\mu_1 + \mu_2)(B)] = \nu(A)[\mu_1(B) + \mu_2(B)] \\
&= \nu(A)\mu_1(B) + \nu(A)\mu_2(B) = (\nu \times \mu_1)(A \times B) + (\nu \times \mu_1)(A \times B) \\
&= [(\nu \times \mu_1) + (\nu \times \mu_2)](A \times B)
\end{aligned}$$

3. $(a\nu) \times \mu = a(\nu \times \mu)$

4. While we generally avoid checking measurability in this text, the following shows that measurability w.r.t. a product σ-field on the range space follows from measurability of the component functions w.r.t. the factor σ-fields. Suppose $f : \Omega \to \Lambda_1 \times \Lambda_2$ is any function. Define the projections $\pi_i : \Lambda_1 \times \Lambda_2 \to \Lambda_i$, $\pi_i(\lambda_1, \lambda_2 = \lambda_i$, and the coordinate or component functions of f by $f_i(\omega) = (\pi_i \circ f)(\omega) = \pi_i(f(\omega))$. So we may write f in ordered pair notation by $f(\omega) = (f_1(\omega), f_2(\omega))$.

5. $\delta_a \times \delta_b = \delta_{(a,b)}$. $(a\mu_1 + b\mu_2) \times \mu_3 = a(\mu_1 \times \mu_3) + b(\mu_2 \times \mu_3)$ for all a, b are nonnegative.

Theorem 1.3.5. *Suppose $f : \Omega \to \Lambda_1 \times \Lambda_2$ where $(\Omega, \mathcal{F}), (\Lambda_1, \mathcal{G}_1)$ and $(\Lambda_2, \mathcal{G}_2)$ are measurable spaces. Then f is measurable from $(\Omega, \mathcal{F})$ to $(\Lambda_1, \mathcal{G}_1) \times (\Lambda_2, \mathcal{G}_2)$ iff each coordinate function f_i is measurable (from $(\Omega, \mathcal{F})$ to $(\Lambda_i, \mathcal{G}_i)$ for $i = 1, 2$.*

1.4 Densities and The Radon-Nikodym Theorem

1.4.1 Absolute Continuity and Singularity

(Billlingsley Section 32) Measures μ and ν on $(\Omega, \mathcal{F})$ are by definition *mutually singular* if they have disjoint supports— that is, if there exist sets S_μ and S_ν such that

$$\begin{cases} \mu(\Omega - S_\mu) = 0, \nu(\Omega - S_\nu) = 0, \\ S_\mu \bigcap S_\nu = \emptyset \end{cases} \tag{1.2}$$

In this case μ is also said to be singular with respect to ν and ν singular with respect to μ. Note that measures are automatically singular if one of them is identically 0.

A finite measure on $\mathbb{R}^1$ with distribution function f is singular with respect to Lebesgue measure in the sense of (1.2) if and only if $f'(x) = 0$ except on a set of Lebesgue measure 0. The latter condition was taken as the definition of singularity, but of course it is the requirement of disjoint supports that can be generalized from $\mathbb{R}^1$ to an arbitrary Ω.

The measure ν is absolutely continuous w.r.t. μ if for each $A \in \mathcal{F}$, $\mu(A) = 0 \implies \nu(A) = 0$. In this case ν is also said to be *dominated* by μ, and the relation is indicated by $\nu \ll \mu$. If $\nu \ll \mu$ and $\mu \ll \nu$, the measures are are equivalent, indicated by $\nu \equiv \mu$.

A finite measure on the line is by Billingsley Theorem 31.7 absolutely continuous in this sense with respect to Lebesgue measure if and only if the corresponding distribution function f satisfies the Billingsley condition (31.28). The latter condition, taken in Billingsley Section 31 as the definition of absolute continuity, is again not the one that generalizes from $\mathbb{R}^1$ to Ω.

There is an $\epsilon - \delta$ idea related to the definition of absolute continuity given above. Suppose that for every ϵ there exists a δ such that $\nu(A) < \epsilon$ if $\mu(A) < \delta$. If this condition holds, $\mu(A) = 0$ implies that $\nu(A) < \epsilon$ for all ϵ, and so $\nu \ll \mu$. Suppose, on the other hand, that this condition fails and that ν is finite. Then for some ϵ there exist sets A_n such that $\mu(A_n) < n^{-2}$ and $\nu(A_n) \geq \epsilon$. If $A = \limsup_n A_n$, then $\mu(A) = 0$ by the first Borel-Cantelli lemma (which applies to arbitrary measures), but $\nu(A) \geq \epsilon > 0$ which applies because ν is finite. Hence $\nu \ll \nu$ fails, and so $\nu(A) < \epsilon$ if $\mu(A) < \delta$ follows if ν is finite and $\nu \ll \mu$. If ν is finite, in order that $\nu \ll \mu$ is therefore necessary and sufficient that for every ϵ there exist a satisfying $\nu(A) < \epsilon$ if $\mu(A) < \delta$. This condition is not suitable as a definition, because it need not follow from $\nu \ll \mu$ if ν is infinite.

(D.D. Cox)

Logical statement $Q(f)$, f is a function. Assume μ, ν fixed and have a statement $Q(f)$: "f is a density of ν w.r.t μ", i.e. $\forall A$, $\nu(A) = \int_A f d\mu, f \geq 0$ e.g. "f is the density for $N(0,1)$ w.r.t. m" $= Q(f)$ is true for $f(x) = \dfrac{1}{\sqrt{2\pi}} \exp\left(\dfrac{-x^2}{2}\right)$, is false for $f(x) = I_{(0,1)}(x)$. Claim that if g is any other density for $N(0,1)$ w.r.t. m, then $g(x) = \dfrac{1}{\sqrt{2\pi}} \exp\left(\dfrac{-x^2}{2}\right)$ m-a.e. More generally, if f is $Q(f)$="f is a density of ν w.r.t. μ". If $Q(f)$ & $Q(g)$, then $f = g$ μ-a.e. Provided μ satisfies the condition of being σ-finite.

Definition 1.4.1. Let μ and ν be measures on $(\Omega, \mathcal{F})$. We say ν is **absolutely continuous** w.r.t. μ and write $\nu \ll \mu$ iff for all $A \in \mathcal{F}$, $\mu(A) = 0$ implies $\nu(A) = 0$. We sometimes say μ dominates ν, or that μ is a dominating measure for ν. We say ν and μ are equivalent (and write $\nu \cong \mu$) iff both $\nu \ll \mu$ and $\mu \ll \nu$.

In words, $\nu \ll \mu$ if the collection of μ-null sets is a subcollection of the collection of ν-null sets, i.e. ν "has more null sets than" μ. If $(\Omega, \mathcal{F}, \mu)$ is a measure space and $f : \Omega \to [0, \infty)$ is Borel measurable, then $\nu(A) = \int f d\mu$ defines a measure ν on the same measurable space $(\Omega, \mathcal{F})$. It is easy to show that $\nu \ll \mu$. It turns out that a converse is true also, provided μ is σ–finite.

Assume $\forall \int_A f d\mu = \int_A g d\mu$, f is unique here. e.g. $A = \{f > g\}$, $\int_A f d\mu = \int_A f d\mu$, $\int_A (f - g) d\mu = 0$ only if $\int_A g d\mu$ is finite. $(f - g)$ is nonnegative on A. If an integral of a nonnegative function is 0, then the function is 0 a.e. Thus, $I_A(f - g) = 0$ is μ-a.e. Similarly, $B = \{g > f\}$, $\Omega = A \cup B \cup \{g = f\}$. $\{g = f\}^c = A \cup B$ is μ measure 0 since $f - g > 0$, $I_A = 0$ μ-a.e.

Using σ-finiteness of μ, we can eliminate the $\int_A g d\mu < \infty$ and $\int_B f d\mu < \infty$

Counterexample: $\mu = \infty \cdot m$, $\mu(A) = \infty$ under $m(A) = 0$ density of μ w.r.t. μ is 2.

$$\int_A 2 d\mu = 2\mu(A) = \begin{cases} 0, & \text{if } m(A) = 0 \\ \infty, & \text{if } m(A) > 0 \end{cases}$$

all so $g(x) \equiv 1$ is a density of this μ w.r.t. itself.

1.4.2 Basic Definition and Result

When does a measure ν have a density w.r.t. μ (σ-finite)? Necessary condition: If $\mu(A) = 0$, then $\nu(A) = 0$. If $\nu(A) = \int_A f d\mu$ for some f, $\nu(A) = \int_A f d\mu = 0$ since $\mu(A) = 0$. ν is absolutely continuous (or is dominated by) μ iff measurable sets A, $\forall A$, $\mu(A) = 0 \implies \nu(A) = 0$

Theorem 1.4.1. ***(Radon-Nikodym Theorem):*** *Let $(\Omega, \mathcal{F}, \mu)$ be a σ-finite measure space and $\nu \ll \mu$. Then there us a nonnegative Borel function f s.t. $\forall A, \int_A f d\mu = \nu(A)$. Furthermore, f is unique μ-a.e. if $\nu(A) = \int g d\mu$ for all $A \in \mathcal{F}$, then $g = f$ μ-a.e.*

The function f is called the **Radon-Nikodym derivative** or **density** of ν w.r.t. μ, and is often denoted $d\nu/d\mu$.

$$\nu(A) = \int I_A d\nu = \int I_A f d\mu \implies d\nu = f d\mu$$

If $\mu = m$ is a Lebesgue measure, then f is called a *Lebesgue density* or a density of the continuous type. We say a random variable X is a continuous random variable iff Law$[X]$ has a Lebesgue density, and we refer to this density at the density of X and will often write

$$f_X(x) = \frac{d\text{Law}[X]}{dm}(x) = \frac{d\nu}{d\mu}(x)$$

Similarly, if $\underline{X}$ is a random n-vector and Law$[\underline{X}] \ll m^n$, then we say X is a *continuous random vector* with a similar notation for its Lebesgue density, which is sometimes also called a density of the continuous type. We have

$$\nu(A) = \int_A 1 d\nu = \int_A \frac{d\nu}{d\mu} d\mu = \int_A \frac{d\nu(\omega)}{d\mu(\omega)} d\mu(\omega)$$

Notice how the $d\mu$'s "cancel" on the r.h.s., that is, $\int_A \frac{d\nu}{d\mu} d\mu = \int_A 1 d\nu = \nu(A)$ by $d\mu$'s "cancel". Also, the Radon-Nikodym derivative is only determined μ-a.e., i.e. we can change its value on a set of μ-measure 0 and not change the measure ν defined by the density. A particular choice for the function $\dfrac{d\nu}{d\mu}$ is called a *version* of the Radon-Nikodym derivative. Two versions of $\dfrac{d\nu}{d\mu}$ are equal μ-a.e. Another way we will sometimes indicate a Radon-Nikodym derivative is the following notation, i.e. $\int \phi d\nu = \int \phi f d\mu$ is true for $\phi = I_A$ / true for simple functions / true for general functions, $\int \cdots d\nu = \int \cdots f d\mu$. That is, $d\nu = f d\mu$.

Example 1.4.1. Let $\Omega = \{a_1, a_2 \cdots\}$ be a discrete set (finite of infinite), and let μ be a measure on $(\Omega, \mathcal{P}(\Omega))$. Put $f(a) = \mu(\{a\})$, then we claim that $d\mu/d\# = f$, $\mu \ll \#$, where $\#$ is counting measure on Ω. By definition of measure,

$$\mu(A) = \sum_{a_i \in A} \mu(\{a_i\}) = \sum_{a_i \in A} f(a_i) = \int f d\#$$

In this context, it is sometimes said that f is a *density of the discrete type for* μ. If μ is a probability measure, the density of the discrete type is also sometimes called the probability mass function (p.m.f.) $f(n) = \mathbb{P}[X = n]$ if $d\mu/d\# = f$ exists. If a random variable has a distribution which is dominated by counting measure, then it is called a discrete random variable.

Recall that a unit point mass measure at ω is given by

$$\delta_\omega(A) = \begin{cases} 1, & \text{if } \omega \in A_i \\ 0, & \text{otherwise} \end{cases}$$

Then a measure can be written as $\mu = \sum_i f(a_i)\delta_{a_i}$. The following example shows that point mass measures can be useful components of dominating measures for distributions which arise in applied statistics.

Example 1.4.2. censored random variable: $X = \min\{Y, C\}$, C is censoring time, $Y \sim Exp(\lambda)$ is a nonnegative random variable. e.g. P_Y has Lebesgue density $f_Y(y|\lambda) = \dfrac{1}{\lambda}\exp\left(-\dfrac{y}{\lambda}\right) I_{(0,\infty)}(y)$. $\mathbb{P}[Y \geq C] = \exp\left(-\dfrac{C}{\lambda}\right) \implies \mathbb{P}[X = C] = P_X(\{C\}) > 0$, but $m(\{C\}) = 0$, $\implies P_X \not\ll m$, that is, there is no Lebesgue measure for P_X. P_X doesn't have Lebesgue density. But X does have a density w.r.t. the dominating measure $\mu = m + \delta_C$, $\delta_C(A) = I_A(C)$.

$$f_X(x|\lambda) = \frac{d\text{Law}[X]}{d\mu}(x) = \frac{dP_X}{d\mu}(x) = \begin{cases} f_Y(x), & \text{if } x < C \\ \exp\left(-\dfrac{C}{\lambda}\right), & \text{if } x = C \\ 0, & \text{otherwise} \end{cases}$$

That is, if $x \geq C$, then $\mathbb{P}[X \leq x] = 1$. To verify this, we will show

$$F_X(x), \text{ c.d.f. of } X = \int_{-\infty}^{x} f_X(z) d\mu(z) = \begin{cases} 0, & \text{if } x \leq 0 \\ 1 - \exp\left(-\dfrac{x}{\lambda}\right) = F_Y(x), & \text{if } 0 \leq x \leq C \\ 1, & \text{if } x \geq C \end{cases}$$

$F_X(\cdot)$ is given by formula $F_X(x) = \mathbb{P}[X \leq x] = \mathbb{P}[X \in (-\infty, x]]$. If $x < C$, then $X = Y$ (not censored). Thus, the r.h.s. $= \mathbb{P}[Y \leq x] = F_Y(x)$.

Now verify that integral gives same formula $\int_{-\infty}^{x} f_X(z)d\mu(z) = \int_{-\infty}^{x} f_X(z)dm(z) + \int_{-\infty}^{x} f_X(z)d\delta_C(z)$ (By homework in 1.2)

$$\int_{-\infty}^{x} f_X(z) dm(z) = \begin{cases} F_Y(x), & \text{if } x < C \text{ (since } f_X(z) = f_Y(z) \text{ if } z < C) \\ F_Y(C) = 1 - \exp\left(-\dfrac{C}{\lambda}\right), & \text{if } x \geq C \text{ (since } f_X(z) = 0 \text{ if } z > C) \end{cases}$$

Note that $f_X(C) = 0$ in Lebesgue density.

$$\int_{-\infty}^{x} f_X(z) d\delta_C(z) = I_{(\infty,x]}(C) f_X(C) = \begin{cases} 0, & \text{if } x < C \\ \exp\left(-\frac{C}{\lambda}\right), & \text{if } x \geq C \end{cases}$$

Thus,

$$\int_{-\infty}^{x} f_X(z) d\mu(z) = \begin{cases} F_Y(x), & \text{if } x < C \\ 1, & \text{if } x \geq C \end{cases}$$

is a c.d.f.

