

Lecture 2: Basics of Measure Theory 1

Lecturer: Yiqiao Zhong

Last update: 15:05:19, 2024-09-11

Disclaimer: materials in this pdf are used for STAT 709 “Mathematical Statistics” at the University of Wisconsin–Madison, but they are not guaranteed to be accurate. Consult textbooks whenever possible.

Do we need measure theory for research in applied statistics or machine learning? The answer is perhaps no for some of you, but nevertheless it is a good idea to know why classical stuff can go wrong and what measure theory can offer.

We will discuss the motivation for measure theory and the relationship between measure theory and statistics.

1 Riemann Integration Recap

Let f be a non-negative real-valued function on the interval $[a, b]$. Riemann integral gives the area of the region of the plane under the graph of the function f and above the interval $[a, b]$.

Definition 1. (Riemann Integral) Given any partition $[a, b] = \bigcup_{i=1}^n [x_i, x_{i+1}]$, let $\Delta = \max_{0 \leq i \leq n} |x_{i+1} - x_i|$, define $I(\Delta) = \sum_{i=0}^n f(t_i)(x_{i+1} - x_i)$ with $t_i \in [x_i, x_{i+1}]$ for each i . If $\lim_{\Delta \rightarrow 0} \sum_{i=1}^n (\sup_{t_i \in [x_i, x_{i+1}]} f(t_i) - \inf_{t_i \in [x_i, x_{i+1}]} f(t_i))(x_{i+1} - x_i) = 0$, define Riemann Integral as follows:

$$I := \lim_{\Delta \rightarrow 0} I(\Delta).$$

Riemann integrals are intuitive but are technically cumbersome when we want to exchange limits. We will show some examples to illustrate the inconvenience of Riemann integrals.

Example 1. Let $\mathbb{Q}$ be the set of rational numbers in $[0, 1]$, and let $A_1, A_2, \dots$ be the sets such that $A_1 \subsetneq A_2 \subsetneq \dots \subsetneq \mathbb{Q}$, $|A_n| = n$ and $\bigcup_{n=1}^{\infty} A_n = \mathbb{Q}$. Define $f_n(x)$ as follows:

$$f_n(x) = \begin{cases} 1, & x \in A_n \\ 0, & \text{otherwise.} \end{cases}$$

Then $\lim_{n \rightarrow \infty} f_n(x) = 1_{\mathbb{Q}}$ is not Riemann integrable; however $\int f_n(x) dx = 0$ for all n . That is to say, $\lim_{n \rightarrow \infty} \int f_n(x) dx$ is well defined, but $\int \lim_{n \rightarrow \infty} f_n(x) dx$ is not well defined.

Example 2. Let $f_n(x) = \begin{cases} 1, & x \in [n, n+1] \\ 0, & \text{otherwise} \end{cases}$, then we may notice the following:

$$\lim_{n \rightarrow \infty} \int_{-\infty}^{\infty} f_n(x) dx = 1 \neq 0 = \int_{-\infty}^{\infty} \lim_{n \rightarrow \infty} f_n(x) dx.$$

The limit and the integration are not exchangeable in these two examples. While the first example seems to be caused by annoying technical details, the second example is quite different. Under the definition of Riemann integrals, there is no obvious indication whether we should or should not expect exchanging limit with integration to be valid—sometimes we have to require cumbersome conditions. How do we find a different way of defining integrals so that exchanging limits is easier?

The high-level idea is to **broaden the space of Riemann integrable functions** by explicitly taking into account our desired goal (exchanging limits, or more simply dealing with summing countably many terms). To that end, we will work with a collection of slightly more irregular sets rather than the interval $[x_i, x_{i+1}]$ so that taking limits is easier. But not too irregular to a point where we lose useful properties (we still love intervals!). We will discuss measure and abstract integration in Section 3 and 4.

2 Motivation for Measure Theory in statistics

Example 3. (*Submanifold support*) Consider a random variable $U \sim \text{Unif}([0, 2\pi])$, let $\begin{cases} X = \cos U \\ Y = \sin U \end{cases}$, where $(X, Y) \in \mathbb{R}^2$, but there is no pmf for (X, Y) . Further, there does not exist a continuous function¹ $f_{X,Y}$ such that (*exercise*):

$$\int_{a_1}^{a_2} \int_{b_1}^{b_2} f_{X,Y}(x, y) \, dx dy = \mathbb{P}(X \in [a_1, a_2], Y \in [b_1, b_2]).$$

Example 4. (*Infinite product space*) Suppose random variables $X_1, X_2, \dots$ are i.i.d. from $\text{Bern}(\frac{1}{2})$. In probability theory and statistics, asymptotics is very important, where we are interested in the average behavior when sample size is large. One example is the following.

$$\mathbb{P} \left(\liminf_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n X_i > \frac{1}{3} \right).$$

The challenge is that how to define the probability if there are infinite random variables.

