STAT 240 Course Notes - Fall 2018

Max Zhu

December 8, 2018

Contents

1 Foundations

Andrey Kolmogorov *(1933, "Foundations of the theory of probability")* put probability on solid mathematical grounds using a model, probability space $(\Omega, \mathcal{F}, P)$. A probability space consists of:

1. Sample space Ω: set of all outcomes ω
2. σ-algebra $\mathcal{F}$: set of all events, i.e. subsets of Ω to which we can assign a probability
3. Probability measure $P : \mathcal{F} \to [0, 1]$: function which assigns probabilities to events

We need measure theory to understand this.

1.1 σ-algebras and measures

A classical problem is to measure the volume $\lambda(A)$ of some $A \subseteq \mathbb{R}^d, d \geq 1$. Consider $d = 1$. Then, λ should:

(i) Assign to intervals its length: $\lambda([a, b]) = b - a$ for all $a, b \in \mathbb{R}, a \leq b$

(ii) Be invariant under translations, rotations, and reflections: $\lambda(A) = \lambda(B)$ for all congruent $A, B \subseteq \mathbb{R}$

(iii) Be σ-additive:
If $\{A_i\}_{i\in\mathbb{N}} \subseteq \mathbb{R}, A_i \cap A_j = \varnothing \; \forall i \neq j$, then $\lambda(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \lambda(A_i)$

In other words, the volume of the union of countably many subsets of $\mathbb{R}$ is equal to the sum of their volumes.

Can we take the power set $\mathcal{P}(\mathbb{R}) = \{A : A \subseteq \mathbb{R}\}$ as $\mathcal{F}$ and find such a λ?

Theorem 1.1: Vitali's Theorem

There exists no λ defined on $\mathcal{P}(\mathbb{R})$ which fulfils (i)-(iii). Furthermore, any measurable set $A \subseteq \mathbb{R} : \lambda(A) > 0$ contains a non-measurable set V (Vitali set).

What about weakening (iii) to finitely many sets? Still no!

Theorem 1.2: Banach-Tarski

Let $d \geq 3$ and $A, B \in \mathbb{R}^d$ be bounded with non-empty interior. Then, there exists $k \in \mathbb{N}$ and partitions:

$$A = \dot{\bigcup}_{i=1}^{k} A_i$$
$$B = \dot{\bigcup}_{i=1}^{k} B_i$$

such that A_i, B_i are congruent $\forall i \in \{1, \dots, k\}$.

For countable Ω, one can define λ (or μ or P) on $\mathcal{P}(\Omega)$ but for uncountable Ω, $\mathcal{P}(\Omega)$ is too large. Therefore, we need to define $\lambda, \mu,$ or P on some proper subset of $\mathcal{P}(\Omega)$ which is closed under certain set operations.

Definition 1.3: σ-algebra

$\mathcal{F} \subseteq \mathcal{P}(\Omega)$ is a σ-algebra on Ω if:

(i) $\Omega \in \mathcal{F}$

(ii) $A \in \mathcal{F} \Rightarrow A^c \in \mathcal{F}$

(iii) $\{A_i\}_{i\in\mathbb{N}} \subseteq \mathcal{F} \Rightarrow \bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$

If (iii) holds for only finitely many sets, $\mathcal{F}$ is an algebra.

Remark 1.4

σ-algebras are closed w.r.t. countable intersection, since

$$\bigcap_{i=1}^{\infty} A_i = (\bigcup_{i=1}^{\infty} A_i^c)^c \in \mathcal{F}$$

by de Morgan.

Example 1.5

1. Trivial σ-algebra: $\{\varnothing, \Omega\} = \mathcal{F}$
2. $\mathcal{F} = \{\dot{\bigcup}_{i=1}^{n} (a_i, b_i] : 0 \leq a_i \leq b_i \leq 1 \ \forall i, n \in \mathbb{N}\}$ is an algebra on $\Omega = (0, 1]$ but not a σ-algebra since $\dot{\bigcup}_{n=0}^{\infty} (\sum_{k=1}^{2n} (\frac{1}{2})^k, \sum_{k=1}^{2n+1} (\frac{1}{2})^k] \notin \mathcal{F}$, while $(\sum_{k=1}^{2n} (\frac{1}{2})^k, \sum_{k=1}^{2n+1} (\frac{1}{2})^k] \in \mathcal{F}$ for all k.

How can σ-algebras be constructed?

Proposition 1.6

Given $A \subseteq \mathcal{P}(\Omega)$, then there exists a unique minimal σ-algebra $\sigma(A)$ which contains all sets of A: a σ-algebra generated by A. $\sigma(A)$ is the intersection of all σ-algebras of which A is a subset.

Proof. Let $\mathcal{F}_A = \{\mathcal{F} : \mathcal{F}$ is a σ-algebra, $A \subseteq \mathcal{F}\}$. Then, $\sigma(A)$ is a σ-algebra since:

(i) $\Omega \in \mathcal{F} \ \forall \mathcal{F} \in \mathcal{F}_A$ since all $\mathcal{F} \in \mathcal{F}_a$ is a σ-algebra

(ii) $A \in \sigma(A) \Rightarrow A \in \mathcal{F} \ \forall \mathcal{F} \in \mathcal{F}_A \Rightarrow A^c \in \mathcal{F} \ \forall \mathcal{F} \in \mathcal{F}_A \Rightarrow A^c \in \sigma(A)$.

(iii) If $\{A_i\}_{i \in \mathbb{N}} \subseteq \sigma(A)$, then $\{A_i\}_{i \in \mathbb{N}} \subseteq \mathcal{F} \ \forall \mathcal{F} \in \mathcal{F}_A$.
So, $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F} \ \forall \mathcal{F} \in \mathcal{F}_A$, so $\bigcup_{i=1}^{\infty} A_i \in \sigma(A)$.

So $\sigma(A)$ is a σ-algebra. Now, $\sigma(A) \supseteq A$ since $\mathcal{F} \supseteq A \ \forall \mathcal{F} \in \mathcal{F}_A$. Also, $\forall$ σ-algebra $\mathcal{F}' \supset A$, we have $\mathcal{F}' \in \mathcal{F}_A$, so $\mathcal{F}' \supseteq \sigma(A)$.

Therefore, $\sigma(A)$ is the minimal σ-algebra containing A. □

Remark 1.7

Unless $|\Omega| < \infty$, a construction of $\sigma(A)$ is typically hopeless.

Example 1.8

$\mathcal{B}(\Omega) := \sigma(\{O : O \subseteq \Omega, O \text{ is open}\})$ is the Borel σ-algebra on Ω. Its elements are called Borel sets. For $\Omega = \mathbb{R}^d$ one can show that

$$\begin{aligned}
\mathcal{B}(\mathbb{R}^d) &= \sigma(\{(a, b] : a \leq b\}) \\
&= \sigma(\{(a, b) : a \leq b\}) \\
&= \sigma(\{[a, b] : a \leq b\}) \\
&= \sigma(\{(-\infty, b]\})
\end{aligned}$$

and so on. Borel sets contain open sets, closed sets, and countable union and intersections of these sets.

Definition 1.9: Measure Space

Let $\mathcal{F}$ be a σ-algebra on Ω. Then, $(\Omega, \mathcal{F})$ is a measurable space, sets in $\mathcal{F}$ are measurable sets. A measure μ on $\mathcal{F}$ is a function such that:

(i) $\mu : \mathcal{F} \to [0, \infty]$

(ii) $\mu(\varnothing) = 0$

(iii) Let $\{A_i\}_{i \in \mathbb{N}} \subseteq \mathcal{F}, A_i \cap A_j = \varnothing \; \forall i \neq j$. Then, $\mu(\dot{\bigcup}_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mu(A_i)$. ($\sigma$-additivity)

$(\Omega, \mathcal{F}, \mu)$ is then called a measure space.

If $\Omega = \bigcup_{i=1}^{\infty} A_i$ for $\{A_i\}_{i \in \mathbb{N}} \subseteq \mathcal{F} : \mu(A_i) < \infty \; \forall i$, then μ is σ-finite.

If $\mu(\Omega) < \infty$, μ is a finite measure.

A measure μ on $\mathcal{B}(\mathbb{R}^d)$ is a Borel measure on $\mathbb{R}^d$.

Remark 1.10

1. σ-additivity (in contrast to finite additivity) allows for limiting processes (pointwise limits of "measurable functions" are measurable). Many fundamental consequences follow, such as Central Limit Theorem and Law of Large Numbers.
2. Uncountable additivity is too strong, since for any $A \subseteq \mathbb{R}$:

$$\begin{aligned} \lambda(A) &= \lambda(\bigcup_{x \in A} \{x\}) \\ &= \sum_{x \in A} \lambda(\{x\}) \\ &= sup_{A' \subseteq A, |A'| < \infty} \sum_{x \in A} \lambda(\{x\}) \\ &= 0 \end{aligned}$$

Example 1.11

If Ω is countable, $\mathcal{F} = \mathcal{P}(\Omega)$, $\forall f : \Omega \to [0, \infty]$, $\mu(A) = \sum_{\omega \in A} f(\omega) \; \forall A \in \mathcal{F}$ defines a measure on $\mathcal{F}$.

If $f(x) = 1 \; \forall x \in \Omega$, μ is called a counting measure.

Suppose for some $\omega_0 \in \Omega, f(\omega) = 1$ if $\omega = \omega_0$, 0 otherwise. Then μ is point mass or Dirac measure.

Proposition 1.12

Let $(\Omega, \mathcal{F}, \mu)$ be a measure space. Then:

1. $A, B \in \mathcal{F}, A \subseteq B \Rightarrow \mu(\bigcup_{i=1}^{\infty} A_i) \leq \sum_{i=1}^{\infty} \mu(A_i)$ (monotonicity)
2. $\{A_i\}_{i \in \mathbb{N}} \subseteq \mathcal{F} \Rightarrow \mu(\bigcup_{i=1}^{\infty} A_i) \leq \sum_{i=1}^{\infty} \mu(A_i)$ (sub additivity)

3.

$$\{A_i\}_{i\in\mathbb{N}} \subseteq \mathcal{F}, A_1 \subseteq \cdots \subseteq A_n \subseteq \cdots \Rightarrow \mu(\bigcup_{i=1}^{\infty} A_i) = \mu(\lim_{n\to\infty} \bigcup_{i=1}^{n} A_i)$$
$$= \lim_{n\to\infty} \mu(A_n)$$

(continuity from below)

4.

$$\{A_i\}_{i\in\mathbb{N}} \subseteq \mathcal{F}, A_1 \supseteq \cdots \supseteq A_n \supseteq \cdots \Rightarrow \mu(\bigcap_{i=1}^{\infty} A_i) = \mu(\lim_{n\to\infty} \bigcap_{i=1}^{n} A_i)$$
$$= \lim_{n\to\infty} \mu(A_n)$$

for $\mu(A_1) < \infty$. (continuity from above)

Proof.

1. $B = A\dot{\cup}(B\backslash A)$
   Therefore, $\mu(B) = \mu(A) + \mu(B\backslash A) \geq 0$
   $\mu(B) \geq \mu(A)$ and if $\mu(A) < \infty$,
   $\mu(B\backslash A) = \mu(B) - \mu(A)$.

2. Let $B_1 = A_1$, $B_n = A_n\backslash\bigcup_{i=1}^{n-1} A_i \subseteq A_n$ $\forall n \geq 2$.
   So, all B_n are pairwise disjoint and $\bigcup_{i=1}^{n} B_i = \bigcup_{i=1}^{n} A_i$ $\forall n \in \mathbb{N}$.
   So, $\mu(\bigcup_{i=1}^{\infty} A_i) = \mu(\bigcup_{i=1}^{\infty} B_i) = \sum_{i=1}^{\infty} \mu(B_i) \leq \sum_{i=1}^{\infty} \mu(A_i)$.

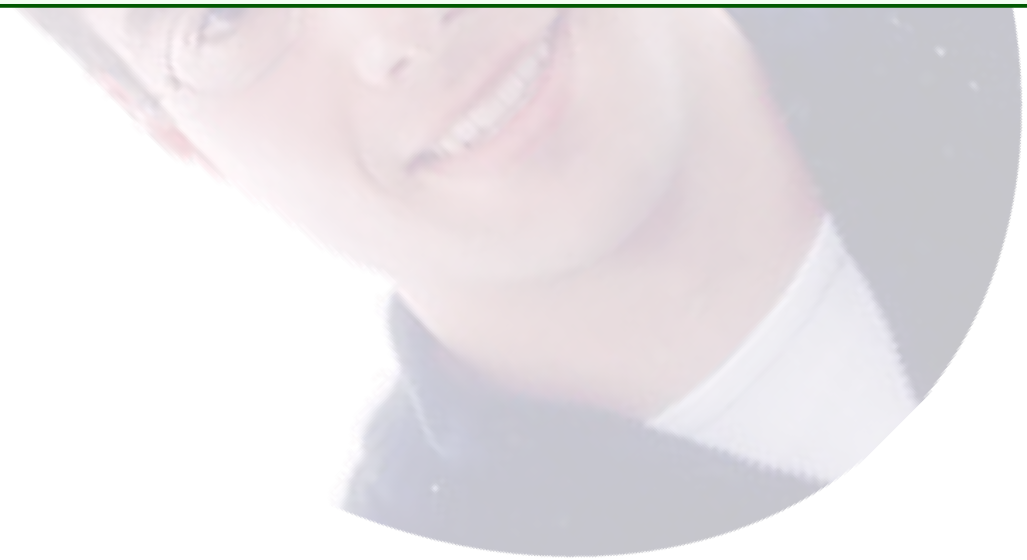
3. Let $A_0 = \varnothing$. Then,

$$\mu(\bigcup_{i=1}^{\infty} A_i) = \mu(\bigcup_{i=1}^{\infty}(A_i\backslash A_{i-1})$$
$$= \sum_{i=1}^{\infty} \mu(A_i\backslash A_{i-1})$$
$$= \lim_{n\to\infty} \sum_{i=1}^{n} \mu(A_i\backslash A_{i-1})$$
$$= \lim_{n\to\infty} \mu(A_n)$$

4. Let $B_i = A_1 \backslash A_i = A_1 \cap A_i^c \; \forall i \in \mathbb{N}$. Then, $B_1 \subseteq B_2 \subseteq \ldots$

$$\begin{aligned}
\bigcup_{i=1}^{\infty} B_i &= \bigcup_{i=1}^{\infty} (A_1 \cap A_i^c) \\
&= A_1 \cap \bigcup_{i=1}^{\infty} A_i^c \\
&= A_1 \cap (\bigcap_{i=1}^{\infty} A_i)^c \\
&= A_1 \backslash \bigcap_{i=1}^{\infty} A_i \Rightarrow \\
\mu(A_i) - \mu(\bigcap_{i=1}^{\infty} &= \mu(A_1 \backslash \bigcap_{i=1}^{\infty} A_i) \\
&= \mu(\bigcup_{i=1}^{\infty} B_i) \\
&= \lim n \to \infty \mu(B_n) \\
&= \lim_{n \to \infty} (\mu(A_1) - \mu(A_n)) \\
&= \mu(A_1) - \lim_{n \to \infty} \mu(A_n) \\
\therefore \mu(\bigcap_{i=1}^{\infty} &= \lim_{n \to \infty} \mu(A_n)
\end{aligned}$$

□



1.2 Probability Measures

Definition 1.13: Probability Measure

Let $(\Omega, \mathcal{F})$ be a measure space. Then, a probability measure P on $\mathcal{F}$ is a function such that:

(i) $P : \mathcal{F} \to [0, 1]$

(ii) $P(\Omega) = 1$

(iii) $\{A_i\}_{i\in\mathbb{N}} \subseteq \mathcal{F}, A_i \cap A_j = \varnothing \; \forall i \neq j \Rightarrow P(\dot{\bigcup}_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty}(A_i)$
(σ-additivity)

$(\Omega, \mathcal{F}, P)$ is a probability space.

Ω is a sample space.

$\omega \in \Omega$ is a sample point.

If Ω is countable/finite, then $(\Omega, \mathcal{F}, P)$ is discrete/finite.

Any $A \in \mathcal{F}$ is an event.

If $A = \{\omega\}$, A is a simple event.

Otherwise, A is a compound event.

Remark 1.14

If $(\Omega, \mathcal{F}, P)$ is discrete, $f(\omega) := P(\{\omega\}), \omega \in \Omega$ defines P via $P(A) = \sum_{\omega\in A} f(\omega), A \in \mathcal{F}$. Then, f is the probability mass function (pmf) on Ω.

Conversely, if Ω is countable, then in $(\Omega, \mathcal{P}(\Omega), P)$ with $P(A) := \sum_{\omega\in A} f(\omega)$, $A \in \mathcal{P}(\Omega)$, P defines a discrete probability measure for any $f : \Omega \to [0, 1]$ such that $\sum_{\omega\in A} f(\omega) = 1$.

Proposition 1.15

Let $(\Omega, \mathcal{F}, P)$ be a probability space. Then,

1. $A \in \mathcal{F} \Rightarrow P(A^c) = 1 - P(A)$
2. $A, B \in \mathcal{F} \Rightarrow P(A \cup B) = P(A) + P(B) - P(A \cap B)$
3. Let $\{A_i\}_{i\in\mathbb{N}} \subseteq \mathcal{F}$
   $S_{k,n} := \sum_{1\leq i_1<\cdots<i_k\leq n} P(A_{i_1} \cap \cdots \cap A_{i_k})$, for $k = 1, \ldots, n$

   Then, $P(\bigcup_{i=1}^{n} A_i) = \sum_{i=1}^{n}(-1)^{k-1} S_{k,n}$

   (inclusion-exclusion principle)

Proof.

1. $1 = P(\Omega) = P(A \dot{\cup} A^c) = P(A) + P(A^c)$

This implies probability measures are measures (take $A = \varnothing$).

2. $A \cup B = (A \backslash (A \cap B)) \dot{\cup} (A \cap B) \dot{\cup} (B \backslash (A \cap B)) \Rightarrow$

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

3. Induction on n. See Exercise 6 in Assignment 1.

□

1.3 Null Sets

Definition 1.16: Null Set

If $(\Omega, \mathcal{F}, \mu)$ is a measure space, every $N \in \mathcal{F}$ such that $\mu(N) = 0$ is a (μ-)null set.

If some property holds $\forall \omega \in \Omega \backslash N$ for null set N, it holds (μ-)almost everywhere.

Or, if μ is a probability measure, (μ-)almost surely.

If $\mathcal{F}$ contains all subsets of null sets, μ is complete.

By sub additivity, any countable union of null sets from $\mathcal{F}$ is a null set of $\mathcal{F}$.

Theorem 1.17: Completion

Let $(\Omega, \mathcal{F}, \mu)$ be a measure space, $\mathcal{N}$ be the set of all null sets.

1. $\bar{\mathcal{F}} := \{F \cup A : F \in \mathcal{F}, A \subseteq N, N \in \mathcal{N}\}$ is a σ-algebra on Ω.
2. $\bar{\mu}(F \cup A) = \mu(F)$ uniquely extends μ to a complete measure on $\bar{\mathcal{F}}$.

1.4 Construction of Measures

Idea: Functions with properties as measures (premeasures) defined on a ring can be extended to complete measures on the σ-algebra generated by the ring.

Definition 1.18: Rings

$\mathcal{R} \subseteq \mathcal{P}(\Omega)$ is a ring on Ω if:

1. $\varnothing \in \mathcal{R}$
2. $A, B \in \mathcal{R} \Rightarrow A \backslash B \in \mathcal{R}$
3. $A, B \in \mathcal{R} \Rightarrow A \cup B \in \mathcal{R}$

A premeasure μ_0 on $\mathcal{R}$ is a function with:

(i) $\mu_0 : \mathcal{R} \to [0, \infty]$

(ii) $\mu_0(\varnothing) = 0$

(iii) $\{A_i\}_{i\in\mathbb{N}} \subseteq \mathcal{R}, A_i \cap A_j = \varnothing\ \forall i \neq j, \dot{\bigcup}_{i=1}^{\infty} A_i \in \mathcal{R} \Rightarrow \mu(\dot{\bigcup}_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mu_0(A_i)$

Theorem 1.19: Caratheodory's extension theorem

If μ_0 is a premeasure on ring $\mathcal{R}$ on Ω, there exists a complete measure μ on $\mathcal{F} := \sigma(\mathcal{R})$ which coincides with μ_0 on $\mathcal{R}$. If μ_0 is σ-finite, then μ is unique.

Remark 1.20

The proof of 1.19 is constructive. The measure μ is constructed:

$$\begin{aligned} \mu(A) &= \inf_{A\subseteq\bigcup_{i=1}^{\infty} A_i, A_i\in\mathcal{R}\ \forall i} \sum_{i=1}^{\infty} \mu_0(A_i) \\ &= A_i \in \mathcal{R}\ \forall i \end{aligned}$$

Theorem 1.21

If $F : \mathbb{R} \to \mathbb{R}$ is right-continuous and increasing ($F(x) \leq F(y)\ \forall x < y$), there exists exactly one Borel measure μ_F such that:

$\mu_F((a, b]) = F(b) - F(a)\ \forall a \leq b$

Proof. $\mathcal{R} := \{\dot{\bigcup}_{k=1}^{n} (a_k, b_k] : -\infty < a_k \leq b_k < \infty\ \forall k, n \in \mathbb{N}\}$ is a ring on $\mathbb{R}$, and $\mu_0(\dot{\bigcup}_{k=1}^{n}(a_k, b_k]) := \sum_{k=1}^{n}(F(b_k) - F(a_k))$ is a premeasure on $\mathcal{R}$. By 1.19, there exists exactly one measure μ_F on $\sigma(\mathbb{R}) = \mathcal{B}(\mathbb{R})$ such that $\mu_F|_{\mathcal{R}} = \mu_0$ (i.e. $\mu_F(A) = \mu_0(A)\ \forall A \in \mathbb{R}$). □

Remark 1.22

1. By 1.19, μ_F is complete, and called the Lebesgue-Stietjes measures associated to F. Its domain, the completion $\bar{\mathcal{B}}(\mathbb{R})$, the Lebesgue σ-algebra, can be shown to strictly contain $\mathcal{B}(\mathbb{R})$. Sets in $\bar{\mathcal{B}}(\mathbb{R}$ are called Lebesgue measurable or Lebesgue sets. By our construction,

$$\mu_F(A) = \inf_{A\subseteq\bigcup_{i=1}^{\infty}(a_i, b_i]} \sum_{i=1}^{\infty} \mu_F((a, b])$$

2. If $F(x) = x$, $\lambda := \mu_F$ is a Lebesgue measure on $\mathbb{R}$. Sets $N \subseteq \bar{\mathcal{B}}(\mathbb{R})$ that are null sets are Lebesgue null sets, $\lambda(N) = 0$.

   By 1.17 (1), $B \in \bar{\mathcal{B}}(\mathbb{R}) \Leftrightarrow B = A \cup N$.

Example 1.23

1. $\{x\} \subseteq \mathbb{R}$ is a null set for all $x \in \mathbb{R}$, since

$$\begin{aligned}\lambda(\{x\}) &= \lambda\big(\bigcap_{n=1}^{\infty}(x - \frac{1}{n}, x]\big)\\ &= \lim_{n\to\infty} \lambda\big((x - \frac{1}{n}, x]\big)\\ &= 0\end{aligned}$$

2. $\mathbb{Q} \subseteq \mathbb{R}$ is a null set since

$$\begin{aligned}\lambda(\mathbb{Q}) &= \lambda(\dot{\bigcup}_{i=1}^{\infty}\{q_i\})\\ &= \sum_{i=1}^{\infty} 0\\ &= 0\end{aligned}$$

3. Cantor set: $C = \bigcap_{i=1}^{\infty} C_i$ where C_i is defined by:

$$\begin{aligned}&C_0 = [0, 1]\\ &C_1 = [0, \frac{1}{3}] \cup [\frac{2}{3}, 1]\\ &C_2 = [0, \frac{1}{9}] \cup [\frac{2}{9}, \frac{1}{3}] \cup [\frac{2}{3}, \frac{7}{9}] \cup [\frac{8}{9}, 1]\\ &\dots\\ &C_i = \frac{C_{i-1}}{3} \cup (\frac{2}{3} + \frac{C_{i-1}}{3})\ \forall i \geq 1\end{aligned}$$

   By Cantor's diagonal argument, C is uncountable. Since $\lambda([0,1]\backslash C) = 2^0\frac{1}{3} + 2^1\frac{1}{9} + \cdots = \sum_{i=1}^{\infty} 2^{i-1}3^{-i} = 1$, therefore $\lambda(C) = 0$.

Remark 1.24

1.21 extends to $F : \mathbb{R}^d \to \mathbb{R}$ which is:

(i) right continuous: $F(\underline{x}) = \lim_{\underline{h}\downarrow 0} F(\underline{x} + \underline{h}) =: F(\underline{x}+)\ \forall x \in \mathbb{R}^d$

(ii) d-increasing: The F-volume $\Delta_{(\underline{a},\underline{b}]}F$ of $(a,b] \geq 0$ for $\underline{a} \leq \underline{b}$, where:

$$\begin{aligned}\Delta_{(\underline{a},\underline{b}]}F :&= \sum_{i\in\{0,1\}^d} (-1)^{\sum_{j=1}^{d} ij} F(a_1^{i_1} b_1^{1-i_1}, \ldots, a_d^{i_d} b_d^{1-i_d}) \\ &= \prod_{j=1}^{d} (b_j - a_j) \\ &= \lambda((a,b])\end{aligned}$$

(iii) If, additionally, $\lim_{x_j \downarrow -\infty} F(\underline{x}) = 0$ for some $j \in \{1, \ldots, d\}$ and $F(\underline{\infty}) = lim_{\underline{x}\uparrow\infty} F(x) = 1$, then μ_F is a probability measure on $\mathcal{B}(\mathbb{R}^d)$. Then, $\Delta_{(a,b]}F$ is the probability of $(\underline{a}, \underline{b}]$.

2 Geometric and Laplace probability spaces

Proposition 2.1

Let $(\Omega, \mathcal{F}, \mu)$ be a measure space such that $0 < \mu(\Omega) < \infty$. Then, $(\Omega, \mathcal{F}, P)$ with $P(A) = \frac{\mu(A)}{\mu(\Omega}$ $\forall A \in \mathcal{F}$, is a probability space.

Proof.

