Cumulative Distribution Function

MATH/STAT 394: Probability I
Summer 2021 A Term

Introduction to Probability
D. Anderson, T.Seppäläinen, B. Valkó

§ 3.2

Aaron Osgood-Zimmerman

Department of Statistics

Recap

Continuous r.v.

- A r.v. is a continuous r.v. if it has a **probability density function** (p.d.f.) f such that
$$\mathbb{P}(X \leq b) = \int_{-\infty}^{b} f(x)dx$$
- This implies that
  - $\mathbb{P}(X = c) = 0$
  - $\mathbb{P}(a \leq X \leq b) = \int_a^b f(x)dx$
  - $\mathbb{P}(X \in B) = \int_{x \in B} f(x)dx$ (for B a Borel set)
- A function is a valid p.d.f. if
  - $f(x) \geq 0$ for all[1] $x \in \mathbb{R}$
  - $\int_{-\infty}^{+\infty} f(x)dx = 1$
- Interpretation:
  f is a **density** so to get the "weight" (i.e. the probability) of some intervals, you integrate the density over the interval.
- Similar to how given that gold has a density of 19.32g per cubic cm, then to know the weight of a gold bar you would integrate this density over the volume of the bar.

[1] Not exactly as shown in this lecture

Recap

Some continuous r.v.

- $X \sim \text{Unif}([a,b])$ if it has a p.d.f

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{if } x \in [a,b] \\ 0 & \text{if } x \notin [a,b] \end{cases}$$

- $X \sim \text{Exp}(\lambda)$ *(waiting time)* if it has a p.d.f.

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}$$

- $X \sim \text{Gaussian}(\mu, \sigma)$ *(aka normal, bell curve)* if it has a p.d.f.

$$f(x) = \begin{cases} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} & \text{if } x \in \mathbb{R} \\ 0 & \text{if } x \notin \mathbb{R} \end{cases}$$

with $\mu \in \mathbb{R}$ and $\sigma \geq 0$

Practice solution

Practice

Determine if there exist some values a, b, c such that the following functions satisfy the p.d.f. requirements

$$f_1(x) = \frac{a}{x^2+1}$$
$$f_2(x) = \begin{cases} b\cos(x) & \text{if } x \in [0, 2\pi] \\ 0 & \text{if } x \notin [0, 2\pi] \end{cases}$$
$$f_3(x) = \begin{cases} cx^{-4} & \text{if } x \geq 1 \\ 0 & \text{if } x \leq 1 \end{cases}$$

Solution

- $f_1(x) \geq 0$ for all x. The anti-derivative of $1/(1+x^2)$ is $\arctan(x)$ (because $\tan'(x) = 1 + \tan^2(x)$ and for f and f^{-1} such that $f \circ f^{-1}(x) = x$, we have that $(f^{-1})'(x) = 1/f'(f^{-1}(x))$). So $\int_{-\infty}^{+\infty} f_1(x)dx = a\left[\arctan(x)\right]_{-\infty}^{+\infty} = a\pi$ so for $a = \frac{1}{\pi}$, $f_1(x)$ is normalized.
- for $b \neq 0$, $f_2(x)$ is negative on non-empty intervals and for $b = 0$ it cannot be normalized.
- $f_3(x) \geq 0$ for all x and $\int_{-\infty}^{+\infty} f_3(x)dx = c\int_1^{+\infty} x^{-4}dx = c\left[-\frac{1}{3}x^{-3}\right]_1^{+\infty} = \frac{c}{3}$. So for $c = 3$, f_3 is normalized.

Outline

Borel algebra

Sets for uncountable sample spaces

Definition (Borel algebra)
The Borel algebra $\mathcal{B}$ on $\mathbb{R}$ is the smallest[2] set of subsets of $\mathbb{R}$ such that

1. $[a, +\infty)$ belongs to $\mathcal{B}$
2. if $A \in \mathcal{B}$, then $A^c \in \mathcal{B}$
3. if $A_1, A_2, \ldots \in \mathcal{B}$, then $\bigcup_{i=1}^{+\infty} A_i \in \mathcal{B}$

Note:

