Applied Statistics

for Computer Science BSc, Exam

Probability Theory and Mathematical Statistics

for Computer Science Engineering BSc, Term grade

István Fazekas
University of Debrecen

2020/21 fall

This work was supported by the construction EFOP-3.4.3-16-2016-00021. The project was supported by the European Union, co-financed by the European Social Fund.

Main topics

1. Probability theory
2. Statistics

Mathematical tools: combinatorics, calculus

Computer tool: Matlab

Book:

Yates, Goodman:

Probability and Stochastic Processes: A Friendly Introduction for Electrical and Computer Engineers

Lecture 6

The general notion of a random variable

Definition of a random variable

Let $(\Omega, \mathcal{F}, P)$ be a probability space.
The function $X : \Omega \to \mathbb{R}$ is called a random variable, if for any fixed $t \in \mathbb{R}$

$$\{\omega : X(\omega) < t\} \in \mathcal{F}.$$

It means that the set $\{\omega : X(\omega) < t\}$ is an event, that is its probability is defined.

Remark. Any discrete random variable is a random variable in the general sense.

Cumulative distribution function, CDF

The cumulative distribution function (CDF) of the random variable X is

$$\boxed{F(t) = P\{\omega : \; X(\omega) < t\}}, \quad t \in \mathbb{R}.$$

Remarks.

1. Any random variable has a CDF.
2. The CDF is a real function.
3. To calculate $F(t)$ first we should fix the value of t and then find the value of $F(t) = P\{\omega : \; X(\omega) < t\}$.

CDF of a discrete random variable

Let X be a discrete random variable with distribution $P(X = x_i) = p_i$, $i = 1, 2, \dots$.
Then its CDF is

$$F(t) = P(X < t) = \sum_{\{i : x_i < t\}} P(X = x_i) = \sum_{\{i : x_i < t\}} p_i.$$

Therefore in this particular case the CDF is a step function 'jumping' p_i at the point x_i.
Example. The CDF of the constant r.v. $X = c$ is

$$F(t) = \begin{cases} 0, & \text{if} \quad t \leq c, \\ 1, & \text{if} \quad t > c. \end{cases}$$

The CDF of a discrete random variable

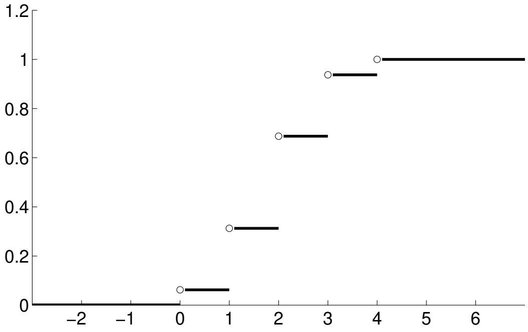

Figure: The CDF of the binomial random variable with parameters $p = 1/2,\ n = 4$

Properties of a CDF

Theorem. The function $F : \mathbb{R} \to \mathbb{R}$ is a CDF if and only if

a) F is monotone increasing,

b) F is left continuous,

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

Proof. I. Let F be the CDF of X. Then

a) use the monotonicity of the probability,

b) use the continuity of the probability,

c) use the continuity of the probability and $P(\Omega) = 1$ and $P(\emptyset) = 1$.

II. Now let F satisfy properties a)-b)-c).

We prove only in the case when F is strictly increasing and continuous. Then it has an inverse F^{-1}.

Let $\Omega = (0, 1)$, let P be the length and let $X = F^{-1}$.

Then the CDF of X is F.