(i) $0 \leq \mu(A) \leq \mu(\Omega) \leq \infty \; \forall A \in \mathcal{F} \Rightarrow P : \mathcal{F} \to [0, 1]$

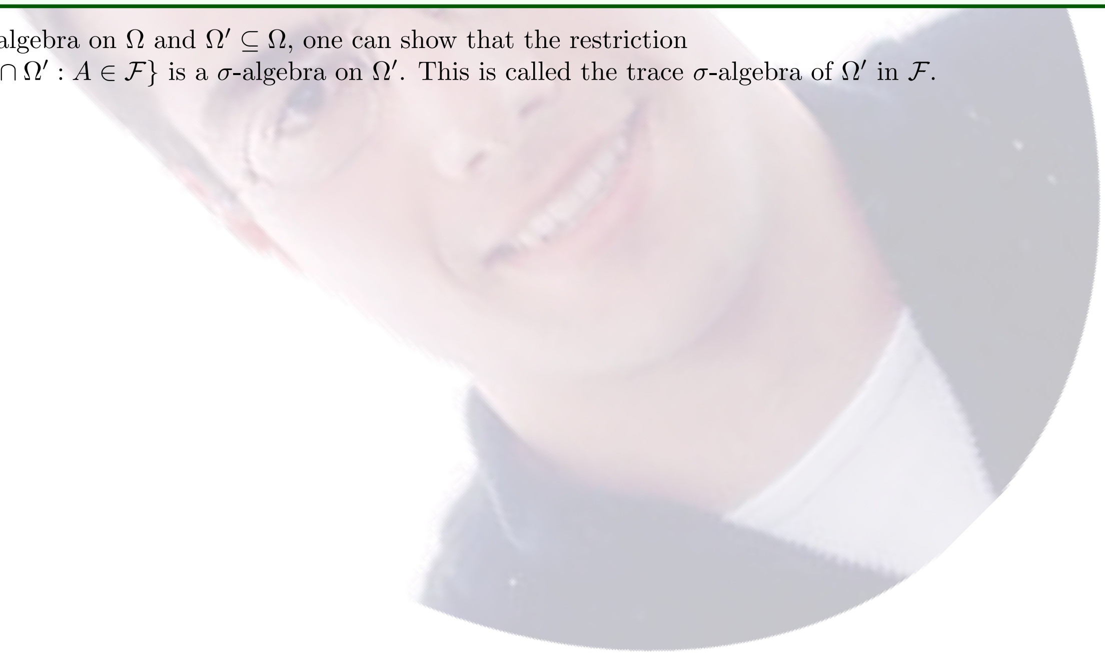
(ii) $P(\Omega) = \frac{P(\Omega)}{P(\Omega)} = 1$

(iii) $\{A_i\}_{i \in \mathbb{N}} \subseteq \mathcal{F}, A_i \cap A_j = \varnothing \; \forall i \neq j \Rightarrow$

$$\begin{aligned} P(\bigcup_{i=1}^{\infty} A_i) &= \frac{\mu(\bigcup_{i=1}^{\infty} A_i)}{\mu(\Omega} \\ &= \sum_{i=1}^{\infty} \frac{\mu(A_i)}{\mu(\Omega} \\ &= \sum_{i=1}^{\infty} P(A_i) \end{aligned}$$

□

If $\mathcal{F}$ is a σ-algebra on Ω and $\Omega' \subseteq \Omega$, one can show that the restriction $\mathcal{F}|_{\Omega'} := \{A \cap \Omega' : A \in \mathcal{F}\}$ is a σ-algebra on Ω'. This is called the trace σ-algebra of Ω' in $\mathcal{F}$.



2.1 Geometric Probability Spaces

Definition 2.2: Geometric Probability Space

If:

$$\begin{aligned} &\Omega \subseteq \mathbb{R}^d : 0 < \lambda(\Omega) < \infty \\ &\mathcal{F} = \mathcal{B}(\Omega) \\ P(A) = &\frac{\lambda(A)}{\lambda(\Omega)} \ \forall A \in \mathcal{F} \end{aligned}$$

then the probability space $(\Omega, \mathcal{F}, P)$ is a geometric probability space.

Example 2.3

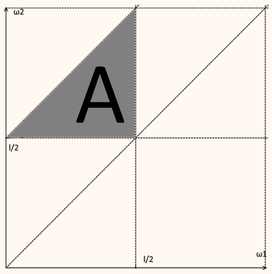
A stick length l is randomly marked and cut at 2 spots. Find the probability that the 3 pieces can form a triangle.

Solution. Let

$$\begin{aligned} \Omega &= \{(\omega_1, \omega_2) \in [0,l]^2, \omega_1 < \omega_2\} \\ \mathcal{F} &= \bar{\mathcal{B}}(\Omega) \\ P(A) &= \frac{\lambda(A)}{\lambda(\Omega)} \end{aligned}$$

We are interested in the set A: "each side < sum of other 2", or

$$\begin{aligned} A = &\{(\omega_1, \omega_2) \in \Omega : \\ &\omega_1 < (\omega_2 - \omega_1) + (l - \omega_2), \\ &\omega_2 - \omega_1 < \omega_1 + l - \omega_2, \\ &l - \omega_2 < \omega_1 + \omega_2 - \omega_1\} \\ = &\{(\omega_1, \omega_2) \in \Omega : \omega_1 < \frac{l}{2}, \frac{l}{2} < \omega_2 < \omega_1 + \frac{l}{2}\} \end{aligned}$$



$P(A) = \frac{1}{4}$, from the picture.

2.2 Laplace Probability Spaces

Here is a similar construction, based on the number $|\Omega|$ of elements in Ω.

Proposition 2.4

Let $1 \leq |\Omega| < \infty$, $\mathcal{F} = \mathcal{P}(A), P(A) = \frac{|A|}{|\Omega|}$ $\forall A \in \mathcal{F}$. Then, $(\Omega, \mathcal{F}, P)$ is a finite probability space called a Laplace probability space.

P is discrete uniform distribution on Ω.

Proof. Apply Prop. 2.1 with $\mu(A) = |A|$ (counting measure). □

Remark 2.5

For Laplace probability spaces, probability mass function on Ω is

$f(\omega) = P(\{\omega\}) = \frac{|\{\omega\}|}{|\Omega|} = \frac{1}{|\Omega|}$ $\forall \omega \in \Omega$
so the discrete uniform distribution assigns equal probability $\frac{1}{|\Omega|}$ to each $\omega \in \Omega$.

Example 2.6

1. Determine probability of obtaining 1 or 5 when rolling a fair, 6-sided die.

$$\Omega = \{1, \dots, 6\}$$
$$\mathcal{F} = \mathcal{P}(\Omega)$$
$$P(A) = \frac{|A|}{|\Omega|} \; \forall A \in \Omega$$

   Let $A =$“rolling 1 or 5”$= \{1, 5\}$. Then, $P(A) = \frac{2}{6} = \frac{1}{3}$.

2. Determine probability of obtaining a sum of 2 and 7 when rolling the die twice.

$$\Omega = \begin{array}{ccc} \{(1,1), & \dots, & (1,6), \\ & \ddots & \\ (6,1), & \dots, & (6,6)\} \end{array}$$
$$\mathcal{F} = \mathcal{P}(\Omega)$$
$$P(A) = \frac{|A|}{|\Omega|} \; \forall A \in \mathcal{F}$$

   So,

   $P(\text{“sum is 2”}) = \frac{1}{36}$

   $P(\text{“sum is 7”}) = \frac{6}{36} = \frac{1}{6}$

3. Determine probability of obtaining at least one 6 when rolling 3 times.

$$\Omega = \{(\omega_1, \omega_2, \omega_3) | \omega_i \in \{1, \dots, 6\} \; \forall i\}$$
$$\mathcal{F} = \mathcal{P}(\Omega)$$
$$P(A) = \frac{|A|}{|\Omega|} \; \forall A \in \mathcal{F}$$

So, $P(\text{“at least one 6”}) = 1 - P(\text{“no 6s”}) = 1 - (\frac{5}{6})^3 = \frac{91}{216}$

3 Probability Counting Techniques

3.1 Basic Rules

Proposition 3.1

1. If $A_1, \ldots, A_n$ are pointwise disjoint finite sets, then $|\bigcup_{i=1}^n A_i| = \sum_{i=1}^n |A_i|$ (addition rule).

2. If $A_1, \ldots, A_n$ are finite sets, then $|\prod_{i=1}^n A_i| = \prod_{i=1}^n |A_i|$ (multiplication rule).

Proof. By induction. □

Example 3.2

Consider an urn with 5 balls labelled 1, ..., 5. Determine the probability of obtaining precisely 1 even ball when drawing twice with replacement. Let $\mathbb{E}$ be the set of even integers.

$$\begin{aligned}
\Omega &= \begin{array}{ccc} \{(1,1), & \ldots, & (1,5), \\ & \ddots & \\ (5,1), & \ldots, & (5,5)\} \end{array} \\
\mathcal{F} &= \mathcal{P}(\Omega) \\
A &= \{(\omega_1, \omega_2) \in \Omega : \omega_1 + \omega_2 \notin \mathbb{E}\} \\
&= A_1 \dot{\cup} A_2
\end{aligned}$$

where

$$\begin{aligned}
A_1 &= \{(\omega_1, \omega_2) : \omega_1 \in \mathbb{E}, \omega_2 \notin \mathbb{E}\} \\
A_2 &= \{(\omega_1, \omega_2) : \omega_1 \notin \mathbb{E}, \omega_2 \in \mathbb{E}\} \\
\therefore P(A) &= \frac{|A|}{|\Omega|} \\
&= \frac{|A_1| + |A_2|}{|\Omega|} \\
&= \frac{2 * 3 + 3 * 2}{25} \\
&= \frac{12}{25}
\end{aligned}$$

3.2 Urn Models

Many counting problems can be associated with drawing k balls from an urn with n balls. Classical models consider drawing:

(I) With order, with replacement

(II) With order, without replacement

(III) Without order, without replacement

(IV) Without order, with replacement

What are the number of possibilities in each of the four setups?

(I)

$$\begin{aligned}\Omega_I &= \{(\omega_1, \dots, \omega_k) : \omega_i \in \{1, \dots, n\}, i \in \{1, \dots, k\}\} \\ &= \prod_{i=1}^{k} \{1, \dots, n\} \\ &= \{1, \dots, n\}^k \\ \Rightarrow |\Omega_I| &= n^k\end{aligned}$$

Example

(a) The number of 53 digit numbers containing only 0-1 is $2^{53} \approx 9 \cdot 10^{15}$.

(b) The number of functions from $A \to B$, where $|A| = k, |B| = n$ is n^k.

(II)

$$\begin{aligned}\Omega_{II} &= \{(\omega_1, \dots, \omega_k) : \omega_i \in \{1, \dots, n\}, \omega_i \neq \omega_j \forall i \neq j\} \\ \Rightarrow |\Omega_{II}| &= n(n-1)(n-2)\dots(n-k+1) \\ &=: (n)_k\end{aligned}$$

Where $(n)_k$ is the “falling factorial” and is read “n to k factors”.

Example

(a) The number of 3 digit numbers with unique digits in $\{1, \dots, 9\}$ is $(9)_3 = 9 \cdot 8 \cdot 7 = 504$.

(b) The number of injective functions from $A \to B$, where $|A| = k, |B| = n$ is $(n)_k$.

(c) If $k = n$, then $(n)_k = n!$, and $0! = 1! = 1$.

$$n! \sim \left(\frac{n}{e}\right)^n \sqrt{2\pi n}$$

(Stirling’s formula)

(III)

$$\Omega_{III} = \{(\omega_1, \dots, \omega_k) : \omega_i \in \{1, \dots, n\} \forall i, \omega_1 < \dots < \omega_k\}$$

Definition: Equivalence Relation

An equivalence relation $\sim$ is a relation on some set S such that:

(i) $x \sim x$ for all $x \in S$. (reflexive)

(ii) $x \sim y \Leftrightarrow y \sim x$ for all $x, y \in S$. (symmetric)

(iii) $x \sim y$ and $y \sim z \Rightarrow x \sim z$ for all $x, y, z \in S$. (transitive)

An equivalence class of some $a \in S$ is $\{x \in S : a \sim x\}$.

Now, define an equivalence relation $\sim$ on Ω via $(\omega_1, \dots, \omega_k) \sim (\omega'_1, \dots, \omega'_k)$ iff there exists a permutation $\pi : \{1, \dots, k\} \to \{1, \dots, k\}$ such that
$(\omega_1, \dots, \omega_k) \sim (\omega'_{\pi(1)}, \dots, \omega'_{\pi(k)})$. Then, Ω_{II} consists of ordered representations of the equivalence classes of $\sim$, and each such class has n! elements. Thus, $|\Omega_{II}| = |\Omega_{III}|k!$, so $|\Omega_{III}| = \binom{n}{k}$.

Example

(a) Lotto 6/49 draws 6 from 49 without order and without replacement. Therefore, there are $\binom{49}{6}$ possible outcomes and the chance of some ticket winning is $\frac{1}{\binom{49}{6}} \approx 7.15 \cdot 10^{-8}$

(b) How many subsets of size k does a set of size n have? $\binom{n}{k}$. Therefore,

$$\sum_{k=0}^{n} \binom{n}{k} = 2^n$$

Note that $\binom{n}{k} = \frac{n!}{k!(n-k)!} = \frac{n!}{(n-k)!(n-(n-k))!} = \binom{n}{n-k}$.

(IV) Here, we can't simply use (I) and divide by k! (eg. If the 2^{nd} ball we draw equals (does not equal) the first, then the permutations don't need to be (do need to be) considered. Identify by distinguishable permutations of $n-1$ "1" and k "0".
$\Rightarrow |\Omega_{IV}| = \frac{number\ of\ n-1+k\ symbols}{(number\ of\ permutations\ of\ n-1\ \text{“1”}s)(number\ of\ permutations\ of\ k\ \text{“0”}s)} = \binom{n-1+k}{k}$

Formally, $\Omega_{IV} = \{(\omega_1, \dots, \omega_k) : \omega_i \in \{1, \dots, n\}\ \forall i, \omega_1 \leq \dots \leq \omega_k\}$. Note that if $f(\omega_1, \dots, \omega_k) = (\omega_1, \omega_2 + 1, \dots, \omega_k + k - 1)$ is a bijection from Ω_{IV} to $\Omega'_{III} = \{(\omega_1, \dots, \omega_k) : \omega_i \in \{1, \dots, n+k-1\}\ \forall i, \omega_1 < \dots < \omega_k\}$.
$\Rightarrow |\Omega_{IV}| = |\Omega'_{III}| = \binom{n+k-1}{k}$.

Example

(a) How many possible domino stones are there? A domino has two squares, each of which can be contain 0-6 dots.
$n = 7, k = 2 \Rightarrow \binom{n+k-1}{k} = 28$

(b) How many different partial derivatives $\frac{\partial^k}{\partial x_{jk} \dots \partial x_{j1}} f$ of $f \in C^k(\mathbb{R}^n)$ exist?

By Scwartz' or Clairaut's theorem, order doesn't matter. Furthermore, we can differentiate with respect to same variable multiple times (so there is replacement).
$\Rightarrow \exists! \binom{n+k-1}{k}$ different partial derivatives.

4 Conditional Probability and Independence

Proposition 4.1

Let $(\Omega, \mathcal{F}, P)$ be a probability space, and $B \in \mathcal{F} : P(B) > 0$. Then, $P(A|B) := \frac{P(A \cap B)}{P(B)}, A \in \mathcal{F}$ is a probability measure on $(\Omega, \mathcal{F})$, called the ordinary conditional probability of A given B. (The vertical bar in the expression $P(A|B)$ means "given".)

Proof.

(i) Let $A \in \mathcal{F}$.

$$\begin{aligned} 0 &\leq P(A \cap B) \leq P(B) \\ &\Rightarrow \frac{P(A \cap B)}{P(B)} \leq 1 \\ &\Rightarrow P(.|B) : \mathcal{F} \to [0,1] \end{aligned}$$

(ii) $P(\Omega|B) = \frac{P(\Omega \cap B)}{P(B)} = \frac{P(B)}{P(B)} = 1$

(iii) If $\{A_i\}_{i \in \mathbb{N}} \in \mathcal{F}, A_i \cap A_j = \varnothing \ \forall i \neq j$, then

$$\begin{aligned} P(\bigcup_{i=1}^{\infty} A_i|B) &= \frac{P((\bigcup_{i=1}^{\infty} A_i) \cap B)}{P(B)} \\ &= \frac{\sum_{i=1}^{\infty} P(A_i \cap B)}{P(B)} \\ &= \sum_{i=1}^{\infty} \frac{P(A_i \cap B)}{P(B)} \\ &= \sum_{i=1}^{\infty} P(A_i|B) \end{aligned}$$

□

Although $P(A|B)$ is only defined if $P(B) > 0$, the convention $P(A|B)P(B) = P(A \cap B) \leq P(B)$ makes sense for any definition of $P(A|B) \in [0,1]$ if $P(B) = 0$.

Theorem 4.2: Law of Total Probability

Let $(\Omega, \mathcal{F}, P)$ be a probability space, $\{B_i\}_{i \in \mathbb{N}} \subseteq \mathcal{F}$ be a partition of Ω. That is,

$$\Omega = \dot{\bigcup}_{i=1}^{\infty} B_i, B_i \cap B_j = \varnothing \ \forall i \neq j$$

Then,

$$\begin{aligned} A \in \mathcal{F} \Rightarrow P(A) &= \sum_{i=1}^{\infty} P(A \cap B_i) \\ &= \sum_{i=1}^{\infty} P(A|B_i)P(B_i) \end{aligned}$$

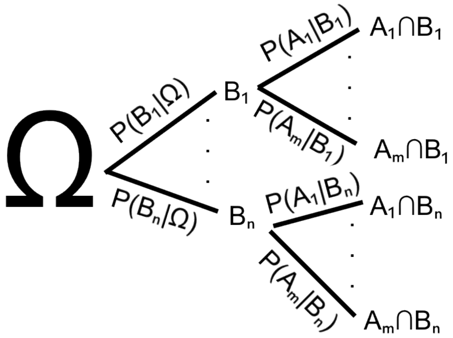
Proof.

$$\begin{aligned} A \in \mathcal{F} \Rightarrow P(A) &= P(A \cap \Omega) \\ &= P(A \cap \bigcup_{i=1}^{\infty} B_i) \\ &= P(\bigcup_{i=1}^{\infty} (A \cap B_i)) \\ &= \sum_{i=1}^{\infty} P(A \cap B_i) \\ &= \sum_{i=1}^{\infty} P(A|B_i)P(B_i) \end{aligned}$$

□

Remark 4.3

It is often helpful to visualize conditional probabilities in a tree diagram. For example, if $\{A_i\}_{i \in \mathbb{N}}^m, \{B_i\}_{i \in \mathbb{N}}^n \subseteq \mathcal{F}$ are permutations of $\mathcal{F}$,



$$\begin{aligned} P(A_1 \cap B_1) &= P(A_1|B_1)P(B_1) \\ P(A_m \cap B_1) &= P(A_m|B_1)P(B_1) \\ P(A_1 \cap B_n) &= P(A_1|B_n)P(B_n) \\ P(A_m \cap B_n) &= P(A_m|B_n)P(B_n) \end{aligned}$$

$$\begin{aligned} \sum_{i,j=1}^{m,n} P(A_i \cap B_j) &= \sum_{i,j=1}^{m,n} P(A_i|B_j)P(B_j) \\ &= \sum_{i=1}^{m} \sum_{j=1}^{n} P(A_i|B_j)P(B_j) \\ &= P(\bigcup_{i=1}^{m} A) \\ &= P(\Omega) \\ &= 1 \end{aligned}$$

Example 4.4

1. Consider an urn with 2 dice, where die_1 has sides (1, 1, 1, 6, 6, 6) and die_2 has sides (1, 2, 3, 4, 5, 6). Draw a die at random and roll it. What is $P(\text{“}roll\ 6\text{”})$?

   Let $\Omega = \{(1,1),(1,6),(2,1),\dots,(2,6)\}$, $\mathcal{F} = \mathcal{P}(\Omega)$. Define P as:

$(\omega_1, \omega 2)$	(1, 1)	(1, 6)	(2, 1)	...	(2, 6)
$P(\{(\omega_1,\omega_2)\})$	$\frac{1}{2}\frac{1}{2}=\frac{1}{4}$	$\frac{1}{2}\frac{1}{2}=\frac{1}{4}$	$\frac{1}{2}\frac{1}{6}=\frac{1}{12}$	$\frac{1}{2}\frac{1}{6}=\frac{1}{12}$	$\frac{1}{2}\frac{1}{6}=\frac{1}{12}$

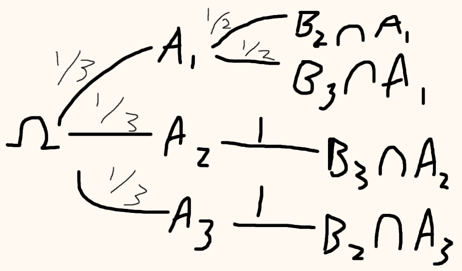
   Let $A = \text{“}Rolling\ 6\text{”} = \{(\omega_1,\omega_2) \in \Omega : \omega_2 = 6\}$
   $B = \text{“}Drawing\ die\ 1\text{”} = \{(\omega_1,\omega_2) \in \Omega : \omega_1 = 1\}$

$$\begin{aligned} P(A) &= P(A|B)P(B) + P(A|B^c)P(B^c) \\ &= \frac{1}{2}\frac{1}{2} + \frac{1}{6}\frac{1}{2} \\ &= \frac{1}{3} \end{aligned}$$

2. Monty Hall problem.

   Suppose you can choose between 3 doors. Behind one is a car, behind the others are goats. You pick a door, wlog, $door_1$.
   The host then, possibly randomly, opens one of the other doors, with a goat behind it. He then gives you a chance to switch to the remaining door. Should you (assuming you want a car more than you want a goat)?

   Let $\Omega = \{(\omega_1,\omega_2) : \omega_1 \in \{1,2,3\}, i = 1,2\}$, where ω_1 represents the door containing a car and ω_2 represents the door the host reveals.
   Let $\mathcal{F} = \mathcal{P}(\Omega)$
   Let $A_i = \{(\omega_1,\omega_2) \in \Omega : \omega_1 = i\}, i = 1,2,3$
   $B_j = \{(\omega_1,\omega_2) \in \Omega : \omega_2 = j\}, i = 2,3$



$$\begin{aligned} \Rightarrow P(A_1 \cap B_2)n &= P(B_2|A_1)P(A_1) = \frac{1}{2}\frac{1}{3} = \frac{1}{6} \\ P(A_1 \cap B_3)n &= P(B_3|A_1)P(A_1) = \frac{1}{2}\frac{1}{3} = \frac{1}{6} \\ P(A_2 \cap B_3)n &= P(B_3|A_2)P(A_2) = 1\frac{1}{3} = \frac{1}{3} \\ P(A_3 \cap B_2)n &= P(B_2|A_3)P(A_3) = 1\frac{1}{3} = \frac{1}{3} \end{aligned}$$

$$\Rightarrow P(\{(\omega_1,\omega_2)\}) = \begin{cases} \frac{1}{6} & \text{if } (\omega_1,\omega_2) = (1,2) \\ \frac{1}{6} & \text{if } (\omega_1,\omega_2) = (1,3) \\ \frac{1}{3} & \text{if } (\omega_1,\omega_2) = (2,3) \\ \frac{1}{3} & \text{if } (\omega_1,\omega_2) = (3,2) \\ 0 & \text{otherwise} \end{cases}$$

So, $P(\text{winning when not switching}) = P(\{(\omega_1,\omega_2) \in \Omega : \omega_1 = 1\}) = \frac{1}{6} + \frac{1}{6} = \frac{1}{3}$
$P(\text{winning when switching}) = P(\{(2,3),(3,2)\}) = \frac{1}{3} + \frac{1}{3} = \frac{2}{3}$ $\therefore$ switching is better.

We can also see this from

$$\begin{aligned} P(A_2|B_3) &= \frac{P(A_2 \cap B_3)}{P(B_3)} \\ &= \frac{P(A_2 \cap B_3}{P(A_1 \cap B_3) + P(A_2 \cap B_3) + P(A_3 \cap B_3)} \\ &= \frac{\frac{1}{3}}{\frac{1}{6} + \frac{1}{3} + 0} \\ &= \frac{2}{3} \end{aligned}$$

Which also equals $P(A_3|B_2)$ by a similar argument.

Theorem 4.5: Bayes' Theorem

Let $(\Omega, \mathcal{F}, P)$ be a probability space,
$A \in \mathcal{F}$ such that $P(A) > 0$.
Then, $P(B|A) = \frac{P(A|B)P(B)}{P(A)}$ for $A, B \in \mathcal{F}$. Also, if $\{B_i\}_{i \in \mathbb{N}} \subseteq \mathcal{F}$ is a partition of Ω, then

$$\begin{aligned} P(B_i|A) &= \frac{P(A_i|B)P(B_i)}{P(A)} \\ &= \frac{P(A|B_i)P(B_i)}{\sum_{j=1}^{\infty} P(A|B_j)P(B_j)} \end{aligned}$$

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

Example 4.6

A phone sends 0s and 1s (ratio = 3:2) to an antenna. With a certain probability $p \in (0,1)$, 0s are received wrongly as 1s. With a certain probability $q \in (0,1)$, 1s are received wrongly as 0s.

(a) Find probability that a 1 is received.

(b) Find probability that a 1 has been sent, given a 1 has been received.

Solution.

(a) Let $\Omega = \{\omega_1, \omega_2) : \omega_i \in \{0,1\}, i = 1,2\}$
$\mathcal{F} = \mathcal{P}(\Omega)$
$S_i =$"i sent"$= \{(i,0),(i,1)\}$ $i = 0,1$

$R_i =$"i received"$= \{(0,i),(1,i)\}$ $i = 0,1$. Then,

$$\begin{aligned} P(R_1) &= P(R_1|S_0)P(S_0) + P(R_1|S_1)P(S_1) \\ &= p\frac{3}{5} + (1-q)\frac{2}{5} \\ &= 0.6p + 0.4(1-q) \end{aligned}$$

(b)

$$\begin{aligned} P(S_1|R_1) &= \frac{P(R_1|S_1)P(S_1)}{P(R_1)} \\ &= \frac{0.4(1-q)}{0.6p+0.4(1-q)} \end{aligned}$$

From (a).

If $P(A|B)$ does not depend on B, then

$$P(A \cap B) = P(A|B)P(B) = P(A)P(B)$$

These events are independent.

