Part 2: Probability Measures and Expectation

Probability Theory: Math 605, Fall 2024

Luc Rey-Bellet
University of Massachusetts Amherst

2024-09-30

1 Dynkin systems

Proving that a property holds for all measurable sets in a σ-algebra may seem a-priori very difficult, often because σ-algebra are defined in a indirect manner, for example the Borel σ-algebra is the smallest σ-algebra generated by open sets. The Dynkin theorem(s) is a technical tool to accomplish this.
If you need to remember only one thing of this section: a probability measure on $\mathbb{R}$ is uniquely determined by its value on the intervals $(a, b]$.

1.1 p-systems and d-systems

Definition 1.1 (p-systems and d-systems)

- A collection of sets $\mathcal{C}$ is a p-system it is closed under (finite) intersections.
- A collection of sets $\mathcal{D}$ is a d-system if
  1. $\Omega \in \mathcal{D}$
  2. $A, B \in \mathcal{D}$ and $A \supset B \implies A \setminus B \in \mathcal{D}$
  3. $A_1, A_2, \cdots \in \mathcal{D}$ with $A_n \nearrow A \implies A \in \mathcal{D}$

The p stands for product (= intersection) and d stands for Eugene Dynkin who introduced that concept.

It is obvious that a σ-algebra is both a p-system and a d-system. The next proposition shows the converse.

Proposition 1.1 $\mathcal{E}$ is a σ-algebra if and only if $\mathcal{E}$ is a p-system and a d-system.

Proof. If $\mathcal{E}$ is a p-system and a d-system then Ω and $\emptyset$ are in $\mathcal{E}$ and $\mathcal{E}$ is closed under complement. All this follows from properties 1. and 2. for d-system. Furthermore $\mathcal{E}$ is then closed under union since $A \cup B = \left(A^c \cap B^c\right)^c$. Finally to extend this to countable unions for pairwise disjoiont A_i define $B_n = \cup_{i=1}^n A_i$ and use the property 3. of d-systems.

1.2 Monotone Class Theorem

The next theorem is a version of many theorems of the same type in probability and measure theory.

Theorem 1.1 (Monotone Class Theorem) If a d-system contains a p-system $\mathcal{C}$ then it contains the σ-algebra generated by $\mathcal{C}$.

Proof. Consider the smallest d-system $\mathcal{D}$ containing $\mathcal{C}$ (intersections of d-systems are d-sytems). It is enough to prove the statement for $\mathcal{D}$, that is, $\mathcal{D} \supset \sigma(\mathcal{C})$. Since $\sigma(\mathcal{C})$ is the smallest σ-algebra containing $\mathcal{C}$ it is enough to show that $\mathcal{D}$ is a σ-algebra itself. By Proposition 1.1 we thus only need to show that $\mathcal{D}$ is a p-system.

Fix $B \in \mathcal{C}$ and consider $\mathcal{D}_1 = \{A \in \mathcal{D} \,:\, A \cap B \in \mathcal{D}\}$.

Note that B belongs to $\mathcal{D}$. We claim that $\mathcal{D}_1$ is a d-system. Clearly $\Omega \in \mathcal{D}_1$. Further if $A_1 \subset A_2$ with both A_1,A_2 in $\mathcal{D}_1$ then $(A_2 \setminus A_1) \cap B = (A_2 \cap B) \setminus (A_1 \cap B)$ which belongs to $\mathcal{D}$. Similarly if $A_n \in \mathcal{D}_1$ and $A_n \nearrow A$ then $(A_n \cap B) \nearrow (A \cap B)$ and so $A \cap B \in \mathcal{D}$ and so $A \in \mathcal{D}_1$.
$\mathcal{D}_1$ is thus a d-system and it contains $\mathcal{C}$ since $B \in \mathcal{C}$ and $\mathcal{C}$ is a p-system. Therefore $\mathcal{D}_1 \supset \mathcal{D}$ and we have shown that if $A \in \mathcal{D}$ and $B \in \mathcal{C}$ then $A \cap B \in \mathcal{D}$.

We now define for fixed $A \in \mathcal{D}$ the set $\mathcal{D}_2 = \{B \in \mathcal{D} \,:\, A \cap B \in \mathcal{D}\}$.

One verifies that $\mathcal{D}_2 =$ is a d-system (just like for $\mathcal{D}_1$) and thus $\mathcal{D}_2 \supset \mathcal{D}$. This proves that $\mathcal{D}$ is a p-system. $\square$

1.3 Uniqueness of Measures

It is usually impossible to compute $P(A)$ for all sets. An important appliction of the the monotone class theorem is that knowing the values of P on p-system generating $\mathcal{A}$ determines P uniquely.

Theorem 1.2 (Uniqueness of probability measures) Suppose P and Q are two probability measures on $(\Omega, \mathcal{A})$. If $P(A) = Q(A)$ for all A in a p-system $\mathcal{C}$ generating $\mathcal{A}$ then $P = Q$.

Proof. We know that $P(A) = Q(A)$ for all $A \in \mathcal{C}$ and . Let us consider $\mathcal{D} = \{B \in \mathcal{A} : P(A) = Q(A)\}$.

Clearly $\mathcal{D} \supset \mathcal{C}$ so to use the Monotone Class Theorem we need to show that $\mathcal{D}$ is a d-system.

- Since $P(\Omega) = Q(\Omega) = 1$ then $\Omega \in \mathcal{D}$ and so property 1. holds.
- For property 2. suppose $A, B \in \mathcal{D}$ with $A \supset B$ then $B \setminus A \in \mathcal{D}$ since

$$P(B \setminus A) = P(B) - P(A) = Q(B) - Q(A) = Q(B \setminus A)$$

- For property 3. if $\{A_n\} \subset \mathcal{D}$ and $A_n \nearrow A$, Then $P(A_n) = Q(A_n)$ for all n and by sequential continuity they must have the same limits and thus $P(A) = Q(A)$ and so $A \in \mathcal{D}$.

Corollary 1.1 If two probability P and Q coincide on the sets of the form $(-\infty, a]$ then they are equal.

2 Measurable maps and random variables

2.1 Motivation

- Given a probability space $(\Omega, \mathcal{A}, P)$ we think of $A \in \mathcal{A}$ as an event and $P(A)$ is the probability to the event A occurs. Think of this an "observation": how likely is it that the A occurs.

- A random variable is a more general kind of observation. Think for example that you are performing some measurement: to an outcome $\omega \in \Omega$ you associate e.g. number $X(\omega) \in \mathbb{R}$. It could also be a vector or even some more general object (e.g. a probability measure!)

- Consider another state space $(F, \mathcal{F})$ (often we will take $(\mathbb{R}, \mathcal{B})$ where $\mathcal{B}$ is the Borel σ-algebra) and a map

$$X : \Omega \to F$$

  We will want to compute

$$P(\{\omega, X(\omega) \in A\}) = P(X \in A) = P(X^{-1}(A))$$

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

- The notation $X^{-1}(A) = \{\omega \, : \, X(\omega) \in A\}$ is for the inverse image and for this to make sense we will need $X^{-1}(A) \in \Omega$.
  All of this motivates the following definitions.

2.2 Measurable functions and random variables

Given a function $f : E \to F$ and $B \subset F$ we write

$$f^{-1}(B) = \{x \in E\,;\, f(x) \in B\}$$

for the inverse image. The following properties are easy to verify

- $f^{-1}(\emptyset) = \emptyset$
- $f^{-1}(A \setminus B) = f^{-1}(A) \setminus f^{-1}(B)$
- $f^{-1}(\bigcup_i A_i) = \bigcup_i f^{-1}(A_i)$
- $f^{-1}(\bigcap_i A_i) = \bigcap_i f^{-1}(A_i)$

Definition 2.1 (Measurable and Borel functions) Given measurable spaces $(E, \mathcal{E})$ and $(F, \mathcal{F})$, a function $f : E \to F$ is *measurable* (with respect to $\mathcal{E}$ and $\mathcal{F}$) if

$$f^{-1}(B) \in \mathcal{E} \text{ for all } B \in \mathcal{F}\,.$$

If $F = \mathbb{R}$ (equipped with the Borel σ-algebra $\mathcal{B}$) a measurable functions is often called a **Borel function**.

Definition 2.2 (Random variable) A *random variable* is a measurable function

$$X : \Omega \to F$$

from a probability space $(\Omega, \mathcal{A}, P)$ to some measurable space $(F, \mathcal{F})$.
Convention: If $F = \mathbb{R}$ then we always take the Borel σ-algebra.

Remarks:

- Using the letter X for a random variable is standard convention from elementary probability.
- The term "random variable" is maybe a bit unfortunate but it is standard. The word "variable" means we have a function and the word "random" means it is defined on some probability space,
- Compare this to the definition of continuity. A function is continuous if, for all open set, $f^{-1}(O)$ is open.
- We just say measurable if there is no ambiguity on the choice of $\mathcal{E}$ and $\mathcal{F}$.

Fortunately it is enough to check the condition for a few sets

Proposition 2.1 $f : E \to F$ is measurable with respect to $\mathcal{E}$ and $\mathcal{F}$ if and only if

$$f^{-1}(B) \in \mathcal{E} \quad \text{for all} \quad B \in \mathcal{C}$$

where $\mathcal{C}$ generates $\mathcal{F}$ (i.e. $\sigma(\mathcal{C}) = \mathcal{F}$).

Proof. Consider the family of sets

$$\mathcal{D} = \left\{B \in \mathcal{F} : f^{-1}(B) \in \mathcal{E}\right\}$$

We now that $\mathcal{D} \supset \mathcal{C}$ and that $\sigma(\mathcal{C}) = \mathcal{F}$.

To conclude it is enough to show that $\mathcal{D}$ is a σ-algebra because if this true $\mathcal{D} \supset \mathcal{C}$ implies $\mathcal{D} \supset \sigma(\mathcal{C}) = \mathcal{F}$.

Showing that $\mathcal{D}$ is a σ-algebra is easy using the rules for inverse images in Section 2.2.

Corollary 2.1 A function from $(E, \mathcal{E})$ to $(\mathbb{R}, \mathcal{B})$ is measurable if and only if

$$f^{-1}((-\infty, a]) = \{x \in E : f(x) \leq a\} \in \mathcal{E}$$

that this, all the level sets of the function f need to be measurable sets

2.3 Operations on measurable functions

Composition of functions

$$f : E \to F, \qquad g : F \to G, \qquad g \circ f : E \to G$$

Like continuity is preserved by composition so is measurability.

Theorem 2.1 (Composition preserves measurability) If $f : E \to F$ is measurable (w.r.t. $\mathcal{E}$ and $\mathcal{F}$) and $g : F \to G$ is measurable (w.r.t. $\mathcal{F}$ and $\mathcal{G}$) then the composition $h = g \circ f$ is measurable (w.r.t $\mathcal{E}$ and $\mathcal{G}$).

Proof. If $C \in \mathcal{G}$ then $(g \circ f)^{-1}(C) = f^{-1}(g^{-1}(C))$. By the measurability of g, $g^{-1}(C) \in \mathcal{F}$ and so by the measurability of f, $f^{-1}(g^{-1}(C)) \in \mathcal{E}$.

Given a function $f : E \to \mathbb{R}$ we define positive/negative parts

$$f_+ = f \vee 0, \quad f_- = -(f \wedge 0) \quad \Longrightarrow \quad f = f_+ - f_-, \quad |f| = f_+ + f_-$$

Theorem 2.2 $f : E \to \mathbb{R}$ is measurable iff and only if f_+ and f_- are measurable.

Proof. It is enough to consider sets of the form $\{x, f(x) \leq a\}$. Proof in your homework.

2.4 Simple functions

Definition 2.3 (Simple functions)

- Given a set $A \in \mathcal{E}$, the **indicator function** 1_A is defined as

$$1_A(x) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{otherwise} \end{cases}$$

- A **simple function** f is a function of the form

$$f(x) = \sum_{i=1}^{n} a_i 1_{A_i}(x)$$

for some finite n, real numbers a_i, and measurable sets A_i.

Remarks

- The A_i are not necessarily disjoint.
- A function is simple if and only if it takes finitely many different values (at most 2^N values including 0)
- The decomposition is not unique.

Definition 2.4 A simple function is in *canonical form* if

$$f(x) = \sum_{i=1}^{m} b_i 1_{B_i}(x)$$

where b_i are all distinct and $(B_i)_{i=1}^m$ form a partition of E.

Remark: One can always rewrite a simple function in canonical form if needed. Just make a list of the values the function takes $b_1, b_2, \cdots, b_m$ and set $B_i = \{x, f(x) = b_i\}$.

Proposition 2.2 If f and g are simple function then so are

$$f + g, \quad f - g, \quad fg, \quad f/g, \quad f \vee g = max\{f, g\}, \quad f \wedge g = \min\{f, g\}$$

Proof. The simplest way to see this is to note that each of these functions takes at most finitely many values if f and g does and therefore they must be simple functions.

2.5 Supremum, infimum, limits

As we see next measurability is preserved by basic operations, in particular taking limits.