A CDF and its inverse

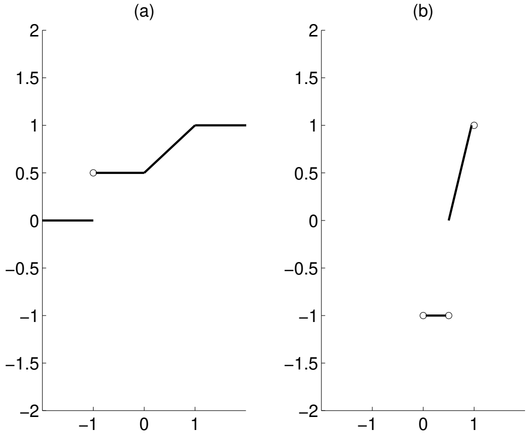

Figure: A CDF and its inverse

The uniform distribution

Exercise. Choose a point randomly from the interval $[a, b]$.
Denote by X the position of the point.
Find the CDF of X.
Solution.
The CDF is

$$F(t) = \begin{cases} 0, & \text{if} \quad t \leq a, \\ \frac{t-a}{b-a}, & \text{if} \quad a < t \leq b, \\ 1, & \text{if} \quad b < t. \end{cases}$$

It is called the uniform distribution on the interval $[a, b]$.
Observe, that F satisfies properties a)-b)-c)

The CDF of the uniform distribution

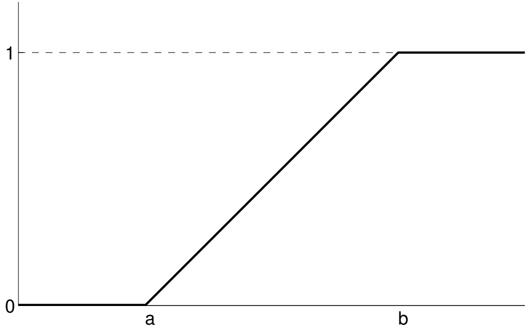

Figure: The CDF of the uniform distribution

Using CDF to calculate probabilities

Proposition. Let F be the CDF of X. Then
a) $P(X \in [a, b)) = F(b) - F(a)$;
b) $P(X = a) = F(a + 0) - F(a)$;
c) $P(X \in [a, b]) = F(b + 0) - F(a)$.
Proof. a) $P(X \in [a, b)) = P(X < b) - P(X < a) = F(b) - F(a)$.
b) $\{X = a\} = \bigcap_{n=1}^{\infty} \{X \in [a, a + 1/n)\}$. Then use continuity of the probability.
c) Use a) and b)

Remark. It follows from b) that F is continuous in t if and only if $P(X = t) = 0$.
Homework. Prove that $P(X \in (a, b)) = F(b) - F(a + 0)$ and $P(X \in (a, b]) = F(b + 0) - F(a + 0)$.

Median, quantiles, quartiles

Definition. μ is called the median of X if

$$P(X < \mu) \leq 1/2 \quad \text{and} \quad P(X > \mu) \leq 1/2.$$

Remark.
1. The median exists, but it is not always unique.
2. If F strictly increasing and continuous then the median is the only solution of the equation $F(t) = 1/2$.

Definition. Let $0 < q < 1$. Then $Q(q)$ is called the q-quantile of X if

$$P(X < Q(q)) \leq q \quad \text{and} \quad P(X > Q(q)) \leq 1 - q.$$

The 0.25-quantile is called the lower quartile,
the 0.75-quantile is called the upper quartile.

Median, quantiles, quartiles

Exercise.
Find the median and the quartiles of the uniform distribution.
Solution.
The median is

$$\mu = \frac{a+b}{2}$$

The lower quartile is

$$\frac{3a+b}{4}$$

The upper quartile is

$$\frac{a+3b}{4}$$

The Cauchy distribution

Exercise. Show that

$$F(t) = (1/\pi)\arctan t + 1/2$$

is a CDF.
It is called Cauchy distribution.
Solution. Check properties a)-b)-c).

Exercise.
Find the median and the quartiles of the Cauchy distribution.