Example 1.4.3. *(An alternative example for Example 1.4.2)*: Suppose a r.v. X is obtained by measuring the concentration of a chemical in water, but because of limitations of the measuring instrument, concentrations less than some amount x_0 are reported as x_0. Suppose Y is the true concentration, then we might think of X as given by $X = \max\{x_0, Y\}$. Suppose Y has Lebesgue density

$$f_Y(y) = \begin{cases} \exp(-y), & \text{if } y > 0 \\ 0, & \text{otherwise} \end{cases}$$

Then X does not have a Lebesgue density because $P[X = x_0] = 1 - \exp(-x_0)$ but $m(\{x_0\}) = 0$, so we do not have $\text{Law}[X] \ll m$. But X does have a density w.r.t. the measure $\mu = m + \delta_{x_0}$ which is given by

$$f_X(x) = \frac{d\text{Law}[X]}{d\mu}(x) = \begin{cases} \exp(-x), & \text{if } x > x_0 \\ 1 - \exp(-x_0), & \text{if } x = x_0 \\ 0, & \text{otherwise} \end{cases}$$

A useful consequence of Radon-Nikodym Theorem is that if f is Borel on $(\Omega, \mathcal{F})$ an d $\int_A f d\mu = 0$ for ant $A \in \mathcal{F}$, then $f = 0$ a.e.

If $\int f d\mu = 1$ for an $f \geq 0$ is μ-a.e., then ν is a probability measure and f is its *probability density function* w.r.t. μ. For any probability measure $\mathbb{P}$ on $(\mathbb{R}^k, \mathcal{B}^k)$ corresponding to a c.d.f. F or a random vector X, if $\mathbb{P}$ has a p.d.f. f w.r.t. a measure μ, then f is also called the p.d.f. of F or X w.r.t. μ.

We said "X does have a density" when we really meant "Law[X] does have a density". This is a common abuse of terminology one sees in probability and statistics.

Example 1.4.4. Let F be a c.d.f. Assume that F is differentiable in the usual sense in calculus. Let f be the derivative of F. From calculus, $F(x) = \int_{-\infty}^{x} f(y)dy$. Let $\mathbb{P}$ be the probability measure corresponding to F. It can be shown that $P(A) = \int_X f dm$ for any $A \in B$, where m is the Lebesgue measure on $\mathbb{R}$. Hence, f is the p.d.f. of $\mathbb{P}$ or F w.r.t. Lebesgue measure. In this case, the Radon-Nikodym derivative is the same as the usual derivative of F in calculus.

A continuous c.d.f. may not have a p.d.f. w.r.t. Lebesgue measure. A necessary and sufficient condition for a c.d.f. F having a p.d.f. w.r.t. Lebesgue measure is that F is **absolute continuous** in the sense that for any $\epsilon > 0$, there exists a $\delta > 0$ such that for each finite collection of disjoint bounded open intervals (a_i, b_i), $\sum(b_i - a_i) < \delta$ implies $\sum[F(b_i) - F(a_i)] < \epsilon$. Absolute continuity is weaker than differentiability, but is stronger than continuity. Thus, any discontinuous c.d.f. (such as a discrete c.d.f.) is not absolute continuous. Note that every c.d.f. is differentiable a.e. Lebesgue measure (Chung Chapter 1). Hence, if f is the p.d.f. of F w.r.t. Lebesgue measure, then f is the usual derivative of F a.e. Lebesgue measure. In such a case probabilities can be computed through integration. It can be shown that the uniform and exponential c.d.f.'s are absolute continuous. A p.d.f. w.r.t. Lebesgue measure is called a Lebesgue p.d.f.

1. Unif$[0,1]$, Lebesgue density $f(x) = I_{(0,1)}(x) = I_{[0,1]}(x)$, m-a.e.

Proposition 1.4.2. ***Calculus with Radon-Nikodym derivatives:*** *Let $(\Omega, \mathcal{F})$ be a measurable space with measures $\mu, \nu, \nu_1, \nu_2, \lambda$. Assume μ and λ are σ-finite.*

1. *If $\nu \ll \mu$ and $f > 0$, then $\int f d\nu = \int f \left(\frac{d\nu}{d\mu}\right) d\mu$*

   Proof. The result is obviously true for indicators. Proceed to simple functions, then take limits using the Monotone Convergence Theorem and Proposition 1.2.6. □

2. *If $\nu_1 \ll \mu$, then $\nu_1 + \nu_2 \ll \mu$ and $\frac{d(\nu_1 + \nu_2)}{d\mu} = \frac{d\nu_1}{d\mu} + \frac{d\nu_2}{d\mu}$, μ-a.e.*

   Proof. Note that $\nu_1 + \nu_2$ is a measure. Now $\nu_i \ll \mu$ for $i = 1, 2$ implies $\nu_1 + \nu_2 \ll \mu$. If $A \in \mathcal{F}$ then

$$\begin{aligned}
\nu_1 + \nu_2(A) &= \nu_1(A) + \nu_2(A) && \text{(the definition of } \nu_1 + \nu_2\text{)} \\
&= \int_A \frac{d\nu_1}{d\mu} d\mu + \int_A \frac{d\nu_2}{d\mu} d\mu && \text{(the definition of } \frac{d\nu_i}{d\mu}\text{)} \\
&= \int_A \left[\frac{d\nu_1}{d\mu} + \frac{d\nu_2}{d\mu}\right] d\mu && \text{(linearity of the integral)}
\end{aligned}$$

   By uniqueness of the Radon-Nikodym derivative, the integrand $\frac{d\nu_1}{d\mu} + \frac{d\nu_2}{d\mu}$ must be a version of $\frac{d(\nu_1 + \nu_2)}{d\mu}$, as required. □

3. ***(Chain Rule)***: *If $\nu \ll \mu \ll \lambda$, then $\frac{d\nu}{d\lambda} = \frac{d\nu}{d\mu}\frac{d\mu}{d\lambda}$, λ-a.e. In particular, if $\mu \cong \nu$, then $\frac{d\nu}{d\mu} = \left(\frac{d\mu}{d\nu}\right)^{-1}$ (homework)*

***Note*.** Part (1) of the this proposition is familiar in the context of probability and statistics in the following way: if X is a continuous r.v. with Lebesgue density f and g is a Borel measurable function $\mathbb{R} \to \mathbb{R}$, then $E[g(X)] = \int_{\mathbb{R}} g d\text{Law}[X] = \int_{-\infty}^{\infty} g(x)f(x)dx$. Note that the first equality is the law of the unconscious statistician (Theorem 1.2.8, change of variables).

***Remark*.**

1. **Statistical Models**: We observe a random vector Y. Assume P_Y is in a family of distributions for a random variable $\{P_\theta : \theta \in \Theta\}$, Θ: parameter space. We could have a dominated family iff there exists a σ-finite μ s.t. $\forall P_\theta$, $P_\theta \ll \mu$, then they have densities w.r.t. μ denoted $f_\theta = \frac{dP_\theta}{d\mu}$ by Radon-Nikodym theorem.

2. **Likelihood**: Densities as a functions of parameters are likelihoods. Observation y of Y, $f_\theta(y) = L(\theta) = L(\theta|y)$ (for bayesian insight). Typically, we will assume something which is continuous and differentiable in θ,

What about the different *versions* of a likelihood? What about dominating measure?

Example 1.4.5. $X \sim N(\mu, 1)$

$$f_\mu(x) = \begin{cases} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2}\right), & \text{if } x \neq 2+\mu \\ 10^{20}, & \text{if } x = 2+\mu \end{cases}$$

It is also a version of Lebesgue density, but I cannot get the MLE. Here plug in a value x, $f_\mu(x)$ is maximized at $x-2$. Thus, typically there is a "regular" version of the density to use. When we start to talk about asymptotic optimality of MLE, we need to have some assumption of *natural* likelihood like second derivative, and we could not use these kinds of "crazy" densities.

Remark. **Change of dominated measure**: Assume $\mu \ll \lambda$, then $P_\theta \ll \lambda$. Thus, $L_\mu(\theta) = \frac{dP_\theta}{d\mu}$, $L_\lambda(\theta) = \frac{dP_\theta}{d\lambda} = \frac{dP_\theta}{d\mu}\frac{d\mu}{d\lambda} = L_\mu(\theta)\frac{d\mu}{d\lambda}$. The second element does not depend on θ. It means that I only care about where the maximum is located ($L_\mu(\theta)$), I do not care about the "value" here ($L_\lambda(\theta)$). Multiplying the likelihood $L(\theta)$ by some functions of x will not change inferences. When I plug in a x, $\frac{d\mu}{d\lambda}$ should be a positive constant. For any realization of observations, we only multiply the likelihood by a constant and doesn't change the location of maximum. If I do a test on ratio of likelihood (LRT), it won't change the ratio of likelihood.

1.4.3 Densities w.r.t. Product Measures

Proposition 1.4.3. *Let $(\Omega_i, \mathcal{F}_i, \mu_i)$, $i = 1, 2$ be the σ-finite measure spaces with $\nu_i \ll \mu_i$. Then $\nu_1 \times \nu_2 \ll \mu_1 \times \mu_2$ and $\frac{d(\nu_1 \times \nu_2)}{d(\mu_1 \times \mu_2)}(\omega_1, \omega_2) = f(\omega_1, \omega_2) = \left[\frac{d\nu_1}{d\mu_1}(\omega_1)\right]\left[\frac{d\nu_2}{d\mu_2}(\omega_2)\right]$, $\mu_1 \times \mu_2$-a.e.*

Proof. Let $A_i \in \mathcal{F}_i$, then

$$\begin{aligned}
(\nu_1 \times \nu_2)(A_1 \times A_2) &= \nu_1(A_1)\nu_2(A_2) \\
&= \int_{A_1} \frac{d\nu_1}{d\mu_1}(\omega_1) d\mu_1(\omega_1) \int_{A_2} \frac{d\nu_2}{d\mu_2}(\omega_2) d\mu_2(\omega_2) \\
&= \int_{A_1}\int_{A_2} \frac{d\nu_1}{d\mu_1}(\omega_1)\frac{d\nu_2}{d\mu_2}(\omega_2) d\mu_2(\omega_2) d\mu_1(\omega_1) \qquad \text{(by Fubini's theorem)} \\
&= \int_{\Omega_1}\int_{\Omega_2} I_{A_1}(\omega_1) I_{A_2}(\omega_2) \frac{d\nu_1}{d\mu_1}(\omega_1)\frac{d\nu_2}{d\mu_2}(\omega_2) d\mu_2(\omega_2) d\mu_1(\omega_1) \\
&= \int_{\Omega_1}\int_{\Omega_2} I_{A_1 \times A_2}(\omega_1, \omega_2) \frac{d\nu_1}{d\mu_1}(\omega_1)\frac{d\nu_2}{d\mu_2}(\omega_2) d\mu_2(\omega_2) d\mu_1(\omega_1) \\
&= \int_{\Omega_1 \times \Omega_2} I_{A_1 \times A_2}(\omega) \frac{d\nu_1}{d\mu_1}(\omega_1)\frac{d\nu_2}{d\mu_2}(\omega_2) d(\mu_1 \times \mu_2)(\omega_1, \omega_2) \\
&= \int_{A_1 \times A_2} \frac{d\nu_1}{d\mu_1}(\omega_1)\frac{d\nu_2}{d\mu_2}(\omega_2) d(\mu_1 \times \mu_2)(\omega_1, \omega_2) \qquad \text{(by Fubini's theorem)}
\end{aligned}$$

By the uniqueness part of the Product Measure Theorem (Theorem 1.3.1), it follows that the measure

$$\nu(C) = \int_C \frac{d\nu_1}{d\mu_1}(\omega_1)\frac{d\nu_2}{d\mu_2}(\omega_2) d(\mu_1 \times \mu_2)(\omega_1, \omega_2)$$

defined on $(\Omega_1, \mathcal{F}_1, \mu_1) \times (\Omega_2, \mathcal{F}_2, \mu_2)$ is in fact $\nu_1 \times \nu_2$. Now $\nu \ll \mu_1 \times \mu_2$ and by the uniqueness part of the Radon-Nikodym theorem, $\frac{d(\nu_1 \times \nu_2)}{d(\mu_1 \times \mu_2)}(\omega_1, \omega_2) = f(\omega_1, \omega_2) = \left[\frac{d\nu_1}{d\mu_1}(\omega_1)\right]\left[\frac{d\nu_2}{d\mu_2}(\omega_2)\right]$, $\mu_1 \times \mu_2$-a.e. □

Remark. The last result implies that if X_1 and X_2 are independent continuous random variables with Lebesgue densities f_1 and f_2, then the joint distribution of (X_1, X_2) is also continuous (i.e. $\text{Law}[(X_1, X_2)] \ll m^2$) and the joint density f w.r.t. m^2 is the product of the marginal densities, i.e. $f(x_1, x_2) = f_1(x_1)f_2(x_2)$. Of course, this remark (and the preceding Proposition) can be extended to more than two random variables and two measures by induction. The converse of this remark is also true (Exercise 1.4.11, homework).

Example 1.4.6. If X, Y are independent random variables,then $P_{XY} = P_X \times P_Y$. $P_X \ll \mu_1$ and $P_Y \ll \mu_2$ then $P_{XY} \ll \mu_1 \times \mu_2$ and $f_{XY} = f_X f_Y$

Under independence, we can construct the joint density w.r.t. the product of the dominating measures from the marginal densities by simple multiplication. In general, there is no such nice relationship between the joint and the marginal densities, but we can always recover the marginal densities from the joint density.

Proposition 1.4.4. ***Marginalization / Marginal density:*** *Let $(\Omega_i, \mathcal{F}_i, \mu_i)$, $i = 1, 2$ be the σ-finite measure spaces with $\nu \ll \mu_1 \times \mu_2$. Let $\pi_1 : \Omega_1 \times \Omega_2 \to \Omega_1$ be the coordinate projection (mapping) given by $\pi_1(\omega_1, \omega_2) = \omega_1$ and similarly for π_2. Then $\nu \circ \pi_i^{-1} \ll \mu_i$ and*

$$\frac{d(\nu \circ \pi_1^{-1})}{d\mu_1}(\omega_1) = \int_{\Omega_2} \frac{d\nu}{d(\mu_1 \times \mu_2)}(\omega_1, \omega_2) d\mu_2(\omega_2)$$

Proof. When I have a measure on a space, the space is the domain of the measurable function. To another space, function creates a new measure in its range space (induced measure). Note that $\nu \circ \pi_1^{-1}$ is a measure on $(\Omega_1, \mathcal{F}_1)$. Our goal in this proof is to show that $\nu_1 \ll \mu_1$, and then that $\frac{d\nu_1}{d\mu_1} = f_1$, μ_1-a.e. If $\mu_1(A) = 0$, then the integral above is 0, so $\nu_1(A) = 0$ and we have that $\nu_1 \ll \mu_1$. Furthermore, since $A \in \mathcal{F}_1$ was arbitrary, we can calculate the ν_1 measure of a set by integrating w.r.t. $d\mu_1$ the function f_1 over the set. Hence, by the uniqueness part of the Radon–Nikodym theorem, $\frac{d\nu_1}{d\mu_1} = f_1$, μ_1-a.e.

