Section 20

Random Variables and Distributions

Po-Ning Chen, Professor

Institute of Communications Engineering

National Chiao Tung University

Hsin Chu, Taiwan 30010, R.O.C.

Random Variables

Definition (random variable) A random variable on a probability space $(\Omega, \mathcal{F}, P)$ is a real-valued function $X = X(\omega)$ (i.e., $X : \Omega \to \Re$) with $\{\omega \in \Omega : X(\omega) \le x\} \in \mathcal{F}$ for every $x \in \Re$.

- The event space $\mathcal{F}$ must be a σ-field. Why? See the next two slides.
- The probability measure P can be treated as a quantitative *mechanism* for the probability of occurrence of events.
- $\{\omega \in \Omega : X(\omega) \le x\}$ must be an event, otherwise we do not know its probability of occurrence.

  Under this definition, the cdf of X is well defined as:

$$\Pr[X \le x] = P\left(\{\omega \in \Omega : X(\omega) \le x\}\right).$$

The Concept of Field/Algebra

Definition (field/algebra) A set $\mathcal{F}$ is said to be a *field* or *algebra* of a *sample space* Ω if it is a *nonempty* collection of subsets of Ω with the following properties:

1. $\emptyset \in \mathcal{F}$ and $\Omega \in \mathcal{F}$;
   - *Interpretation*: A mechanism to determine whether an *outcome* lies in the empty set (impossible) or the sample space (certain).
2. (closure under complement action) $A \in \mathcal{F} \Rightarrow A^c \in \mathcal{F}$;
   - *Interpretation*: "having a mechanism to determine whether an outcome lies in A" is equivalent to "having a mechanism to determine whether an outcome lies in A^c."
3. (closure under finite union) $A \in \mathcal{F}$ and $B \in \mathcal{F} \Rightarrow A \cup B \in \mathcal{F}$.
   - *Interpretation*: If one has a mechanism to determine whether an outcome lies in A, and a mechanism to determine whether an outcome lies in B, then he can surely determine whether the outcome lies in the union of A and B.

- Elements of $\mathcal{F}$ is referred to as *events*.

σ-field/algebra

- To work on a *field* may cause some problems when a person is dealing with "*limit*".

  E.g., $\Omega = \Re$ (the real line) and $\mathcal{F}$ is a collection of all *open*, *semi-open* and *closed* intervals whose two endpoints are rational numbers, including $\Re$ itself.

  Let

  $$A_i = [0, 1.\underbrace{010010001\ldots1}_{i \text{ of them}}).$$

  Then, does the infinite union $\cup_{i=1}^{\infty} A_i$ belong to $\mathcal{F}$? The answer is NO!

- We therefore need an extension definition of field, which is named σ-field.

Definition (σ-field/σ-algebra) A set $\mathcal{F}$ is said to be a *σ-field* or *σ-algebra* of a sample space Ω if it is a nonempty collection of subsets of Ω with the following properties:

1. $\emptyset \in \mathcal{F}$ and $\Omega \in \mathcal{F}$;
2. (closure under complement action) $A \in \mathcal{F} \Rightarrow A^c \in \mathcal{F}$;
3. (closure under countable union) $A_i \in \mathcal{F}$ for $i = 1, 2, 3, \ldots \Rightarrow \bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$.

Probability measure

Definition (probability measure) A set function P on a measurable space $(\Omega, \mathcal{F})$ is a *probability measure*, if it satisfies:

1. $0 \leq P(\mathcal{A}) \leq 1$ for $\mathcal{A} \in \mathcal{F}$;
2. $P(\emptyset) = 0$ and $P(\Omega) = 1$.
3. **(countable additivity)** if $\mathcal{A}_1$, $\mathcal{A}_2$, $\ldots$ is a disjoint sequence of sets in $\mathcal{F}$, then

$$P\left(\bigcup_{k=1}^{\infty} \mathcal{A}_k\right) = \sum_{k=1}^{\infty} P(\mathcal{A}_k).$$

Merit by defining random variables based on $(\Omega, \mathcal{F}, P)$?

Answer 1: $(\Omega, \mathcal{F}, P)$ is what truly occurring internally, but possibly **non-observable**.

- In order to infer what really occurs for this **non-observable** random outcome ω, an experiment that results in **observable values** x that depends on this **non-observable** outcome must be performed.
- So x that takes real values is a function of $\omega \in \Omega$.
- Since ω is random with respect to probability measure P, the probability of the **observation** X is defined over the probability on Ω.
- Some books therefore state that $X : (\Omega, \mathcal{F}, P) \rightarrow (X(\Omega), \mathcal{B}, Q)$ yields an **observation probability space** $(X(\Omega), \mathcal{B}, Q)$, where

  $X(\mathcal{A}) = \{X(\omega) \in \Re : \omega \in \mathcal{A}\}, \mathcal{B} = \{X(\mathcal{A}) : \mathcal{A} \subset \mathcal{F}\}$ and $Q(X(\mathcal{A})) = P(\mathcal{A})$.

Merit by defining random variables based on $(\Omega, \mathcal{F}, P)$?

Example An atom may spin counterclockwisely or clockwisely, which is not directly observable. The original true probability space $(\Omega, \mathcal{F}, P)$ for this atom is

$$\Omega = \{\text{counterclockwise}, \text{clockwise}\},$$
$$\mathcal{F} = \Big\{\emptyset, \{\text{counterclockwise}\}, \{\text{clockwise}\}, \{\text{counterclockwise}, \text{clockwise}\}\Big\},$$

and

$$\begin{cases} P(\emptyset) = 0, \\ P(\{\text{counterclockwise}\}) = 0.4, \\ P(\{\text{clockwise}\}) = 0.6, \\ P(\{\text{counterclockwise}, \text{clockwise}\}) = 1. \end{cases}$$

Now an experiment that uses some advanced facility is performed to examine the spin direction of this atom. (Suppose there is no **observation noise** in this experiment; so a 1-1 correspondence mapping from Ω to $\Re$ can be obtained.) This results in an observable two-value random variable X, namely,

$$X(\text{counterclockwise}) = 1 \quad \text{and} \quad X(\text{clockwise}) = -1.$$

Merit by defining random variables based on $(\Omega, \mathcal{F}, P)$?

Answer 2: $(\Omega, \mathcal{F}, P)$ may be too abstract and lack of the required mathematical structure for manipulation, such as *ordering* (that is required for cdf).

Example

$$\begin{aligned}
\Omega &= \{\blacktriangle, \blacktriangledown, \square, \blacksquare, \lozenge, \blacklozenge\} \\
\mathcal{F} &= \text{A } \sigma\text{-field collection of subsets of } \Omega \\
P &= \text{Some assigned probability measure on } \mathcal{F}
\end{aligned}$$

Define a random variable X on $(\Omega, \mathcal{F}, P)$ as:

$$\begin{aligned}
X(\blacktriangle) &= 1 \\
X(\blacktriangledown) &= 2 \\
X(\square) &= 3 \\
X(\blacksquare) &= 4 \\
X(\lozenge) &= 5 \\
X(\blacklozenge) &= 6
\end{aligned}$$

Merit by defining random variables based on $(\Omega, \mathcal{F}, P)$?

Examine what subsets should be included in $\mathcal{F}$.

$$
\begin{array}{lll}
\text{For } x < 1, & \{\omega \in \Omega : X(\omega) \le x\} & = \emptyset \\
\text{For } 1 \le x < 2, & \{\omega \in \Omega : X(\omega) \le x\} & = \{\blacktriangle\} \\
\text{For } 2 \le x < 3, & \{\omega \in \Omega : X(\omega) \le x\} & = \{\blacktriangle, \blacktriangledown\} \\
\text{For } 3 \le x < 4, & \{\omega \in \Omega : X(\omega) \le x\} & = \{\blacktriangle, \blacktriangledown, \square\} \\
\text{For } 4 \le x < 5, & \{\omega \in \Omega : X(\omega) \le x\} & = \{\blacktriangle, \blacktriangledown, \square, \blacksquare\} \\
\text{For } 5 \le x < 6, & \{\omega \in \Omega : X(\omega) \le x\} & = \{\blacktriangle, \blacktriangledown, \square, \blacksquare, \lozenge\} \\
\text{For } x \ge 6, & \{\omega \in \Omega : X(\omega) \le x\} & = \{\blacktriangle, \blacktriangledown, \square, \blacksquare, \lozenge, \blacklozenge\} = \Omega
\end{array}
$$

By definition, $\mathcal{F}$ must be a σ-field containing the above seven events or sets.

Note that we can **sort** $1, 2, 3, 4, 5, 6$ (to yield the cdf), but we may not be able to sort $\blacktriangle, \blacktriangledown, \square, \blacksquare, \lozenge, \blacklozenge$, not to mention the manipulation of $(\blacktriangle + \blacktriangledown)$ or $(\square - \blacksquare)$.

Merit by defining random variables based on $(\Omega, \mathcal{F}, P)$?

Example of a random variable Y without inverse.

Define a random variable Y on $(\Omega, \mathcal{F}, P)$ as:

$$\begin{aligned} Y(\blacktriangle) = Y(\blacktriangledown) = Y(\square) &= 1 \\ Y(\blacksquare) = Y(\lozenge) = Y(\blacklozenge) &= 2 \end{aligned}$$

Examine what subsets must be included in $\mathcal{F}$.

$$\begin{array}{lll} \text{For } y < 1, & \{\omega \in \Omega : Y(\omega) \le y\} & = \emptyset \\ \text{For } 1 \le y < 2, & \{\omega \in \Omega : Y(\omega) \le y\} & = \{\blacktriangle, \blacktriangledown, \square\} \\ \text{For } y \ge 2, & \{\omega \in \Omega : Y(\omega) \le y\} & = \{\blacktriangle, \blacktriangledown, \square, \blacksquare, \lozenge, \blacklozenge\} = \Omega \end{array}$$

Hence, $\mathcal{F}$ must be a σ-field containing the above three sets for Y.

The third merit by defining random variables based on $(\Omega, \mathcal{F}, P)$ will be deferred until the introduction of the definition of random processes.

Existence of mean

- In mathematics, $(+\infty) + (-\infty)$ is undefined.
- To avoid the occurrence of $(+\infty) + (-\infty)$, Real Analysis first defines the *integration of a non-negative function* $f(x)$, namely, $f(x) \geq 0$ for $x \in \Re$ (either based on Riemann integral or based on Lebesgue integral). See the next two slides for details.
- Then, the integration of a general (possibly negative) function is defined through

$$\int f(x)dx = \int f^+(x)dx - \int f^-(x)dx,$$

where $f^+(x) = \max\{f(x), 0\}$ and $f^-(x) = \max\{-f(x), 0\}$.

- An integral does not exist when $\displaystyle\int f^+(x)dx = \int f^-(x)dx = \infty$.
- Similar rule applies to random variable X. Notably, the expectation of X is indeed an integration $\left(E[X] = \displaystyle\int_{-\infty}^{\infty} x\,dP_X(x) = \int_0^{\infty} x\,dP_X(x) - \int_{-\infty}^{0} (-x)dP_X(x)\right)$.
- This is the reason why people say the mean of a Cauchy distribution does not exist (and is undefined)!

$$\int_0^{\infty} x \left(\frac{1}{\pi}\frac{1}{(1+x^2)}\right) dx = \int_{-\infty}^{0} (-x) \left(\frac{1}{\pi}\frac{1}{(1+x^2)}\right) dx = \infty.$$

Riemann Integral versus Lebesgue Integral

Riemann integral:

Let $s(x)$ represent a step function on $[a, b)$, which is defined as that there exists a partition $a = x_0 < x_1 < \cdots < x_n = b$ such that $s(x)$ is constant during (x_i, x_{i+1}) for $0 \leq i < n$.
If a function $f(x)$ is Riemann integrable,

$$\int_a^b f(x)dx \triangleq \sup_{\left\{s(x)\;:\;s(x)\leq f(x)\right\}} \int_a^b s(x)dx = \inf_{\left\{s(x)\;:\;s(x)\geq f(x)\right\}} \int_a^b s(x)dx.$$

Example of a non-Riemann-integrable function:
$f(x) = 0$ if x is irrational; $f(x) = 1$ if x is rational.
Then

$$\sup_{\left\{s(x)\;:\;s(x)\leq f(x)\right\}} \int_a^b s(x)dx = 0,$$

but

$$\inf_{\left\{s(x)\;:\;s(x)\geq f(x)\right\}} \int_a^b s(x)dx = (b-a).$$

Riemann Integral versus Lebesgue Integral

Lebesgue integral:

Let $t(x)$ represent a **simple** function, which is defined as the linear combination of indicator functions for finitely many, mutually disjoint partitions.

