Introduction to Probability Theory

Unless otherwise noted, references to Theorems, page numbers, etc. from Casella-Berger, chap 1.

Statistics: draw conclusions about a population of objects by sampling from the population

1 Probability space

We start by introducing mathematical concept of a **probability space**, which has three components $(\Omega, \mathcal{B}, P)$, respectively the *sample space*, *event space*, and *probability function*. We cover each in turn.

■■■

Ω: sample space. Set of *outcomes* of an experiment.

Example: tossing a coin twice. $\Omega = \{HH, HT, TT, TH\}$

■■■

An *event* is a subset of Ω. Examples: (i) "at least one head" is $\{HH, HT, TH\}$; (ii) "no more than one head" is $\{HT, TH, TT\}$. &etc.

In probability theory, the event space $\mathcal{B}$ is modelled as a σ-*algebra* (or σ-field) of Ω, which is a collection of subsets of Ω with the following properties:

(1) $\emptyset \in \mathcal{B}$
(2) If an event $A \in \mathcal{B}$, then $A^c \in \mathcal{B}$ (closed under complementation)
(3) If $A_1, A_2, \ldots \in \mathcal{B}$, then $\cup_{i=1}^{\infty} A_i \in \mathcal{B}$ (closed under countable union). A countable sequence can be indexed using the natural integers.

Additional properties:
(4) (1)+(2) $\rightarrow \Omega \in \mathcal{B}$
(5) (3)+De-Morgan's Laws[1] $\rightarrow \cap_{i=1}^{\infty} A_i \in \mathcal{B}$ (closed under coutable intersection)

■■■

Consider the two-coin toss example again. Even for this simple sample space $\Omega = \{HH, HT, TT, TH\}$, there are multiple σ-algebras:

1. $\{\emptyset, \Omega\}$: "trivial" σ-algebra

[1] $(A \cup B)^c = A^c \cap B^c$

2. The "powerset" $\mathcal{P}(\Omega)$, which contains all the subsets of Ω

■■■

In practice, rather than specifying a particular σ-algebra from scratch, there is usually a class of events of interest, $\mathcal{C}$, which we want to be included in the σ-algebra. Hence, we wish to "complete" $\mathcal{C}$ by adding events to it so that we get a σ-algebra.

For example, consider 2-coin toss example again. We find the smallest σ-algebra containing $(HH), (HT), (TH), (TT)$; we call this the σ-algebra "generated" by the fundamental events $(HH), (HT), (TH), (TT)$. It is...

Formally, let $\mathcal{C}$ be a collection of subsets of Ω. The *minimal σ-field generated by $\mathcal{C}$*, denoted $\sigma(\mathcal{C})$, satisfies: (i) $\mathcal{C} \subset \sigma(\mathcal{C})$; (ii) if $\mathcal{B}'$ is any other σ-field containing $\mathcal{C}$, then $\sigma(\mathcal{C}) \subset \mathcal{B}'$.

■■■

Finally, a *probability function* P assigns a number ("probability") to each event in $\mathcal{B}$. It is a function mapping $\mathcal{B} \to [0, 1]$ satisfying:

1. $P(A) \geq 0$, for all $A \in \mathcal{B}$.
2. $P(\Omega) = 1$
3. Countable additivity: If $A_1, A_2, \cdots \in \mathcal{B}$ are pairwise disjoint (i.e., $A_i \cap A_j = \emptyset$, for all $i \neq j$), then $P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$.

Define: *Support* of P is the set $\{A \in \mathcal{B} : P(A) > 0\}$.

Example: Return to 2-coin toss. Assuming that the coin is fair (50/50 chance of getting heads/tails), then the probability function for the σ-algebra consisting of all subsets of Ω is

Event A	$P(A)$
HH	$\frac{1}{4}$
HT	$\frac{1}{4}$
TH	$\frac{1}{4}$
TT	$\frac{1}{4}$
$\emptyset$	0
Ω	1
(HH, HT, TH)	$\frac{3}{4}$ (using pt. (3) of Def'n above)
(HH,HT)	$\frac{1}{2}$
$\vdots$	$\vdots$

1.1 Probability on the real line

In statistics, we frequently encounter probability spaces defined on the real line (or a portion thereof). Consider the following probability space: $([0,1], \mathbb{B}([0,1]), \mu)$