Refresher on $\limsup$ and $\liminf$ of sequences: Recall the definitions of $\liminf$ and $\limsup$ for sequences of real numbers (they always exists if we allow the values $\pm\infty$.)

$$\liminf_n a_n = \sup_n \inf_{m\geq n} a_m = \lim_n \inf_{m\geq n} a_m = \text{ smallest accumulation point of } \{a_n\}$$

$$\limsup_n a_n = \inf_n \sup_{m\geq n} a_m = \lim_n \sup_{m\geq n} a_m = \text{ largest accumulation point of } \{a_n\}$$

$$\lim_n a_n \text{ exists } \iff \liminf_n a_n = \limsup_n a_n$$

We have then

Theorem 2.3 Suppose $f_n : E \to \overline{\mathbb{R}}, n = 1, 2, \cdots$ is a sequence of measurable functions (with respect to $\mathcal{E}$ and the Borel σ-algebra). Then the functions

$$\inf_n f_n, \qquad \sup_n f_n\,, \qquad \liminf_n f_n\,, \qquad \limsup_n f_n,$$

are measurable.
If $f = \lim_n f_n$ exsts then f is measurable

Proof.

- Let us write $g = \sup_n f_n$. It is enough to check that $\{g \leq a\}$ is measurable for any a. We have

$$\{g \leq a\} = \{f_n \leq a \text{ for all } n\} = \bigcap_n \{f_n \leq a\}.$$

  So $\inf_n f_n$ is measurable if each f_n is measurable.

- For $g = \inf_n f_n$ we could use that the Borel σ-algebra is generated by the collection $\{[a, +\infty) : a \in \mathbb{R}\}$ and

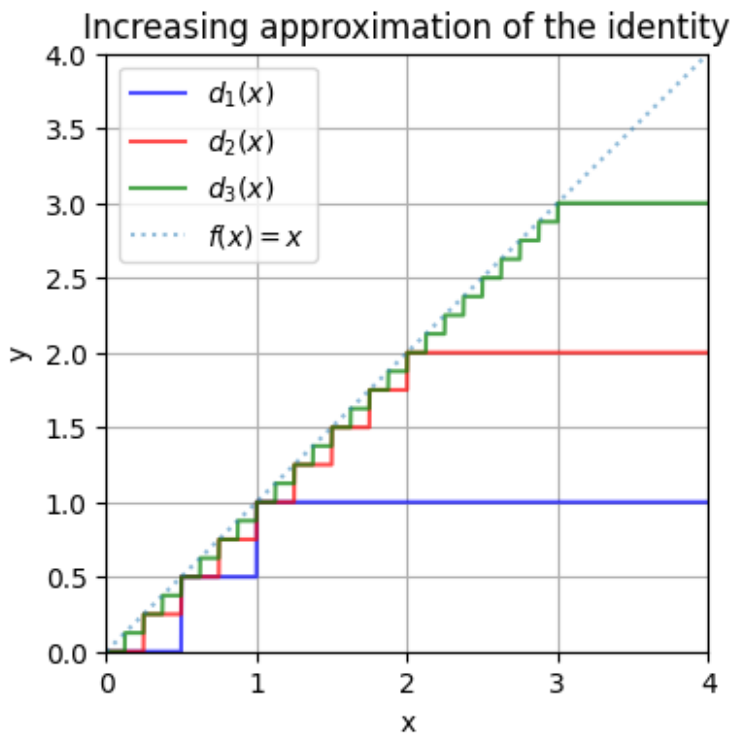
$$\{g \geq a\} = \{f_n \geq a \text{ for all } n\} = \bigcap_n \{f_n \geq a\}.$$

- Since $\limsup$ and $\liminf$ are written in terms of $\inf$ and $\sup$ they do preserve measurability.
- If $f = \lim_n f_n$ exists then $f = \lim_n f_n = \limsup_n f_n = \liminf_n f_n$ and thus is measurable.

2.6 Approximation by simple functions

The following theorem is very important, because it reduces many a computation about measurable function to a computation about a simple function and then taking a limit. In that context one also uses all the time that any measurable f is the difference of two non-negative measurable functions.

Theorem 2.4 (Approximation by simple functions) A nonnegative function $f : E \to \mathbb{R}_+$ is measurable $\iff$ f is the limit of an increasing sequence of positive simple functions.



$$d_n = \sum_{k=1}^{n2^n} \frac{k-1}{2^n} 1_{[\frac{k-1}{2^n}, \frac{k}{2^n})} + n1_{[n,\infty)}$$

Simple function, right continuous, $d_n(x) \nearrow x$ on $[0, \infty)$

Proof. It is not difficult to see that that the function d_n given in the previous page is increasing (due to the dyadic decomposition) and $d_n(x) \nearrow x$ as $n \to \infty$ since if $x \in [\frac{k-1}{2^n}, \frac{k}{2^n})$ then $|x - d_n(x)| \leq \frac{1}{2^n}$.

Let f be a non-negative measurable function then the function

$$g_n = d_n \circ f$$

is a measurable functions (as a composition of measurable functions) and it is a simple function because $d_n \circ f$ takes only finitely many values. Since d_n is increasing and $f(x) \geq 0, d_n(f(x)) \nearrow f(x)$. $\square$

Corollary 2.2 (Approximation by simple functions) A function $f : E \to \mathbb{R}$ is measurable if and only if it can be written as the limit of sequence of simple functions.

Proof. Write $f = f_+ - f_-$ and apply Theorem 2.4 to $f_\pm$. $\square$

Theorem 2.5 Suppose f and g are measurable then

$$f + g, \quad f - g, \quad fg, \quad f/g(\text{ if } g(x) \neq 0)$$

are measurable

Proof. Homework.

2.7 Extended real-valued function

- Write $\overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty, \infty\}$.
- Often it is useful to consider function which are allowed to take values $\pm\infty$.
- The Borel σ-algebra on $\overline{\mathbb{R}}$ consists of all sets of the form A, $A \cup \{-\infty\}$, $A \cup \{\infty\}$, $A \cup \{-\infty, \infty\}$.
- This Borel σ-algebra is generated by the intervals of the form $\{[-\infty, r]\}$.
- All properties of measurable functions on $f : E \to \mathbb{R}$ extend to functions $f : E \to \overline{\mathbb{R}}$: approximation by simple functions, supremeum, infimum, etc...
- We will use all this whenever we need it.

2.8 Homework problems

Exercise 2.1 Show that f is measurable if and only if $f+$ and f_- are measurable.

Exercise 2.2 A function $f : \mathbb{R} \to \mathbb{R}$ is continuous at x if for any $\epsilon > 0$ there exists $\delta > 0$ such that

$$|x - y| < \delta \implies |f(x) - f(y)| < \epsilon \,.$$

A function $f : \mathbb{R} \to \mathbb{R}$ is continuous if it is continuous at all $x \in \mathbb{R}$.

- Show that f is continuous if and only if for every open set O, $f^{-1}(O)$ is open.
- Show that every continuous function is measurable if we equiped $\mathbb{R}$ with the Borel σ-algebra.

Remark: This also holds for any continuous function between arbitrary metric space.

Exercise 2.3

- Suppose $f : \mathbb{R} \to \mathbb{R}$ (both equipped with Borel σ algebra) is a right-continuous step function, if there exists a (finite or countable) collection of intervals $I_n = [t_n, s_n)$ such that f is constant on I_n and $\cup_n I_n = \mathbb{R}$. Show that such a function is measurable.
- A function $f : \mathbb{R} \to \mathbb{R}$ is right continuous if $f(x_n) \to f(x)$ for any decreasing sequence $x_n \searrow x$ and this holds for every x. Show that such a function is measurable.
  Hint: Set $c_n = \sum_{k=1}^{\infty} \frac{k}{2^n} 1_{[\frac{k-1}{2^n}, \frac{k}{2^n})}$ and $f_n = f \circ c_n$.

Exercise 2.4 Suppose $f : \mathbb{R} \to \mathbb{R}$ is increasing. Show that f is measurable.

Exercise 2.5 Given two measurable function f, g from $(E, \mathcal{E})$ to $(\mathbb{R}, \mathcal{B})$. Show that the sets

$$\{f \leq g\}, \quad \{f < g\}, \quad \{f = g\}, \quad \{f \neq g\}$$

are all measurable.

Exercise 2.6 Suppose $(E, \mathcal{E})$ and $(F, \mathcal{F})$ are two measurable spaces. A (measurable) rectangle in $E \times F$ is a set of the form

$$A \times B \quad A \in \mathcal{E}, B \in \mathcal{F}.$$

The product σ-algebra $\mathcal{E} \otimes \mathcal{G}$ is defined as the σ-algebra generated by all measurable rectangles.

- Suppose $f : E \to F$ is measurable (with respect to $\mathcal{E}$ and $\mathcal{F}$) and $g : E \to G$ is measurable (with respect to $\mathcal{E}$ and $\mathcal{G}$). Show that the function $h : E \to F \times G$ given by $h(x) = (f(x), g(x))$ is measurable (with respect to $\mathcal{E}$ and $\mathcal{F} \otimes \mathcal{G}$).
- Suppose $f : E \times F \to G$ is measurable (with respect to $\mathcal{E} \otimes \mathcal{F}$ and $\mathcal{G}$). For any fixed $x_0 \in E$ define the section of f as the function

  $$h : F \to G \quad \text{with } h(y) = f(x_0, y)$$

  Show that h is measurable. *Hint:* Show first that the map $g : Y \to X \times Y$ given by $g(y) = (x_0, y)$ is measurable.

3 Distribution functions and quantile functions

3.1 Random variables

Let us apply what we have learned in the last sections to random variables

$$X : \Omega \to \mathbb{R}$$

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

Theorem 3.1 Suppose $f : E \to F$ is a measurable map between the measurable spaces $(E, \mathcal{E})$ and $(F, \mathcal{F})$ and P a probability measure on $(E, \mathcal{E})$. Then

1. $f^{-1}(F) = \left\{f^{-1}(B), B \in \mathcal{F}\right\}$ is a σ-algebra, in general a sub σ-algebra of $\mathcal{E}$.
2. $P \circ f^{-1}(B)$ which is defined has $P \circ f^{-1}(B) = P(f^{-1}(B)) = P(\{x : f(x) \in B\})$ is a probability measure on $(F, \mathcal{F})$.

Proof. Check the axioms.

Definition 3.1 (Image of a measure) The measure $P \circ f^{-1}$ is called the **image of the measure P under f**. Various other notations are used (such as $f_{\#}P$, etc...)

Adding some terminology

Definition 3.2 (The σ-algebra generated by a random variable X) Given a random variable $X : \Omega \to \mathbb{R}$ defined on the probability space $(\Omega, \mathcal{A}, P)$, the **σ-algebra generated by a random variable X** is the σ-algebra $X^{-1}(\mathcal{B}) \subset \mathcal{A}$.

The interpretation is that this σ-algebra contains all the "information" you can extract from the probability measures P simply by using the random variable X. This will play an increasingly important role in the future!

Definition 3.3 (Distribution of a random variable X) Given a random variable $X : \Omega \to \mathbb{R}$ defined on the probability space $(\Omega, \mathcal{A}, P)$, the **distribution of the random variable X** is the probability measure P^X given by

$$P^X \equiv P \circ X^{-1}$$

defined on $(\mathbb{R}, \mathcal{B})$. That is we have

$$P^X(B) = P(X \in B).$$

3.2 Cumulative distribution function

By Corollary 1.1, probability on $\mathbb{R}$ are uniquely defined by their values on the intervals $(-\infty, x]$, this justify the following definition

Definition 3.4 (Cumulative distribution function) The cumulative distribution function (CDF) of a random variable X is the function $F : (-\infty, \infty) \to [0, 1]$ defined by

$$F_X(t) = P\{X \leq t\} = P^X((-\infty, t])$$

Theorem 3.2 (Properties of CDF) If the function $F(t)$ is the CDF for some random variable X, then F has the following properties

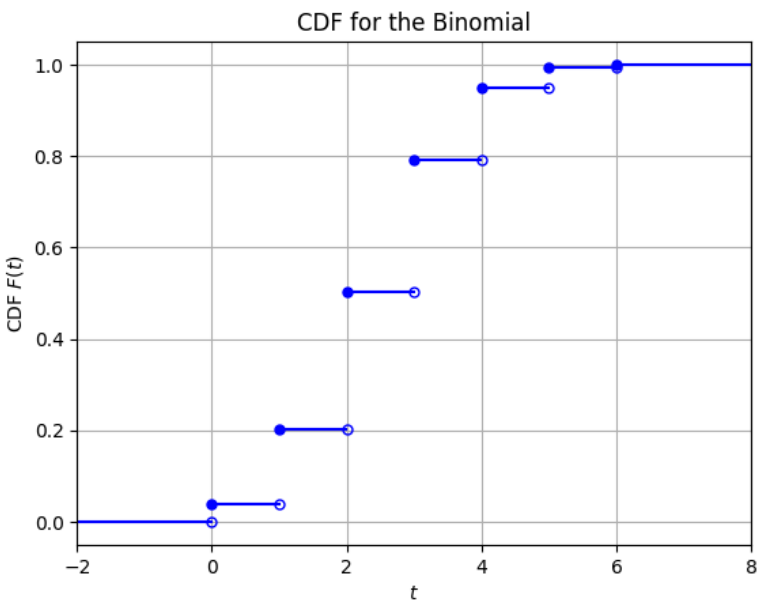
1. F is increasing.
2. $\lim_{t\to-\infty} F(t) = 0$ and $\lim_{t\to+\infty} F(t) = 1$
3. F is right-continuous: for every t, $F(t) = F(t+) \equiv \lim_{s\searrow t} F(s)$.

