Measure and probability

Peter D. Hoff

September 26, 2013

This is a very brief introduction to measure theory and measure-theoretic probability, designed to familiarize the student with the concepts used in a PhD-level mathematical statistics course. The presentation of this material was influenced by Williams [1991].

Contents

1 Algebras and measurable spaces

A measure μ assigns positive numbers to sets A: $\mu(A) \in \mathbb{R}$

- A a subset of Euclidean space, $\mu(A)$ = length, area or volume.
- A an event, $\mu(A)$ = probability of the event.

Let $\mathcal{X}$ be a space. What kind of sets should we be able to measure?

$\mu(\mathcal{X})$ = measure of whole space. It could be ∞, could be 1.
If we can measure A, we should be able to measure A^C.
If we can measure A and B, we should be able to measure $A \cup B$.

Definition 1 (algebra)**.** *A collection $\mathcal{A}$ of subsets of $\mathcal{X}$ is an algebra if*

1. $\mathcal{X} \in \mathcal{A}$;
2. $A \in \mathcal{A} \Rightarrow A^c \in \mathcal{A}$;
3. $A, B \in \mathcal{A} \Rightarrow A \cup B \in \mathcal{A}$.

$\mathcal{A}$ is closed under finitely many set operations.

For many applications we need a slightly richer collection of sets.

Definition 2 (σ-algebra)**.** *$\mathcal{A}$ is a σ-algebra if it is an algebra and for $A_n \in \mathcal{A}$, $n \in \mathbb{N}$, we have $\cup A_n \in \mathcal{A}$.*

$\mathcal{A}$ is closed under countably many set operations.

Exercise: Show $\cap A_n \in \mathcal{A}$.

Definition 3 (measurable space)**.** *A space $\mathcal{X}$ and a σ-algebra $\mathcal{A}$ on $\mathcal{X}$ is a measurable space $(\mathcal{X}, \mathcal{A})$.*

2 Generated σ-algebras

Let $\mathcal{C}$ be a set of subsets of $\mathcal{X}$

Definition 4 (generated σ-algebra)**.** *The σ-algebra generated by $\mathcal{C}$ is the smallest σ-algebra that contains $\mathcal{C}$, and is denoted $\sigma(\mathcal{C})$.*

Examples:

1. $\mathcal{C} = \{\phi\} \to \sigma(\mathcal{C}) = \{\phi, \mathcal{X}\}$
2. $\mathcal{C} = C \in \mathcal{A} \to \sigma(C) = \{\phi, C, C^c, \mathcal{X}\}$

Example (Borel sets):

Let $\mathcal{X} = \mathbb{R}$

$$\begin{aligned} \mathcal{C} &= \{C : C = (a,b), a < b, (a,b) \in \mathbb{R}^2\} = \text{open intervals} \\ \sigma(\mathcal{C}) &= \text{smallest } \sigma\text{-algebra containing the open intervals} \end{aligned}$$

Now let

$$\begin{aligned} G \in \mathcal{G} = \text{open sets} \quad &\Rightarrow \quad G = \cup C_n \text{ for some countable collection } \{C_n\} \subset \mathcal{C}. \\ &\Rightarrow \quad G \in \sigma(\mathcal{C}) \\ &\Rightarrow \quad \sigma(\mathcal{G}) \subset \sigma(\mathcal{C}) \end{aligned}$$

Exercise: Convince yourself that $\sigma(\mathcal{C}) = \sigma(\mathcal{G})$.

Exercise: Let $\mathcal{D}$ be the closed intervals, $\mathcal{F}$ the closed sets. Show

$$\sigma(\mathcal{C}) = \sigma(\mathcal{G}) = \sigma(\mathcal{F}) = \sigma(\mathcal{D})$$

Hint:

- $(a,b) = \cup_n [a + c/n, b - c/n]$
- $[a,b] = \cap_n (a - 1/n, b + 1/n)$

The sets of $\sigma(\mathcal{G})$ are called the "Borel sets of $\mathbb{R}$."
Generally, for any topological space $(\mathcal{X}, \mathcal{G})$, $\sigma(\mathcal{G})$ are known as the Borel sets.

3 Measure

Definition 5 (measure)**.** *Let $(\mathcal{X}, \mathcal{A})$ be a measurable space. A map $\mu : \mathcal{A} \to [0, \infty]$ is a measure if it is countably additive, meaning if $A_i \cap A_j = \phi$ for $\{A_n : n \in \mathbb{N}\} \subset \mathcal{A}$, then*

$$\mu(\cup_n A_n) = \sum_n \mu(A_n).$$

A measure is finite if $\mu(\mathcal{X}) < \infty$ (e.g. a probability measure)
A measure is σ-finite if $\exists \{C_n : n \in \mathbb{N}\} \subset \mathcal{A}$ with

1. $\mu(C_n) < \infty$,
2. $\cup_n C_n = \mathcal{X}$.