Now if $A \in \mathcal{F}_1$, then $\pi_1^{-1}(A) = A \times \Omega_2$ is a rectangle set. Thus,

$$\begin{aligned}
\nu \circ \pi_1^{-1}(A) &= \nu(\pi_1^{-1}(A)) = \nu(A \times \Omega_2) \\
&= \int_{A \times \Omega_2} \frac{d\nu}{d(\mu_1 \times \mu_2)}(\omega_1, \omega_2) d(\mu_1 \times \mu_2)(\omega_1, \omega_2) \\
&= \int_A \left[\int_{\Omega_2} \frac{d\nu}{d(\mu_1 \times \mu_2)}(\omega_1, \omega_2) d\mu_2(\omega_2) \right] d\mu_1(\omega_1) \qquad \text{(by Fubini's theorem)}
\end{aligned}$$

□

Example 1.4.7. X, Y are random variables, $P_{XY} \ll (\mu_1 \times \mu_2)$, $dP_{XY} = f_{XY} d(\mu_1 \times \mu_2)$. Then the marginal density for X is $f_X(x) = \int f_{XY} d\mu_2(y)$

Example 1.4.8. Suppose $\mu_1 = m$ and $\mu_2 = \#$ on $\{1, \cdots, k\}$ in a discrete set $\Omega = \{a_1, a_2, \cdots\}$. $f_{XY}(x, y) = \pi(y) \frac{1}{\sqrt{2\pi\sigma^2(y)}} \exp\left[-\frac{(x - \mu(y))^2}{2\sigma^2(y)} \right]$, $f_{XY}(x, j) = \pi_j \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp\left[-\frac{(x - \mu_j)^2}{2\sigma_j^2} \right] = \pi_j \phi(x|\mu_j, \sigma_j^2)$ & $\pi_j \geq 0, \sum_{j=1}^{k} \pi_j = 1$.

Find the marginal for X: $\int f_{XY}(x, y) d\mu_2(y) = \sum_{j=1}^{k} f_{XY}(x, j) = \sum_{j=1}^{k} \pi_j \phi(x|\mu_j, \sigma_j^2)$, a mixture normal density.

Verify f_{XY} is a probability density w.r.t. $m \times \mu_2$. $\int f_{XY} d(\mu_1 \times \mu_2) = \int_{-\infty}^{\infty} \sum_{j=1}^{k} \pi_j \phi(x|\mu_j, \sigma_j^2) dm(x) = \sum_{j} \pi_j \int \phi dm = \sum_{j=1}^{k} \pi_j = 1$

Example 1.4.9. Continuing Example 1.4.2 $Y_1, \cdots, Y_n$ are i.i.d. with $Exp(\lambda)$, $\lambda > 0$. Observations $X_i = \min\{Y_i, C_i\}$. The joint density of $X_1, \cdots, X_n$ w.r.t. $\mu = \prod_{i=1}^{n}(m \times \delta_{C_i})$ if $f_\mu(x_1, \cdots, x_n) = \prod_{i=1}^{n} \left[\frac{1}{\lambda} \exp\left(-\frac{x_i}{\lambda}\right)\right]^{1-\delta_i} \left[\exp\left(-\frac{C_i}{\lambda}\right)\right]^{\delta_i} = \prod_{i=1}^{n} g_\lambda(x_i)^{1-\delta_i} \bar{G}_\lambda(C_i)^{\delta_i}$, where $g_\lambda \sim Exp(\lambda)$ is Y densities, $\bar{G}_\lambda(y) = 1 - G_\lambda(y) = \mathbb{P}[Y > y]$, and

$$\delta_i = \begin{cases} 1, & \text{if } Y_i \geq C \\ 0, & \text{else} \end{cases}$$

1.4.4 Support of a Measure

Before introducing the next important concept from measure theory, we briefly review the topology of Euclidean spaces. This is discussed at much greater length in Rudin's book, Principles of Mathematical Analysis. Let $x \in \mathbb{R}^n$, then a **neighborhood** of x is any ball (or sphere) of positive radius ϵ centered at x. A ball of positive radius ϵ centered at x is a set of the form

$$B(x, \epsilon) = \{y \in \mathbb{R}^n : \|x - y\| < \epsilon\}.$$

Here $\|\cdot\|$ denotes the **Euclidean distance (norm)** on $\mathbb{R}^n$ given by $\|x\| = \|(x_1, \cdots, x_n)\| = \sqrt{x_1^2, \cdots, x_n^2}$

A set $A \subset \mathbb{R}^n$ is called *open* iff for every $x \in A$, there is some $\epsilon > 0$ s.t. $B(x, \epsilon) \subset A$. A set $C \subset \mathbb{R}^n$ is called *closed* iff it is the complement of an open set. One can show that a union of open sets is also open, and hence that an intersection of closed sets is also closed. Also, the sets $\mathbb{R}^n$ and $\emptyset$ are both open and closed. Thus, any set $D \subset \mathbb{R}^n$ is contained in some closed set (namely $\mathbb{R}^n$), and the intersection of all closed sets which contain D is also a closed set, namely the smallest closed set containing D. This set is called the closure of D and denoted $\bar{D}$. $\bar{D}$ is also given by the following

$$\bar{D} = \{\lim_n x_n : x_1, \cdots, x_n \cdots \text{ is a sequence of points in } D \text{ for which the } \lim_n \text{ exists}\}$$

Otherwise said, $\bar{D}$ is the set of limit points of D. Now we briefly explore a concept related to absolutely continuity.

Definition 1.4.2. Suppose ν is a measure on $(\mathbb{R}^n, \mathcal{B}_n)$. The **support** of ν is the set

$$supp(\nu) = \{x \in \mathbb{R}^n : \nu(B(x, \epsilon)) > 0 \text{ for all } \epsilon > 0\}$$

One can show that *supp*(ν) is a closed set, and if ν is a probability measure, then *supp*(ν) is the smallest closed set with probability 1.

Proposition 1.4.5. *Suppose μ and ν are Borel measures on $\mathbb{R}^n$, μ is σ-finite, and $\nu \ll \mu$. Then supp$(\nu) \subset \bar{S}$ where $S = \{x \in supp(\mu) : \frac{d\nu}{d\mu}(x) > 0\}$*

Proof. Let $x \in supp(\nu)$, then for any $\epsilon > 0$ we have $\nu(B(x,\epsilon)) = \int_{B(x,\epsilon)} \frac{d\nu}{d\mu} d\mu > 0$. In particular, the nonnegative function $I_{B(x,\epsilon)}(y) \cdot \left(\frac{d\nu}{d\mu}\right)(y)$ cannot be identically 0 on $B(x,\epsilon)$, i.e. $\left(\frac{d\nu}{d\mu}\right)(y) > 0$ for some $y \in B(x,\epsilon)$.

Now let A_n be the sequence of balls $B\left(x, \frac{1}{n}\right)$ and $y_n \in A_n$ s.t. $\left(\frac{d\nu}{d\mu}\right)(y_n) > 0$. One checks that $y_n \to x$, i.e. x is a limit point of S, so $x \in \bar{S}$, as asserted. □

Remark. $\nu \ll \mu$ σ-finite implies $supp(\nu) \subset supp(\mu)$. The converse is false, i.e. $supp(\nu) \subset supp(\mu)$ does not imply $\nu \ll \mu$. Also, we cannot in general claim $supp(\nu) = \bar{S}$ in Proposition 1.4.5. One does however have the next result.

Proposition 1.4.6. *Let $U \subset \mathbb{R}^n$ be open. Suppose*

1. *μ is LEbesgue measure restricted to U, i.e. $\mu(B) = m(B \cap U)$ for all $B \in \mathcal{B}_n$*
2. *$\nu \ll \mu$*
3. *the **version** of $f = \frac{d\nu}{d\mu}$ is continuous on U*

Then $supp(\nu) = \bar{S}$ where $S = \{x \in U : f(x) > 0\}$

Proof. Now f is continuous on U and $f(x) > 0$ for some $x \in U$ implies there is a $\epsilon > 0$ s.t. $f(y) > \epsilon$ for all y in some neighborhood $B(x,\delta_0)$ of x. Hence, for $x \in S$, we have for all $\delta > 0$ that

$$\nu(B(x,\delta)) \geq \epsilon m^n(B(x, \min\{\delta, \delta_0\})).$$

Since the r.h.s. above is positive, it follows that $S \subset supp(\nu)$. On the other hand, if $x \in supp(\nu)$, then for all $\delta > 0$,

$$0 < \int_{B(x,\delta)} f(y) dm(y)$$

so in particular, for all δ there is a $y \in B(x,\delta)$ with $f(y) > 0$ and we can find a sequence $y_n \in S$ with $y_n \to x$. Thus, $x \in \bar{S}$, and we have shown that $supp(\nu) \subset \bar{S}$. Since $\bar{S}$ is the smallest closed set containing S and $supp(\nu)$ is a closed set containing S by the first part of the proof, it follows that $supp(\nu) = \bar{S}$ □

Example 1.4.10. Consider the exponential distribution with Lebesgue density

$$f(x) = \begin{cases} \exp(-x), & \text{if } x \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

We cannot apply the previous proposition to this version of the density, but we can apply it to

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

which is another version (that agrees with the first version except on the set $\{0\}$, which has Lebesgue measure 0). In this second version, the density is positive on the open set $(0,\infty)$, and so the support is the closed set $[0,\infty)$.

1.5 Conditional Expectation

Kolmogorov, 1930 wrote a book for conditional probability for mathematician, we need it for Borel paradox handouts.

First start with machine learning problem. Suppose $Y : (\Omega, \mathcal{F}, \mathbb{P}) \to (\mathbb{R}, \mathcal{B})$ (tumor) is a random variable and $X : (\Omega, \mathcal{F}, \mathbb{P}) \to (\Lambda, \mathcal{G})$ (image) is any random element. We want to predict Y using X. Knowing the value of Y tells us something about the particular outcome ω which occurred, and hence possibly also something about the value of X, i.e. $X(\omega)$. It is often of interest to find the "best predictor" or "estimator" of X based on the observed value of Y . By "based on the observed value of Y ", we mean this predictor is a function of $\hat{Y} = h(X)$, . For mathematical convenience, we take "best" to mean "minimizes the mean squared prediction error (MSPE)," which is defined to be

$$\text{Criterion function: } \text{MSPE}(h(X)) = E[(Y - h(X))^2]$$

, we use calculations of variations methods to find necessary conditions for $h^*(X)$, where h^* is the optimal predictor to minimize MSPE. How do we take derivatives w.r.t. $h(X)$ and set to 0? Suppose we have

$$\begin{aligned} m(t; g(x)) &= \text{MSPE}[(h^*(X) + tg(X))] \\ &= E[(Y - h^*(X) - tg(X))^2] = E[(Y - h^*(X))^2] - 2tE[g(X)(Y - h^*(X))] + t^2 E[g(X)^2] \end{aligned}$$

Set $\frac{dm}{dt} = 0$, $m(t)$ has its minimum at $t = \frac{E[g(X)(Y - h^*(X))]}{E[g(X)^2]} = 0$ since $h^*(X)$ is optimal. Now g function is arbitrary of $X \implies E[g(X)(Y - h^*(X))] = 0$, $\forall g(X)$ normal equations. $\forall g(X)$, $E[g(X)h^*(X)] = E[g(X)Y]$ is a necessary condition for $h^*(X)$ to be optimal. Note that if X, Y both discrete, there are finite equations. It suffices to hold for all indicators, i.e. $\forall A \subseteq$ Range X is measurable, since indicators trivially have finite second moments and also for simple functions by MCT (simple function approximation, etc.). If $h^*(X)$ minimizes MSPE$(h(X))$, then $\forall A \subseteq \mathit{Range}(X)$, $E[I_A(X)h^*(X)] = E[I_A(X)Y]$. This follows by taking a sequence of simple functions on $\mathbb{R}$ converging to h. Note that $I_A(X) - I_{X^{-1}(A)}(\omega)$ and $Y^{-1}(A)$ is a generic element of $\sigma(X)$. Now $E[I_A(X)h^*(X)] = E[I_A(X)Y] \implies E[I_C h^*(X)] = E[I_C Y], \forall C \in \sigma(X)$ provides us with a possibly useful characterization of the "best" predictor of Y which is a function of X, we call this $h^*(X)$ the conditional expectation $E[Y|X]$ defined by measurability and satisfying $E[I_A(X)h^*(X)] = E[I_A(X)Y]$.

Example 1.5.1. Elementary conditional probability: $\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$, $A, B \subseteq \Omega$. Define $\mathbb{P}[A|X] = E[I_A|X]$, what is $\mathbb{P}[A|I_B]$? Derive by joint distribution of (I_A, I_B) two indicator (bernoulli variable). Note that we use measure and event notation here. $P_{I_A I_B} = \mathbb{P}(A \cap B)\delta_{(1,1)} + \mathbb{P}(A \cap B^c)\delta_{(1,0)} + \mathbb{P}(A^c \cap B)\delta_{(0,1)} + \mathbb{P}(A^c \cap B^c)\delta_{(0,0)}$

Joint density w.r.t. $(\mu_1 \times \mu_2)$, $\mu_1 = \mu_2 =$ counting measure on $\{0, 1\} = \delta_0 + \delta_1$

$$f_{I_A I_B}(x, y) = \begin{cases} \mathbb{P}(A \cap B), & (x, y) = (1, 1) \\ \mathbb{P}(A \cap B^c), & (x, y) = (1, 0) \\ \mathbb{P}(A^c \cap B), & (x, y) = (0, 1) \\ \mathbb{P}(A^c \cap B^c), & (x, y) = (0, 0) \end{cases}$$

Marginal density for indicator I_B

$$f_{I_B}(x) = \begin{cases} \mathbb{P}(B), & x = 1 \\ \mathbb{P}(B^c), & x = 0 \end{cases}$$

$$f_{I_A|I_B}(y|x) = \frac{f_{I_A I_B}(x,y)}{f_{I_B}(x)} = \begin{cases} \mathbb{P}(A|B), & \text{if } x = 1\&y = 1 \\ \mathbb{P}(A^c|B), & \text{if } x = 1\&y = 0 \\ \mathbb{P}(A|B^c), & \text{if } x = 0\&y = 1 \\ \mathbb{P}(A^c|B^c), & \text{if } x = 0\&y = 0 \end{cases}$$

Thus, $E[I_A|I_B] = \int y f_{I_A|I_B}(y|I_B)d\mu_2(y) = 0 \cdot f_{I_A|I_B}(0|I_B) + 1 \cdot f_{I_A|I_B}(1|I_B) = P(A|B)I_B + P(A|B^c)I_{B^c}$

Example 1.5.2. mixture model: joint density w.r.t. $m \times \mu$, μ is a counting measure. $f_{XY}(x,y) = \sum_{i=1}^{k} \pi_i \phi(x|\mu_i, \sigma_i^2) I_{\{i\}}(y) = \pi_y \phi(x|\mu_y, \sigma_y^2)$. $f_X(x) = \sum_{i=1}^{k} \pi_i \phi(x|\mu_i, \sigma_i^2)$. $f_{Y|X}(j|x) = \dfrac{\pi_j \phi(x|\mu_j, \sigma_j^2)}{f_X(x)} = \mathbb{P}[Y = j|X = x]$. We have data and given a new observation $X = x$, $P[Y = j|X = x]$ to predict the classes of this new observation. It is useful for prediction, and similar to the clustering and unsupervised learning (because we don't know what y is.)