Proof. Item 1. is the monotonicity property for the probability measure P^X. Item 2. follows from sequential continuity and from the fact that $(-\infty, t] \searrow \emptyset$ as $t \searrow -\infty$ and so $F(t) \searrow P^X(\emptyset) = 0$. A similar argument works for $t \nearrow \infty$. Item 3. follows also from sequential continuity since as $s \searrow t$, $(-\infty, s] \searrow (-\infty, t]$.

Remarks:

Note that F is in general not (left)-continuous. Indeed if $s \nearrow t$ then $(-\infty, s] \nearrow (-\infty, t)$ and $P^X((-\infty, t]) = P^X((-\infty, t)) + P^X(\{t\})$. We denote the left limit by $F(t-)$.

- One can compute probabilities using the CDF. For example
  - $P(a < X \leq b) = F(b) - F(a)$
  - $P(a \leq X \leq b) = F(b) - F(a-)$
  - $P(X = b) = F(b) - F(b-)$
- A atom for a probability measure P on a set Ω is an element $\omega \in \Omega$ such that $P(\{\omega\}) > 0$.
- The distribution P^X of the random variable X has atoms whenever the CDF is discontinuous (i.e. $F_X(t-) \neq F_X(t)$).
- The distribution P^X of the random variable X has at most countably many atoms. (Why? see homework)
- A discrete random variable X taking values $\{x_n\}$ has a purely atomic distribution P^X The CDF $F_X(t)$ is piecewise constantand we have $F_X(t) = \sum_{n\,:\,x_n \leq t} P(\{x_n\})$



3.3 Continuous random variables

Another way to define a CDF is to use a PDF (=probability density function).

Definition 3.5 (Probability density function) A **probability density function** (PDF) is a function $f : \mathbb{R} \to \mathbb{R}$ such that

- $f(t) \geq 0$, f is non-negative
- $\int_{-\infty}^{\infty} f(t)dt = 1$, f is normalized

The corresponding CDF is then given by the integral

$$F(t) = \int_{-\infty}^{t} f(x)dx$$

For now think of the integral as a Riemann integral (e.g. f is piecewise continuous). In particular by the fundamental theorem of Calculus we have

$$F'(t) = f(t)$$

We will revisit this later when equipped with better integration tools. Many of the classical distributions in probability are given by densities. Here are some examples which will come back.

Examples of PDF:

1. **Uniform RV on $[a,b]$:** Wikipedia page on uniform distribution.
   This random variable takes values uniformly distributed in the interval $[a,b]$. It has a density given by

$$f(x) = \begin{cases} \frac{1}{b-a} & a \leq x \leq b \\ 0 & \text{otherwise} \end{cases} \qquad F(x) = \begin{cases} 0 & x \leq a \\ \frac{x-a}{b-a} & a \leq x \leq b \\ 1 & x \geq b \end{cases}$$

2. **Exponential RV with parameter β:** Wikipedia page on exponential.
   The distribution is parametrized by $\lambda > 0$ and the ODF and CDF are given by

$$f(x) = \begin{cases} 0 & x \leq 0 \\ \beta e^{-\beta x} & x \geq 0 \end{cases} \qquad F(x) = \begin{cases} 0 & x \leq 0 \\ 1 - e^{-\beta x} & x \geq 0 \end{cases}$$

3. **Gamma RV with parameters (α, β):** Wikipedia page on gamma distribution.
   The random variables is parametrized by $\alpha > 0$ and $\beta > 0$ and the density is given by

$$f(x) = \begin{cases} 0 & x \leq 0 \\ \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} & x \geq 0 \end{cases}$$

   where $\Gamma(\alpha)$ is the **gamma function** given by $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx$.

4. **Weibull distribution with parameters (α, β):**

5. **Normal distribution with parameters (μ, σ^2):** The normal distribution has parameter $\mu \in \mathbb{R}$

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad F(x) = \int_{-\infty}^{x} f(t)\, dt$$

6. **Log-normal distribution parameters (μ, σ^2):**

7. **Laplace distribution with parameters (α, β):** This is a 2-sided and shifted version of the exponential distribution.

$$f(x) = \frac{\beta}{2} e^{-\beta|x-\alpha|}$$

8. **Cauchy distribution with parameters (α, β):** This is an example of distribution without a finite mean

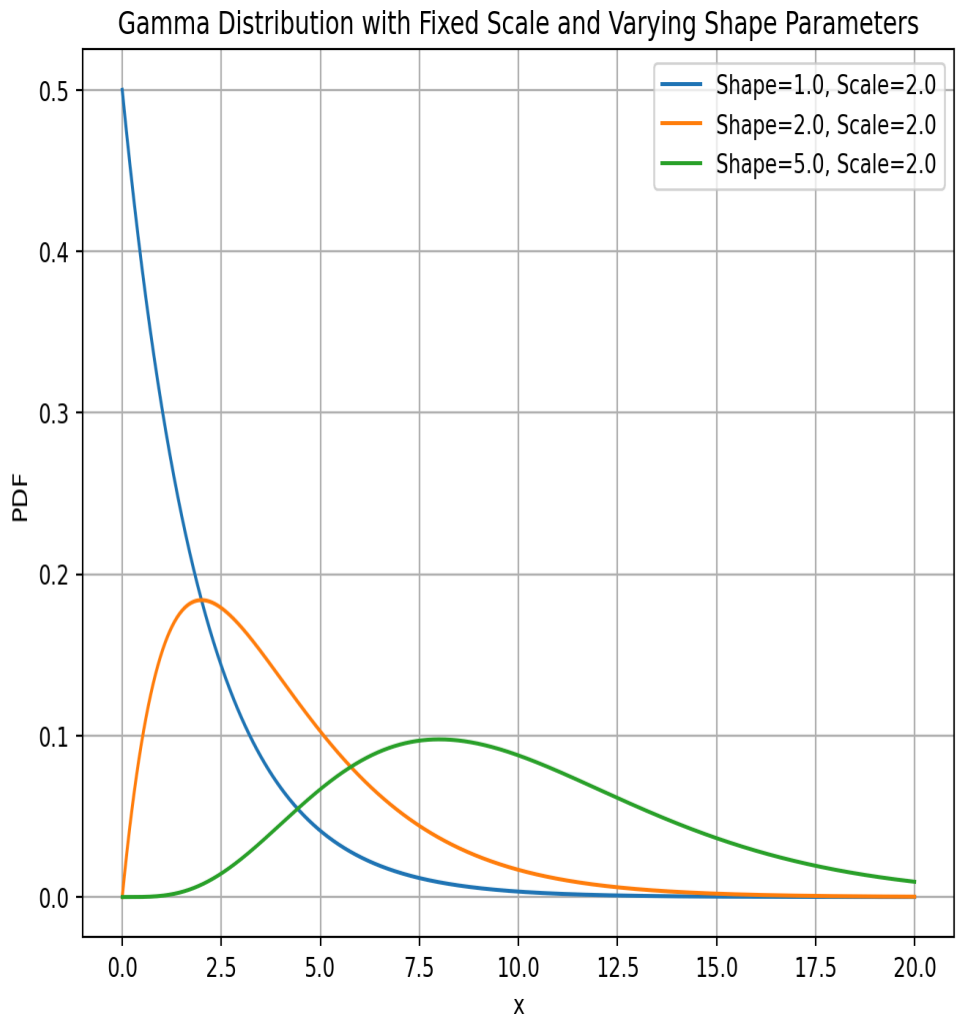
$$f(x) = \frac{1}{\beta\pi} \frac{1}{1 + (x-\alpha)^2/\beta^2} \qquad F(x) = \frac{1}{\pi} \arctan\left(\frac{x-\alpha}{\beta}\right) + \frac{1}{2}$$

9. **Pareto distribution with paramters (x_0, α):**

It is always a good idea to map the density of a random variables (ask ChatGPT for help). Note that the Gamma random variables is often paramterized by $\theta = 1/\beta$.

▼ Code

```python
import numpy as np
import matplotlib.pyplot as plt
from scipy.stats import gamma

# Fixed scale parameter
scale = 2.0  # Scale parameter (theta)

# Define a range of shape parameters
shape_parameters = [1.0, 2.0, 5.0]

# Generate x values (range)
x = np.linspace(0, 20, 1000)

# Plot PDFs for different shape parameters
plt.figure(figsize=(8, 6))
for shape in shape_parameters:
    pdf_values = gamma.pdf(x, a=shape, sca
    plt.plot(x, pdf_values, label=f'Shape=

# Add labels and title
plt.title('Gamma Distribution with Fixed S
plt.xlabel('x')
plt.ylabel('PDF')
plt.legend()
plt.grid(True)
plt.show()
```



3.4 Random variables with mixed distribution.

It easy to build a random variable whose distribution is neither discrete nor continous.

Example: Flip a fair coin. If the coins lands on tail you win a prize X uniformly distributed on $[0,1]$ and if the coins lands on tail you loose. Then X has an atom at 0 and

$$F(x) = \begin{cases} 0 & x < 0 \\ \frac{1}{2} + \frac{1}{2}x & 0 \leq x < 1 \\ 1 & x \geq 1 \end{cases}$$

More generally we can use the concept of mixture

Definition 3.6 (Mixtures of Random variables) Suppose $X_1, X_2, \cdots, X_m$ are random variables with CDF $F_{X_1}(t)$ and $\alpha = (\alpha_1, \cdots, \alpha_m)$ is such that $\alpha_i \geq 0$ and $\sum_{i=1}^m \alpha_i = 1$. Then

$$\sum_{i=1}^m \alpha_i F_{X_i}(t)$$

is a CDF of a random variable X which is called the $(\alpha_1, \cdots, \alpha_m)$ mixture of $X_1, X_2, \cdots, X_m$.

In the previous example we had a $(1/2, 1/2)$ mixture of $X_1 = 0$ (a discrete RV) and X_2 a uniform RV on $[0,1]$ (a continuous RV).

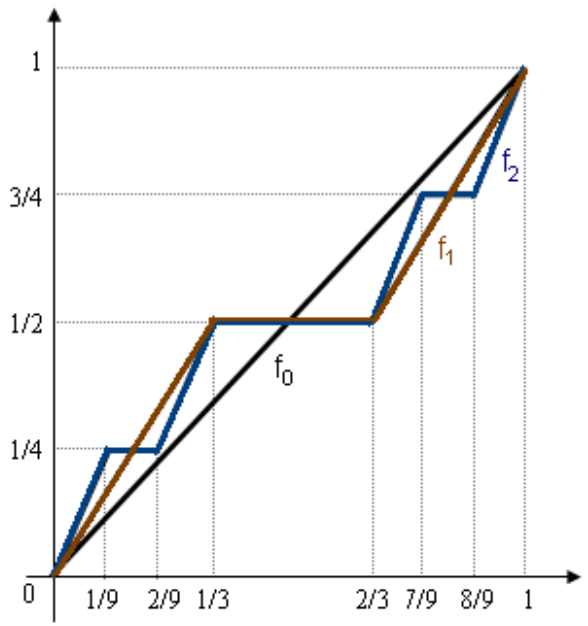
3.5 Devil's staircase

We construct here a CDF with remarkable properties

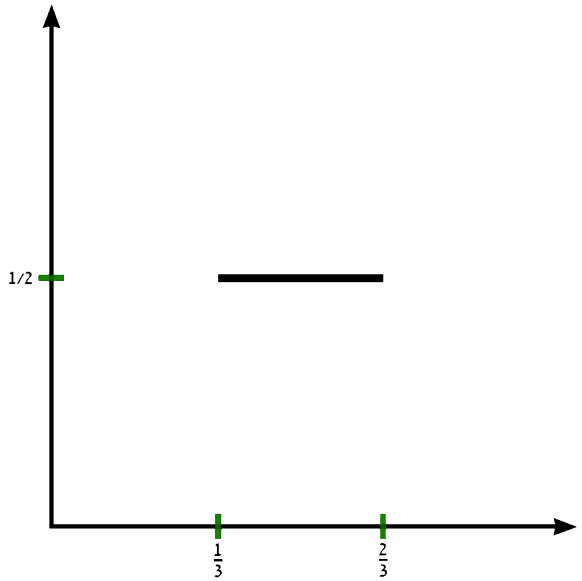
- $F(t)$ has no discontinuities (no atoms)
- $F(t)$ does not have a density, that is $F(t)$ cannot be written as $F(t) = \int_0^x f(t)dt$.

The construction is based on the Cantor set and F is defined iteratively.