Definition 6 (measure space)**.** *The triple $(\mathcal{X}, \mathcal{A}, \mu)$ is called a measure space.*

Examples:

1. Counting measure: Let $\mathcal{X}$ be countable.
   - $\mathcal{A}$ = all subsets of $\mathcal{X}$ (show this is a σ-algebra)
   - $\mu(A)$ = number of points in A
2. Lebesgue measure: Let $\mathcal{X} = \mathbb{R}^n$

- $\mathcal{A}$ = Borel sets of $\mathcal{X}$
- $\mu(A) = \prod_{k=1}^{n}(a_k^H - a_k^L)$, for rectangles $A = \{x \in \mathbb{R}^n : a_k^L < x_k < a_k^H, k = 1, \ldots, n\}$.

The following is the foundation of the integration theorems to come.

Theorem 1 (monotonic convergence of measures)**.** *Given a measure space* $(\mathcal{X}, \mathcal{A}, \mu)$*,*

1. *If* $\{A_n\} \subset \mathcal{A},\ A_n \subset A_{n+1}$ *then* $\mu(A_n) \uparrow \mu(\cup A_n)$*.*
2. *If* $\{B_n\} \subset \mathcal{A},\ B_{n+1} \subset B_n$*, and* $\mu(B_k) < \infty$ *for some* k*, then* $\mu(B_n) \downarrow \mu(\cap B_n)$*.*

Exercise: Prove the theorem.

Example (what can go wrong):
Let $\mathcal{X} = \mathbb{R},\ \mathcal{A} = \mathcal{B}(\mathbb{R}),\ \mu = \text{Leb}$
Letting $B_n = (n, \infty)$, then

- $\mu(B_n) = \infty\ \forall n$;
- $\cap B_n = \phi$.

---

4 Integration of measurable functions

Let $(\Omega, \mathcal{A})$ be a measurable space.
Let $X(\omega) : \Omega \to \mathbb{R}$ (or $\mathbb{R}^p$, or $\mathcal{X}$)

Definition 7 (measurable function)**.** *A function* $X : \Omega \to \mathbb{R}$ *is measurable if*

$$\{\omega : X(\omega) \in B\} \in \mathcal{A}\ \forall B \in \mathcal{B}(\mathbb{R}).$$

So X is measurable if we can "measure it" in terms of $(\Omega, \mathcal{A})$.
Shorthand notation for a measurable function is "$X \in m\mathcal{A}$".

Exercise: If X, Y measurable, show the following are measurable:

- $X + Y$, XY, X/Y
- $g(X)$, $h(X, Y)$ if g, h are measurable.

Probability preview: Let $\mu(A) = \Pr(\omega \in A)$
Some $\omega \in \Omega$ "will happen." We want to know

$$\begin{aligned}\Pr(X \in B) &= \Pr(w : X(\omega) \in B) \\ &= \mu(X^{-1}(B))\end{aligned}$$

For the measure of $X^{-1}(B)$ to be defined, it has to be a measurable set,
i.e. we need $X^{-1}(B) = \{\omega : X(\omega) \in B\} \in \mathcal{A}\}$

We will now define the abstract Lebesgue integral for a very simple class of measurable functions, known as "simple functions." Our strategy for extending the definition is as follows:

1. Define the integral for "simple functions";
2. Extend definition to positive measurable functions;
3. Extend definition to arbitrary measurable functions.

Integration of simple functions

For a measurable set A, define its indicator function as follows:

$$I_A(\omega) = \begin{cases} 1 \text{ if } \omega \in A \\ 0 \text{ else} \end{cases}$$

Definition 8 (simple function). *$X(\omega)$ is simple if $X(\omega) = \sum_{k=1}^{K} x_k I_{A_k}(\omega)$, where*

- $x_k \in [0, \infty)$
- $A_j \cap A_k = \phi$, $\{A_k\} \subset \mathcal{A}$

Exercise: Show a simple function is measurable.

Definition 9 (integral of a simple function). *If X is simple, define*

$$\mu(X) = \int X(\omega)\mu(d\omega) = \sum_{k=1}^{K} x_k \mu(A_k)$$

Various other expressions are supposed to represent the same integral:

$$\int X d\mu \quad , \quad \int X d\mu(\omega) \quad , \quad \int X d\omega.$$

We will sometimes use the first of these when we are lazy, and will avoid the latter two.

Exercise: Make the analogy to expectation of a discrete random variable.

Integration of positive measurable functions

Let $X(\omega)$ be a measurable function for which $\mu(\omega : X(\omega) < 0) = 0$

- we say "$X \geq 0$ a.e. μ"
- we might write "$X \in (m\mathcal{A})^+$".

Definition 10. *For $X \in (m\mathcal{A})^+$, define*

$$\mu(X) = \int X(\omega)\mu(d\omega) = \sup\{\mu(X^*) : X^* \textit{is simple}, X^* \leq X\}$$

Draw the picture

Exercise: For $a, b \in \mathbb{R}$, show $\int (aX + bY)d\mu = a \int X d\mu + b \int Y d\mu$.

Most people would prefer to deal with limits rather than sups over classes of functions. Fortunately we can "calculate" the integral of a positive function X as the limit of the integrals of functions X_n that converge to X, using something called the monotone convergence theorem.

