Notes 1 : Measure-theoretic foundations I

Math 733 - Fall 2013 *Lecturer: Sebastien Roch*

References: [Wil91, Section 1.0-1.8, 2.1-2.3, 3.1-3.11], [Fel68, Sections 7.2, 8.1, 9.6], [Dur10, Section 1.1-1.3].

1 Overview of the semester

1.1 Foundations

Recall the basic probabilistic framework in the discrete case. Say we are tossing two unbiased coins:

- Sample space (finite space): $\Omega = \{\mathtt{HH}, \mathtt{HT}, \mathtt{TH}, \mathtt{TT}\}$.
- Outcome (point): $\omega \in \Omega$.
- Probability measure (normalized measure): $\mathbb{P}[\omega] = 1/4$, $\forall \omega \in \Omega$.
- Random variable (function on Ω): $X(\omega) =$ # of heads. E.g. $X(\mathtt{HH}) = 2$.
- Event (subset of Ω): $A = \{\omega \in \Omega \,:\, X(\omega) = 1\} = \{\mathtt{HT}, \mathtt{TH}\}$; $\mathbb{P}[A] = \mathbb{P}[\mathtt{HT}] + \mathbb{P}[\mathtt{TH}] = 1/2$.
- Expectation ($\mathbb{P}$-weighted sum): $\mathbb{E}[X] = \sum_{\omega \in \Omega} \mathbb{P}[\omega] X(\omega) = 1$.
- Distribution: X is distributed according to a binomial distribution $B(2, 1/2)$.

The previous example generalizes to countable sample spaces. But problems quickly arise when uncountably many outcomes are considered. Two examples:

EX 1.1 (Infinite sequence of Bernoulli trials) *Consider an infinite sequence of unbiased coin flips. The state space is*

$$\Omega = \{\omega = (\omega_n)_n \,:\, \omega_n \in \{\mathtt{H}, \mathtt{T}\}\}.$$

It is natural to think of this example as a sequence of finite probability spaces $(\Omega^{(n)}, \mathbb{P}^{(n)})$, $n \geq 1$, *over the first* n *trials. But the full infinite space is necessary to consider many interesting events, e.g.,*

$$B = \{\alpha \text{ consecutive } \mathtt{H}\text{'s occur before } \beta \text{ consecutive } \mathtt{T}\text{'s}\}.$$

How to define the probability of B in that case? (Note that B is uncountable.)

The problem is that we cannot assign a probability to all events (i.e., subsets) in a consistent way. More precisely, note that probabilities have the same properties as (normalized) volumes. In particular the probability of a disjoint union of subsets is the sum of their probabilities:

- $0 \leq \mathbb{P}[A] \leq 1$ for all $A \subseteq \Omega$.
- $\mathbb{P}[A \cup B] = \mathbb{P}[A] + \mathbb{P}[B]$ if $A \cap B = \emptyset$.
- $\mathbb{P}[\Omega] = 1$.

But Banach and Tarski have shown that there is a subset F of the sphere S^2 such that, for $3 \leq k \leq \infty$, S^2 is the disjoint union of k rotations of F

$$S^2 = \bigcup_{i=1}^{k} \tau_i^{(k)} F,$$

so the "volume" of F must be $4\pi/3, 4\pi/4, \ldots, 0$ simultaneously. So if we were picking a point uniformly at random from S^2, the event F could not be assigned a "meaningful" probability.

Measure theory provides a solution. The idea is to restrict the set of events $\mathcal{F} \subset 2^\Omega$ to something more manageable—but big enough to contain all sets that arise "naturally" from simpler sets. E.g., in the example above,

EX 1.2 (Infinite Bernoulli trials) *In the infinite Bernoulli trials example above, let p_n be the probability that B occurs by time n. Then clearly $p_n \uparrow$ by monotonicity and we can define the probability of B in the full space as the limit p_∞.*

In this course, we will start by developing such a rigorous measure-theoretic framework for probability.

1.2 Limit laws

Whereas a major focus of undergraduate probability is on computing various probabilities and expectations, the emphasis of this course is on deriving general laws such as the law of large numbers (LLN).