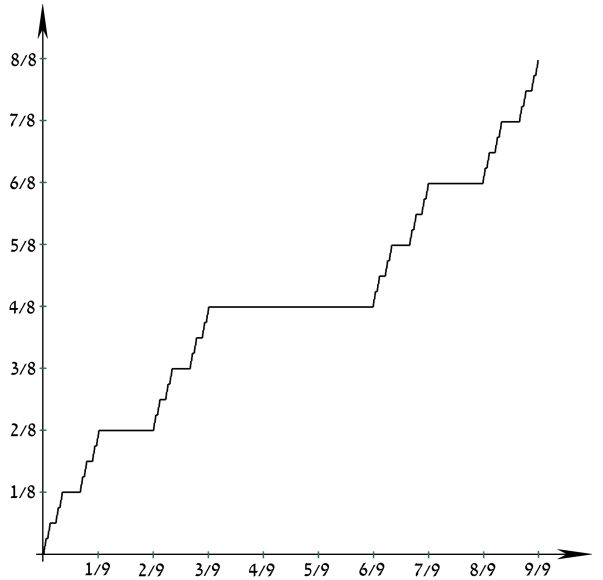
- Set $F_0(t) = t$
- Define the function F_1 to be equal to $\frac{1}{2}$ on $[1/3, 2/3]$ continuous and linear $[0, 1]$ with $F(0) = 0$ and $F(1) = 1$. Then we have $|F_1(t) - F_0(t)| < \frac{1}{2}$.
- In the second step, let F_2 to be equal to $\frac{1}{4}$ on $[1/9, 2/9]$, unchanged on $[1/3, 2/3]$, $\frac{3}{4}$ on $[1/9, 2/9]$, continuous and piecewise linear $[0, 1]$ with $F(0) = 0$ and $F(1) = 1$. We have $|F_2(t) - F_1(t)| < \frac{1}{4}$.
- Repeat the procedure now on the interval $[1/27, 2/27]$, $[7/27, 8/27]$, $[19/27, 20/27]$, $[25/27, 26/27]$....
- It is not diificult to see, by induction, that $|F_n(t) - F_{n-1}(t)| \leq \frac{1}{2^n}$ and thus the sequence F_n converges uniformly to a continuous function $F(t)$ which is increasing on $[0, 1]$



The functions F_0, F_1, F_2, F_3



The iterative construction



The function $F(t)$

The function $F(t)$ is CDF in good standing. We have $P([1/3, 2/3]) = 0$ as well as $P([1/9, 2/9]) =$P([7/9,8/9])=0$ and so on. In particular there are 2^{n-1} intervals of lengths $\frac{1}{3^n}$ whose probability vanishes. The total lenghts of all the interval on which the probability vanishes is thus $\frac{1}{3} + 2 \times \frac{1}{9} + 4\frac{1}{27} = \sum_{n=0}^{\infty} \frac{2^{n-1}}{3^n} = 1$. Thus it cannot have a density!

A random variable X with CDF $F(t)$ is neither continuous (in the sense of having a density), nor discrete and it is called sometimes a singular continous dostribution. The CDF is called the Cantor's function or sometime, more poetically, the devil's staircase.

3.6 Quantile functions

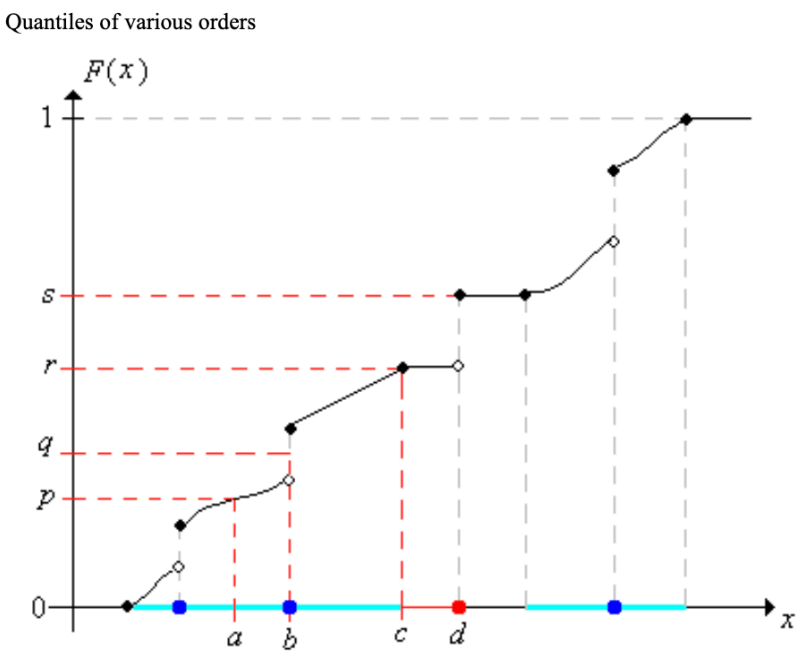
Intuitively a p-quantile for a RV X, where $p \in (0,1)$, is a value $t \in \mathbb{R}$ where the probability that $F_X(t) = P(X \leq t)$ reaches (or crosses over) p. For $p = \frac{1}{2}$ it is usually referred to as the **median**. More formally

Definition 3.7 (Quantiles of a RV X.) For $p \in (0,1)$, a p-quantile for the RV X is a value $t \in \mathbb{R}$ such that

$$P(X < t) = F_X(t-) \leq p \quad \text{and} \quad P(X \leq t) = F_X(t) \geq p$$

Remark: Various cases are possible

- a is the unique p-quantile for p (F_X is strictly increasing at a)
- b is the unique q-quantile (but there is an whole interval of q which share the same quantile b!).
- The interval $[c, d]$ are all r- quantiles (because F_X is locally constant).



We now make a choice to make it unique (other conventions occur in the literature).

Definition 3.8 (Quantile function for a random variable X) For a RV X with CDF $F(t)$ we define the **quantile function of** X, $Q : [0,1] \to \mathbb{R}$ as

$$Q(p) = \min\{t \,:\, F(t) \geq p\}$$

with the convention that $\inf \emptyset = +\infty$

Remark:

- Q is well defined since F being increasing and right-continuous implies that

$$\{t \,:\, F(t) \geq p\} = [a, \infty)$$

  and thus the mimimum exists.

- $Q(p)$ is a p-quantile since if $s = Q(p)$ then $F(s) \geq p$ and, for any $t < s$, $F(t) < p$. Therefore $F(s-) \leq p$. In fact this shows that $Q(p)$ is the **smallest** p-quantile of X.

- If we had picked $\widetilde{Q}(t) = \inf\{t \,:\, F(t) > p\}$ this would have given us the largest p-quantile (a fine, and common, choice as well).

Theorem 3.3 (Properties of the quantile function) The quantile function $Q(p)$ satisfies the following properties

1. $Q(p)$ is increasing.
2. $Q(F(t)) \leq t$.
3. $F(Q(p)) \geq p$.
4. $Q(p-) = Q(p)$ and $Q(p+)$ exists. That is Q is left continuous.

Proof.

1. If $p \leq q$ then F increasing implies that $\{t : F(t) \geq q\} \subset \{t : F(t) \geq p\}$ and this implies that $Q(q) \geq Q(p)$.
2. By definition $Q(F(t))$ is the smallest s such that $F(s) \geq F(t)$. Thus $Q(F(t)) \leq t$.
3. $Q(p)$ is a value s such that $F(s) \geq p$ and thus $F(Q(p)) \geq p$.
4. This holds because F is right-continuous.

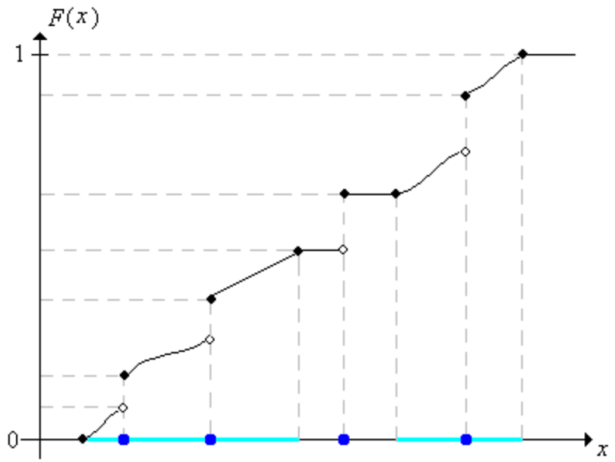


Figure 3.6.7: Graph of the distribution function

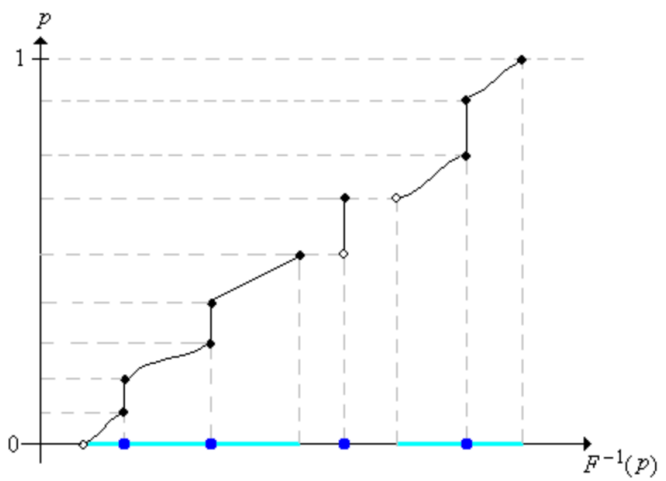


Figure 3.6.8: Graph of the quantile function

Flat portions of $F(t)$ become jump for $Q(p)$ and vice-versa.

3.7 Application: Construction of probability measures on $\mathbb{R}$

The most important property of quantile is the following property which shows that Q is a form of functional inverse for F.

Theorem 3.4 We have

$$Q(p) \leq t \iff p \leq F(t)$$

Proof.

- If $Q(p) \leq t$ then since F is increasing $F(Q(p)) \leq F(t)$. But by Theorem 3.3, item 2. $F(Q(p)) \geq p$ and thus $p \leq F(t)$.
- Conversely if $p \leq F(t)$ then, since Q is increasing, $Q(p) \leq Q(F(p)) \leq p$ where the last inequality is from Theorem 3.3, item 3.

□.

We turn next to constructing all probabilities on $\mathbb{R}$. To do this we first need to construct at least one.

Theorem 3.5 (Lebesgue measure on $[0,1]$) There exists a unique probability measure P_0 on $[0,1]$ with its Borel σ-algebra such that

$$P([a,b]) = b - a$$

The measure P_0 is the distribution of the uniform random variable on $[0,1]$ with PDF

$$f(t) = \begin{cases} 1 & 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

and CDF

$$F(t) = \begin{cases} 0 & x \leq 0 \\ x & 0 \leq x \leq 1 \\ 1 & x \geq 1 \end{cases}$$

Proof. Go and take Math 623....

Equipped with this we can now prove

Theorem 3.6 Any probability measure P on $\mathbb{R}$ has the form

$$P = P_0 \circ Q^{-1}$$

where P_0 is the Lebesque measure on $[0, 1]$ and Q is the quantile function for $F(t) = P((-\infty, t])$.

Proof. By definition of the image measure (see Theorem 3.1), P is a probaility measure, and from the fact that $P_0([0, a]) = a$ we get, using Theorem 3.4

$$\begin{aligned} P_0 \circ Q^{-1}((-\infty, t]) &= P_0(\{p : Q(p) \leq t\}) \\ &= P_0(\{p : p \leq F(t)\}) \\ &= F(t) \end{aligned}$$

and we are done since the CDF determines the measure P uniquely. $\square$

Another way to interpret this result is that we have constructed a probability space for any RV with a given CDF. Namely we constructed a probability space (here $(\Omega, \mathcal{A}, P) = ([0, 1], \mathcal{B}, P_0)$) (here P_0) is the Lebesgue measure on $[0, 1]$ and a map $X = Q$ (the quantile function) with $Q : [0, 1] \to \mathbb{R}$.

3.8 Simulation

In computers are built-in random number generators which generate a uniform RV on $[0, 1]$, that a RV whose distribution is P_0.

Inverse method to generate Random Variables:

To generate a RV X with PDF $F(t)$:

- Generate a random number U. If $U = u$
- If $U = u$ set $X = Q(u)$ where Q is the quantile function for X.

