# Statistical Methods for Data Science

## Lesson 05 - Continuous random variables

Salvatore Ruggieri

Department of Computer Science
University of Pisa
salvatore.ruggieri@unipi.it

# Discrete random variables

DEFINITION. Let $\Omega$ be a sample space. A *discrete random variable* is a function $X : \Omega \to \mathbb{R}$ that takes on a finite number of values $a_1, a_2, \ldots, a_n$ or an infinite number of values $a_1, a_2, \ldots$.

DEFINITION. The *probability mass function* $p$ of a discrete random variable $X$ is the function $p : \mathbb{R} \to [0,1]$, defined by

$$p(a) = \mathrm{P}(X = a) \quad \text{for } -\infty < a < \infty.$$

Support finite or countable $\{a_1, \ldots, a_n, \ldots\}$

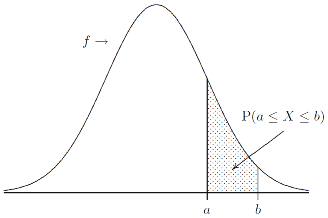
- $p(a_i) > 0$ for $i = 1, 2, \ldots$
- $p(a) = 0$ if $a \notin \{a_1, a_2, \ldots\}$
- $\sum_i p(a_i) = 1$

What happens when the support is uncountable? E.g., $[0,1]$ or $\mathbb{R}^+$ or $\mathbb{R}$

- $p(a_i)$ must be 0 because $|\mathbb{R}| = 2^{\aleph_0} > \aleph_0 = |\mathbb{N}|$
- hence, $\sum_i p(a_i) = 0$

# Continuous random variables

We cannot assign a "mass" to a real number, but we can assign it to an interval!




DEFINITION. A random variable $X$ is *continuous* if for some function $f : \mathbb{R} \rightarrow \mathbb{R}$ and for any numbers $a$ and $b$ with $a \le b$,

$$P(a \le X \le b) = \int_a^b f(x)\,dx.$$

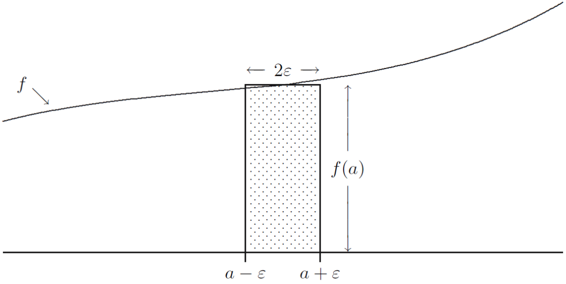
The function $f$ has to satisfy $f(x) \ge 0$ for all $x$ and $\int_{-\infty}^{\infty} f(x)\,dx = 1$. We call $f$ the *probability density function* (or *probability density*) of $X$.

$F(a) = P(X \le a) = \int_{-\infty}^{a} f(x)dx$ *[Cumulative Distribution Function]*

# Density function

$$P(X = a) \le P(a - \varepsilon \le X \le a + \varepsilon) = \int_{a-\varepsilon}^{a+\varepsilon} f(x)dx = F(a + \varepsilon) - F(a - \varepsilon)$$

for $\varepsilon \to 0$, $P(a - \varepsilon \le X \le a + \varepsilon) \to 0$, hence $P(X = a) = 0$
What is the meaning of the density function $f(x)$?
$f(a)$ is a (relative) measure of how likely is $X$ will be near $a$



# $X \quad U( \; ; \; )$

DEFINITION. A continuous random variable has a *uniform distribution* on the interval $[\alpha, \beta]$ if its probability density function $f$ is given by $f(x) = 0$ if $x$ is not in $[\alpha, \beta]$ and

$$f(x) = \frac{1}{\beta - \alpha} \qquad \text{for } \alpha \leq x \leq \beta.$$

We denote this distribution by $U(\alpha, \beta)$.

$F(x) = \int_{1}^{x} f(x)dx = \frac{1}{\quad} \int^{x} 1dx = \frac{x}{\quad}$ for $\qquad x$

**See R script**

# $X \quad Exp(\quad)$

For $X \quad Geo(p)$, we have: $F(x) = P(X \quad x) = 1 \quad (1 \quad p)^{bxc}$
extend to reals: $F(X) = P(X \quad x) = 1 \quad (1 \quad p)^{x} = 1 \quad e^{x\,log(1 \quad p)} = 1 \quad e^{\quad x}$ for $\quad = log(\frac{1}{1 \quad p})$
$f(x) = \frac{dF}{dx}(x) = \quad e^{\quad x}$

DEFINITION. A continuous random variable has an *exponential distribution* with parameter $\lambda$ if its probability density function $f$ is given by $f(x) = 0$ if $x < 0$ and

$$f(x) = \lambda e^{-\lambda x} \quad \text{for } x \geq 0.$$

We denote this distribution by $Exp(\lambda)$.

is the rate of events, e.g.,

- $\quad = \frac{1}{10}$ number of bus arrivals per minute, or $\frac{1}{\quad} = 10$ minutes to wait for bus arrival
- $P(X > 1) = 1 \quad P(X \quad 1) = e^{\quad} = 0{:}9048$ probability of waiting more than 1 minute.

Exponential is *memoryless*: $P(X > s + t jX > s) = \frac{e^{-\lambda\cdot(s+t)}}{e^{-\lambda\cdot s}} = e^{\quad t} = P(X > t)$

**See R script and seeing-theory.brown.edu**

# $X \quad N( \; ; \; ^2)$

DEFINITION. A continuous random variable has a *normal distribution* with parameters $\mu$ and $\sigma^2 > 0$ if its probability density function $f$ is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \quad \text{for } -\infty < x < \infty.$$

We denote this distribution by $N(\mu, \sigma^2)$.