Remark. (Other conditional expectation notation:) $E[Y|X]$ is a random variable $h(x)$ satisfying $\forall A \subseteq$ *Range*(X), $E[I_A(x)h(x)] = E[I_A(X)Y]$. $h :$ *Range*$(X) \to \mathbb{R}$, $h(x) = E[Y|X = x]$ as a function of x. Note $E[Y|X] = E[Y|X = x]$ why? see 1.5.1

Conditional distribution: In the previous framework, $f_{Y|X}(y|x) = \dfrac{f_{XY}(x,y)}{f_X(x)}$. Define a family of probability measures on Range(Y): $P_{Y|X}(A|x) = E[I_A(Y)|X = x]$. If we fix x, $P_{Y|X}(\cdot|x)$ is a probability measure on Range(Y). Further, $E[g(X,Y)|X = x] = \int g(x,y) f_{Y|X}(y|x)d\mu_2(y) = \int g(x,y)dP_{Y|X}(y|x)$.

1.5.1 Characterization of Measurable Transformations of a Random Element

Recall that we wanted $E[Y|X]$ to be a function of X satisfying other conditions (namely $(E[I_C h^*(X)] = E[I_C Y], \forall C \in \sigma(X))$. The next result is a very useful characterization of the class of r.v.'s which are functions of X.

Theorem 1.5.1. *Suppose $X : (\Omega, \mathcal{F}) \to (\Lambda, \mathcal{G})$ is a random variable and $Z : (\Omega, \mathcal{F}) \to (\mathbb{R}^n, \mathcal{B}_n)$. Then Z is $\sigma(X)$-measurable iff there is a Borel finction $h : (\Lambda, \mathcal{G}) \to (\mathbb{R}^n, \mathcal{B}_n)$ such that $Z = h(X)$*

Remark. To say "Z is $\sigma(X)$-measurable " means $Z : (\Omega, \mathcal{F}) \to (\mathbb{R}^n, \mathcal{B}_n)$, i.e. $\sigma(Z) = Z^{-1}(\mathcal{B}_n) \subset \sigma(X)$. Note that $\sigma(X)$ is a sub–σ–field of $\mathcal{F}$. The theorem may be summarized pictorially as follows:

$$\begin{tikzcd}
(\Lambda, \mathcal{G}) \arrow[rr, "h \text{ exists iff } \sigma(Z) \subset \sigma(X)"] & & (\mathbb{R}^n, \mathcal{B}_n) \\
& (\Omega, \mathcal{F}) \arrow[ul, "X"] \arrow[ur, "Z"'] &
\end{tikzcd}$$

Proof. Assume that $\sigma(Z) \subset \sigma(X)$ and we will show the existence of such an h. Also, assume for now $n = 1$. We proceed in steps, as usual.

1. If Z is a simple function $\sum_{i=1}^{m} a_i I_{A_i}$, where the sets A_i are disjoint and coefficient a_i are distinct and nonzero, i.e. $a_i \neq a_j$, if $i \neq j$. Then $A_i = Z^{-1}(\{a_i\}) \in \sigma(Z)$ and hence also $A_i \in \sigma(X), i \leq i \leq m$, i.e. $A_i = X^{-1}(C_i)$ for some $C_i \in \mathcal{G}$ since all $A_i \in \sigma(X)$ are of this form by definition of $\sigma(X)$. Put $h = \sum_{i=1}^{m} a_i I_{C_i}$. Then $h(X(\omega)) = \sum_{i=1}^{m} a_i I_{C_i}(X(\omega)) = \sum_{i=1}^{m} a_i I_{X^{-1}(C_i)}(\omega) = \sum_{i=1}^{m} a_i I_{A_i}(\omega)$. This completes the proof if Z is a simple function.

2. If Z is not simple, then there exist simple functions Z_n such that $Z_n(\omega) \to Z(\omega)$, $\forall \omega \in \Omega$ by simple function approximation. By step 1, each $Z_n = g_n(Y)$ for some $g_n : (\Lambda, \mathcal{G}) \to (\mathbb{R}^n, \mathcal{B}_n)$. Now put $L = \{\lambda \in \Lambda : \lim_n g_n(\lambda) \text{ exists } \}$. Let $h_n = g_n I_L$. Clearly there is a function $h = \lim_n h_n$ (since if $\lambda \in L$ then $h_n(\lambda) = g_n(\lambda))$ and the sequence of real numbers $g_n(\lambda)$ has a limit by definition of L, and if $\lambda \notin L$ then $g_n(\lambda) = 0$, which has the limit 0 as $n \to \infty$), and h is measurable by Proposition 1.2.1 (c).

   We will show $Z(\omega) = h(X(\omega))$ $\forall \omega \in \Omega$. Note that $X(\omega) \in L$ because $g(X(\omega)) = Z_n(\omega) \to Z(\omega)$. By definition of h_n, $h_n(X(\omega)) = g_n(X(\omega)) \to Z(\omega)$, but $h_n(X(\omega)) \to h(X(\omega))$ by definition of h, so $Z(\omega) = h(X(\omega))$. This finishes Step 2.

3. Finally, to remove the restriction $n = 1$, use the result for $n = 1$ on each component of $Z = (Z_1, \cdots, Z_n)$ and apply Theorem 1.3.5 to conclude that Z is $\sigma(Y)$ measurable when each component is $\sigma(X)$ measurable. To prove the converse, assuming $Z = h(X) = h \circ X$ for some $h : (\Lambda, \mathcal{G}) \to (\mathbb{R}^n, \mathcal{B}_n)$, we have $Z^{-1}(B) = (h \circ X)^{-1}(B) = X^{-1}(h^{-1}(B))$. If $\mathcal{B} \in \mathcal{B}_n$, then $h^{-1}(B) \in \mathcal{G}$, so it follows that $X^{-1}(h^{-1}(B)) \in \sigma(X)$. This shows $\sigma(Z) \subset \sigma(X)$.

□

The function h in $E[Y|X] = h \circ X$ is a Borel function on $(\Lambda, \mathcal{G})$. Let $x \in \Lambda$, $E[Y|X = x] = h(x)$ is a function on Λ, whereas $E[Y|X] = h \circ X$ is a function on Ω.

1.5.2 Formal Definition of Conditional Expectation

We have shown that $E[I_C h^*(X)] = E[I_C Y], \forall C \in \sigma(X)$ is necessary for h^* to be the "optimal" predictor of Y based on X . One can show that it is also sufficient. Realizing that $E[I_C h^*(X)] = E[I_C Y], \forall C \in \sigma(X)$ characterizes the "optimal" such predictor when Y has finite second moment allows us to generalize this notion of "optimal" predictor when X has only first moment. Also, notice that it only depends on the σ-field $\sigma(Y)$, so we can generalize the definition of conditional expectation to the situation where the given "information" is in the form of a σ-field (which may not often be the case in practical applications).

Definition 1.5.1. Let Y be an integrable r.v. on $(\Omega, \mathcal{F}, \mathbb{P}) \to (\mathbb{R}, \mathcal{B})$.

1. Suppose $\mathcal{G}$ be a sub-σ-field of $\mathcal{F}$. The conditional expectation of Y given $\mathcal{G}$ denoted by $E[Y|\mathcal{G}]$ is the a.s.-unique r.v. satisfying:

   (a) $E[Y|\mathcal{G}]$ is $\mathcal{G}$-measurable from $(\Omega, \mathcal{G}) \to (\mathbb{R}, \mathcal{B})$

   (b) $\int_A E[Y|\mathcal{G}] d\mathbb{P} = \int_A Y d\mathbb{P}$ for any $A \in \mathcal{G}$

2. Let $B \in \mathcal{F}$. The conditional probability of B given $\mathcal{G}$ is defined to be $\mathbb{P}[B|\mathcal{G}] = E[I_B|\mathcal{G}]$

3. Let X be measurable random element on $(\Omega, \mathcal{F})$, then $E[Y|X] = E[Y|\sigma(X)]$

Remark. 1. Note that $E[Y|\mathcal{G}]$ is a r.v., i.e. a mapping from $(\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B})$. Thus, $E[Y|\mathcal{G}](\omega) \in \mathbb{R}$ for each $\omega \in \Omega$.

2. Since $E[|Y|] < \infty$, we also have $E[|I_A Y|] < \infty$ for all $A \in \mathcal{G}$. So the r.h.s. of 1-(b) is defined and is a finite real number.

3. From a probabilistic point of view, one can say that $\sigma(X)$ "contains the information in X" useful for prediction of any r.v. Y. Note that from the observed value $X(\omega)$ one can only determine whether or not $\omega \in A$ if $A \in \sigma(X)$

Theorem 1.5.2. ***(Existence and uniqueness of conditional expectation)*** *Suppose Y is real valued and $E[|Y|] < \infty$, then $h(X) = E[Y|X]$ iff h is Borel and $\forall A \subset Range(X)$ and A is measurable, normal equation $E[I_A(X)h(X)] = E[I_A(X)Y]$. There exists an essentially unique $h(X)$ satisfying the previous condition. (that is, satisfying Definition 1.5.1)*

Before we prove the Theorem 1.5.2, we need the following definition for conditional distribution (and it will be used again later.) This definition requires too much details here and not so easily to be violated and find out a counterexample.

Definition 1.5.2. Let $Y : (\Omega, \mathcal{F}, \mathbb{P}) \to (\Lambda_1, \mathcal{G}_1)$ and $X : (\Omega, \mathcal{F}, \mathbb{P}) \to (\Lambda_2, \mathcal{G}_2)$ be random elements. A **family of conditional distribution** for Y given $X = x$ is a function $P_{Y|X} : \mathcal{G}_1 \times \Lambda \to [0, 1]$ satisfying

1. For all $x \in \Lambda_2$, the range of X, $P_{Y|X}(\cdot|x)$ is a probability measure on $(\Lambda_1, \mathcal{G}_1)$
2. For all $A \subset \mathcal{G}_1$, the range of Y, $P_{Y|X}(A|x)$ is a is a version of $P[Y \in A|X = \cdot] = E[I_A(Y)|X = x]$

When such a $P_{Y|X}(A|x)$ exists, we shall write it as $P_{Y|X}(A|X = x)$.

Proof. **Proof of Theorem 1.5.2**: The goals are to show that $\exists h : Range(X) \to \mathbb{R}$ s.t. $\forall$ measurable $\lambda \subseteq Range(X)$, $E[I_A(X)h(X)] = E[I_A(X)Y]$. First assume $Y \geq 0$, define a measure ν on $(\Omega, \mathcal{F})$ by $\nu(A) = \int_A Y d\mathbb{P} = E[I_A Y]$. Then ν is a measure on Ω with density Y w.r.t. $\mathbb{P}$. That is, $\nu \ll \mathbb{P}$ and $\dfrac{d\nu}{d\mathbb{P}} = Y$ almost surely. Here note that we will get a function from range of X to the real number, Y is from underlying probability space to real number.

$$\begin{tikzcd}
(\Lambda, \mathcal{G}) \arrow[rr, "{E[Y|X = \cdot]}"] & & (\mathbb{R}, \mathcal{B}) \\
& (\Omega, \mathcal{F}) \arrow[ul, "X"] \arrow[ur, "{E[Y|X]}"'] &
\end{tikzcd}$$

First, suppose $A \subseteq Range(X)$, we have $P_X(A) = 0 = \mathbb{P}[X^{-1}(A)]$. $\nu \circ X^{-1}(A) = \int_{X^{-1}(A)} Y d\mathbb{P} = \int I_A(X) Y d\mathbb{P} = 0$. Thus, $\nu \circ X^{-1}(A) = 0$. We can check the claim that the induced measure is dominated by the distribution of X ($(\nu \circ X^{-1}) \ll P_X$). Therefore the R-N derivative exists that $\exists h$ s.t. $hd(\nu \circ X^{-1}) = hdP_X$. Now $E[I_A(X)h(X)] = \int I_A(x)h(x)dP_X(x) = \int I_A(x)h(x)d(\nu \circ X^{-1})(x) = \int I_{X^{-1}(A)} d\nu = \int I_{X^{-1}(A)} Y d\mathbb{P} = \int I_A(X) Y d\mathbb{P} = E[I_A(X)Y]$. So $h(X) = E[Y|X]$.

For general Y, we apply another version below. Let ν_0 and $\mathbb{P}_0$ denote the restrictions of ν and $\mathbb{P}$ to $\mathcal{G}$, i.e. ν_0 is the measure on $(\Omega, \mathcal{G})$ given by $\nu_0(A) = \nu(A)$ for all $A \in \mathcal{G}$. Then we still have $\nu_0 \ll \mathbb{P}_0$, but not necessarily that $\dfrac{d\nu_0}{d\mathbb{P}_0} = Y$ since Y is not necessarily $\mathcal{G}$-measurable, i.e. we may not have $\sigma(Y) \subset \mathcal{G}$. However, by the Radon-Nikodym theorem (note that $\mathbb{P}_0$ is trivially σ–finite) we have that there is a r.v. $h^*(X) = \dfrac{d\nu_0}{d\mathbb{P}_0}$, $d\mathbb{P}_0$-a.s. such that $h^*(X)$ is $\mathcal{G}$-measurable (i.e. property (i) of the definition holds) and

$$\nu_0(A) = \int_A h^*(X) d\mathbb{P}_0, \forall A \in \mathcal{G}$$

Since $\nu_0(A) = \nu(A) = \int_A h^*(X) d\mathbb{P}$, we have

$$\int_A Y d\mathbb{P} = \int_A h^*(X) d\mathbb{P}_0, \forall A \in \mathcal{G} \tag{1.3}$$

Now we claim that for any r.v. W on $(\Omega, \mathcal{G}, \mathbb{P}_0)$, $\int W d\mathbb{P}_0 = \int W d\mathbb{P}$. (Note that W is automatically a r.v. on $(\Omega, \mathcal{F}, \mathbb{P})$.) This is certainly true if W is an indicator by definition of $\mathbb{P}_0$, and then it follows immediately for simple functions by linearity of integrals. For $W \geq 0$, consider a sequence of $\mathcal{G}$-measurable simple functions $0 \leq \phi_n \uparrow W$ as in Proposition 1.2.6 and apply MCT. Finally, the general case (which we do not actually need here) follows from linearity and the decomposition of W into its positive and negative parts.