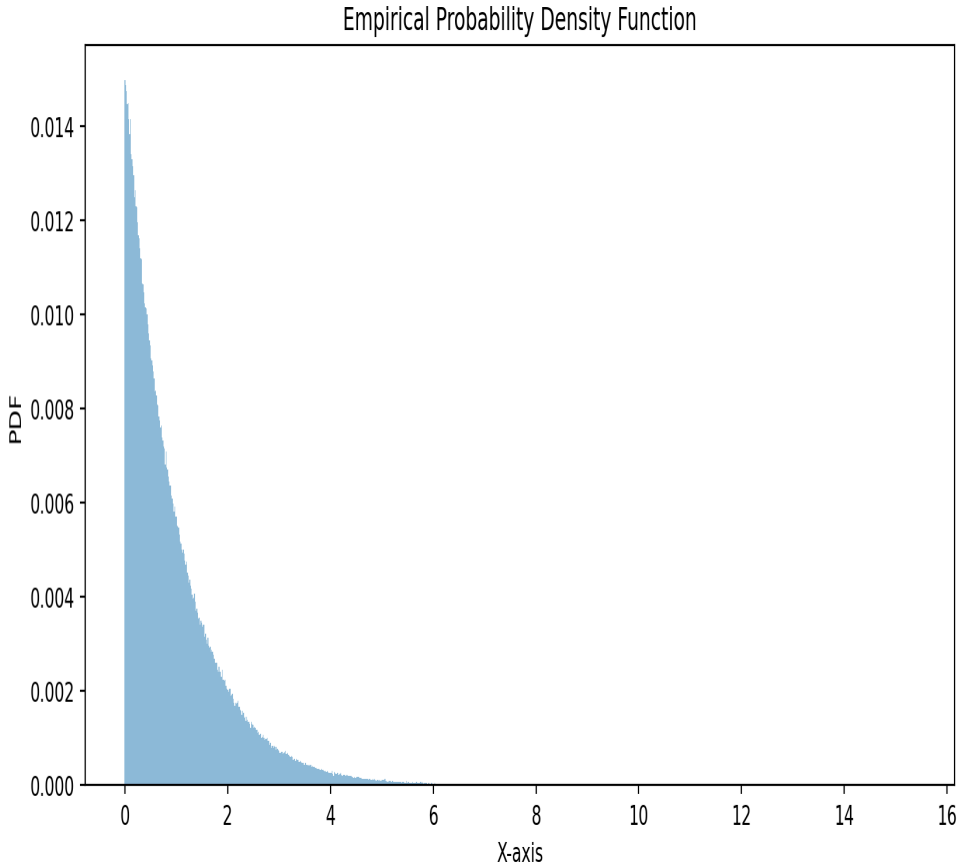
Example:

- If X has an exponential distribution, then $F(t) = \int_0^t \lambda e^{-\lambda s} ds = 1 - e^{-\lambda t}$ and $Q(p) = -\frac{1}{\lambda} \ln(1-p)$
- If X is uniform on $\{1, 2, \cdots, n\}$ then the quantile function is the function $Q(p) = \lceil np \rceil$. Recall $\lceil x \rceil$ is the smallest integer equal or greater than x.
- If X is a normal RV then the CDF is $F(t) = \int_{-\infty}^t \frac{e^{-x^2/2}}{\sqrt{2\pi}} dx$. The quantile $Q = F^{-1}$ has no closed form, but there exists excellent numerical routine to compute it. This can be used to generate normal random variables.

The inverse methods has its limitation and we will learn other simulation methods later on.

▼ Code

```
import numpy as np
import matplotlib.pyplot as plt
from scipy.special import ndtri # quantile

uniform = np.random.rand(1000000)  # gener
dataexponential = - np.log(1-uniform) # qu
datanormal =  ndtri(uniform)   # quantile
datadiscreteuniform10 = np.ceil (10*unifor

# Create a histogram
hist, bin_edges = np.histogram(dataexponen
# Adjust the number of bins as needed

# Calculate the PDF from the histogram
bin_width = bin_edges[1] - bin_edges[0]
pdf = hist * bin_width

# Plot the empirical PDF
plt.bar(bin_edges[:-1], pdf, width=bin_wid
plt.xlabel('X-axis')
plt.ylabel('PDF')
plt.title('Empirical Probability Density F
plt.show()
```



3.9 Homework problems

Exercise 3.1

1. Suppose Y is a real-valued random variable with a **continuous** cdf $F_Y(t)$ and probability distribution P^Y on $(\mathbb{R}, \mathcal{B})$. Show that the random variable $U = F_Y(Y)$ has a uniform distribution P_0 on $[0, 1]$ (i.e. Lebesgue measure).
2. In Theorem 3.6, using the quantile function Q for a given a CDF F we constructed a random variable X : $([0, 1], \mathcal{A}, P_0) \to (\mathbb{R}, \mathcal{B})$ (P_0 is Lebesgue measure) whose CDF is F. In other words we showed $Q(U)$ has CDF F.
   Use this fact and part 1. to construct a random variable $X' : ([0, 1], \mathcal{B}, P_Y) \to (\mathbb{R}, \mathcal{B})$ such that is CDF is F.

Exercise 3.2 Show that the function $X(\omega) = \left\lceil \frac{\ln(1-\omega)}{\ln(1-p)} \right\rceil$ defines a geometric random variable with success probability p on the probability space $(\Omega, \mathcal{A}, P_0)$ (where P_0 is Lebsegue measure. (Or in other words if U is uniform on $[0, 1]$ then $\left\lceil \frac{\ln(1-U)}{\ln(1-p)} \right\rceil$ has a geometric distribution, which provides an easy way to generate geometric random variables on a computer). Provide a code to illustrate this, including the empirical distribution.

Hint: There is a natural relation between the CDF of exponential and geometric random variables.

Exercise 3.3 Some notations:

- A probability measure P on the measurable space $(\Omega, \mathcal{A})$ is called diffuse if P has no atoms.
- Two probability measures P and Q on $(\Omega, \mathcal{A})$ are called singular if we can partition $\Omega = \Omega_P \cup \Omega_Q$ (with $\Omega_P \cap \Omega_Q) = \emptyset$) such that $P(\Omega_P) = 1$ and $Q(\Omega_Q) = 1$.
- The set of all probbaility measures on $(\Omega, \mathcal{A})$ is denoted by $\mathcal{P}(\Omega)$. It is a convex set: if $P, Q \in \mathcal{P}(\Omega)$ then $R = \alpha P + (1-\alpha)Q \in \mathcal{P}(\Omega)$ for any $\alpha \in [0,1]$. We say then that R is a mixture of P and Q.

Show the following

1. Show that any probability measure P can be decomposed as a mixture of two singular atomic measure P_a and diffuse measure P_d.
2. Suppose P is a probability measure on $(\mathbb{R}, \mathcal{B})$ with CDF $F(t)$. Describe the decomposition of the measure P into an atomic and diffuse measure in terms of the CDF F, that is write $F = F_a + F_d$.
3. Suppose P is a diffuse measure on $(\mathbb{R}, \mathcal{B})$ and $A \subset \mathbb{R}$ is any subset with $P(A) > 0$. Show that for any $0 \leq t \leq 1$ there exists a set $B_t \subset A$ such that $P(B_t) = tP(A)$.
   Hint: Let $B_t = A \cap (-\infty, t]$. Study the function $h(t) = P(B_t)$.

Exercise 3.4

- Prove that the Cantor function (a.k.a devil's staircase) given in Section 3.5 is continuous and that this defines a diffuse probability measure P.
- Let C be the Cantor set obtained by removing from $[0, 1]$ the intervals $(1/3, 2/3)$ and $(1/9, 2/9)$ $(7/9, 8/9)$ and so on. If P_0 is the Lebesgue measure on $[0, 1]$, show that $P_0(C) = 0$ and that yet C has the same cardinality as $[0, 1]$. *Hint:* One option is to use the Cantor function.
- Show that the Lebesgue measure on $[0, 1]$, the Cantor measure, and any atomic measure are all singular.

Exercise 3.5 In this problem you should write a code, run it, including a visualization of your result. (The use of ChatGPT or similar tools to help you wirte the code is encouraged.) We suppose the quantile function of the normal random variable with parameter $(\mu, \sigma) = (0, 1)$ is known. For example in python

```
1 from scipy.special import ndtri # quantile for the normal RV
```

Calling random numbers (as many as needed) and using the quantile function ndtri write a code which generates a mixture of 3 normal random variables with parameters

$$(\mu_1, \sigma_1) = (-2, .4), \quad (\mu_2, \sigma_2) = (0, .3), \quad (\mu_3, \sigma_3) = (3, 1)$$

with mixing parameters $(2/7, 4/7, 1/7)$.

->

4 Integration with respect to a probability measure

Given a probability space $(\Omega, \mathcal{A}, P)$ and a random variable $X : \Omega \to \mathbb{R}$ how do we define the expectation of X for general random variables?

There are 2 parts in the theory. A general theory using the measure P from which we deduce a more practical way which uses the probability P^X on $\mathbb{R}$ (the only thing we really know how to handle ...)

4.1 Definition of the expectation (a.k.a the integral)

We start by giving a definition of expectation for an arbitrary random variables. The definition is a bit rigid and may seem at first sight slighlty arbitrary but subsequent analysis will show that this is a good choice.

Definition 4.1 (Definition of expectation) Let $(\Omega, \mathcal{A}, P)$ be a probability space.

1. Suppose X is a simple RV (i.e., it takes finitely many values) then $X = \sum_{j=1}^{M} b_j 1_{B_j}$ (in canonical form!). We define

$$E[X] = \sum_{j=1}^{M} b_j P(B_j) \tag{4.1}$$

2. Suppose X is an arbitrary non-negative RV (i.e. $X(\omega) \geq 0$ for all $\omega \in \Omega$.) Then using the functions d_n given in Theorem 2.4 consider the simple RV $X_n = d_n \circ X$ and define

$$E[X] = \lim_{n\to\infty} E[X_n] \quad \text{where the limit allowed to be } +\infty \tag{4.2}$$

3. For an arbitrary RV X, write $X = X_+ - X_-$ and define

$$E[X] = \begin{cases} E[X_+] - E[X_-] & \text{if } E[X_+] < \infty \text{ or } E[X_-] < \infty \\ \text{undefined} & \text{if } E[X_+] = \infty \text{ and } E[X_-] = \infty \end{cases} \tag{4.3}$$

Remarks Let us make a number of comments on the definition.

1. If the simple RV is *not* in canonical form, i.e. $X = \sum_{i=1}^{N} a_i 1_{A_i}$, then $E[X] = \sum_n a_i P(A_i)$. The argument is tedious but not difficult, take $N = 2$ then consider the sets

$$B_0 = A_1^c \cap A_2^c, B_1 = A_1 \cap A_2^c, B_2 = A_1^c \cap A_2, B_3 = A_1 \cap A_2$$

and the values

$$b_0 = 0, b_1 = a_1, b_2 = a_2, b_3 = a_1 + a_2$$

Then

$$\begin{aligned} E[X] &= b_1 P(B_1) + b_2 P(B_2) + b_2 P(B_3) \\ &= a_1 P(A_1 \cap A_2^c) + a_2 P(A_1^c \cap A_2) + (a_1 + a_2) P(A_1 \cap A_2) = a_1 P(A_1) = a_2 P(A_2) \end{aligned}$$

You can do a similar proof for arbitrary N by an inductive argument.

2. The preceeding remark implies that if X and Y are simple random variables then $E[X + Y] = E[X] + E[Y]$, this is immediate form the the formula which does not use the canonical form and so we have linearity of expectation at least for simple random variables.

3. If Z is a nonnegative random variable then $Z \geq 0$ implies that $E[Z] \geq 0$. Indeed if $Z = \sum_i b_i 1_{B_i}$ is in canonical form then $b_i \geq 0$ and to $E[Z] \geq 0$.

4. If X and Y are simple and nonnegative and $X \leq Y$ then $E[X] \leq E[Y]$. This follows from the linearity by writing $Y = X + (Y - X)$ and so $E[Y] = E[X] + E[Y - X]$. Since $Y - X \geq 0$ then $E[Y - X] \geq 0$ and so $E[X] \leq E[Y]$.

5. The function d_n are increasing in n, $d_n(x) \leq d_{n+1}(x)$ and this implies that $X_n \leq X_{n+1}$ and thus by monotonicity $E[X_n] \leq E[X_{n+1}]$.
   Therefore the limit in Equation 4.2 always exists but could well be equal to $+\infty$.

6. The definition in item 2. seems somewhat arbitrary since it is using a particualr choice of simple function d_n. We will show soon that this choice actually does not matter.

7. For general X we allow the expectation to equal to $+\infty$ (if $E[X_+] = \infty]$ and $E[X_-] < \infty]$) or ($-\infty$ if $E[X_+] < \infty]$ and $E[X_-] = \infty]$). If both $E[X_+] = \infty]$ and $E[X_-] = \infty]$ the expectation is undefined.

8. If $X : \Omega \to \overline{\mathbb{R}}$ is is extended real-valued (the values $\pm\infty$ are aalso allowed) we can still define expectation in the same way. If X is infinite on a set of positive measure then expectation will be infinite or not defined.

Definition 4.2 A measurable function is **integrable** if $E[X]$ is finite or equivalently if $E[|X|] < \infty$ or equivalently if $E[X_\pm] < \infty$.

The set of integrable RV is denote by $\mathcal{L}^1 = \mathcal{L}^1(\Omega, \mathcal{A}, P)$.

4.2 Monotone Convergence

We extend monotonicity to general non-negative RVs.

Theorem 4.1 (Monotonicity) If $X \geq 0$ then $E[X] \geq 0$. If $0 \leq X \leq Y$ then $E[X] \leq E[Y]$.

Proof. If $X \geq 0$ so is $X_n = d_n \circ X$ and therefore $E[X] \geq 0$. If $0 \leq X \leq Y$ then $X_n \leq Y_n$ and so $E[X_n] \leq E[Y_n]$ and thus $E[X] \leq E[Y]$.

The next theorem (Monotone convergence Theorem) is very useful in itself and, in addition, the other convergence theorems for expectations derive from it.

Theorem 4.2 (Monotone Convergence Theorem) Suppose X_n are *non-negative and increasing*: $0 \leq X_n(\omega) \leq X_{n+1}(\omega)$. Then $X(\omega) = \lim_{n\to\infty} X_n(\omega)$ exists and

$$\lim_{n\to\infty} E[X_n] = E[X] = E[\lim_{n\to\infty} X_n]$$

Proof. Since $X_n(\omega)$ is an increasing the sequence, the limit $X(\omega) \in \overline{\mathbb{R}}$ exists and $E[X]$ exists. By monotonicity, see Theorem 4.1, we have $X_n \leq X_{n+1} \leq X$ and therefore $\lim_{n\to\infty} E[X_n]$ exists and we have

$$\lim_{n\to\infty} E[X_n] \leq E[X]\,.$$

We need to show the reverse inequality: $\lim_{n\to\infty} E[X_n] \geq E[X]$. To prove this we need to show the following claim.