Example 5. (*Random measure*) Consider a sequence of matrices $(X_n)_{n=1}^\infty$ ², where $[X_n]_{ij} \stackrel{i.i.d}{\sim} N(0, 1)$ and $[X_n]_{ii} \stackrel{i.i.d}{\sim} N(0, 2)$. The empirical spectral distribution (ESD) of $Y_n = \frac{1}{\sqrt{n}} X_n$ is

$$\mu_{Y_n} := \frac{1}{n} \sum_{i=1}^n \delta_{\lambda_i(Y_n)},$$

where λ_i 's are the eigenvalues and δ_a denotes the point mass (see Section 3). For finite n , the above quantity is essentially equivalent to a histogram. What happens if n is very large? In a certain limit, μ_{Y_n} converges to a deterministic measure μ_{SC} where μ_{SC} has density

$$\frac{d\nu}{dx} = \frac{1}{2\pi} \sqrt{4 - x^2} \mathbf{1}_{\{|x| \leq 2\}}.$$

The above formula gives the density of the limiting “histrogram” and is known as Wigner’s semicircle law. The challenge is to define clearly what convergence means in this context—convergence of a more abstract mathematical object than a simple random variable.

¹We can even strengthen this to Borel measurable function in $\mathbb{R}^2$.

²Note: X_n is called *Gaussian Orthogonal Ensemble*.

3 Measures

Given a sample space Ω , $\mathcal{F}$ is a collection of subsets of Ω .

Definition 2. (σ -algebra) $\mathcal{F}$ is called a σ -algebra if and only if it has the following properties.

- (i) The empty set $\emptyset \in \mathcal{F}$.
- (ii) If $A \in \mathcal{F}$, then the complement $A^c \in \mathcal{F}$.
- (iii) If $A_i \in \mathcal{F}$, $i = 1, 2, \dots$, then their union $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$.

A useful observation: (1) countably intersection is implied by (ii) and (iii), (2) if $\mathcal{F}$ is a σ -algebra and $(a, b_k] \in \mathcal{F}$ where b_k is a nondecreasing sequence converging to b , then $(a, \lim_k b_k) \in \mathcal{F}$.

Definition 3. (Measurable space) A pair $(\Omega, \mathcal{F})$ consisting of a set Ω and a σ -algebra $\mathcal{F}$ of subsets of Ω is called a measurable space. The elements of $\mathcal{F}$ are called measurable sets in measure theory or events in probability and statistics.

In general, the smallest σ -algebra containing $\mathcal{C}$, a collection of subsets of Ω , is denoted by $\sigma(\mathcal{C})$ and is called the σ -algebra generated by $\mathcal{C}$. In the d -dimensional Euclidean space $\mathbb{R}^d$, let $\mathcal{C}$ be the collection of all open sets in $\mathbb{R}^d$. Then $\mathcal{B} = \sigma(\mathcal{C})$ is called the Borel σ -algebra.

Remark 1. Borel sets are the de facto sets in the Euclidean space we will work with³. It is no less and no more: all open sets, close sets (by complement), countable unions/intersections of Borel sets are still Borel; also, Borel sets do not include every subset of $\mathbb{R}^d$ (if we had defined that way, we wouldn't be conveniently defining measures).⁴

Remark 2. It is easy to show that sets like $(a, b]$, $[a, b]$ are Borel sets. In $\mathbb{R}^d$, $(a_1, b_1] \times (a_d, b_d]$ is a Borel set. In fact, sets of this form generate Borel sets [1, Exercise 1.1.5] ([exercise](#)).

The length of an interval $(a, b]$ is $b - a$, and the volume of a hypercube $(a_1, b_1] \times (a_d, b_d]$ is $\prod_{k \leq d} (b_k - a_k)$. We introduce measures to generalize this notion to abstract measurable spaces.

Definition 4. (Measure) Let $(\Omega, \mathcal{F})$ be a measurable space. A set function ν defined on $\mathcal{F}$ is called measure if and only if it has the following properties.