Our intuitive understanding of probability is closely related to empirical frequencies. If one flips a coin sufficiently many times, heads is expected come up roughly half the time. The LLN gives a mathematical justification of this fact.

THM 1.3 (Weak Law of Large Numbers (Bernoulli Trials)) *Let $(\Omega^{(n)}, \mathbb{P}^{(n)})$, $n \geq 1$, be the sequence of probability spaces corresponding to finite sequences of Bernoulli trials. If X_n is the number of heads by trial n, then*

$$\mathbb{P}^{(n)}\left[\left|\frac{X_n}{n} - \frac{1}{2}\right| \geq \varepsilon\right] \to 0, \tag{1}$$

for all $\varepsilon > 0$.

Proof: Although we could give a direct proof for the Bernoulli case, we give a more general argument.

LEM 1.4 (Chebyshev's Inequality (Discrete Case)) *Let Z be a discrete random variable on a probability space $(\Omega, \mathbb{P})$. Then for all $\varepsilon > 0$*

$$\mathbb{P}[|Z| \geq \varepsilon] \leq \frac{\mathbb{E}[Z^2]}{\varepsilon^2}. \tag{2}$$

Proof: Let $\mathcal{A}$ be the range of Z. Then

$$\mathbb{P}[|Z| \geq \varepsilon] = \sum_{z \in \mathcal{A}: |z| \geq \varepsilon} \mathbb{P}[Z = z] \leq \sum_{z \in \mathcal{A}: |z| \geq \varepsilon} \frac{z^2}{\varepsilon^2}\mathbb{P}[Z = z] \leq \frac{1}{\varepsilon^2}\mathbb{E}[Z^2].$$

■

We return to the proof of (1). Let

$$Z_n = \frac{X_n}{n} - \frac{1}{2}\left(= \frac{X_n - \frac{1}{2}n}{n} = \frac{X_n - \mathbb{E}^{(n)}[X_n]}{n}\right).$$

Then

$$\mathbb{P}^{(n)}[|Z_n| \geq \varepsilon] \leq \frac{\mathbb{E}^{(n)}[Z_n^2]}{\varepsilon^2} = \frac{1}{\varepsilon^2}\frac{n(\frac{1}{2})(\frac{1}{2})}{n^2} = \frac{1}{4\varepsilon^2 n} \to 0,$$

using the variance of the binomial. ■

Among other things, we will prove a much stronger result on the infinite space where the probability and limit are interchanged.

In fact a finer result can be derived.

THM 1.5 (DeMoivre-Laplace Theorem (Symmetric Case)) *Let X_n be a binomial $B(n, 1/2)$ on the space $(\Omega^{(n)}, \mathbb{P}^{(n)})$ and assume $n = 2m$. Then for all $z_1 < z_2$*

$$\mathbb{P}^{(n)}\left[z_1 \leq \frac{X_n - n/2}{\sqrt{n}/2} \leq z_2\right] \to \frac{1}{\sqrt{2\pi}}\int_{z_1}^{z_2} e^{-x^2/2}\mathrm{d}x.$$

Proof: Let

$$a_k = \mathbb{P}^{(n)}[X_n = m+k] = \frac{n!}{(m-k)!(m+k)!}2^{-n} = a_0 \frac{(m-k+1)\cdots m}{(m+1)\cdots(m+k)},$$

where recall that $m = n/2$. Then

$$\mathbb{P}^{(n)}\left[z_1 \leq \frac{X_n - n/2}{\sqrt{n}/2} \leq z_2\right] = \sum_{z_1\sqrt{m/2} \leq k \leq z_2\sqrt{m/2}} a_k.$$

By Stirling's formula:

$$n! \sim n^n e^{-n}\sqrt{2\pi n},$$

we have

$$a_0 = \frac{n!}{m!m!}2^{-n} \sim \frac{1}{\sqrt{\pi m}}.$$

Divide the numerator and denominator by m^k and use that, for j/m small,

$$1 + \frac{j}{m} = e^{\frac{j}{m} + O(\frac{j^2}{m^2})}.$$

Noticing that

$$0 - \frac{2}{m}[1 + \cdots + k - 1] - \frac{k}{m} = -\frac{2}{m}\frac{k(k-1)}{2} - \frac{k}{m} = -\frac{k^2}{m},$$

and bounding the error term in the exponent by $O(k^3/m^2)$, we get

$$a_k \sim \frac{1}{\sqrt{\pi m}}e^{-k^2/m} = \frac{1}{\sqrt{2\pi}}\sqrt{\frac{2}{m}}e^{-(k\sqrt{\frac{2}{m}})^2/2},$$

as long as $k^3/m^2 \to 0$ or $k = o(m^{2/3})$.