Claim: Suppose Y is simple and $Y \leq X$ then $\lim_{n\to\infty} E[X_n] \geq E[Y]$.

Indeed if the claim is true $\lim_{n\to\infty} E[X_n] \geq E[d_k \circ X]$ for all k and taking the limit $k \to \infty$ concludes the proof.

To prove the claim take $b \geq 0$ and consider the set $B = \{X > b\}$ and set $B_n = \{X_n > b\}$. Since $B_n \nearrow B$ we have $P(B_n) \to P(B)$ by sequential continuity. Furthermore

$$X_n 1_B \geq X_n 1_{B_n} \geq b 1_{B_n}$$

which implies, by monotonicity, that $E[X_n 1_B] \geq bP(B_n)$ and taking $n \to \infty$ we obtain

$$\lim_{n\to\infty} E[X_n 1_B] \geq bP(B). \tag{4.4}$$

Now this inequality remains true if we consider the set $\overline{B} = \{X \geq b\}$ instead of B. To see this, take an increasing sequence $b_m \nearrow b$ so that $\{X > b_m\} \nearrow \{X \geq b\}$. Indeed apply Equation 4.4 (with b replaced by b_m) and then used monotonicity.

To conclude note that if $Y = \sum_{i=1}^m a_i 1_{A_i}$ (in canonical form) and $X \geq Y$ then $X \geq a_i$ on A_i. By finite additivity, using Equation 4.4, we have

$$\lim_{n\to\infty} E[X_n] = \sum_{i=1}^m \lim_{n\to\infty} E[X_n 1_{A_i}] \geq \sum_{i=1}^m a_i P(A_i) = E[Y]$$

and this concludes the proof. $\square$

4.3 Further properties of the expectation

Remark: The monotone convergence theorem shows that if X_n is any sequence of simple function increasing to X then $E[X] = \lim_n E[X_n]$.

Theorem 4.3 (Linearity of Expectation) If X and Y are integrable nonnegative random variable then for any $a \geq 0$ and $b \geq 0$ we have

$$E[aX + bY] = aE[X] + bE[Y]$$

Proof. If X and Y are simple this is true by the remarks after Definition 4.1. For general X and Y pick X_n and Y_n simple functions which increase to X and Y respectively (e.g. $X_n = d_n \circ X$ or $Y_n = d_n \circ X$). Then

$$E[aX_n + bY_n] = aE[X_n] + bE[Y_n].$$

Now by the Monotone Convergence Theorem $aX_n + bY_n$ increases to $aX + bY$ and thus taking $n \to \infty$ concludes the proof. $\square$

We will extend the linearity of expectation to general function later after we have developed more theory.

4.4 Negligible sets and completion of a measure space

Let us discuss here a bit carefully sets of probability 0.

Definition 4.3

- A measurable set $A \in \mathcal{A}$ is **negligible with respect to** P (or a **null set for** P) if $P(A) = 0$.
- A set A (not necessarily measurable) is **negligible with respect to** P if there exists $B \in \mathcal{A}$ such that $A \subset B$ and $P(B) = 0$ (i.e. A is a subset of set of meaasure 0).

It is a fine point of measure theory that negligible set need not be measurable. This is true for example for the Borel σ-algebra and Lebesgue measure (see your Math 623 class for more details) and this related to the existence of non- Borel measurable sets.

There is a standard procedure, which is called the **completion of a probability space** to deal with such issue. The idea is to extends the σ-algebra and the probability measure P in such a way all negligible sets are measurable and without changing the probability assigned to sets of positive probability.

The idea is to define, with $\mathcal{N}$ denoting all the null sets of $\mathcal{A}$, a new σ-algebra

$$\overline{\mathcal{A}} = \{A \cup N \,:\, A \in \mathcal{A}, N \in \mathcal{N}\}$$

and a new probability measure

$$\overline{P}(A \cup N) = P(A)\,.$$

It is not terribly difficult to check that $\overline{\mathcal{A}}$ is a σ-algebra and $\overline{P}$ is a probability measure. The probability space $(\Omega, \overline{\mathcal{A}}, \overline{P})$ is called the completion of $(\Omega, \mathcal{A}, P)$.

For example the completion of the Borel σ-algebra on $[0, 1]$ with the Lebesgue measure is called the Lebesgue σ-algebra. This does not play much of a practical role in probability, but at a few occasions it may be convenient to assume that the space is complete.

4.5 Almost sure properties

Generally speaking, almost sure properties are property which are true except possibly on a set of measure 0 (or on a neglgible set).
For example

- We say that two RVs X and Y are equal almost surely if

$$P(X = Y) = P(\{\omega \,:\, X(\omega) = Y(\omega)\}) = 1$$

  that is X and Y differ on a negligible set. We write $X = Y$ a.s.

- If $X = Y$ almost surely then $E[X] = E[Y]$. Indeed then the simple approximations satisfies $X_n = Y_n$ almost surely. If two simple random variables are equal almost surely then their expectations are equal (use their canonical form to see this).

- We say, for example, that $X \geq Y$ a.s if $P(\{\omega \,:\, X(\omega) \geq Y(\omega)\} = 1$.

- We say X_n converges to X almost surely if there exists a set of measure 0, N, such that for all $\omega \in \Omega \setminus N$ we have $\lim_n X_n(\omega) = X(\omega)$.

An example where almost sure property occur naturally is the follwoing result

Theorem 4.4 Suppose $X \geq 0$. Then $E[X] = 0$ if and only $X = 0$ a.s

Proof. If $X = 0$ a.s. then $E[X] = 0$ because $E[0] = 0$. Conversely let $A_n = \left\{\omega : X(\omega) \geq \frac{1}{n}\right\}$. Then $X \geq X1_{A_n} \geq \frac{1}{n}1_{A_n}$ and thus by monotonicity

$$0 = E[X] \geq E[X1_{A_n}] \geq \frac{1}{n}P(A_n)$$

and thus $P(A_n) = 0$ for all n. But $A_n \nearrow \{X > 0\}$ and thus by sequential continuity $P(X = 0) = 1$. $\square$

Some other examples will be used later, see in particular Exercise 5.1.

4.6 Fatou's Lemma

Our first convergence theorem was the monotone convergence theorem Theorem 4.2. Our second convergence theorem still deals with non-negative function random variables and is called the Fatou's lemma.

Theorem 4.5 (Fatou's Lemma) Suppose X_n are non-negative random variables. Then

$$E[\liminf_n X_n] \leq \liminf_n [X_n]$$

Proof. Set $Y_n = \inf_{m \geq n} X_m$. Then $Y_n \leq Y_{n+1}$ and $\liminf_n X_n = \lim_n \inf_{m \geq n} X_m = \lim_n Y_n$. We can use the monotone convergence theorem for the sequence Y_n to get

$$E[\liminf_n X_n] = \lim_n E[Y_n]\,. \tag{4.5}$$

Also for $m \geq n$ we have $Y_n = \inf_{k \geq n} X_k \leq X_m$ and so by monotonicity $E[Y_n] \leq E[X_m]$ and thus

$$E[Y_n] \leq \inf_{m \geq n} E[X_m]\,. \tag{4.6}$$

Combining Equation 4.7 and Equation 4.6 we find

$$E[\liminf_n X_n] \leq \lim_n \inf_{m \geq n} E[X_m] = \liminf_n E[X_n] \tag{4.7}$$

□

Variation on Fatou's Lemma: One can deduce directly from Fatou's Lemma the following results

1. If $X_n \geq Y$ and Y is an integrable RV then $E[\liminf_n X_n] \leq \liminf_n E[X_n]$.
   Proof: Apply Fatou's Lemma to the RV $Y_n = X_n - Y$ which is nonnegative.

2. If $X_n \leq Y$ and Y is an integrable RV $E[\limsup_n X_n] \geq \limsup_n E[X_n]$. *Proof*: Apply Fatou's Lemma to the RV $Y_n = Y - X_n$ which is nonnegative.

3. We shall use these versions of Fatou's Lemma to prove our next big result, the Dominated Convergence Theorem.

4. Intuitively the Fatou's Lemma tells us that "probability can leak away at infinity"" but you can never "create" it. For example cosnider the following example with $\Omega = [0, 1]$ and P the Lebesgue measure.

$$X_n(\omega) = n1_{[0,\frac{1}{n}]}(\omega)$$

   Then we have $X_n \to 0$ a.s. but also

$$E[X_n] = nP([0, \frac{1}{n}]) = 1 \text{ for all } n.$$

   and thus so $E[\lim_n X_n] = 0 \neq 1 = \lim_n E[X_n]$.

4.7 Dominated convergence Theorem

Theorem 4.6 (Dominated convergence theorem) Suppose $\{X_n\}$ is a collection of random variable such that

1. $\lim_n X_n(\omega) = X(\omega)$ for all ω
2. There exists an integrable random variable Y such that $|X_n| \leq Y$ for all n. Then

$$\lim_n E[X_n] = E[X] = E[\lim_n X_n]$$

Proof. We derive it from Fatou's Lemma. The condition $|X_n| \leq Y$ means that $-Y \leq X_n \leq Y$.
Applying Fatou's lemma to $Y - X_n \geq 0$ we find that

$$E[\liminf_n (Y - X_n)] \leq \liminf E[Y - X_n]$$

Using that $\liminf_n(-a_n) = -\limsup_n a_n$ and $\lim_n X_n = X$ we find

$$E[\liminf_n (Y - X_n)] = E[Y] + E[\liminf_n (-X_n)] = E[Y] - E[\limsup_n X_n)] = E[Y] - E[X]$$

and $\liminf E[Y - X_n] = E[Y] - \limsup_n E[X_n]$ and thus we have $\limsup_n E[X_n] \leq E[X]$. Applying Fatou's to $X_n + Y \geq 0$ yields in a similar manner $E[X] \leq \liminf_n E[X_n]$ (check this). Therefore we have $\limsup_n E[X_n] \leq E[X] \leq \liminf_n E[X_n]$. This proves that $\lim_n E[X_n] = E[X]$. $\square$.

A special case of the dominated convergence theorem is frequently used

Theorem 4.7 (Bounded convergence theorem) Suppose $\{X_n\}$ is a collection of random variable such that

1. $\lim_n X_n(\omega) = X(\omega)$ for all ω
2. There exists an integrable random variable c such that $|X_n| \leq c$ for all n. Then

$$\lim_n E[X_n] = E[X] = E[\lim_n X_n]$$

Proof. $Y = c$ is integrable so the result follows from the dominated convergence theorem. $\square$.

Remark on almost sure versions:

Monotone convergence theorem, Fatou's lemma and dominated convergence theorem has also almost sure versions. For example if Y is integrable and $|X_n| \leq Y$ almost surely and $X_n(\omega) \to X$ almost surely then $\lim_n E[X_n] = E[X]$. To see this define

$$N = \{\omega \,:\, |X_n(\omega)| \leq Y(\omega) \text{for all } n\} \quad \text{and} \quad M = \{\omega \,:\, \lim_n X_n(\omega) = X(\omega)\}$$

Then $P(M^c) = P(N^c) = 0$. We can modify the RV on sets of measures of 0 in such a way that the statements hold for all ω: set $X_n = 0, X = 0, Y = 0$ on $M^c \cup N^c$. Then the properties holds for all ω and since the expectations do not change we are done.

4.8 The Expectation rule (very useful for computations)

Computing the expectation of RV X (or $h(X)$) can be done using either P (good for proofs) or P^X (good for computations). As we will see this is an abstract version of the change of variable formula from Calculus!

Notation Another widely used (and convenient) notation for the expectation is $E[X] = \int_\Omega X(\omega) dP(\omega)$.

Theorem 4.8 (Expectation rule) Suppose X is a RV on $(\Omega, \mathcal{A}, P)$ taking value in $(F, \mathcal{F})$ and with distribution P^X. Let $h : (F, \mathcal{F}) \to (\mathbb{R}, \mathcal{B})$ be measurable.

1. $h(X) \in \mathcal{L}^1(\Omega, \mathcal{A}, P)$ if and only if $h \in \mathcal{L}^1(F, \mathcal{F}, P^X)$.

2. If either $h \geq 0$ or h satisifies the equivalent conditions in 1. we have

$$E[h(X)] = \int_\Omega h(X(\omega)) dP(\Omega) = \int_F h(x) dP^X(x) \quad (4.8)$$

3. Conversely suppose Q is a probability measure on $(F, \mathcal{F})$ such that

$$E[h(X)] = \int_F h(x) dQ(x)$$

for all non-negative measurable h. Then $Q = P^X$.

Proof. The probability distribution of X, P^X, is defined by $P^X(B) = P(X^{-1}(B))$. Therefore

$$E[1_B(X)] = P(X \in B) = P^X(B) = \int_F 1_B(x) dP^X(x)$$

This prove Equation 4.8 for chracteristic functions, and by linearity Equation 4.8 hold for simple functions h.