Definition 4.7: Independence

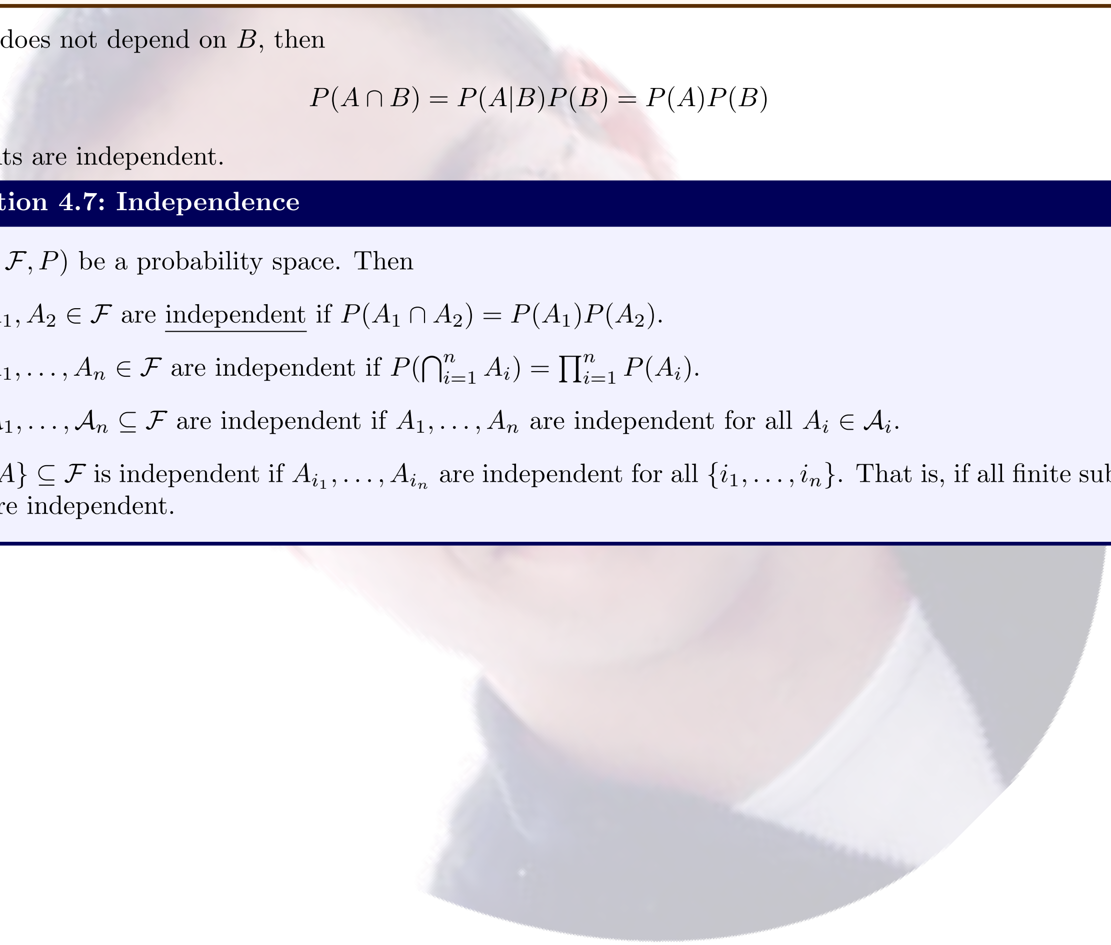
Let $(\Omega, \mathcal{F}, P)$ be a probability space. Then

(i) $A_1, A_2 \in \mathcal{F}$ are independent if $P(A_1 \cap A_2) = P(A_1)P(A_2)$.

(ii) $A_1, \ldots, A_n \in \mathcal{F}$ are independent if $P(\bigcap_{i=1}^n A_i) = \prod_{i=1}^n P(A_i)$.

(iii) $\mathcal{A}_1, \ldots, \mathcal{A}_n \subseteq \mathcal{F}$ are independent if $A_1, \ldots, A_n$ are independent for all $A_i \in \mathcal{A}_i$.

(iv) $\{A\} \subseteq \mathcal{F}$ is independent if $A_{i_1}, \ldots, A_{i_n}$ are independent for all $\{i_1, \ldots, i_n\}$. That is, if all finite subsets are independent.



5 Random Variables and Distributions

When predicting outcomes of experiments, it is useful to consider mappings $X : \Omega \to \Omega'$ for some measurable set $(\Omega', \mathcal{F}')$, often $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ or $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$.

5.1 Random Variables

Definition 5.1: Preimage

The preimage of $X : \Omega \to \Omega'$ is defined by $X^{-1}(A') = \{\omega \in \Omega : X(\omega) \in A'\}, A' \subseteq \Omega'$.

Lemma 5.2

(i) $X^{-1}(\varnothing) = \varnothing, X^{-1}(\Omega') = \Omega$

(ii) $(X^{-1}(A'))^c = X^{-1}(A'^c) \;\forall A' \subseteq \Omega'$

(iii) $\bigcup_{i\in I} X^{-1}(A'_i) = X^{-1}(\bigcup_{i\in I} A'_i)$,
$\bigcap_{i\in I} X^{-1}(A'_i) = X^{-1}(\bigcap_{i\in I} A'_i)$

If $\mathcal{F}$ is a σ-algebra on Ω', then $\sigma(X) := \{X^{-1}(A') : A' \in \mathcal{F}\}$ is a σ-algebra on Ω, the σ-algebra generated by X.

Proof.

(i)

(ii)

$$\begin{aligned} &\omega \in (X^{-1}(A'))^c \\ \Leftrightarrow &\omega \notin X^{-1}(A') \\ \Leftrightarrow &X(\omega) \notin A' \\ \Leftrightarrow &X(\omega) \in A'^c \\ \Leftrightarrow &\omega \in X^{-1}(A'^c) \end{aligned}$$

(iii) Similar to (ii).

□

Definition 5.3

Let $(\Omega, \mathcal{F}), (\Omega', \mathcal{F}')$ be measurable. Then, $X : \Omega \to \Omega'$ is called $((\mathcal{F}, \mathcal{F}')-)$ measurable if $\sigma(X) \subseteq \mathcal{F}$. I.e. $X^{-1}(A') \in \mathcal{F} \;\; \forall A' \in \mathcal{F}$.

If $(\Omega', \mathcal{F}')$ is $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ or $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, X is called a random variable (ω) or random vector in this case.

Example 5.4

1. Let $(\Omega, \mathcal{F})$ be measurable and $V \subseteq \Omega : V \notin \mathcal{F}$ be a non-measurable (Vitali) set. Then, $X : \Omega \to \mathbb{R}$ (so $(\Omega', \mathcal{F}') = (\mathbb{R}, \mathcal{B}(\mathbb{R})))$ with
$$X(\omega) = 1_v(\omega) = \begin{cases} 1 \text{ if } \omega \in V \\ 0 \text{ if } \omega \notin V \end{cases}$$
is a non-measurable function since
$$X^{-1}((\frac{1}{2}, 1]) = \{\omega \in \Omega : X(\omega) \in (\frac{1}{2}, 1]\} = \{\omega \in \Omega : X(\omega) = 1\} = V \notin \mathcal{F}$$

2. If $\Omega \neq \varnothing$, no non-constant $X^{-} : \Omega \to \mathbb{R}$ is $(\{\varnothing, \Omega\}, \mathcal{B}(\mathbb{R}))$-measurable.

3. If $(\Omega, \mathcal{F})$ is a measurable space, the following functions from $\Omega \to \mathbb{R}$ are measurable:
   (i) Let $c \in \mathbb{R}, X = c$. Then,
$$X^{-1}(B) = \begin{cases} \Omega, & c \in \mathcal{B} \\ \varnothing, & c \notin \mathcal{B} \end{cases}$$
$$\Rightarrow \sigma(X) = \{\varnothing, \Omega\} \subseteq \mathcal{F}$$
   for $B \in \mathcal{B}(\Omega)$.

   (ii) Let $A \in \mathcal{F}, X = \mathbb{1}_A$. Then,
$$X^{-1}(B) = \begin{cases} \Omega, & 0 \in B \text{ and } 1 \in B \\ A, & 0 \in B \text{ and } 1 \notin B \\ A^c, & 0 \notin B \text{ and } 1 \in B \\ \varnothing, & 0 \notin B \text{ and } 1 \notin B \end{cases}$$
$$\Rightarrow \sigma(X) = \{\varnothing, A, A^c, \Omega\} \subseteq \mathcal{F}$$

   (iii) Let $\{A_i\}_{i \in \{1, \ldots, n\}} \subseteq \mathcal{F}$ be a partition of Ω, and
$$X_n = \sum_{i=1}^{n} x_i \mathbb{1}_{A_i}$$
   for $x_i \neq x_j \ \forall i \neq j$. Then, as in parts (i) and (ii), we obtain
$$\sigma(X_n) = \{\bigcup_{i \in I} A_i : I = \{1, \ldots, n\}\} \subseteq \mathcal{F}$$

Remark 5.5

1. We typically write X instead of $X(\omega)$. For example, $X = 1_A$ instead of $X(\omega) = 1_A(\omega)$. This is to not confuse the study of X (a function) with $X(\omega)$ (a single value often denoted x, referred to as a "realization" of X, after an experiment has been conducted and ω (state of the world) is known).

2. Part 3 (iii) of 5.4 hints at the fact that we can study sequences of $x_1, \ldots, x_n, \ldots$. Such sequences play a major role in major limiting results.

   Random variables X_n as in (iii) are called simple random variables, and are the building blocks of defining an integral $EX = \int_\Omega X(\omega)dP(\{\omega\})$ known as "expectation of X" through a process called algebraic induction.

3. One can show measurability is preserved by many operations, eg compositions of measurable functions are measurable. Also

$$\begin{gathered}\inf_{k \geq n} X_k \\ \sup_{k \geq n} X_k \\ \lim_{n\to\infty} \inf X_n \\ \lim_{n\to\infty} \sup X_n \\ \lim_{n\to\infty} X_n\end{gathered}$$

all preserve measurability. Continuous X are measurable and whenever we use such operations, assume resulting mappings to be measurable.

4. One can show random vectors are vectors of random variables. We typically write random vectors as $\underline{X} = (X_1, \ldots, X_d)$ where $X_1, \ldots, X_d$ are random variables.

Proposition 5.6

If $(\Omega, \mathcal{F}, P)$ is a probability space, $(\Omega', \mathcal{F}')$ a measurable space, $X : \Omega \to \Omega'$ a measurable function, then

$$P_X = P \circ X^{-1}$$

is a probability measure on $(\Omega', \mathcal{F}')$, called the distribution of X (or "image measure of P with respect to X" or "push-forward measure").

Proof. Check the three properties of probability measures.

(i) $P_X : \mathcal{F}' \to [0, 1]$

(ii)

$$\begin{aligned}P_X(\Omega') &= P(X^{-1}(\Omega')) \\ &= P(\Omega) \\ &= 1\end{aligned}$$

(iii)

$$\begin{aligned}&\{A_i\}_{i\in\mathbb{N}} \subseteq \mathcal{F}', A_i' \cap A_j' = \varnothing \text{ for } i \neq j \\ \Rightarrow P_X(\bigcup_{i=1}^{\infty} A_i') &= P(X^{-1}(\bigcup_{i=1}^{\infty} A_i')) \\ &= P(\bigcup_{i=1}^{\infty} X^{-1}(A_i')) \\ &= \sum_{i=1}^{\infty} P(X^{-1}(A_i')) \\ &= \sum_{i=1}^{\infty} P_X(A_i')\end{aligned}$$

□

Remark 5.7

1. P_X assigns probabilities to events involving measurable X since $X^{-1}(A') \in \mathcal{F}$ for $A' \in \mathcal{F}'$ and P assigns probabilities to such events.

2. We often write

$$\begin{aligned} P(X \in A') &:= P(\{\omega \in \Omega : X(\omega) \in A'\}) \\ &= P(X^{-1}(A')) \\ &= P_X(A') \end{aligned}$$

   for $A' \in \mathcal{F}'$.

3. If X is a random variable, distribution of X is a Borel probability measure on $\mathbb{R}$. We call

$$\begin{aligned} F(X) &:= P_X((-\infty, x]) \\ &= P(x \in (-\infty, x]) & &= P(\{\omega \in \Omega : X(\omega) \leq x\}) \end{aligned}$$

   for $x \in \mathbb{R}$ the distribution function (df) of X, and write $X \sim F$. For $P(X \in (-\infty, x])$, we write $P(X \leq x)$, so $X \sim F$ iff $F(x) = P(X \leq x)$ for $x \in \mathbb{R}$.

The following result characterizes all distribution functions on $\mathbb{R}$.

Theorem 5.8

$F : \mathbb{R} \to [0, 1]$ is the distribution function of a unique Borel probability measure P_F on $\mathbb{R}$ iff:

(i) $F(-\infty) = \lim_{x \downarrow -\infty} F(x) = 0$ and $F(\infty) = \lim_{x \uparrow \infty} F(x) = 1$.

(ii) F is increasing.

(iii) F is right continuous.

Here, P_F is P_X for $X \sim F$.

Proof. $(\Rightarrow)$ Suppose $F : \mathbb{R} \to [0, 1]$ is the distribution function of a unique Borel probability measure P_F on $\mathbb{R}$. Then,

(i)

$$\begin{aligned} F(-\infty) &= \lim_{n \to \infty} F(-n) \\ &= \lim_{n \to \infty} P_F((-\infty, -n]) \\ &= P_F(\bigcap_{n=1}^{\infty} (-\infty, -n]) \text{ by continuity from above} \\ &= P_F(\varnothing) \\ &= 0 \end{aligned}$$

Also,

$$\begin{aligned} F(\infty) &= \lim_{n\to\infty} F(n) \\ &= \lim_{n\to\infty} P_F((-\infty, n]) \\ &= P_F(\bigcup_{i=1}^{\infty}(-\infty, n]) \text{ by continuity from below} \\ &= P_F(\mathbb{R}) \\ &= 1 \end{aligned}$$

(ii) If $x, y \in \mathbb{R}$ such that $x < y$ then $F(x) = P_F((-\infty, x]) \leq P_F((-\infty, y]) = F(y)$ since $(-\infty, x] \subseteq (-\infty, y]$.

(iii) Let $\{h_n\}$ be a sequence such that $h_n \downarrow 0$ as $n \to \infty$. Then,

$$\begin{aligned} \lim_{n\uparrow\infty} F(x+h_n) &= \lim_{n\uparrow\infty} P_F((-\infty, x+h_n]) \\ &= P_F(\bigcap_{i=1}^{\infty}(-\infty, x+h_n]) \\ &= P_F((-\infty, x]) \\ &= F(x) \end{aligned}$$

for $x \in \mathbb{R}$.

($\Leftarrow$) By theorem 1.21, (ii) and (iii) imply there exists exactlt one Borel measure μ_F on $\mathbb{R}$ such that $\mu_F((a,b]) = F(b) - F(a)$ for $a \leq b$. Thus,

$$\begin{aligned} \mu_F(\mathbb{R}) &= \lim_{n\to\infty} \mu_F((-n, n]) \\ &= \lim_{n\to\infty} (F(-n) - F(n)) \\ &= 1 \end{aligned}$$

so μ_F is a probability measure on $\mathbb{R}$. Its distribution function is F since for all $x \in \mathbb{R}, n \in \mathbb{N}$ such that $n > -x$, $\mu_F((-n, x]) = F(x) - F(-n)$ which implies

$$\begin{aligned} F(x) &= \lim_{n\to\infty} (\mu_F((-n, x]) + F(-n)) \\ &= \mu_F((-\infty, x]) \end{aligned}$$

by continuity from below. So μ_F is the P_F as claimed. (Also P_F satisfies $P_F((a,b]) = F(b) - F(a)$ for all $a \leq b$.)

□

Remark 5.9

($\Rightarrow$) implies any distribution function F has those properties.

($\Leftarrow$) implies any such F induces a unique distribution with distribution function F. Here are some examples of such distributions.

1. $F(x) = \min\{\max\{x, 0\}, 1\}$ for $x \in \mathbb{R}$ is the distribution function of the standard uniform distribution, $U(0,1)$.

2. $F(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-1}{2}(\frac{z-\mu}{\sigma})^2} dz$ for $x, \mu, \sigma \in \mathbb{R}, \sigma > 0$ is the distribution function of the normal distribution, $N(\mu, \sigma^2)$
3. $F(x) = (1-p)\mathbb{1}_{[0,\infty)}(x) + p\mathbb{1}_{[1,\infty)}(x)$ for $x \in \mathbb{R}, p \in [0,1]$ is the distribution function for the Bernoulli distribution, $B(1,p)$.

   How do we construct $X \sim F$ in this case?

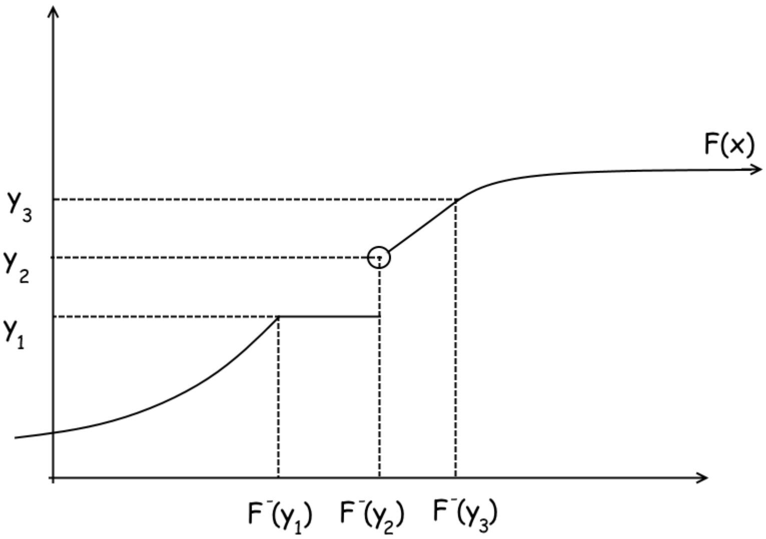
Definition 5.10

If $F : \mathbb{R} \to \mathbb{R}$ is increasing, the generalized inverse F^- of F is defined by $F^-(y) = \inf\{x \in \mathbb{R} : F(x) \geq y\}$ for $y \in \mathbb{R}$.

If F is a distribution function, F^- is the quantile function of F.

Remark 5.11

1. Sketch:



2. Some facts:
   - If F is strictly increasing and continuous, then $F^- = F^{-1}$, the ordinary inverse.
   - F^- is increasing and left-continuous.
   - One can often work with F^- as F^{-1} (which we will do), but be careful when it matters, such as in the sketch where $F(F^-(y_2)) > y2$.

   Quantile functions allow us to answer remark 5.9.

Proposition 5.12

If $F : \mathbb{R} \to [0,1]$ satisfies (i), (ii), (iii) of theorem 5.8 then there is a probability space $(\Omega, \mathcal{F}, P)$ and random variable $X : \Omega \to \mathbb{R}$ such that $X \sim F$.

Proof. Consider $(\Omega, \mathcal{F}, P) = ((0,1), \bar{\mathcal{B}}((0,1)), \lambda)$ and $X(\omega) := F^{-}(\omega)$ for $\omega \in \mathbb{R}$. Then,

$$\begin{aligned} P(X \leq x) &= P(\{\omega \in \Omega : X(\omega) \leq x\}) \\ &= P(\{\omega \in \Omega : \omega \leq F(x)\}) \\ &= P((0, F(x)) = \lambda((0, F(x)) \\ &= F(x) \end{aligned}$$

for all $x \in \mathbb{R}$. We can also use the notation P_X for P_F. □

Proposition 5.12 can be exploited to generate (pseudo-)random numbers. Pseudorandom numbers are numbers which resemble realizations of $X \sim F$ on a computer based on following a result known as inversion method for sampling (generating random numbers). Note that there are various ways to sample from $U(0,1)$ (distribution function is $F_U(x) = x \; \forall x \in [0,1]$) on a computer.

Proposition 5.13

Let F be a distribution function, $U \sim U(0,1)$. Then $X : F^{-}(U) \sim F$.

Proof.

$$\begin{aligned} P(X \leq x) &= P(F^{-}(U) \leq x) \\ &= P(U \leq F(x)) \\ &= F_U(F(x)) \text{ since } U \sim U(0,1) \\ &= F(x) \end{aligned}$$

□

Remark 5.14

1. If $X \sim F$, proof of theorem 5.8 ($\Leftarrow$) implies that

$$\begin{aligned} P(X \in (a,b]) &= P_X((a,b]) \\ &= P_F((a,b]) \\ &= F(b) - F(a) \end{aligned}$$

2. If $X \sim F$, then

$$\begin{aligned} P(X = x) &= P(X \in \bigcap_{i=1}^{\infty} (x - \frac{1}{n}, x]) \\ &= \lim_{n \to \infty} P(X \in (x - \frac{1}{n}, x]) \\ &= F(x) - \lim_{n \to \infty} F(x - \frac{1}{n}) \\ &= F(x) - F(x-) \end{aligned}$$

for $x \in \mathbb{R}$.

- If F is continuous, then $P(X = x) = 0$.

- If F is constant over $(a,b]$, then $P(x \in (a,b]) = F(b) - F(a+) = 0$.
- If F jumps in x, then $P(X = x) = F(x) - F(x-)$, so the probability equals the jump height of F in x.
- Note that in each "jump gap" $(F(x-), F(x))$ is a rational number and since F is increasing, they are all distinct. Therefore, the number of jumps is at most equal to the number of rationals, so distribution functions can have at most countably many jumps.

Definition 5.15

Let X be a random variable, with distribution P_X and distribution function F.

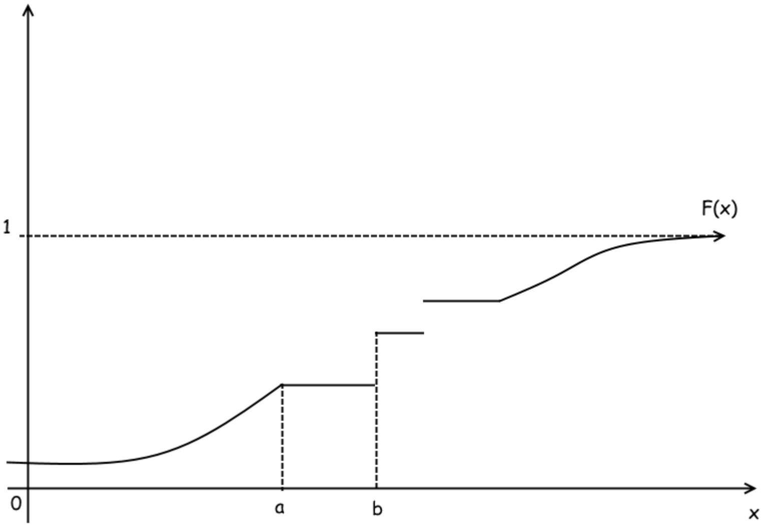
1. $X(P_X, F)$ is a discrete random variable (discrete distribution, discrete distribution function) if there is $S = \{x_1, \dots, x_n, \dots\} \subseteq \mathbb{R}$ such that $P(X \in S) = 1$. Equivalently, if F is a step function or if $\text{range}(F) := \{F(x) : x \in \mathbb{R}\}$ is at most countable.
2. $X(P_X, F)$ is a continuous(ly distributed) random variable (continuous distribution, continuous distribution function) if F is continuous.
3. $X(P_X, F)$ is an absolutely continuous(ly distributed) random variable (absolutely continuous distribution, absolutely continuous distribution function) if

$$F(x) = \int_{-\infty}^{x} f(z)dz$$

for some function $f(z) : \mathbb{R} \to [0, \infty)$ which is integrable, such that $\int_{-\infty}^{\infty} f(z)dz = 1$. Then, f is called the density of $X(P_X, F)$.

Remark 5.16

1. Not all distribution functions F are discrete, continuous or absolutely continuous. They can have mixed type, that is, be discrete on one nonzero interval and continuous on another. Here is a sketch of one such distribution function.



If $X \sim F$, then

$$\begin{aligned} P(x \in (a,b]) &= F(b) - F(a) = \text{ "jump height"} \\ P(x = a) &= F(a) - F(a-) = 0 \\ P(x \in (a,b)) &= 0 \end{aligned}$$

2. If F is differentiable everywhere on $\mathbb{R}$, F' is integrable, and the fundamental theorem of calculus implies F is absolutely continuous with density $f = F'$.

   Every absolutely continuous function F is continuous since

$$\begin{aligned} |F(x_h) - F(x)| &= |\int_x^{x+h} f(z)dz| \\ &\leq \int_x^{x+h} f(z)dz \\ &\leq M \int_x^{x+h} 1dz \text{ for some } M \in \mathbb{R} \\ &= M(x+h-x) \\ \Rightarrow \lim_{h\to 0} Mh = 0 \end{aligned}$$

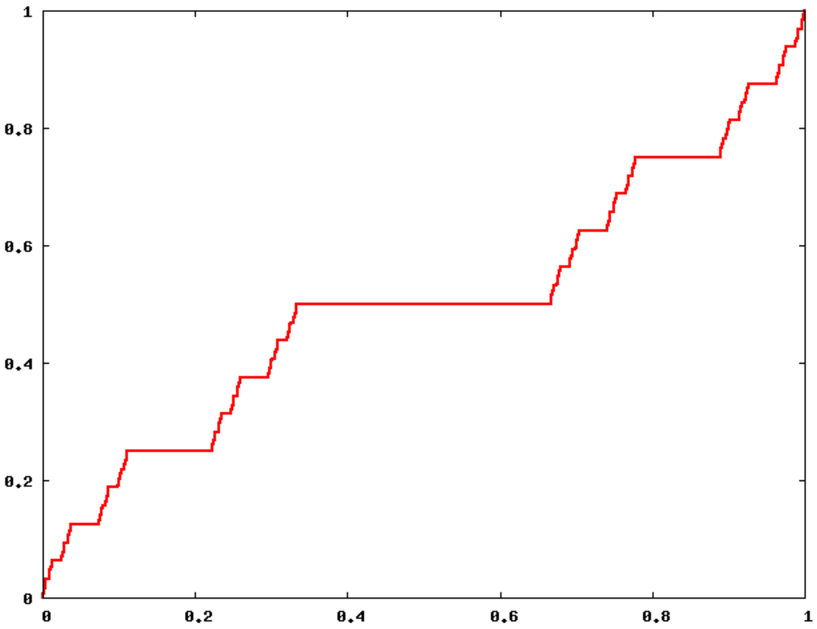
   which implies continuity. Also if $X \sim F$ then $P(X = x) = 0$ for all $x \in \mathbb{R}$. Hence $P(X = x) \neq f(x)$ in general. However,

$$\begin{aligned} P(X \in (x-\varepsilon, x]) &= F(x) - F(x-\varepsilon) \\ &= \int_{x-\varepsilon}^{x} f(z)dz \\ &\approx \varepsilon f(x) \end{aligned}$$

   for small $\varepsilon > 0$. (interpolation of densities)

3. Not all continuous distribution functions are absolutely continuous. Here is an example of a distribution function which is continuous but not absolutely continuous.

   For each $x = \sum_{i=1}^{\infty} a_i 3^{-i} \in C$ where $a_i \in \{0,2\}$ and C is the Cantor set, define $F(x) := \sum_{i=1}^{\infty} \frac{a_i}{2} 2^{-i}$. That is, a base 2 expansion of a unique number in $[0,1]$. Thus F maps $C \to [0,1]$. Now extend the domain such that F is constant in intervals missing from C.

   So F is increasing, range$(F) = [0,1]$, and F is continuous. This function is also called Cantor's distribution or the devil's staircase. Since F is constant on $[0,1]\backslash C$, a density candidate of F would have to be 0 on $[0,1]\backslash C$. Let f be such a density candidate. Then,

$$\begin{aligned}\int_{[0,1]} f(z)dz &= \int_C f(z)dz + \int_{[0,1]\backslash C} f(z)dz \\ &= 0 \neq 1\end{aligned}$$

   which shows F is not absolutely continuous.

4. For probability mass functions or densities, always provide a domain. For example $f(x) = \frac{1}{c}$ is a density on $[0,c]$ (namely $U(0,c)$) but not on $[0,\infty)$ since $\int_0^{\infty} \frac{1}{c} \neq 1$ for all $c \in \mathbb{R}$.