- (i) $0 \leq \nu(A) \leq \infty$ for any $A \in \mathcal{F}$.
- (ii) $\nu(\emptyset) = 0$
- (iii) If $A_i \in \mathcal{F}$, $i = 1, 2, \dots$, and A_i 's are disjoint, i.e., $A_i \cap A_j = \emptyset$ for any $i \neq j$, then

$$\nu\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \nu(A_i).$$

The triple $(\Omega, \mathcal{F}, \nu)$ is called a measure space. If $\nu(\Omega) = 1$, then ν is called a probability measure and we usually denote it by P instead of ν . In which case $(\Omega, \mathcal{F}, P)$ is called a probability space.

Some examples. The following examples provide three important measures in probability and statistics.

Example 6. (Point mass) Let Ω be a sample space, $\mathcal{F}$ the collection of all subsets. Given $x \in \Omega$, define point mass at x by

$$\delta_x(A) = \begin{cases} 1, & x \in A \\ 0, & \text{otherwise} \end{cases} \quad \forall A \in \mathcal{F}.$$

³Does any (applied) mathematical area ever use sets other than Borel sets?? I am not aware of one.

⁴It is known that the cardinality of the Borel sets is equal to that of the reals $\mathbb{R}$, which is smaller than the power set of the reals $\mathbb{R}$; see Section 5.

Example 7. (Counting measure) Let Ω be a sample space, $\mathcal{F}$ the collection of all subsets, and $\nu(A)$ the number of elements in $A \in \mathcal{F}$ ($\nu(A) = \infty$ if A contains infinitely many elements). Then ν is a measure on $\mathcal{F}$ and is called the counting measure.

Example 8. (Lebesgue measure) There is a unique measure m on $(\mathbb{R}, \mathcal{B})$ that satisfies

$$m([a, b]) = b - a$$

for every finite interval $[a, b]$, $-\infty < a \leq b < \infty$. This is called the Lebesgue measure. If we restrict m to the measurable space⁵ $([0, 1], \mathcal{B}_{[0,1]})$, then m is a probability measure.

Product measure space. Let us assume a very mild assumption on the measure space (in fact we assume it throughout this course). A measure ν on $(\Omega, \mathcal{F})$ is σ -finite if there exists a sequence $A_1, A_2, \dots$ such that $\cup A_i = \Omega$ and $\nu(A_i) < \infty$ for all i .

Theorem 1. (Product space) Let $(\Omega_i, \mathcal{F}_i, \nu_i)$, $i \in \{1, 2, \dots, K\} = \mathcal{I}$, be measure spaces with σ -finite measures, where $K \geq 2$ is an integer. Then there exists a unique σ -finite measure on the product σ -algebra $\sigma(\prod_{i \in \mathcal{I}} \mathcal{F}_i)$, called the product measure and denoted by $\prod_{i \in \mathcal{I}} \nu_i$, such that

$$\prod_{i \in \mathcal{I}} \nu_i \left(\prod_{i \in \mathcal{I}} A_i \right) = \prod_{i \in \mathcal{I}} \nu_i(A_i)$$

for all $A_i \in \mathcal{F}_i$, $i \in \mathcal{I}$.

It is easy to extend the product of measures to sets like $\prod_i A_i$ uniquely, but there are many other sets in $\sigma(\prod_{i \in \mathcal{I}} \mathcal{F}_i)$. The above theorem states that extension is possible and unique.

CDF. Given a probability P on space $(\mathbb{R}^d, \mathcal{B}^d)$, the cumulative density function (cdf) is

$$F(x_1, \dots, x_d) = P((-\infty, x_1] \times \dots \times (-\infty, x_d]).$$

The basic properties of CDF are stated in [2] ([exercise](#)).

4 Measurable functions

The next step is to broaden the space of (piecewise) continuous functions to a larger space. The idea is to reduce the definition to set theory. An important device is the inverse image. For any function or mapping $f : \Omega \rightarrow \Gamma$, the inverse image (or pre-image) of a subset $B \subset \Gamma$ under f is

$$f^{-1}(B) = \{\omega \in \Omega : f(\omega) \in B\}.$$

We give the basic properties of the inverse image.

- (i) $f^{-1}(B^c) = (f^{-1}(B))^c$ for any $B \in \mathcal{G}$,
- (ii) $f^{-1}(\cup_{i \in \mathcal{I}} B_i) = \cup_{i \in \mathcal{I}} f^{-1}(B_i)$ for any $B_i \in \mathcal{G}$.

