Lecture Notes: Continuous random variables

When we consider a discrete uniform distribution with a finite number of outcomes, each outcome has an equal probability of occurrence, say $1/n$, where n is the total number of outcomes. As we increase the number of outcomes n, the probability of each outcome decreases, but the distribution begins to resemble a continuous uniform distribution over the interval.

So we know how to generate random numbers in $\mathbb{N}$ from an uniform distribution in $\mathbb{N}$, but in many applications, from statistical simulations to machine learning, we often require random numbers in the interval $[0, 1]$. These numbers are particularly useful due to their easy transformation into other numerical ranges and their direct association with probabilities in probabilistic experiments or simulations.

Let's take the example of the Linear Congruential Generator (LCG) introduced in the previous section. Recall that an LCG generates a sequence of random numbers between 0 and $m-1$. To get a random number between 0 and 1, we simply need to divide the output of the LCG by m. This operation maps the range from 0 to $m-1$ to the range from 0 to 1 (excluding 1), making it suitable for the applications we discussed.

Mathematically, this can be represented as:

$$R_{n+1} = \frac{X_{n+1}}{m} \tag{1}$$

where X_{n+1} is the output of the LCG and R_{n+1} is the corresponding random number between 0 and 1.

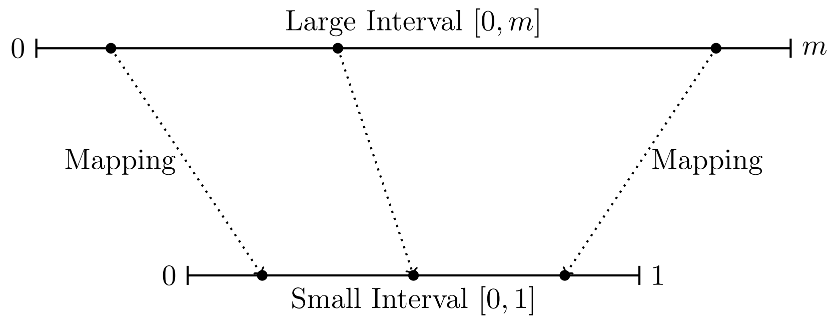

Figure 1: Mapping of random numbers from a large interval to a small interval.

Imagine we have a discrete uniform distribution where the probability for each of the n outcomes is $1/n$. This means the 'width' of each outcome, or the range of values it represents, is $1/n$. The 'height' of each outcome, or its probability, is also $1/n$. Therefore, the area of each 'bar' in the histogram (representing the product of the width and height) is $(1/n) * (1/n) = 1/n^2$.

As n approaches infinity ($n \to \infty$), the width and the height of each outcome tend to zero. However, the product of the width and height, which represents the probability density of each outcome, remains constant at $1/n^2$.

$$\lim_{n\to\infty} \left(\frac{1}{n} \cdot \frac{1}{n}\right) = \frac{1}{n^2} = 0 \tag{2}$$

The above equation signifies that as we increase n, each individual outcome's probability density goes to zero. However, the integral (or the sum in the discrete case) of the density over any interval of length $1/n$ remains constant at $1/n$,

$$\int_a^{a+1/n} \frac{1}{n} dt = \frac{1}{n} \tag{3}$$

Where a is any point in the interval $[0, 1]$. The above equation signifies that the probability of observing an outcome within any subinterval of equal length in the continuous uniform distribution remains constant, irrespective of the location of the subinterval within the [0,1] interval.

This mathematical reasoning effectively demonstrates how we transition from a discrete uniform distribution to a continuous uniform distribution as the number of outcomes becomes infinitely large. Each individual point in the continuous distribution has a probability of zero, but the density or the probability per unit length is constant across the interval, embodying the essence of a continuous uniform distribution.

Definition 1. *Probability Density Function (pdf) The probability density function (pdf) of a continuous random variable X provides a measure for the likelihood of X being near a particular value such that:*

1. *For every x in the real numbers, $0 \leq f_X(x)$, and there may exist some x where $f_X(x) > 1$.*
2. *The integral of the pdf over the entire real line is 1, i.e., $\int_{-\infty}^{\infty} f_X(x)\,dx = 1$.*

Mathematically, the pdf of X is represented as a function $f_X : \mathbb{R} \to [0, \infty)$, defined such that the probability that X lies in the interval $[a, b]$ is

$$\int_a^b f_X(x)\,dx$$

for any real numbers a and b with $a \leq b$.

Example 1 (Continuous Uniform distribution)**.** *A continuous random variable X follows a uniform distribution over the interval $[a, b]$ if it has the same probability density for all values within this interval. The probability density function (PDF) of a uniformly distributed random variable X is given by:*

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{for } a \leq x \leq b, \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

This means that for any two subintervals of equal length within the interval $[a, b]$, the probability that X falls into either subinterval is the same.