Hence, from (1.3) we have

$$\int_A Y d\mathbb{P} = \int_A h^*(X) d\mathbb{P}, \forall A \in \mathcal{G}$$

which is Definition 1.5.1-1(b).

If $h'(X)$ is any other r.v. satisfying 1.5.1-1 (a) and (b), then $h'(X) = \dfrac{d\nu_0}{d\mathbb{P}_0} = h^*(X)$, $\mathbb{P}_0$-a.s. by the essential uniqueness of Radon-Nikodym derivatives. Note that $\mathbb{P}_0$-a.s. implies $\mathbb{P}$-a.s. since a $\mathbb{P}_0$-null set is just a $\mathbb{P}$-null set which happens to belong to $\mathcal{G}$.

If we drop the restriction that $Y \geq 0$ but require $E[|Y|] < \infty$, then apply the previous argument to Y_+ and Y_- to obtain essentially unique r.v.'s $h^*_+(X)$ and $h^*_-(X)$ which are $\mathcal{G}$-measurable and satisfy

$$\int_A Y_+ d\mathbb{P} = \int_A h^*_+(X) d\mathbb{P}, \int_A Y_- d\mathbb{P} = \int_A h^*_-(X) d\mathbb{P}, \forall A \in \mathcal{G}$$

We claim $h^*_+(X)$ and $h^*_-(X)$ are both finite a.s. so that the r.v. $h^*(X) = h^*_+(X)$ and $h^*_-(X)$ is defined a.s. (i.e. it can be of the form $\infty - \infty$ only on a null set, and we may define it arbitrarily there). Now Y_+ and Y_- are both finite a.s., and if say $A = [h^*_+(X) = \infty]$ satisfied $\mathbb{P}(A) > 0$, then since $A = Z^{-1}(\{\infty\}) \in \mathcal{G}$, $\int_A Y_+ d\mathbb{P} = \int_A h^*_+(X) d\mathbb{P} = \infty$. However, since X is integrable, $\int_A Y_+ d\mathbb{P} \leq \int_A h^*_+(X) d\mathbb{P} < \infty$, a contradiction. This establishes the claim for $h^*_+(X)$ and the claim that $h^*_-(X) < \infty$ a.s. follows similarly.

Verification of properties (a) and (b) is easy. If $h'(X)$ is any other r.v. satisfying (a) and (b), then let $D = h^*(X) - h(X)$. Then D is $\mathcal{G}$–measurable, so $A = [D \geq 0]$ is in $\mathcal{G}$. Since both $h^*(X)$ and $h'(X)$ satisfy (b)

$$\int_\Omega I_A D d\mathbb{P} = \int_A h^*(X) d\mathbb{P} - \int_A h'(X) d\mathbb{P} = \int_A Y d\mathbb{P} - \int_A Y d\mathbb{P} = 0$$

However, $I_A D$ is a nonnegative function, so by Proposition 1.2.4-5, $I_A D = 0$ a.s. A similar argument shows $I_{A^c} D = 0$, a.s., and hence $h^*(X) = h'(X)$, a.s., which completes the proof. □

Remark. "Essentially unique" can always change $h(X) = E[Y|X]$ on sets of probability $\mathbb{P}$ 0, and can change $h(x) = E[Y|X = x]$ on a set of P_X measure 0. Similarly, conditional probability $P_{Y|X}(\cdot|x)$ can be changed on a set of x values having P_X measure 0.

Let $X : (\Omega, \mathcal{F}, \mathbb{P}) \to (\Lambda, \mathcal{G})$ be any random elements, and let X be an integrable r.v. $E[Y|X](\omega), \omega \in \Omega$ is a r.v. on Ω which is $\sigma(X)$-measurable by definition. Hence, by Theorem 1.5.1, there is a function $h : (\Lambda, \mathcal{G}) \to (\mathbb{R}, \mathcal{B})$ such that $h(X(\omega)) = (h \circ X)(\omega)$. Furthermore, this function is Law$[X]$-essentially unique in the sense that for some other $h' : (\Lambda, \mathcal{G}) \to (\mathbb{R}, \mathcal{B})$ implies that $h' = h$ Law$[X]$-a.s., i.e. $P[h(X) - h'(X)] = 1$. Any such version is defined to be the conditional expectation of Y given $X = x$, and denoted $E[Y|X = x] = h(x), x \in \Lambda$. The following picture may help the student keep matters clear:

$$\begin{tikzcd}
(\Lambda, \mathcal{G}) \arrow[rr, "{E[Y|X=\cdot]}"] & & (\mathbb{R}, \mathcal{B}) \\
& (\Omega, \mathcal{F}) \arrow[ul, "X"] \arrow[ur, "{E[Y|X]}"'] &
\end{tikzcd}$$

The notations here are very confusing for many students, so we will try to explain some of the subtleties. One difficulty is that $E[Y|X=x]$ is a function of $x \in \Lambda$ in our setup, and the argument of the function x does not appear in a convenient place. Indeed, in the defining equation above $E[Y|X=x]=h(x)$ where h is the function such that $E[Y|X=x]=h(x)$, if we substitute the random object X for x we obtain the seemingly nonsensical "$E[Y|X=X]=E[Y|X]$" The following may be a little clearer:

$$E[Y|X](\omega) = E[Y|X=X(\omega)] \tag{1.4}$$

The argument of the function $E[Y|X=\cdot]$ is whatever appears on the r.h.s. of the equals sign "=" after the conditioning bar "|". We do not call $E[Y|X=\cdot]$ a random variable in general since it is not a function defined on the underlying probability space $(\Omega, \mathcal{F}, \mathbb{P})$, although it is a function on the probability space $(\Lambda, \mathcal{G}, Law[X])$, so technically we could call it a random variable.

1.5.3 Examples of Conditional Expectations.

The definition of $E[XY|\mathcal{G}]$ is very unsatisfactory from an intuitive point of view, although it turns out to be very convenient from a formal mathematical point of view. In order to make it more appealing intuitively, we shall verify that it gives the "right answer" in a number of circumstances with which the student is already familiar.

Example 1.5.3. Suppose $A_1, \cdots, A_n$ are events which partition Ω (i.e. the A_i are mutually exclusive and $\Omega = \bigcup_{i=1}^{n} A_i$). Suppose $P(A_i) > 0$ for each i and $a_1, a_2, \cdots$ an are distinct real numbers. Let $X = \sum_{i=1}^{n} a_i I_{A_i}$ be a simple r.v. If Y is an integrable r.v., then

$$E[Y|X] = \sum_{i=1}^{n} \frac{\int_{A_i} Y d\mathbb{P}}{P(A_i)} I_{A_i}, a.s.$$

Consider the elementary case $n=2$ and $Y = I_B$ for some event B. Write $A = A_1$ and $A^c = A_2$. The values of a_1 and a_2 are irrelevant, as long as they are distinct, since any such X contains the same "information", namely the σ-field $\sigma(X) = \{\emptyset, A, A^c, \Omega\}$. We may take $X = I_A$ for simplicity. Then, according to the previous equation,

$$E[Y|X] = P[B|X] = \frac{\int_A I_B d\mathbb{P}}{P(A)} I_A + \frac{\int_{A^c} I_B d\mathbb{P}}{P(A^c)} I_{A^c}, a.s.$$

That is, almost surely

$$P[B|X](\omega) = \begin{cases} \dfrac{P(A \cap B)}{P(A)}, & \text{if } \omega \in A \\ \dfrac{P(A^c \cap B)}{P(A^c)}, & \text{if } \omega \in A^c \end{cases}$$

Note that for $\omega \in A$, $P[B|I_A](\omega) = P[B|A] = \dfrac{P(A \cap B)}{P(A)}$, with probability 1, where $P[B|A]$ denotes the "classical" or "elementary" conditional probability of B given A. Similarly, for $\omega \in A^c$, $P[B|I_A](\omega) = P[B|A^c]$ a.s. Thus, we have $P[B|I_A] = P[B|A]I_A + P[B|A^c]I_{A^c}$. Note that we have mixed meanings for conditional probability in the last display. The l.h.s. is the "sophisticated" type of conditional probability defined in Definition 1.5.1, whereas both conditional probabilities on the r.h.s. are of the elementary variety. It will always be clear when we intend elementary conditional probability (which is a fixed number) rather than our more sophisticated kind (which is a random variable) since the second member of the conditional probability operator will be a set in the case of elementary conditional probability but will be a random variable or a σ-field for the more sophisticated variety.

We may also express in terms of the "other" kind of conditional expectation. The reader should be able to check that $E[Y|X = x] = \sum_{i=1}^{n} \frac{\int_{A_i} Y d\mathbb{P}}{P(A_i)} I_{\{a_i\}}(x), Law[X]$-a.s.

Proof. Let Z denote the proposed $E[Y|X]$. Since $A_i = X^{-1}(\{a_i\})$, it follows that Z is $\sigma(X)$-measurable. In fact, one can show that $\sigma(X)$ is the collection of all unions of the A_i. For instance, if $B \in \mathcal{B}$, then $X^{-1}(B) = \bigcup_{\{i:a_i \in B\}} A_i$. Hence, if $A \in \sigma(X)$, say $A = X^{-1}(B)$ for $B \in \mathcal{B}$, then

$$\int_A Y d\mathbb{P} = \int_{X^{-1}(B)} Y d\mathbb{P} = \sum_{\{i:a_i \in B\}} \int_{A_i} Y d\mathbb{P}$$

Now,

$$\begin{aligned}\int_A Z d\mathbb{P} &= \sum_{\{i:a_i \in B\}} \int_{A_i} \sum_{j=1}^{n} \frac{\int_{A_j} Y(\omega_i) d\mathbb{P}(\omega_1)}{P(A_j)} I_{A_j}(\omega) d\mathbb{P}(\omega) \\ &= \sum_{\{i:a_i \in B\}} \sum_{j=1}^{n} \frac{\int_{A_j} Y(\omega_i) d\mathbb{P}(\omega_1)}{P(A_j)} \int_{A_i} I_{A_j}(\omega) d\mathbb{P}(\omega)\end{aligned}$$

Note in the last expression that when i is fixes in the outer summation, then $\int_{A_i} I_{A_j} d\mathbb{P}$ is nonzero only when $j = 1$ since otherwise A_i and A_j are disjoint. If $i = j$ then this integral is $P(A_i)$. Hence,

$$\int_A Z d\mathbb{P} = \sum_{i:a_i \in B} \frac{\int_{A_i} Y d\mathbb{P}}{P(A_i)} \int_{A_i} I_{A_i}(\omega) d\mathbb{P}(\omega) = \sum_{i:a_i \in B} \int_{A_i} Y d\mathbb{P}$$

These shows the Definition 1.5.1-1 (a) and (b). One virtue of the abstract definition of conditional expectation is that it allows us to make sense of $P[B|Y]$ even when $P[Y = y] = 0$ for any single value y. The next result makes this clearer. □

Proposition 1.5.3. *Suppose $X : (\Omega, \mathcal{F}, \mathbb{P}) \to (\Lambda_1, \mathcal{G}_1)$ and $Y : (\Omega, \mathcal{F}, \mathbb{P}) \to (\Lambda_2, \mathcal{G}_2)$ be random elements and μ_i is a σ-finite measure on $(\Lambda_i, \mathcal{G}_i)$ for $i = 1, 2$ s.t. $Law[X, Y] \ll \mu_1 \times \mu_2$. Let $f(x, y)$ denote the corresponding joint density. Let $g(x, y)$ be any Borel function $\Lambda_1 \times \Lambda_2 \to \mathbb{R}$ s.t, $E[g(X, Y)] < \infty$. Then*

$$E[g(X, Y)|X] = \frac{\int_{\Lambda_1} g(X, y) f(X, y) d\mu_2(y)}{\int_{\Lambda_1} f(X, y) d\mu_2(y)}, a.s.$$

Remark. Note that the denominator is $f_X(X)$, which is the marginal density of X w.r.t. μ_1. $P_{XY} \ll \mu_1 \times \mu_2$, μ_1 on *Range*(X), μ_2 on *Range*(Y), $f_{XY}(x, y) d(\mu_1 \times \mu_2)(x, y) = dP_{XY}(x, y)$ is the joint (Lebesgue) density. Define the conditional density of Y given X by

$$f_{Y|X}(y|x) = \begin{cases} \dfrac{f_{XY}(x, y)}{f_X(x)}, & \text{if } f_X(x) > 0 \\ 0, & \text{otherwise} \end{cases}$$

where $f_X(x) = \int f_{XY}(x,y)d\mu_2(y)$.

Note that for fixed x this is a density w.r.t. μ_2 for a probability measure on Λ_1, where X is a random element taking values in Λ_2. Proposition 1.5.3 may be rewritten as

$$E[g(X,Y)|X] = \int g(X,y)f_{Y|X}(y|X)d\mu_2(y)$$
$$E[g(X,Y)|X=x] = \int g(x,y)f_{Y|X}(y|x)d\mu_2(y)$$

Proof. It follows from Fubini's theorem that both of the functions $\int_{\Lambda_2} g(x,y)f(x,y)d\mu_2(y)$ and $\int_{\Lambda_2} f(x,y)d\mu_2(y)$ are measurable functions of x, and the second is positive Law$[X]$-a.s. (the set of x values where it is 0 has Law$[X]$ measure 0). If we define $h(x)$ to be the quotient of the first over the second (i.e. $h(X)$ is the function of X on the r.h.s. of Proposition 1.5.3), then $h(x)$ is defined Law$[Y]$-a.s. and is measurable from $\Lambda_1 \to \mathbb{R}$. As the r.h.s. of Proposition 1.5.3 equals $h(X) = h \circ X)$, it follows that the r.h.s. is $\sigma(X)$-measurable. This is property 1 of Definition 1.5.1.