Theorem 2 (monotone convergence theorem)**.** *If $\{X_n\} \subset (m\mathcal{A})^+$ and $X_n(\omega) \uparrow X(\omega)$ as $n \to \infty$ a.e. μ, then*

$$\mu(X_n) = \int X_n \mu(d\omega) \uparrow \int X \mu(d\omega) = \mu(X) \text{ as } n \to \infty$$

With the MCT, we can explicitly construct $\mu(X)$: *Any* sequence of SF $\{X_n\}$ such that $X_n \uparrow X$ pointwise gives $\mu(X_n) \uparrow \mu(X)$ as $n \to \infty$.

Here is one in particular:

$$X_n(\omega) = \begin{cases} 0 & \text{if } X(\omega) = 0 \\ (k-1)/2^n & \text{if } (k-1)/2^n < X(\omega) < k/2^n < n, k = 1, \ldots, n2^n \\ n & \text{if } X(\omega) > n \end{cases}$$

Exercise: Draw the picture, and confirm the following:

1. $X_n(\omega) \in (m\mathcal{A})^+$;

2. $X_n \uparrow X$;

3. $\mu(X_n) \uparrow \mu(X)$ (by MCT).

Riemann versus Lebesgue

Draw picture

Example:

Let $(\Omega, \mathcal{A}) = ([0,1], \mathcal{B}([0,1]))$

$$X(\omega) = \begin{cases} 1 & \text{if } \omega \text{ is rational} \\ 0 & \text{if } \omega \text{ is irrational} \end{cases}$$

Then

$$\int_0^1 X(\omega) d\omega \ \text{ is undefined, but } \int_0^1 X(\omega)\mu(d\omega)$$

Integration of integrable functions

We now have a definition of $\int X(\omega)\mu(d\omega)$ for positive measurable X. What about for measurable X in general?

Let $X \in m\mathcal{A}$. Define

- $X^+(\omega) = X(\omega) \vee 0 > 0$, the positive part of X;

- $X^-(\omega) = (-X(\omega)) \vee 0 > 0$, the negative part of X.

Exercise: Show

- $X = X^+ - X^-$

- X^+, X^- both measurable

Definition 11 (integrable, integral)**.** *$X \in m\mathcal{A}$ is integrable if $\int X^+ d\mu$ and $\int X^- d\mu$ are both finite. In this case, we define*

$$\mu(X) = \int X(\omega)\mu(d\omega) = \int X^+(\omega)\mu(d\omega) - \int X^-(\omega)\mu(d\omega).$$

Exercise: Show $|\mu(X)| \leq \mu(|X|)$.

---

5 Basic integration theorems

Recall $\liminf_{n\to\infty} c_n = \lim_{n\to\infty}(\inf_{k\geq n} c_k)$
$\limsup_{n\to\infty} c_n = \lim_{n\to\infty}(\sup_{k\geq n} c_k)$

Theorem 3 (Fatou's lemma)**.** *For* $\{X_n\} \subset (m\mathcal{A})^+$,

$$\mu(\liminf X_n) \leq \liminf \mu(X_n)$$

Theorem 4 (Fatou's reverse lemma)**.** *For* $\{X_n\} \subset (m\mathcal{A})^+$ *and* $X_n \leq Z\ \forall n$, $\mu(Z) < \infty$,

$$\mu(\limsup X_n) \geq \limsup \mu(X_n)$$

I most frequently encounter Fatou's lemmas in the proof of the following:

Theorem 5 (dominated convergence theorem)**.** *If* $\{X_n\} \subset m\mathcal{A}$, $|X_n| < Z$ *a.e.* μ, $\mu(Z) < \infty$ *and* $X_n \to X$ *a.e.* μ, *then*

$$\mu(|X_n - X|) \to 0, \quad \textit{which implies} \;\; \mu(X_n) \to \mu(X).$$

Proof.
$|X_n - X| \leq 2Z$, $\mu(2Z) = 2\mu(Z) < \infty$
By reverse Fatou, $\limsup \mu(|X_n - X|) \leq \mu(\limsup |X_n - X|) = \mu(0) = 0$.
To show $\mu(X_n) \to \mu(X)$, note

$$|\mu(X_n) - \mu(X)| = |\mu(X_n - X)| \leq \mu(|X_n - X|) \to 0.$$

□

Among the four integration theorems, we will make the most use of the MCT and the DCT:

MCT : If $\{X_n\} \in (m\mathcal{A})^+$ and $X_n \uparrow X$, then $\mu(X_n) \to \mu(X)$.

DCT : If $\{X_n\}$ are dominated by an integrable function and $X_n \to X$, then $\mu(X_n) \to \mu(X)$.

6 Densities and dominating measures

One of the main concepts from measure theory we need to be familiar with for statistics is the idea of a family of distributions (a model) the have densities with respect to a common dominating measure.

Examples:

- The normal distributions have densities with respect to Lebesgue measure on $\mathbb{R}$.
- The Poisson distributions have densities with respect to counting measure on $\mathbb{N}_0$.

