Week 6: Random number generators, exponential distribution

Armenak Petrosyan

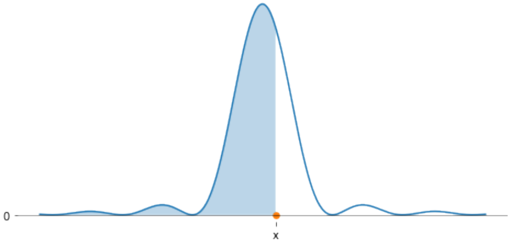
Last time

Definition (Continuous random variable, pdf)

X is called a **continuous random variable** if there exists a function $f(x)$ such that

$$F(x) = \int_{-\infty}^{x} f(y)\,\mathrm{d}y.$$

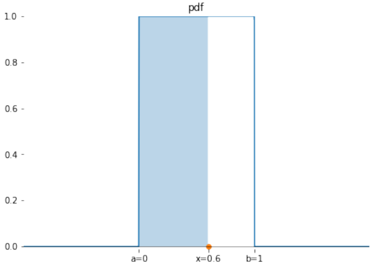
$f(x)$ is called the **probability density function** or **pdf** of continuous random variable X.

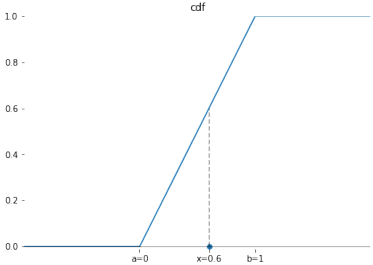


Uniform distribution (reminder)

$$f(x) = F'(x) = \begin{cases} 0 & x < a \\ \frac{1}{b-a} & a \leq x \leq b \\ 0 & x > b \end{cases}$$

$$F(x) = \begin{cases} 0 & x < a \\ \frac{x-a}{b-a} & a \leq x \leq b \\ 1 & x \geq b \end{cases}$$





Exercise 1

Problem (3.1-1 in the textbook)

Compute the variance and mgf of the uniform distribution on $[a, b]$

Solution

- $E[X^2] = \int_a^b x^2 \frac{1}{b-a}\,\mathrm{d}x = \left.\frac{1}{b-a}\frac{x^3}{3}\right|_a^b = \frac{a^2+ab+b^2}{3}.$
- $\mathrm{Var}(X) = E[X^2] - E[X]^2 = \frac{a^2+ab+b^2}{3} - \left(\frac{a+b}{2}\right)^2 = \frac{(b-a)^2}{12}.$
- $M(t) = \int_a^b e^{tx}\frac{1}{b-a}\,\mathrm{d}x = \left.\frac{e^{tx}}{t(b-a)}\right|_a^b = \frac{e^{tb}-e^{ta}}{t(b-a)}$ *when* $t \neq 0.$
- $M(0) = \int_a^b \frac{1}{b-a}\,\mathrm{d}x = 1.$

Percentiles

Definition (Median)

Median of the continuous random variable X is a number $m \in \mathbb{R}$ such that

$$P(X \leq m) = 0.5.$$

Definition (Percentiles)

Let $0 \leq p \leq 1$. The $100 \cdot p$-**th percentile** of the X is called a number π_p such that

$$P(X \leq \pi_p) = p.$$

$\pi_{0.25}, \pi_{0.5}, \pi_{0.75}$ are called **first, second and third quartiles**.

Problem (3.1-14 from the textbook)

Let $f(x) = 1/2$, $0 < x < 1$ *or* $2 < x < 3$, *zero elsewhere, be the pdf of* X

a *Sketch the graph of this pdf.*

b *Define the cdf of X and sketch its graph.*

c *Find* $q_1 = \pi_{0.25}$.

d *Find* $m = \pi_{0.5}$. *Is it unique?*

e *Find* $q_3 = \pi_{0.75}$.

Solution

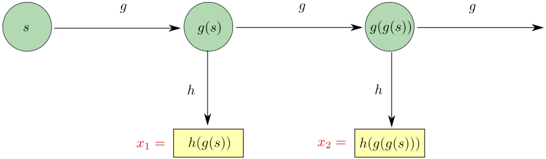
Thursday class

Generating random numbers

1. **True random number generator (TRNG):** uses physical process (e.g. atmospheric noise).
2. **Pseudo-random number generator (PRNG):** it starts with a value s (called seed) and applies operations to generate a sequence of numbers that has the following form

$$s,\ h(g(s)),\ h(g(g(s))),\ \ldots$$

The seed s is chosen "randomly" (usually uses system time).



- **Linear Congruential Generator** PRNG generates random numbers in $[0, 1)$

$$s_0 = s, \quad s_{n+1} = as_n + b$$

$$\boxed{x_n = s_n \mod 1}$$

- The Python standard library uses **Mersenne's Twister** PRNG algorithm.

Random sampling from a discrete distribution

- We want to sample from a discrete distribution that has cdf $F(x)$.
- Sample a value y from uniform distribution on $[0, 1)$.
- Let x_i be the largest value for which $F(x_i) \leq y$.

Theorem

Let $F(x)$ be a continuous cdf and let Y be a uniformly distributed random variable on $[0,1]$. Let $F^{-1}(y)$ be the inverse of F. Then the random variable $X = F^{-1}(Y)$ is distributed as F:

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

Inverse transform method: (based on above theorem)

- We want to sample from a continuous random variable with cdf $F(x)$.
- Sample a value y from uniform distribution on $[0,1)$.
- Let x be the largest value for which $F(x) \leq y$ or, equivalently, $x = F^{-1}(y)$.