- By 2., $(-\infty, a) \in \mathcal{B}$ for any a
- By 3. and 2. using $(A \cup B)^c = A^c \cap B^c$
  for any $A_1, A_2, \ldots \in \mathcal{B}$, $\bigcap_{i=1}^{+\infty} A_i \in \mathcal{B}$
- Therefore
  $$[a, b) = [a, +\infty) \cap (-\infty, b) \in \mathcal{B} \text{ for any } a < b$$
- Since $\{b\} = \cap_{n=1}^{+\infty} [b, b + 1/n)$, $\{b\} \in \mathcal{B}$ and $[a, b] = [a, b) \cup \{b\} \in \mathcal{B}$

[2] smallest in the sense of inclusion. Namely a set A is smaller than a set B if $A \subseteq B$

Outline

Cumulative Distribution Function

Motivation

- For discrete r.v. we have the p.m.f.
- For continuous r.v. we have the p.d.f.
- Can we have a function that would give us all the information we need while being defined in both cases?
- Reminder: all we want is to be able to compute

$$\mathbb{P}(X \in B) \quad \text{for any Borel set } B \subset \mathbb{R}$$

- The Borel sets are built by taking unions/intersections/complements of sets $(-\infty, t]$ (or $[t, \infty)$) for any $t \in \mathbb{R}$
- So if we know

$$\mathbb{P}(X \leq t) = \mathbb{P}(X \in (-\infty, t])$$

  we sould be able to compute $\mathbb{P}(X \in B)$ for any Borel set B by the usual rules of prob. (using e.g. inclusion-exclusion principle)

Cumulative distribution function

Definition

Let X be a r.v., its ***cumulative distribution function*** *(c.d.f.) is defined as*

$$F(t) = \mathbb{P}(X \leq t) \quad \textit{for any} \quad t \in \mathbb{R}$$

Properties

- The c.d.f. of a r.v. is always **non-decreasing**, for $a \leq b$, clearly

$$F(a) = \mathbb{P}(X \leq a) \leq \mathbb{P}(X \leq b) = F(b)$$

- Moreover

$$\lim_{t \to -\infty} F(t) = 0, \quad \lim_{t \to +\infty} F(t) = 1$$

(because $\mathbb{P}(X \leq -\infty) = 0$ and $\mathbb{P}(X \leq +\infty) = 1$)

Outline

Cumulative distribution function of discrete r.v.

Example

Consider $X \sim \text{Bin}(2, 2/3)$. What is the c.d.f. of X?

Solution

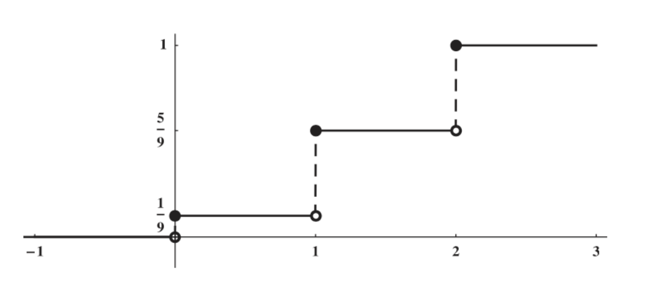
- First recall that

$$\mathbb{P}(X=0) = \left(\frac{1}{3}\right)^2 = \frac{1}{9}, \quad \mathbb{P}(X=1) = 2 \cdot \frac{2}{3} \cdot \frac{1}{3} = \frac{4}{9}, \quad \mathbb{P}(X=2) = \left(\frac{2}{3}\right)^2 = \frac{4}{9}$$

- Now for $t < 0$, $F(t) = \mathbb{P}(X \leq t) = 0$
- For $0 \leq t < 1$, $F(t) = \mathbb{P}(X = 0) = \frac{1}{9}$
- For $1 \leq t < 2$, $F(t) = \mathbb{P}(X = 0) + \mathbb{P}(X = 1) = \frac{5}{9}$
- For $2 \leq t$, $F(t) = \mathbb{P}(X = 0) + \mathbb{P}(X = 1) + \mathbb{P}(X = 2) = 1$
- So

$$F(t) = \begin{cases} 0 & \text{if } t < 0 \\ 1/9 & \text{if } 0 \leq t < 1 \\ 5/9 & \text{if } 1 \leq t < 2 \\ 1 & \text{if } 2 \leq t \end{cases}$$

Cumulative distribution function of discrete r.v.



c.d.f. of $X \sim \text{Bin}(2, 2/3)$

Figure from Introduction to probability, D. Anderson, T. Seppäläinen, B. Valkò

Cumulative distribution function of discrete r.v.