Example 5.17

1. Let $X \sim U(0,1) \Rightarrow F(x) = P(X \leq x) = x$ for $x \in [0,1]$. For any pairwise disjoint intervals $(a_i, b_i] \subseteq [0,1], i \in \mathbb{N}$ with $\sum_{i=1}^{\infty}(b_i - a_i) = h$, we have $P(X \in \sum_{i=1}^{\infty}(a_i, b_i]) = h$ by σ-additivity.

   In particular, the probability of X to fall in an interval length h is h. This can also be seen from the density, $f(x) = F'(x) = 1$ for $x \in (0,1]$ since $P(X \in (x, x+h]) = \int_x^{x+h} f(z)dz = x + h - x = h$.

2. Let $X \sim N(\mu, \sigma^2)$. Then, $F(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-1}{2}(\frac{z-\mu}{\sigma})^2} dz$ and the density candidate is $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-1}{2}(\frac{z-\mu}{\sigma})^2}$. It can be shown that this is indeed a density but it's just calculus and not on the test and fuck copying that proof into LaTeX.

A random number generator from $N(\mu, \sigma^2)$ can be done with inversion based method based on fast and accurate approximation of F^-.

5.2 Random Vectors

Many notions introduced so far can be extended to random vectors $\underline{X} = (X_1, \dots, X_d)$.

Definition 5.18

Let $(\Omega, \mathcal{F}, P)$ be a probability space, $\underline{X} = (X_1, \dots, X_d)$ a random vector. Then $P_{\underline{X}} = P \circ \underline{X}^{-1}$ (with

$$P_{\underline{X}}(B) = P(\underline{X}^{-1}(B)) = P(\{\omega \in \Omega : \underline{X}(\omega) \in B\})$$

for $B \in \mathcal{B}(\mathbb{R})$) is called the distribution of $\underline{\text{X}}$ (a Borel probability measure on $\mathbb{R}^d$) and

$$\begin{aligned} F(\underline{x}) &:= P_{\underline{X}}((-\infty, x]) \\ &= P(\underline{X} \in (-\underline{\infty}, \underline{x}]) \\ &= P(\{\omega \in \Omega : X(\omega) \le \underline{x}\}) \\ &= P(\{\omega \in \Omega : X_j(\omega) \le x_j \ \forall j = 1, \dots, d\}) \\ &= P(\bigcap_{j=1}^{d} \{\omega \in \Omega : X_j(\omega) \le x_j\}) \end{aligned}$$

for $\underline{X} = (x_1, \dots, x_d) \in \mathbb{R}^d$ is the distribution function of $\underline{\text{X}}$, as before. We write $\underline{X} \sim F$ in this case, and write $P(\underline{X} \le \underline{x})$ or $P(X_1 \le x_1, \dots, X_d \le x_d)$ for $P(\underline{X} \in (-\underline{\infty}, \underline{x}])$.

Thus, $X \sim F \Leftrightarrow F(\underline{x}) = P(\underline{X} \le \underline{x})$ for all $\underline{x} \in \mathbb{R}^d$.

We call

$$\begin{aligned} F_j(x_j) &= P(X_j \le x_j) \\ &= P(X_1 \le \infty, \dots, X_{j-1} \le \infty, X_j \le x_j, X_{j+1} \le \infty, \dots, X_d \le \infty) \\ &= F(\infty, \dots, \infty, x_j, \infty, \dots, \infty) \end{aligned}$$

the j-th margin of F or the j-th marginal distribution function of $\underline{\text{X}}$.

Similar to theorem **??** (based on remark 1.24 (1)), we obtain this characterization of all distribution functions F.

Theorem 5.19

A function $F : \mathbb{R}^d \to [0, 1]$ is the distribution function of a unique Borel probability measure $P_F (= P_X)$ on $\mathbb{R}^d$ iff:

(i) $\lim_{x_j \to -\infty} F(\underline{x}) = 0$ for at least one $j = 1, \dots, d$ and $\lim_{\underline{x} \uparrow \underline{\infty}} F(\underline{x}) = 1$.

(ii) F is d-increasing, that is, $\Delta_{(\underline{a}, \underline{b}]} F \ge 0$ for all $a \le b$ (see remark 1.24).

(iii) F is right continuous: $\lim_{\underline{h} \downarrow 0} F(\underline{x} + \underline{h}) = F(\underline{x})$ for all $\underline{x} \in \mathbb{R}^d$.

If we have a distribution function F, with $X \sim F$, we have $\Delta_{(\underline{a}, \underline{b}]} F = \mu_F((\underline{a}, \underline{b}) = P_X(\underline{a}, \underline{b}]) = P(x \in (\underline{a}, \underline{b}])$. So, the F-volume of $(\underline{a}, \underline{b}]$ is the probability of $x \in (a_j, b_j]$ for all $j = 1, \dots, d$. This probability can be computed as in remark 1.24 (2).

Definition 5.20

Let $\underline{X} = (X_1, \dots, X_d)$ be a random vector with distribution P_X and distribution function F.

1. $\underline{X}(P_{\underline{X}}, F)$ is discrete (discrete distribution, discrete df) if there exists $S = \{\underline{x}_1, \underline{x}_2, \dots\} \subseteq \mathbb{R}^d$ such that $P(\underline{X} \in S) = 1$. Alternatively, if F is a step function. Or, if range$(F) := \{F(\underline{x}) : \underline{x} \in \mathbb{R}^d\}$ is at most countable. In this case, $f(\underline{x}_i) = P(\underline{X} = \underline{x}_i)$ for $i \in \mathbb{N}$ is the probability mass function of $\underline{X}(P_{\underline{X}}, F)$.

2. $\underline{X}(P_{\underline{X}}, F)$ is a continuously distributed random vector (continuous distribution, continuous df) if F is continuous.

3. $\underline{X}(P_{\underline{X}}, F)$ is an absolutely continuous random vector (abs. cont. distribution, abs. cont. df) if
$$F(\underline{x}) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_d} f(z_1, \dots, z_d) dz_1 \dots dz_d$$
for all $\underline{x} \in \mathbb{R}^d$, for some $f : \mathbb{R}^d \to [0, \infty)$ such that
$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(z_1, \dots, z_d) dz_1 \dots dz_d = 1$$
.

In this case we call f the (joint) density of $\underline{X}(P_{\underline{X}}, F)$.

Remark 5.21

1. If F is abs. cont., then
$$\begin{aligned} F_j(x_j) &= F(\infty, \dots, \infty, x_j, \infty, \dots, \infty) \\ &= \int_{-\infty}^{x_j} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(z_1, \dots, z_d) \end{aligned}$$
which implies F_j is abs. cont. with density
$$\begin{aligned} f_j(x_j) &= \frac{d}{dx_j} F_j(x_j) \\ &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(z_1, \dots, z_{j-1}, x, z_{j+1}, z_d) dz_1 \dots dz_{j-1} dz_{j+1} \dots dz_d \end{aligned}$$
being the j-th marginal density of F. So for $j \in \{1, \dots, d\}$, the j-th margin F_j of F can be obtained by taking limits ($X_k \to \infty \; \forall k \neq j$) and its density f_j by integrating out the joint density of f with respect to all components except the j-th.

2. We have just shown if F is abs. cont., so are its margins. The converse is not true in general.

6 Independence and dependence

Independence is the most crucial assumption in stochastic models and results.

Definition 6.1

Let $(\Omega, \mathcal{F}, P)$ be a probability space, $(\Omega', \mathcal{F}')$ a measurable space, and $X_i : \Omega \to \Omega'$ be $(\mathcal{F}, \mathcal{F}')-$measurable for all $i \in I \subseteq \mathbb{R}$.

Then, $X_i, i \in I$ are independent if $\sigma(X_i), i \in I$ are independent. That is, if

$$X_{i_1}^{-1}(A'_{i_1}), \ldots, X_{i_n}^{-1}(A'_{i_n})$$

are independent for all

$$A'_{i_1}, \ldots, A'_{i_1} \in \mathcal{F}'$$
$$\{i_1, \ldots, i_n\} \in I$$
$$n \in \mathbb{N}$$

Remark 6.2

1. d random variables $X_1, \ldots, X_d$ are thus independent iff

$$\begin{aligned} &X^{-1}(B_1), \ldots, X^{-1}(B_d) \text{ are independent for all } B_1, \ldots, B_d \in \mathcal{B}(\mathbb{R}) \\ \Leftrightarrow &P(\bigcap_{j \in J} X_j^{-1}(B_j)) = \prod_{j \in J} P(X_j^{-1}(B_j)) \ \forall J \subseteq \{1, \ldots, d\} \\ \Leftrightarrow &P(\bigcap_{j=1}^{d} X_j^{-1}(B_j)) = \prod_{j=1}^{d} P(X_j^{-1}(B_j)) \end{aligned}$$

for all $B_1, \ldots, B_d \in \mathcal{B}(\mathbb{R})$.

One can show that it suffices to consider $B_j \in \{(-\infty, x) : x \in \mathbb{R}\}$ for $j = 1, \ldots, d$.

Therefore, $X_1, \ldots, X_d$ are independent iff

$$\begin{aligned} F(\underline{x}) &= P(\underline{X} \leq \underline{x}) \\ &= P(\bigcap_{j=1}^{d} X_j^{-1}(-\infty, x_j]) \\ &= \prod_{j=1}^{d} P(X_j^{-1}((-\infty, x_j])) \\ &= \prod_{j=1}^{d} P(X_j \leq x_j) \\ &= \prod_{j=1}^{d} F_j(x_j) \end{aligned}$$

2. If F is abs. cont., $X_1, \dots, X_d$ are independent iff

$$\begin{aligned} f(\underline{x}) &= \frac{\partial^d}{\partial x_1 \dots \partial x_d} F(\underline{x}) \\ &= \prod_{j=1}^{d} \frac{\partial}{\partial x_j} F_j(x_j) \\ &= \prod_{j=1}^{d} f_j(x_j) \end{aligned}$$

for $\underline{x} \in \mathbb{R}^d$.

Similarly for discrete random vectors $\underline{X} = (X_1, \dots, X_d)$ with support $S = \{\underline{x}_1, \underline{x}_2, \dots\}$, $X_1, \dots, X_d$ are independent iff $f(\underline{x}) = \prod_{j=1}^{d} f_j(x_j)$ for all $\underline{x} \in S$, or equivalently, $P(\underline{X} = \underline{x}) = \prod_{j=1}^{d} P(X_j = x_j)$ for all $\underline{x} \in S$.

3. One can show if $X_{j_k} : j \in \mathbb{N}, k \in \{1, \dots, d_j\}$ are independent random variables and $h_j : \mathbb{R}^{d_j} \to \mathbb{R}$ are $(\mathcal{B}(\mathbb{R}^{d_j}), \mathcal{B}(\mathbb{R}))-$measurable maps, then

$$Y_j = h_j(X_{j_1}, \dots, X_{j_{d_j}}), j \in \mathbb{N}$$

are also independent random variables. In other words, measurable functions of independent random variables are also random variables.

Example 6.3

1. Let $\underline{x} = (x_1, x_2) \sim F$ where F is the distribution function of the $U((0,1)^2)$ distribution with density 1 on $(0,1)^2$. Are X_1, X_2 independent?

$$\begin{aligned} F(\underline{x}) &= \int_0^{x_2} \int_0^{x_1} 1 dz_1 dz_2 \\ &= 1(x_1 - 0)(x_2 - 0) \\ &= x_1 x_2 \end{aligned}$$

for all $\underline{x} = (x_1, x_2) \in (0,1)^2$. Therefore,

$$\begin{aligned} F_1(x_1) &= F(x_1, 1) \\ &= x_1 \end{aligned}$$

$$\begin{aligned} F_2(x_2) &= F(1, x_2) \\ &= x_2 \end{aligned}$$

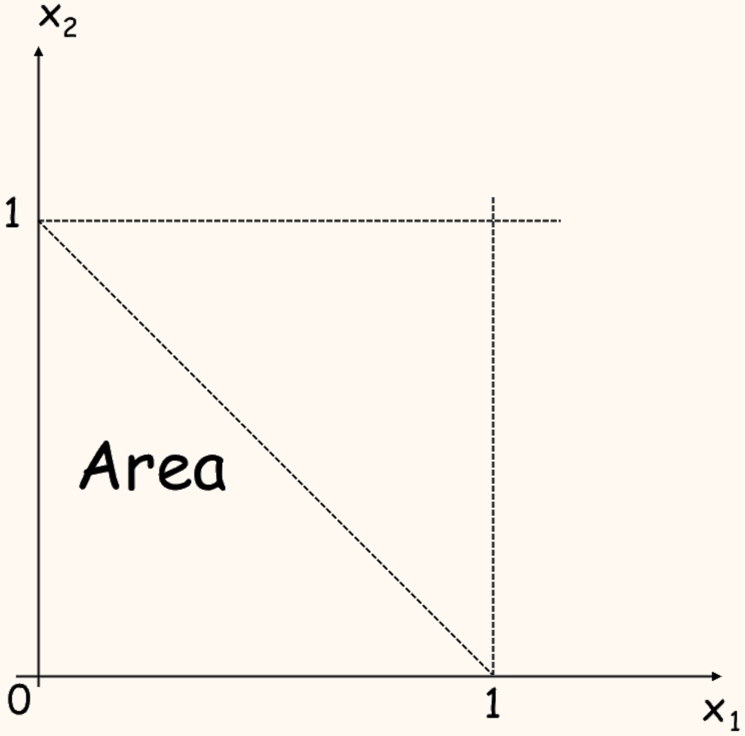
for all $\underline{x} = (x_1, x_2) \in (0,1)^2$. So, $F(x_1, x_2) = F_1(x_1) F_2(x_2)$ and X_1, X_2 are independent.

Also note that $f(\underline{x}) = 1 = f_1(x_1) f_2(x_2)$ for all $x_1, x_2 \in (0,1)$ in this case.

2. Let $\underline{X} = (X_1, X_2) \sim F$ for F being the distribution function of $U(S^2)$ distribution with density 2 on $S^2 = \{\underline{x} \in \mathbb{R}^2_+ : x_1 + x_2 \leq 1\}$. Are X_1, X_2 independent?

$$
\begin{aligned}
F(\underline{x}) &= \int_0^{x_2}\int_0^{x_1} 2dz_1dz_2 \\
&= 2(x_1-0)(x_2-0) \\
&= 2x_1x_2
\end{aligned}
$$

for all $\underline{x} \in S^2$.



Therefore, $F_1(x_1) = 2$"area" $= 2(\frac{1}{2}(1)(1) - \frac{1}{2}(1-x_1)(1-x_1)) = x_1(2-x_1)$ for all $x_1 \in (0,1)$. By symmetry, $F_2(x_2) = x_2(2-x_2)$ for all $x_2 \in (0,1)$.

Thus, $F(\frac{1}{2},\frac{1}{2}) = \frac{1}{2} \neq \frac{9}{16} = F_1(\frac{1}{2})F_2(\frac{1}{2})$, so X_1, X_2 are not independent.

Also note that $f(\underline{x}) = 2 \neq 2(1-x_1)2(1-x_2) = f_1(x_1)f_2(x_2)$ in general.

Proposition 6.4

For distribution functions $F_1, \ldots, F_d$, there is a probability space $(\Omega, \mathcal{F}, P)$ and independent random variables $X_1, \ldots, X_d$ such that $X_j \sim F_j$ for all $j \in \{1, \ldots, d\}$.

Proof. Consider $\Omega = \mathbb{R}^d, \mathcal{F} = \mathcal{B}(\mathbb{R}^d), P(\prod_{j=1}^d B_j) = \prod_{j=i}^d P_{F_j}(B_j), X_j(\underline{\omega}) = \omega_j$ for all $\underline{\omega} \in \mathbb{R}^d, j \in \{1, \ldots, d\}$. Then, for all j, X_j is measurable and

$$\begin{aligned}
P(\bigcap_{j=1}^{d} X_j^{-1}(B_j)) &= P(\bigcap_{j=1}^{d}\{\omega \in \Omega : X_j(\omega) \in B_j\}) \\
&= P(\prod_{j=1}^{d} B_j) \\
&= \prod_{j=1}^{d} P(B_j) \\
&= \prod_{j=1}^{d} P(X_j^{-1}(B_j))
\end{aligned}$$

for all $B_1, \ldots, B_d \in \mathcal{B}(\mathbb{R})$ which implies $X_1, \ldots, X_d$ are independent.

Furthermore,

$$\begin{aligned}
P(X_j \leq x_j) &= P(X_j^{-1}((-\infty, x_j])) \\
&= \prod_{j=1}^{d} P_{F_j}(B_j) \\
&= P_{F_j}((-\infty, x_j]) \\
&= F_j(x_j)
\end{aligned}$$

So, $X_j \sim F_j$. □

Infinite sequences of independent random variables $X_1, X_2, \ldots$ can be constructed by involving "Kolmogorov's extension theorem". In particular, can construct independent and identically distributed (iid) random variables $X_1, X_2, \ldots$ from a distribution function F (universal df). In short, $X_1, X_2, \cdots \sim F$.

Interlude: Independence implies pairwise independence but not conversely.

Definition 6.5

A copula C is a distribution function with $U(0,1)$ margins.

Proposition 6.6

A function $C : [0,1]^d \to [0,1]$ is a d dimensional copula iff:

(i) $C(\underline{u}) = 0$ if $u_j = 0$ for at least one $j \in \{1, \ldots, d\}$. **(Groundedness)**

(ii) $C(1, \ldots, 1, u_j, 1, \ldots, 1) = u_j$ for all $u_j \in [0,1], j \in \{1, \ldots, d\}$. **($U(0,1)$ margins)**

(iii) $\Delta_{(\underline{a}, \underline{b}]} C \geq 0$ for all $(\underline{a}, \underline{b}]) \in [0,1]^d : \underline{a} \leq \underline{b}$. **(d-increasing)**

Example 6.7

1. $C(\underline{u}) = \prod(\underline{u}) := \prod_{j=1}^{d} u_j$ for $\underline{u} \in [0,1]^d$ is a copula (independence copula). To show it's a density, show it's non-negative and differentiable.
2. $C(\underline{u}) = M(\underline{u}) = \min\{u_1, \ldots, u_d\}, \underline{u} \in [0,1]^d$ is a copula (comonotone copula) since $\underline{U} \sim C$ for $\underline{U} =$

$(U,\dots,U)$ with $U \sim U(0,1)$.

Proof.

$$\begin{aligned} P(\underline{U} \leq \underline{u}) &= P(U \leq u_1, \dots, U \leq u_d) \\ &= P(U \leq \min\{u_1,\dots,u_d\}) \\ &= \min\{u_i,\dots,u_d\} \\ &= C(\underline{u}) \end{aligned}$$

for all $\underline{u} \in [0,1]^d$. □

3. $C(\underline{u}) = W(\underline{u}) := \max\{(\sum j = 1^d u_j) - d + 1, 0\}, \underline{u} \in [0,1]^d$ is a copula (counter-monotone copula) iff $d = 2$.

   Proof. Suppose $d = 2$. Let $\underline{U} = (U, 1-U)$ for $U \sim U(0,1)$. Then

$$\begin{aligned} P(\underline{U} \leq \underline{u}) &= P(\underline{U} \leq u_1, 1-U \leq u_2) \\ &= P(1-u_2 \leq U \leq u_1) \\ &= \max\{u_1 - (1-u_2), 0\} \\ &= W(u_1,u_2) \end{aligned}$$

   for all $u_1, u_2 \in [0,1]$.

   Suppose $d > 2$. Then,

$$\begin{aligned} \Delta_{(\underline{\frac{1}{2}},\underline{1}]}W = \sum_{i\in\{0,1\}^d} & (-1)^{\sum_{j=1}^d i_j} W((\frac{1}{2})^{i_1} \cdot 1^{1-i_1}, \dots, ((\frac{1}{2})^{i_d} \cdot 1^{1-i_d}) \\ =& max\{1(d) - d + 1, 0\} \\ & - d \max\{1(d-1) + \frac{1}{2} - d + 1, 0\} \\ & + (2^d) \max\{1(d-2) + \frac{1}{2} + \frac{1}{2} - d + 1, 0\} \\ & - \cdots + (-1)^d \max\{\frac{1}{2}(d) - d + 1, 0\} \\ =& 1 - \frac{d}{2} + 0 + \cdots + 0 < 0 \end{aligned}$$

   for $d > 3$. Therefore, W is not d-increasing (though it is componentwise increasing), so it is not a distribution. □

Theorem 6.8: Sklar's Theorem

1. For any distribution function F with continuous margins $F_1, \dots, F_d$, there exists exactly one copula C:

$$F(\underline{x}) = C(F_1(x_2), \dots, F_d(x_d))$$

   for all $\underline{x} \in \mathbb{R}^d$. (*) This copula is given by

$$C(\underline{u}) = F(F_1^-(u_1), \dots, F_d^-(u_d))$$

   for all $u \in (0,1)^d$.

2. Given any copula C and any univariate distribution functions $F_1, \dots, F_d$, F defined by (*) is a distribution function with margins $F_1, \dots, F_d$.

Proof.

1. Let $\underline{X} \sim F$ and $U := (F_1(X_1), \dots, F_d(U_d))$. Then, $\underline{U}$ has $U(0,1)$ margins, since

$$\begin{aligned} P(U_j \leq u_j) &= P(F_j(X_j) \leq u_j) \\ &= P(F_j^-(F_j(X_j)) \leq F_j^-(u_j)) \\ &= P(X_j \leq F^-(u_j)) \\ &= F_j(F_j^-(u_j)) \\ &= u_j \end{aligned}$$

for all $u_j \in (0,1), j \in \{1, \dots, d\}$. So U has a copula, say C, as its distribution function. Since

$$\begin{aligned} X_j &= F_j^-(F_j(X_j)) \\ &= F_j^-(u_j) \end{aligned}$$

for all j, therefore

$$\begin{aligned} F(\underline{x}) &= P(X_j \leq x_j \ \forall j) \\ &= P(F_j^-(u_j) \leq x_j \ \forall j) \\ &= P(U_j \leq F_j(x_j) \ \forall j) \\ &= C(F_1(x_1), \dots, F_d(x_d)) \end{aligned}$$

for all $\underline{x} \in \mathbb{R}^d$, so C satisfies (*). It is uniquely given by $C(u) = C(F_1(F_1^-(u_1)), \dots, F_d(F_d^-(u_d))) = F(F_1^-(u_1), \dots, F_d^-(u_d))$ for all $u \in (0,1)^d$.

2. Let $\underline{U} \sim C$, and $\underline{X} := (F_1^-(U_1), \dots, F_d^-(U_d))$. Since

$$\begin{aligned} P(\underline{X} \leq \underline{x}) &= = P(F_1^-(U_1) \leq x_1, \dots, F_d^-(U_d) \leq x_d) \\ &= P(U_1 \leq F_1(x_1), \dots, U_d \leq F_d(x_d)) \\ &= C(F_1(x_1), \dots, F_d(x_d)) \end{aligned}$$

for all $x \in \mathbb{R}^d$, therefore F defined by (*) is a distribution function, namely that of $\underline{X}$, with margins $F_1, \dots, F_d$ (by prop. 5.13).

□

Remark 6.9

1. Part (1) allows us to decompose any continuous F into its copula and margins $F_1, \dots, F_d$. The copula is this precisely the function containing the information about the dependence between $X_1, \dots, X_d$ of $\underline{X}$.

   Part (2) allows us to construct flexible multivariate distribution functions.

2. Two other important results:

   - Frécbet-Hoeffding bounds: $W(\underline{u}) \leq C(\underline{u}) \leq M(\underline{u})$ for all $\underline{u} \in [0,1]^d$ and copulas C.

- invariance principle: If $T_1, \dots, T_d$ are strictly increasing, then $(T_1(x_1), \dots, T_d(x_d))$ also has copula C.

7 Summary Statistics

In applications we often want to summarize a distribution function F (or $X \sim F$) in terms of a real number, a summary statistic, such as

- α-quantile $F^{-}(\alpha)$ (describes a location) is called the median of F for $\alpha = \frac{1}{2}$.
- Mode of $F = \operatorname{avg}\sup f(x)$ if F has density f or pmf f (for the latter, the mode denotes the x-value most likely to appear).
- Mean of F (describes the location of the average outcome or expected outcome), the variance (describes scale/variation/dispersion), if they exist.

7.1 Expectation and moments

What value μ do you expect when rolling a fair die?

Let $\Omega = \{1, \dots, 6\}, \mathcal{F} = \mathcal{P}(\Omega), P(A) = \frac{|A|}{|\Omega|}$ for $A \in \mathcal{F}$. Then,

$$\begin{aligned}
\mu &= \sum_{\text{all outcomes}} \text{“outcome”} \cdot P(\text{“outcome”}) \\
&= \sum_{\omega \in \Omega} \omega P(\{\omega\}) \\
&= \sum_{i=1}^{6} i \cdot \frac{1}{6} \\
&= \frac{7}{2}
\end{aligned}$$

Let $X(\omega) = \omega, X : \Omega \to \mathbb{R}$. Then, $\mu = \sum_{\omega \in \Omega} X(\omega) P(\{\omega\})$ is the mean or expectation of X computed over Ω. In the dice example, $P(X = x) = \frac{1}{6}$ for all $x \in S = \{1, \dots, 6\}$. We compute μ over $\mathbb{R}$ via

$$\begin{aligned}
\mu &= \sum_{\omega \in S} x P(X = x) \\
&= \sum_{\omega \in S} x f(x)
\end{aligned}$$

If X is absolutely continuous with density f, analogous formulas are

$$\begin{aligned}
\mu &= \int_{\Omega} X(\omega) dP(\{\omega\}) \\
&= \int_{\Omega} x dP \text{ for shorthand} \qquad\qquad = \int_{\mathbb{R}} x f(x) dx
\end{aligned}$$

if these integrals exist, in which case one also writes $E(X)$ for μ.

Consider a general probability space $(\Omega, \mathcal{F}, P)$ with random variable X. In general, $E(X)$ can be constructed in 3 steps.

1. If $X = \sum_{i=1}^n x_i = \mathbb{1}_{A_i}$ is simple, one can always define mean of X as
$$E(X) := \sum_{i=1}^n x_i P(A_i) = \sum_{i=1}^n x_i P(X = x_i)$$
2. If $X \geq 0$, mean of X is defined by
$$E(X) := \sup E(Y)$$
for simple Y such that $0 \leq Y \leq X$.
3. In general, for $X = X^+ - X^-$ where $X^+ = \max\{X, 0\}$ and $X^- = -\min\{X, 0\}$, one defines
$$E(X) = E(X^+) - E(X^-)$$
if $E(X^+) < \infty$ or $E(X^-) < \infty$. $E(X)$ is finite $\Leftrightarrow$ both $E(X^+)$ and $E(X^-)$ are finite $\Leftrightarrow$ $E(|X|) < \infty$ for $|X| = X^+ + X^-$.

This process is known as algebraic induction and leads to this definition.

Definition 7.1

Let $(\Omega, \mathcal{F}, P)$ be a probability space with random variable X. If $E(X^+) < \infty$ or $E(X^-) < \infty$, X is quasi-integrable and
$$\begin{aligned} E(X) &= \int_\Omega X dP \\ &= \int_\Omega X(\omega) dP(\{\omega\}) \end{aligned}$$
If $E(|X|) < \infty$ (notation: $X \in L^1(\Omega, \mathcal{F}, P)$), then X is integrable and $E(X)$ is the mean or expectation of X (or its distribution or df).