Then, letting $\Delta = \sqrt{\frac{2}{m}}$,

$$\begin{aligned}\mathbb{P}^{(n)}\left[z_1 \leq \frac{X_n - n/2}{\sqrt{n}/2} \leq z_2\right] \quad &\sim \quad \sum_{z_1\Delta^{-1} \leq k \leq z_2\Delta^{-1}} \frac{1}{\sqrt{2\pi}}\Delta e^{-(k\Delta)^2/2} \\ &\to \quad \frac{1}{\sqrt{2\pi}}\int_{z_1}^{z_2} e^{-x^2/2}\mathrm{d}x.\end{aligned}$$

■

The previous proof was tailored for the binomial. We will prove this theorem in much greater generality using more powerful tools.

1.3 Martingales

Finally, time permitting, we will give an introduction to martingales, which play an important role in the theory of stochastic processes, developed further in the Spring semester.

2 Measure spaces

2.1 Basic definitions

Let S be a set. We discussed last time an example showing that we cannot in general assign a probability to every subset of S. Here we discuss "well-behaved" collections of subsets. First, an algebra on S is a collection of subsets stable under finitely many set operations.

DEF 1.6 (Algebra on S) *A collection Σ_0 of subsets of S is an* algebra on S *if*

1. $S \in \Sigma_0$*;*
2. $F \in \Sigma_0$ *implies* $F^c \in \Sigma_0$*;*
3. $F, G \in \Sigma_0$ *implies* $F \cup G \in \Sigma_0$.

That, of course, implies that the empty set and intersections are in Σ_0. The collection Σ_0 is an actual algebra (i.e., a vector space with a bilinear product) with symmetric difference as the sum, intersection as product and the underlying field being the field with two elements.

EX 1.7 *On $\mathbb{R}$, sets of the form*

$$\bigcup_{i=1}^{k} (a_i, b_i]$$

where the union is disjoint with $k < +\infty$ and $-\infty \leq a_i \leq b_i \leq +\infty$ form an algebra.

Finite set operations are not enough for our purposes. For instance, we want to be able to take limits. A σ-algebra is stable under countably many set operations.

DEF 1.8 (σ-Algebra on S) *A collection Σ of subsets of S is a* σ-algebra on S *if*

1. $S \in \Sigma$*;*
2. $F \in \Sigma$ *implies* $F^c \in \Sigma_0$*;*

3. *$F_n \in \Sigma, \forall n$ implies $\cup_n F_n \in \Sigma$.*

EX 1.9 *2^S is a trivial example.*

To give a nontrivial example, we need the following definition. We begin with a lemma.

LEM 1.10 (Intersection of σ-algebras) *Let $\mathcal{F}_i$, $i \in I$, be σ-algebras on S where I is arbitrary. Then $\cap_i \mathcal{F}_i$ is a σ-algebra.*

Proof: We prove only one of the conditions. The other ones are similar. Suppose $A \in \mathcal{F}_i$ for all i. Then A^c is in $\mathcal{F}_i$ for all i. ■

DEF 1.11 (σ-Algebra generated by $\mathcal{C}$) *Let $\mathcal{C}$ be a collection of subsets of S. Then we let $\sigma(\mathcal{C})$ be the smallest σ-algebra containing $\mathcal{C}$, defined as the intersection of all such σ-algebras (including in particular 2^S).*