For example, let $\mathcal{U}_1, \ldots, \mathcal{U}_m$ be the mutually-disjoint partitions of the domain $\mathcal{X}$ and $\cup_{i=1}^m \mathcal{U}_i = \mathcal{X}$. The indicator function of $\mathcal{U}_i$ is $\mathbf{1}(x; \mathcal{U}_i) = 1$ if $x \in \mathcal{U}_i$, and 0, otherwise.

Then $t(x) = \sum_{i=1}^m a_i \mathbf{1}(x; \mathcal{U}_i)$ is a simple function (and $\int t(x) = \sum_{i=1}^m a_i \cdot \lambda(\mathcal{U}_i)$, where $\lambda(\cdot)$ is a Lebesgue measure (cf. Slide 20-36).)
If a function $f(x)$ is Lebesgue integrable, then

$$\int_a^b f(x)dx = \sup_{\left\{t(x)\ :\ t(x) \leq f(x)\right\}} \int_a^b t(x)dx = \inf_{\left\{t(x)\ :\ t(x) \geq f(x)\right\}} \int_a^b t(x)dx.$$

The previous example is actually Lebesgue integrable, and its Lebesgue integral is equal to zero.

X^+ versus X^-

- Define $X^+ = \max\{X, 0\}$ and $X^- = \max\{-X, 0\}$.
- Then $X = X^+ - X^-$, and both X^+ and X^- are non-negative random variables (so their expectation or integration over its probability measure can be Riemann-evaluable or Lebesgue-evaluable).
- Define $E[X] = E[X^+] - E[X^-]$.
- Similar definition applies to $E[X^3]$, $E[X^5]$, $E[X^7]$, $E[f(X)]$, etc.

Definition (simple function) A simple function is defined as the linear combination of indicator functions for finitely many, mutually disjoint partitions.

Definition ($\mathcal{F}/\mathcal{B}$-measurable) A real-valued function $f : (\Omega, \mathcal{F}) \to (\Re, \mathcal{B})$ is $\mathcal{F}/\mathcal{B}$-measurable, if $\{\omega \in \Omega : f(\omega) \in \mathcal{H}\} \in \mathcal{F}$ for every $\mathcal{H} \in \mathcal{B}$, where $\mathcal{B}$ is the one-dimensional Borel set.

- Since it is widely adopted that the Borel set is the σ-field for real line $\Re$, $\mathcal{F}/\mathcal{B}$-measurable is sometimes abbreviated as $\mathcal{F}$-measurable (or measurable $\mathcal{F}$).
- A Borel set is one that consists of all (countable) intersections and unions of open, semi-open and closed intervals. In other words, the element in a Borel set can be obtained by repeating countable (or finite) set-theoretic operations starting from intervals. So the Borel set is a σ-field.
- The Borel set is usually large enough for all "practical" purposes. However, it does not contain every subset of the real line! (See the example in the next two slides.)

A set of real numbers outside the Borel set (due to Vitali)

We now consider the Borel set of $[0,1)$, which is obtained by repeating countable (or finite) set-theoretic operations starting from intervals contained in $[0,1)$.

Denote by $\circledcirc$ the operator under $[0,1)$ as:

$$x \circledcirc y = (x+y) \bmod 1 \text{ for two real numbers } x, y \in [0,1)$$

and

$$\mathcal{A} \circledcirc x = \{r \in \Re : r = a \circledcirc x \text{ for some } a \in \mathcal{A}\}.$$

Two real numbers, x and y in $[0,1)$, are said to be equivalent, denoted by $x \sim y$, if $x \circledcirc r = y$ for some rational r in $[0,1)$. Notably, $x \sim y$ and $y \sim z$ imply $x \sim z$, because $x \circledcirc r_1 = y$ and $y \circledcirc r_2 = z$ imply $z = x \circledcirc (r_1 \circledcirc r_2)$.

Form a set that consists of all the equivalent points in $[0,1)$ (*there are countable many elements in this set because rational numbers in* $[0,1)$ *are countable*), and for convenience, name such set an equivalent class.

Let $\mathcal{H}$ be a subset of $[0,1)$, consisting of exactly one point from each equivalent class.

List the countably many sets of $\mathcal{H} \circledcirc r$ for each rational number r.

Claim: $\mathcal{H} \circledcirc r_1$ and $\mathcal{H} \circledcirc r_2$ are disjoint for any rational numbers, $r_1 \neq r_2$.

Proof: If there exists $a \in [0,1)$ satisfying $a \in \mathcal{H} \circledcirc r_1$ and $a \in \mathcal{H} \circledcirc r_2$, then $a \sim h_1$ for some $h_1 \in \mathcal{H}$ and $a \sim h_2$ for some $h_2 \in \mathcal{H}$, since r_1 and r_2 are rational numbers (and $h_1 \neq h_2$ since $a = h_1 \circledcirc r_1 = h_2 \circledcirc r_2$). Hence, h_1 and h_2 belong to the same equivalent class, contradicting to the construction of $\mathcal{H}$. □

As a consequence of the above claim, all the sets in the list of $\mathcal{H} \circledcirc r$ are disjoint.

Observe that $[0,1) = \bigcup_{r\in\mathbb{Q}} \mathcal{H} \circledcirc r$, where $\mathbb{Q}$ is the set of all rational numbers in $[0,1)$, since every point in $[0,1)$ belongs to some equivalent class.

Now if $\mathcal{H}$ is contained in the Borel set $\mathcal{B}[0,1)$ of $[0,1)$, then by forming a probability space $([0,1), \mathcal{B}[0,1), P)$ with $P(\mathcal{A}) = \int_{x\in\mathcal{A}} dx$ for $\mathcal{A} \in \mathcal{B}[0,1)$, we yield:

$$\int_{x\in[0,1)} dx = \sum_{r\in\mathbb{Q}} \int_{x\in\mathcal{H}\circledcirc r} dx. \tag{20.1}$$

Apparently, $\int_{x\in\mathcal{H}} dx = \int_{x\in\mathcal{H}\circledcirc r} dx$ for any rational $r \in [0,1)$. Thus, if $\int_{x\in\mathcal{H}} dx = \lambda > 0$, then (20.1) gives $1 = \sum_{r\in\mathbb{Q}} \lambda = \infty$, and if $\int_{x\in\mathcal{H}} dx = 0$, (20.1) gives $1 = \sum_{r\in\mathbb{Q}} 0 = 0$. Accordingly, $\mathcal{H}$ is not contained in $\mathcal{B}[0,1)$.

Can we form a set $\mathcal{H}$ that consists of exactly one point from each of the uncountably many equivalent classes?

- Give a sequence of set $\{A_\epsilon\}_{\epsilon>0}$, where the union of them is the sample space.
- $\bigcap_{\epsilon>0}\bigcup_{\alpha>\epsilon} A_\alpha$ consists of all the elements that appears in $\{A_\epsilon\}_{\epsilon>0}$ infinitely many times (could be countably or uncountably many).

  I.e., $\omega \in \bigcap_{\epsilon>0}\bigcup_{\alpha>\epsilon} A_\alpha$ implies $\omega \in \bigcup_{\alpha>\epsilon} A_\alpha$ for every $\epsilon > 0$.
  Hence, for any $\epsilon > 0$, there exists $\alpha > \epsilon$ such that this $\omega \in A_\alpha$.
  (Such choices of ϵ as well as α can be of uncountably many.)

- $\bigcup_{\epsilon>0}\bigcap_{\alpha>\epsilon} A_\alpha^c$ consists of all the elements that appears in $\{A_\epsilon\}_{\epsilon>0}$ finitely many times.
- Now if $\{A_\epsilon\}_{\epsilon>0}$ are mutually disjoint, then $\bigcup_{\epsilon>0}\bigcap_{\alpha>\epsilon} A_\alpha^c$ consists of all the elements that appears in $\{A_\epsilon\}_{\epsilon>0}$ exactly one time.

Theorem 13.5 If a real-valued function $f : \Omega \to \Re$ is $\mathcal{F}$-measurable, then there exists a sequence of $\mathcal{F}$-measurable *simple functions* $f_n : \Omega \to \Re$ such that

$$\begin{cases} 0 \le f_n(\omega) \uparrow f(\omega) & \text{for those } \omega \text{ satisfying } f(\omega) \ge 0; \\ 0 \ge f_n(\omega) \downarrow f(\omega) & \text{for those } \omega \text{ satisfying } f(\omega) \le 0. \end{cases}$$

Proof: Define mutually disjoint sets as:

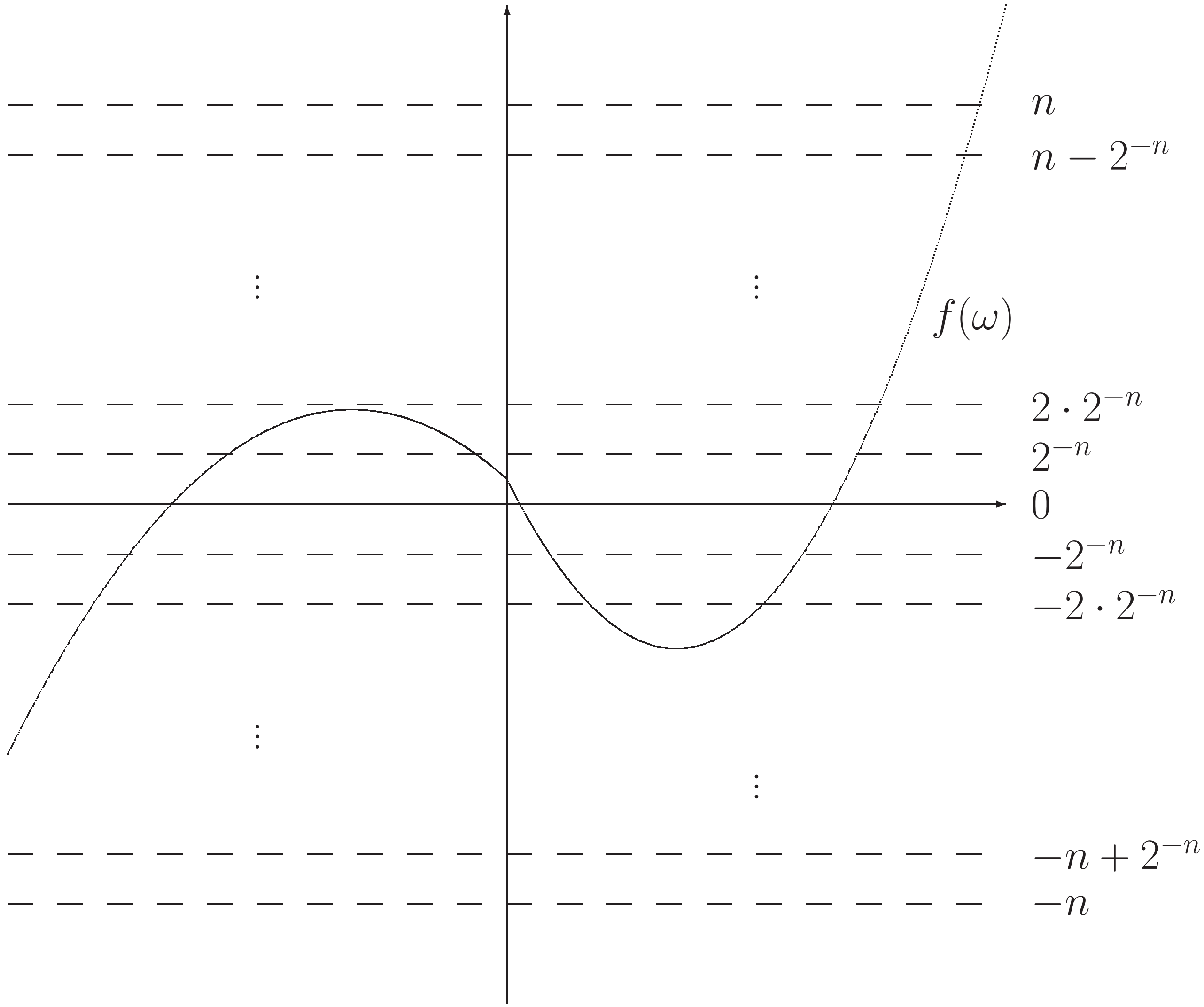
$$\mathcal{A}_k = \begin{cases} \{\omega \in \Omega : n \le f(\omega)\}, & \text{if } k = n2^n + 1; \\ \{\omega \in \Omega : (k-1)2^{-n} \le f(\omega) < k2^{-n}\}, & \text{if } -n2^n + 1 \le k \le n2^n; \\ \{\omega \in \Omega : f(\omega) < -n\}, & \text{if } k = -n2^n. \end{cases}$$