We rarely compute expectations over Ω, but rather over $\mathbb{R}$, in terms of df F of X (or its density/pmf). This is classified by this result (by algebraic induction).

Theorem 7.2: Change of Variables

Let $(\Omega, \mathcal{F}, P)$ be a probability space with random variable $X \sim F$. If $h : \mathbb{R} \to \mathbb{R}$ is measurable and $E(|h(X)|) < \infty$, then
$$\begin{aligned} \int_\Omega h(x) dP &= E(h(x)) \text{ by definition} \\ &= \int_\mathbb{R} h(x) dF(x) \text{ is the claim} \\ &:= \begin{cases} \sum_{x \in S} h(x) f(x) & \text{if } F \text{ is discrete, support is } S \text{, pmf is } f \\ \int_{x \in \mathbb{R}} h(x) f(x) dx & \text{if } F \text{ is abs. cont. with density } f \end{cases} \end{aligned}$$
(!!!)

Remark 7.3

1. $E(|h(x)|) < \infty$ can be verified by showing $\int_\mathbb{R} |h(x)| dF(x) < \infty$. Often this is clear, such as if h is

bounded then

$$\begin{aligned} E(|h(x)|) &= \int_{\mathbb{R}} |h(x)| dF(x) \\ &\leq M \int_{\mathbb{R}} dF(x) < \infty \end{aligned}$$

where M is a real number.

2. $E(h(x))$ can be computed over $\mathbb{R}$ via
   (i) $\int_{\mathbb{R}} h(x) dF_X(x)$ where F_X is df of X; or by substitution via $h(x) = y$:
   (ii) $\int_{\mathbb{R}} y dF_y(y)$ where F_y is df of $y = h(x)$.
3. If $h(x) = x^p, p > 0$, then p-th moment of $X \sim F$, if it exists, is

   $$E(x^p) = \int_{\text{“}\mathbb{R}\text{”}} x^p dF(x)$$

   One can use “Holder's inequality” to show if $E(|x|^q) < \infty$ for some $q > p$, then $E(|x|^p) < \infty$ for all $0 < p < q$. That is, higher moments existing imply lower moments existing.
4. Proof of theorem 7.2 readily extends to random vectors $\underline{X} \sim F$, measurable $\mathbb{R}^d \to \mathbb{R}$ such that $E(|h(x)|) < \infty$. In this case,

   $$\begin{aligned} \int_{\Omega} h(\underline{x}) dP &= E(h(\underline{X})) \\ &= \int_{\mathbb{R}^d} h(\underline{x}) dF(\underline{x}) \end{aligned}$$

   If $h(\underline{x}) = x_1 \cdot \dots \cdot x_d$, and $X_1, \dots, X_d$ are independent, then

   $$\begin{aligned} E(X_1, \dots, X_d) &= \int_{\mathbb{R}^d} x_1 \cdot \dots \cdot x_d dF(\underline{x}) \\ &= \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} x_1 \dots x_d dF_1(x_1) \dots dF_d(x_d) \\ &= \prod_{j=1}^{d} \int_{\mathbb{R}} x_j dF_j(x_j) \\ &= E(x_1) \dots E(x_d) \end{aligned}$$

Proposition 7.4

Let $X, Y \in L^1(\Omega, \mathcal{F}, P)$. Then,

1. $aX + bY \in L^1(\Omega, \mathcal{F}, P)$ with $E(aX + bY) = aE(X) + bE(Y)$. **(linearity)**
2. $A \in \mathcal{F} \Rightarrow E(\mathbb{1}_A) = P(A)$
3. $X \geq 0$ almost surely $\Rightarrow E(X) \geq 0$.
4. $X \leq Y$ almost surely $\Rightarrow E(X) \leq E(Y)$.
5. $X \geq 0$ almost surely, $E(X) = 0 \Rightarrow X = 0$ almost surely.
6. $|E(X)| \leq E(|X|)$

7. If $X \sim F$, then
    (i) $E(X) = \int_0^\infty \bar{F}(x)dx - \int_{-\infty}^0 F(x)dx$ where $\bar{F}(x) = 1 - F(x)$.
    (ii) $E(x) = \int_0^1 F^{-}(\alpha)d\alpha$

Proof.

1. algebraic induction.
2. $E(\mathbb{1}_A) = 1 \cdot P(A) + 0 \cdot P(A_c) = P(A)$ since $\mathbb{1}_A$ is simple
3. $E(X) = E(X^+) \geq 0$ since $X \geq 0$ and by algebraic induction.
    (i) If X is simple, $x_j \geq 0 \Rightarrow E(x) = \sum_{i=1}^n x_i P(A_i) \geq 0$.
    (ii) Else, $\sup E(Y) \geq 0$ over simple Y such that $0 \leq Y \leq X$.
4.

$$\begin{aligned} E(Y) &= E(Y - X + X) \\ &= E(Y - X) + E(X) \\ &\geq E(X) \end{aligned}$$

since $Y - X \geq 0$ almost surely, and thus $E(Y - X) \geq 0$.

5. Assume $P(X \geq 0)$. Then,

$$\begin{aligned} \lim_{n\uparrow\infty} P(X > \frac{1}{n}) &= \lim_{n\uparrow\infty} P(X \in (\frac{1}{n}, \infty]) \\ &= P(X \in \bigcup_{i=1}^{\infty} (\frac{1}{n}, \infty]) \\ &= P(X > 0) \\ &> 0 \end{aligned}$$

Therefore, there exists $n_0 \in \mathbb{N} : P(X \geq \frac{1}{n}) > 0$ (*) for all $n \geq n_0$. For $\varepsilon = \frac{1}{n_0} > 0$ (**), we have

$$\begin{aligned} 0 = E(X) &\geq E(X \cdot \mathbb{1}_{\{X \geq \varepsilon\}}) \\ &\geq E(\varepsilon \mathbb{1}_{\{X \geq \varepsilon\}}) \\ &= \varepsilon E(\mathbb{1}_{\{X \geq \varepsilon\}}) \\ &= \varepsilon P(X \geq \varepsilon) \\ &> 0 \end{aligned}$$

by (**) and (*).

6.

$$\begin{aligned} |E(X)| &= |E(X^+) - E(X^-)| \\ &\leq |E(X^+) + |E(X^-)| \\ &= E(X^+) + E(X^-) \\ &= E(X^+ + X^-) \\ &= E(|X|) \end{aligned}$$

7. (i) Use integration by parts formula

$$\int_a^b h(x)dF(x) = [h(x)F(x)]_a^b - \int_a^b F(x)dh(x)$$

and the fact that

$$\int_{\mathbb{R}} h(x)dF(x) = -\int_{\mathbb{R}} h(x)d\bar{F}(x).$$

We also need that

(a)

$$\begin{aligned} 0 &\leq x\bar{F}(x) \\ &= x(1 - F(x)) \\ &= x(F(\infty) - F(x)) \\ &= x\int_x^\infty dF(z) \\ &\geq \int_x^\infty zdF(z) \to 0 \end{aligned}$$

as $x \uparrow \infty$.

(b) $0 \geq xF(x) = \int_{-\infty}^x xdF(z) \geq \int_{-\infty}^x zdF(z) \to 0$ as $x \downarrow -\infty$, therefore,

$$\begin{aligned} E(X) &= \int_{\mathbb{R}} xdF(x) \\ &= \int_0^\infty xdF(x) + \int_{-\infty}^0 xdF(x) \\ &= -\int_0^\infty xd\bar{F}(x) + \int_{-\infty}^0 xdF(x) \\ &= -([x\bar{F}(x)]_0^\infty - \int_0^\infty \bar{F}(x)dx) + [xF(x)]_{-\infty}^0 - \int_{-\infty}^0 F(x)dx \end{aligned}$$

□

Proposition 7.5

If X is quasi-integrable, with $P(A) = 0$, then

$$\int_A XdP := \int_\Omega X\mathbb{1}_A dP = E(X\mathbb{1}_A) = 0$$

Proof. Let $X \geq 0$. Then $X\mathbb{1}_A \geq 0$. Let $Y = \sum_{i=1}^n y_i \mathbb{1}_{B_i}$ be simple, such that $0 \leq Y \leq X\mathbb{1}_A$ for all ω. Then,

$$\begin{aligned}
&0 \leq y_i \mathbb{1}_{B_i} \leq X\mathbb{1}_A \text{ for } i = 1, \dots, n \\
\Rightarrow & y_i = 0 \text{ or } B_i \subseteq A \text{ for } i = 1, \dots, n \\
\Rightarrow & y_i = 0 \text{ or } P(B_i) \leq P(A) = 0 \\
\Rightarrow & y_i = 0 \text{ or } P(B_i) = 0 \\
\Rightarrow & E(Y) = \sum_{i=1}^n y_i P(B_i) \\
\Rightarrow & E(X) = \sup_{0 \leq Y \leq X, Y \text{ simple}} E(Y) = 0
\end{aligned}$$

for all Y. For general X,

$$E(X\mathbb{1}_A) = E(X^+\mathbb{1}_A) - E(X^-\mathbb{1}_A) = 0$$

□

7.2 Variance and Covariance

Definition 7.6

Let $(\Omega, \mathcal{F}, P)$ be a probability space with random variables X, Y such that $E(X^2) < \infty, E(Y^2) < \infty$ (notation: $X, Y \in L^2(\Omega, \mathcal{F}, P)$). Then,

$$\text{Var}(X) := E((X - EX)^2)$$

is called the variance of X (or its distribution or df), and $\sqrt{\text{Var}(X)}$ is the standard deviation.

The covariance of X, Y is defined by

$$\text{Cov}(X, Y) := E((X - EX)(Y - EY))$$

and the correlation of X, Y by

$$\text{Cor}(X, Y) := \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}$$

Remark 7.7

Here are some properties of variance and covariance.

1.

$$\begin{aligned}
\text{Var}(X) &= E(X^2 - 2XE(X) + E(X)^2) \\
&= E(X^2) - 2E(X)E(X) + E(X)^2 \\
&= E(X^2) - E(X)^2
\end{aligned}$$

Similarly,

$$\text{Cov}(X, Y) = E(XY) - E(X)E(Y)$$

Also note that

$$\begin{aligned}\text{Cov}(X,X) &= \text{Var}(X)\\ \text{Cov}(X,c) &= 0\ \forall c \in \mathbb{R}\\ \text{Cov}(X,Y) &= \text{Cov}(Y,X)\end{aligned}$$

2.

$$\begin{aligned}\text{Var}(X) = 0 &\Leftrightarrow E((X-EX)^2) = 0\\ &\Leftrightarrow (X-EX)^2 = 0 \text{ almost surely}\\ &\Leftrightarrow X - EX = 0 \text{ almost surely}\\ &\Leftrightarrow X = EX \text{ almost surely}\end{aligned}$$

3.

$$\begin{aligned}\text{Var}(aX+bY) &= E(((aX+bY)-E(aX+bY))^2)\\ &= E((a(X-EX))^2) + 2E(a(X-EX)b(Y-EY)) + E((b(Y-EY))^2)\\ &= a^2\text{Var}(X) + 2ab\text{Cov}(X,Y) + B^2\text{Var}(Y)\end{aligned}$$

for $a, b \in \mathbb{R}$. In particular if $Y = 1$ almost surely, $\text{Var}(aX+b) = a^2\text{Var}(X)$.

4. If X, Y are independent,

$$E(XY) = E(X)E(Y)$$

which implies

$$\text{Cov}(X,Y) = 0 = Cor(X,Y)$$

So, independence implies uncorrelatedness. The converse is not true in general:

Let $X \sim U(-1,1), Y = X^2$. Then,

$$\begin{aligned}\text{Cov}(X,Y) &= E(XY) - E(X)E(Y)\\ &= E(X^3) - 0E(X^2)\\ &= 0\end{aligned}$$

by symmetry, since X^3 is an odd function. So X, Y are uncorrelated but dependent.

Proposition 7.8

Let $X, Y \in L^2(\Omega, \mathcal{F}, P)$. Then,

$$\begin{aligned}&\rho := \text{Cor}(X,Y) \in [-1,1] \text{ and } |\rho| = 1\\ \Leftrightarrow &Y \text{ is a linear function of } X \text{ with slope } \lessgtr 0\\ \Leftrightarrow &\rho = \pm 1\end{aligned}$$

Proof. (⇒) Let $z_t = tX + Y, t \in \mathbb{R}$. Let $a = t^2\mathrm{Var}(X), b = 2t\mathrm{Cov}(X,Y), c = Var(Y)$. Then,

$$\begin{aligned} 0 &\leq \mathrm{Var}(z_t) = t^2\mathrm{Var}(X) + 2t\mathrm{Cov}(X,Y) + Var(Y) \\ \Rightarrow & 0 \geq b^2 - 4ac \\ &= (2\mathrm{Cov}(X,Y))^2 - 4\mathrm{Var}(X)\mathrm{Var}(Y) \\ \Rightarrow & |\rho| = \sqrt{\rho^2} = \sqrt{\frac{\mathrm{Cov}(X,Y)^2}{\mathrm{Var}(X)\mathrm{Var}(Y)}} \leq 1 \end{aligned}$$

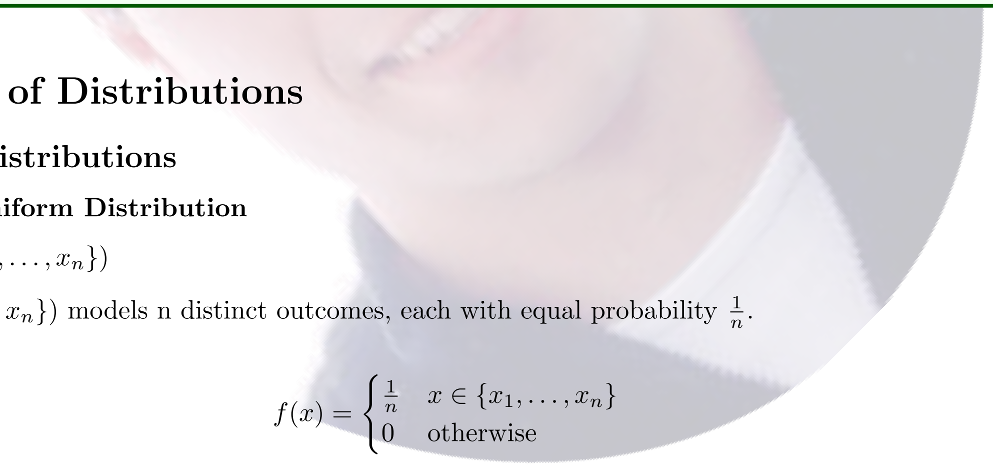
Furthermore, $|\rho| = 1$ iff $b^2 - 4ac = 0$

$$\begin{aligned} \Rightarrow & \exists! \tilde{t} \in \mathbb{R} : \mathrm{Var}(z_{\tilde{t}}) = 0 \\ \Rightarrow & \tilde{t}X + Y = c \in \mathbb{R} \text{ almost surely} \\ \Rightarrow & Y = c - \tilde{t}X \text{ almost surely} \end{aligned}$$

In this case,

$$\begin{aligned} \rho &= \frac{\mathrm{Cov}(X, c - \tilde{t}X}{\sqrt{\mathrm{Var}(X)Var(c - \tilde{t}X}} \\ &= \frac{-\tilde{t}\mathrm{Cov}(X,X)}{\sqrt{\mathrm{Var}(X)\tilde{t}^2\mathrm{Var}(X)}} \\ &= \frac{\tilde{t}\mathrm{Var}(X)}{|\tilde{t}|\mathrm{Var}(X)} \\ &= \frac{-\tilde{t}}{|\tilde{t}|} \end{aligned}$$

So, $-\tilde{t} \lessgtr 0$ iff $\rho = \pm 1$. (⇐) is trivial. □



8 Examples of Distributions

8.1 Discrete Distributions

8.1.1 Discrete Uniform Distribution

- Notation:$U(\{x_1, \ldots, x_n\})$
- $X \sim U(\{x_1, \ldots, x_n\})$ models n distinct outcomes, each with equal probability $\frac{1}{n}$.
- PMF:

$$f(x) = \begin{cases} \frac{1}{n} & x \in \{x_1, \ldots, x_n\} \\ 0 & \text{otherwise} \end{cases}$$

Check: $f(x) \geq 0$ for $x \in \mathbb{R}$, and $\sum_{i=1}^{n} f(x_n) = \frac{1}{n}n = 1$

- Distribution function:

$$\begin{aligned} F(x) &= P(X \leq x) \\ &= \sum_{k \in \{1, \ldots, n\} : x_k \leq x} P(X = x_k) \\ &= \frac{1}{n} |\{x_k : x_k \leq x\}| \end{aligned}$$

  If $x_k = k$, we have

$$F(x) = \begin{cases} 0 & x < 1 \\ \frac{\lfloor x \rfloor}{n} & x \in [1, n) \\ 1 & x \geq n \end{cases}$$

- Mean: We have $E(X) = \sum_{k=1}^n x_k P(X = x_k) = \frac{1}{n} \sum_{k=1}^n x_k$. If $x_k = k$, then $E(X) = \frac{n+1}{2}$.
- Variance: We have $\text{Var}(X) = E(X^2) - E(X)^2 = \frac{1}{n} \sum_{k=1}^n x_k^2 - (\frac{1}{n} \sum_{k=1}^n x_k)^2$. If $x_k = k$, then $\text{Var}(X) = \frac{1}{n} \frac{n(n+1)(2n+1)}{6} - (\frac{n+1}{2})^2 = \frac{n^2-1}{12}$.

Remark

If $U \sim U(0, 1)$, then $\lceil nU \rceil \sim U(\{1, \ldots, n\})$ since

$$\begin{aligned} P(nU = k) &= P(k - 1 < nU \leq k) \\ &= P(U \in (\frac{k-1}{n}, \frac{k}{n}]) \\ &= \frac{k}{n} - \frac{k-1}{n} \\ &= \frac{1}{n} \end{aligned}$$

for $k = 1, \ldots, n$. In particular, $X = x_{\lceil nU \rceil} \sim U(\{1, \ldots, n\})$

8.1.2 Binomial Distribution

- Notation: $B(n, p)$ where $n \geq 1, p \in (0, 1)$.
- $X \sim B(n, p)$ models the number of successes when independently repeating same experiment with outcomes success or failure n times, where $P(\text{success}) = p$. These experiments are Bernoulli trials. Then, $X = \sum_{k=1}^n X_k$ for $X_1, \ldots, X_n \sim B(1, p)$.
- PMF:

$$f(x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & x \in \{0, \ldots, n\} \\ 0 & \text{otherwise} \end{cases}$$

  Check: $f(x) \geq 0$ for $x \in \mathbb{R}$, and

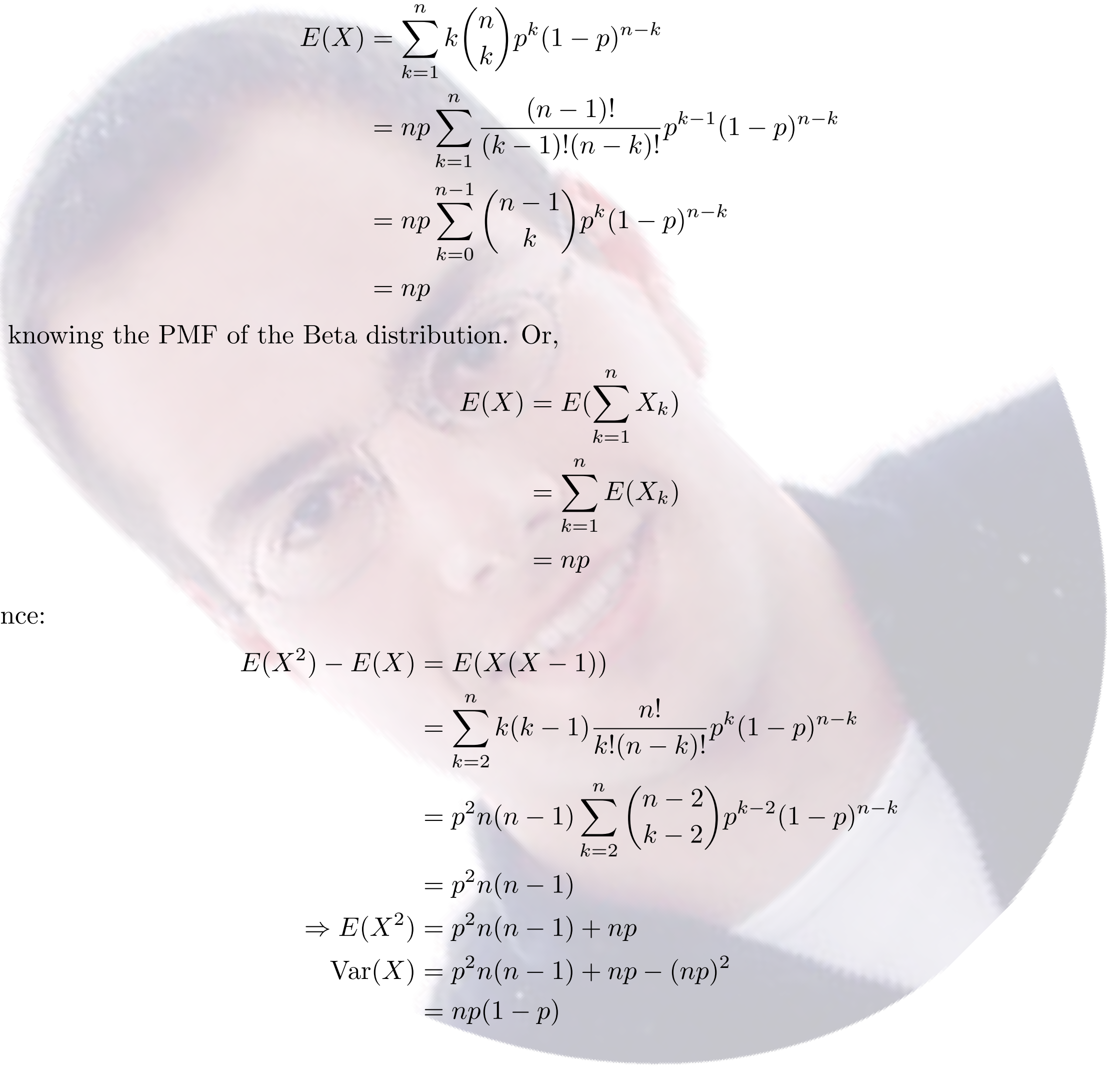
$$\begin{aligned} \sum_{k=0}^n \binom{n}{x} p^k (1-p)^k &= (p + (1-p))^n \\ &= 1 \end{aligned}$$

- Distribution function:

$$\begin{aligned}F(X) &= P(X \leq x)\\ &= P(X \leq \lfloor x \rfloor)\\ &= \sum_{k=0}^{\lfloor x \rfloor} \binom{n}{x} p^k (1-p)^{n-k}\\ &= \int_p^1 f_{x,n}(z)dz\end{aligned}$$

where $f_{x,n}$ is the density of the Beta$(x+1, n-x)$ distribution. By letting $p = 0$, this gives 1 for $x \in [0, n]$.

- Mean:



$$\begin{aligned}E(X) &= \sum_{k=1}^{n} k \binom{n}{k} p^k (1-p)^{n-k}\\ &= np \sum_{k=1}^{n} \frac{(n-1)!}{(k-1)!(n-k)!} p^{k-1} (1-p)^{n-k}\\ &= np \sum_{k=0}^{n-1} \binom{n-1}{k} p^k (1-p)^{n-k}\\ &= np\end{aligned}$$

From knowing the PMF of the Beta distribution. Or,

$$\begin{aligned}E(X) &= E(\sum_{k=1}^{n} X_k)\\ &= \sum_{k=1}^{n} E(X_k)\\ &= np\end{aligned}$$

- Variance:

$$\begin{aligned}E(X^2) - E(X) &= E(X(X-1))\\ &= \sum_{k=2}^{n} k(k-1) \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}\\ &= p^2 n(n-1) \sum_{k=2}^{n} \binom{n-2}{k-2} p^{k-2} (1-p)^{n-k}\\ &= p^2 n(n-1)\\ \Rightarrow E(X^2) &= p^2 n(n-1) + np\\ \mathrm{Var}(X) &= p^2 n(n-1) + np - (np)^2\\ &= np(1-p)\end{aligned}$$

Or,

$$\begin{aligned}\mathrm{Var}(X) &= \mathrm{Var}(\sum_{k=1}^{n} X_k)\\ &= \sum_{k=1}^{n} Var(X_k) + 2 \sum_{1 \leq k \leq l \leq n} \mathrm{Cov}(X_k, X_l)\\ &= np(1-p)\end{aligned}$$

since $\mathrm{Var}(X_k) = E(X_k^2) - E(X_k)^2 = p - p^2$

Remark

If you participate in weekly lottery, with 2% probability of winning each week, then the number of wins in one year is $B(52, 0.02)$ distributed.

8.1.3 Geometric Distribution

- Notation: $Geo(p), p \in (0, 1)$
- $X \sim Geo(p)$ models the number of independent Bernoulli trials with success probability p until first success.
- PMF:

$$f(x) = \begin{cases} p(1-p)^{x-1} & x \in \mathbb{N} \\ 0 & \text{otherwise} \end{cases}$$

  Check: $f(x) \geq 0$ for $x \in \mathbb{R}$. $\sum_{k=1}^{\infty} p(1-p)^{k-1} = p \sum_{k=0}^{\infty} (1-p)^{k-1} = p\frac{1}{1-(1-p)} = 1$

- Distribution function:

$$\begin{aligned} F(x) &= P(X \leq \lfloor x \rfloor) \\ &= \sum_{k=1}^{\lfloor x \rfloor} p(1-p)^{k-1} \\ &= p \sum_{k=0}^{\lfloor x-1 \rfloor} p(1-p)^{k-1+1} \\ &= p\frac{1-(1-p)^{\lfloor x \rfloor}}{1-(1-p)} \\ &= 1-(1-p)^{\lfloor x \rfloor} \end{aligned}$$

  for $x \in [1, \infty)$.