1. The sample space is the real interval $[0,1]$

2. $\mathbb{B}([0,1])$ denotes the "Borel" σ-algebra on [0,1]. This is the minimal σ-algebra generated by the elementary events $\{[0,b), 0 \leq b \leq 1\}$. This collection contains things like $[\frac{1}{2}, \frac{2}{3}]$, $[0, \frac{1}{2}] \cup (\frac{2}{3}, 1]$, $\left\{\frac{1}{2}\right\}$, $[\left\{\frac{1}{2}\right\}, \left\{\frac{2}{3}\right\}]$.

   - To see this, note that closed intervals can be generated as countable intersections of open intervals (and vice versa):

$$\begin{aligned} \lim_{n\to\infty} [0, 1/n) &= \cap_{n=1}^{\infty} [0, 1/n) = \{0\}, \\ \lim_{n\to\infty} (0, 1/n) &= \cap_{n=1}^{\infty} (0, 1/n) = \emptyset, \\ \lim_{n\to\infty} (a - 1/n, b + 1/n) &= \cap_{n=1}^{\infty} (a - 1/n, b + 1/n) = [a, b] \\ \lim_{n\to\infty} [a + 1/n, b - 1/n] &= \cup_{n=1}^{\infty} [a + 1/n, b - 1/n] = (a, b) \end{aligned} \tag{1}$$

     (Limit has unambiguous meaning because the set sequences are monotonic.)
   - Thus, $\mathbb{B}([0,1])$ can equivalently be characterized as the minimal σ-field generated by: (i) the open intervals (a,b) on $[0,1]$; (ii) the closed intervals $[a,b]$; (iii) the closed half-lines $[0,a]$, and so on.
   - Moreover: it is also the minimal σ-field containing all the open sets in $[0,1]$:
$$\mathbb{B}([0,1]) = \sigma(\text{open sets on } [0,1]).$$
   - This last characterization of the Borel field, as the minimal σ-field containing the open subsets, can be generalized to any metric space (ie. so that "openness" is defined). This includes $\mathbb{R}$, $\mathbb{R}^k$, even functional spaces (eg. $\mathcal{L}^2[a,b]$, the space of square-integrable functions on $[a,b]$).

3. $\mu(\cdot)$, for all $A \in \mathcal{B}$, is *Lebesgue measure*, defined as the sum of the lengths of the intervals contained in A. Eg.: $\mu([\frac{1}{2}, \frac{2}{3}]) = \frac{1}{6}$, $\mu([0, \frac{1}{2}] \cup (\frac{2}{3}, 1]) = \frac{5}{6}$, $\mu([\frac{1}{2}]) = 0$.

■■■

More examples: Consider the measurable space $([0,1], \mathcal{B})$. Are the following probability measures?

- for some $\delta \in [0,1]$, $A \in \mathcal{B}$,

$$P(A) = \begin{cases} \mu(A) & \text{if } \mu(A) \leq \delta \\ 0 & \text{otherwise} \end{cases}$$

- $$P(A) = \begin{cases} 1 & \text{if } A = [0,1] \\ 0 & \text{otherwise} \end{cases}$$

- $P(A) = 1$, for all $A \in \mathcal{B}$.

Can you figure out an appropriate σ-algebra for which these functions are probability measures?

For third example: take σ-algebra as $\{\emptyset, [0,1]\}$.

1.2 Additional properties of probability measures

(CB Thms 1.2.8-11) For prob. fxn P and $A, B \in \mathcal{B}$:

- $P(\emptyset) = 0$;
- $P(A) \leq 1$;
- $P(A^c) = 1 - P(A)$.
- $P(B \cap A^c) = P(B) - P(A \cap B)$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$;
- Subadditivity (Boole's inequality): for events $A_i, i \geq 1$,

$$P\left(\cup_{i=1}^{\infty} A_i\right) \leq \sum_{i=1}^{\infty} P(A_i).$$

- Monotonicity: if $A \subset B$, then $P(A) \leq P(B)$

- $P(A) = \sum_{i=1}^{\infty} P(A \cap C_i)$ for any partition $C_1, C_2, \ldots$

By manipulating the above properties, we get

$$\begin{aligned} P(A \cap B) &= P(A) + P(B) - P(A \cup B) \\ &\geq P(A) + P(B) - 1 \end{aligned} \tag{2}$$

which is called the *Bonferroni* bound on the joint event $A \cap B$. (Note: when $P(A)$ and $P(B)$ are small, then bound is < 0, which is trivially correct. Also, bound is always ≤ 1.)