Then

$$f_n(\omega) = \begin{cases} n, & \text{if } \omega \in A_{n2^n+1}; \\ (k-1)2^{-n}, & \text{if } \omega \in A_k \text{ for } 1 \le k \le n2^n; \\ k2^{-n}, & \text{if } \omega \in A_k \text{ for } -n2^n + 1 \le k \le 0; \\ -n, & \text{if } \omega \in A_{-n2^n} \end{cases}$$

is one of the required function sequences. □

Limits of Random Variables



Definition (simple random variables) A simple random variable takes only finitely many values.

Theorem 13.5a If a random variable $X : \Omega \to \Re$ is $\mathcal{F}$-measurable under probability space $(\Omega, \mathcal{F}, P)$, then there exists a sequence of $\mathcal{F}$-measurable *simple random variables* $X_n : \Omega \to \Re$ such that $X_n^+ \uparrow X^+$ and $X_n^- \uparrow X^-$.

$$\left(\text{Equivalently, } \begin{cases} 0 \le X_n(\omega) \uparrow X(\omega) & \text{for those } \omega \text{ satisfying } X(\omega) \ge 0; \\ 0 \ge X_n(\omega) \downarrow X(\omega) & \text{for those } \omega \text{ satisfying } X(\omega) \le 0. \end{cases} \right)$$

Proof: Defining $X_n = f_n$ according to $X = f$ in Theorem 13.5 satisfies the need. (Since Ω may be an abstract and non-real-valued set, we cannot partition Ω using "$\geq$" or "$\leq$". This is another reason why we have to partition over $X(\omega) \in \Re$.) □

- This theorem provides a merit of generalizing theorems that originally apply to simple random variables.
- Notably, a simple random variable only takes finitely many values; hence, all moments of a simple random variable are bounded (and hence, exist). Therefore, all moments of a simple random variable can be determined by taking the derivatives of its moment generating function.

Definition (random vectors) A random vector on a probability space $(\Omega, \mathcal{F}, P)$ is a real-valued function $X : \Omega \to \Re^k$ with $\{\omega \in \Omega : X(\omega) = x^k\} \in \mathcal{F}$.

- A random vector is a finite collection of random variables. In fact, each dimension of $X(\omega) = (X_1(\omega), X_2(\omega), \ldots, X_k(\omega))$ is itself a random variable.
- Hence, an equivalent definition of random vectors is:

  Definition (random vectors) A random vector is a *finite* collection of random variables, each of which is defined on *the same* probability space.

- Another equivalent definition that can be seen in the literature is:

  Definition (random vectors) A random vector is an indexed family of random variables $\{X_i, i \in \mathcal{I}\}$, in which each X_i is defined on the same probability space, and the index set $\mathcal{I}$ is finite.

- **Why requiring each X_i to be defined on the same probability space?**

  Because, it is through "*the same probability space*" that the joint distribution of two (or three, four,..., etc) random variables can be well-defined.

$$\begin{aligned}&\Pr[X_i \le x_i \text{ and } X_j \le x_j]\\ &\quad = P\left(\{\omega \in \Omega : X_i(\omega) \le x_i \text{ and } X_j(\omega) \le x_j\}\right)\\ &\quad = P\left(\{\omega \in \Omega : X_i(\omega) \le x_i\} \cap \{\omega \in \Omega : X_j(\omega) \le x_j\}\right).\end{aligned}$$

  Then, it can be proved that:

$$\begin{aligned}A_i \triangleq \{\omega \in \Omega : X_i(\omega) \le x_i\} &\in \mathcal{F} && \text{because } X_i \text{ defined over } (\Omega, \mathcal{F}, P)\\ A_j \triangleq \{\omega \in \Omega : X_j(\omega) \le x_j\} &\in \mathcal{F} && \text{because } X_j \text{ defined over } (\Omega, \mathcal{F}, P)\\ A_i^c &\in \mathcal{F} && \mathcal{F} \text{ closure under complement action}\\ A_j^c &\in \mathcal{F} && \mathcal{F} \text{ closure under complement action}\\ A_i^c \cup A_j^c &\in \mathcal{F} && \mathcal{F} \text{ closure under countable union}\\ (A_i^c \cup A_j^c)^c = A_i \cap A_j &\in \mathcal{F} && \mathcal{F} \text{ closure under complement action}\end{aligned}$$

  Hence, $P(A_i \cap A_j)$ is probabilistically measurable (for any x_i and x_j).

It can be proved from closures under complement action and countable union that $\mathcal{F}$ is closure under countable intersection.

Example

$$\begin{aligned}\Omega &= \{\blacktriangle, \blacktriangledown, \square, \blacksquare, \lozenge, \blacklozenge\}\\ \mathcal{F} &= \text{A } \sigma\text{-field collection of subsets of } \Omega\\ P &= \text{A probability measure on } \mathcal{F}\end{aligned}$$

Define a random vector $\{X_i, i \in \{1,2\}\}$ as:

$$\begin{array}{llll} X_1(\blacktriangle) = 1; & X_1(\blacksquare) = 2; & X_2(\blacktriangle) = 1; & X_2(\blacksquare) = 2\\ X_1(\blacktriangledown) = 2; & X_1(\lozenge) = 1; & X_2(\blacktriangledown) = 1; & X_2(\lozenge) = 2\\ X_1(\square) = 1; & X_1(\blacklozenge) = 2; & X_2(\square) = 1; & X_2(\blacklozenge) = 2 \end{array}$$

Examine what subsets should be included in $\mathcal{F}$.

$$\begin{array}{lll} \text{For } x_1 < 1, & \{\omega \in \Omega : X_1(\omega) \le x_1\} & = \emptyset\\ \text{For } 1 \le x_1 < 2, & \{\omega \in \Omega : X_1(\omega) \le x_1\} & = \{\blacktriangle, \square, \lozenge\}\\ \text{For } x_1 \ge 2, & \{\omega \in \Omega : X_1(\omega) \le x_1\} & = \{\blacktriangle, \blacktriangledown, \square, \blacksquare, \lozenge, \blacklozenge\} = \Omega\\ \text{For } x_2 < 1, & \{\omega \in \Omega : X_2(\omega) \le x_2\} & = \emptyset\\ \text{For } 1 \le x_2 < 2, & \{\omega \in \Omega : X_2(\omega) \le x_2\} & = \{\blacktriangle, \blacktriangledown, \square\}\\ \text{For } x_2 \ge 2, & \{\omega \in \Omega : X_2(\omega) \le x_2\} & = \{\blacktriangle, \blacktriangledown, \square, \blacksquare, \lozenge, \blacklozenge\} = \Omega \end{array}$$

Hence, $\mathcal{F}$ must be a σ-field containing the above six sets.

We can further extend the random vector to a possibly *(uncountably) infinite* collection of random variables all defined on the same probability space.

Definition (random process) A random process is an indexed family of random variables $\{X_t, t \in \mathcal{I}\}$, in which each X_t is defined on the same probability space.

- Under such definition, all finite dimensional distributions are well-defined because

$$
\begin{aligned}
&\left[X_{t_1} \leq x_1 \text{ and } X_{t_2} \leq x_2 \text{ and } \cdots \text{ and } X_{t_k} \leq x_k\right] \\
&= \left\{\omega \in \Omega : X_{t_1}(\omega) \leq x_1 \text{ and } X_{t_2}(\omega) \leq x_2 \text{ and } \cdots \text{ and } X_{t_k}(\omega) \leq x_k\right\} \\
&= \bigcap_{i=1}^{k} \{\omega \in \Omega : X_{t_i}(\omega) \leq x_i\}
\end{aligned}
$$

  is surely an event by the σ-field properties, and hence, is probabilistically measurable.

- Third merit by defining random processes based on $(\Omega, \mathcal{F}, P)$:
  - All finite dimensional joint distributions are well-defined without the tedious process of listing all of them.
- **The converse however is not true**, i.e., it is not necessarily valid that the statistical properties of a real random process are completely determined by providing all finite-dimensional distributions for samples.
  - See the counterexample in the next slide.

Example Define random processes $\{X_t, t \in [0,1)\}$ and $\{Y_t, t \in [0,1)\}$ as

$$X_t(\omega) = \begin{cases} 1, & \omega \neq t; \\ 0, & \omega = t, \end{cases} \quad \text{and} \quad Y_t(\omega) = 1,$$

where $\omega \in \Omega = [0,1)$. Let $P(A) = \int_A d\alpha$ for any $A \in \mathcal{F}$. Then,

$$\Pr\left[\min_{t\in[0,1)} X_t < 1\right] \;=\; P\left(\left\{\omega \in \Omega : \min_{t\in[0,1)} X_t(\omega) < 1\right\}\right) = P(\Omega) = 1,$$

but

$$\Pr\left[\min_{t\in[0,1)} Y_t < 1\right] \;=\; P\left(\left\{\omega \in \Omega : \min_{t\in[0,1)} Y_t(\omega) < 1\right\}\right) = P(\emptyset) = 0.$$

Thus, X_t and Y_t have **different** statistical properties; however, X_t and Y_t have **exactly the same** multi-dimensional distribution for any samples at $t_1, t_2, \ldots, t_k$ and any k:

$$\begin{aligned}
&\Pr[X_{t_1} \le x_1 \text{ and } X_{t_2} \le x_2 \text{ and } \cdots \text{ and } X_{t_k} \le x_k] \\
&\quad = P\left(\bigcap_{i=1}^{k} \{\omega \in \Omega : X_{t_i}(\omega) \le x_i\}\right) = \begin{cases} 1, & \min_{1\le i\le k} x_i \ge 1; \\ 0, & \text{otherwise} \end{cases} \\
&\quad = \Pr[Y_{t_1} \le x_1 \text{ and } Y_{t_2} \le x_2 \text{ and } \cdots \text{ and } Y_{t_k} \le x_k].
\end{aligned}$$

σ-field generated by a random variable

We are more interested in the **associated sub-event space** of X, which is named the **σ-field generated by random variable** X and is denoted by $\sigma(X)$.

(*Remember that the event space has to be a σ-field; otherwise, the "limit" may not be probabilistically measured.*)

Note that those events that are not "probabilistically measurable" are of no interest through the random experiment X.