- Mean: Note that

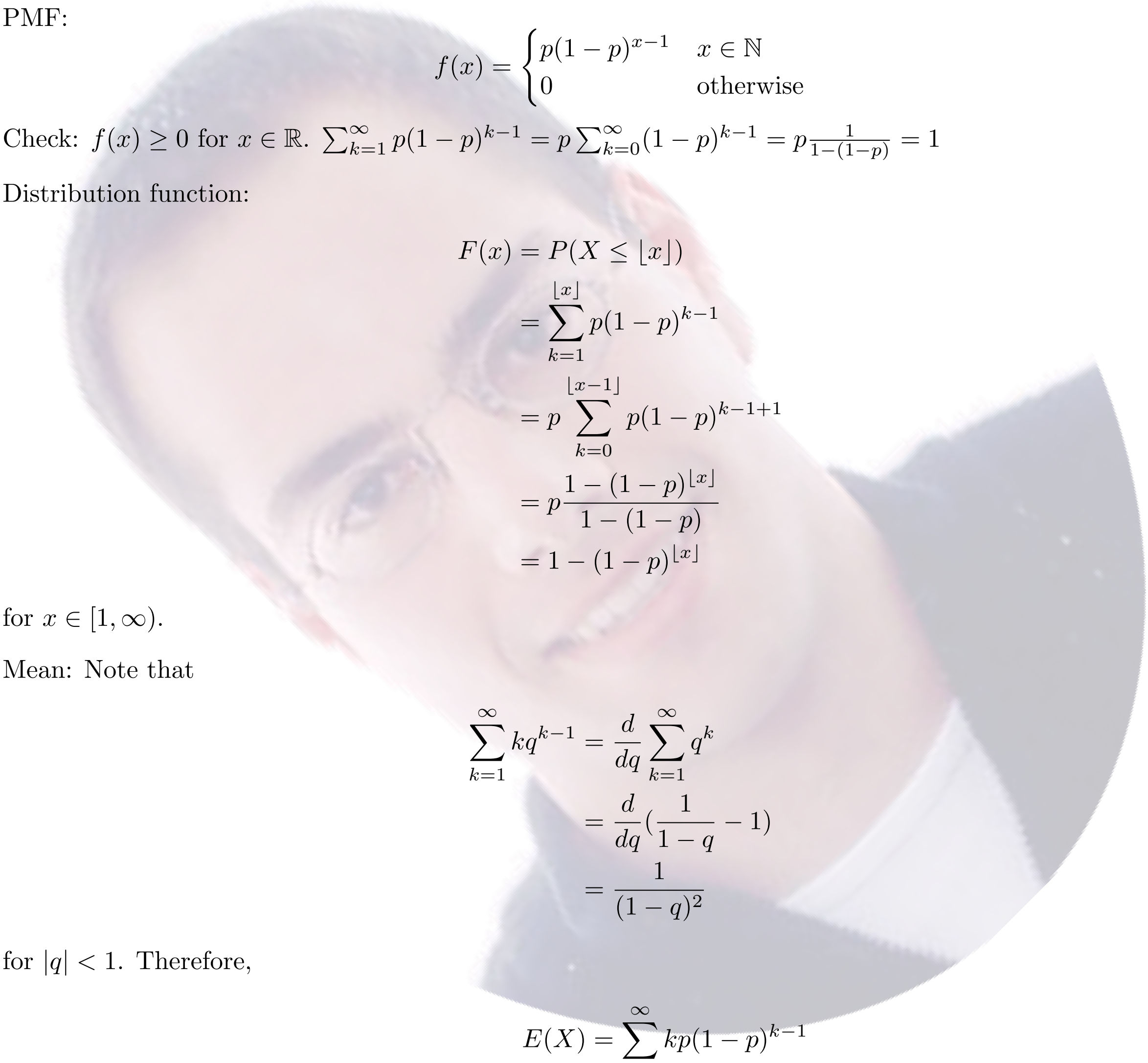
$$\begin{aligned} \sum_{k=1}^{\infty} kq^{k-1} &= \frac{d}{dq} \sum_{k=1}^{\infty} q^k \\ &= \frac{d}{dq}(\frac{1}{1-q} - 1) \\ &= \frac{1}{(1-q)^2} \end{aligned}$$

  for $|q| < 1$. Therefore,

$$E(X) = \sum_{k=1}^{\infty} kp(1-p)^{k-1}$$

$$p\frac{1}{1-(1-p)^2}$$

$$= \frac{1}{p}$$

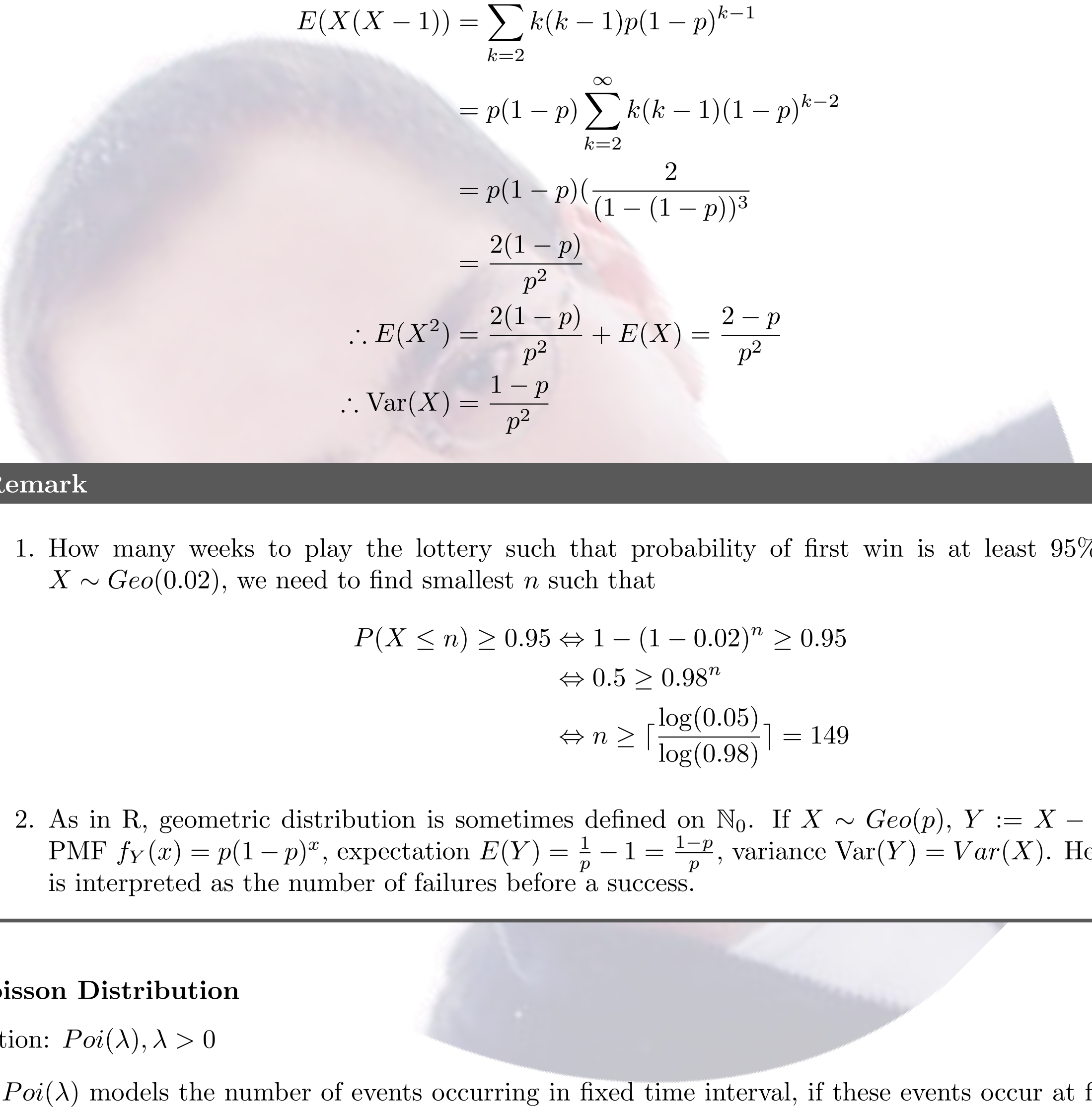


- Variance: Note that

$$\begin{aligned}\sum_{k=2}^{\infty} k(k-1)q^{k-2} &= \sum_{k=0}^{\infty} \frac{d^2}{dq^2} q^k \\ &= \frac{d^2}{dq^2}(\frac{1}{1-q} - q - 1) \\ &= \frac{2}{(1-q)^3}\end{aligned}$$

for $|q| < 1$. Therefore,

$$\begin{aligned}E(X(X-1)) &= \sum_{k=2}^{\infty} k(k-1)p(1-p)^{k-1} \\ &= p(1-p)\sum_{k=2}^{\infty} k(k-1)(1-p)^{k-2} \\ &= p(1-p)(\frac{2}{(1-(1-p))^3} \\ &= \frac{2(1-p)}{p^2} \\ \therefore E(X^2) &= \frac{2(1-p)}{p^2} + E(X) = \frac{2-p}{p^2} \\ \therefore \mathrm{Var}(X) &= \frac{1-p}{p^2}\end{aligned}$$



Remark

1. How many weeks to play the lottery such that probability of first win is at least 95%? If $X \sim Geo(0.02)$, we need to find smallest n such that

$$\begin{aligned}P(X \leq n) \geq 0.95 &\Leftrightarrow 1 - (1-0.02)^n \geq 0.95 \\ &\Leftrightarrow 0.5 \geq 0.98^n \\ &\Leftrightarrow n \geq \lceil \frac{\log(0.05)}{\log(0.98)} \rceil = 149\end{aligned}$$

2. As in R, geometric distribution is sometimes defined on $\mathbb{N}_0$. If $X \sim Geo(p)$, $Y := X - 1$ has PMF $f_Y(x) = p(1-p)^x$, expectation $E(Y) = \frac{1}{p} - 1 = \frac{1-p}{p}$, variance $\mathrm{Var}(Y) = Var(X)$. Here, Y is interpreted as the number of failures before a success.

8.1.4 Poisson Distribution

- Notation: $Poi(\lambda), \lambda > 0$
- $X \sim Poi(\lambda)$ models the number of events occurring in fixed time interval, if these events occur at fixed rate λ and independently of time of last event.
- PMF:

$$f(x) = \begin{cases} \frac{\lambda^x}{x!}e^{-\lambda} & x \in \mathbb{N}_0 \\ 0 & \text{otherwise} \end{cases}$$

Check: $f(x) \geq 0$ for $x \in \mathbb{R}$. $\sum_{x=0}^{\infty} f(x) = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} e^{-\lambda} = e^{\lambda} e^{-\lambda} = 1$

- Distribution function:

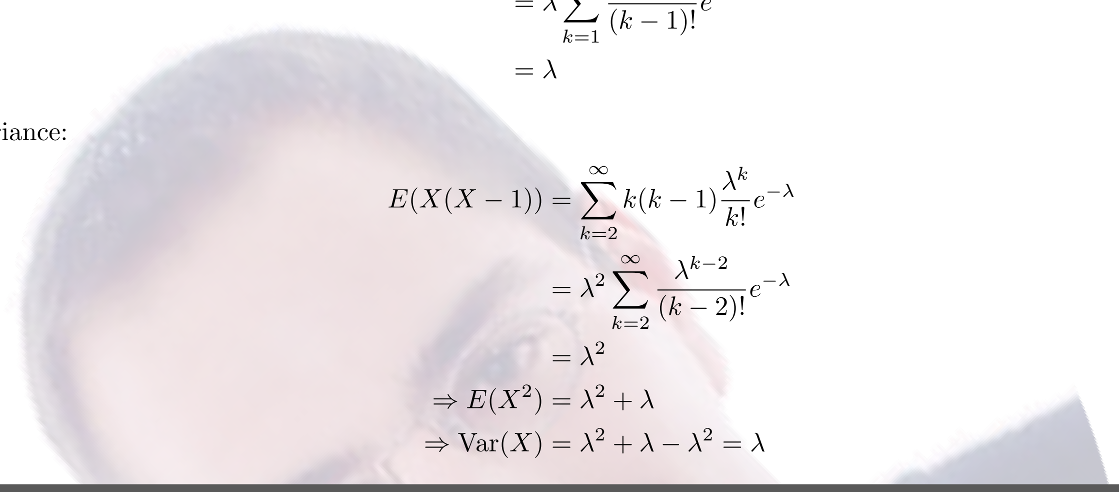
$$F(x) = \sum_{k=0}^{\lfloor x \rfloor} \frac{\lambda^k}{k!} e^{-\lambda}$$

for $x \in \mathbb{R}$. R uses $F(x) = \frac{\Gamma(\lfloor x \rfloor + 1, \lambda)}{\lfloor x \rfloor !}$, where $\Gamma(s,z) = \int_z^\infty t^{s-1} e^{-t} dt$ is the upper incomplete gamma function.

- Mean:

$$\begin{aligned} E(X) &= \sum_{k=1}^{\infty} k \frac{\lambda^k}{k!} e^{-\lambda} \\ &= \lambda \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} e^{-\lambda} \\ &= \lambda \end{aligned}$$

- Variance:



$$\begin{aligned} E(X(X-1)) &= \sum_{k=2}^{\infty} k(k-1) \frac{\lambda^k}{k!} e^{-\lambda} \\ &= \lambda^2 \sum_{k=2}^{\infty} \frac{\lambda^{k-2}}{(k-2)!} e^{-\lambda} \\ &= \lambda^2 \\ \Rightarrow E(X^2) &= \lambda^2 + \lambda \\ \Rightarrow \mathrm{Var}(X) &= \lambda^2 + \lambda - \lambda^2 = \lambda \end{aligned}$$

Remark

1. What is probability of obtaining more than 12 emails per hour if you receive 240 emails per day on average?

$$\begin{aligned} \lambda &= E(X) \\ &= \frac{240}{24} \\ &= 10 \Rightarrow P(X > 12) && = \sum_{k=13}^{\infty} \frac{10^k}{k!} e^{-10} \\ &\approx 22.84\% \end{aligned}$$

In R, this can be calculated using

```
ppois(12, lambda=10, lower.tail=FALSE)
```

2. Suppose we divide the time interval in (1) into n equal parts, and assume one only receives 0 or 1 emails in each of the n intervals (1 email with probability p). Then for $n \to \infty, p \downarrow 0$ such that $np \to \lambda$, we obtain

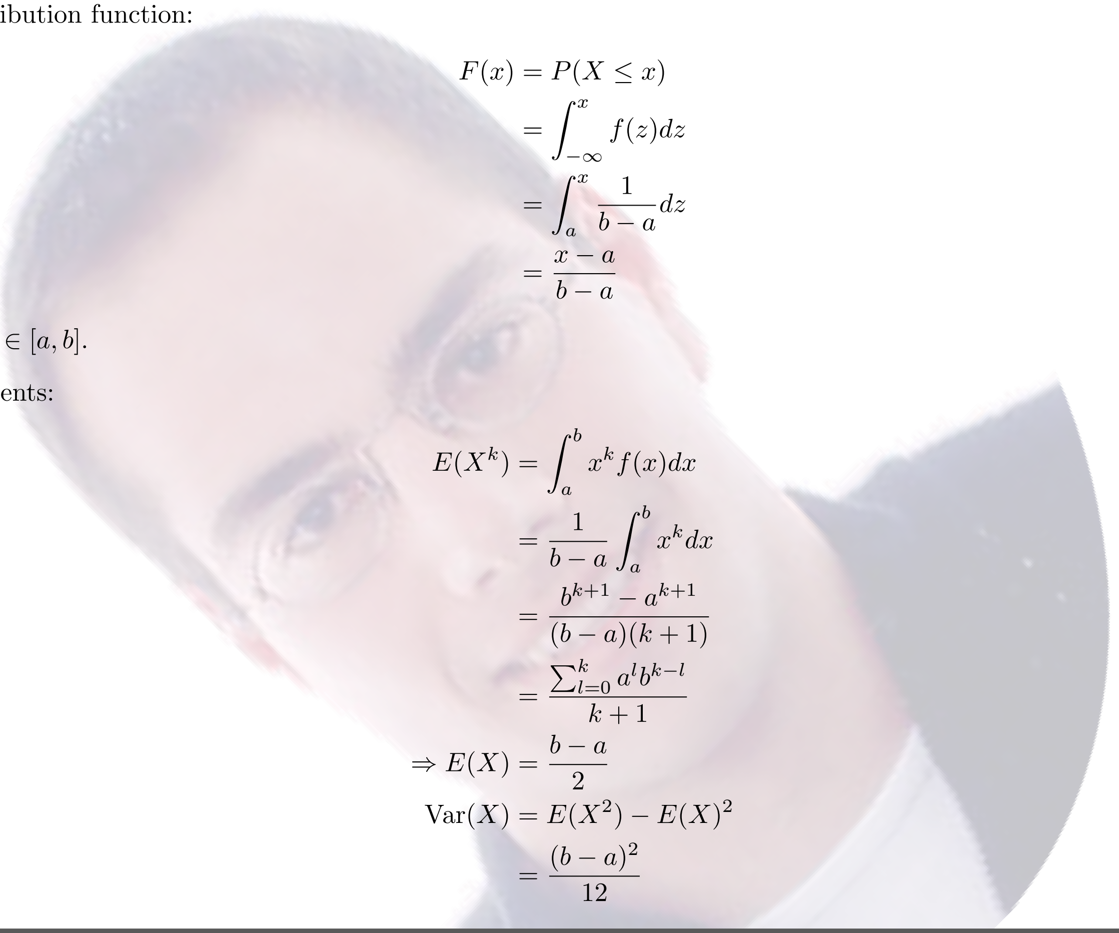
$$\begin{aligned} P(X = x) &= \binom{n}{x} p^x (1-p)^{n-x} \\ &= \frac{n(n-1)\dots(n-x+1)}{n^x} \frac{(np)^x}{x!} (1 - \frac{np}{n})(1-p)^{-x} \end{aligned}$$

(***) Which is the $Poi(\lambda)$ distribution as the limit of the $B(n,p)$ distribution. Note: if $a_n \to a$ as $n \to \infty$, $(1 + \frac{a_n}{n})^n \to e^a$ as $n \to \infty$.

8.2 Absolutely Continuous Distributions

8.2.1 Continuous Uniform Distribution

- Notation: $U(a,b)$ for $a, b \in \mathbb{R}, a < b$.
- $X \sim U(a,b)$ models outcomes uniformly distributed over (a,b), that is, $P(X \in (x, x+h])$ is constant and equals $\frac{h}{b-a}$ for all $x \in [a, b-h]$. See Example 5.17 (1).
- Density:
$$f(x) = \frac{1}{b-a}\mathbb{1}_{(a,b]}(x)$$
for $x \in \mathbb{R}$. Check: $f(x) \geq 0$ for $x \in \mathbb{R}$. Also, $\int_a^b \frac{1}{b-a}dz = \frac{1}{b-a}\int_a^b (1)dz = \frac{b-a}{b-a} = 1$.



- Distribution function:
$$\begin{aligned} F(x) &= P(X \leq x) \\ &= \int_{-\infty}^{x} f(z)dz \\ &= \int_a^x \frac{1}{b-a}dz \\ &= \frac{x-a}{b-a} \end{aligned}$$
for $x \in [a,b]$.
- Moments:
$$\begin{aligned} E(X^k) &= \int_a^b x^k f(x)dx \\ &= \frac{1}{b-a}\int_a^b x^k dx \\ &= \frac{b^{k+1}-a^{k+1}}{(b-a)(k+1)} \\ &= \frac{\sum_{l=0}^{k} a^l b^{k-l}}{k+1} \\ \Rightarrow E(X) &= \frac{b-a}{2} \\ \text{Var}(X) &= E(X^2) - E(X)^2 \\ &= \frac{(b-a)^2}{12} \end{aligned}$$

Remark

1. $U(0,1)$ is the standard uniform distribution.
2. It is an important building block for other distributions. For example, it is used in
   - Inversion method ($F^{-}(U) \sim F$)
   - Sklar's Theorem (where we used $F(X) \sim U(0,1)$ if F is continuous and $X \sim F$)
   - Sampling from $U(\{1, \dots, n\})$
3. $U \sim U(0,1) \Rightarrow X : a + (b-a)U \sim U(a,b)$ is the stochastic representation of $U(a,b)$. (***)

8.2.2 Gamma Distribution

- Notation: $\Gamma(\alpha, \beta)$ where $\alpha > 0$ is the shape, $\beta > 0$ is the rate.
- Special cases:
  - Exponential distribution
  - Erlang distribution
  - Chi-squared distribution
- Density:

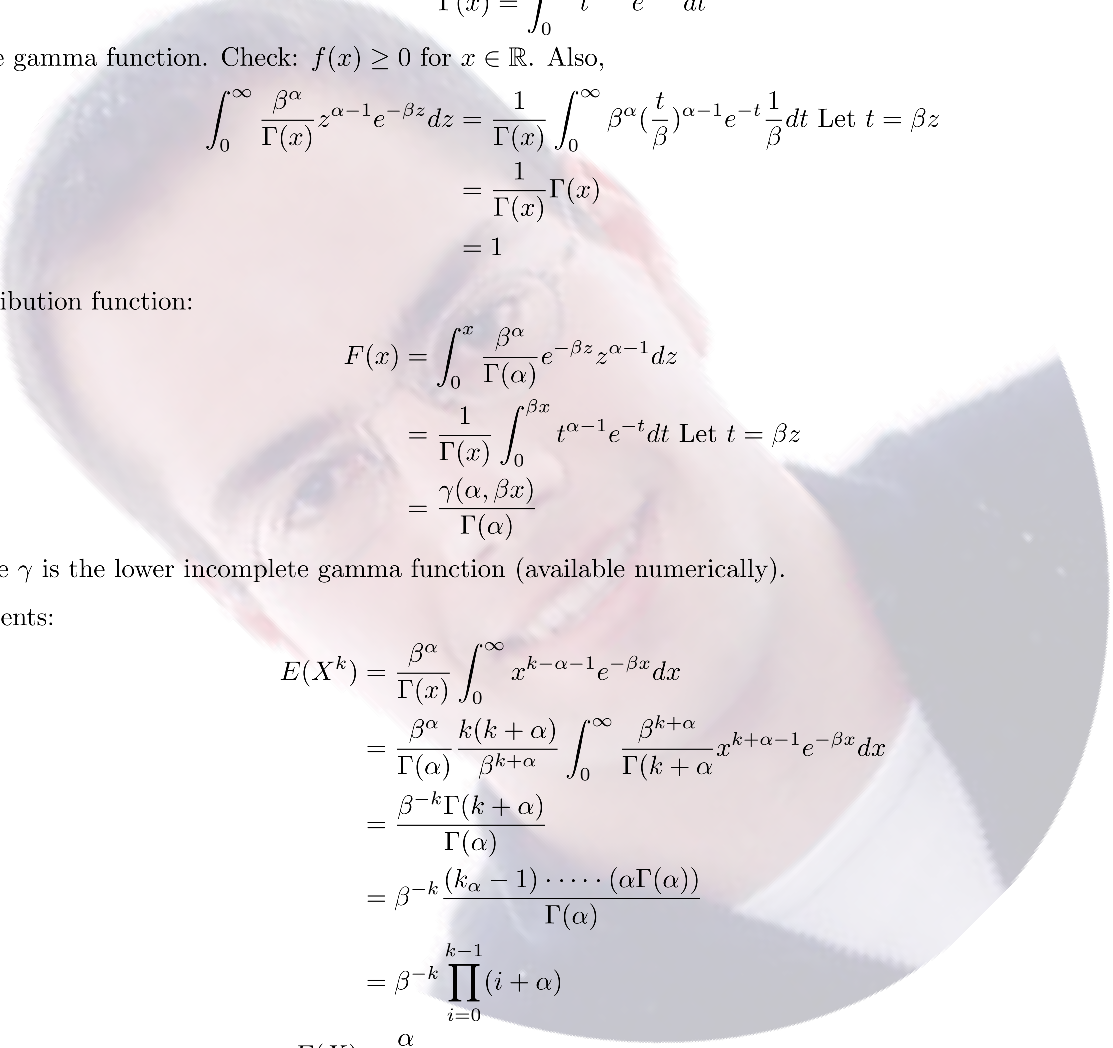
$$f(x) = \frac{\beta^x}{\Gamma(x)} x^{\alpha-1} e^{-\beta x}$$

for $x > 0$ where

$$\Gamma(x) = \int_0^\infty t^{\alpha-1} e^{-\infty} dt$$

is the gamma function. Check: $f(x) \geq 0$ for $x \in \mathbb{R}$. Also,

$$\begin{aligned} \int_0^\infty \frac{\beta^\alpha}{\Gamma(x)} z^{\alpha-1} e^{-\beta z} dz &= \frac{1}{\Gamma(x)} \int_0^\infty \beta^\alpha (\frac{t}{\beta})^{\alpha-1} e^{-t} \frac{1}{\beta} dt \text{ Let } t = \beta z \\ &= \frac{1}{\Gamma(x)} \Gamma(x) \\ &= 1 \end{aligned}$$



- Distribution function:

$$\begin{aligned} F(x) &= \int_0^x \frac{\beta^\alpha}{\Gamma(\alpha)} e^{-\beta z} z^{\alpha-1} dz \\ &= \frac{1}{\Gamma(x)} \int_0^{\beta x} t^{\alpha-1} e^{-t} dt \text{ Let } t = \beta z \\ &= \frac{\gamma(\alpha, \beta x)}{\Gamma(\alpha)} \end{aligned}$$

where γ is the lower incomplete gamma function (available numerically).