With three events, the above properties imply:

$$P(\cup_{i=1}^{3} A_i) = \sum_{i=1}^{3} P(A_i) - \sum_{i<j}^{3} P(A_i \cap A_j) + P(A_1 \cap A_2 \cap A_3)$$

and with n events, we have

$$\begin{aligned} P(\cup_{i=1}^{n} A_i) = &\sum_{i=1}^{n} P(A_i) - \sum_{1 \leq i < j \leq n}^{n} P(A_i \cap A_j) + \sum_{1 \leq i < j < k \leq n}^{n} P(A_i \cap A_j \cap A_k) + \\ &\cdots + (-1)^{n+1} P(A_1 \cap A_2 \cap \cdots \cap A_n). \end{aligned}$$

This equality, the *inclusion-exclusion* formula, can be used to derive a wide variety of bounds (depending on what is known and unknown).

2 Updating information: conditional probabilities

Consider a given probability space $(\Omega, \mathcal{B}, P)$.

Definition 1.3.2: if $A, B \in \mathcal{B}$, and $P(B) > 0$, then the conditional probability of A given B, denoted $P(A|B)$ is

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

If you interpret $P(A)$ as "the prob. that the outcome of the experiment is in A", then $P(A|B)$ is "the prob. that the outcome is in A, given that you know it is in B".

- If A and B are disjoint, then $P(A|B) = 0/P(B) = 0$.

- If $A \subset B$, then $P(A|B) = P(A)/P(B) < 1$. Here "B is necessary for A".
- If $B \subset A$, then $P(A|B) = P(B)/P(B) = 1$. Here "B implies A"

As CB point out, when you condition on the event B, then B becomes the sample space of a new probability space, for which the $P(\cdot|B)$ is the appropriate probability measure:

$\Omega \to B$
$\mathcal{B} \to \{A \cap B,\ \forall A \in \mathcal{B}\}$
$P(\cdot) \to P(\cdot|B)$

■■■

From manipulating the conditional probability formula, you can get that

$$\begin{aligned} P(A \cap B) &= P(A|B) \cdot P(B) \\ &= P(B|A) \cdot P(A) \\ \Rightarrow P(A|B) &= \frac{P(B|A) \cdot P(A)}{P(B)}. \end{aligned}$$

For a partition of disjoint events $A_1, A_2, \ldots$ of Ω: $P(B) = \sum_{i=1}^{\infty} P(B \cap A_i) = \sum_{i=1}^{\infty} P(B|A_i)P(A_i)$. Hence:

$$P(A_i|B) = \frac{P(B|A_i) \cdot P(A_i)}{\sum_{i=1}^{\infty} P(B|A_i)P(A_i)}$$

which is **Baye's Rule**.

■■■

Example: Let's Make a Deal.

- There are three doors (numbered 1,2,3). Behind one of them, a prize has been randomly placed.
- You (the contestant) bet that prize is behind door 1
- Monty Hall opens door 2, and reveals that there is no prize behind door 2.
- He asks you: "do you want to switch your bet to door 3?"

Informally: MH has revealed that prize is not behind 2. There are two cases: either (a) it is behind 1, or (b) behind 3. In which case is MH's opening door 2 more probable? In case (a), MH could have opened either 2 or 3; in case (b), MH is forced to open 2 (since he cannot open door 1, because you chose that door). MH's opening of door 2 is more probable under case (b), so you should switch. (This is actually a "maximum likelihood" argument.)

More formally, define two random variables D (for door behind which the prize is) and M (denoted the door which Monty opens). Consider a comparison of the conditional probabilities $P(D = 1|M = 2)$ vs. $P(D = 3|M = 2)$. Note that these two sum to 1, so you will switch $D = 3$ if $P(D = 3|M = 2) > 0.5$.