Next we check the second property of Definition 1.5.1. Let $B \in \mathcal{B}_n$ so $X^{-1}(B)$ is a generic element of $\sigma(X)$. Then

$$\begin{aligned}\int_{X^{-1}(B)} h(X)d\mathbb{P} &= \int_B h(x)dP_X(x) = \int_B h(x)f_X(x)d\mu_1(x)\\ &= \int_B \frac{\int g(x,y)f(x,y)d\mu_2(y)}{f_X(x)} f_X(x)d\mu_1(x) = \int_B \int g(x,y)f(x,y)d\mu_2(y)d\mu_1(x)\\ &= \int_{\Lambda_2 \times B} g(x,y)f(x,y)d(\mu_1 \times \mu_2)(x,y) = \int_{X^{-1}(B)} g(X,Y)d\mathbb{P}\end{aligned}$$

□

We need that the normal equation holds. Take $A \subseteq Range(X)$,

$$\begin{aligned}E[I_A(x)h(x)] &= \int I_A(x)h(x)f_X(x)d\mu_1(x)\\ &= \int \left[\int g(x,y)f_{Y|X}(y|x)d\mu_2(y)\right] f_X(x)I_A(x)d\mu_1(x)\\ &= \int\int I_A(x)g(x,y)f_{Y|X}(y|x)f_X(x)d\mu_1(x)d\mu_2(y) \qquad \text{(by Fubini)}\end{aligned}$$

where $f_{Y|X}(y|x)f_X(x) = f_{XY}(x,y)$ is joint density , $\mu_1 \times \mu_2$-a.e. Since $N = \{(x,y) : f_X(x) = 0\}$ satisfies $P_{XY}(N) = 0$, $\int I_N f_{XY} d(\mu_1 \times \mu_2) = \int I_N f_X d\mu_1 = 0$. If we are looking the set $\{(x,y) : f_{XY}(x,y) > 0 \& f_X(x) = 0\}$ is a subset of N, $P_{XY}(\{(x,y) : f_{XY}(x,y) > 0 \& f_X(x) = 0\}) = 0$. $\int\int I_A(x)g(x,y)f_{Y|X}(y|x)f_X(x)d\mu_1(x)d\mu_2(y) = E[I_A(x)g(X,Y)]$ similar to integrate $h(X)$ previously defined. Thus, normal equation for $h(x)$ as given.

1.5.4 Conditional Distributions

We now consider conditional distributions in general cases where we may not have any p.d.f. Let X and Y be two random vectors defined on a common probability space. It is reasonable to consider $P[Y^{-1}(B)|X = x]$ as a candidate for the conditional distribution of Y, given $X = x$, where B is any Borel set. However, since conditional probability is defined almost surely, for any fixed x,$P[Y^{-1}(B)|X = x]$ may not be a probability measure. The first part of the following theorem (whose proof can be found in Billingsley (1986, pp. 460-461)) shows that there exists a version of conditional probability such that $P[Y^{-1}(B)|X = x]$ is a probability measure for any fixed x.

We need to recall this definition to proof the following

Definition 1.5.3. Let $X : (\Omega, \mathcal{F}, \mathbb{P}) \to (\Lambda_1, \mathcal{G}_1)$ and $Y : (\Omega, \mathcal{F}, \mathbb{P}) \to (\Lambda_2, \mathcal{G}_2)$ be random elements. A **family of conditional distribution** for X given $Y = y$ is a function $P_{Y|X} : \mathcal{G}_1 \times \Lambda \to [0, 1]$ satisfying

1. For all $x \in \Lambda_2$, the range of X, $P_{Y|X}(\cdot|x)$ is a probability measure on $(\Lambda_1, \mathcal{G}_1)$
2. For all $A \subset \mathcal{G}_1$, the range of Y, $P_{Y|X}(A|x)$ is a is a version of $P[Y \in A|X = \cdot] = E[I_A(Y)|X = x]$

When such a $P_{Y|X}(A|x)$ exists, we shall write it as $P_{Y|X}(A|X = x)$.

Proposition 1.5.4. *Suppose that the assumptions of Proposition 1.5.3 hold. Then we have*

1. *the family of regular conditional distributions $Law[Y|X = x]$ exists*
2. *$Law[Y|X = x] \ll \mu_2$ for $Law[X]$–almost all values of x*
3. *the Radon-Nikodym derivatives are given by*

$$\frac{dLaw[Y|X = x]}{d\mu_1}(x) = f_{Y|X}(y|x), \mu_1 \times \mu 2 \text{ - } a.e.,$$

where $f_{Y|X}(y|x)$ is the conditional density given in Proposition 1.5.3.

Proof. For all $B \in \mathcal{G}_1$,

$$P[Y \in B|X = x] = \int I_B(y) f_{Y|X}(y|x) d\mu_2(y)$$

This verifies (i) of Definition 1.5.2. Condition (ii) of the definition follows since $f_{Y|X}(y|x)$ is a probability density w.r.t. $d\mu_2(y)$ for each fixed $x \in \Lambda_1$, i.e. $f_{Y|X}(y|x) \geq 0$ for all x and y, and for all y, $\int f_{Y|X}(y|x) d\mu_2(y) = 1$. □

Remark. The reader may find the definition and previous result very puzzling. After all, is it not obvious that conditional probability distributions exist? The answer is, "No," but it is also not obvious why they should not automatically exist. To explain, suppose $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space and $\mathcal{G}$ is a sub-σ-field of F. Then for each event $A \in \mathcal{F}$, the conditional probability $P[A|\mathcal{G}] = E[I_A|\mathcal{G}]$ is an almost surely uniquely defined r.v. Fix $\omega \in \Omega$. Does it follow that $P[A|\mathcal{G}](\omega)$ is a probability measure when considered as a function of the event A? Given that $P[A|\mathcal{G}](\cdot)$ may be modified arbitrarily on P-null sets (as long as it is done in a $\mathcal{G}$-measurable way), clearly we may not use any version of the family of r.v.'s $\{P[A|\mathcal{G}](\cdot) : A \in \mathcal{F}\}$ and obtain a family of probability measures $\{P[\cdot|\mathcal{G}](\omega) : \omega \in \Omega\}$. In general, such versions of $P[A|\mathcal{G}](\cdot)$ may not exist. Like a number of issues in measure theory, (e.g. the existence of subsets of $\mathbb{R}$ which are not Borel measurable) the nonexistence of conditional probability distributions is a technical detail which is of little importance in statistics. The next theorem shows that conditional distributions exist for the settings we shall encounter in this book. For further discussion of the difficulties involved with obtaining a family of conditional probability distributions (including counterexamples wherein they don't exist), see Ash or Brieman. Exercise 33.13, p. 464 of Billingsley provides a specific example.

Theorem 1.5.5. 1. *(Existence of conditional distribution): Let Y be a random n-vector on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and $\mathcal{A}$ be a sub-σ-field of $\mathcal{F}$. Then there exists a function $P(B, \omega)$ on $\mathcal{B}_n \times \Omega$ such that* *(a)* $P(B,\omega) = P[Y^{-1}(B)|\mathcal{A}]$ *a.s. for any fixed* $B \in \mathcal{B}_n$

(b) $P(\cdot,\omega)$ *is a probability measure on* $(\mathbb{R}^n, \mathcal{B}_n)$ *for any fixed* $\omega \in \Omega$

Let Y *be measurable from* $\Omega, \mathcal{F}, \mathbb{P}) \to (\Lambda, \mathcal{G})$. *Then there exists a family* $P_{Y|X}(B|x)$ *such that*

(a) $P_{Y|X}(B|x) = P[Y^{-1}(B)|X = x]$ *a.s.* P_X *for any fixed* $B \in \mathcal{B}_n$

(b) $P_{Y|X}(\cdot|x)$ *is a probability measure on* $(\mathbb{R}^n, \mathcal{B}_n)$ *for any fixed* $y \in \Lambda$

Furthermore, if $E[g(X,Y)] < \infty$ *with a Borel function* g, *then*

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

2. ***(Two stage experiment theorem):*** *Let* $(\Lambda, \mathcal{G}, \mathbb{P}_1)$ *be a probability space. Suppose that* $\mathbb{P}_2$ *is a function from* $\mathcal{B}_n \times \Lambda \to \mathbb{R}$ *and satisfies*

   (a) $\mathbb{P}_2(x,\cdot)$ *is a probability measure on* $(\mathbb{R}^n, \mathcal{B}_n)$ *for any* $x \in \Lambda$

   (b) $\mathbb{P}_2(\cdot, A)$ *is Borel for* $A \in \mathcal{B}_n$

   There is a unique probability measure $\mathbb{P}$ *on* $(\mathbb{R}^n \times \Lambda, \sigma(\mathcal{B}_n \times \mathcal{G}))$ *s.t. for* $A \in \mathcal{B}_n$ *and* $B \in \mathcal{G}$, $\mathbb{P}(A \times B) = \int_B \mathbb{P}_2(x,A)d\mathbb{P}_1(x)$. *Furthermore, if* $(\Lambda, \mathcal{G}) = (\mathbb{R}^m, \mathcal{B}^m)$, *and* $X(x,y) = x$ *and* $Y(x,y) = y$ *define the coordinate random vectors, then* $P_X = P_1, P_{Y|X}(x|\cdot) = P_2(x,\cdot)$, *and the probability measure above is the joint distribution of* (X,Y), *which has the following joint c.d.f.* $F(x,y) = \int_{(-\infty,x]} P_{Y|X}((-\infty,y)|z)dP_X(z)$, $x \in \mathbb{R}^n, y \in \mathbb{R}^m$

The proof of this theorem may be found in Breiman. It also follows from Theorems 33.3, p. 460, and Theorem 34.5, p. 471 of Billingsley. Comparison of Proposition 1.5.5 and Theorem 1.5.6 demonstrates the usual situation in statistics: in spite of the difficulty of proving a general result like Theorem 1.5.6, with a few more "concrete" assumptions as in Proposition 1.5.5, one can "barehandedly" construct the conditional distribution.

Remark. 1. Now we rewrite the previous Theorem 1.5.5-2 and outline the usual procedure for rigorously "deriving" a conditional distribution. One typically has a "candidate" for the conditional distribution $P_{Y|X}$, and it is necessary to verify that it satisfies the defining properties. The "candidate" comes from previous experience with elementary conditional probabilities or conditional densities, or from intuition. A candidate for $P_{Y|X}$ must be a function of the form $p(x,A)$ where A varies over measurable sets in the range of Y and x varies over elements in the range of X . Then there are basically three conditions that must be verified:

(a) $\forall x \in \Lambda$, $p(x,\cdot)$ is a probability measure on $(\Lambda_1, \mathcal{G}_1)$

(b) $\forall A \in \mathcal{G}_1$, $p(\cdot, A)$ is Borel measurable $(\Lambda_2, \mathcal{G}_2) \to (\mathbb{R}, \mathcal{B})$ for each fixed $B \in \mathcal{B}_n$

(c) $\forall B \in \mathcal{G}_2$ and $\forall A \in \mathcal{G}_1$, $P[Y \in B \& X \in A] = \int_A p(x,A)dLaw[X](x)$.

Now condition (a) here is simply a restatement of condition 2 in Definition 1.5.2, and conditions (b) and (c) together amount to condition 1 in Definition 1.5.2. Note that (b) means that $p(X,A)$ is a $\sigma(X)$ measurable r.v. as required in item 1 of the definition of Definition 1.5.1. We will show that (c) here is simply a restatement of the integral condition in item 2 of Definition 1.5.1. Now according to that condition in Definition 1.5.1, we should have

$$\forall B \in \mathcal{G}_2, \int_{[Y\in B]} p(X(\omega),A)dP(\omega) = \int_{[Y\in B]} I_{[X\in A]}(\omega)dP(\omega)$$

To explain, $[X \in A]$, which is another way of denoting $\{\omega \in \Omega : X(\omega) \in A\} = X^{-1}(A)$ is a generic element of $\sigma(X)$. Also, recall that $P[C|\mathcal{G}] = E[I_C|\mathcal{G}]$, so we use an indicator for Y in Definition 1.5.1. Now

$$\begin{aligned}\int_{[Y\in B]} I_{[X\in A]}(\omega)dP(\omega) &= \int I_{[Y\in B]}I_{[X\in A]}(\omega)dP(\omega) = \int I_{[Y\in B]\cap[X\in A]}(\omega)dP(\omega)\\ &= P[Y \in B \& X \in A]\end{aligned}$$

Also, by the change of variables,

$$\int_{[X\in A]} p(X(\omega), A)dP(\omega) = \int_A p(x, A)dLaw[X](x)$$

This completes the verification that (3) here is the same as condition 2 in Definition 1.5.1.

Condition (a) here is usually easy to check. We generally regard condition (b) as automatic – any function $p(x, A)$ that you can "write down" (e.g. as a formula in x) is measurable. So, any difficulties usually come in verification of condition (c).

2. Note that y is the only variable of integration in Theorem 1.5.5-1(b), and both sides are functions of x. This should be clear because y occupies the site in the function where the measurable set would go when evaluating its measure. The notation is not entirely desirable, and it is perhaps preferable to write

$$E[h(X,Y)|X = x] = h(x, y)P_{Y|X}(dy|X = x)$$

This makes clearer the variable of integration, and it is more consistent perhaps that a "differential set" dy should occupy the set argument than a regular variable. However, putting the d in front of the measure is much more convenient for the mnemonics of Radon-Nikodym derivatives, which is why we chose this convention. We shall use the convention as the above equation for clarity on occasion.

For a fixed x, $P_{Y|X=x} = P_{Y|X=x}(x|\cdot)$ is called the conditional distribution of Y given $X = x$. Under the conditions in Theorem 1.5.5-1, if X is a random m-vector and (X, Y) has a p.d.f. w.r.t. $\mu_1 \times \mu_2$ (μ_1 and μ_2 are σ-finite measures on $(\mathbb{R}^n, \mathcal{B}_n)$ and $(\mathbb{R}^m, \mathcal{B}^m)$, respectively), then $f_{Y|X}(y|x)$ is the p.d.f. of $P_{Y|X=x}$ w.r.t. μ_1 for any fixed x.

The second part of Theorem 1.5.5-1 states that given a distribution on one space and a collection of conditional distributions (which are conditioned on values of the first space) on another space, we can construct a joint distribution in the product space. It is sometimes called the "two-stage experiment theorem" for the following reason. If $X \in \mathbb{R}^n$ is selected in stage 1 of an experiment according to its marginal distribution $P_X = P_1$, and Y is chosen afterward according to a distribution $P_1(x, \cdot)$, then the combined two-stage experiment produces a jointly distributed pair (X, Y) with distribution $P(X, Y)$ and $P(Y|X = x) = P_1(x, \cdot)$. This provides a way of generating dependent random variables. The following is an example.

Example 1.5.4. A market survey is conducted to study whether a new product is preferred over the product currently available in the market (old product). The survey is conducted by mail. Questionnaires are sent along with the sample products (both new and old) to N customers randomly selected from a population, where N is a positive integer. Each customer is asked to fill out the questionnaire and return it. Responses from customers are either 1 (new is better than old) or 0 (otherwise). Some customers, however, do not return the questionnaires. Let Y be the number of ones in the returned questionnaires. What is the distribution of Y?