Example
Let $X \sim \text{Ber}(p)$. What is the c.d.f. of X?

Solution

$$F(t) = \begin{cases} 0 & \text{if } t < 0 \\ 1 - p & \text{if } 0 \leq t < 1 \\ 1 & \text{if } 1 \leq t \end{cases}$$

Cumulative distribution function of discrete r.v.

Exercise

Let $X \sim \mathrm{Geom}(p)$. What is the c.d.f. of X?

Solution

- We have for k an integer,

$$\mathbb{P}(X \leq k) = \sum_{\ell=1}^{k} (1-p)^{\ell-1} p = p \frac{1-(1-p)^k}{1-(1-p)} = 1-(1-p)^k$$

- So

$$F(t) = \begin{cases} 0 & \text{if } t < 1 \\ 1-(1-p)^k & \text{if } k \leq t < k+1 \quad \text{for } k \in \mathbb{N} \setminus \{0\} \end{cases}$$

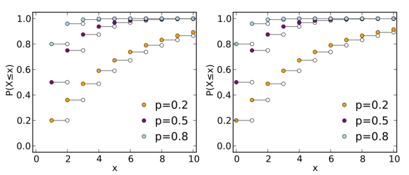


Figure from Wikipedia

Cumulative distribution function of discrete r.v.

From p.d.f. to c.d.f.

For a discrete r.v. X with p.m.f. p, its c.d.f. is

$$\boxed{F(t) = \mathbb{P}(X \leq t) = \sum_{\substack{k \leq t \\ k \in \mathcal{X}}} p(k)}$$

Properties of the c.d.f. of a discrete r.v.

- Piece-wise constant (i.e. constant on some intervals $[a_i, a_{i+1}]$ that cover $\mathbb{R}$)
- The points of discontinuity of F are the points of non-zero probability mass of X
- Namely, $F(t)$ jumps at $t = k$ iff k is a possible value, and

$$\mathbb{P}(X = k) = p(k) = \text{height of the jump at k}$$

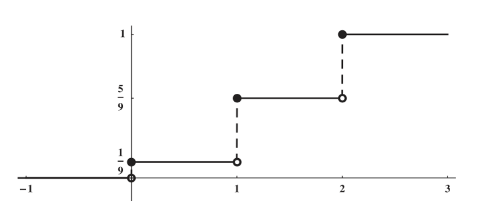
Cumulative distribution function of discrete r.v.

Right and left limits of a function

For a function F, we denote, if it exists,

$$F(t+) = \lim_{s \to t^+} F(s) := \lim_{\substack{s \to t \\ s > t}} F(s) \qquad \text{(Right limit)}$$

$$F(t-) = \lim_{s \to t^-} F(s) := \lim_{\substack{s \to t \\ s < t}} F(s) \qquad \text{(Left limit)}$$



c.d.f. of $X \sim \text{Bin}(2, 2/3)$

In black the right limit, in white the left limit at the discontinuity points.

Figure from Introduction to probability, D. Anderson, T. Seppäläinen, B. Valkò

Cumulative distribution function of discrete r.v.

Lemma
If X has a c.d.f. F then[3]

$$F(t+) = F(t) = \mathbb{P}(X \leq t) \qquad \textit{(Right continuity)}$$
$$F(t-) = \mathbb{P}(X < t)$$

From c.d.f. to p.m.f.

For any $k \in \mathbb{R}$,

$$\boxed{\mathbb{P}(X = k) = \mathbb{P}(X \leq k) - \mathbb{P}(X < k) = F(k) - F(k-)}$$

If, in addition, F is piecewise-constant, then X is a discrete r.v. with a p.m.f. defined above.

[3] See supp. slides for a proof

Cumulative distribution function of discrete r.v.

Exercise

Assume that a r.v. X has the following c.d.f.

$$F(t) = \begin{cases} 0 & \text{if } t < 1 \\ 1/5 & \text{if } 1 \leq t < 3 \\ 2/3 & \text{if } 3 \leq t < 4 \\ 1 & \text{if } 4 \leq t \end{cases}$$

What is the p.m.f. of X?