EX 1.12 *The smallest σ-algebra containing all open sets in $\mathbb{R}$, denoted $\mathcal{B}(\mathbb{R}) = \sigma(\text{Open Sets})$, is called the* Borel σ-algebra*. This is a non-trivial σ-algebra in the sense that it can be proved that there are subsets of $\mathbb{R}$ that are not in $\mathcal{B}$, but any "reasonable" set is in $\mathcal{B}$. In particular, it contains the algebra in EX 1.7.*

EX 1.13 *The σ-algebra generated by the algebra in EX 1.7 is $\mathcal{B}(\mathbb{R})$. This follows from the fact that all open sets of $\mathbb{R}$ can be written as a countable union of open intervals. (Indeed, for $x \in O$ an open set, let I_x be the largest open interval contained in O and containing x. If $I_x \cap I_y \neq \emptyset$ then $I_x = I_y$ by maximality (take the union). Then $O = \cup_x I_x$ and there are only countably many disjoint ones because each one contains a rational.)*

We now define measures.

DEF 1.14 (Additivity) *A non-negative set function on an algebra Σ_0*

$$\mu_0 : \Sigma_0 \to [0, +\infty],$$

is additive *if*

1. *$\mu_0(\emptyset) = 0$;*
2. *$F, G \in \Sigma_0$ with $F \cap G = \emptyset$ implies $\mu_0(F \cup G) = \mu_0(F) + \mu_0(G)$.*

Moreover, μ_0 is σ-additive if the latter is true for $F_n \in \Sigma_0$, $n \geq 0$, disjoint with $\cup_n F_n \in \Sigma_0$.

EX 1.15 *For the algebra in the EX 1.7, the set function*

$$\lambda_0\left(\bigcup_{i=1}^{k}(a_i,b_i]\right)=\sum_{i=1}^{k}(b_i-a_i)$$

is additive. (In fact, it is also σ-additive. We will show this later.)

DEF 1.16 (Measure space) *Let Σ be a σ-algebra on S. A σ-additive function μ on Σ is called a* measure. *Then (S,Σ,μ) is called a* measure space.

DEF 1.17 (Probability space) *If $(\Omega,\mathcal{F},\mathbb{P})$ is a measure space with $\mathbb{P}(\Omega)=1$ then $\mathbb{P}$ is called a* probability measure *and $(\Omega,\mathcal{F},\mathbb{P})$ is called a* probability triple.

2.2 Extension theorem

To define a measure on $\mathcal{B}(\mathbb{R})$ we need the following tools from abstract measure theory.

THM 1.18 (Caratheodory's extension theorem) *Let Σ_0 be an algebra on S and let $\Sigma=\sigma(\Sigma_0)$. If μ_0 is σ-additive on Σ_0 then there exists a measure μ on Σ that agrees with μ_0 on Σ_0. (If in addition μ_0 is finite, the next lemma implies that the extension is unique.)*

LEM 1.19 (Uniqueness of extensions) *Let $\mathcal{I}$ be a π-*system *on S, i.e., a family of subsets stable under finite intersections, and let $\Sigma=\sigma(\mathcal{I})$. If μ_1,μ_2 are finite measures on (S,Σ) that agree on $\mathcal{I}$, then they agree on Σ.*

EX 1.20 *The sets $(-\infty,x]$ for $x\in\mathbb{R}$ are a π-system generating $\mathcal{B}(\mathbb{R})$.*

2.3 Lebesgue measure

Finally we can define Lebesgue measure. We start with $(0,1]$ and extend to $\mathbb{R}$ in the obvious way. We need the following lemma.

LEM 1.21 (σ-Additivity of λ_0) *Let λ_0 be the set function defined above, restricted to $(0,1]$. Then λ_0 is σ-additive.*

DEF 1.22 (Lebesgue measure on unit interval) *The unique extension of λ_0 to $(0,1]$ is denoted λ and is called* Lebesgue measure.

3 Measurable functions

3.1 Basic definitions

Let (S, Σ, μ) be a measure space and let $\mathcal{B} = \mathcal{B}(\mathbb{R})$.