Density

Theorem 6. *Let $(\mathcal{X}, \mathcal{A}, \mu)$ be a measure space, $f \in (m\mathcal{A})^+$. Define*

$$\nu(A) = \int_A f d\mu = \int 1_A(x) f(x) \mu(dx)$$

Then ν is a measure on $(\mathcal{X}, \mathcal{A})$.

Proof. We need to show that ν is countably additive. Let $\{A_n\} \subset \mathcal{A}$ be disjoint. Then

$$\begin{aligned}
\nu(\cup A_n) &= \int_{\cup A_n} f d\mu \\
&= \int 1_{\cup A_n}(x) f(x) \mu(dx) \\
&= \int \sum_{n=1}^{\infty} f(x) 1_{A_n}(x) \mu(dx) \\
&= \int \lim_{k\to\infty} g_k(x) \mu(dx),
\end{aligned}$$

where $g_k(x) = \sum_{n=1}^{k} f(x) 1_{A_n}(x)$. Since $0 \leq g_k(x) \uparrow 1_{\cup A_n}(x) f(x) \equiv g(x)$, by the MCT

$$\begin{aligned}
\nu(\cup A_n) = \int \lim_{k\to\infty} g_k d\mu = \lim_{k\to\infty} \int g_k d\mu \\
&= \lim_{k\to\infty} \int \sum_{n=1}^{k} f(x) 1_{A_n}(x) d\mu \\
&= \lim_{k\to\infty} \sum_{n=1}^{k} \int_{A_n} f d\mu = \sum_{n=1}^{\infty} \nu(A_n)
\end{aligned}$$

□

Definition 12 (density)**.** *If $\nu(A) = \int_A f d\mu$ for some $f \in (m\mathcal{A})^+$ and all $A \in \mathcal{A}$, we say that the measure ν has density f with respect to μ.*

Examples:

- $\mathcal{X} = \mathbb{R}$, μ is Lebesgue measure on $\mathbb{R}$, f a normal density $\Rightarrow \nu$ is the normal distribution (normal probability measure).
- $\mathcal{X} = \mathbb{N}_0$, μ is counting measure on $\mathbb{N}_0$, f a Poisson density $\Rightarrow \nu$ is the Poisson distribution (Poisson probability measure).

Note that in the latter example, f is a density even though it isn't continuous in $x \in \mathbb{R}$.

Radon-Nikodym theorem
For $f \in (m\mathcal{A})^+$ and $\nu(A) = \int_A f d\mu$,

- ν is a measure on $(\mathcal{X}, \mathcal{A})$,
- f is called the density of ν w.r.t. μ (or "ν has density f w.r.t. μ).

Exercise: If ν has density f w.r.t. μ, show $\mu(A) = 0 \Rightarrow \nu(A) = 0$.

Definition 13 (absolutely continuous)**.** *Let μ, ν be measures on $\mathcal{X}, \mathcal{A}$. The measure ν is absolutely continuous with respect to μ if $\mu(A) = 0 \Rightarrow \nu(A) = 0$.*

If ν is absolutely continuous w.r.t. μ, we might write either

- "ν is dominated by μ" or
- "$\nu \ll \mu$."

Therefore, $\mu(A) = \int_A f d\mu \Rightarrow \nu \ll \mu$.
What about the other direction?

Theorem 7 (Radon-Nikodym theorem)**.** *Let $(\mathcal{X}, \mathcal{A}, \mu)$ be a σ-finite measure space, and suppose $\nu \ll \mu$. Then there exists an $f \in (m\mathcal{A})^+$ s.t.*

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

In other words

$$\nu \ll \mu \Leftrightarrow \nu \text{ has a density w.r.t. } \mu$$

Change of measure Sometimes we will say "f is the RN derivative of ν w.r.t. μ", and write $f = \frac{d\nu}{d\mu}$.
This helps us with notation when "changing measure:"

$$\int g \; d\nu = \int g \left[\frac{d\nu}{d\mu}\right] \; d\mu = \int g f d\mu$$

You can think of ν as a probability measure, and g as a function of the random variable. The expectation of g w.r.t. ν can be computed from the integral of gf w.r.t μ.

Example:

$$\int x^2 \sigma^{-1} \phi([x-\theta]/\sigma) dx$$

- $g(x) = x^2$;
- μ is Lebesgue measure, here denoted with "dx";
- ν is the normal(θ, σ^2) probability measure;
- $d\nu/d\mu = f = \sigma^{-1}\phi([x-\theta]/\sigma)$ is the density of ν w.r.t. μ.

---

7 Product measures

We often have to work with joint distributions of multiple random variables living on potentially different measure spaces, and will want to compute integrals/expectations of multivariate functions of these variables. We need to define integration for such cases appropriately, and develop some tools to actually do the integration.