Also called Gaussian distribution

Standard Normal/Gaussian is $N(0; 1)$

- $f(x) = \frac{1}{\sqrt{2}} e^{\frac{x^2}{2}}$ sometimes written as $(x)$
- No closed form for $F(a) = \Phi(a) = \int_{1}^{a} (x) dx$

# CCDF of $Z$ $N(0;1)$

**Table B.1.** Right tail probabilities $1 - \Phi(a) = \mathrm{P}(Z \geq a)$ for an $N(0,1)$ distributed random variable $Z$.

| $a$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0.0 | 5000 | 4960 | 4920 | 4880 | 4840 | 4801 | 4761 | 4721 | 4681 | 4641 |
| 0.1 | 4602 | 4562 | 4522 | 4483 | 4443 | 4404 | 4364 | 4325 | 4286 | 4247 |
| 0.2 | 4207 | 4168 | 4129 | 4090 | 4052 | 4013 | 3974 | 3936 | 3897 | 3859 |
| 0.3 | 3821 | 3783 | 3745 | 3707 | 3669 | 3632 | 3594 | 3557 | 3520 | 3483 |
| 0.4 | 3446 | 3409 | 3372 | 3336 | 3300 | 3264 | 3228 | 3192 | 3156 | 3121 |
| 0.5 | 3085 | 3050 | 3015 | 2981 | 2946 | 2912 | 2877 | 2843 | 2810 | 2776 |
| 0.6 | 2743 | 2709 | 2676 | 2643 | 2611 | 2578 | 2546 | 2514 | 2483 | 2451 |
| 0.7 | 2420 | 2389 | 2358 | 2327 | 2296 | 2266 | 2236 | 2206 | 2177 | 2148 |
| 0.8 | 2119 | 2090 | 2061 | 2033 | 2005 | 1977 | 1949 | 1922 | 1894 | 1867 |
| 0.9 | 1841 | 1814 | 1788 | 1762 | 1736 | 1711 | 1685 | 1660 | 1635 | 1611 |
| 1.0 | 1587 | 1562 | 1539 | 1515 | 1492 | 1469 | 1446 | 1423 | 1401 | 1379 |

E.g., $P(Z > 1{:}04) = 0{:}1492$

**See R script**

# Quantiles

DEFINITION. Let $X$ be a continuous random variable and let $p$ be a number between 0 and 1. The $p$th *quantile* or 100$p$th *percentile* of the distribution of $X$ is the smallest number $q_p$ such that

$$F(q_p) = \mathrm{P}(X \leq q_p) = p.$$

The *median* of a distribution is its 50th percentile.

If $F()$ is *strictly* increasing, $q_p = F \;\, ^1(p)$

E.g., for $Exp(\;\;)$, $F(a) = 1 \quad e \quad ^x$, hence $F \;\, ^1(p) = \frac{1}{\,} \log \frac{1}{(1 \;\; p)}$

**See R script**

# Simulation

Not all problems can be solved with calculus!

Complex interactions among random variables can be simulated

Generated random values are called *realizations*

Basic issue: *how to generate realizations?*

- in R: $rnorm(5); rexp(2); rbinom(\ldots); \ldots$

Ok, but how do they work?

**Assumption**: we are only given $runif()$!

**Problem**: derive all the other random generators

# Simulation: discrete distributions

**Bernoulli random variables**

Suppose $U$ has a $U(0,1)$ distribution. To construct a $Ber(p)$ random variable for some $0 < p < 1$, we define

$$X = \begin{cases} 1 & \text{if } U < p, \\ 0 & \text{if } U \geq p \end{cases}$$

so that

$$\begin{aligned} P(X = 1) &= P(U < p) = p, \\ P(X = 0) &= P(U \geq p) = 1 - p. \end{aligned}$$

This random variable $X$ has a Bernoulli distribution with parameter $p$.

For $X_1, \ldots, X_n \sim Ber(p)$ i.i.d., we have: $\sum_{i=1}^{n} X_i \sim Binom(n, p)$

**See R script**

# Simulation: continuous distributions

$F : \mathbb{R} \to [0;1]$ and $F^{-1} : [0;1] \to \mathbb{R}$

- E.g., $F$ *strictly* increasing
- N.B., the textbook notation for $F^{-1}$ is $F^{inv}$

For $X \sim U(0;1)$ and $0 \leq b \leq 1$

$$P(X \leq b) = b$$

then, for $b = F(x)$

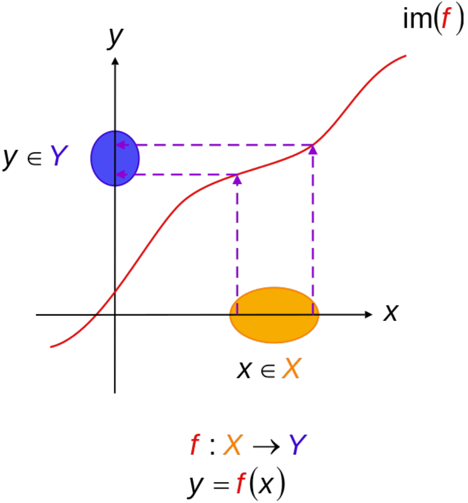
$$P(X \leq F(x)) = F(x)$$

and then by inverting

$$P(F^{-1}(X) \leq x) = F(x)$$

In summary:

$$F^{-1}(X) \sim F \text{ for } X \sim U(0;1)$$




**See R script**