DEF 1.23 (Measurable function) *Suppose $h : S \to \mathbb{R}$ and define*

$$h^{-1}(A) = \{s \in S : h(s) \in A\}.$$

The function h is Σ-measurable *if $h^{-1}(B) \in \Sigma$ for all $B \in \mathcal{B}$. We denote by $m\Sigma$ (resp. $(m\Sigma)^+$, $b\Sigma$) the Σ-measurable functions (resp. that are non-negative, bounded).*

3.2 Random variables

In the probabilistic case:

DEF 1.24 *A* random variable *is a measurable function on a probability triple $(\Omega, \mathcal{F}, \mathbb{P})$.*

The behavior of a random variable is characterized by its distribution function.

DEF 1.25 (Distribution function) *Let X be a RV on a triple $(\Omega, \mathcal{F}, \mathbb{P})$. The* law *of X is*

$$\mathcal{L}_X = \mathbb{P} \circ X^{-1},$$

which is a probability measure on $(\mathbb{R}, \mathcal{B})$. By LEM 1.19, $\mathcal{L}_X$ is determined by the distribution function *of X*

$$F_X(x) = \mathbb{P}[X \leq x], \quad x \in \mathbb{R}.$$

EX 1.26 *The distribution of a constant is a jump of size 1 at the value it takes. The distribution of Lebesgue measure on $(0, 1]$ is linear between 0 and 1.*

Distribution functions are characterized by a few simple properties.

PROP 1.27 *Suppose $F = F_X$ is the distribution function of a RV X on $(\Omega, \mathcal{F}, \mathbb{P})$. Then*

1. *F is non-decreasing.*
2. $\lim_{x\to+\infty} F(x) = 1$, $\lim_{x\to-\infty} F(x) = 0$.
3. *F is right-continuous.*

Proof: The first property follows from monotonicity.

For the second property, note that the limit exists by the first property. That the limit is 1 follows from the following important lemma.

LEM 1.28 (Monotone-convergence properties of measures) *Let (S, Σ, μ) be a measure space.*

1. *If $F_n \in \Sigma$, $n \geq 1$, with $F_n \uparrow F$, then $\mu(F_n) \uparrow \mu(F)$.*
2. *If $G_n \in \Sigma$, $n \geq 1$, with $G_n \downarrow G$ and $\mu(G_k) < +\infty$ for some k, then $\mu(G_n) \downarrow \mu(G)$.*

Proof: Clearly $F = \cup_n F_n \in \Sigma$. For $n \geq 1$, write $H_n = F_n \backslash F_{n-1}$ (with $F_0 = \emptyset$). Then by disjointness

$$\mu(F_n) = \sum_{k \leq n} \mu(H_k) \uparrow \sum_{k < +\infty} \mu(H_k) = \mu(F).$$

The second statement is similar. A counterexample when the finite assumption is violated is given by taking $G_n = (n, +\infty)$. ■

Similarly, for the third property, by LEM 1.28

$$\mathbb{P}[X \leq x_n] \downarrow \mathbb{P}[X \leq x],$$

if $x_n \downarrow x$. ■

It turns out that the properties above characterize distribution functions in the following sense.

THM 1.29 (Skorokhod representation) *Let F satisfy the three properties above. Then there is a RV X on*

$$(\Omega, \mathcal{F}, \mathbb{P}) = ([0,1], \mathcal{B}[0,1], \lambda),$$

with distribution function F. The law of X is called the Lebesgue-Stieltjes measure associated to F.

The result says that all real RVs can be generated from uniform RVs.

Proof: Assume first that F is continuous and strictly increasing. Define $X(\omega) = F^{-1}(\omega)$, $\omega \in \Omega$. Then, $\forall x \in \mathbb{R}$,

$$\mathbb{P}[X \leq x] = \mathbb{P}[\{\omega : F^{-1}(\omega) \leq x\}] = \mathbb{P}[\{\omega : \omega \leq F(x)\}] = F(x).$$

In general, let

$$X(\omega) = \inf\{x : F(x) \geq \omega\}.$$

It suffices to prove that

$$X(\omega) \leq x \quad \Longleftrightarrow \quad \omega \leq F(x). \tag{3}$$

The $\Leftarrow$ direction is obvious by definition. On the other hand,

$$x > X(\omega) \quad \Rightarrow \quad \omega \leq F(x).$$

By right-continuity of F, we have further $\omega \leq F(X(\omega))$ and therefore

$$X(\omega) \leq x \quad \Rightarrow \quad \omega \leq F(X(\omega)) \leq F(x).$$

■

Turning measurability on its head, we get the following important definition.