If every customer returns the questionnaire, then (from elementary probability) Y has the binomial distribution $Bi(p, N)$ (assuming that the population is large enough so that customers respond independently), where $p \in (0, 1)$ is the overall rate of customers who prefer the new product. Now, let X be the number of customers who respond. Then X is random. Suppose that customers respond independently with the same probability $\pi \in (0,1)$. Then P_Y is the binomial distribution $Bi(\pi, N)$. Given $X = x$ (an integer between 0 and N), $P_{Y|X=x}$ is the binomial distribution $Bi(p,x)$ if $x \geq 1$ and the point mass at 0 if $y = 0$. Binomial distributions have p.d.f.'s w.r.t. counting measure, we obtain that the joint c.d.f. of (X,Y) is

$$\begin{aligned} F(x,y) &= \sum_{k=0}^{x} P_{Y|X=k}((-\infty, y]) \binom{N}{k} \pi^k (1-\pi)^{N-k} \\ &= \sum_{k=0}^{x} \sum_{j=0}^{\min(y,k)} \binom{k}{j} p^j (1-p)^{k-j} \binom{N}{k} \pi^k (1-\pi)^{N-k} \end{aligned}$$

for $x = 0, 1, \cdots, y$, and $y = 0, 1, \cdots, N$. The marginal c.d.f. $F_Y(y) = F(\infty, y) = F(N, y)$. The p.d.f. of Y w.r.t. counting measure is

$$\begin{aligned} f_Y(y) &= \sum_{k=y}^{x} \binom{k}{y} p^y (1-p)^{k-y} \binom{N}{k} \pi^k (1-\pi)^{N-k} \\ &= \binom{N}{y} (\pi p)^y (1-\pi p)^{N-x} \sum_{k=y}^{N} \binom{N-y}{k-y} \left(\frac{\pi - \pi p}{1 - \pi p} \right)^{k-y} \left(\frac{1-\pi}{1-\pi p} \right)^{N-k} \\ &= \binom{N}{y} (\pi p)^y (1-\pi p)^{N-x} \end{aligned}$$

is the binomial distribution of $Bi(\pi p, N)$

1.5.5 Results on Conditional Expectation

Theorem 1.5.6. *Let X, Y, Y_1, and Y_2 be integrable r.v.'s on $(\Omega, \mathcal{F}, \mathbb{P})$, and let $\mathcal{G}$ be a fixed sub–σ–field of $\mathcal{F}$.*

1. ***(Conditional expectation of constant r.v.):*** *If $Y = k$ a.s. $k \in \mathbb{R}$, then $E[Y|\mathcal{G}] = k$ a.s.*

   Proof. follows 6. □

2. ***(Monotonicity):*** *If $Y_1 \leq Y_2$ a.s., then $E[Y_1|\mathcal{G}] \leq E[Y_2|\mathcal{G}]$ a.s.*

   Proof. It suffices to show that $Y \geq 0$ a.s. implies $E[Y|\mathcal{G}] \geq 0$ a.s. by taking $Y = Y_2 - Y_1$, this was shown in the proof of Theorem 1.5.2. Suppose $Y \geq 0$ a.s., and we will show that $E[Y|X] \geq 0$ a.s. Let $A = \{x : E[Y|X = x] < 0\}$. Then by the "normal equations" for $E[Y|X]$, $E[I_A(X)E[Y|X]] = E[I_A(X)Y] \geq 0$, where the last inequality follows since $I_A(X)Y \geq 0$ a.s. Since $I_A(X)E[Y|X] \leq 0$ by definition of A, it follows from Prop. 1.2.4(b) ($f \geq 0, \int f dm = 0 \implies f = 0, \mu$-a.e.) that $I_A(X)E[Y|X] = 0$ a.s. Since $E[Y|X] < 0$ when $I_A(X) > 0$ by definition of A, it follows $I_A(X) = 0$ a.s., which implies $E[Y|X] \geq 0$ a.s. If $Y_1 \leq Y_2$ a.s., then applying $Y_2 - Y_1$ and using linearity, we get $E[Y_2 - Y_1|X] \geq 0 \implies E[Y_1|X] \leq E[Y_2|X]$ a.s. □

3. ***(Linearity):*** *If $a_1, a_2 \in \mathbb{R}$, then $E[a_1Y_1 + a_2Y_2|\mathcal{G}] = a_1E[Y_1|\mathcal{G}] + a_2E[Y_2|\mathcal{G}]$ a.s.*

   Proof. First show $E[aY|\mathcal{G}] = aE[Y|\mathcal{G}]$ a.s. Clearly $aE[Y|\mathcal{G}]$ is $\mathcal{G}$-measurable, and for $A \in \mathcal{G}$, $\int aE[Y|\mathcal{G}]d\mathbb{P} = a\int E[Y|\mathcal{G}]d\mathbb{P} = a\int Y d\mathbb{P} = \int_A (aY)d\mathbb{P}$ Then, verify "normal equation", $E[I_C(X)(a_1E[Y_1|X]+a_2E[Y_2|X])] = a_1E[I_C(X)E[Y_1|X]] + a_2E[I_C(X)E[Y_2|X]] = a_1E[I_C(X)Y_1] + a_2E[I_C(X)Y_2] = E[I_C(X)(a_1Y_1 + a_2Y_2)]$ □

4. ***(Law of Total Expectation):*** $E[E[Y|\mathcal{G}]] = E[Y]$

   Proof. Taking $A = \Omega \in \mathcal{G}$ in of Definition 1.5.1-1(b) □

5. ***(Conditional Expectation given degenerated r.v.):*** $E[Y|\{\emptyset, \Omega\}] = E[Y]$

6. *If $\sigma(Y) \subset \mathcal{G}$, then $E[Y|\mathcal{G}] = Y$ a.s.*

7. ***(Law of Successive Conditioning):*** *If $\mathcal{G}_\infty$ is a sub–σ–field of $\mathcal{G}$, then $E[E[X|\mathcal{G}]|\mathcal{G}_\infty] = E[E[X|\mathcal{G}_\infty]|\mathcal{G}] = E[X|\mathcal{G}_\infty]$ a.s.*

   Proof. Taking $A \subset Range(\phi)$, then $E[I_A(\phi(X))E[Y|X]] = E[I_{\phi^{-1}(A)}(X)E[Y|X]] = E[I_{\phi^{-1}(A)}(X)Y] = E[I_A(\phi(X))Y] = E[I_A(\phi(X))]E[Y|\phi(X)]$ The last calculation shows that $E[Y|\phi(X)]$ satisfies the normal equations that define $E[E[Y|X]|\phi(X)]$. The other equation follows from 6. since $E[Y|\phi(X)]$ is already a function of X. □

8. *If $\sigma(Y_1) \subset \mathcal{G}$ and $E|Y_1Y_2| < \infty$, then $E[Y_1Y_2|\mathcal{G}] = Y_1E[Y_2|\mathcal{G}]$ a.s.*

   Proof. If Y_1 is $\mathcal{G}$-measurable, then we may treat it the same as a constant when computing $E[Y_1Y_2|\mathcal{G}]$. Clearly $Y_1E[Y_2|\mathcal{G}]$ is $\mathcal{G}$-measurable. We will verify property 2 of the definition only when X_1 is an $\mathcal{G}$-measurable simple function, say $Y_1 = \sum a_i I_{A_i}$ for $A_i \in \mathcal{G}$. In this case, for $A_i \in \mathcal{G}, \int_A Y_1E[Y_2|\mathcal{G}]d\mathbb{P} = \sum a_i \int_{A\cap A_i} E[Y_2|\mathcal{G}]d\mathbb{P} = \sum a_i \int_{A\cap A_i} Y_2 d\mathbb{P} = \int_A Y_1Y_2 d\mathbb{P}$. The second equality follows since $A \cap A_i \in \mathcal{G}$.

   First show $E[\psi(X)Y] = E[\psi(X)E[Y|X]]$ by starting with ψ simple. $E[\psi(X)Y|X] = \psi(X)E[Y|X]$ a.s. true for $I_A(X) = \psi(X)$. The normal equations for $E[I_A(X)Y|X]$, $\forall B$, $E[I_B(X)I_A(X)Y] = E[I_B(X)E[I_A(X)Y|X]] = E[I_B(X)I_A(X)E[Y|X]]$. Thus, $I_A(X)E[Y|X]$ satisfies for $E[I_A(X)Y|X]$, Also, the linearity holds for simple function. Then apply MCT & simple function approximation for $\psi \geq 0$, then $\psi = \psi_+ - \psi_-$. After this it is easy to check that $\psi(X)E[Y|X]$ satisfies the normal equations for $E[\psi(X)Y|X]$. □

9. *If X and Y are independent and $E|g(X,Y)| < \infty$ for a Borel function g, then $E[g(X,Y)|X = x] = E[g(x,Y)]$ a.s. P_X.*

10. *If $E[Y^2] < \infty$, then $E[Y|\mathcal{G}]^2 \leq E[Y^2|\mathcal{G}]$ a.s.*

11. ***(Monotone Convergence Theorem):*** *If $Y_n \geq 0$ for any n, then $E[\liminf_n Y_n|\mathcal{G}] \leq \liminf_n E[Y_n|\mathcal{G}]$ a.s. If $0 \leq Y_i \uparrow Y$ a.s. then $E[Y_i|\mathcal{G}] \uparrow E[Y|\mathcal{G}]$ a.s.*

    Proof. Clearly $\lim E[Y_i|\mathcal{G}]$ is a $\mathcal{G}$-measurable r.v. by Proposition 1.2.1 (c). If $A \in \mathcal{G}$ then $I_AE[Y_i|\mathcal{G}]$ is a nonnegative increasing sequence of functions so by two applications of the ordinary Monotone Convergence Theorem, $\int_A \lim E[Y_i|\mathcal{G}]d\mathbb{P} = \lim \int_A E[Y_i|G]d\mathbb{P} = \lim \int_A Y_i d\mathbb{P} = \int_A Y d\mathbb{P}$. The result follows from the essential uniqueness of conditional expectations. □

12. ***(Dominated Convergence Theorem):*** *Suppose there is an integrable r.v. X s.t. $Y_i \leq X$ a.s. for all i and suppose that $Y_i \to Y$ a.s. Then $E[Y_i|\mathcal{G}] \to E[Y|\mathcal{G}]$*

We can replace all the σ-field above with a random variable / random vector. It can also be shown that Holder's inequality, Liapounov's inequality, Minkowski's inequality, and Jensen's inequality hold a.s. with the expectation E replaced by the conditional expectation $E(\cdot|\mathcal{G})$.

Example 1.5.5. Recall the MSPE in the beginning of this section. Let Y be a random variable on $(\Omega, \mathcal{F}, \mathbb{P})$ with $E[Y^2] < \infty$ and let X be a measurable function from $(\Omega, \mathcal{F}, \mathbb{P}) \to (\Lambda, \mathcal{G})$. One may wish to predict the value of Y based on an observed value of X. Let $g(X)$ be a predictor, i.e., $g \in \mathcal{N} = \{\text{all Borel functions } g \text{ with } E[g(Y)]^2 < \infty\}$. Each 2 predictor is assessed by the "mean squared prediction error" $E[Y - g(X)]$. We now show that $E(Y|X)$ is the best predictor of Y in the sense that

$$E[Y - E(Y|X)] = \min_{g \in \mathcal{N}} E[Y - g(X)]^2$$

First, Theorem 1.5.6-8 implies $E(Y|X) \in \mathcal{N}$. Next, for any $g \in \mathcal{N}$,

$$\begin{aligned}
E[Y - g(X)]^2 &= E[Y - E(Y|X) + E(Y|X) - g(X)]^2 \\
&= E[Y - E(Y|X)]^2 + E[E(Y|X) - g(X)]^2 + 2E\{[Y - E(Y|X)][E(Y|X) - g(X)]\} \\
&= E[Y - E(Y|X)]^2 + E[E(Y|X) - g(X)]^2 + 2E\{E\{[Y - E(Y|X)][E(Y|X) - g(X)]|Y\}\} \\
&\quad \text{(by Theorem 1.5.6-4 Law of Total Expectation)} \\
&= E[Y - E(Y|X)]^2 + E[E(Y|X) - g(X)]^2 + 2E\{[[E(Y|X) - g(X)]E[Y - E(Y|X)|Y]\} \\
&\quad \text{(by Theorem 1.5.6-8)} \\
&= E[Y - E(Y|X)]^2 + E[E(Y|X) - g(X)]^2 \\
&\quad \text{(by Theorem 1.5.6-1,3,4: Conditional expectation of constant r.v.; Linearity; Law of Total Expectation)} \\
&\geq E[Y - E(Y|X)]^2,
\end{aligned}$$

Theorem 1.5.7. ***(Conditional Expectation and Independence):*** *Suppose Y is an integrable r.v. and X_1 and X_2 are random vectors with (Y, X_1) independent of Y_2. Then $E[Y|X_1, X_2] = E[Y|X_1]$ a.s. In particular, $E[Y|X_2] = E[Y]$ a.s. From an intuitive point of view, X_2 provides no information about Y if they are independent, so it is reasonable that the conditional expectation of Y given X_2 not depend on X_2.*

Proof. First, $E(Y|X_1)$ is Borel on $(\Omega, \sigma(X_1, X_2))$, since $\sigma(X_1) \subset \sigma(X_1, X_2)$. Next, we need to show that for any Borel set $B \in \mathcal{B}_{k_1+k_2}$,

$$\int_{(Y_1,Y_2)^{-1}(B)} X d\mathbb{P} = \int_{(Y_1,Y_2)^{-1}(B)} E[Y|X_1] d\mathbb{P}$$

If $B = B_1 \times B_2$, where $B_i \in \mathcal{B}_k$, then $(Y_1, Y_2)^{-1}(B) = Y_1^{-1}(B_1) \cap Y_2^{-1}(B_2)$ and

$$\begin{aligned}
\int_{Y_1^{-1}(B_1) \cap Y_2^{-1}(B_2)} E[Y|X_1] d\mathbb{P} &= \int I_{Y_1^{-1}(B_1)} I_{Y_2^{-1}(B_2)} E[Y|X_1] d\mathbb{P} \\
&= \int I_{Y_1^{-1}(B_1)} E[Y|X_1] d\mathbb{P} \int I_{Y_2^{-1}(B_2)} d\mathbb{P} = \int I_{Y_1^{-1}(B_1)} X d\mathbb{P} \int I_{Y_2^{-1}(B_2)} d\mathbb{P} \\
&= \int I_{Y_1^{-1}(B_1)} I_{Y_2^{-1}(B_2)} X d\mathbb{P} = \int_{Y_1^{-1}(B_1) \cap Y_2^{-1}(B_2)} X d\mathbb{P}
\end{aligned}$$

This shows that $\int_{(Y_1,Y_2)^{-1}(B)} X d\mathbb{P} = \int_{(Y_1,Y_2)^{-1}(B)} E[Y|X_1] d\mathbb{P}$ holds for $B = B_1 \times B_2$. We can show that the collection $H = \{B \subset \mathbb{R}^{k_1+k_2} : Bsatisfies \int_{(Y_1,Y_2)^{-1}(B)} X d\mathbb{P} = \int_{(Y_1,Y_2)^{-1}(B)} E[Y|X_1] d\mathbb{P}\}$ is a σ-field. Since we have already shown that $\mathcal{B}_{k_1} \times \mathcal{B}_{k_2} \subset H, \mathcal{B}_{k_1+k_2} = \sigma(\mathcal{B}_{k_1} \times \mathcal{B}_{k_2}) \subset H$ and thus the result follows. □

Theorem 1.5.8. ***(Bayes Formula):*** *Suppose $\Theta : (\Omega, \mathcal{F}, \mathbb{P}) \to (\Lambda_2, \mathcal{G}_2)$ is a random element and let λ be a σ-finite measure on $(\Lambda_2, \mathcal{G}_2)$ such that $Law[\Theta] \ll \lambda$. Denote the corresponding density (**Prior Density**) by $\pi(\theta) = \dfrac{dLaw[\Theta]}{d\lambda}(\theta)$ where π is the probability measure on range of Θ (parameter space), $\pi \ll \lambda$. Let μ be a σ-finite measure on $(\Lambda_1, \mathcal{G}_1)$. Suppose that for each $\theta \in \Lambda_1$ there is given a probability density function w.r.t.*

μ denoted $f(\cdot|\theta)$. Denote by X a random element taking values in Λ_1 with $\frac{dLaw[X|\Theta=\theta]}{d\mu} = \frac{dP_{X|\Theta}(\cdot|\theta)}{d\mu} = f(\cdot|\theta)$. Then there is a version of ***posterior density*** *$Law[\Theta|X=x]$ given by*

$$\pi(\theta|x) = \frac{dLaw[\Theta|X=x]}{d\lambda}(\theta) = \frac{f(x|\theta)\pi(\theta)}{\int_{\Lambda_2} f(x|\tilde{\theta})\pi(\tilde{\theta})d\lambda(\tilde{\theta})}$$

Proof. By Section 1.5.4 Conditional distribution with $h(x,\theta) = f(x|\theta)$, there is a joint distribution for (X,Θ) for which $\pi(\theta)$ is the marginal density of θ w.r.t. λ, $f(x|\theta)\pi(\theta)$ is the joint density for (X,Θ), and $f(x|\theta)$ is the conditional density for X given $\Theta=\theta$. There only remains to verify the formula for the conditional density of Θ given $X=x$. But the marginal density for X is the joint density with θ integrated out, i.e. $\int_{\Lambda_2} f(x|\theta)\pi(\theta)d\lambda(\theta)$. Thus, one recognizes the r.h.s. of the formula as the joint density divided by the marginal for X, i.e. the conditional density for Θ given $X=x$. □

1.5.6 Discussion

Why do we call normal equation? How to solve a general prediction loss function? (It will be useful in learning problem that we have a conditional expectation on objective function, and in CAAM Optimization course.) How to compute $E[Y|X]$ and $P_{Y|X}$ with nontrivial example?