D	M	$Prob$
1	1	0
1	2	$\frac{1}{3} * \frac{1}{2} = \frac{1}{6}$
1	3	$\frac{1}{3} * \frac{1}{2} = \frac{1}{6}$
2	1	0
2	2	0
2	3	$\frac{1}{3} * 1 = \frac{1}{3}$
3	1	0
3	2	$\frac{1}{3} * 1 = \frac{1}{3}$
3	3	0

(Note that Monty will never open door 1, because you bet on door 1.)

Before Monty opens door 2, you believe that the $Pr(D = 3) = \frac{1}{3}$. After Monty opens door 2, you can update to

$$Pr(D = 3|M = 2) = Pr(D = 3, M = 2)/Pr(M = 2) = \frac{1}{3}/(\frac{1}{3} + \frac{1}{6}) = \frac{2}{3}.$$

So you should switch.

3 Independence

Two events $A, B \in \mathcal{B}$ are *statistically independent* iff

$$P(A \cap B) = P(A) \cdot P(B).$$

(Two disjoint events are not independent.)

Independence implies that

$$P(A|B) = P(A), \quad P(B|A) = P(B) :$$

knowing that outcome is in B does not change your perception of the outcome's being in A.

Example: in 2-coin toss: the events "first toss is heads" (HH,HT) and "second toss is heads" (HH,TH) are independent. (Note that independence of two events does not mean that the two events have zero intersection in the sample space.)

Some trivial cases for independence of two events A_1 and A_2: (i) $P(A_1) \leq P(A_2) = 1$; $P(A_1) = 0$.

When there are more than two events (i.e., $A_1, \ldots, A_n$), we use concept of *mutual independence*: $A_1, \ldots, A_n$ are mutually independent iff

$$\text{for any subcollection } A_{i_1}, \ldots, A_{i_k} : \quad P(\cap_{j=1}^{k} A_{i_j}) = \prod_{j=1}^{k} P(A_{i_j}).$$

This is very strong: it is stronger than $P(\cap_{i=1}^{n} A_i) = \prod_{i=1}^{n} P(A_i)$, and also stronger than $P(A_i \cap A_j) = P(A_i)P(A_j)$, $\forall i \neq j$. Indeed, it involves $\sum_{k=2}^{n} \left(\begin{array}{c} n \\ k \end{array} \right) = 2^n - n - 1$ equations (which are the number of subcollections of $A_{i_1}, \ldots, A_{i_k}$).

4 Random variables

A random variable is a *function* from the sample space Ω to the real numbers

Examples: 2-coin toss. Let $t_i = \begin{cases} 1 & \text{if } H \text{ in } i\text{-th toss} \\ 2 & \text{if } T \text{ in } i\text{-th toss} \end{cases}$ for $i = 1, 2$.

1. One RV is $x \equiv t_1 + t_2$.

Ω	x
HH	2
HT	3
TH	3
TT	4

Note that RV need not be one-to-one mapping from Ω to $\mathbb{R}$.

2. Another RV is x equal to the number of heads

Ω	$P(\cdot)$	x	P_x
HH	$\frac{1}{4}$	2	$\frac{1}{4}$
HT	$\frac{1}{4}$	1	$\frac{1}{4}$
TH	$\frac{1}{4}$	1	$\frac{1}{4}$
TT	$\frac{1}{4}$	0	$\frac{1}{4}$

implying

$$x = \begin{cases} 0 & \text{w/prob } \frac{1}{4} \\ 1 & \text{w/prob } \frac{1}{2} \\ 2 & \text{w/prob } \frac{1}{4}. \end{cases}$$

This example illustrates how we use $(\Omega, \mathcal{B}, P_\omega)$, the original probability space, to define (induce) a probability space for a random variable: here $(\{0,1,2\}, \text{all subsets of } \{0,1,2\}, P_x)$.

We were able to do this only because we assumed that the event space for the original experiment was rich enough such that the probabilities $P(HH)$, $P(HT)$, so on, are well-defined. (What is we had assumed that the event space was the trivial σ-field?)

This is the simplest example of a *discrete* random variable: one with a countable range.