Let $(\mathcal{X}, \mathcal{A}_x, \mu_x)$ and $(\mathcal{Y}, \mathcal{B}_y, \mu_y)$ be σ-finite measure spaces. Define

$$\mathcal{A}_{xy} = \sigma(F \times G : F \in \mathcal{A}_x, G \in \mathcal{A}_y)$$
$$\mu_{xy}(F \times G) = \mu_x(F)\mu_y(G)$$

Here, $(\mathcal{X} \times \mathcal{Y}, \mathcal{A}_{xy})$ is the "product space", and $\mu_x \times \mu_y$ is the "product measure."

Suppose $f(x, y)$ is an $\mathcal{A}_{xy}$-measurable function. We then might be interested in

$$\int_{\mathcal{X}\times\mathcal{Y}} f(x, y)\mu_{xy}(dx \times dy).$$

The "calculus" way of doing this integral is to integrate first w.r.t. one variable, and then w.r.t. the other. The following theorems give conditions under which this is possible.

Theorem 8 (Fubini's theorem)**.** *Let $(\mathcal{X}, \mathcal{A}_x, \mu_x)$ and $(\mathcal{Y}, \mathcal{A}_y, \mu_y)$ be two complete measure spaces and f be $\mathcal{A}_{xy}$-measurable and $\mu_x \times \mu_y$-integrable. Then*

$$\int_{\mathcal{X}\times\mathcal{Y}} f \; d(\mu_x \times \mu_y) = \int_X \left[\int_Y f \; d\mu_y\right] \; d\mu_x = \int_Y \left[\int_X f \; d\mu_x\right] \; d\mu_y$$

Additionally,

1. *$f_x(y) = f(x, y)$ is an integrable function of y for x a.e. μ_x.*

2. *$\int f(x, y) \; d\mu_x(x)$ is μ_y-integrable as a function of y.*

Also, items 1 and 2 hold with the roles of x and y reversed.

The problem with Fubini's theorem is that often you don't know of f is $\mu_x \times \mu_y$-integrable without being able to integrate variable-wise. In such cases the following theorem can be helpful.

Theorem 9 (Tonelli's theorem)**.** *Let $(\mathcal{X}, \mathcal{A}_x, \mu_x)$ and $(\mathcal{Y}, \mathcal{A}_y, \mu_y)$ be two σ-finite measure spaces and f in$(m\mathcal{A}_{xy})^+$. Then*

$$\int_{\mathcal{X}\times\mathcal{Y}} f \; d(\mu_x \times \mu_y) = \int_X \left[\int_Y f \; d\mu_y\right] \; d\mu_x = \int_Y \left[\int_X f \; d\mu_x\right] \; d\mu_y$$

Additionally,

1. *$f_x(y) = f(x, y)$ is a measurable function of y for x a.e. μ_x.*

2. *$\int f(x, y) \; d\mu_x(x)$ is $\mathcal{A}_y$-measurable as a function of y.*

Also, 1 and 2 hold with the roles of x and y reversed.

8 Probability measures

Definition 14 (probability space)**.** *A measure space* (Ω, A, P) *is a probability space if* $P(\Omega) = 1$*. In this case, P is called a probability measure.*

Interpretation: Ω is the space of all possible outcomes, $\omega \in \Omega$ is a possible outcome.

Numerical data X is a function of the outcome ω: $X = X(\omega)$
Uncertainty in the outcome leads to uncertainty in the data.
This uncertainty is referred to as "randomness", and so $X(\omega)$ is a "random variable."

Definition 15 (random variable)**.** *A random variable* $X(\omega)$ *is a real-valued measurable function in a probability space.*

Examples:

- multivariate data: $X : \Omega \to \mathbb{R}^p$
- replications: $X : \Omega \to \mathbb{R}^n$
- replications of multivariate data: $X : \Omega \to \mathbb{R}^{n \times p}$

Suppose $X : \Omega \to \mathbb{R}^k$.
For $B \in \mathcal{B}(\mathbb{R}^k)$, we might write $P(\{\omega : X(\omega) \in B\})$ as $P(B)$.

Often, the "Ω-layer" is dropped and we just work with the "data-layer:"
$(\mathcal{X}, \mathcal{A}, P)$ is a measure space , $P(A) = \Pr(X \in A)$ for $A \in \mathcal{A}$.

Densities
Suppose $P \ll \mu$ on $(\mathcal{X}, \mathcal{A})$. Then by the RN theorem, $\exists p \in (m\mathcal{A})^+$ s.t.

$$P(A) = \int_A p \, d\mu = \int_A p(x)\mu(dx).$$

Then p is the probability density of P w.r.t. μ.
(probability density = Radon-Nikodym derivative)
Examples:

1. Discrete:
   $\mathcal{X} = \{x_k : k \in \mathbb{N}\}$, $\mathcal{A}$ = all subsets of $\mathcal{X}$.
   Typically we write $P(\{x_k\}) = p(x_k) = p_k,\ 0 \le p_k \le 1,\ \sum p_k = 1$.

2. Continuous:
   $\mathcal{X} = \mathbb{R}^k$, $\mathcal{A} = \mathcal{B}(\mathbb{R}^k)$.
   $P(A) = \int_A p(x)\mu(dx)$, μ =Lebesgue measure on $\mathcal{B}(\mathbb{R}^k)$.

