Master Program in *Data Science and Business Informatics*

Statistics for Data Science

Lesson 12 - Simulation

Salvatore Ruggieri

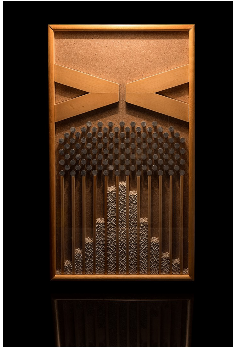
Department of Computer Science
University of Pisa, Italy
salvatore.ruggieri@unipi.it

Simulation

- Not all problems can be solved with calculus!
- Complex interactions among random variables can be simulated
- Generated random values are called *realizations*

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: $rnorm(5), rexp(2), rbinom(\ldots), \ldots$

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 $runif()$!

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: $\mathit{rnorm}(5), \mathit{rexp}(2), \mathit{rbinom}(\ldots), \ldots$
- Ok, but how do they work?
- **Assumption**: we are only given $\mathit{runif}()$!
- **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

$X \sim Cat(\mathbf{p})$

DEFINITION. A discrete random variable X has a *Bernoulli distribution* with parameter p, where $0 \leq p \leq 1$, if its probability mass function is given by

$$p_X(1) = \mathrm{P}(X = 1) = p \quad \text{and} \quad p_X(0) = \mathrm{P}(X = 0) = 1 - p.$$

We denote this distribution by $Ber(p)$.

- Alternative definition: $p_X(a) = p^a \cdot (1-p)^{1-a}$ for $a \in \{0, 1\}$
- Categorical distribution generalizes to $n \geq 2$ possible values

Categorical distribution

A discrete random variable X has a Categorical distribution with parameters $p_0, \ldots, p_{n_C - 1}$ where $\sum_i p_i = 1$ and $p_i \in [0, 1]$ if its p.m.f. is given by:

$$p_X(i) = P(X = i) = p_i \quad \text{for } i = 0, \ldots, n_C - 1$$

- Alternative definition: $p_X(a) = \prod_i p_i^{\mathbb{1}_{a==i}}$ for $a = 0, \ldots, n_C - 1$

$X \sim Mult(n, \mathbf{p})$

- $X \sim Bin(n, p)$ models the number of successes in n Bernoulli trials
- **Intuition**: for $X_1, X_2, \dots, X_n$ i.i.d. $X_i \sim Ber(p)$: $X = \sum_{i=1}^{n} X_i \sim Bin(n, p)$
- $X \sim Mult(n, \mathbf{p})$ models the number of categories in n Categorical trials
- **Intuition**: for $X_1, X_2, \dots, X_n$ such that $X_i \sim Cat(\mathbf{p})$ and independent (**i.i.d.**), define:

$$Y_1 = \sum_{i=1}^{n} \mathbb{1}_{X_i == 0} \sim Bin(n, p_0), \dots, Y_{n_C - 1} = \sum_{i=1}^{n} \mathbb{1}_{X_i == n_C - 1} \sim Bin(n, p_{n_C - 1})$$

$$X = (Y_1, \dots, Y_{n_C - 1}) \sim Mult(n, \mathbf{p})$$

Multinomial distribution

A discrete random variable $X = (Y_1, \dots, Y_{n_C - 1}$ has a Multinomial distribution with parameters $p_0, \dots, p_{n_C - 1}$ where $\sum_i p_i = 1$ and $p_i \in [0, 1]$ if its p.m.f. is given by:

$$p_X(i_0, \dots, i_{n_C - 1}) = P(X = (i_0, \dots, i_{n_C - 1})) = \frac{n!}{i_0! i_1! \dots i_{n_C - 1}!} p_0^{i_0} p_1^{i_1} \dots p_{n_C - 1}^{i_{n_C - 1}}$$

$X \sim Mult(n, \mathbf{p})$

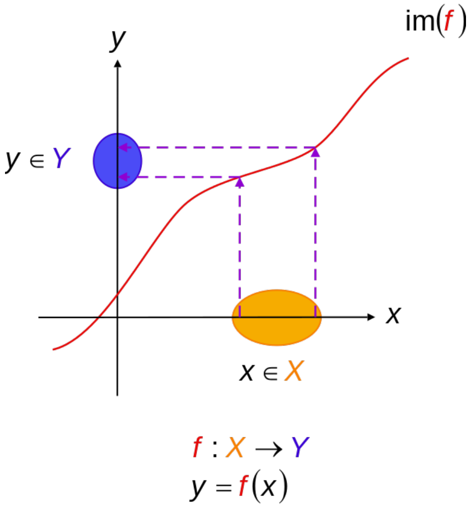
- Example: student selection from a population with:
  - 60% undergraduates
  - 30% graduate
  - 10% PhD students
- Assume $n = 20$ students are randomly selected
- $X \sim (Y_1, Y_2, Y_3)$ where:
  - Y_1 number of undergraduate students
  - Y_2 number of graduate students
  - Y_3 number of PhD students
- $P(X = (10, 6, 4)) = \frac{20!}{10!6!4!}(0.6)^{10}(0.3)^6(0.1)^4 = 9.6\%$