Definition (σ-field generated by X) The σ-field generated by X is the smallest σ-field with respect to which it is probabilistically measurable.

Theorem 20.1 Let $\boldsymbol{X} = (X_1, \ldots, X_k)$ be a random vector.

1. The σ-field $\sigma(\boldsymbol{X}) = \sigma(X_1, \ldots, X_k)$ consists exactly of all the sets $\{\omega \in \Omega : \boldsymbol{X}(\omega) \in \mathcal{H}\}$ for $\mathcal{H} \subset \mathcal{B}^k$.

2. A random variable Y is $\sigma(\boldsymbol{X})$-measurable if, and only if, there exists a $\mathcal{B}^k/\mathcal{B}$-measurable function $f : \Re^k \to \Re$ (cf. Slides 20-14 and 20-30 for the definition of $\mathcal{B}^k/\mathcal{B}$-measurable function) such that $Y(\omega) = f(X_1(\omega), \ldots, X_k(\omega))$ for all $\omega \in \Omega$, where $\mathcal{B}^k$ and $\mathcal{B}$ are k-dimensional and 1-dimensional Borel sets, respectively.

σ-field generated by a random variable

Example

1. Define a real-valued function $X(a) = X(b) = 1$ and $X(c) = -1$. Is X a random variable defined on $(\Omega, \mathcal{F}, P)$, where

   - $\Omega = \{a, b, c\}$,
   - $\mathcal{F} = \Big\{\emptyset, \{a\}, \{b, c\}, \{a, b, c\}\Big\}$
   - $P(\emptyset) = 1 - P(\{a, b, c\}) = 0$, $P(\{a\}) = 0.4$ and $P(\{b, c\}) = 0.6$?

Answer: No, because $\{\omega \in \{a, b, c\} : X(\omega) = 1\} = \{a, b\}$ is not an event.

2. Is X a random variable defined on $(\Omega, \mathcal{F}, P)$, where

   - $\Omega = \{a, b, c\}$,
   - $\mathcal{F} = \Big\{\emptyset, \{a\}, \{b\}, \{c\}, \{a, b\}, \{b, c\}, \{c, a\}, \{a, b, c\}\Big\}$
   - $P(\emptyset) = 1 - P(\{a, b, c\}) = 0$, $P(\{a\}) = P(\{b\}) = P(\{c\}) = 1/3$ and $P(\{a, b\}) = P(\{b, c\}) = P(\{c, a\}) = 2/3$?

Answer: Yes, since $\{\omega \in \{a, b, c\} : X(\omega) \in \mathcal{H}\} \in \mathcal{F}$ for every $\mathcal{H} \subset \mathcal{B}$.

σ-field generated by a random variable

An easier way to validate that X is a random variable defined on $(\Omega, \mathcal{F}, P)$ is to examine $\{\omega \in \{a, b, c\} : X(\omega) \le x\} \in \mathcal{F}$ for every $x \in \Re$, namely, to examine whether the cdf of X exists or not.

$$\{\omega \in \{a, b, c\} : X(\omega) \le x\} = \begin{cases} \emptyset, & \text{if } x < -1; \\ \{c\}, & \text{if } -1 \le x < 1; \\ \{a, b, c\}, & \text{if } x \ge 1. \end{cases}$$

3. What is $\sigma(X)$ in Problem 2?

Answer from Theorem 20.1: Any $\mathcal{H}$ containing neither 1 nor -1 gives $\{\omega \in \Omega : \boldsymbol{X}(\omega) \in \mathcal{H}\} = \emptyset$. Any $\mathcal{H}$ containing 1 but not -1 gives $\{\omega \in \Omega : \boldsymbol{X}(\omega) \in \mathcal{H}\} = \{a, b\}$. Any $\mathcal{H}$ contains -1 but not 1 gives $\{\omega \in \Omega : \boldsymbol{X}(\omega) \in \mathcal{H}\} = \{c\}$. Any $\mathcal{H}$ containing both 1 and -1 gives $\{\omega \in \Omega : \boldsymbol{X}(\omega) \in \mathcal{H}\} = \{a, b, c\}$.

Alternative answer: A set consists of all complements, intersections and unions of $\{\omega \in \{a, b, c\} : X(\omega) \le x\}$, i.e.,

$$\Big\{\emptyset, \{c\}, \{a, b\}, \{a, b, c\}\Big\},$$

since $\{a, b\} = \{c\}^c$.

σ-field generated by a random variable

Definition ($\mathcal{F}/\mathcal{G}$-measurable) A function $f : (\Omega, \mathcal{F}) \to (\Theta, \mathcal{G})$ is $\mathcal{F}/\mathcal{G}$-measurable, if $\{\omega \in \Omega : f(\omega) \in G\} \in \mathcal{F}$ for every $G \in \mathcal{G}$.

Definition ($\mathcal{B}^k/\mathcal{B}$-measurable) A function $f : (\Re^k, \mathcal{B}^k) \to (\Re, \mathcal{B})$ is $\mathcal{B}^k/\mathcal{B}$-measurable, if $\{x^k \in \Re^k : f(x^k) \in G\} \in \mathcal{B}^k$ for every $G \in \mathcal{B}$.

Proof of Theorem 20.1

1. Any set of the form $\mathcal{A}_{\mathcal{H}} = \{\omega \in \Omega : \boldsymbol{X}(\omega) \in \mathcal{H}\}$ must be an element of $\sigma(\boldsymbol{X})$. Hence, $\mathcal{S} \triangleq \{B \subset \Omega : B = \mathcal{A}_{\mathcal{H}} \text{ for some } \mathcal{H} \in \mathcal{B}^k\} \subset \sigma(\boldsymbol{X})$.

   By definition, $\sigma(\boldsymbol{X})$ is the smallest σ-field with respect to which X is probabilistically measurable, and $\mathcal{S}$ is apparently a σ-field for which X is probabilistically measurable. Thus, $\mathcal{S} = \sigma(\boldsymbol{X})$.

   (a) $\emptyset \in \mathcal{S}$ by taking $\mathcal{H} = \emptyset$; $\Omega \in \mathcal{S}$ by taking $\mathcal{H} = \Re^k$.

   (b) $\mathcal{A}_{\mathcal{H}} \in \mathcal{S} \Rightarrow \mathcal{A}_{\mathcal{H}}^c = \mathcal{A}_{\mathcal{H}^c} \in \mathcal{S}$.

   (c) $\bigcup_{i=1}^{\infty} \mathcal{A}_{\mathcal{H}_i} = \mathcal{A}_{\cup_{i=1}^{\infty} \mathcal{H}_i} \in \mathcal{S}$.

2. If there exists such $\mathcal{B}^k/\mathcal{B}$-measurable f, then for every $G \in \mathcal{B}$, we have:

$$\mathcal{H} \triangleq \{x^k \in \Re^k : f(x^k) \in G\} \in \mathcal{B}^k,$$

and

$$\begin{aligned}\{\omega \in \Omega : Y(\omega) \in G\} &= \{\omega \in \Omega : f(X_1(\omega), \ldots, X_k(\omega)) \in G\}\\ &= \{\omega \in \Omega : (X_1(\omega), \ldots, X_k(\omega)) \in \mathcal{H}\}\\ &= \{\omega \in \Omega : \boldsymbol{X}(\omega) \in \mathcal{H}\}\\ &\in \sigma(\boldsymbol{X}).\end{aligned}$$

To prove the necessity, suppose Y is a simple random variable, and is $\sigma(\boldsymbol{X})$-measurable. Let $y_1, y_2, \ldots, y_m$ be all the distinct values that Y can take.

Then by the first part of Theorem 20.1, $\mathcal{A}_i = \{\omega \in \Omega : Y(\omega) = y_i\} \in \sigma(\boldsymbol{X}) = \mathcal{S}$ implies that $\mathcal{A}_i = \{\omega \in \Omega : \boldsymbol{X}(\omega) \in \mathcal{H}_i\}$ for some $\mathcal{H}_i \in \mathcal{B}^k$.

Define $f(x^k) \triangleq \sum_{i=1}^m y_i \, \mathbf{1}(x^k; \mathcal{H}_i)$, where $\mathbf{1}(x^k; \mathcal{H}_i)$ equals one if $x^k \in \mathcal{H}_i$, and zero, otherwise. Apparently, $\{x^k \in \Re^k : f(x^k) \in G\}$ for a given $G \in \mathcal{B}$ can be formed by finite unions of $\{\mathcal{H}_i\}$, and hence, is contained in $\mathcal{B}^k$ for every $G \in \mathcal{B}$, which indicates f is $\mathcal{B}^k/\mathcal{B}$-measurable.

Since $\{\mathcal{A}_i\}_{i=1}^m$ are disjoint, no $\boldsymbol{X}(\omega)$ lies in more than one $\mathcal{H}_i$. Accordingly, $f(\boldsymbol{X}(\omega)) = Y(\omega)\big(= y_i$ if $\boldsymbol{X}(\omega) \in \mathcal{H}_i\big)$.

σ-field generated by a random variable

Theorem 13.5a If a random variable $X : \Omega \to \Re$ is $\mathcal{F}$-measurable under probability space $(\Omega, \mathcal{F}, P)$, then there exists a sequence of $\mathcal{F}$-measurable *simple random variables* $X_n : \Omega \to \Re$ such that $X_n^+ \uparrow X^+$ and $X_n^- \uparrow X^-$.

Now suppose Y is not necessary simple, but is still $\sigma(\boldsymbol{X})$-measurable. Then by Theorem 13.5a, there exists a sequence of simple random variables Y_n such that $Y_n(\omega) \to Y(\omega)$ for every $\omega \in \Omega$.

The previous proof shows that there exists $\mathcal{B}^k/\mathcal{B}$-measurable f_n such that $Y_n(\omega) = f_n(\boldsymbol{X}(\omega))$ for all $\omega \in \Omega$.

Define $f(x^k) = \limsup_{n\to\infty} f_n(x^k)$. Then f is also $\mathcal{B}^k/\mathcal{B}$-measurable.

The limit of $\mathcal{B}^k/\mathcal{B}$-measurable functions is also $\mathcal{B}^k/\mathcal{B}$-measurable, since we can surely form the "limit set" by countably many set-theoretical operations, and $\mathcal{B}^k$ and $\mathcal{B}$ are σ-fields. (Cf. Theorem 13.4)

$$\limsup_{n\to\infty} f_n(x^k) = \lim_{n\to\infty} \sup_{q\geq n} f_q(x^k) = \inf_{n\geq 1} \sup_{q\geq n} f_q(x^k).$$

As a result, $f(\boldsymbol{X}(\omega)) = \limsup_{n\to\infty} f_n(\boldsymbol{X}(\omega)) = \limsup_{n\to\infty} Y_n(\omega) = Y(\omega)$. $\square$

Distributions

In this discussion, we drop the mathematical notion of line measure μ that is used in Billingsley's book, and focus more on the engineering notion of "distributions".

Definition (distribution function) The distribution function of a random variable X is defined as $F_X(x) \triangleq \Pr[X \leq x]$.

Properties of $F_X(\cdot)$

- non-decreasing
- right-continuous
- the number of discontinuous points is countable

  Since $F_X(\infty) = 1$ and $F_X(-\infty) = 0$, the number of "jumps" that exceeds $1/2$ is at most 2 (index them by 1 and 2); the number of "jumps" that exceeds $1/3$ but are less than $1/2$ is at most 3 (index them by 3, 4 and 5); the number of "jumps" that exceeds $1/4$ but are less than $1/3$ is at most 4 (index them by 6, 7, 8 and 9); $\cdots$. So we can index these discontinuous points countably.

Theorem 14.1 If a function $F(\cdot)$ is non-decreasing, right-continuous and satisfies $\lim_{x \downarrow -\infty} F(x) = 0$ and $\lim_{x \uparrow \infty} F(x) = 1$, then there exists a random variable and a probability space such that the cdf of the random variable defined over the probability space is equal to $F(\cdot)$.