3. Mixed discrete and continuous:

$$Z \sim N(0,1),\ X = \begin{cases} Z & \text{w.p. } 1/2 \\ 0 & \text{w.p. } 1/2 \end{cases}$$

   Define P by $P(A) = \Pr(X \in A)$ for $A \in \mathcal{B}(\mathbb{R})$. Then

   (a) $P \not\ll \mu_L$ $\ (\mu_L(\{0\}) = 0,\ P(\{0\}) = 1/2\ )$

   (b) $P \ll \mu = \mu_L + \mu_0$, where $\mu_0 = \#(A \cap \{0\})$ for $A \in \mathcal{A}$.

   Exercise: Verify $(\mathbb{R}, \mathcal{B}(\mathbb{R}), P)$ is a measure space and $P \ll \mu$.

The following is a concept you are probably already familiar with:

Definition 16 (support)**.** *Let $(\mathcal{X}, \mathcal{G})$ be a topological space, and $(\mathcal{X}, \mathcal{B}(\mathcal{G}), P)$ be a probability space. The support of P is given by*

$$\text{supp}(P) = \{x \in \mathcal{X} : P(G) > 0 \textit{ for all } G \in \mathcal{G} \textit{ containing } x\}$$

Note that the notion of support requires a topology on $\mathcal{X}$.
Examples:

- Let P be a univariate normal probability measure. Then $\text{supp}(P) = \mathbb{R}$.
- Let $X = [0,1]$, $\mathcal{G}$ be the open sets defined by Euclidean distance, and $P(\mathbb{Q} \cap [0,1]) = 1$.
  1. $(\mathbb{Q}^c \cap [0,1]) \subset \text{supp}(P)$ but
  2. $P(\mathbb{Q}^c \cap [0,1]) = 0$.

---

9 Expectation

In probability and statistics, a weighted average of a function, i.e. the integral of a function w.r.t. a probability measure, is (unfortunately) referred to as its expectation or expected value.

Definition 17 (expectation)**.** *Let* $(\mathcal{X}, \mathcal{A}, P)$ *be a probability space and let* $T(X)$ *be a measurable function of* X *(i.e. a statistic). The expectation of* T *is its integral over* $\mathcal{X}$*:*

$$\mathrm{E}[T] = \int T(x)P(dx).$$

Why is this definition unfortunate? Consider a highly skewed probability distribution. Where do you "expect" a sample from this distribution to be?

Jensen's inequality

Recall that a convex function $g : \mathbb{R} \to \mathbb{R}$ is one for which

$$g(pX_1 + (1-p)X_2) \leq pg(X_1) + (1-p)g(X_2), \quad X_1, X_2 \in \mathbb{R},\ p \in [0,1],$$

i.e. "the function at the average is less than the average of the function."

Draw a picture.

The following theorem should therefore be no surprise:

Theorem 10 (Jensen's inequality)**.** *Let* $g : \mathbb{R} \to \mathbb{R}$ *be a convex function and* X *be a random variable on* $(\mathbb{R}, \mathcal{B}(\mathbb{R}), P)$ *such that* $\mathrm{E}[|X|] < \infty$ *and* $\mathrm{E}[|g(X)|] < \infty$*. Then*

$$g(\mathrm{E}[X]) \leq \mathrm{E}[g(X)].$$

i.e. "the function at the average is less than the average of the function."

The result generalizes to more general sample spaces.

Schwarz's inequality

Theorem 11 (Schwarz's inequality)**.** *If* $\int X^2\ dP$ *and* $\int Y^2\ dP$ *are finite, then* $\int XY\ dP$ *is finite and*

$$|\int XY\ dP| \leq \int |XY|\ dP \leq (\int X^2\ dP)^{1/2}(\int Y^2\ dP)^{1/2}.$$

In terms of expectation, the result is

$$\mathrm{E}[XY]^2 \leq \mathrm{E}[|XY|]^2 \leq \mathrm{E}[X^2]\mathrm{E}[Y^2].$$

One statistical application is to show that the correlation coefficient is always between -1 and 1.

Hölders inequality

A more general version of Schwarz's inequality is Hölder's inequality.

Theorem 12 (Hölder's inequuality)**.** *Let*

- $w \in (0,1)$,
- $\mathrm{E}[X^{1/w}] < \infty$ *and*
- $\mathrm{E}[Y^{1/(1-w)}] < \infty$.

Then $\mathrm{E}[|XY|] < \infty$ *and*

$$|\mathrm{E}[XY]| \leq \mathrm{E}[|XY|] \leq \mathrm{E}[X^{1/w}]^w \mathrm{E}[Y^{1/(1-w)}]^{1-w}.$$

Exercise: Prove this inequality from Jensen's inequality.