Solution Values with non-zero weight are $\{1, 3, 4\}$ and we have

$$\mathbb{P}(X = 1) = F(1) - F(1-) = \frac{1}{5} - 0 = \frac{1}{5}$$

$$\mathbb{P}(X = 3) = F(3) - F(3-) = \frac{2}{3} - \frac{1}{5} = \frac{7}{15}$$

$$\mathbb{P}(X = 4) = F(4) - F(4-) = 1 - \frac{2}{3} = \frac{1}{3}$$

Outline

Borel Sets

Cumulative distribution function

c.d.f. of discrete r.v.

c.d.f. of continuous r.v.

Proofs

Cumulative distribution function for continuous r.v.

From p.d.f. to c.d.f.

For a continuous r.v. X with p.d.f. f, the c.d.f. of X is (by definition)

$$F(t) = \mathbb{P}(X \leq t) = \int_{-\infty}^{t} f(t)dt$$

Properties

- If X is continuous, then its c.d.f. is continuous
  (because it is defined through an integral)

Cumulative distribution function for continuous r.v.

Example
Let $X \sim \text{Unif}([a,b])$ what is the c.d.f. of X?

Recall that $f(x) = \begin{cases} \frac{1}{b-a} & \text{if } x \in [a,b] \\ 0 & \text{if } x \notin [a,b] \end{cases}$

Solution By definition,

$$F(t) = \int_{-\infty}^{t} f(x)dx = \begin{cases} 0 & \text{if } t \leq a \\ \frac{t-a}{b-a} & \text{if } a \leq t \leq b \\ 1 & \text{if } b \leq t \end{cases}$$

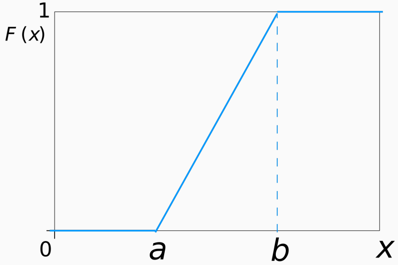


Figure from Wikipedia

Cumulative distribution function for continuous r.v.

Example

Let $X \sim \text{Exp}(\lambda)$ what is the c.d.f. of X?

Recall that $f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}$.

Solution By definition, if $t \leq 0$, $\int_{-\infty}^{t} f(x)dx = 0$. If $t > 0$,

$$F(t) = \int_{-\infty}^{t} f(x)dx = \int_{0}^{t} \lambda e^{-\lambda x} dx = \left[-e^{-\lambda x}\right]_0^t = 1 - e^{-\lambda t}$$

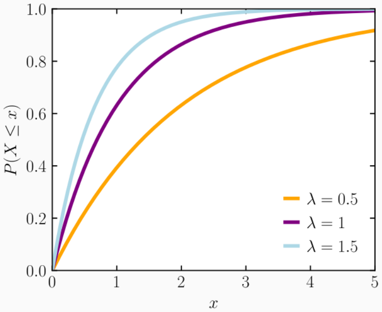


Figure from Wikipedia

Cumulative distribution function, continuous r.v.

From c.d.f. to p.d.f.

If X has a c.d.f. F that is differentiable everywhere, then the p.d.f. of X is given as

$$f(x) = F'(x)$$

Proof Because we then have by definition

$$\mathbb{P}(X \leq t) = F(t) - \underbrace{F(-\infty)}_{0} = \int_{-\infty}^{t} f(x)dx$$

Note:

- In fact, it is sufficient for F to be differentiable at all but a countable (finite or infinite) number of points.
- Why? Because when integrating f, the points of discontinuity of f won't matter.

Non-unicity of the p.d.f.

Let f be the p.d.f. of a r.v. X. Define for $d \in \mathbb{R}$, $c \in \mathbb{R}$,

$$g(x) = \begin{cases} f(x) & \text{if } x \neq c \\ d & \text{if } x = c \end{cases}$$

Then for an $a \leq c \leq b$,

$$\int_a^b g(x)dx = \int_a^c g(x)dx + \int_c^b g(x)dx = \int_a^c f(x)dx + \int_c^b f(x)dx = \int_a^b f(x)dx$$

Therefore g can also be used to compute the prob. of X

Takeaway:

- The p.d.f. of a continuous r.v. is not unique!
- It is uniquely defined **almost everywhere** (a.w), meaning that except from a set $A \subseteq \mathbb{R}$ (e.g. $A = \{k_1, \ldots, k_n\}$) s.t.

  $$\int_{x \in A} dx = 0$$

  it is uniquely defined.