Now, let's consider the case when we want to generate realizations of non-uniformly distributed random variable X with a sample space $S \subseteq \mathbb{N}$, with a probability mass function $f_X(x)$.

To do that we first consider the definition of another crucial concept in probability, the cumulative distribution.

Thus, like in the previous chapter we can define the cumulative distribution function

Definition 2 (Cumulative Distribution Function (CDF) for Continuous Random Variables)**.** *Let X be a continuous random variable with a probability density function (pdf) $f_X(x)$. The cumulative distribution function (CDF) of X is defined for every number x by the following integral:*

$$F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f_X(t)\,dt$$

where $F_X(x)$ represents the probability that the random variable X takes on a value less than or equal to x.

Example 2 (CDF of Continuous Uniform Distribution)**.** *Let's continue with our example of a continuous random variable X that follows a uniform distribution over the interval $[a, b]$. We previously defined its probability density function (PDF) as:*

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{for } a \leq x \leq b, \\ 0 & \text{otherwise.} \end{cases}$$

Now, we'll find the Cumulative Distribution Function (CDF) of X, denoted $F_X(x)$. The CDF is defined as the integral of the PDF from the lower bound of the distribution up to x:

$$F_X(x) = \int_{-\infty}^{x} f(t)\, dt$$

After calculating, we obtain:

$$F_X(x) = \begin{cases} 0 & \text{for } x < a, \\ \frac{x-a}{b-a} & \text{for } a \leq x \leq b, \\ 1 & \text{for } x > b. \end{cases}$$

This CDF tells us the probability that the random variable X will take a value less than or equal to x. Note that $F_X(x)$ increases linearly from 0 to 1 as x ranges from a to b, consistent with our intuition of the uniform distribution.

Example 3 (Exponential Distribution)**.** *A continuous random variable X follows an Exponential distribution with a rate parameter $\lambda > 0$ if its probability density function (PDF) is given by:*

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{for } x \geq 0, \\ 0 & \text{for } x < 0. \end{cases}$$

Verifying it's a PDF: *First, let's verify that $f(x)$ is a valid PDF. For it to be a PDF, the integral over its entire domain must be equal to 1.*

$$\int_{-\infty}^{\infty} f(x)\, dx = \int_{0}^{\infty} \lambda e^{-\lambda x}\, dx$$

After integrating, we find that:

$$\int_{0}^{\infty} \lambda e^{-\lambda x}\, dx = -e^{-\lambda x}\Big|_{0}^{\infty} = 0 - (-1) = 1$$

Calculating the CDF: *Now, let's find the Cumulative Distribution Function (CDF) of X, denoted $F_X(x)$. The CDF is defined as the integral of the PDF from the lower bound of the distribution up to x:*

$$F_X(x) = \int_{-\infty}^{x} f(t)\, dt$$

After calculating, we obtain:

$$F_X(x) = \begin{cases} 0 & \text{for } x < 0, \\ 1 - e^{-\lambda x} & \text{for } x \geq 0. \end{cases}$$

This CDF gives us the probability that the random variable X will take a value less than or equal to x. As x approaches infinity, $F_X(x)$ approaches 1, consistent with our intuition of the Exponential distribution.

Expectation and Variance

Before delving into the Central Limit Theorem, it's crucial to have a firm grasp on two statistical fundamentals—Expectation and Variance. These are more than just average values or measures of spread; they serve as key mathematical pillars that support much of probability theory and statistics.

Expectation

The expectation of a random variable X, often denoted $E[X]$, intuitively represents its average or "expected" value. This can be understood as the weighted average of all possible outcomes, where the weight corresponds to the probability of each outcome.

For a discrete random variable, the mathematical expression for expectation is:

$$E[X] = \sum_x x \cdot P(X = x)$$

For a continuous random variable, the integral replaces the sum:

$$E[X] = \int x f(x)\, dx$$

Expectation serves as the center of mass of the distribution. To see this, consider that each term $x \cdot P(X = x)$ in the discrete formula represents the "contribution" of that specific outcome x to the overall average, scaled by its likelihood $P(X = x)$.

1.1.1 Example with Uniform Distribution

Consider a fair six-sided die. Each face x appears with a probability of $\frac{1}{6}$. The expectation can be calculated as:

$$\begin{aligned} E[X] &= \frac{1}{6} \times 1 + \frac{1}{6} \times 2 + \frac{1}{6} \times 3 + \frac{1}{6} \times 4 + \frac{1}{6} \times 5 + \frac{1}{6} \times 6 \\ &= \frac{1}{6}(1 + 2 + 3 + 4 + 5 + 6) \\ &= 3.5 \end{aligned}$$

The number 3.5 serves as the "balancing point" of the distribution.