If $h : F \to \mathbb{R}$ is positive then pick a sequence of simple function h_n such that $h_n \nearrow h$. Then

$$\begin{aligned} E[h(X)] = E[\lim_{n\to\infty} h_n(X)] &= \lim_{n\to\infty} E[h_n(X)] \quad \text{by the MCT in } \Omega \\ &= \lim_{n\to\infty} \int_F h_n(x) dP^X(x) \quad \text{because } h_n \text{ is simple.} \\ &= \int_F \lim_{n\to\infty} h_n(x) dP^X(x) \quad \text{by MCT in } F \\ &= \int_F h(x) dP^X(x) \end{aligned}$$

This proves Equation 4.8 for h non-negative. If we apply this to $|h|$ this proves part 1. of the Theorem. For general h, write $h = h_+ - h_-$ and deduce the result by substraction.

For the converse in item 3. just take $f = 1_A$ to be a characteristic function. Then

$$P(X \in A) = E[1_A(X)] = \int_F 1_A(x) dQ(x) = Q(A)\,.$$

Since A is arbitrary, the distribution of X is Q.

Consequences:

- If X is a real-valued random variable, we can compute its expectation as doing an integral on $\mathbb{R}$
- If X is real-valued and $h : \mathbb{R} \to \mathbb{R}$ is a (measurable) function (e.g. X^n, or $e^{i\alpha X}$, or $\cdots$. Then we have

$$E[X] = \int_{\mathbb{R}} x dP^X(x) \quad E[X^n] = \int_{\mathbb{R}} x^n dP^X(x) \quad \cdots$$

  An alternative would to compute the distribution P^Y of the $Y = X^n$ and then we have

$$E[X^n] = E[Y] = \int_{\mathbb{R}} y dP^Y(y)$$

- Generally we will compute $E[h(X)]$ using the distribution of X ....
- But often we will work backward. We will use the change of variable formula to compute the distribution of Y (see item 3. in Theorem 4.8). Checking the equality for all non-negative function or all characteristic function is not always easy so we will show that one can restrict onesleves to just nice functions! (Later..)

4.9 Examples

Example: gamma random variable. The gamma random variable X has density

$$f(x) = \begin{cases} \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

and the Gamma function $\Gamma(\alpha)$ given by

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx \,.$$

Let us compute $E[X^\delta]$ for some $\delta > 0$. Using the expectation rule we find

$$E[X^\delta] = \int_0^\infty x^\delta \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} dx = \frac{\Gamma(\alpha+\delta)}{\Gamma(\alpha)\beta^\delta} \underbrace{\int_0^\infty \frac{\beta^{\alpha+\delta}}{\Gamma(\alpha+\delta)} x^{\alpha+\delta-1} e^{-\beta x} dx}_{=1}$$

If $\delta = n$ is an integer then we can use that $\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$ and so $E[X^n] = \frac{\alpha(\alpha+1)\cdots\alpha+(n-1)}{\beta^\alpha}$

Example: power of an exponential random variable and Weibull random variables

Let us compute next the distribution of $Y = X^\delta$ when X is an exponential random variable (i.e. a Gamma random variable $\alpha = 1$). For any non-negative function h we have, by the expectation ———-

$$E[h(Y)] = E[h(X^\delta)] = \int_0^\infty h(x^\delta)\beta e^{-\beta x} dx$$

and with the change of variable $y = x^\delta, dy = \delta x^{\delta-1} dx$ we find

$$E[h(Y)] = \int_0^\infty h(y)\frac{\beta}{\delta} y^{\frac{1}{\delta}-1} e^{-\beta y^{\frac{1}{\delta}}} dy$$

from which we learn that power of exponential random variables are Weibull random variables.

By a similar computation we see that taking a random vairbales to some poistive power transform the family of Weibull random variables into itself.

4.10 Location and scale

We investigate how the pdf of a random variable transform under a linear transformation $X \mapsto Y = aX + b$.

Theorem 4.9 Suppose the real-value RV X has the pdf $f_X(t)$ then $Y = aX + b$ has the pdf $f_Y(y) = \frac{1}{|a|} f(\frac{y-b}{a})$

Proof. For a change we prove it directly using the expectation rule. The pdf $f_Y(y)$ must satisfy, for any nonnegative h,

$$E[h(Y)] = \int h(y) f_Y(y) dy$$

We rewrite this using the pdf of X using the expectation rule again and the change of variable $y = ax + b$

$$E[h(Y)] = E[h(aX+b)] = \int_{-\infty}^{\infty} h(ax+b) f_X(x) dy = \int_{-\infty}^{\infty} h(y) f_X\left(\frac{y-b}{a}\right) \frac{1}{|a|} dx$$

and therefore we must have $f_Y(y) = \frac{1}{|a|} f_X\left(\frac{y-b}{a}\right) \frac{1}{|a|}$.

Remark Alternatively you can prove this using the CDF, for example for $a > 0$

$$F_Y(t) = P(Y \le t) = P(aX + b \le t) = P\left(X \le \frac{y-b}{a}\right) = F_X\left(\frac{y-b}{a}\right)$$

and then differentiate. Do the case $a < 0$.

Location-scale family of random variables: A family of random variables parametrized by parameter $\alpha \in \mathbb{R}$ (=location) and $\beta \in (0, \infty)$ (=scale) is called a location-scale family if X belonginging to the family implies that $Y = aX + b$ also belong to the family for any parameter a and b. If f has a density this is equivalent to require that the densities have the form

$$f_{\alpha,\beta}(x) = \frac{1}{\beta} f(\frac{x-\alpha}{\beta})$$

for some fixed function $f(x)$.

- Normal RV are scale/location family with parameters μ (=location) and $\sigma > 0$ (=scale) and $f(x) = \frac{e^{-(x-\mu)^2/2\sigma^2}}{\sqrt{2\pi}\sigma}$
- The Cauchy distribution with pdf $\frac{1}{\beta\pi} \frac{1}{1+(\frac{x-\alpha}{\beta})^2}$ is also a location scale family.
- Some family of distribution are only a scale family. For example the exponential random variables with density $f(x) = \frac{1}{\beta} e^{-x/\beta}$ are a scale family with scaling parameter β.

4.11 Homework problems

Exercise 4.1 Show the following facts:

1. Show that if $X = 0$ almost surely if and only if $E[X1_A] = 0$ for all measurable sets A.
2. Suppose X is a random variable with $E[X] < \infty$. Show that $X < \infty$ almost surely.
   Hint: Consider the set $B_n = \{X \geq n\}$.

Exercise 4.2 (infinite sum of random variables) Suppose X_n is a collection of random variables defined on the probability space $(\Omega, \mathcal{A}, P)$.

1. Prove that if the X_n are all nonnegative then $E[\sum_{k=1}^{\infty} X_k] = \sum_{k=1}^{\infty} E[X_k]$.
   Hint: Use the monotone convergence theorem.
2. Prove that if $\sum_{n=1}^{\infty} E[|X_k|]$ is finite then $E[\sum_{k=1}^{\infty} X_n] = \sum_{k=1}^{\infty} E[X_k]$.
   Hint: Consider the RV $Y = \sum |X_k|$ and use the dominated convergence theorem and Exercise 5.1, part 2.

Exercise 4.3 (Building new probability measures using densities) Suppose Y is a random variable on the probability space $(\Omega, \mathcal{A}, P)$ with $Y \geq 0$ almost surely and $E[Y] = 1$.

1. Define $Q : \mathcal{A} \to \mathbb{R}$ by $Q(A) = E[Y 1_A]$. Show that Q is probability measure on $(\Omega, \mathcal{A}, P)$. We denote by E_Q the expectation with respect to Q.
2. Show, using the definition of the integral, that $E_Q[X] = E[XY]$.
3. Show if $B \in \mathcal{A}$ is such that $P(B) = 0$ then we have $Q(B) = 0$. (We say then that Q is absolutely continuous with respect to P.)
4. Show that, in general $Q(B) = 0$ does not imply $P(B) = 0$ but that if $Y > 0$ almost surely then $Q(B) = 0$ does imply $P(B) = 0$.
5. Assuming $Y > 0$ almost surely show that $\frac{1}{Y}$ is integrable with respect to Q and show that the measure R defined by $R(A) = E_Q[\frac{1}{Y} 1_A]$ is equal to P.

Exercise 4.4 (the log normal distribution)

1. Suppose X is a normal random variable. Show that the random variable $Y = e^X$ has the distribution with the following density

$$f(x) = \begin{cases} \frac{1}{x}\frac{1}{\sqrt{2\pi}\sigma} e^{-(\log(x)-\mu)^2/2\sigma^2} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

   The random variable Y is called the log-normal distribution with parameter μ and σ^2.

2. Show that $E[Y^r] = e^{r\mu + \frac{1}{2}\sigma^2 r^2}$. *Hint:* Do the change of variables $y = \log(x) - \mu$ in the integral for $E[Y^r]$.

Exercise 4.5 (Cauchy distribution)

1. Suppose X is a random variable with density $f_X(x)$. Express the density f_Y of $Y = \frac{a}{X}$ in terms of f_X.

2. A Cauchy RV with parameters (α, β) has the pdf $f(x) = \frac{1}{\beta\pi}\frac{1}{1+(x-\alpha)^2/\beta^2}$.

   - Show that if X is a Cauchy RV so is $Y = aX + b$ and find how the parameters transform.
   - Show that if X has a Cauchy distribution with $\alpha = 0$ then $\frac{1}{X}$ has again a Cauchy distribution.
   - Show that the mean and the variance of a Cauchy RV are undefined.

Exercise 4.6 Consider the RV X with CDF given by

$$F(t) = \begin{cases} 0 & t \leq -1 \\ 1/4 + \frac{1}{3}(t+1)^2 & -1 \leq t < 0 \\ 1 - \frac{1}{4}e^{-2t} & t \geq 0 \end{cases}$$

Compute $E[X]$ and $\mathrm{Var}(X)$.

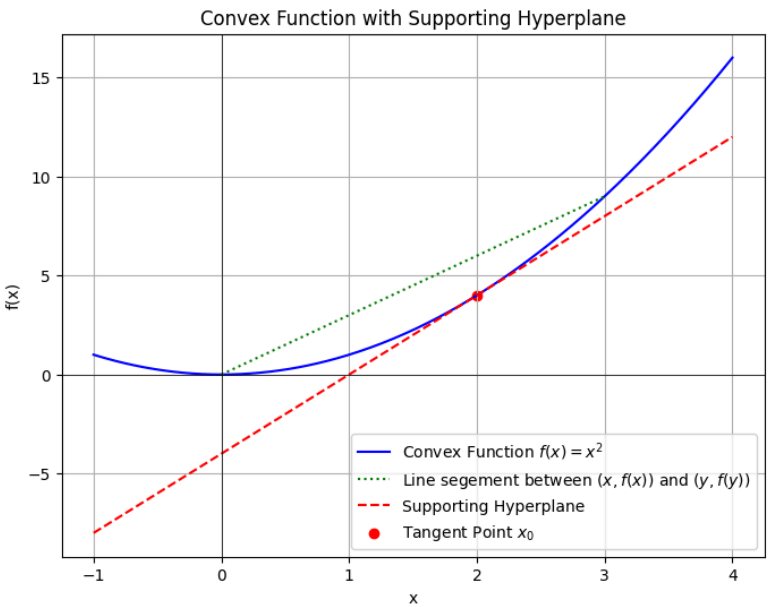
5 Inequalities

5.1 Jensen inequality

Some facts about convex functions: Recall that a function h on $\mathbb{R}^d$ is convex if

$$\phi(\alpha x + (1-\alpha)y) \leq \alpha\phi(x) + (1-\alpha)\phi(y)$$

for all x, y and all $0 \leq \alpha \leq 1$. This means that the line segment between $(x, \phi(x))$ and $(y, \phi(y))$ lies above the graph of $\phi(z)$ for z lying on the line segement between x and y.



- An equivalent description of a convex fuction ("obvious" from a picture, with a proof in the homework) is that at any point x_0 we can find a plane $l(x)$ in $\mathbb{R}^n \times \mathbb{R}$ such that $\phi(x) \geq l(x) = f(x_0) + c \cdot (x - x_0)$ (the graph of f lies above l for all x) and $\phi(x_0) = l(x_0)$. If f is *differentiable* at x_0 the plane is given by the tangent plane to the graph at x_0, we have

$$f(x) \geq f(x_0) + \nabla f(x_0) \cdot (x - x_0)$$

- If f is *twice continuously differentiable* then f is convex if and only if the matrix of second derivative $D^2 f(x_0)$ is positive definite.

Theorem 5.1 (Jensen inequality) If $\phi : \mathbb{R} \to \mathbb{R}$ is a convex function then

$$E[\phi(X)] \geq \phi(E[X])$$

provided both expectations exist, i.e. $E[|X|] < \infty$ and $E[|\phi(X)|] < \infty$.