Distributions

Theorem 14.1 releases us with the burden of referring to a probability space before our defining a random variable. We can indeed define a random variable X directly by its cdf, i.e., $\Pr[X \le x]$. Nevertheless, it is better to keep in mind (and learn) that a formal mathematical notion of random variables is defined over some probability space.

Notably, Theorem 14.1 only proves the "existence" but not the "uniqueness".

Definition (support) The support of a random variable X is a Borel set $\mathcal{H}$ for which $\Pr[X \in \mathcal{H}] = 1$.

- Since we can only well-define the probability of an event in Borel set, the support of X must be a Borel set (cf. Slide 20-36).

Definition (discrete random variables) If the support of a random variable X is discrete, then X is called a discrete random variable.

Examples of discrete random variables

1. Binomial distribution: $\Pr[X = r] = \binom{n}{r} p^r (1-p)^{n-r}$ for $r = 0, 1, \ldots, n$.

Example of the probability space which X is defined on.

$$\begin{cases} \Omega = \{0, 1, 2, \ldots, n\}, \\ \mathcal{F} = 2^{\Omega}, \\ P(\mathcal{A}) = \sum_{i \in \mathcal{A}} P[X = i], \\ X(\omega) = \omega. \end{cases}$$

2. Poisson distribution: $\Pr[X = r] = e^{-\lambda} \frac{\lambda^r}{r!}$ for $r = 0, 1, 2, \ldots$.

How about the support of X cannot be made discrete? Then the probability density function of X may exist.

Probability density function

Definition (Lebesgue measure) A Lebesgue measure λ over the Borel set $\mathcal{B}$ is that for any $\mathcal{A} \in \mathcal{B}$,

$$\lambda(\mathcal{A}) = \sum_{i=1}^{\infty} \lambda(\mathcal{I}_i),$$

and $\{\mathcal{I}_i\}_{i=1}^{\infty}$ are disjoint intervals satisfying $\mathcal{A} = \cup_{i=1}^{\infty}\mathcal{I}_i$, and $\lambda(\mathcal{I})$ is equal to the right-margin of interval $\mathcal{I}$ minus the left-margin of the same interval.

Definition (probability density function) A random variable and its distribution (cdf) have *density* f with respect to Lebesgue measure, if f is $\mathcal{B}$-measurable (i.e., $f : (\Re, \mathcal{B}) \mapsto (\Re, \mathcal{B})$) non-negative function that satisfies

$$\Pr[X \in \mathcal{A}] = \int_{\mathcal{A}} f(x)\,\lambda(dx) = \int_{\mathcal{A}} f(x)dx \quad \text{for every } \mathcal{A} \in \mathcal{B}.$$

Proposition (uniqueness of density) If $\lambda(\{x \in \Re : f(x) \neq g(x)\}) = 0$, then both f and g can be the density of the same random variable.

- It may not be easy to examine the existence of density by means of the above definition. So we need an equivalent examinable condition.

Probability density function

Theorem Let $F_X(x) = \Pr[X \le x]$. Then

$$\Pr[X \in \mathcal{A}] = \int_{\mathcal{A}} f(x)\,\lambda(dx) = \int_{\mathcal{A}} f(x)dx \quad \text{for every } \mathcal{A} \in \mathcal{B}.$$

if, and only if,

$$F_X(b) - F_X(a) = \int_a^b f(x)dx \quad \text{for every } a, b \in \Re.$$

- $F_X(x)$ is differentiable for every $x \in \Re$ except for a set of Lebesgue measure zero (cf. Slide 20-36).
- Note that $F_X(x)$ is not necessarily differentiable on every x for such $f(\cdot)$ to exist.
- If such f exists, then $F'_X(x) = f(x)$ for every $x \in \Re$ except for a set of Lebesgue measure zero.
- If such f exists, and is continuous, then $F'_X(x) = f(x)$ for every $x \in \Re$.

Examples of (continuous) pdf

1. **Exponential distribution:** The probability density function (pdf) of exponential distribution with parameter $\alpha > 0$ is:

$$f(x) = \begin{cases} \alpha e^{-\alpha x}, & \text{if } x \geq 0; \\ 0, & \text{otherwise.} \end{cases}$$

Its cdf is then equal to:

$$F(x) = \int_{-\infty}^{x} f(y)dy = \begin{cases} 1 - e^{-\alpha x}, & \text{if } x \geq 0; \\ 0, & \text{otherwise.} \end{cases}$$

2. **Normal distribution:** The pdf of normal distribution with parameters m and $\sigma > 0$ is:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-m)^2/(2\sigma^2)} \text{ for } x \in \Re.$$

No close-form formula exists for its cdf.

3. **Standard normal distribution:** Normal distribution with $m = 0$ and $\sigma = 1$.

Examples of (continuous) pdf

4. Uniform distribution A uniform distribution has pdf equal to:

$$f(x) = \begin{cases} \dfrac{1}{b-a}, & \text{if } a \leq x < b; \\ 0, & \text{otherwise.} \end{cases}$$

Its cdf is then equal to:

$$F(x) = \int_{-\infty}^{x} f(y)dy = \begin{cases} 0, & \text{if } x < a; \\ \dfrac{x-a}{b-a}, & \text{if } a \leq x < b; \\ 1, & \text{otherwise.} \end{cases}$$

A continuous cdf without pdf

It is easy to construct a random variable whose support cannot be made discrete.

But a more interesting case would be the random variable having continuous cdf but no pdf!

In other words, $F(b) - F(a) > 0$ for $b > a$ for every a and b (in the smallest closure of the set that gathers unit probability mass), but there exists no $f(\cdot)$ such that $F(b) - F(a) = \int_a^b f(x)dx$.

Example Let $X = \sum_{n=1}^{\infty} B_n 2^{-n}$, where $\{B_i\}_{i=1}^{\infty}$ are i.i.d. with $\Pr[B_n = 0] = p$ and $\Pr[B_n = 1] = 1 - p$ and $p \in (0, 1/2)$. Since X is a binary representation of a number in $[0, 1)$, the support of X is equal to $[0, 1)$.
For any $x = .b_1 b_2 b_3 \ldots$, where $b_j \in \{0, 1\}$,

$$\begin{aligned}
F_X(x) &= \Pr[X \leq x] \\
&= \Pr\left[(B_1 < b_1) \vee (B_1 = b_1 \wedge B_2 < b_2) \vee (B_1 = b_1 \wedge B_2 = b_2 \wedge B_3 < b_3) \vee \cdots\right] \\
&= \Pr[B_1 < b_1] + \Pr[B_1 = b_1 \wedge B_2 < b_2] + \Pr[B_1 = b_1 \wedge B_2 = b_2 \wedge B_3 < b_3] + \cdots \\
&= \Pr[B_1 < b_1] + \Pr[B_1 = b_1] \Pr[B_2 < b_2] + \Pr[B_1 = b_1] \Pr[B_2 = b_2] \Pr[B_3 < b_3] + \cdots \\
&= b_1 p + [(1-2p)b_1 + p] b_2 p + [(1-2p)b_1 + p][(1-2p)b_2 + p] b_3 p + \cdots \\
&= \sum_{k=1}^{\infty} b_k p \left(\prod_{\ell=1}^{k-1} [(1-2p)b_\ell + p] \right), \text{ where we assume } \prod_{\ell=1}^{0} [(1-2p)b_\ell + p] = 1.
\end{aligned}$$

A continuous cdf without pdf

For any $\hat{x} > x = .b_1 b_2 b_3 \cdots$, let j be the smallest integer such that $\bar{b}_j = 1$, where $.\bar{b}_1\bar{b}_2\bar{b}_3\cdots$ is the binary representation of $(\hat{x} - x)$. Denote i the largest integer such that $b_i = 0$ among $\{b_0, b_1, b_2, \cdots, b_j\}$, where $b_0 = 1$. Put $x + 2^{-j} = .\tilde{b}_1\tilde{b}_2\tilde{b}_3\cdots$.

Note that $b_i b_{i+1}\cdots b_j = \begin{cases} 011\cdots 11, & i < j; \\ 0, & i = j, \end{cases}$ $\tilde{b}_i\tilde{b}_{i+1}\cdots\tilde{b}_j = \begin{cases} 100\cdots 00, & i < j; \\ 1, & i = j, \end{cases}$

$b_k = \tilde{b}_k$ for $k < i$ and $k > j$, and $\hat{x} \geq x + 2^{-j} > x$.

Example (Assume binary number system.)
$x = 0.11$ and $\hat{x} = 0.11101 \Rightarrow (\hat{x} - x) = 0.00101 = .\bar{b}_1\bar{b}_2\bar{b}_3\cdots \Rightarrow j = 3$
$\{b_0, b_1, b_2, \cdots, b_j\} = \{b_0, b_1, b_2, b_3\} = \{1, 1, 1, 0\} \Rightarrow i = 3$
$x + 2^{-j} = .\tilde{b}_1\tilde{b}_2\tilde{b}_3\cdots = 0.11 + 2^{-3} = 0.111$
$\hat{x} = 0.11101 \geq x + 2^{-j} = 0.111 > x = 0.11$ and $b_i b_{i+1}\cdots b_j = b_3 = 0$, $\tilde{b}_i\tilde{b}_{i+1}\cdots\tilde{b}_j = \tilde{b}_3 = 1$, and $b_k \neq \tilde{b}_k$ only for $3 = i \leq k \leq j = 3$.

Example (Assume binary number system.)
$x = 0.011$ and $\hat{x} = 0.100 \Rightarrow (\hat{x} - x) = 0.001 = .\bar{b}_1\bar{b}_2\bar{b}_3\cdots \Rightarrow j = 3$
$\{b_0, b_1, b_2, \cdots, b_j\} = \{b_0, b_1, b_2, b_3\} = \{1, 0, 1, 1\} \Rightarrow i = 1$
$x + 2^{-j} = .\tilde{b}_1\tilde{b}_2\tilde{b}_3\cdots = 0.011 + 2^{-3} = 0.100$
$\hat{x} = 0.100 \geq x + 2^{-j} = 0.100 > x = 0.011$ and $b_i b_{i+1}\cdots b_j = b_1 b_2 b_3 = 011$, $\tilde{b}_i\tilde{b}_{i+1}\cdots\tilde{b}_j = \tilde{b}_1\tilde{b}_2\tilde{b}_3 = 100$, and $b_k \neq \tilde{b}_k$ for $1 = i \leq k \leq j = 3$.

A continuous cdf without pdf

We then derive

$$
\begin{aligned}
&F_X(\hat{x}) - F_X(x) \geq F_X(x + 2^{-j}) - F_X(x) \\
&= \sum_{k=1}^{\infty} \tilde{b}_k p \left(\prod_{\ell=1}^{k-1} [(1-2p)\tilde{b}_\ell + p] \right) - \sum_{k=1}^{\infty} b_k p \left(\prod_{\ell=1}^{k-1} [(1-2p) b_\ell + p] \right) \\
&= \sum_{k=i}^{\infty} \tilde{b}_k p \left(\prod_{\ell=1}^{k-1} [(1-2p)\tilde{b}_\ell + p] \right) - \sum_{k=i}^{\infty} b_k p \left(\prod_{\ell=1}^{k-1} [(1-2p) b_\ell + p] \right) \\
&= \left(\prod_{\ell=1}^{i-1} [(1-2p) b_\ell + p] \right) \left[\sum_{k=i}^{\infty} \tilde{b}_k p \left(\prod_{\ell=i}^{k-1} [(1-2p)\tilde{b}_\ell + p] \right) - \sum_{k=i}^{\infty} b_k p \left(\prod_{\ell=i}^{k-1} [(1-2p) b_\ell + p] \right) \right] \\
&= \left(\prod_{\ell=1}^{i-1} [(1-2p) b_\ell + p] \right) \left[p + \sum_{k=j+1}^{\infty} \tilde{b}_k p \left((1-p)p^{j-i} \prod_{\ell=j+1}^{k-1} [(1-2p)\tilde{b}_\ell + p] \right) \right. \\
&\qquad \left. - \sum_{k=0}^{j-i-1} p^2 (1-p)^k - \sum_{k=j+1}^{\infty} b_k p \left(p(1-p)^{j-i} \prod_{\ell=j+1}^{k-1} [(1-2p) b_\ell + p] \right) \right] \\
&= \left(\prod_{\ell=1}^{i-1} [(1-2p) b_\ell + p] \right) \left[p(1-p)^{j-i} + \sum_{k=j+1}^{\infty} b_k p \left(\prod_{\ell=j+1}^{k-1} [(1-2p) b_\ell + p] \right) \left((1-p)p^{j-i} - p(1-p)^{j-i} \right) \right]
\end{aligned}
$$

A continuous cdf without pdf

where we assume $\prod_{\ell=i}^{k-1}[(1-2p)b_\ell + p] = 1$ for $k = i$. This implies that $F_X(x)$ is strictly increasing for $x \in [0,1)$.