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 $0 \leq b \leq 1$

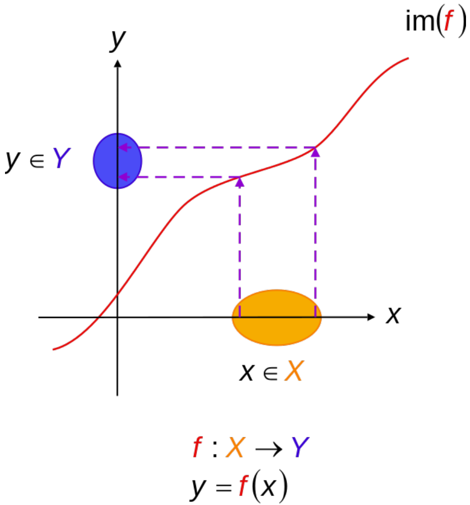
  $P(X \leq b) = b$



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



See R script

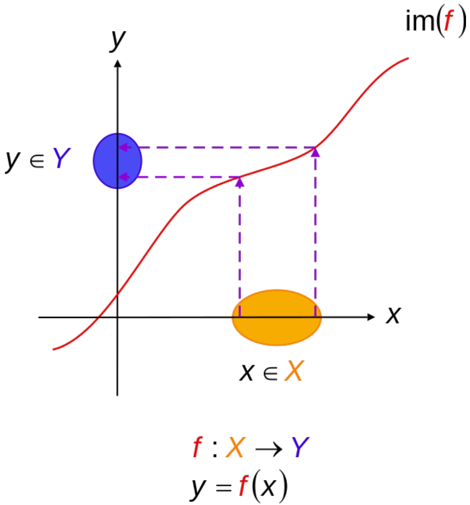
Simulation: continuous distributions

- $F : \mathbb{R} \to [0, 1]$ 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

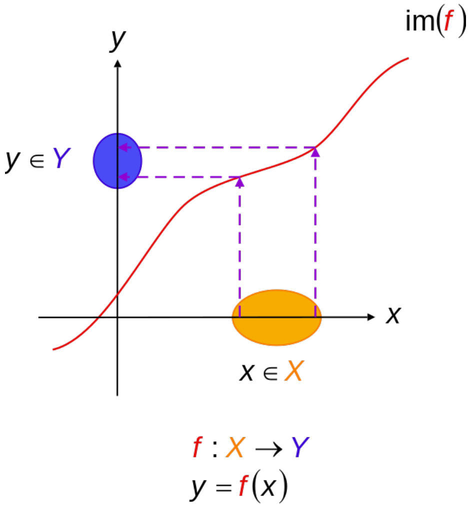
  $$P(F^{-1}(X) \leq x) = F(x)$$



See R script

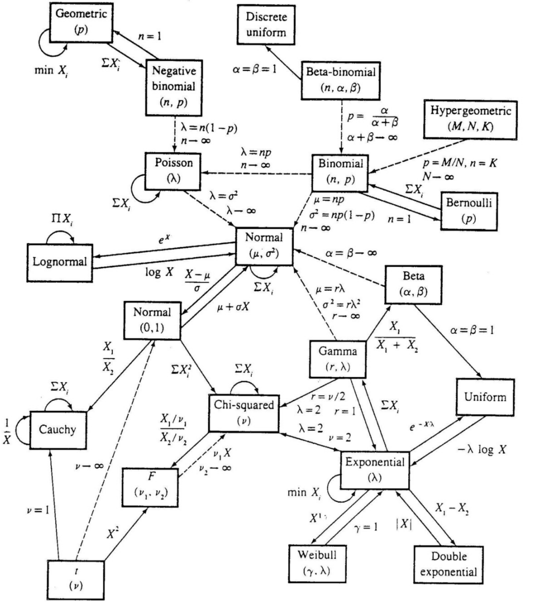
Simulation: continuous distributions

- $F : \mathbb{R} \to [0,1]$ 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 $0 \le b \le 1$
  $$P(X \le b) = b$$
- then, for $b = F(x)$
  $$P(X \le F(x)) = F(x)$$
- and then by inverting
  $$P(F^{-1}(X) \le x) = F(x)$$
- 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