- Moments:

$$\begin{aligned} E(X^k) &= \frac{\beta^\alpha}{\Gamma(x)} \int_0^\infty x^{k-\alpha-1} e^{-\beta x} dx \\ &= \frac{\beta^\alpha}{\Gamma(\alpha)} \frac{k(k+\alpha)}{\beta^{k+\alpha}} \int_0^\infty \frac{\beta^{k+\alpha}}{\Gamma(k+\alpha} x^{k+\alpha-1} e^{-\beta x} dx \\ &= \frac{\beta^{-k} \Gamma(k+\alpha)}{\Gamma(\alpha)} \\ &= \beta^{-k} \frac{(k_\alpha - 1) \cdot \cdots \cdot (\alpha \Gamma(\alpha))}{\Gamma(\alpha)} \\ &= \beta^{-k} \prod_{i=0}^{k-1} (i+\alpha) \\ \Rightarrow E(X) &= \frac{\alpha}{\beta} \\ \mathrm{Var}(X) &= E(X^2) - E(X)^2 \\ &= \frac{\alpha(1+\alpha)}{\beta^2} - (\frac{\alpha}{\beta})^2 \\ &= \frac{\alpha}{\beta} \end{aligned}$$

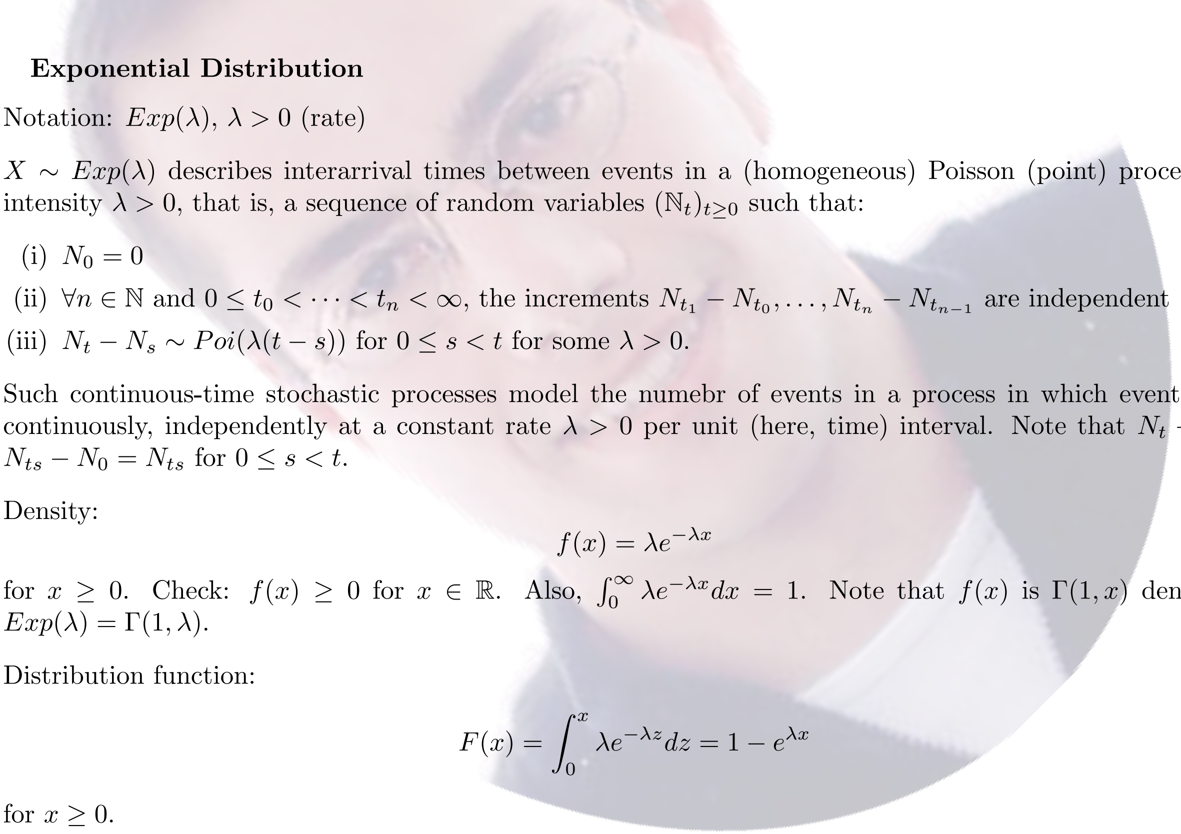
Remark

1. $Y \sim \Gamma(\alpha, 1) \Rightarrow X = \frac{Y}{\beta} \sim \Gamma(\alpha, \beta)$ since

$$\begin{aligned} F_X(x) &= P(X \leq x) \\ &= P(\frac{Y}{\beta} \leq x) \\ &= P(Y \leq \beta x) \\ &= F_Y(x) \\ &= \int_0^{\beta x} \frac{1}{\Gamma(x)} t^{\alpha-1} e^{-t} dt \text{ Let } t = \beta z \\ &= \int_0^{x} \frac{\beta^\alpha}{\Gamma(x)} z^{\alpha-1} e^{-\beta z} dz \end{aligned}$$

2. Parametrization: $f(x) = \frac{1}{\Gamma(x)s^\alpha} x^{\alpha-1} e^{-\frac{x}{s}}$ for $s > 0$ is also used frequently. See also R's

```
dgamma(x, shape, rate=1, scale=1/rate)
```



8.2.3 Exponential Distribution

- Notation: $Exp(\lambda)$, $\lambda > 0$ (rate)
- $X \sim Exp(\lambda)$ describes interarrival times between events in a (homogeneous) Poisson (point) process with intensity $\lambda > 0$, that is, a sequence of random variables $(\mathbb{N}_t)_{t \geq 0}$ such that:

  (i) $N_0 = 0$

  (ii) $\forall n \in \mathbb{N}$ and $0 \leq t_0 < \cdots < t_n < \infty$, the increments $N_{t_1} - N_{t_0}, \ldots, N_{t_n} - N_{t_{n-1}}$ are independent

  (iii) $N_t - N_s \sim Poi(\lambda(t - s))$ for $0 \leq s < t$ for some $\lambda > 0$.

  Such continuous-time stochastic processes model the numebr of events in a process in which events occur continuously, independently at a constant rate $\lambda > 0$ per unit (here, time) interval. Note that $N_t - N_s = N_{ts} - N_0 = N_{ts}$ for $0 \leq s < t$.

- Density:

$$f(x) = \lambda e^{-\lambda x}$$

  for $x \geq 0$. Check: $f(x) \geq 0$ for $x \in \mathbb{R}$. Also, $\int_0^\infty \lambda e^{-\lambda x} dx = 1$. Note that $f(x)$ is $\Gamma(1, x)$ density, so $Exp(\lambda) = \Gamma(1, \lambda)$.

- Distribution function:

$$F(x) = \int_0^x \lambda e^{-\lambda z} dz = 1 - e^{\lambda x}$$

  for $x \geq 0$.

- Moments:

$$\begin{aligned} E(X^k) &= \lambda^{-k} \prod_{i=0}^{k-1}(i+1) \\ &= \frac{k!}{\lambda^k} \\ \Rightarrow E(X) &= \frac{1}{\lambda} \\ \mathrm{Var}(X) &= \frac{2}{\lambda^2} - (\frac{1}{\lambda})^2 = \frac{1}{\lambda^2} \end{aligned}$$

Remark

1. $U \sim U(0,1) \Rightarrow X = \frac{-1}{\lambda}\log(U) \sim Exp(\lambda)$ is the stochastic representation.
2. $X \sim Exp(\lambda) \Rightarrow Y = \lceil X \rceil \sim Geo(1-e^{-\lambda})$ since

$$\begin{aligned} f_Y(k) &= P(\lceil X \rceil = k) \\ &= P(X \in (k-1, k]) \\ &= F_X(k) - F_X(k-1) \\ &= 1 - e^{-\lambda k} - (1 - e^{-\lambda(k-1)}) \\ &= (e^{-\lambda})^{k-1}(1-e^{-\lambda}) \end{aligned}$$

   for $k \in \mathbb{N}$.
3. $Exp(\lambda)$ is the only continuous memoryless distribution in the sense that $X \sim Exp(\lambda)$ satisfies

$$P(X > s + t | X > s) = P(X > t)$$

   for $s, t \geq 0$. Similarly, $Geo(p)$ is the only discrete memoryless distribution.
4. If $(N_t)_{t\geq 0}$ is a Poisson process with intensity $\lambda > 0$, let

$$T_k := \min\{t \geq 0 : N_t = k\}$$

   denote arrival time of k-th event, $k \in \mathbb{N}_0$. Then,

$$\begin{aligned} &P(\text{"interarrival time between k-th and (k+1)-th event is} > x\text{"}) \\ &\qquad = P(T_{k+1} - T_k > x) \\ &\qquad = P(N_{T_{k+x}} = k) \\ &\qquad = P(N_{T_k+x} = N_{T_k}) \\ &\qquad = P(N_{T_k+x} - N_{T_k} = 0) \\ &\qquad = P(N_x = 0) \\ &\qquad = \frac{(\lambda x)^0}{0!}e^{-\lambda x} \\ &\qquad = e^{-\lambda x} \\ &\Rightarrow T_{k+1} - T_k \sim Exp(\lambda) \end{aligned}$$

   for $k \in \mathbb{N}_0$. So, interarrival times in a Poisson process with intensity $\lambda > 0$ are $Exp(\lambda)$ distributed from this stochastic process.

5. For a sequence $(Y_k)_{k\in\mathbb{N}}$ of non-negative random variables, the compound Poisson process

$$X_t = \sum_{k=1}^{N_t} Y_k$$

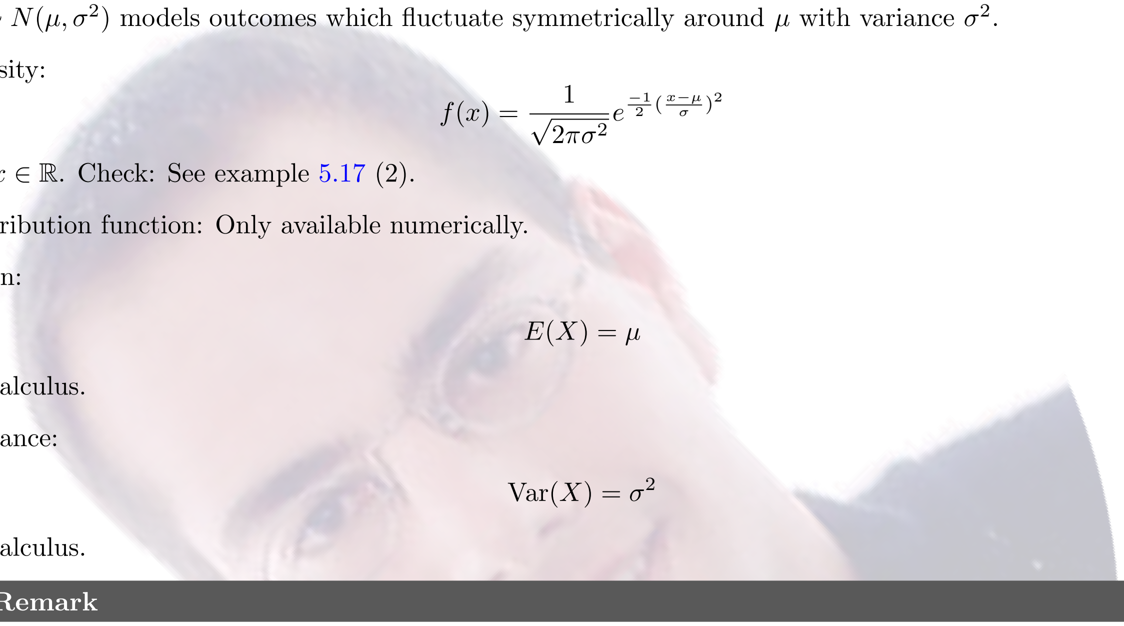
is an important stochastic process in applications.

8.2.4 Normal Distribution

- Notation: $N(\mu, \sigma^2)$ where $\mu \in \mathbb{R}$ is the mean/location, and $\sigma > 0$ is the standard deviation/scale.
- $X \sim N(\mu, \sigma^2)$ models outcomes which fluctuate symmetrically around μ with variance σ^2.
- Density:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-1}{2}(\frac{x-\mu}{\sigma})^2}$$

for $x \in \mathbb{R}$. Check: See example 5.17 (2).



- Distribution function: Only available numerically.
- Mean:

$$E(X) = \mu$$

by calculus.

- Variance:

$$\mathrm{Var}(X) = \sigma^2$$

by calculus.

Remark

1. $N(0,1)$ is known as the standard normal distribution, and its df and density are denoted by Φ and ϕ respectively.
2. $Z \sim N(0,1) \Leftrightarrow X := \mu + \sigma Z \sim N(\mu, \sigma^2)$ since

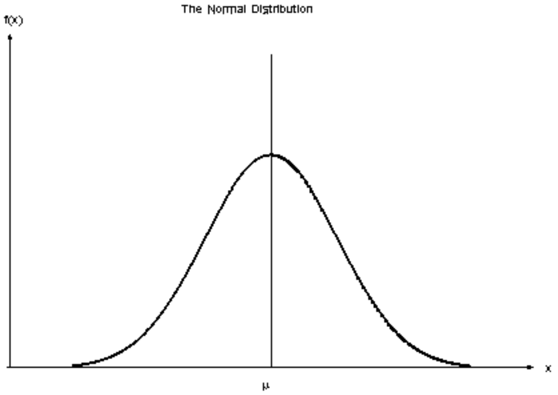
$$\begin{aligned} F_X(x) &= P(X \leq x) \\ &= P(\mu + \sigma Z \leq x) \\ &= P(z \leq \frac{x-\mu}{\sigma}) \\ &= \Phi(\frac{x-\mu}{\sigma}) \end{aligned}$$

for $x \in \mathbb{R}$

$$\begin{aligned} \Leftrightarrow f_X(x) &= \phi(\frac{x-\mu}{\sigma})\frac{1}{\sigma} \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-1}{2}(\frac{x-\mu}{\sigma})^2} \end{aligned}$$

which is the density of $N(\mu, \sigma^2)$. Since $E(Z) = 0$, $\mathrm{Var}(Z) = 1$, we obtain $E(X) = \mu + \sigma Z$ and $\mathrm{Var}(X) = \sigma^2 \mathrm{Var}(Z) = \sigma^2$

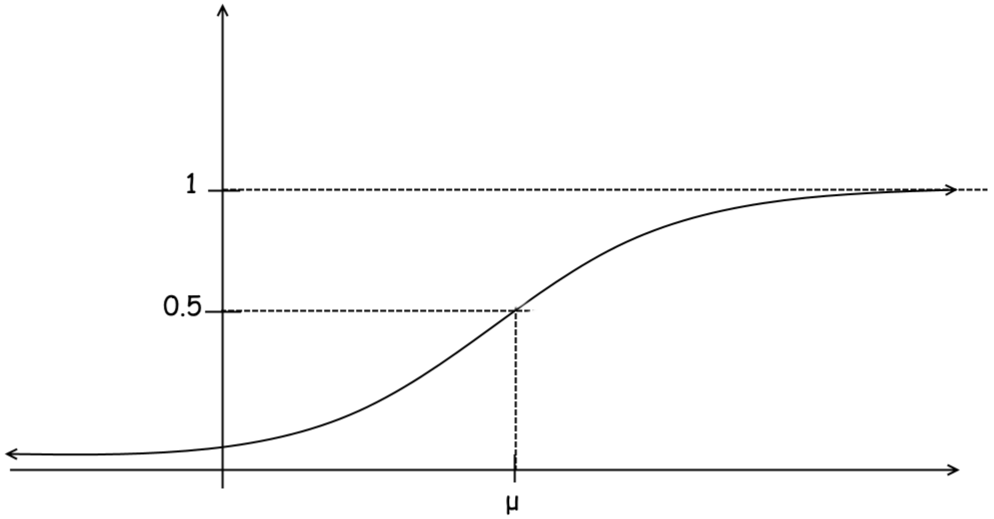
3. Sketch of f (bell curve):



Since $f(\mu + x) = f(\mu - x)$ for $x \in \mathbb{R}$,

$$F(\mu - x) = 1 - F(\mu + x)$$

for $x \in \mathbb{R}$ by calculus. We can sketch $F(x)$:



Therefore to evaluate distribution function of F of $N(\mu, \sigma^2)$ it suffices to know $\Phi(x)$ for all $x \in \mathbb{R}$.

8.3 Multivariate Distributions

8.3.1 Mean vector, Covariance and Correlation Matrices

Definition 8.1

Let $\underline{X} = (X_1, \ldots, X_d)$. If $E(|X_j|) \leq \infty$ for all j, the mean vector or expectation of $\underline{X}$(or its distribution function or distribution) is defined

$$\underline{\mu} = E(\underline{X}) = (E(X_1), \ldots, E(X_d))$$

If $E(X_j) < \infty$ for all j, the covariance and correlation matrices is defined by

$$\Sigma = \mathrm{Cov}(\underline{X}) = (\mathrm{Cov}(X_i, X_j))_{i,j=1,\ldots,d}$$
$$P = \mathrm{Cor}(\underline{X}) = (\mathrm{Cor}(X_i, X_j))_{i,j=1,\ldots,d}$$

Lemma 8.2

1. $E(A\underline{X} + \underline{b}) = AE(\underline{X}) + \underline{b}$. In particular, $E(\underline{a}^T \underline{X}) = \underline{a}^T E(\underline{X})$.
2. $\mathrm{Cov}(A\underline{X} + \underline{b}) = A\mathrm{Cov}(\underline{X})A^T$. In particular, $\mathrm{Var}(\underline{a}^T \underline{X}) = \mathrm{Cor}(\underline{a}^T \underline{X}) = \underline{a}^T\mathrm{Cov}(X)\underline{a}$.

Proof. 1.

$$\begin{aligned}(E(A\underline{x} + \underline{b}))_j &= E(\sum_{k=1}^{d} a_{jk}X_k + b_j)\\ &= \sum_{k=1}^{d} a_{jk}E(X_k) + b_j\\ &= (AE(X) + b)_j\end{aligned}$$

for all j.

2.

$$\begin{aligned}\mathrm{Cov}(A\underline{X} + \underline{b})_{ij} &= \mathrm{Cov}((A\underline{X} + \underline{b})_i, (A\underline{X} + \underline{b})_j)\\ &\mathrm{Cov}(\sum_{k=1}^{d} a_{ik}X_k + b_i, \sum_{l=1}^{d} a_{jl}X_l + b_j)\\ &= E((\sum_{k=1}^{d} a_{ik}(X_k - E(X_k)))(\sum_{l=1}^{d} a_{jl}(X_l - E(X_l))))\\ &= E(\sum_{k=1}^{d}\sum_{l=1}^{d} a_{ik}a_{jl}(X_k - E(X_k))(X_l - E(X_l)))\\ &= \sum_{k=1}^{d}\sum_{l=1}^{d} a_{ik}a_{jl}\mathrm{Cov}(X_k, X_l)\\ &= A\mathrm{Cov}(\underline{X})A^T\end{aligned}$$

In particular,

$$\begin{aligned}\mathrm{Var}(\sum_{j=1}^{d} X_k) &= \sum_{i,j=1}^{d} \mathrm{Cov}(X_i, X_j) \\ &= \sum_{j=1}^{d} \sigma_j^2 + 2 \sum_{1\le i<j\le d} \mathrm{Cov}(X_i, X_j) \\ &= (\sum_{j=1}^{d} \sigma_j)^2\end{aligned}$$

□

Proposition 8.3

A real, symmetric matrix Σ is a covariance matrix iff Σ is positive semidefinite.

Proof. ($\Rightarrow$) $a^T \Sigma a = \mathrm{Var}(\underline{a^T}\, \underline{X}) \geq 0$ for $a \in \mathbb{R}^d$.

($\Leftarrow$) A real, symmetric, positive semidefinite Σ allows for a Cholesky decomposition: $\Sigma = AA^T$ with Cholesky factor A which is lower triangular, and has nonnegative diagonal entries. Let $Z_1, \ldots, Z_d \overset{ind.}{\sim} N(0,1)$ and define $\underline{X} = A\underline{Z}$. Then,

$$\begin{aligned}\mathrm{Cov}(\underline{X}) &= A\mathrm{Cov}(\underline{Z})A^T \\ &= AIA^T \\ &= \Sigma\end{aligned}$$

Since $Z_1, \ldots, Z_d$ are independent, so $\mathrm{Cov}(\underline{Z} = I$. Thus Σ is a covariance matrix of $\underline{X}$. □

One can show if Σ is positive definite, Σ is invertible.

8.3.2 Normal Distribution

- Notation: $N(\underline{\mu}, \Sigma)$ for $\underline{\mu} \in \mathbb{R}^d, \Sigma \in \mathbb{R}^{d\times d}$, a covariance matrix.
- $X \sim N(\underline{\mu}, \Sigma) \Leftrightarrow X = \underline{\mu} + A\underline{Z}$ where A is the Cholesky factor of Σ and $\underline{Z} = (Z_1, \ldots, Z_d)$ for $Z_j \overset{ind.}{\sim}$ for $j = 1, \ldots, d$. In other words, $\underline{X}$ is a linear transform of independent standard normal random variables. $\underline{X}$ models outcomes which fluctuate around $\underline{\mu}$ with covariance matrix Σ.
- Density: $\underline{Z}$ has density

$$\begin{aligned}f_{\underline{Z}}(\underline{z}) &= \prod_{j=1}^{d} f_{Z_j}(z_j) \\ &= \prod_{j=1}^{d} \frac{1}{\sqrt{2\pi}} e^{\frac{-1}{2} z_j^2} \\ &= \frac{e^{\frac{-1}{2} z^T z}}{(2\pi)^{\frac{d}{2}}}\end{aligned}$$

The density $f_{\underline{X}}(\underline{x}$ of $\underline{X} = T(\underline{Z})$ for $T(\underline{Z}) = Az + \mu$ can be determined by the density transformation theorem:

If T is injective and differentiable (and therefore continuous), $|\det T'(z)| > 0$ for all z, then $\underline{X} = T(z)$. Thus,

$$f_{\underline{x}}(\underline{X}) f_{\underline{Z}}(T^{-1}(\underline{X})) \frac{1}{|\det T'(T^{-1}(\underline{x}))|}$$

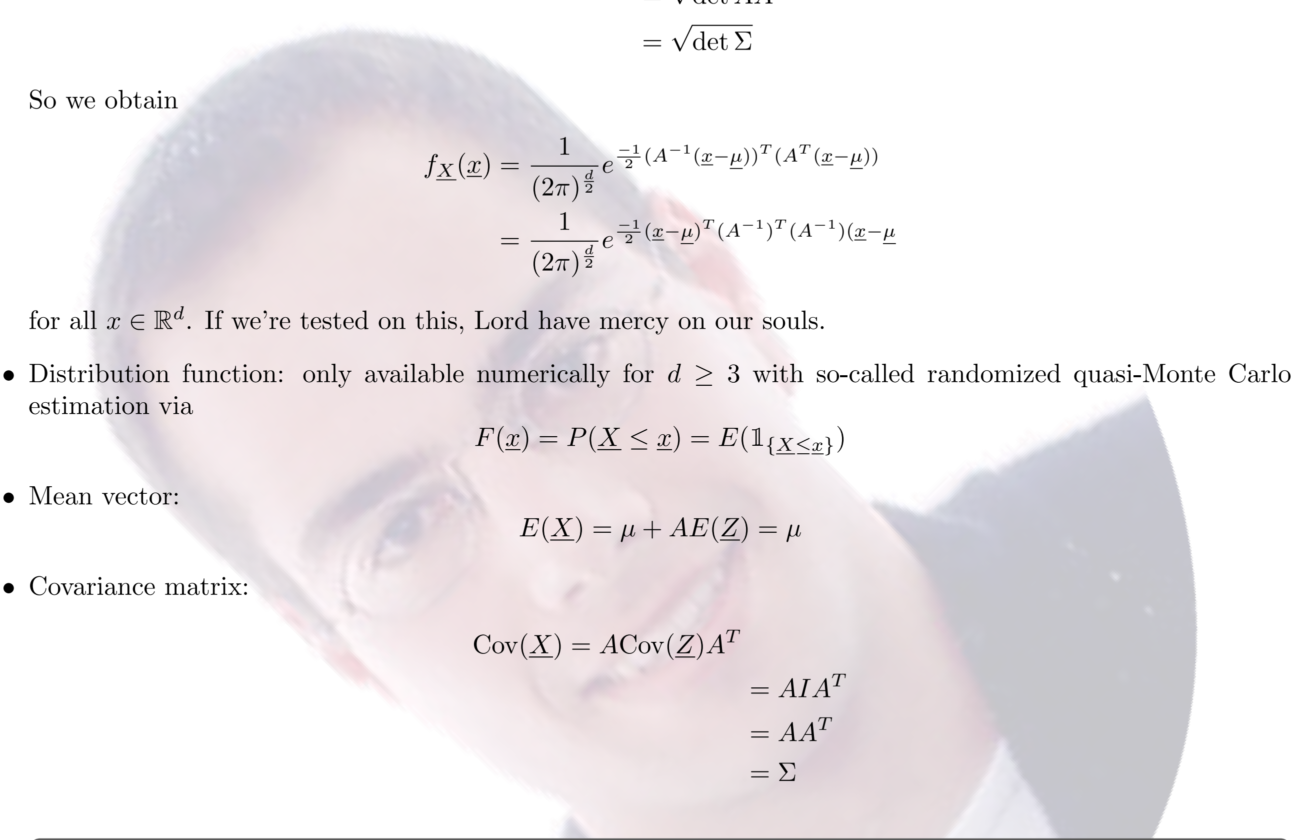
for all $\underline{x} \in \mathbb{R}^d$. With $T^{-1}(\underline{X}) = A^{-1}(\underline{x} - \underline{\mu}), T'(\underline{z}) = A$ and

$$\begin{aligned} |\det T'(T^{-1}(\underline{x}))| &= |\det A| \\ &= \sqrt{(\det A)^2} \\ &= \sqrt{(\det A)(\det A^T)} \\ &= \sqrt{\det AA^T} \\ &= \sqrt{\det \Sigma} \end{aligned}$$

So we obtain

$$\begin{aligned} f_{\underline{X}}(\underline{x}) &= \frac{1}{(2\pi)^{\frac{d}{2}}} e^{\frac{-1}{2}(A^{-1}(\underline{x}-\underline{\mu}))^T (A^T(\underline{x}-\underline{\mu}))} \\ &= \frac{1}{(2\pi)^{\frac{d}{2}}} e^{\frac{-1}{2}(\underline{x}-\underline{\mu})^T (A^{-1})^T (A^{-1})(\underline{x}-\underline{\mu}} \end{aligned}$$

for all $x \in \mathbb{R}^d$. If we're tested on this, Lord have mercy on our souls.



- Distribution function: only available numerically for $d \geq 3$ with so-called randomized quasi-Monte Carlo estimation via
$$F(\underline{x}) = P(\underline{X} \leq \underline{x}) = E(\mathbb{1}_{\{\underline{X} \leq \underline{x}\}})$$
- Mean vector:
$$E(\underline{X}) = \mu + AE(\underline{Z}) = \mu$$
- Covariance matrix:
$$\begin{aligned} \text{Cov}(\underline{X}) &= A\text{Cov}(\underline{Z})A^T \\ &= AIA^T \\ &= AA^T \\ &= \Sigma \end{aligned}$$

Remark

1. One can show $\underline{X} \sim N(\underline{\mu}, \Sigma)$ iff $\underline{a}^T\underline{X} \sim N(\underline{a}^T\underline{\mu}, \underline{a}\Sigma\underline{a})$ for all $a \in \mathbb{R}^d$. In particular, if $X_j \sim N(\mu_j, \sigma_j^2)$ for all $j = 1, \ldots, d$ then we have
$$\sum_{j=1}^d X_j \sim N(\sum_{j=1}^d \mu_j, \sum_{i,j=1}^d \sigma_{ij})$$

where $\sigma_{ij} := \Sigma_{ij}$, by taking $\underline{a} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$ or $\underline{a} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$.

2. $f_{\underline{X}}$ is constant if $(\underline{x} - \underline{\mu})^T \Sigma^{-1} (\underline{x} - \underline{\mu})$ is constant, that is, if the level curves of the density are ellipsoids.

3. $X_1, \ldots, X_d$ are uncorrelated implies

$$\begin{aligned} \Sigma &= \mathrm{Cov}(\underline{X}) \\ &= \begin{bmatrix} \sigma_1^2 & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & \sigma_d^2 \end{bmatrix} \Rightarrow \Sigma^{-1} \qquad &= \begin{bmatrix} \frac{1}{\sigma_1^2} & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & \frac{1}{\sigma_d^2} \end{bmatrix} \end{aligned}$$

So,

$$\begin{aligned} f_{\underline{X}}(\underline{x}) &= \frac{1}{(2\pi)^d \prod_{j=1}^d \sigma_j^2} e^{(\frac{-1}{2} \sum_{j=1}^{d} (\frac{x_j - \mu_j}{\sigma_j})^2} \\ &= \prod_{i=1}^{d} f_{X_j}(x_j) \end{aligned}$$

which implies $X_1, \ldots, X_d$ are independent. (Only here does uncorrelatedness imply independence!!!)