In addition, for any $x \in [0,1)$,

$$\begin{aligned}
\Pr[X = x] &= \Pr\left[B_1 = b_1 \wedge B_2 = b_2 \wedge B_3 = b_3 \wedge \cdots\right] \\
&= \Pr\left[B_1 = b_1\right]\Pr\left[B_2 = b_2\right]\Pr\left[B_3 = b_3\right]\cdots \\
&= \prod_{i=1}^{\infty}[(1-2p)b_i + p] \\
&\le \prod_{i=1}^{\infty}\max(p, 1-p) \\
&= 0,
\end{aligned}$$

which indicates $F_X(\cdot)$ is continuous over $[0,1)$.

Theorem 31.2 A nondecreasing function is differentiable almost everywhere (i.e., except on a set of Lebesgue measure 0), and its derivative is non-negative.

A continuous cdf without pdf

Thus, from Theorem 31.2, $F_X(\cdot)$ is differentiable almost everywhere, namely,

$$\begin{aligned}&\limsup_{\varepsilon\downarrow 0}\frac{F_X(x+\varepsilon)-F_X(x)}{\varepsilon}=\liminf_{\varepsilon\downarrow 0}\frac{F_X(x+\varepsilon)-F_X(x)}{\varepsilon}\\ =\ &\limsup_{\varepsilon\downarrow 0}\frac{F_X(x)-F_X(x-\varepsilon)}{\varepsilon}=\liminf_{\varepsilon\downarrow 0}\frac{F_X(x)-F_X(x-\varepsilon)}{\varepsilon}=F'(x),\end{aligned}$$

and $F'_X(x)\geq 0$.

Theorem 31.3 If f is non-negative and integrable, and if $F(x)=\int_{-\infty}^{x}f(t)dt$, then $F'(x)=f(x)$ except on a set of Lebesgue measure 0.

Now suppose there exists $f(x)$ (nonnegative and integrable) such that $F_X(x)=\int_{-\infty}^{x}f(t)dt$. Then by Theorem 31.3, $F'_X(x)=f(x)$ except on a set of Lebesgue measure 0. However, $F'_X(x)=0$ **for every** $x\in[0,1)$ **at which $F_X(x)$ is differentiable** (See the proof below). Hence, $f(x)=0$ except on a set of Lebesgue measure 0; a contradiction is thus obtained since $F_X(x)>0$ but $\int_{-\infty}^{x}f(y)dy=0$ for any $x\in[0,1)$. □

A continuous cdf without pdf

Suppose $F'_X(x) = c > 0$ for some $x \in [0, 1)$. Let $k_n \triangleq \lfloor x2^n \rfloor$, $k_n 2^{-n} \triangleq \sum_{i=1}^{n} u_i 2^{-i}$, where $u_i \in \{0, 1\}$, and

$$\begin{aligned} a_n &\triangleq F_X((k_n + 1)2^{-n}) - F_X(k_n 2^{-n}) \\ &= \Pr\left[k_n 2^{-n} < X \le (k_n + 1)2^{-n}\right] \\ &= \Pr\left[k_n 2^{-n} \le X < (k_n + 1)2^{-n}\right] \\ &\qquad \text{(since } \Pr(X = k_n 2^{-n}) = \Pr(X = (k_n + 1)2^{-n}) = 0) \\ &= \prod_{i=1}^{n} [(1 - 2p)u_i + p] \end{aligned}$$

Then,

$$F'_X(x) = \lim_{n\to\infty} \frac{F_X((k_n + 1)2^{-n}) - F_X(k_n 2^{-n})}{2^{-n}} = \lim_{n\to\infty} \frac{a_n}{2^{-n}} = c,$$

which implies $a_{n+1}/a_n \to 1/2$. However, $a_{n+1}/a_n = [(1 - 2p)u_{n+1} + p] \not\to 1/2$, a contradiction.

A continuous cdf without pdf

- It can be shown that

$$F_X(x) = \begin{cases} pF_X(2x), & \text{if } 0 \le x \le 1/2; \\ p + (1-p)F_X(2x-1), & \text{if } 1/2 \le x \le 1 \end{cases}$$

So by $F_X(0) = 0$ and $F_X(1) = 1$, we can obtain $F_X(1/2) = p$, $F_X(1/4) = p^2$ and $F_X(3/4) = p + (1-p)p, \ldots$

Distribution of a function of random variable

- If $g(\cdot)$ is strictly increasing , then $T(\cdot) = g^{-1}(\cdot)$ exists.

  Then
  $$\Pr[g(X) \leq x] = \Pr[X \leq T(x)] = F_X(T(x)).$$

  Now if $F_X(\cdot)$ has continuous derivative $f(\cdot)$, and $T(\cdot)$ is differentiable, then the pdf of $g(X)$ is equal to:
  $$f(T(x))T'(x).$$

- Now how about general $g(\cdot)$? We can still have:
  $$\Pr[g(X) \leq x] = \Pr[X \in \mathcal{G}_x]$$
  where $\mathcal{G}_x = \{r \in \Re : g(r) \leq x\}$.

  – It is always suggestive to derive cdf first. Then examine whether the cdf has **continuous derivative** or not. If it does, then a pdf can be obtained by taking the derivative of the cdf.

Multidimensional distributions

- The (joint) cdf of a multidimensional random vector $\boldsymbol{X} = (X_1, \ldots, X_k)$ is defined as:

$$F_{\boldsymbol{X}}(x^k) = \Pr\left[X_1 \leq x_1 \wedge X_2 \leq x_2 \wedge \cdots \wedge X_k \leq x_k\right].$$

Multidimensional distributions

Theorem 12.5 (a variant version) Suppose that a function $F : \Re^k \to \Re$ satisfies:

- (continuous from above) $F(\cdot)$ is continuous in the sense that $\lim_{h\downarrow 0} F(x_1 + hy_1, \ldots, x_k + hy_k) = F(x_1, \ldots, x_k)$ for all $x^k \in \Re^k$ and $y^k \in \Re^k$ with each $y_i > 0$.
- $F(\Delta_A)$ is non-negative for any *bounded rectangle* Δ_A, namely, for any a^k and b^k with $a_i \le b_i$ for $1 \le i \le k$,

$$F(\Delta_A) = \sum_{\text{For each } i,\ x_i = \text{either } a_i \text{ or } b_i} (-1)^{s(x^k)} F(x^k) \ge 0,$$

  where $s(x^k)$ is the number of x_i equal to a_i (an equivalent extension to non-decreasingness for one dimensional random variable),
- $\lim_{h\downarrow -\infty} F(hy_1, \ldots, hy_k) = 0$ for $y^k \in \Re^k$ with each $y_i > 0$;
- $\lim_{h\uparrow\infty} F(hy_1, \ldots, hy_k) = 1$ for $y^k \in \Re^k$ with each $y_i > 0$.

Then there exists a unique probability measure whose resultant k-dimensional cdf is equal to $F(\cdot)$.

Multidimensional distributions

- The previous theorem indicates the sufficiency of defining multidimensional cdf for a multidimensional random vector (or probability measure).
- Difference between one-dimensional cdf and multidimensional cdf.
  - A one-dimensional cdf can only have **countably many** discontinuities;
  - but a more-than-one-dimensional cdf can have **uncountably many** discontinuities.
- Similarity between one-dimensional cdf and multidimensional cdf.
  - Only countably many discontinuous points can have positive probability mass.

Support and density

Definition (Lebesgue measure) A Lebesgue measure λ over the Borel set $\mathcal{B}^k$ is that for any $\mathcal{A} \in \mathcal{B}^k$,

$$\lambda(\mathcal{A}) = \sum_{i=1}^{\infty} \lambda(\Delta_i),$$

and $\{\Delta_i\}_{i=1}^{\infty}$ are disjoint bounded rectangles satisfying $\mathcal{A} = \cup_{i=1}^{\infty}\Delta_i$, and $\lambda(\Delta)$ is equal the volume of the bounded rectangle Δ (namely, $\prod_{i=1}^{k}(b_i - a_i)$, where a^k and b^k define the bounded rectangle).

Definition (support) The support of a multidimensional random vector $\boldsymbol{X}$ is a Borel set $\mathcal{H} \in \mathcal{B}^k$ for which $\Pr[\boldsymbol{X} \in \mathcal{H}] = 1$.

Definition A multidimensional random vector is *discrete* if it has countable support.

Definition (probability density function) A random variable $\boldsymbol{X}$ and its distribution (cdf) have *density* f with respect to Lebesgue measure, if f is $\mathcal{B}^k$-measurable non-negative function that satisfies

$$\Pr[\boldsymbol{X} \in \mathcal{A}] = \int_{\mathcal{A}} f(x^k)\lambda(dx^k) = \int_{\mathcal{A}} f(x^k)dx^k \quad \text{for every } \mathcal{A} \in \mathcal{B}^k.$$

Density for a mapping

Suppose that

- $\boldsymbol{X}$ has density f;
- $\boldsymbol{X}$ has support V;
- $g = (g_1, \ldots, g_k)$ is a 1-to-1 mapping from V onto U, where V and U are open sets in $\Re^k$;
- $\partial g_i / \partial x_j$ exists and is continuous in V for every i, j (i.e., g is *continuous, differentiable*);
- $\partial T_i / \partial x_j$ exists and is continuous in U for every i, j (i.e., T is *continuous, differentiable*), where T is the inverse mapping to g;
- the Jocobian determinant

$$J(\boldsymbol{x}) = \mathrm{Det} \begin{bmatrix} \frac{\partial T_1}{\partial x_1} & \frac{\partial T_1}{\partial x_2} & \cdots & \frac{\partial T_1}{\partial x_k} \\ \frac{\partial T_2}{\partial x_1} & \frac{\partial T_2}{\partial x_2} & \cdots & \frac{\partial T_2}{\partial x_k} \\ \vdots & \vdots & \cdots & \vdots \\ \frac{\partial T_k}{\partial x_1} & \frac{\partial T_k}{\partial x_2} & \cdots & \frac{\partial T_k}{\partial x_k} \end{bmatrix} (\boldsymbol{x}) \neq 0 \quad \text{for every } \boldsymbol{x} \in U.$$

Then $g(\boldsymbol{X})$ has density $f(T(\boldsymbol{x}))|J(\boldsymbol{x})|$.

Independence

Definition (independence) Random variables $X_1, \ldots, X_k$ are independent if

$$\Pr\left[X_1 \in \mathcal{H}_1 \wedge \cdots \wedge X_k \in \mathcal{H}_k\right] = \Pr\left[X_1 \in \mathcal{H}_1\right] \cdots \Pr\left[X_k \in \mathcal{H}_k\right]$$

for all **linear** (See blow for the definition of linearity) Borel set $\mathcal{H}_j \in \mathcal{B}$.