DEF 1.30 *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability triple. Let Y_γ, $\gamma \in \Gamma$, be a collection of maps from Ω to $\mathbb{R}$. We let*

$$\sigma(Y_\gamma, \gamma \in \Gamma)$$

be the smallest σ-algebra on which the Y_γ's are measurable.

In a sense, the above σ-algebra corresponds to the partial information available when the Y_γ's are observed.

EX 1.31 *Suppose we flip two unbiased coins and let X be the number of heads. Then*

$$\sigma(X) = \sigma(\{\{\mathtt{HH}\}, \{\mathtt{HT}, \mathtt{TH}\}, \{\mathtt{TT}\}\}),$$

which is coarser than 2^Ω.

3.3 Properties of measurable functions

Note that h^{-1} preserves all set operations. E.g., $h^{-1}(A \cup B) = h^{-1}(A) \cup h^{-1}(B)$. This gives the following important lemma.

LEM 1.32 (Sufficient condition for measurability) *Suppose $\mathcal{C} \subseteq \mathcal{B}$ with $\sigma(\mathcal{C}) = \mathcal{B}$. Then $h^{-1} : \mathcal{C} \to \Sigma$ implies $h \in m\Sigma$. That is, it suffices to check measurability on a collection generating $\mathcal{B}$.*

Proof: Let $\mathcal{E}$ be the sets such that $h^{-1}(B) \in \Sigma$. By the observation above, $\mathcal{E}$ is a σ-algebra. But $\mathcal{C} \subseteq \mathcal{E}$ which implies $\sigma(\mathcal{C}) \subseteq \mathcal{E}$ by minimality. ■

As a consequence we get the following properties of measurable functions.

PROP 1.33 (Properties of measurable functions) *Let h, h_n, $n \geq 1$, be in $m\Sigma$ and $f \in m\mathcal{B}$.*

1. *$f \circ h \in m\Sigma$.*
2. *If S is topological and h is continuous, then h is $\mathcal{B}(S)$-measurable, where $\mathcal{B}(S)$ is generated by the open sets of S.*
3. *The function $g : S \to \mathbb{R}$ is in $m\Sigma$ if for all $c \in \mathbb{R}$,*
$$\{g \leq c\} \in \Sigma.$$
4. *$\forall \alpha \in \mathbb{R}$, $h_1 + h_2, h_1 h_2, \alpha h \in m\Sigma$.*
5. *$\inf h_n$, $\sup h_n$, $\liminf h_n$, $\limsup h_n$ are in $m\Sigma$.*
6. *The set*
$$\{s : \lim h_n(s) \text{ exists in } \mathbb{R}\},$$
is measurable.

Proof: We sketch the proof of a few of them.

(2) This follows from LEM 1.32 by taking $\mathcal{C}$ as the open sets of $\mathbb{R}$.

(3) Similarly, take $\mathcal{C}$ to be the sets of the form $(-\infty, c]$.

(4) This follows from (3). E.g., note that, writing the LHS as $h_1 > c - h_2$,
$$\{h_1 + h_2 > c\} = \cup_{q \in \mathbb{Q}}[\{h_1 > q\} \cap \{q > c - h_2\}],$$
which is a countable union of measurable sets by assumption.

(5) Note that
$$\{\sup h_n \leq c\} = \cap_n \{h_n \leq c\}.$$
Further, note that $\lim\inf$ is the $\sup$ of an $\inf$. ■

Further reading

More background on measure theory [Dur10, Appendix A].

References

[Dur10] Rick Durrett. *Probability: theory and examples*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 2010.

[Fel68] William Feller. *An introduction to probability theory and its applications. Vol. I.* Third edition. John Wiley & Sons Inc., New York, 1968.

[Wil91] David Williams. *Probability with martingales*. Cambridge Mathematical Textbooks. Cambridge University Press, Cambridge, 1991.