The Cauchy distribution

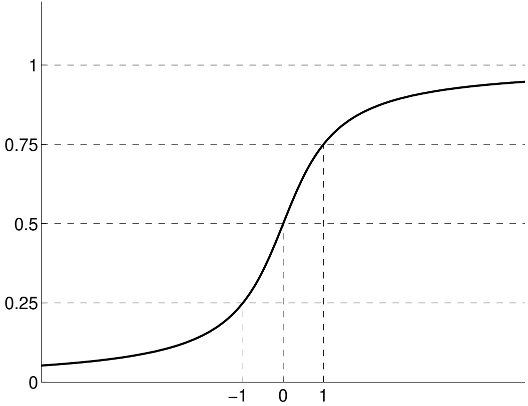

Figure: The CDF, the median, and the quartiles of the Cauchy distribution

Homework

Exercise 1.

Let X be uniformly distributed on the interval $(0, 1)$. Find the distribution function of

a) $Y = X + 1$;

b) $Y = 3X$;

c) $Y = X^3$;

d) $Y = \sqrt{X}$;

e) $Y = |X - 1/2|$.

Hint.

$$F_Y(t) = P(Y < t) = P(X+1 < t) = P(X < t-1) = F_X(t-1) = \dots$$

The probability density function, PDF

Definition.
Let F be the CDF of X.
We say that the distribution of X is absolutely continuous, if there exists a function $f : \mathbb{R} \to \mathbb{R}$ so that

$$\boxed{F(t) = \int_{-\infty}^{t} f(s)\, ds}, \quad \forall t \in \mathbb{R},$$

f is called the probability density function (PDF) of X.

Remark.
1. Usually we use

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

2. If the PDF exists, then the CDF should be continuous.
Therefore a discrete random variable has no PDF.

PDF of the uniform distribution

Example. Let

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

One can show that $\int_{-\infty}^{t} f(s)\,ds = F(t)$ with

$$F(t) = \begin{cases} 0, & \text{if} \quad t \leq a, \\ \frac{t-a}{b-a}, & \text{if} \quad a < t \leq b, \\ 1, & \text{if} \quad b < t. \end{cases}$$

So f is the PDF of the uniform distribution.

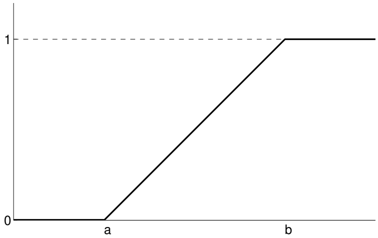

Figure: The CDF of the uniform distribution

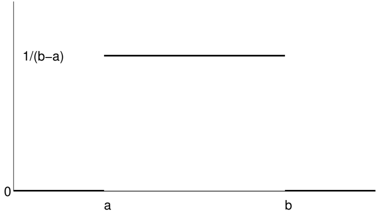

Figure: The PDF of the uniform distribution

Exercise for PDF

Choose a point from the unit square at random.
Let X denote the distance of the point from the nearest side of the square.
Find the CDF and the PDF of X. Visualize the CDF and the PDF.
Solution.

$$F(x) = \begin{cases} 0, & \text{if} \quad x \le 0, \\ 1 - (1-2x)^2 = 4x - 4x^2, & \text{if} \quad 0 < x \le 1/2, \\ 1, & \text{if} \quad x > 1/2. \end{cases}$$

The PDF is

$$f(x) = \begin{cases} 4 - 8x, & \text{if} \quad x \in [0, 1/2], \\ 0, & \text{if} \quad x \notin [0, 1/2]. \end{cases}$$

Using PDF

Theorem. Let f be the PDF of X. Then

$$\boxed{P(X \in B) = \int_B f(t)\, dt}\,, \tag{1}$$

for any Borel measurable set B on the real line.
In particular if X has PDF, then $P(X = t) = 0$ for any $t \in \mathbb{R}$.