4.1 Example: Continuous random variable

For continuous random variables $x : \Omega \to \mathbb{R}$, we define the probability space:

- Sample space is real line $\mathbb{R}$
- Event space is $\mathcal{B}(\mathbb{R})$, the "Borel" σ-algebra on the real line, which is generated by the half-lines $\{(-\infty, a], a \in \mathbb{R}\}$.
- Probability measure P_x defined so that, for $A \in \mathcal{B}(\mathbb{R})$,

  $$P_x(A) = P_\omega(\omega \in \Omega : x(\omega) \in A) \equiv P_\omega(x^{-1}(A)).$$

  Implicit assumption: for all $A \in \mathcal{B}(\mathbb{R})$, $x^{-1}(A) \in \mathcal{B}(\Omega)$. Otherwise, $P_\omega(x^{-1}(A))$ may not be well-defined, since the domain of the $P(\cdot)$ function is $\mathcal{B}(\Omega)$. This is the requirement that the random variable $x(\cdot)$ is "Borel-measurable".

Example: consider $X(\omega) = |\omega|$, with ω from the probability space $([-1,1], \mathcal{B}[-1,1], \mu/2)$. Then the probability space for $X(\cdot)$ is (i) sample space $[0,1]$; (ii) event space $\mathcal{B}[0,1]$, and (iii) probability measure P_x such that

$$P_x(A) = P_\omega(x(\omega) \in A) = P_\omega(\omega : \omega \in A, -\omega \in A) = \mu(A).$$

For example, $P_x([\frac{1}{3}, \frac{2}{3}]) = \mu([\frac{1}{3}, \frac{2}{3}])/2 + \mu([-\frac{2}{3}, -\frac{1}{3}])/2 = \mu([\frac{1}{3}, \frac{2}{3}])$.

4.2 CDF and PDF

For a random variable X on $(\mathbb{R}, \mathcal{B}(\mathbb{R}), P_x)$, we define its *cumulative distribution function* (CDF)

$$F_X(x) \equiv P_x(X \leq x), \text{for all } x.$$

(note that all the sets $X \leq x$ are in $\mathcal{B}(\mathbb{R})$).

For a discrete random variable: step function which is continuous from the right (graph)

For a continuous random variable:

Thm 1.5.3: F(x) is a CDF iff

1. $\lim_{x\to\infty} F(x) = 1$ and $\lim_{x\to-\infty} F(x) = 0$
2. $F(x)$ is nondecreasing
3. $F(x)$ is right-continuous: for every x_0, $\lim_{x\downarrow x_0} F(x) = F(x_0)$

Any random variable X is "tight": For every $\epsilon > 0$ there exists a constant $M < \infty$ such that $P(|X| > M) < \epsilon$. Does not have a probability "mass" at ∞.

■■■

Definition 1.5.8: the random variables X and Y are *identically distributed* if for every set $A \in \mathcal{B}(\mathbb{R})$, $P_X(X \in A) = P_Y(Y \in A)$.

Note that X, Y being identically distributed does not mean than $X = Y$! (Example: 2-coin toss, with X being number of heads and Y being number of tails)

But **Thm 1.5.10**: X and Y are identically distributed $\iff F_X(z) = F_Y(z)$ for every z.

■■■

Definition 1.6.1: Probability mass function (pmf) for a discrete random variable X is

$$f_X(x) \equiv P_X(X = x).$$

Recover from CDF as the distance (on the y-axis) between the "steps".

Definition 1.6.3: Probability density function (pdf) for a continuous random variable X is $f_X(x)$ which satisfies

$$F_X(x) = \int_{-\infty}^{x} f_X(t)dt. \tag{3}$$

Thm 1.6.5 A function $f_x(x)$ is a pmf or pdf iff

- $f_X(x) \geq 0$ for all x
- For discrete RV: $\sum_x f_X(x) = 1$; for continuous RV: $\int_{-\infty}^{\infty} f_X(x)dx = 1$.

■■■

By Eq. (3), and the fundamental theorem of calculus, if $f_X(\cdot)$ is continuous, then $f_X(\cdot) = F'_X(\cdot)$ (i.e., F_X is the anti-derivative of f_X).

■■■

4.3 Conditional CDF/PDF

Random variable $X \sim (\mathbb{R}, \mathcal{B}(\mathbb{R}), P_X)$

What is $Prob(X \leq x | X \in A)$ (conditional CDF)?

