MA 6.101
Probability and Statistics

Tejas Bodas

Assistant Professor, IIIT Hyderabad

Agenda for the next two lectures

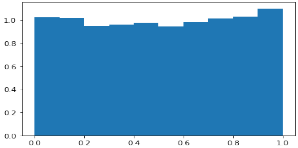
- Intro to Stochastic Simulation
  - We will generate samples from discrete or continuous r.v's using samples from uniform distribution.
- Limit theorems for Convergence of random variables
  - Sure convergence
  - Almost sure convergence & SLLN
  - Convergence in probability
  - Convergence in r^{th} mean
  - Weak Convergence or Convergence in distribution & CLT

Generate samples using uniform distribution

Our aim: Obtain samples from a discrete random variable

- Suppose you have access to samples from a uniform random variable U over support $[0, 1]$.

- ```
  import numpy as np
  import matplotlib.pyplot as plt
  uni_samples = np.random.uniform(0, 1, 5000)
  plt.hist(uni_samples, bins = 10, density = True)
  plt.show()
  ```



- *uni_samples* is a vector of 5000 realizations of uniform random variable U.

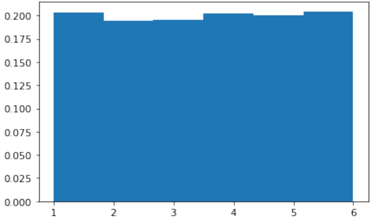
- You can also see it as a realization of $U_1, U_2, \ldots U_{5000}$ i.i.d uniform variables.
  ```

How to simulate a dice using these samples?

- Can you use these 5000 samples and convert them into outcomes of a dice ?

```
 t=0
dice_samples=np.zeros(5000)
for u in uni_samples:
  if u < 1/6:
    dice_sample = 1
  if 1/6 < u < 2/6:
    dice_sample = 2
  if 2/6 < u < 3/6:
    dice_sample = 3
  if 3/6 < u < 4/6:
    dice_sample = 4
  if 4/6 < u < 5/6:
    dice_sample = 5
  if 5/6 < u < 6/6:
    dice_sample = 6
  dice_samples[t] = dice_sample
  t = t+1
plt.hist(dice_samples, bins = 6, density = True)
```

- [0.02, 0.8, 0.6, 0.03]
- [1, 5, 4, 1]



Our aim: Obtain samples from a discrete random variable

- Consider a discrete random variable X with support set $\{x_0, x_1, \ldots\}$ and pmf $p_X(x_j) = p_j$ for $j = 0, 1, \ldots$ such that $\sum_j p_j = 1$.
- Cardinality of the support set of X could be finite or infinite.
- Our aim: Create i.i.d. samples of r.v. X using i.i.d. random samples of U.
- We shall now formally see the inverse transform method to do this.

The inverse transform method

- Aim: We wish to create i.i.d. samples of a discrete r.v. X with $p_X(x_j) = p_j$ using i.i.d. samples of a uniform r.v. U over $[0, 1]$.

- Let $u \in [0, 1]$ be a realization of r.v. U. Then the corresponding sample of X is generated as follows

$$X = \begin{cases} x_0 \text{ if } u < p_0 \\ x_1 \text{ if } p_0 \leq u < p_0 + p_1 \\ x_2 \text{ if } p_0 + p_1 \leq u < p_0 + p_1 + p_2 \\ \vdots \\ x_j \text{ if } \sum_{i=0}^{j-1} p_i \leq u < \sum_{i=0}^{j} p_i \\ \vdots \end{cases}$$

- Why is this method correct?

The inverse transform method

- A sample of X is generated using the sample of U as follows

$$X = x_j \quad \text{if} \quad \sum_{i=0}^{j-1} p_i \leq U < \sum_{i=0}^{j} p_i$$

- Why the name "inverse transform method"?
  - Recall that $\{x_0, x_1, x_2, \ldots\}$ is the support set of X and without loss of generality (WLOG), suppose $x_0 < x_1 < x_2 < \ldots$.
  - Let $F_X(x) := \mathbb{P}[X \leq x]$ denote the cdf of X and thus we have

$$F_X(x_j) = \sum_{i=0}^{k} p_i$$

  - This implies that

$$X = x_j \quad \text{if} \quad F_X(x_{j-1}) \leq U < F_X(x_j) \text{ (implying } p_X(x_j) = p_j)$$

  - After generating a random number U, we determine the value of X by finding the interval $\left[F_X(x_{j-1}), F_X(x_j)\right)$ in which u lies.
  - **We are thus finding the inverse of $F_X(U)$!**

How to generate samples of a continuous random variable

(Using samples of a continuous uniform variable over $[0, 1]$)

Our aim: Obtain samples from a continuous random variable

- Suppose you have access to samples from a uniform random variable U over support $[0, 1]$.(We will not study how to generate such samples.)

- Consider a continuous random variable X with support set $\mathcal{X}$ and let $F_X(x)$ denotes its cdf.

- Support set of X could be arbitrary.

- Our aim: Create i.i.d. samples of r.v. X using i.i.d. random samples of U.

- We shall again see the inverse transform method to do this.

Sampling from continuous random variables

Lemma

Let U be uniform random variable over $[0,1]$. Consider any continuous cdf $F(.)$. Consider a random variable X defined as follows

$$X := F^{-1}(U)$$

Then the cdf of X is $F(.)$.

Proof:

- Let $F_X(x)$ be the cdf of X, i.e., $F_X(x) := \mathbb{P}[X \leq x]$. Then

$$F_X(x) = \mathbb{P}[F^{-1}(U) \leq x]$$

$$= \mathbb{P}[U \leq F(x)]$$

$$= F(x)$$

Sampling from continuous random variables

Lemma

Let U be uniform random variable over $[0, 1]$. Consider any continuous cdf $F(.)$. Consider a random variable X defined as follows

$$X := F^{-1}(U)$$

Then the cdf of X is $F(.)$.

- Using this lemma, how to generate samples of a continuous random variable X using samples of uniform random variable U?
- **Answer:** Draw $u \sim U$ and obtain $F^{-1}(u)$. This is a sample from X.
- Do you observe anything "special" about this lemma?

Application in data analysis

- Lemma: $X = F^{-1}(U)$ has distribution $F(.)$.
- What will be cdf of a random variable $Y = F(X)$? **Uniform!**
- A conseqence of this lemma is that $F(X)$ is a uniform distribution.
- This property is known as "probability integral transform or universality of uniform".
- This property is used to test whether a set of observations can be modelled as arising from a specified distribution $G(.)$ or not.
  - Given set of data samples $s_1, s_2, \ldots, s_n$, plot $G(s_i)$ for different samples.
  - If these points are spread uniformly over the interval $[0, 1]$ then it indicates that the samples are indeed from $G(.)$.

Stochastic Simulation

- This was a brief introduction to this area of stochastic simulation.
- Refer the book Simulation by Sheldon Ross!
- Some popular techniques in simulation are:
-
  - The inverse transform method
  - Accept-Reject method (rejection sampling)
  - Importance sampling
  - Markov Chain Monte Carlo (MCMC) methods
    - Hasting-Metropolis algorithm
    - Gibbs sampling
    - Slice sampling