Practice next lecture

Practice

John has an insurance policy on his car with a 200\$ deductible, meaning that if an accident occur, he would pay the cost of the repair up to 200\$ with the insurance policy paying the rest. So if he has an accident worth 123\$ he would pay 123\$ but if the accident is worth 345\$ he would only pay 200\$.

Assume that the cost of an accident is uniformly distributed over $[50, 1000]$. Denote X the amount that John pays.

- What is the c.d.f. of X?
- Is X continuous, discrete or neither discrete nor continuous?
- What is the prob. that $X = 200$?

Outline

Right continuity proof*

Recall the Prop. Slide 14 of Lecture 7:

Proposition

Let $A_1, A_2, \dots$ be a sequence of subsets, s.t. $A_1 \subseteq A_2 \subseteq \dots$.

Let $A_\infty = \cup_{k=1}^{+\infty} A_k$. Then

$$\lim_{n\to+\infty} \mathbb{P}(A_n) = \mathbb{P}(A_\infty)$$

This implies the following

Corollary
Let $A_1, A_2, \dots$ be a sequence of subsets, s.t. $A_1 \supseteq A_2 \supseteq \dots$.

Let $A_\infty = \cap_{k=1}^{+\infty} A_k$. Then

$$\lim_{n\to+\infty} \mathbb{P}(A_n) = \mathbb{P}(A_\infty)$$

Proof Just consider the complements in the previous proposition and use De Morgan's law.

Right continuity proof*

Proposition

The c.d.f. of a r.v. is right continuous, i.e.

$$F(t+) = \lim_{\substack{s \to t \\ s > t}} F(s) = F(t)$$

Proof

- We want to show that for any non-increasing sequence $(x_n)_{n=0}^{+\infty}$ s.t. $x_n > t$ for all n and $\lim_{n \to +\infty} x_n = t$,
$$\lim_{n \to +\infty} F(x_n) = \lim_{n \to +\infty} \mathbb{P}(X \leq x_n) = \mathbb{P}(X \leq t) = F(t)$$
- Let $(x_n)_{n=0}^{+\infty}$ be such a sequence.
- Define $A_n = \{\omega : X(\omega) \leq x_n\}$. Clearly $A_{n+1} \subseteq A_n$ since x_n is non-increasing
- Now denote $A_\infty = \cap_{k=1}^{+\infty} A_k$, we have that if $\omega \in A_\infty$, then since x_n is non-increasing and converges to t, $\omega \in \{\omega : X(\omega) \leq t\} := A_t$
- On the other hand, clearly if $\omega \in A_t$ then $\omega \in A_\infty$ since $x_n > t$.
- Therefore using the previous corollary,
$$\lim_{n \to +\infty} F(x_n) = \lim_{n \to +\infty} \mathbb{P}(A_n) = \mathbb{P}(A_\infty) = \mathbb{P}(X \leq t) = F(t)$$

Left limit*

Proposition

The c.d.f. of a r.v. satisfies

$$F(t-) = \lim_{\substack{s \to t \\ s < t}} F(s) = \mathbb{P}(X < t)$$

Proof

- Let $(x_n)_{n=1}^{+\infty}$ non-decreasing, with $x_n > t$ and $\lim_{n \to +\infty} x_n = t$.
- Define $A_n = \{\omega : X(\omega) \leq x_n\}$. Clearly $A_{n+1} \supseteq A_n$ since x_n is non-decreasing
- Let $A_\infty = \cup_{k=1}^{+\infty} A_k$ and $A_t = \{\omega : X(\omega) < t\}$.
- If $\omega \in A_t$ then $X(\omega) = s < t$ and $\exists x_n$ s.t. $x_n \geq s$ so $\omega \in A_\infty$
- On the other hand if $\omega \in A_\infty$ then $\exists A_p$ s.t. $\omega \in A_p$ so $X(\omega) \leq x_n < t$
- The result follows from the first Prop. presented in the previous slides.