Master Program in *Data Science and Business Informatics*

Statistics for Data Science

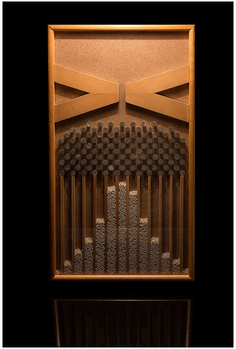
Lesson 12 - Simulation

Salvatore Ruggieri

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

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?*
  - The **Galton Board**



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

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

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

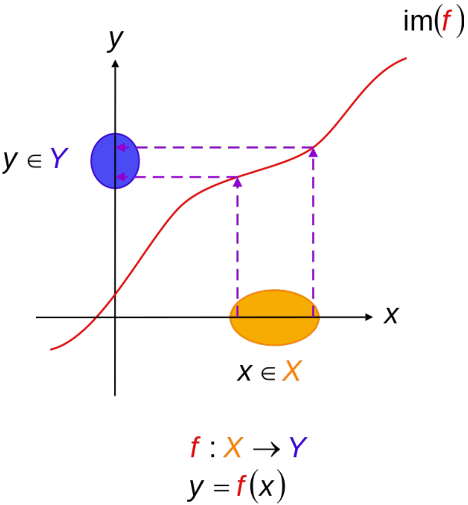
- $F : \mathbb{R} \to [0,1]$ and $F^{-1} : [0,1] \to \mathbb{R}$
  - E.g., F *strictly* increasing
  - N.B., the textbook notation for F^{-1} is F^{inv}
- For $X \sim U(0,1)$ and $0 \leq b \leq 1$

  $$P(X \leq b) = b$$
- then, for $b = F(x)$

  $$P(X \leq F(x)) = F(x)$$
- and then by inverting

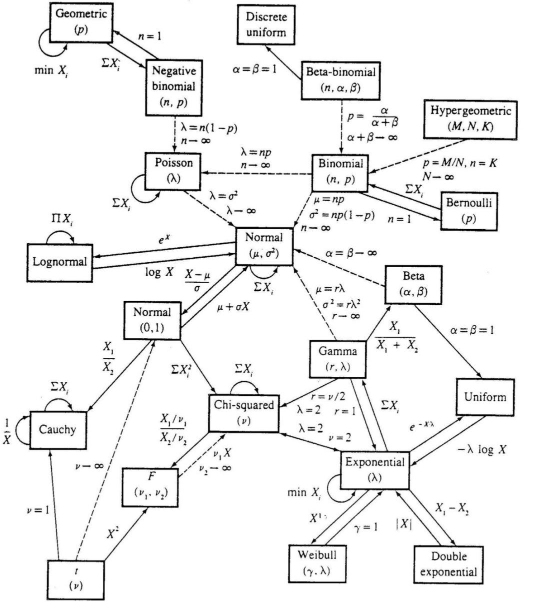
  $$P(F^{-1}(x) \leq x) = F(x)$$
- In summary:

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



See R script

Common distributions



Relationships among common distributions. Solid lines represent transformations and special cases, dashed lines represent limits. Adapted from Leemis (1986).

Optional reference

William H. Press, Saul A. Teukolsky, William T. Vetterling, Brian P. Flannery (2007)
Numerical Recipes - The Art of Scientific Computing
Chapter 7: Random Numbers
online book