4. It follows from proof of Sklar's theorem (6.8) that if $\underline{Y} \sim N(\underline{0}, \rho)$ for correlation matrix C, then $\underline{U} = (F_{Y_1}(Y_1), \ldots, F_{Y_d}(Y_d)) = (\Phi(Y_1), \ldots, \Phi(Y_d))$ follows copula of $N(\underline{0}, \rho)$ which is the normal or Gaussian copula C. By first part of Sklar's theorem, $C(\underline{u}) = \phi_\rho(\phi^{-1}(u_1), \ldots, \phi^{-1}(u_d))$ for all $\underline{u} \in [0,1]^d$, where ϕ_ρ is the distribution function of $N(\underline{0}, \rho)$. Note, it suffices to consider more general $\underline{X} = \underline{\mu} + A\underline{Z}$ for $AA^T = \Sigma$, since $\underline{X} = (T_1(Y_1), \ldots, T_d(Y_d))$ where $T_j(Y_j) = \mu_j + \sigma_j Y_j$ for all $j = 1, \ldots, d$, which are strictly increasing marginal transformations, so by invariance principle the copula of $\underline{X}$ is the copula of $\underline{Y}$.

9 Limit Theorems

9.1 Modes of convergence

A sequence $\{X_n\}_{n\in\mathbb{N}}$ of random variables can converge in different ways.

Definition 9.1

Let $(\Omega, \mathcal{F}, P)$ be a probability space, $X, X_1, \ldots, X_n : \Omega \to \mathbb{R}$ be random variables. Then $\{X_n\}_{n\in\mathbb{N}}$ converges to X almost surely (notation: $X_n \underset{(n\to\infty)}{\overset{a.s.}{\to}} X$) if

$$P(\lim_{n\to\infty} X_n = X) = 1$$

.

$\{X_n\}_{n\in\mathbb{N}}$ converges to X in probability ($X_n \underset{(n\to\infty)}{\overset{p}{\to}} X$) if

$$\forall \varepsilon > 0, \lim_{n\to\infty} P(|X_n - X| > \varepsilon) = 0$$

.

If $F, F_1, F_2, \ldots$ are distribution functions with $X_n \sim F_n$ for all $n \in \mathbb{N}$, then X_n converges in distribution ($X_n \underset{(n\to\infty)}{\overset{d}{\to}} X$) to $X \sim F$ if $\lim_{n\to\infty} F_n(x) = F(x)$ for all x such that F is continuous at x.

Remark 9.2

1. One can show $X_n \underset{(n\to\infty)}{\overset{a.s.}{\to}} X \Rightarrow X_n \underset{(n\to\infty)}{\overset{p}{\to}} X \Rightarrow X_n \underset{(n\to\infty)}{\overset{d}{\to}} X$. Converses do not hold in general without further conditions.
2. To each of these modes of convergence is associated a limit theorem.

9.2 Weak and Strong Laws of Large Numbers

Lemma 9.3

Let $h : [0, \infty) \to [0, \infty)$ be strictly increasing and X be a random variable such that $E(H(|X|)) < \infty$. Then

$$P(|X| \geq x) \leq \frac{E(h(|x|))}{h(x)}$$

for all $x > 0$.

Proof. Let $x > 0$. Then

$$\begin{aligned} P(|X| \geq x) &= P(|h(X)| \geq h(x)) \\ &= E(\mathbb{1}_{\{h(|X|)\geq h(x)\}}) \\ &\leq E(\frac{h(|X|)}{h(x)} \mathbb{1}_{\{h(|X|)\geq h(x)\}}) \\ &\leq \frac{E(h(|X|))}{h(x)} \end{aligned}$$

□

For $h(x) = x$, $P(|X| \geq x) \leq \frac{E(X)}{x}$ for all $x > 0$ is called Markov's inequality. For $h(x) = x^2$, $P(|X| \geq x) \leq \frac{E(X^2)}{x^2}$ for all $x > 0$ is called Chebyshev's inequality.

Proposition 9.4: Weak Law of Large Numbers

If $\{X_n\}_{n\in\mathbb{N}}$ is a sequence of iid random variables with $\mu = EX$, and $\sigma^2 = \text{Var}(X) < \infty$, then

$$\overline{X_n} := \frac{1}{n}\sum_{i=1}^{n} X_i \underset{(n\to\infty)}{\overset{p}{\to}} \mu$$

Proof. Let $\varepsilon > 0$. Then,

$$\begin{aligned} P(|\overline{X_n} - \mu| > \varepsilon) \leq P(|\overline{X_n} - \mu| \geq \varepsilon) \\ &\leq \frac{E((\overline{X_n} - \mu)^2)}{\varepsilon^2} \\ &= \frac{\text{Var}(\overline{X_n}}{\varepsilon^2} \\ &= \frac{(\frac{1}{n})^2 n\text{Var}(X_1)}{\varepsilon^2} \\ &= \frac{\sigma^2/n}{\varepsilon^2} \underset{(n\to\infty)}{\overset{p}{\to}} 0 \end{aligned}$$

□

By remark 7.7 (2), this following result seems intuitive, but the proof requires more work due to missing assumption of finite second moments.

Theorem 9.5: Strong Law of Large Numbers

If $\{X_n\}_{n\in\mathbb{N}}$ is a sequence of iid random variables with $\mu = E(X)$, then

$$\overline{X_n} \underset{(n\to\infty)}{\overset{a.s.}{\to}} \mu$$

9.3 Central Limit Theorem

What if we need an approximate distribution/df of $\overline{X_n}$ or $\sum_{i=1}^{n} X_i$? Working with characteristic functions often makes working with sums of (independent) random variables easier.

9.3.1 Characteristic Functions

Definition 9.6

The characteristic function (cf) $\phi_{\underline{X}} : \mathbb{R}^d \to \mathbb{C}$ of $\underline{X} \sim F$ is defined by

$$\phi_{\underline{X}}(\underline{t}) = E(e^{i\underline{t}^T\underline{X}}), t \in \mathbb{R}^d$$

$$\text{For } d = 1, \phi_X(t) = E(e^{itx}), t \in \mathbb{R}$$

Remark 9.7

1. By Euler's formula $e^{ix} = \cos(x) + i\sin(x)$,

$$\phi_{\underline{X}}(\underline{t}) = E(\cos(\underline{t}^T\underline{X})) + iE(\sin(\underline{t}^T\underline{X}))$$

   Therefore, $E(|e^{i\underline{t}^T\underline{X}}|) = E(\sqrt{\cos^2(\underline{t}^T\underline{X}) + \sin^2(\underline{t}^T\underline{X})}) = 1$. In particular $\phi_{\underline{X}}$ always exists, $|\phi_{\underline{X}}| \leq 1$, $\phi_{\underline{X}}(0) = 1$. Furthermore $\phi_{\underline{X}}$ is real iff

$$\begin{aligned}\phi_{\underline{X}}(\underline{t}) &= \overline{\phi_{\underline{X}}(\underline{t})}\\ &= E(\cos(\underline{t}^T\underline{X})) - iE(\sin(\underline{t}^T\underline{X}))\\ &= \phi_{\underline{X}}(-\underline{t})\\ &= \phi_{\underline{-X}}(\underline{t})\end{aligned}$$

   for all $\underline{t} \in \mathbb{R}^d$. That is, if $\phi_{\underline{X}}$ is point-symmetric about $\underline{0}$, or by uniqueness, if $\underline{X} \overset{d}{=} \underline{-X}$. (Note: $\overset{d}{=}$ means distributed equally.)

2. One can show $\phi_{\underline{X}}$ is continuous.

3. If A is an $d \times d$ matrix and $\underline{b} \in \mathbb{R}^d$, then for random vector $\underline{X} = (X_1, \ldots, X_d)$ we have

$$\begin{aligned}\phi_{A\underline{X}+\underline{b}}(\underline{t}) &= E(e^{i\underline{t}^T(A\underline{X}+\underline{b})})\\ &= e^{i\underline{t}^T\underline{b}}E(e^{i\underline{t}^TA\underline{X}})\\ &= e^{i\underline{t}^T\underline{b}}\phi_{\underline{X}}(\underline{t}^TA)\end{aligned}$$

4. If $X_1, \ldots, X_d$ are independent, then

$$\begin{aligned}\phi_{X_1+\cdots+X_d}(t) &= E(e^{it^T\sum_{i=1}^d X_i})\\ &= E(\prod_{j=1}^d e^{it^TX_j})\\ &= \prod_{j=1}^d E(e^{it^TX_j})\\ &= \prod_{j=1}^d \phi_{X_j}(t)\end{aligned}$$

Example 9.8

1. Let $t \in \mathbb{R}$. Then for $Z \sim N(0,1)$, we have

$$\begin{aligned}
\phi_Z(t) &= E(\cos(tZ)) + iE(\sin(tZ)) \\
&= \int_{\mathbb{R}} \cos(tz)\varphi(z)dz \\
\frac{d}{dt}\phi_Z(t) &= \int_{\mathbb{R}} (-z)\sin(tz)\varphi(z)dz \\
&= \int_{\mathbb{R}} \sin(tz)\varphi'(z)dz \\
&= [\sin(tz)\varphi(z)]_{-\infty}^{\infty} - t\int_{\mathbb{R}} \cos(tz)\varphi(z)dz \\
&= -t\phi_Z(t)
\end{aligned}$$

   Therefore,

$$\begin{aligned}
(\log\phi_Z(t))' &= \frac{\phi_Z'(t)}{\phi_Z(t)} = -t \\
\Rightarrow \log\phi_Z(t) &= \frac{-t^2}{2} + C \\
\Rightarrow \phi_Z(t) &= e^{\frac{-t^2}{2}+C} \\
\Rightarrow \phi_Z(t) &= e^{\frac{t^2}{2}}
\end{aligned}$$

   So for $X = \mu + \sigma Z \sim N(\mu, \sigma^2)$, we have

$$\phi_{\underline{X}}(t) = e^{it\mu - \frac{1}{2}\sigma^2 t^2}$$

2. If $\underline{X} = \underline{\mu} + A\underline{Z} \sim N(\underline{\mu}, \Sigma)$, then

$$\begin{aligned}
\phi_{\underline{X}}(\underline{t} &= e^{i\underline{t}^T\underline{\mu}}\phi_{\underline{Z}}(\underline{t}^T A) \\
(\text{Let } \tilde{t}^T &= \underline{t}^T A) \\
&= e^{i\underline{t}^T\underline{\mu}} \prod_{j=1}^{d} e^{\frac{-1}{2}\tilde{t}_j^2} \\
&= e^{i\underline{t}^T\underline{\mu} - \frac{1}{2}\sum_{j=1}^{d}\tilde{t}_j^2} \\
&= e^{i\underline{t}^T\underline{\mu} - \frac{1}{2}\underline{t}^T\Sigma\underline{t}}
\end{aligned}$$

Theorem 9.9

1. Uniqueness: $\phi_{\underline{X}}(\underline{t}) = \phi_{\underline{Y}}(\underline{t})$ for all $\underline{t} \in \mathbb{R}^d$ iff $\underline{X} \overset{d}{=} \underline{Y}$.
2. Continuity:
   (i) $X_n \underset{(n\to\infty)}{\overset{d}{\to}} X \Rightarrow \phi_{X_n}(t) \to \phi_X(t)$ for all $t \in \mathbb{R}$.
   (ii) If pointwise for all $t \in \mathbb{R}$ $\phi(t) := \lim_{n\to\infty}\phi_{X_n}(t)$ exists and is continuous at 0 then $X_n \underset{(n\to\infty)}{\overset{d}{\to}} X$ for a random variable X, with cf ϕ.

Example 9.10

1. If $\underline{X} \sim N(\underline{\mu}, \Sigma)$, then

$$\begin{aligned} \phi_{\underline{a}^T \underline{X}}(t) &= E(e^{i(t\underline{a})^T \underline{X}}) \\ &= \phi_{\underline{X}}(t\underline{a}) \\ &= e^{i(t\underline{a})^T \underline{\mu} - \frac{1}{2}(t\underline{a})^T \Sigma (t\underline{a})} \\ &= e^{it(\underline{a}^T \underline{\mu}) - \frac{1}{2}t^2 \underline{a}^T \Sigma \underline{a}} \end{aligned}$$

which is the cf of the $N(\underline{a}^T\underline{\mu}, \underline{a}^T\Sigma\underline{a})$ distribution. Therefore $\underline{a}^T\underline{X} \sim N(\underline{a}^T\underline{\mu}, \underline{a}^T\Sigma\underline{a})$ by uniqueness. In particular if $\underline{a} = (1, \ldots, 1)$, then $\sum_{j=1}^d X_j \sim N(\sum_{j=1}^d \mu_j, \sum_{i,j=1}^d \sigma_{ij})$ and if $X_1, \ldots, X_d$ are uncorrelated (thus independent) then $\sum_{j=1}^d X_j \sim N(\sum_{j=1}^d \mu_j, \sum_{j=1}^d \sigma_j^2)$.

2. If $X_1 \sim N(3, 5)$ and $X_2 \sim N(6, 14)$ are independent, then

$$\begin{aligned} P(X_1 > X_2) &= P(X_1 - X_2 > 0) \\ &= 1 - \Phi(\frac{0 - (-3)}{\sqrt{19}}) \\ &\approx 0.2456 \end{aligned}$$

since $X_1 - X_2 \sim N(\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2 + 0) = N(-3, 19)$.

9.3.2 Main Result

One can show the following lemma:

Lemma 9.11

1. If $a_n \to a$ as $n \to \infty$ then $(1 + \frac{a_n}{n})^n = e^a$ as $n \to \infty$.

2. If $E(|X|^m) < \infty$ for some $m \in \mathbb{N}$, then as $t \to 0$,

$$\phi_X(t) = \sum_{k=0}^{m} \frac{(it)^k}{k!} E(X^k) + o(|t|^m)$$

Note: $h(t) \in o(g(t))$ as $t \to 0$ means $\frac{|h(t)|}{|g(t)|} \to 0$ as $t \to 0$.

Theorem 9.12: Central Limit Theorem (CLT)

If $\{X_n\}_{n\in\mathbb{N}}$ is a sequence of iid random variables with $\mu_1 = E(X_1)$ and $\sigma^2 = \mathrm{Var}(X_1) < \infty$ then

$$\sqrt{n}\frac{\overline{X_n} - \mu}{\sigma} = \frac{\sum_{i=1}^n X_i - n\mu}{\sqrt{n}\sigma} \underset{(n\to\infty)}{\overset{d}{\to}} N(0,1)$$

Proof. If $Y_k := \frac{X_k - \mu}{\sigma}$ for all $k \in \mathbb{N}$ (z-scores), then

$$\begin{aligned}
\phi_{\sqrt{n}\frac{\overline{X_n}-\mu}{\sigma}}(t) &= \phi_{\sqrt{n}\overline{Y_n}}(t) \\
&= \phi_{\frac{1}{\sqrt{n}}\sum_{i=1}^n Y_k}(t) \\
&= E(e^{i\frac{t}{\sqrt{n}}\sum_{i=1}^n Y_k}) \\
&= \prod_{k=1}^n E(e^{\sqrt{n}Y_k}) \\
&= (\phi_{Y_1}(\frac{t}{\sqrt{n}}))^n \\
&= (1 + i\frac{t}{\sqrt{n}}E(Y_1) + (i\frac{t}{\sqrt{n}})^2 \cdot \frac{1}{2!}E(Y_1^2) + o(|\frac{t}{\sqrt{n}}|^2)) \\
&= (1 + \frac{-\frac{t^2}{2} + o(\frac{t^2}{2}) \div \frac{1}{n}}{n})^n \to e^{\frac{-t^2}{2}}
\end{aligned}$$

as $n \to \infty$, for all $t \in \mathbb{R}$. This is the df of $N(0,1)$ so this proves the claim. □

Remark 9.13

1. For iid random variables with finite second moments, CLT implies $\overline{X_n} \sim N(\mu, \frac{\sigma^2}{n})$ for large n. Or, $\sum_{j=1}^n X_j \sim N(n\mu, n\sigma^2)$ for large n. Compare with example 9.10 (1) but without assumption of joint normality. If the distribution of X_1 is very different from $N(\mu, \sigma^2)$, a large n should be chosen.
2. If, additionally, $E(|X_1|^3) < \infty$, the Berry-Essen theorem states the existence of $c \in (\frac{1}{\sqrt{2\pi}}, \frac{1}{2})$ such that $\sup_{x\in\mathbb{R}} |F_{\sqrt{n}\frac{\overline{X_n}-\mu}{\sigma}}(x) - \Phi(x)| \leq c\frac{E(|\frac{X_1-\mu}{\sigma}|^3)}{\sqrt{n}}$ for all $n \in \mathbb{N}$.

Example 9.14

1. Find probability that the proportion of rolled 6's when rolling a fair die $n = 100$ times is within $\varepsilon = 0.02$ of $p = \frac{1}{6}$ based on

   (a) Chebyshev inequality

   (b) CLT

   Let $X_i = \mathbb{1}_{\{\text{i-th roll is 6}\}} \sim B(1, p = \frac{1}{6})$ for $i \in \mathbb{N}$.

   (a)

$$\begin{aligned} P(|\overline{X_n} - p| \leq \varepsilon) &= 1 - P(|\overline{X_n} - p| > \varepsilon) \\ &\geq 1 - \frac{p(1-p)}{n\varepsilon^2} \\ &\approx -2.4722 \end{aligned}$$

   This is clearly not helpful. Our choice of n was too small.

   (b)

$$\begin{aligned} P(|\overline{X_n} - p| \leq \varepsilon) &= P(-\varepsilon \leq \overline{X_n} - p \leq \varepsilon) \\ &= P(-\varepsilon\frac{\sqrt{n}}{\sqrt{p(1-p)}} \leq \sqrt{n}\frac{\overline{X_n} - p}{\sqrt{p(1-p)}} \leq \varepsilon\frac{\sqrt{n}}{\sqrt{p(1-p)}}) \\ \text{(If } n \text{ is large, by CLT)} \quad &\approx \Phi(\varepsilon\frac{\sqrt{n}}{\sqrt{p(1-p)}}) - \Phi(-\varepsilon\frac{\sqrt{n}}{\sqrt{p(1-p)}}) \\ &= 2\Phi(\varepsilon\frac{\sqrt{n}}{\sqrt{p(1-p)}}) - 1 \\ &\approx 0.4085 \end{aligned}$$

   Note that both (a) and (b) imply that for all $\varepsilon > 0$, $P(|\overline{X_n} - p| \leq \varepsilon) \to 1$ as $n \to \infty$ which confirms the Weak Law of Large Numbers.

   (c) Compute the smallest number of times n you have to roll a fair die such that the proportion of "6" being within $\varepsilon = 0.02$ of $p = \frac{1}{2}$ is at least 95% based on the

   (a) Chebyshev inequality

   (b) CLT

   (a)

$$\begin{aligned} P(|\overline{X_n} - p| \leq \varepsilon) &\geq 1 - \frac{p(1-p)}{n\varepsilon^2} \\ n &\geq \lceil \frac{p(1-p)}{0.05\varepsilon^2} \rceil = 6945 \end{aligned}$$

   (b)

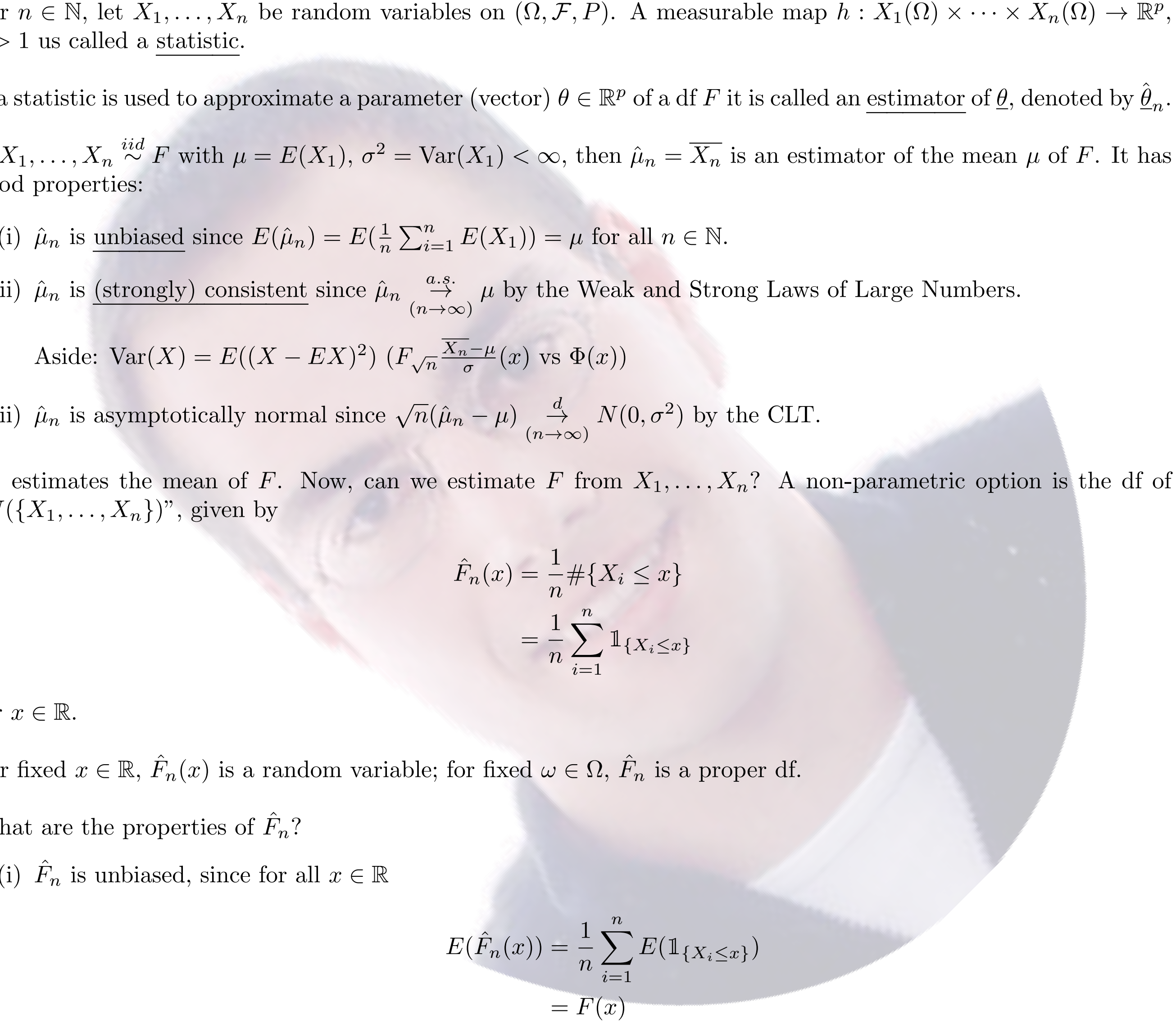
$$\begin{aligned} P(|\overline{X_n} - p| \leq \varepsilon) &= 2\Phi(\varepsilon\frac{\sqrt{n}}{\sqrt{p(1-p)}}) - 1 \\ &\geq 0.95 \Rightarrow n \geq \lceil (\frac{\sqrt{p(1-p)}}{\varepsilon}\Phi^{-1}(\frac{1.95}{2}))^2 \rceil = 1334 \end{aligned}$$

   In other words, if $n \geq 1334$, then $[\overline{X_n} - \varepsilon, \overline{X_n} + \varepsilon]$ is an asymptotic 95%-confidence interval for p $(= EX_n)$.

Remark 9.15

If X_1 is discrete (say with support $S \subseteq \mathbb{Z}$) and n is small, one often applies a continuity correction by computing $P(a - c \leq \sum_{i=1}^{n} X_i \leq b + c)$ instead of $P(a \leq \sum_{i=1}^{n} X_i \leq b)$ e.g. for $c = \frac{1}{2}$. If $a = b$, this is necessary for all n.

10 Empirical df



$(F_{\sqrt{n}\frac{\overline{X_n}-\mu}{\sigma}}(x)$ vs $\Phi(x))$
For $n \in \mathbb{N}$, let $X_1, \dots, X_n$ be random variables on $(\Omega, \mathcal{F}, P)$. A measurable map $h : X_1(\Omega) \times \cdots \times X_n(\Omega) \to \mathbb{R}^p$, $p > 1$ us called a statistic.

If a statistic is used to approximate a parameter (vector) $\theta \in \mathbb{R}^p$ of a df F it is called an estimator of $\underline{\theta}$, denoted by $\underline{\hat{\theta}}_n$.

If $X_1, \dots, X_n \overset{iid}{\sim} F$ with $\mu = E(X_1)$, $\sigma^2 = \mathrm{Var}(X_1) < \infty$, then $\hat{\mu}_n = \overline{X_n}$ is an estimator of the mean μ of F. It has good properties:

(i) $\hat{\mu}_n$ is unbiased since $E(\hat{\mu}_n) = E(\frac{1}{n}\sum_{i=1}^{n} E(X_1)) = \mu$ for all $n \in \mathbb{N}$.

(ii) $\hat{\mu}_n$ is (strongly) consistent since $\hat{\mu}_n \underset{(n\to\infty)}{\overset{a.s.}{\to}} \mu$ by the Weak and Strong Laws of Large Numbers.

Aside: $\mathrm{Var}(X) = E((X - EX)^2)$ $(F_{\sqrt{n}\frac{\overline{X_n}-\mu}{\sigma}}(x)$ vs $\Phi(x))$

(iii) $\hat{\mu}_n$ is asymptotically normal since $\sqrt{n}(\hat{\mu}_n - \mu) \underset{(n\to\infty)}{\overset{d}{\to}} N(0, \sigma^2)$ by the CLT.

$\hat{\mu}_n$ estimates the mean of F. Now, can we estimate F from $X_1, \dots, X_n$? A non-parametric option is the df of "$U(\{X_1, \dots, X_n\})$", given by

$$\begin{aligned}\hat{F}_n(x) &= \frac{1}{n}\#\{X_i \leq x\} \\ &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{X_i \leq x\}}\end{aligned}$$

for $x \in \mathbb{R}$.

For fixed $x \in \mathbb{R}$, $\hat{F}_n(x)$ is a random variable; for fixed $\omega \in \Omega$, $\hat{F}_n$ is a proper df.

What are the properties of $\hat{F}_n$?

(i) $\hat{F}_n$ is unbiased, since for all $x \in \mathbb{R}$

$$\begin{aligned}E(\hat{F}_n(x)) &= \frac{1}{n}\sum_{i=1}^{n} E(\mathbb{1}_{\{X_i \leq x\}}) \\ &= F(x)\end{aligned}$$

(ii) $\hat{F}_n$ is (strongly) consistent since for all $x \in \mathbb{R}$, $\hat{F}_n(x) \underset{(n\to\infty)}{\overset{a.s.}{\to}} E(\mathbb{1}_{\{}X_1 \leq x\}) = F(x)$ by the Weak or Strong Law of Large Numbers.

(iii) $\hat{F}_n$ is asymptotically normal since, for all $x \in \mathbb{R}$, $\sqrt{n}(\hat{F}_n(x) - F(x)) \underset{(n\to\infty)}{\overset{d}{\to}} N(0, F(x)(1 - F(X))$ by CLT. Here, $F(x)(1 - F(X)$ is the variance of $B(1, F(x))$.

Note:

1. $n\hat{F}_n(x) = \sum_{i=1}^n \mathbb{1}_{\{X_i \leq x\}} \sim B(n, F(x))$
2. Everything also applies to iid $\underline{X_1}, \ldots, \underline{X_n}$ from a multivariate df F.