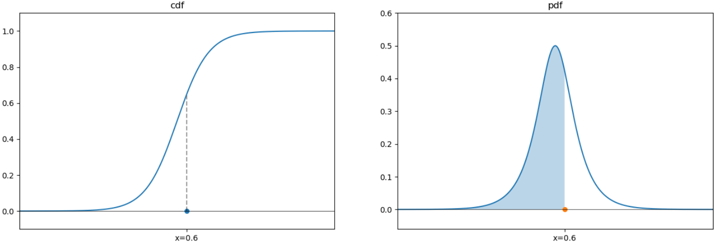
Rejection method: (uses area argument)

- We want to sample from a random variable with pdf $f(x)$.
- Let $[a, b)$ be large enough to contain most of histogram of $f(x)$.
- Sample a value (x, y) from uniform distribution on $[0, 1) \times [a, b)$.
- If $y \leq f(x)$, let X be a sample of X. Otherwise, reject it and try another point (x, y).

Example

Logistic distribution

$$\boxed{F(x) = \frac{e^x}{1+e^x}} \quad \Rightarrow \quad \boxed{f(x) = \frac{e^x}{(1+e^x)^2}}$$



- $y = \frac{e^x}{1+e^x} \Rightarrow e^x = \frac{1-y}{y}$.
- $F^{-1}(y) = \log(1-y) - \log(y)$.

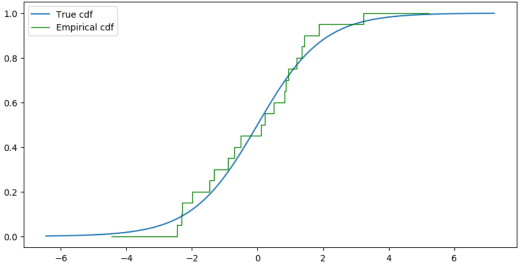
```python
import numpy as np

def logistic_inv(y):
    return np.log(1-y)-np.log(y)

#  random sample from uniform distribution on [0,1)
y = np.random.sample((20,))

# random sample from new distribution
x = logistic_inv(y)
```

$$\boxed{F_{\text{emp}}(x) = \frac{\text{number of values } \leq x}{\text{size of data}}}$$ **(Empirical cdf)**



Definition (Poisson process)

Suppose we have an action that happens regularly over time (bus arrival, visits of a website by users, etc) that satisfies the following properties:

1. The occurrence of the action at any time interval in the future is independent of past occurrences.
2. The average number of actions happening in a given time interval is proportional to the interval length.

Denote by λ the average number of times the action happens in the unit time interval $[0, 1]$. Then we call this type of action a **Poisson process with rate** λ.

- Poisson process is the continuous version of Bernoulli trials.
- If we denote by Y the number of actions in a Poisson process that happened in a certain fixed interval, $[a, b]$, then Y will have a Poisson distribution with parameter $\lambda \cdot (b - a)$.

Exponential distribution: motivation

Let X be the time when the action happens for the first time in a Poisson process with rate λ.

- X is called **waiting time** (it is the time we need to wait for the action to happen).
- X is a continuous random variable as we see below.
- $\mathrm{Range}(X) = [0, \infty)$.
- $F(X \leq x) = 0$ when $x < 0$.
- For $x \geq 0$,
$$F(x) = P(X \leq x) = 1 - P(X > x).$$
- $P(X > x)$ is the probability that there are no actions in the interval $[0, x]$. Since the number of actions in the interval $[0, x]$ is a Poisson random variable with parameter $\lambda \cdot x$,
$$P(X > x) = e^{-\lambda x}\frac{(\lambda x)^0}{0!} = e^{-\lambda x}.$$
- Hence,
$$F(x) = 1 - e^{-\lambda x}.$$
- For the pdf of X, we have
$$f(x) = F'(x) = \begin{cases} 0 & x < 0 \\ \lambda e^{-\lambda x} & x \geq 0 \end{cases}.$$

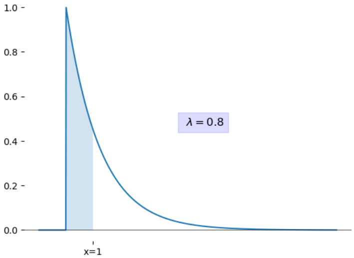
Exponential distribution: definition

Definition (Exponential distribution)

We say that a continuous random variable X has **exponential distribution** with parameter θ if its pdf is given by

$$f(x) = \begin{cases} 0 & x < 0 \\ \frac{1}{\theta} e^{-\frac{x}{\theta}} & x \geq 0 \end{cases}.$$

- We prefer $\theta = \frac{1}{\lambda}$ parametrization because it represents the average waiting time as will be apparent soon.
- Exponential distribution is the continuous version of the geometric distribution.



Mean and variance of exponential distribution

- From definition
$$M(t) = \int_0^\infty e^{tx}\frac{1}{\theta}e^{-\frac{x}{\theta}}\,\mathrm{d}x = \frac{1}{\theta}\int_0^\infty e^{x(t-\frac{1}{\theta})}\,\mathrm{d}x = \frac{1}{\theta}\cdot\frac{1}{t-\frac{1}{\theta}}\cdot e^{x(t-\frac{1}{\theta})}\Big|_{x=0}^{\infty} = \frac{1}{1-\theta t} \text{ when } t < \frac{1}{\theta}.$$
- Hence
$$M'(t) = \frac{\theta}{(1-\theta t)^2}, \quad M''(t) = \frac{2\theta^2}{(1-\theta t)^3}.$$
- Therefore
$$E[X] = M'(0) = \theta$$
$$\mathrm{Var}(X) = E[X^2] - E[X]^2 = M''(0) - E[X]^2 = \theta^2.$$