---

10 Conditional expectation and probability

Conditioning in simple cases:

$X \in \{x_1, \ldots, x_K\} = \mathcal{X}$
$Y \in \{y_1, \ldots, y_M\} = \mathcal{Y}$
$\Pr(X = x_k | Y = y_m) = \Pr(X = x_k, Y = y_m) / \Pr(Y = y_m)$
$\mathrm{E}[X|Y = y_m] = \sum_{k=1}^{K} x_k \Pr(X = x_k | Y = y_m)$

This discrete case is fairly straightforward and intuitive. We are also familiar with the extension to the continuous case:

$$\mathrm{E}[X|Y = y] = \int x p(x|y)\ dx = \int x \left[\frac{p(x,y)}{p(y)}\right]\ dx$$

Where does this extension come from, and why does it work? Can it be extended to more complicated random variables?

Introduction to Kolmogorov's formal theory:

Let $\{\Omega, \mathcal{A}, P\}$ be a probability space and X, Y random variables with finite supports $\mathcal{X}$, $\mathcal{Y}$.
Suppose $\mathcal{A}$ contains all sets of the form $\{\omega : X(\omega) = x, Y(\omega) = y\}$ for $(x, y) \in \mathcal{X} \times \mathcal{Y}$.
Draw the picture.
Let $\mathcal{F}$, $\mathcal{G}$ be the σ-algebras consisting of all subsets of $\mathcal{X}$ and $\mathcal{Y}$, respectively.
Add $\mathcal{F}$, $\mathcal{G}$ to the picture (rows and columns of $\mathcal{X} \times \mathcal{Y}$-space)
In the Kolmogorov theory, $\mathrm{E}[X|Y]$ is a random variable Z defined as follows:

$$Z(\omega) = \begin{cases} \mathrm{E}[X|Y = y_1] & \text{if } Y(\omega) = y_1 \\ \mathrm{E}[X|Y = y_2] & \text{if } Y(\omega) = y_1 \\ & \vdots \\ \mathrm{E}[X|Y = y_M] & \text{if } Y(\omega) = y_M \end{cases}$$

We say that $Z = \mathrm{E}[X|Y]$ is a (version) of the conditional expectation of X given Y. Note the following:

1. $\mathrm{E}[X|Y]$ is a random variable;
2. $\mathrm{E}[X|Y]$ is a function of ω only through $Y(\omega)$.

This latter fact makes $\mathrm{E}[X|Y]$ "$\sigma(Y)$-measurable", where

$$\sigma(Y) = \sigma(\{\omega : Y(\omega) \in F\}, F \in \mathcal{F}\})$$

$\sigma(Y)$ is the smallest σ-algebra on Ω that makes Y measurable.
This means we don't need the whole σ-algebra $\mathcal{A}$ to "measure" $\mathrm{E}[X|Y]$, we just need the part that determines Y.
<u>Defining properties of conditional expectation</u>

$$\begin{aligned} \int_{Y=y} \mathrm{E}[X|Y]\, dP = \mathrm{E}[X|Y = y] P(Y = y) &= \sum_x x P(X = x|Y = y) P(Y = y) \\ &= \sum_x x \Pr(X = x, Y = y) = \int_{Y=y} X\, dP \end{aligned}$$

In words, the integral of $\mathrm{E}[X|Y]$ over the set $Y = y$ equals the integral of X over $Y = y$.

In this simple case, it is easy to show

$$\int_A \mathrm{E}[X|Y]\ dP = \int_A X\ dP\ \forall G \in \sigma(Y)$$

In words, the integral of $\mathrm{E}[X|Y]$ over any $\sigma(Y)$-measurable set is the same as that of X. Intuitively, $\mathrm{E}[X|Y]$ is "an approximation" to X, matching X in terms of expectations over sets defined by Y.

Kolmogorov's fundamental theorem and definition

Theorem 13 (Kolmogorov,1933)**.** *Let* $(\Omega, \mathcal{A}, P)$ *be a probability space, and* X *a r.v. with* $\mathrm{E}[|X|] < \infty$. *Let* $\mathcal{G} \subset \mathcal{A}$ *be a sub-*σ *algebra of* $\mathcal{A}$. *Then* $\exists$ *a r.v.* $\mathrm{E}[X|\mathcal{G}]$ *s.t.*

1. $\mathrm{E}[X|\mathcal{G}]$ *is* $\mathcal{G}$*-measurable*
2. $\mathrm{E}[|\mathrm{E}[X|\mathcal{G}]|] < \infty$
3. $\forall G \in \mathcal{G}$,
$$\int_G \mathrm{E}[X|\mathcal{G}] dP = \int_G X dP.$$

Technically, a random variable satisfying 1, 2 and 3 is called "a version of $\mathrm{E}[X|\mathcal{G}]$", as the conditions only specify things a.e. P.

From 1,2 and 3, the following properties hold

(a) $\mathrm{E}[\mathrm{E}[X|\mathcal{G}]] = \mathrm{E}[X]$.

(b) If $X \in m\mathcal{G}$, then $\mathrm{E}[X|\mathcal{G}] = X$.

(c) If $\mathcal{H} \subset \mathcal{G}$, $\mathcal{H}$ a σ-algebra, then $\mathrm{E}[\mathrm{E}[X|\mathcal{G}]|\mathcal{H}] = \mathrm{E}[X|\mathcal{H}]$

(d) If $Z \in m\mathcal{G}$ and $|ZX|$ is integrable, $\mathrm{E}[ZX|\mathcal{G}] = Z\mathrm{E}[X|\mathcal{G}]$.