- A Borel set is **linear** if it is one-dimensional.
- Again,

  $\Pr\left[X_1 \in \mathcal{H}_1 \wedge \cdots \wedge X_k \in \mathcal{H}_k\right] = \Pr\left[X_1 \in \mathcal{H}_1\right] \cdots \Pr\left[X_k \in \mathcal{H}_k\right]$ for all linear Borel set $\mathcal{H}_j$

  if, and only if,

  $\Pr\left[X_1 \leq x_1 \wedge \cdots \wedge X_k \leq x_k\right] = \Pr\left[X_1 \leq x_1\right] \cdots \Pr\left[X_k \leq x_k\right]$ for all real number x_j.

Independence for multidimensional random vector

Definition (independence) Random vectors $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_u$ are independent if

$$\Pr[\boldsymbol{X}_1 \in \mathcal{H}_1 \wedge \cdots \wedge \boldsymbol{X}_k \in \mathcal{H}_k] = \Pr[\boldsymbol{X}_1 \in \mathcal{H}_1] \cdots \Pr[\boldsymbol{X}_k \in \mathcal{H}_k]$$

for all Borel set $\mathcal{H}_j$.

- Notably, the dimension of each random variable needs not be identical.
- Again,

  $\Pr[\boldsymbol{X}_1 \in \mathcal{H}_1 \wedge \cdots \wedge \boldsymbol{X}_k \in \mathcal{H}_k] = \Pr[\boldsymbol{X}_1 \in \mathcal{H}_1] \cdots \Pr[\boldsymbol{X}_k \in \mathcal{H}_k]$ for all Borel set $\mathcal{H}_j$

  if, and only if,

  $\Pr[\boldsymbol{X}_1 \leq \boldsymbol{x}_1 \wedge \cdots \wedge \boldsymbol{X}_k \leq \boldsymbol{x}_k] = \Pr[\boldsymbol{X}_1 \leq \boldsymbol{x}_1] \cdots \Pr[\boldsymbol{X}_k \leq \boldsymbol{x}_k]$ for all real vector $\boldsymbol{x}_j$.

Theorem 20.3 (a rephrased version) If $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent, then

$$\Pr\left[(\boldsymbol{X},\boldsymbol{Y}) \in \mathcal{B}\right] = \int_{\mathcal{X}} \Pr\left[(\boldsymbol{x},\boldsymbol{Y}) \in \mathcal{B}\right] dF_{\boldsymbol{X}}(\boldsymbol{x}),$$

where $\mathcal{X}$ is the support of $\boldsymbol{X}$.

An exemplified application of Theorem 20.3

Suppose that X and Y are independent and are exponentially distributed random variables with parameters α and β, respectively. Then

$$\begin{aligned}
\Pr[Y/X \ge z] &= \int_0^\infty \Pr[Y/x \ge z] dF_X(x) \\
&= \int_0^\infty \Pr[Y \ge xz] \alpha e^{-\alpha x} dx \\
&= \int_0^\infty \left(\int_{xz}^\infty \beta e^{-\beta y} dy \right) \alpha e^{-\alpha x} dx \\
&= \begin{cases} \dfrac{\alpha}{\alpha + \beta z}, & \text{if } z > 0; \\ 1, & \text{if } z \le 0. \end{cases}
\end{aligned}$$

Partial integration for general random vector

If $\boldsymbol{X}$ and $\boldsymbol{Y}$ are not necessarily independent, then

$$\Pr\left[(\boldsymbol{X},\boldsymbol{Y})\in\mathcal{B}\right]=\int_{\mathcal{X}}\Pr\left[(\boldsymbol{x},\boldsymbol{Y})\in\mathcal{B}\middle|\boldsymbol{X}=\boldsymbol{x}\right]dF_{\boldsymbol{X}}(\boldsymbol{x}),$$

where $\mathcal{X}$ is the support of $\boldsymbol{X}$.

Convolution for independent random variables

For independent random variables X and Y,

$$\begin{aligned} F_{X+Y}(z) &= \Pr[X+Y \le z] \\ &= \int_{-\infty}^{\infty} \Pr[x+Y \le z] dF_X(x) \\ &= \int_{-\infty}^{\infty} \Pr[Y \le z-x] dF_X(x) \\ &= \int_{-\infty}^{\infty} F_Y(z-x) dF_X(x). \end{aligned}$$

If Y has density $f_Y(\cdot)$, then for fixed x,

$$F_Y(z-x) = \int_{-\infty}^{z-x} f_Y(s)ds = \int_{-\infty}^{z} f_Y(t-x)dt,$$

which implies

$$\begin{aligned} F_{X+Y}(z) &= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{z} f_Y(t-x)dt \right) dF_X(x) \\ &= \int_{-\infty}^{z} \left(\int_{-\infty}^{\infty} f_Y(t-x)dF_X(x) \right) dt. \end{aligned}$$

So $Z = X + Y$ has density $\int_{-\infty}^{\infty} f_Y(z-x)dF_X(x)$.

Convolution for independent random variables

In other words, if one of independent random variables X and Y has density, then $X+Y$ has density.

If, in addition to Y, X has density,

$$f_{X+Y}(z) = \int_{-\infty}^{\infty} f_Y(z-x)dF_X(x) = \int_{-\infty}^{\infty} f_Y(z-x)f_X(x)dx = (f_X * f_Y)(z),$$

where "$*$" denotes the *convolution* operator.

Example 20.5 Let $X_1, \ldots, X_k$ be independent random variables, each with exponential density with parameter α. Then $X_1 + \cdots + X_k$ has density satisfying:

$$f_{X_1+\cdots+X_k} = f_{X_2+\cdots+X_k} * f_{X_1}.$$

Then

$$f_{X_1+\cdots+X_k}(z) = \alpha \frac{(\alpha z)^{k-1}}{(k-1)!} e^{-\alpha z},$$

which can be proved by induction.

Theorem 20.5

1. If X_n converges to X with probability 1, then X_n converges to X in probability.
   - Convergence in probability is sometimes denoted as $\text{p-}\lim_{n\to\infty} X_n = X$.
2. X_n converges to X in probability if, and only if, for each subsequence $\{X_{n_k}\}_{k=1}^{\infty}$ of $\{X_n\}_{n=1}^{\infty}$, there exists a further subsequence $\{X_{n_{k_i}}\}_{i=1}^{\infty}$ such that $X_{n_{k_i}}$ converges to X with probability 1 as i goes to infinity.

Proof of Theorem 20.5-2.:

(a) *Only-if* **part:** X_n converges to X in probability means that for any $\varepsilon > 0$,

$$\lim_{n\to\infty} \Pr\left[|X_n - X| \geq \varepsilon\right] = 0.$$

Hence, for any i fixed and any sequence $\{n_k\}_{k=1}^{\infty}$,

$$\lim_{k\to\infty} \Pr\left[|X_{n_k} - X| \geq \varepsilon\right] = 0.$$

Therefore, with initially $k_0 = 0$,

$$k_i \triangleq \min\left\{k > k_{i-1} : \Pr\left[|X_{n_k} - X| \geq \varepsilon\right] < \frac{1}{2^i}\right\}$$

exists.

We then obtain that $k_1 < k_2 < k_3 < \cdots$, and

$$\Pr\left[\left|X_{n_{k_i}} - X\right| \geq \varepsilon\right] < \frac{1}{2^i}.$$

By the first Borel-Cantelli lemma, $\sum_{i=1}^{\infty} \Pr\left[\left|X_{n_{k_i}} - X\right| \geq \varepsilon\right] < \infty$ implies that $X_{n_{k_i}}$ converges to X with probability 1.

(b) ***If* part:** If X_n does not converge to X in probability, then there exists $\varepsilon > 0$ such that

$$\limsup_{n\to\infty} \Pr\left[|X_n - X| \geq \varepsilon\right] > 0,$$

or equivalently, there exists a subsequence $\{n_k\}_{k=1}^{\infty}$,

$$\lim_{k\to\infty} \Pr\left[|X_{n_k} - X| \geq \varepsilon\right] > 0.$$

Since $\{X_{n_k}\}_{k=1}^{\infty}$ does not converge to X in probability, no subsequence of $\{X_{n_k}\}_{k=1}^{\infty}$ can converge to X with probability 1. $\square$

Some useful observations

Observations

1. If $\{X_n\}_{n=1}^{\infty}$ converges to X in probability, and $\{X_n\}_{n=1}^{\infty}$ converges to Y in probability, then $\Pr[X = Y] = 1$.

2. If $\{X_n\}_{n=1}^{\infty}$ converges to X in probability, then $\{f(X_n)\}_{n=1}^{\infty}$ converges to $f(X)$ in probability for any continuous function f.

Glivenko-Cantelli Theorem

- Relation between an empirical distribution and the respective true distribution is an essential question to engineers!
- The *empirical distribution* (cdf) for an i.i.d. sequence X_1, X_2, X_3, $\ldots$ is a **random variable**, defined as:

$$\boldsymbol{F}_n(x) = \frac{1}{n}\sum_{k=1}^{n} I_{(-\infty,x]}(X_k),$$

where $I_{(-\infty,x]}(u) = \begin{cases} 1, & \text{if } u \le x; \\ 0, & \text{if } u > x. \end{cases}$

- Since $I_{(-\infty,x]}(X_1), I_{(-\infty,x]}(X_2), \ldots$ are also i.i.d., the strong law of large numbers indicates $\frac{1}{n}\sum_{k=1}^{n} I_{(-\infty,x]}(X_k)$ converges to its marginal mean, that is

$$E\left[I_{(-\infty,x]}(X_1)\right] = F(x),$$

with probability 1.

- Can the relation between $\boldsymbol{F}_n(x)$ and $F(x)$ be stronger than that implied by the strong law of large numbers? Yes, answered by Glivenko-Cantelli.

Glivenko-Cantelli Theorem

- Glivenko-Cantelli Theorem says that the random variable

$$D_n = \sup_{x \in \Re} |\boldsymbol{F}_n(x) - F(x)|$$

converges to 0 with probability 1, a much stronger statement!

In other words,

$$\Pr\left[\lim_{n\to\infty} \sup_{x \in \Re} |\boldsymbol{F}_n(x) - F(x)| = 0\right] = 1$$

is a stronger statement than

$$\Pr\left[\lim_{n\to\infty} |\boldsymbol{F}_n(x) - F(x)| = 0\right] = 1 \text{ for any } x \in \Re.$$

- We will discuss how fast D_n converges to 0 after the introduction of the Berry-Esseen Theorem.