Go back to basics: $Prob(X \leq x | X \in A) = \frac{Prob(\{X \leq x\} \cap A)}{Prob(A)}$

This expression can be differentiated to obtain the conditional PDF.

- *Example:* $X \sim U[0,1]$, with conditioning event $X \geq z$.

  Conditional CDF:

  $$Prob(X \leq x | X \geq z) = \begin{cases} 0 & \text{if } x \leq z \\ (x-z)/(1-z) & \text{if } x > z \end{cases}$$

  Hence, the conditional pdf is $1/(1-z)$, for $x > z$.

- *Example:* (truncated wages)

  Wage offers $X \sim U[0,10]$, and $X > 5.25$ (only observe wages when they lie above minimum wage)

  Conditional CDF:

  $$\begin{aligned} F_X(x | X \geq 5.25) &= \frac{Prob(\{X \leq x\} \cap \{X \geq 5.25\})}{Prob(X \geq 5.25)} \\ &= \frac{\frac{1}{10}(x - 5.25)}{\frac{1}{10}4.75} \text{ for } x \in [5.25, 10] \end{aligned}$$

  Hence,

  $$f_X(x | X \geq 5.25) = \frac{1}{4.75} \text{ for } x \in [5.25, 10].$$

5 Lebesgue integral

Consider the measure space $(\mathbb{R}, \mathcal{B}, P)$, where P is any measure (not necessarily Lesbesgue measure). Then we define the *Lebesgue-Stieltjes integral*:

$$\mathbb{E}_P f = \int f dP \equiv \sup_{\{E_i\}} \left\{ \sum_i \left(\inf_{\omega \in E_i} f(\omega)\right) P(E_i) \right\} \tag{4}$$

where the "sup" is taken over all finite partitions $\{E_1, E_2, \ldots\}$ of $\mathbb{R}$. Assign value of $+\infty$ when this "sup" does not exist.

The definition above is not constructive, and typically when one needs to compute a LS integral, one proceeds by converting it into the usual Riemann integral by replacing $dP(x)$ by $p(x)dx$, where $p(x)$ denotes the density function of P (wrt. Lebesgue measure). Then $\int f(x)p(x)dx$ is a Riemann integral which can be computed in the usual way.

5.1 Lebesgue vs. Riemann integral

These partitions can be defined quite generally. Consider a (bounded) function f and a sequence of numbers y_k with $y_1 < y_2 < y_3 < \cdots < y_K$. Define the partition by

$$E_k = \{\omega : \ y_k \leq f(\omega) < y_{k+1}\}, \quad k = 1, \ldots K-1.$$

Letting P be Lesbesgue measure, we see that the Lesbesgue integral for this partition is equal to (roughly) the areas of the rectangles when you "slice" along the y-axis. Indeed this distinction between "slicing the range" vs. "slicing the domain" of the function appears to have been a distinctive feature of his integration approach (vs. Riemann's approach) to Lebesgue himself.[2]

Using Lebesgue integration, one can consider the integral of a wider class of functions than Riemann integration. (By "integrable" here, we are mean that the integral *exists*, ie. is not undefined. Sometimes we make the additional restriction that it is *finite*.) Indeed, we have

Theorem:[3] Let f be a bounded real-valued function on $[a, b]$.
(a) The function f is Riemann-integrable on $[a, b]$ iff f is continuous almost everywhere

[2] It also suggests that the integral of a function should be the same no matter if you "reorder" its domain

[3] Ash, *Measure, Integration, and Functional Analysis*, Academic Press, 1972. Theorem 1.7.1.

on $[a, b]$ (w.r.t Lebesgue measure)
(b) If f is Riemann-integrable on $[a, b]$, then f is integrable w.r.t Lebesgue measure on $[a, b]$, and the two integrals are identical. ■

A well-known example is the *Dirichlet function*:

$$f(\omega) = \begin{cases} 1 & \omega \text{ is rational} \\ 0 & \text{otherwise;} \end{cases} \qquad \omega \in [0, 1].$$

The Lebesgue integral $\int_0^1 f d\mu = 0$, but this function is nowhere continuous on [0,1], and hence not Riemann-integrable.