Proving (a) and (b) should be trivial.
For (c), we need to show that $\mathrm{E}[X|\mathcal{H}]$ "is a version of" $\mathrm{E}[Z|\mathcal{H}]$, where $Z = \mathrm{E}[X|\mathcal{G}]$
This means the integral of $\mathrm{E}[X|\mathcal{H}]$ over any $\mathcal{H}$-measurable set H must equal that of Z over H. Let's check:

$$\begin{aligned}\int_H \mathrm{E}[X|\mathcal{H}]\ dP &= \int_H X\ dP, \quad \text{by definition of } \mathrm{E}[X|\mathcal{H}] \\ &= \int_H \mathrm{E}[X|\mathcal{G}]\ dP, \quad \text{since } H \in \mathcal{H} \subset \mathcal{G}.\end{aligned}$$

Exercise: Prove (d).

Independence

Definition 18 (independent σ-algebras)**.** *Let $(\Omega, \mathcal{A}, P)$ be a probability space. The sub-σ-algebras $\mathcal{G}$ and $\mathcal{H}$ are independent if $P(A \cap B) = P(A)P(B)$ $\forall A \in \mathcal{G}, B \in \mathcal{H}$.*

This notion of independence allows us to describe one more intuitive property of conditional expectation.

(e) If $\mathcal{H}$ is independent of $\sigma(X)$, then $\mathrm{E}[X|\mathcal{H}] = \mathrm{E}[X]$.

Intuitively, if X is independent of $\mathcal{H}$, then knowing where you are in $\mathcal{H}$ isn't going to give you any information about X, and so the conditional expectation is the same as the unconditional one.

Interpretation as a projection

Let $X \in m\mathcal{A}$, with $\mathrm{E}[X^2] < \infty$.

Let $\mathcal{G} \subset A$ be a sub-σ-algebra.

Problem: Represent X by a $\mathcal{G}$-measurable function/r.v. Y s.t. expected squared error is minimized, i.e.

$$\text{minimize}\mathrm{E}[(X-Y)^2]\text{among } Y \in m\mathcal{G}$$

Solution: Suppose Y is the minimizer, and let $Z \in m\mathcal{G}$, $\mathrm{E}[Z^2] < \infty$.

$$\begin{aligned}\mathrm{E}[(X-Y)^2] &\le \mathrm{E}[X - Y - \epsilon Z)^2] \\ &= \mathrm{E}[(X-Y)^2] - 2\epsilon\mathrm{E}[Z(X-Y)] + \epsilon^2\mathrm{E}[Z^2].\end{aligned}$$

This implies

$$\begin{aligned}2\epsilon\mathrm{E}[Z(X-Y)] &\le \epsilon^2\mathrm{E}[Z^2] \\ 2\mathrm{E}[Z(X-Y)] &\le \epsilon\mathrm{E}[Z^2] \text{ for } \epsilon > 0 \\ 2\mathrm{E}[Z(X-Y)] &\ge \epsilon\mathrm{E}[Z^2] \text{ for } \epsilon < 0\end{aligned}$$

which implies that $\mathrm{E}[Z(X-Y)] = 0$. Thus if Y is the minimizer then it must satisfy

$$\mathrm{E}[ZX] = \mathrm{E}[ZY]\ \forall Z \in m\mathcal{G}.$$

In particular, let $Z = 1_G(\omega)$ for any $G \in \mathcal{G}$. Then

$$\int_G X\ dP = \int_G Y\ dP,$$

so Y must be a version of $\mathrm{E}[X|\mathcal{G}]$.

---

11 Conditional probability

Conditional probability

For $A \in \mathcal{A}$, $\Pr(A) = \mathrm{E}[1_A(\omega)]$.

For a σ-algebra $\mathcal{G} \subset \mathcal{A}$, define $\Pr(A|\mathcal{G}) = \mathrm{E}[1_A(\omega)|\mathcal{G}]$.

Exercise: Use linearity of expectation and MCT to show

$$\Pr(\cup A_n|\mathcal{G}) = \sum \Pr(A_n|\mathcal{G})$$

if the $\{A_n\}$ are disjoint.

Conditional density

Let $f(x, y)$ be a joint probability density for X, Y w.r.t. a dominating measure $\mu \times \nu$, i.e.

$$P((X,Y) \in B) = \iint_B f(x,y)\mu(dx)\nu(dy).$$

Let $f(x|y) = f(x,y)/\int f(x,y)\nu(dy)$

Exercise: Prove $\int_A f(x|y)\mu(dx)$ is a version of $\Pr(X \in A|Y)$.

This, and similar exercises, show that our "simple" approach to conditional probability generally works fine.

References

David Williams. *Probability with martingales.* Cambridge Mathematical Textbooks. Cambridge University Press, Cambridge, 1991. ISBN 0-521-40455-X; 0-521-40605-6.