Glivenko-Cantelli Theorem

Theorem 20.6 (Glivenko-Cantelli Theorem) Suppose that $X_1, X_2, \ldots$ are i.i.d. Then

$$D_n = \sup_{x \in \Re} |\boldsymbol{F}_n(x) - F(x)|$$

converges to 0 with probability 1, where

$$\boldsymbol{F}_n(x) = \frac{1}{n} \sum_{k=1}^{n} I_{(-\infty, x]}(X_k),$$

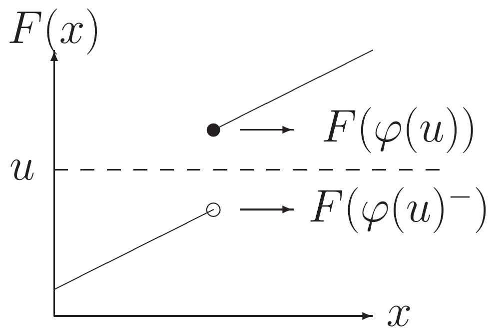
and $I_{(-\infty,x]}(u) = \begin{cases} 1, & \text{if } u \le x; \\ 0, & \text{if } u > x. \end{cases}$

Proof:

- By the strong law of large numbers, the event $A_x = \left[\lim_{n\to\infty} \boldsymbol{F}_n(x) = F(x)\right]$ has probability 1. In addition, the event $B_x = \left[\lim_{n\to\infty} \boldsymbol{F}_n(x^-) = F(x^-)\right]$ has probability 1, where

$$F(x^-) = \Pr[X < x] \text{ and } \boldsymbol{F}_n(x^-) = \frac{1}{n} \sum_{k=1}^{n} I_{(-\infty, x)}(X_k).$$

Glivenko-Cantelli Theorem



- Define the *quantile function* $\varphi(u) = \inf[x \in \Re : u \le F(x)]$ for $0 \le u \le 1$. Then

$$F(\varphi(u)^-) \le u \le F(\varphi(u)).$$

By the right-continuity of $F(\cdot)$, $\{x \in \Re : u \le F(x)\} = [\varphi(u), \infty)$.
So if $F(\cdot)$ is continuous at $x = \varphi(u)$, then $F(\varphi(u)^-) = u = F(\varphi(u))$;
else if $F(\cdot)$ is only right-continuous and has a jump at $x = \varphi(u)$, we have $F(\varphi(u)^-) \le u \le F(\varphi(u))$. ($F(x^-) \triangleq \lim_{\delta \downarrow 0} F(x - \delta)$)

- Let $x_{m,k} = \varphi(k/m)$ for $m \ge 1$ and $1 \le k \le m$, where the infimum of an empty set is infinity, and $x_{m,0} = -\infty$. Hence,

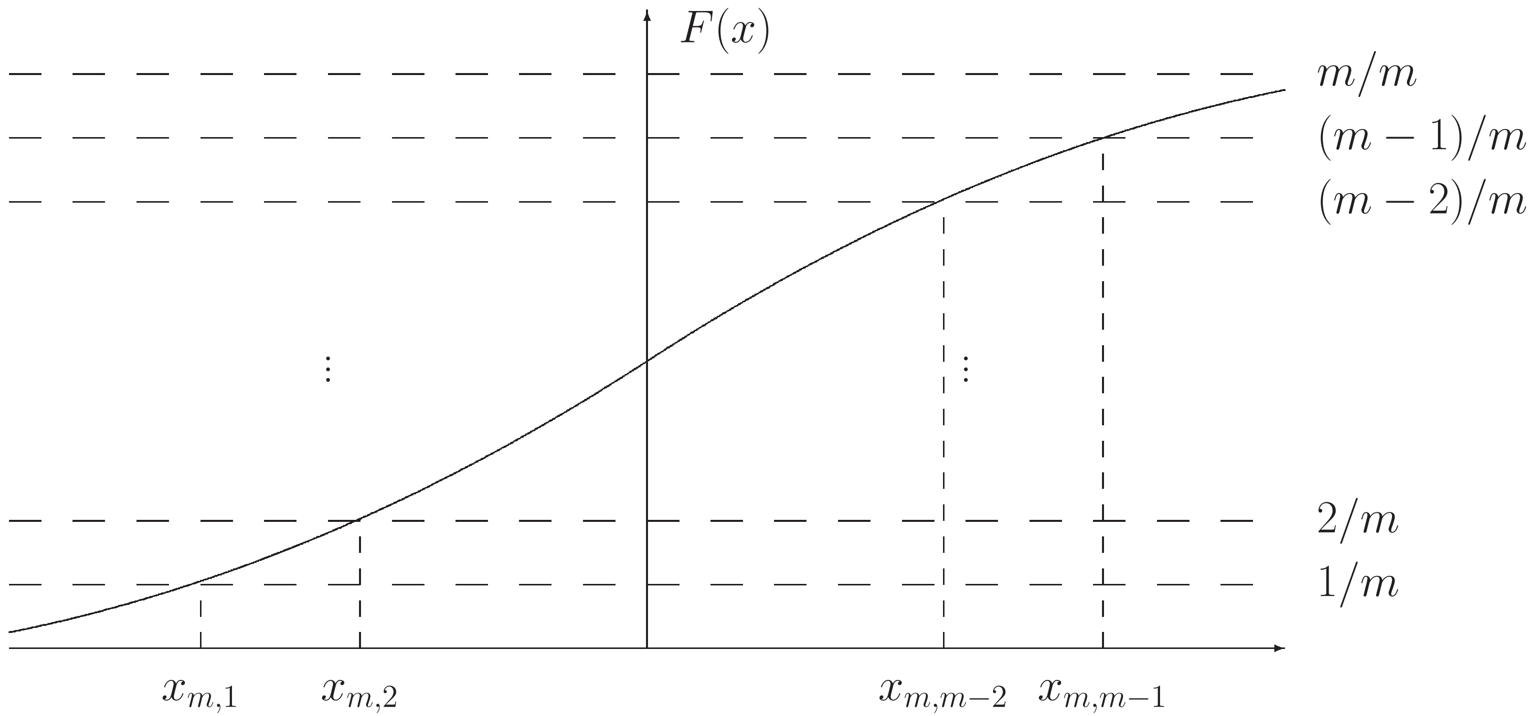
$$F(x_{m,k}^-) \le \frac{k}{m} \quad \big(\le F(x_{m,k})\big) \quad \text{and} \quad \big(F(x_{m,k-1}^-) \le \big) \quad \frac{k-1}{m} \le F(x_{m,k-1}),$$

which implies $F(x_{m,k}^-) - F(x_{m,k-1}) \le \dfrac{1}{m}$. (What if $x_{m,k-1} = x_{m,k}$?)

So for $x_{m,k-1} \le x < x_{m,k}$ (Hence, $F(x_{m,k-1}) \le F(x) \le F(x_{m,k}^-)$.),

$$F(x_{m,k}^-)\left(\le F(x_{m,k-1}) + \frac{1}{m}\right) \le F(x) + \frac{1}{m} \quad \text{and} \quad F(x_{m,k-1})\left(\ge F(x_{m,k}^-) - \frac{1}{m}\right) \ge F(x) - \frac{1}{m}.$$

Glivenko-Cantelli Theorem



Glivenko-Cantelli Theorem

- Define $D_{m,n} \triangleq \max_{k=1,\ldots,m} \left\{ \left| \boldsymbol{F}_n(x_{m,k}) - F(x_{m,k}) \right|, \left| \boldsymbol{F}_n(x_{m,k}^-) - F(x_{m,k}^-) \right| \right\}$.

  Then

$$D_n \leq D_{m,n} + \frac{1}{m},$$

because for $x_{m,k-1} \leq x < x_{m,k}$,

$$\boldsymbol{F}_n(x) \leq \boldsymbol{F}_n(x_{m,k}^-) \leq F(x_{m,k}^-) + D_{m,n} \leq F(x) + \frac{1}{m} + D_{m,n}$$

and

$$\boldsymbol{F}_n(x) \geq \boldsymbol{F}_n(x_{m,k-1}) \geq F(x_{m,k-1}) - D_{m,n} \geq F(x) - \frac{1}{m} - D_{m,n}.$$

Hence,

$$-\frac{1}{m} - D_{m,n} \leq \boldsymbol{F}_n(x) - F(x) \leq \frac{1}{m} + D_{m,n},$$

or equivalently,

$$|\boldsymbol{F}_n(x) - F(x)| \leq \frac{1}{m} + D_{m,n}.$$

Accordingly,

$$D_n = \sup_{x \in \Re} |\boldsymbol{F}_n(x) - F(x)| \leq \frac{1}{m} + D_{m,n}.$$

Glivenko-Cantelli Theorem

- $0 \le D_n \le D_{m,n} + \frac{1}{m}$ indicates that:
  - if $D_{m,n}$ converges to zero with probability 1 $\left(\text{i.e., } \Pr\left[\lim_{n\to\infty} D_{m,n} = 0\right] = 1\right)$, then $\Pr\left[0 \le \limsup_{n\to\infty} D_n \le 1/m\right] = 1$ for any m.
  - Therefore, $\Pr\left[\lim_{n\to\infty} D_n = 0\right] = 1$.
- *Proof of $D_{m,n}$ converging to 0 with probability 1*:

  As $\Pr[A_{x_{m,k}}] = \Pr[B_{x_{m,k}}] = 1$ for $1 \le k \le m$,

$$\Pr\left[\bigcap_{k=1}^{m}\left(A_{x_{m,k}} \bigcap B_{x_{m,k}}\right)\right] = 1.$$

  The above statement is equivalent to saying that for each $1 \le k \le m$,

$$\lim_{n\to\infty} |\boldsymbol{F}(x_{m,k}) - F(x_{m,k})| = \lim_{n\to\infty} \left|\boldsymbol{F}(x_{m,k}^-) - F(x_{m,k}^-)\right| = 0 \text{ with probability 1}$$

  implies that the **finite** maximum

$$D_{m,n} = \max_{k=1,\ldots,m} \left\{|\boldsymbol{F}(x_{m,k}) - F(x_{m,k})|, \left|\boldsymbol{F}(x_{m,k}^-) - F(x_{m,k}^-)\right|\right\}$$

  certainly converges to zero with probability 1. □

Glivenko-Cantelli Theorem

- No one can directly claim without proof that $D_n = \sup_{x\in\Re}|\boldsymbol{F}_n(x) - F(x)|$ converges to 0 with probability 1 simply because for each $x \in \Re$, $|\boldsymbol{F}_n(x) - F(x)|$ converges to 0 with probability 1. So, the proof of the Glivenko-Cantelli theorem is not trivial!

  Let $f_n(x) = 0$ for $x < n$, and 1, for $x \geq n$.
  Then $\lim_{n\to\infty} f_n(x) = 0$ for each $x \in \Re$.
  But $\lim_{n\to\infty} \sup_{x\in\Re} f_n(x) = 1$.

- In our slides, I intentionally avoid using the *inherited probability space*, and only rely on the *observation probability space*, namely, the cdf itself, since the *observation probability space* is what we engineers are more familiar with.

- From this, you learn that to rely on the *observation probability space* is sufficient for most problems of engineering interest; however, I would like to point out that it is advantageous to learn the role of the intrinsic, inherited probability space on which a random variable is originally defined.

More properties on cdf

Theorem (decomposition of distribution) Any cdf F can be decomposed into cdf's of three distinct types: **discrete** F_d, **absolutely continuous** F_{ac} and **singular** F_s, each of which is itself a cdf. In other words, for $x \in \Re$,

$$F(x) = \alpha_1 F_d(x) + \alpha_2 F_{ac}(x) + \alpha_3 F_s(x),$$

where $\alpha_1 + \alpha_2 + \alpha_3 = 1$.

Definition F_d is a discrete cdf, if it is a cdf for a discrete random variable.

Definition F_{ac} is an absolutely continuous cdf, if it has density, namely, there exists f_{ac} such that

$$F_{ac}(x) = \int_{-\infty}^{x} f_{ac}(t)dt.$$

(Notably, a discrete cdf does not have density from (rigorous) mathematical standpoint, unless the Dirac delta function $\delta(\cdot)$ is acceptable to be a legitimate density for engineering convenience, which satisfies

$$\delta(t) = \begin{cases} 0, & \text{if } t \neq 0; \\ \infty, & \text{if } t = 0 \end{cases},$$

and $\int_{\mathcal{X}} \delta(t)dt = 1$ if $0 \in \mathcal{X}$.)

More properties on cdf

Definition F_s is a singular (and continuous) cdf, if it is not discrete (so the number of jumps is uncountably many, if it has jumps) and does not have density.

In other words, it is a continuous cdf without density. You may refer to Slide 20-40 for a specific one-dimensional example.

The singular cdf is in fact more easily to construct for multidimensional random variables. For example, $\Pr[X_1 + X_2 = 0] = 1$, and each of X_1 and X_2 is Gaussian distributed with mean 0 and variance 1.
Then, the pdf $f_{X_1,X_2}(x_1, x_2)$ for (X_1, X_2) does not exist (even if the cdf $F_{X_1+X_2}(z)$ for $X_1 + X_2$ exists)!