5.2 Properties of Lesbesgue integral

- Indicator function: $P(B) = \int \mathbb{1}_B(\omega) dP \equiv \int_B dP$
- For disjoint sets A, B: $\int_{A \cup B} dP = \int_A dP + \int_B dP$.
- Scalar multiplication: $\int kf dP = k \int f dP$
- Additivity: $\int (f + g) dP = \int f dP + \int g dP$.
- Monotonicity: if $f \leq g$ P-almost everywhere, then $\int f dP \leq \int g dP$.

5.3 [skip] Convergence results for Lebesgue integrals

Consider a non-negative bounded function $f(\omega)$, and a sequence of partitions $\mathcal{E}^1 \subset \mathcal{E}^2 \subset \mathcal{E}^3$, etc. For a partition $\mathcal{E}^i$, consider the *simple function* defined as

$$\forall \omega \in E_k \in \mathcal{E}^i : \; f^i(\omega) = \inf_{\omega' \in E_k} f(\omega').$$

This function is constant on each element of the partition $\mathcal{E}^i$, and equal to the infimum of the function $f(\omega)$ in each element.

Because the range of each function $f^i(\omega)$ is finite, the Lesbesgue integral $\int f^i dP$ is just a finite sum and exists.[4]

Furthermore, we see that $f^i(\omega) \uparrow f(\omega)$, for P-almost all ω. Intuitively $\int f dP$ should be the "limit" of $\int f^i dP$:

[4]Indeed, the LS integral, as defined in Eq. (4), can be defined as the sup over all simple function dominated by f, i.e., $\sup_{\text{g simple}} \int g dP$ for $g \leq f$, P-everywhere.

Monotone convergence theorem: If $\{f_n\}$ is a *non-decreasing* sequence of measurable *non-negative* functions, with $f_n(\omega) \uparrow f(\omega)$, then

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

(As stated, don't require boundedness of f_n, f: so both LHS and RHS can be ∞.)

For general functions f which may take both positive and negative values, we break it up into the positive $f^+ = \max\{f, 0\}$ and negative $f^- = (f^+ - f)$ parts. Both f^+ and f^- are non-negative functions. We define

$$\int f dP = \int f^+ dP - \int f^- dP$$

and use the Monotone Convergence Theorem for each integral separately.

Additional convergence results for Lebesgue integrals:

- **Fatou's lemma**: for (possibly non-convergent) sequence of *non-negative* functions f_n:

  $$\text{liminf}_{n\to\infty} \int f_n dP \geq \int (\text{liminf}_{n\to\infty} f_n) dP.$$

  (On the RHS, the "liminf" is taken pointwise in ω.) This is for sequences of functions which need not converge.

- **Dominated convergence theorem:** If $f_n(\omega) \to f(\omega)$ for P-almost all ω, and there exists a function $g(\omega)$ such that $|f_n(\omega)| \leq g(\omega)$ for P-almost all ω and for all n, and g is integrable ($\int g dP < \infty$), then $\int f_n dP \to \int f dP$. That is, $Ef_n \to Ef$.

- **Bounded convergence theorem:** If $f_n(\omega) \to f(\omega)$ for P-almost all ω, and there exists constant $B < \infty$ such that $|f_n(\omega)| \leq B$ for P-almost all ω and for all n, then $\int f_n dP \to \int f dP$.

5.4 Converting Lebesgue to Riemann integral

As we remarked above, for computational purposes we usually convert a Lebesgue integral into the usual Riemann integral by replacing $dP(x)$ by $p(x)dx$, where $p(x)$ denotes the density function of P (wrt. Lebesgue measure). Under what conditions on the original probability function $P(x)$ can be do this?

By the *Radon-Nikodym Theorem*, a (necessary and sufficient) condition for the existence of a density function $p(x)$ corresponding to the probability function $P(x)$ is that the probability measure $P(x)$ of the real-valued random variable X be *absolutely continuous with respect to Lebesgue measure*.

Absolutely continuous w.r.t. Lebesgue measure means that all sets in the support of X (which is a part of the real line) which have zero Lebesgue measure must also have zero probability under $P(x)$; i.e., for all $A \in \mathbb{R}$ such that $\mu(A) = 0 \rightarrow P(A) = 0$.

Since only singletons (and countable sets of singletons) have zero Lebesgue measure, this condition essentially rules out random variables which have a "point mass" at some points. Intuitively, this implies jumps in the CDF F