Proof. Choose $x_0 = E[X]$ and pick a supporting hyperplane $l(x)$ at x_0 so that for any x

$$\phi(x) \geq \phi(E[X]) + l(E[X])(x - E[X])$$

By the motonicity of expectation we obtain

$$E[\phi(X)] \geq \phi(E[X]) + l(E[X])E[(X - E[X])] = \phi(E[X])\,.$$

Examples

- Since $f(x) = x^2$ is convex we have $E[X]^2 \leq E[X^2]$.
- Since $f(x) = e^{\alpha x}$ is convex for any $\alpha \in \mathbb{R}$ we have $E[e^{\alpha X}] \geq e^{\alpha E[X]}$.

Remark The theory of convex functions is very rich and immensely useful!

We will need the following slight generalization of Jensen inequality

Theorem 5.2 If $\phi : \mathbb{R}^d \to \mathbb{R}$ is a convex function and $X = (X_1, \cdots, X_d)$ is a RV taking values in $\mathbb{R}^d$. Then we have

$$E[\phi(X)] \geq \phi((E[X_1], \cdots, E[X_d]))$$

provided both expectations exist.

Proof. Same proof as Jensen.

5.2 L^p-norms

Suppose $(\Omega, \mathcal{A}, P)$ is a probability space and X is a real-valued random variable.

Definition 5.1 (L^p-norms) Given a random variable X and $1 \leq p \leq \infty$ we define

$$\|X\|_p = E[|X|^p]^{\frac{1}{p}} \quad \text{for } 1 \leq p < \infty$$

and

$$\|X\|_\infty = \inf\{b \in \mathbb{R}_+ \, : \, |X| \leq b \text{ a.s}\}$$

and $\|X\|_p$ is called the L^p norm of a RV X.

Remarks

- It is easy to check that

$$\|X\|_p = 0 \implies X = 0 \text{almost surely},$$

$$\|cX\|_p = c\|X\|_p.$$

- $\|X\|_p < \infty$ means that $|X|^p$ is integrable (if $1 \leq p < \infty$) and that X is almost surely bounded (if $p = \infty$). Often $\|X\|_\infty$ is called the essential supremum of X.

5.3 Cauchy-Schwartz, Hölder, Minkowski

Theorem 5.3 (Hölder and Minkowski inequalities)

1. **Hölder:** Suppose $1 \leq p, q \leq \infty$ are such that $\frac{1}{p} + \frac{1}{q} = 1$ then we have

$$\|XY\|_1 \leq \|X\|_p \|Y\|_q \,.$$

   Special case is the **Cauchy-Schwartz** inequality $p = q = 2$

$$\|XY\|_1 \leq \|X\|_2 \|Y\|_2 \,.$$

2. **Minkowski:** For $1 \leq p \leq \infty$ we have

$$\|X + Y\|_p \leq \|X\|_p + \|Y\|_p \,.$$

   (a.k.a triangle inequality)

Proof. The proof is ultimately a consequence of Jensen inequality (there are many different proofs but all relies in one way or the other on convexity). To start one shows that the functions

$$f(u, v) = u^b v^{1-b} \quad \text{and } g(u, v) = (u^b + v^b)^{\frac{1}{b}} \tag{5.1}$$

are concave on $\mathbb{R}^2_+$ for $b \in (0, 1]$) (that is $-f$ and $-g$ are convex). To see this compute for example $D^2 f$ and $D^2 g$.

Once this is done let us turn to Hölder inequality:

- If $p = 1$ and $q = \infty$ then we have $|XY| \leq |X| \|Y\|_\infty$ almost surely and thus $\|XY\|_1 \leq \|X\|_1 \|Y\|_\infty$.
- The convexity of f in Equation 5.1 implies that for non negative random variables U and V we have

$$E[U^b V^{1-b}] \leq E[U]^b E[V]^{1-b}.$$

If $1 < p < q < \infty$ then we set $b = \frac{1}{p}$ and $1 - b = \frac{1}{q}$ and $U = |X|^p$ and $V = |Y|^q$. Then

$$E[|X||Y|] = E\left[(|X|^p)^{\frac{1}{p}} (|Y|^q)^{\frac{1}{q}}\right] \leq E[|X|^p]^{\frac{1}{p}} E[|Y|^q]^{\frac{1}{q}}$$

For Minkowski

- If $p = \infty$ Minkowski inequality is easy to check.
- The convexity of f in Equation 5.1 implies that for non negative random variables U and V we have

$$E[U^b + V^b] \leq \left(E[U]^b + E[V]^b\right)^{1/b}.$$

which implies Minkovski if we take $b = \frac{1}{p}$ and $U = |X|^p$ $V = |Y|^q$.

□.

Definition 5.2 (L^p spaces) For $1 \leq p \leq \infty$ we define

$$\mathcal{L}^p(\Omega, \mathcal{A}, P) = \{X : \Omega \to \mathbb{R}, \|X\|_p < \infty\}$$

and the quotient space

$$L^p(\Omega, \mathcal{A}, P) = \mathcal{L}^p(\Omega, \mathcal{A}, P)/\sim$$

where $X \sim Y$ means $X = Y$ a.s is an equivalence relation.
The space $L^p(\Omega, \mathcal{A}, P)$ is a normed vector space.

Theorem 5.4 The map $p \mapsto \|X\|_p$ is an increasing map from $[1, \infty)$ to $[0, \infty)$.

- If $\|X\|_\infty < \infty$ then $p \mapsto \|X\|_p$ is continuous on $[1, \infty]$.
- If $\|X\|_\infty = \infty$ there exists q such that $\|X\|_p$ is continuous on $[1, q]$ (that is left-continuous at q) and $\|X\|_p = +\infty$ on (q, ∞).

Proof. Homework

Examples:

- If X has a Pareto distribution with parameter α and x_0 then its the CDF is

$$F(t) = 1 - \left(\frac{t}{x_0}\right)^{\alpha} \qquad \text{for } t \geq x_0$$

and $F(t) = 0$ for $t \leq x_0$.
The pdf is $f(x) = \frac{\alpha x_0^{\alpha}}{x^{\alpha+1}}$ and we have

$$E[|X|^p] = E[X^p] = \int_{x_0}^{\infty} \alpha x_0^{\alpha} x^{-\alpha-1+p} dx = \begin{cases} \frac{\alpha}{\alpha-\beta} x_0^{\beta} & p < \alpha \\ +\infty & \beta \geq p \end{cases}$$

- If X has a normal distribution (or an exponential, gamma, etc...} then $X \in L^p$ for all $1 \leq p < \infty$ but $X \notin L^{\infty}$.
- Other norms exists (Orlicz norms) to capture the tail of random variables.

5.4 Markov, Chebyshev, and Chernov

Another very important inequality is the so-called Markov equality. Very simple and very useful.

Theorem 5.5 (Markov inequality) If $X \geq 0$ then for any $a > 0$

$$P(X \geq a) \leq \frac{E[X]}{a}$$

Proof. Using that X is non-negative we have

$$X \geq X1_{X \geq a} \geq a1_{X \geq a}$$

and taking expectation and monotonicity gives the result.

Theorem 5.6 (Chebyshev inequality) We have

$$P(|X - E[X]| \geq \varepsilon) \leq \frac{\mathrm{Var}[X]}{\varepsilon^2}$$

where $\mathrm{Var}(X) = E[X - E[X]]$ is the variance of X.

Proof. Apply Markov inequality to the random variable $(X - E[X])^2$ whose expectation is $\mathrm{Var}[X]$:

$$P(|X - E[X]| \geq \varepsilon) = P((X - E[X])^2 \geq \varepsilon^2) \leq \frac{E[(X - E[X])^2]}{\varepsilon^2} = \frac{\mathrm{Var}[X]}{\varepsilon^2}$$

- Chebyshev inequality suggests measuring deviation from the mean in multiple of the standard deviation $\sigma = \sqrt{\mathrm{Var}[X]}$:

$$P(|X - E[X]| \geq k\sigma) \leq \frac{1}{k^2}$$

- Chebyshev inequality might be extremly pessimistic
- Chebyshev is sharp. Consider the RV X with distributionc$P(X = \pm 1) = \frac{1}{2k^2} \quad P(X = 0) = 1 - \frac{1}{k^2}$ Then $E[X] = 0$ and $\mathrm{Var}[X] = \frac{1}{k^2}$

$$P(|X| \geq k\sigma) = P(|X| \geq 1) = \frac{1}{k^2}$$

Theorem 5.7 (Chernov inequality) We have for any a

$$P(X \geq a) \leq \inf_{t \geq 0} \frac{E[e^{tX}]}{e^{ta}} \qquad P(X \leq a) \leq \inf_{t \leq 0} \frac{E[e^{tX}]}{e^{ta}}$$

Proof. This is again an application of Markov inequality. If $t \geq 0$ since the function e^{tx} is increasing

$$P(X \geq a) = P(e^{tX} \geq e^{ta}) \leq \frac{E[e^{tX}]}{e^{ta}}$$

Since this holds for any $t \geq 0$ we can then optimize over t. The second inequality is proved in the same manner. □

- Chernov inequality is a very sharp inequality as we will explore later on when studying the law of large numbers. The optimization over t is the key ingredient which ensures sharpness.
- The function $M(t) = E[e^{tX}]$ is called the moment generating function for the RV X and we will meet again.
- Example Suppose X is a standard normal random variable $\mu = 0$ and σ^2. Then, completing the square, we have

$$E[e^{tX}] = \frac{1}{\sqrt{2\pi}} \int e^{tx} e^{-x^2/2\sigma^2} dx = \int e^{-(x-\sigma^2 t)^2/2} e^{\sigma^2 t^2/2} = e^{\sigma^2 t^2/2}$$

and Chernov bound gives for $a \geq 0$ that $P(X \geq a) \leq \sup_{t \geq 0} e^{\sigma^2 t^2/2 - ta} = e^{-\inf_{t \geq 0}(ta - \sigma^2 t^2/2)} = e^{-a^2/2\sigma^2}$ which turns out to be sharp up to a prefactor (see exercises).

5.5 Homework problems

Exercise 5.1

- Prove the *one-sided Chebyshev inequality*: if X is a random variable and $\epsilon > 0$ then

$$P(X - E[X] \geq \epsilon) \leq \frac{\sigma^2}{\sigma^2 + \epsilon^2}$$

  where $\sigma^2 = \mathrm{Var}(X)$.
  Hint: Set $Y = X - E[X]$ and use Markov inequality for $P(Y \geq \epsilon) = P((Y + \alpha)^2 \geq (\epsilon + \alpha)^2)$ and optimize over α

- Prove that
  - The one-sided Chebyshev inequality is sharper than the Chebyshev inequality for one sided bounds $P(X - E[X] \geq \epsilon)$.
  - The Chebyshev inequality is is sharper than the one-sided Chebyshev inquality for two sided bound $P(|X - E[X]| \geq \epsilon)$

Exercise 5.2 Prove Theorem 5.4. For the monotonicity use Hölder or Jensen. For the continuity let $p_n \nearrow q$ and use the dominated convergence theorem.

Exercise 5.3 The Chernov bound has the form

$$P(X \geq a) \leq e^{-\sup_{t \geq 0}\{ta - \ln E[e^{tX}]\}}.$$

Show that this bound is useful only if $a > E[X]$. To do this use Jensen inequality to show that if $a \leq E[X]$ the Chernov bound is trivial.

Exercise 5.4 Consider an exponential RV X with parameter λ and density $\lambda e^{-\lambda t}$ for $t \geq 0$.

- Compute $M(t) = e^{tX}$ as well as all moments $E[X^n]$
- To do a Chernov bound compute $\sup_{t \geq 0}\{ta - \ln E[e^{tX}]\}$ (see also Exercise 5.3).
- For $a > \frac{1}{\lambda}$ estimate $P(X > a)$ (which of course is equal to $e^{-\lambda a}$!) using
  a. Markov inequality.
  b. Chebyshev
  c. One-sided Chebyshev
  d. Chernov

Exercise 5.5 (mean and median of a RV) A *median* m for a RV X is a value m such that $P(X \leq m) \geq \frac{1}{2}$ and $P(X \geq m) \geq 1/2$ (see the quantile). For example if the CDF is one-to-one then the median $m = F^{-1}(\frac{1}{2})$ is unique.

The median m and the mean $\mu = E[X]$ are two measures (usually distinct) of the "central value" of the RV X.

- Consider the minimum square deviation $\min_{a \in \mathbb{R}} E[(X-a)^2]$. Show that the minimum is attained when $a = E[X]$.
- Consider the minimum absolute deviation $\min_{a \in \mathbb{R}} E[|X-a|]$. Show that the minimum is attained when a is a median m.
  Hint: Suppose $a > m$ then we have

$$|z-a| - |z-m| = \begin{cases} m-a & \text{if } z \geq a \\ a+m-2z \quad (\geq m-a) & \text{if } m < z < a \\ a-m & \text{if } z \leq m \end{cases}$$

Exercise 5.6 (mean and median of a RV, continued) Our goal is to prove a bound on how far the mean and the median can be apart from each other, namely that

$$|\mu - m| \leq \sigma$$

where σ is the standard deviation of X. I am asking for two proofs:

- First proof: Use the characterization of the median in Exercise 5.5 and Jensen inequality (twice) starting from $|\mu - m|$.
- Second proof: Use the one-sided Chebyshev for X and $-X$ with $\epsilon = \sigma$.

The quantity

$$S = \frac{\mu - m}{\sigma} \in [-1, 1]$$

is called the non-parametric skew of X and measures the assymetry of the distribution of X.