Example 1.5.6. Let X be a $\mathbb{R}$-valued r.v. and put $Y = X^2$. Assume P_X has a Lebesgue density $f(x)$. Find $P-X|Y$. It may not be obvious, but there is no joint density for P_{XY} w.r.t. a product measure, so we can't do the "usual" type of calculation. Note that $P_{XY}(Q) = 1$, where Q is the parabola $Q = \{(x,y) : y = x^2\}$. Now $m^2(Q) = 0$ so we know P_{XY} doesn't have a density w.r.t. m. Indeed, with a lot more work, one can show it doesn't have a joint density w.r.t. any product measure (with the factor measures being σ-finite, of course). So, what are we to do? Use our intuitive understanding of conditioning to guess the result, then verify that it works.

Given $Y = y$, we know X took on one of two values, $\pm\sqrt{y}$. One way of thinking about conditioning and Lebesgue densities goes back to Feller's Volume 2 in his classic introduction to probability. We will present an "engineering" version of that argument. It is similar to treatments of conditioning based on non-standard analysis, where differentials are respectable mathematical objects.

Assume for now the density $f(x)$ is continuous. (Once we have the right answer, the continuity will turn out not to matter.) In reality, we can never observe a continuous r.v.: that would mean having an infinite number of digits after the decimal point. Let X denote the rounded off version of X and $\frac{\delta}{2}$ the maximum roundoff error. For example, if we round off to 2 decimal digits, then $\frac{\delta}{2} = 0.005$, i.e., $\delta = 0.01$.

Note that X is a discrete r.v. since it must take values in $\delta \times \mathbb{Z}$. If $\tilde{x}$ is a possible value of $\tilde{X}$, then

$$\mathbb{P}[\tilde{X} = \tilde{x}] = \int_{\tilde{x}-\frac{\delta}{2}}^{\tilde{x}+\frac{\delta}{2}} f(w)dw \doteq f(\tilde{x})\delta$$

This is a crude rectangle-rule approximation to the integral which is valid as long as δ is small enough. This is where we use the assumed continuity of $f(x)$.

Now, $\tilde{X}$ and $\tilde{Y} = \tilde{X}^2$ are both discrete and we can use elementary conditional probability to solve the problem for these r.v.'s. Note that even when X is continuous, $P_{X|Y}(\cdot|y)$ must be a discrete distribution, since all the probability is concentrated on the 2 points $\pm\sqrt{y}$.

Computing the conditional probability mass function (p.m.f.) of $\tilde{X}$ given $\tilde{Y} = y$ using elementary conditional probability, it suffices to compute for one of the possible values:

$$\begin{aligned}\mathbb{P}[\tilde{X} = \sqrt{y}|\tilde{Y} = y] = \frac{\mathbb{P}[\tilde{X} = \sqrt{y}\&\tilde{Y} = y]}{\mathbb{P}[\tilde{Y} = y]} &= \frac{\mathbb{P}[\tilde{X} = \sqrt{y}]}{\mathbb{P}[\tilde{X} = \sqrt{y} \text{ or } \tilde{X} = -\sqrt{y}]} \\ \doteq \frac{f(\sqrt{y})\delta}{(f(\sqrt{y}) + f(-\sqrt{y}))\delta} &= \frac{f(\sqrt{y})}{f(\sqrt{y}) + f(-\sqrt{y})}\end{aligned}$$

Now the "$\doteq$" approximation becomes exact in the limit as $\delta \to 0$. Let's conjecture this is the right answer in general and see if we can prove it works.

In the previous calculation, we implicitly assumed $y > 0$, but the cases $y = 0$ and $y < 0$ are easy to deal with. For convenience, define for $y > 0$,

$$\begin{aligned}p(y) &= \frac{f(\sqrt{y})}{f(\sqrt{y}) + f(-\sqrt{y})} \\ q(y) &= \frac{f(-\sqrt{y})}{f(\sqrt{y}) + f(-\sqrt{y})]}\end{aligned}$$

For general X with $dP_X = fdm$, we conjecture that $P_{X|Y}$ (where $Y = X^2$) is given by

$$P_{X|Y}(\cdot|y) = \begin{cases} p(y)\delta_{\sqrt{y}} + q(y)\delta_{-\sqrt{y}} & \text{if } y > 0 \\ \delta_0 & \text{if } y \leq 0 \end{cases}$$

In the last line, we could have used any probability distribution on the appropriate space ($\mathbb{R}$, in this case), since $\mathbb{P}[Y \leq 0] = 0$.

Let's check that this works. That means, checking the defining properties for $P_{X|Y}$. We want to check that for each y, $P_{X|Y}(\cdot|y)$ is a probability measure on Range(X). This is obvious. We want to check that for each measurable $B \subset Range(X)$, $P[X \in B|Y = y] = P_{X|Y}(B|y)$ for P_Y-almost all y. To check this, we observe it is a function of y and we need to show that for all (measurable.) $A \subset \mathbb{R}$, $E[I_A(Y)I_B(X)] = E[I_A(Y)P_{X|Y}(B|Y)]$. This is just the normal equations for $E[I_B(X)|Y] = P[X \in B|Y]$. Note the the "$P_{X|Y}$" should be interpreted as our proposed version of $P_{X|Y}$. The statement is true for the real $P_{X|Y}$, and we want to verify that our proposed $P_{X|Y}$ satisfies it.

Since $P[Y \leq 0] = 0$,we can replace A with $A \cap (0, \infty)$, i.e., assume $A \subset (0, \infty)$. Now the $E[I_A(Y)I_B(X)]$ is simply $P[Y \in A\&X \in B] = P[X^2 \in A\&X \in B]$. To work with the $E[I_A(Y)P_{X|Y}(B|Y)]$, it is convenient to write

$$P_{X|Y}(B|Y) = p(Y)I_B(\sqrt{Y}) + q(Y)I_B(-\sqrt{Y}) = p(X^2)I_B(|X|) + q(X^2)I_B(-|X|)$$

valid for $Y > 0$. The first equality follows from the general fact that $I_C(z) = \delta_z(C)$.

In order to work with the event $[X^2 \in A]$, define $C_+ = \{x : x > 0, \&x^2 \in A\}$, $C_- = \{x : x < 0, \&x^2 \in A\}$, then $[X^2 \in A] = [X \in C_+] \cup [X \in C_-]$, and the two events are disjoint. Thus, $I_A(Y) = I_A(X^2) = I_{C_+}(X) + I_{C_-}(X)$. Note that $C_- = -C_+$, by which we mean -1 times every element in C_+. Put another way, $I_{C_+}(-x) = I_{C_-}(x)$.

Working on $E[I_A(Y)P_{X|Y}(B|Y)]$,

$$\begin{aligned}E[I_A(Y)P_{X|Y}(B|Y)] &= E[(I_{C_+}(X) + I_{C_-}(X))(p(X^2)I_B(|X|) + q(X^2)I_B(-|X|))] \\ &= E[I_{C_+}(X)p(X^2)I_B(X)] + E[(I_{C_+}(X)q(X^2)I_B(-X)] \\ &\quad + E[I_{C_-}(X)p(X^2)I_B(-X)] + E[I_{C_-}(X)q(X^2)I_B(X)]\end{aligned}$$

In the above, when we remove the absolute value signs around X, we make use of the sign of X implied by the indicator of $C_{\pm}$. For instance, $I_{C_-}(X)I_B(|X|) = I_{C_-}(X)I_B(-X)$ since $X \in C_- \implies X < 0 \implies |X| = -X$.

We can hopefully work with $E[I_{C_+}(X)p(X^2)I_B(X)]+E[(I_{C_+}(X)q(X^2)I_B(-X|]+E[I_{C_-}(X)p(X^2)I_B(-X)]+E[I_{C_-}(X)q(X^2)I_B(X)]$ to get $E[I_A(Y)I_B(X)]$. We tried various ways to combine them before finding the right approach. We will combine the first and fourth terms, but first note

$$\begin{aligned} I_{C_+}(x)p(x^2) + I_{C_-}(x)q(x^2) &= I_{C_+}(x)\frac{f(x)}{f(x)+f(-x)} + I_{C_-}(x)q(x^2)\frac{f(x)}{f(x)+f(-x)} \\ &= I_A(x^2)\frac{f(x)}{f(x)+f(-x)} \end{aligned}$$

The first equation follows since since the numerator of $I_{C_-}(x)q(x^2)$ is $I_{C_-}(x)f(-\sqrt{x^2}) = I_{C_-}(x)f(x)$ because $I_{C_-(x)} = 1 \implies x < 0 \implies x = -\sqrt{x^2}$. Similarly, to combine $E[(I_{C_+}(X)q(X^2)I_B(-X|] + E[I_{C_-}(X)p(X^2)I_B(-X)]$ we will use $I_{C_+}(x)q(x^2) + I_{C_-}(x)p(x^2) = I_A(x^2)\dfrac{f(-x)}{f(x)+f(-x)}$

Now we compute the expectations using the distribution of X (with its Lebesgue density)

$$\begin{aligned} E[I_{C_+}(X)p(X^2)I_B(X)] + E[I_{C_-}(X)q(X^2)I_B(X)] &= \int I_A(x^2)I_B(x)\frac{f(x)}{f(x)+f(-x)}f(x)dm(x) \\ E[I_{C_+}(X)q(X^2)I_B(X)] + E[I_{C_-}(X)p(X^2)I_B(X)] &= \int I_A(x^2)I_B(-x)\frac{f(-x)}{f(x)+f(-x)}f(x)dm(x) \\ = \int I_A(z^2)I_B(z)\frac{f(z)}{f(-z)+f(z)}f(-z)dm(z) \end{aligned}$$

The last step follows by changing variables $-x \to z$. Recall that $dm(x)$ is basically the dx you are used to.

Now we combine the results from this last computation. After substituting "x" for "z" as a dummy variable of integration, the sum of the two previous results is

$$\begin{aligned} \int I_A(x^2)I_B(x)\frac{f(x)}{f(x)+f(-x)}[f(x)+f(-x)]dm(x) &= \int I_A(x^2)I_B(x)f(x)dm(x) \\ &= E[I_A(X^2)I_B(X)] \end{aligned}$$

Hence, we have verified that our proposed form for $P_{X|Y}(\cdot|y)$ is a version of the correct answer. Note that we can change it on a set of y values having P_Y measure zero, and the answer would still be correct.

Remark. Suppose (instead of squared error of loss) we want a prediction loss where $L(y.\hat{y})$,

$$\begin{aligned} E[L(Y,h(X))] &= E[E[L(Y|h(X)|X)]] \qquad \text{(by total expectation)} \\ &= \int \left[\int L(y,h(x))dP_{Y|X}(y|x)\right] dP_X(x) \qquad (dP_{Y|X}(y|x) = d\mu(y)) \end{aligned}$$

If $h^*(x) = \underset{a}{\operatorname{argmin}} \int L(y,a)dP_{Y|X}(y|x)$, with a exists and unique, we can update the "data" x here and derive the argmin and this minimize $E[L(Y,h(X))] \geq E[L(Y,h^*(X))]$ over $h(X)$

Finally, we have already termed our conditional expectation as "**normal equations**" so many times. Why? Because we want to find $h(X)$ to minimize $E[(Y-h(X))^2]$, we can set the L_2 norm of probability measure $L_2(\mathbb{P}) = \{W : E[W^2] < \infty\}$, and this is a linear space. By defining an inner product on L_2 with $\langle U, V\rangle = E[UV]$, and the properties

1. $\langle U, V\rangle = \langle V, U\rangle$
2. $\langle a_1 U_1 + a_2 U_2, V\rangle = a_1\langle U_1, V\rangle + a_2\langle U_2, V\rangle$
3. $\langle U, U\rangle \geq 0$, and it equals to 0 iff $U = 0$ a.s.

is complete. We suppose $M = \{h(X) : h : Range(X) \to \mathbb{R} \& E[h(X)^2] < \infty\}$, this is a closed linear subspace. In L_2, $||W|| = \sqrt{\langle W, W\rangle}$, MSPE$(Y, h(X)) = ||Y - h(X)||^2$, $Y - h^*(X)$ is normal to linear subspace M i.e. $\forall g(X) \in M$, $\langle Y - h(X), g(X)\rangle = 0$, and $h^*(X) = E[Y|X]$ i.e. $E[(Y - h^*(X))g(X)] = 0, \forall g(X) \in M$

References

[1] D.D. Cox , “Mathematical Statistics for Data Scientist.” Chapter 1.

[2] J. Shao , “Mathematical Statistics.” Chapter 1.

[3] P. Billingsley , “Probability and Measure.” Section 2, 15, 16, 18, 32-34.

[4] R.B. Ash , “Real Analysis and Probability.” Chapter 1, 2.2

[5] K.L. Chung , “A Course in Probability Theory.” Chapter 1, 2, 9.1