Definition 5. (Measurable function) Let $(\Omega, \mathcal{F})$ and $(\Lambda, \mathcal{G})$ be measurable spaces and f a function from Ω to Λ . The function f is called a measurable function from $(\Omega, \mathcal{F})$ to $(\Lambda, \mathcal{G})$ if and only if $f^{-1}(\mathcal{G}) \subset \mathcal{F}$.

If $f : \mathbb{R} \rightarrow \mathbb{R}$ is continuous, then an equivalent definition (from the topological perspective) states that $f^{-1}(O)$ is an open set for any open set $O \subset \mathbb{R}$ ([exercise](#)). It then follows from the properties of inverse image that all continuous functions are measurable functions from $(\mathbb{R}, \mathcal{B})$ to $(\mathbb{R}, \mathcal{B})$ ([exercise](#)).

⁵ $\mathcal{B}_{[0,1]}$ means any of the set of the form $[0, 1] \cap B$ with $B \in \mathcal{B}$.

Random variables. We call $X : (\Omega, \mathcal{F}) \rightarrow (\mathbb{R}, \mathcal{B})$ a random variable and $X : (\Omega, \mathcal{F}) \rightarrow (\mathbb{R}^d, \mathcal{B}^d)$ a random vector⁶. It maps a potentially large and complicated sample space Ω to real-valued observations.

Distribution. A random variable X induces a measure defined by

$$P \circ X^{-1}(B) = P(X^{-1}(B)), \quad \text{for any } B \in \mathcal{B}.$$

Usually, we denote this measure by μ_X , and it is called the distribution or law of X . Basically, it measures the probability of the event that the observed value of X falls in the set B .

Borel functions. How to define functions of random variables? When both $\mathcal{F}$ and $\mathcal{G}$ are Borel σ -algebra such as $\mathcal{B}^d$, we call the measurable function f as *Borel function*⁷. Practically we only care about functions that are Borel, so we assume functions are Borel unless specified otherwise.

Here are some important definitions and discussions.

1. X is random variable iff $X^{-1}((-\infty, a]) \in \mathcal{F}$ for all $a \in \mathbb{R}$.
2. Composing two measurable functions $f : (\Omega, \mathcal{F}) \rightarrow (\Lambda, \mathcal{G})$ and $g : (\Lambda, \mathcal{G}) \rightarrow (\Xi, \mathcal{H})$ results in a measurable function $g \circ f : (\Omega, \mathcal{F}) \rightarrow (\Xi, \mathcal{H})$.
3. If $X_1, \dots, X_n$ are random variables, and $f : (\mathbb{R}^d, \mathcal{B}^d) \rightarrow (\mathbb{R}, \mathcal{B})$ is Borel, then $f(X_1, \dots, X_n)$ is a random variable.
4. If $X_1, X_2, \dots$ are random variable, then

$$\inf_n X_n, \quad \sup_n X_n, \quad \limsup_n X_n, \quad \liminf_n X_n$$

are all random variables.

Some properties such as 2 inherit the “nicety” of continuous functions, but other properties such as 3 already show the advantage of measurable functions ([exercise](#)).

Remark 3. *To conclude this lecture, the reason for mathematical abstraction in this section is to find a nice space so that mathematical operations are rigorous. The measurable space should be sufficiently rich so that basic sets such as intervals are included, but not too complicated (e.g., $\Omega = \mathbb{R}, \mathcal{F} = 2^\Omega$). Borel sets and Borel measurable functions are standard notions to work with.*

5 Further readings

To familiarize basic measure theory, it is highly recommended to read and check the properties of measures, e.g., Pages 4 and 5 of [2] or Chap 1.1–1.2 of [1] and run through some exercises. This lecture note is largely based on these two references.

Completion of measure. Lebesgue measurable sets are a bigger space that includes Borel measurable sets. It essentially adds into the space all modifications of Borel sets that have made no difference under the Lebesgue measure. In particular, all subsets of zero-measure sets are Lebesgue measurable, but not necessarily Borel measurable.

There are many more Lebesgue measurable sets compared with Borel measurable sets. See this link and this link for details.

⁶Sometimes people are lazy, calling random vector as random variable as well, or simply write r.v.

⁷More general definition of Borel functions requires Ω and Λ to have topological structure (meaning that we can talk about open sets), but we only need $\mathcal{B}^d$ for this course.

References

- [1] Rick Durrett. *Probability: theory and examples*, volume 49. Cambridge university press, 5th edition, 2019.
- [2] Jun Shao. *Mathematical Statistics*. Springer Science & Business Media, 2nd edition, 2013.