Proof. $P(X \in [a, b)) = F(b) - F(a) = \int\limits_{-\infty}^{b} f(x)dx - \int\limits_{-\infty}^{a} f(x)dx$

$= \int\limits_{a}^{b} f(x)dx.$

So (1) is proved for $B = [a, b)$.
Moreover

$$P(X = x) = \lim_{n\to\infty} P\big(X \in [x, x + 1/n)\big) = \lim_{n\to\infty} \int\limits_{x}^{x+1/n} f(t)dt = 0$$

The PDF of the Cauchy distribution

Exercise. Show that

$$f(t) = 1/(\pi(1+t^2)), \quad t \in \mathbb{R}$$

is the PDF of the Cauchy distribution.
Visualize the PDF and $P(X \in [a, b])$.
Solution. $\int_{-\infty}^{x} f(t)dt = F(x)$ with $F(x) = (1/\pi)\arctan x + 1/2$.

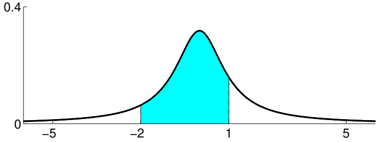

Figure: The PDF of the Cauchy distribution. The area of the blue domain is $P(X \in [a, b])$

The properties of the PDF

Theorem. $f : \mathbb{R} \to \mathbb{R}$ is PDF if and only if

1. f is Borel measurable,
2. $f(x) \geq 0$ for almost all x,
3. $\int_{-\infty}^{\infty} f(x)dx = 1$.

Proof.

One one hand, if f satisfies the above three conditions, then

$$F(t) = \int\limits_{-\infty}^{t} f(s)ds$$

satisfies the properties of a CDF.

On the on the other hand, if f is a PDF, then

$\int_{-\infty}^{\infty} f(x)dx = F(\infty) = 1$.

Moreover $\int_a^b f(x)dx = F(b) - F(a) \geq 0$.

So f is non-negative.

Exercise for PDF

Let

$$f(x) = \begin{cases} \sin x, & \text{if} \quad x \in [0, \pi/2], \\ 0, & \text{if} \quad x \notin [0, \pi/2]. \end{cases}$$

Visualize f.

Show that f is a PDF.

Find the corresponding CDF.

Solution.

$f(x)$ is continuous excluding one point, so it is measurable.

$f(x) \geq 0 \quad \forall x \in \mathbb{R}$.

$$\int\limits_{-\infty}^{\infty} f(x)dx = \int\limits_{0}^{\pi/2} \sin x dx = -\cos(\pi/2) + \cos 0 = 1.$$

Therefore f is a PDF.

The PDF of the exponential distribution

Exercise. Show that

$$f(x) = \begin{cases} 0, & x \leq 0, \\ \lambda e^{-\lambda x}, & x > 0. \end{cases}$$

is a PDF, where λ is a positive parameter.
Visualize the PDF.
Find the corresponding CDF.
Solution.
Check properties 1-2-3.

$$F(x) = \begin{cases} 0, & x \leq 0, \\ 1 - e^{-\lambda x}, & x > 0. \end{cases}$$

is the CDF.

The PDF of the exponential distribution

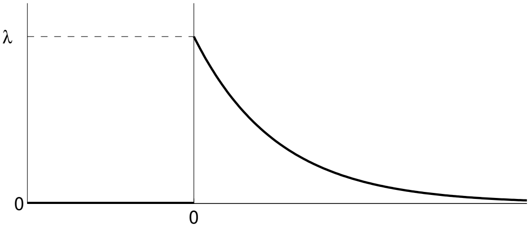

Figure: The PDF of the exponential distribution

The exponential distribution is memoryless

Let X have exponential distribution.
Show that

$$P(X < t + s | X \geq t) = P(X < s), \qquad t > 0,\ s > 0.$$

If a random variable X has continuous CDF and satisfies the above equation, then X is exponentially distributed.

Exercise. Let X be exponential with parameter $\lambda = 1$. Show that $Y = 1 - e^{-X}$ is uniformly distributed on the interval $[0, 1]$.