Variance

The variance, denoted by $\mathrm{Var}[X]$, measures how much a distribution deviates from its mean. The concept can be understood as the average of the squared differences from the mean $E[X]$.

For a discrete random variable:

$$\mathrm{Var}[X] = \sum_x (x - E[X])^2 \cdot P(X = x)$$

For a continuous random variable:

$$\mathrm{Var}[X] = \int (x - E[X])^2 f(x)\, dx$$

The squared term $(x - E[X])^2$ captures the 'distance' of each outcome from the mean, and squaring ensures that all distances are non-negative. The probability $P(X = x)$ or $f(x)\, dx$ then gives the 'weight' to each squared distance, similar to how it worked in the case of expectation.

1.2.1 Example with Uniform Distribution

For a fair six-sided die, using $E[X] = 3.5$ as calculated earlier:

$$\begin{aligned}
\mathrm{Var}[X] &= \frac{1}{6}((1-3.5)^2 + (2-3.5)^2 + (3-3.5)^2 + (4-3.5)^2 + (5-3.5)^2 + (6-3.5)^2) \\
&= \frac{1}{6}(6.25 + 2.25 + 0.25 + 0.25 + 2.25 + 6.25) \\
&= 2.92
\end{aligned}$$

Here, $\mathrm{Var}[X] = 2.92$ encapsulates the average "spread" around the mean $E[X] = 3.5$.

1.2.2 Example with Exponential Distribution

The Exponential distribution is often used to model the time between events in a Poisson process. It is characterized by its rate parameter λ, and its probability density function (PDF) is:

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

To find the expectation $E[X]$, we compute:

$$\begin{aligned}
E[X] &= \int_0^\infty x\lambda e^{-\lambda x}\, dx \\
&= \lambda \int_0^\infty x e^{-\lambda x}\, dx
\end{aligned}$$

We can solve this integral by parts, letting $u = x$ and $dv = e^{-\lambda x}dx$. Then $du = dx$ and $v = -\frac{1}{\lambda}e^{-\lambda x}$.

$$\begin{aligned}
E[X] &= -\frac{x}{\lambda}e^{-\lambda x}\Big|_0^\infty + \frac{1}{\lambda}\int_0^\infty e^{-\lambda x}\, dx \\
&= 0 + \frac{1}{\lambda}\left[-\frac{1}{\lambda}e^{-\lambda x}\right]\Big|_0^\infty \\
&= \frac{1}{\lambda}
\end{aligned}$$

Next, we find the variance $\mathrm{Var}[X]$:

$$\begin{aligned}
\mathrm{Var}[X] &= \int_0^\infty (x - E[X])^2 \lambda e^{-\lambda x}\, dx \\
&= \int_0^\infty (x - \frac{1}{\lambda})^2 \lambda e^{-\lambda x}\, dx \\
&= \lambda \int_0^\infty x^2 e^{-\lambda x}\, dx - 2\int_0^\infty x e^{-\lambda x}\, dx + \frac{1}{\lambda^2}\int_0^\infty e^{-\lambda x}\, dx
\end{aligned}$$

To find the variance $\mathrm{Var}[X]$, we compute:

$$\begin{aligned}
\mathrm{Var}[X] &= E[X^2] - (E[X])^2 \\
&= \int_0^\infty x^2 \lambda e^{-\lambda x}\, dx - \left(\frac{1}{\lambda}\right)^2
\end{aligned}$$

Firstly, let's find $E[X^2]$, which is the first term in the variance equation.

$$\begin{aligned} E[X^2] &= \int_0^\infty x^2 \lambda e^{-\lambda x}\, dx \\ &= \lambda \int_0^\infty x^2 e^{-\lambda x}\, dx \end{aligned}$$

To solve this integral, we'll use integration by parts twice. We start by setting $u = x^2$ and $dv = e^{-\lambda x} dx$. Then $du = 2x dx$ and $v = -\frac{1}{\lambda} e^{-\lambda x}$.

$$\begin{aligned} E[X^2] &= -\frac{x^2}{\lambda} e^{-\lambda x}\Big|_0^\infty + \frac{1}{\lambda}\int_0^\infty 2x e^{-\lambda x}\, dx \\ &= 0 + \frac{2}{\lambda}\int_0^\infty x e^{-\lambda x}\, dx \\ &= 0 + \frac{2}{\lambda}\left(\frac{1}{\lambda}\right) \\ &= \frac{2}{\lambda^2} \end{aligned}$$

Finally, we plug this back into the variance formula to get:

$$\mathrm{Var}[X] = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}$$

Thus, we have proven that $\mathrm{Var}[X] = \frac{1}{\lambda^2}$ for an Exponential distribution, completing the mathematical reasoning for this example.

Notice that as λ increases, both $E[X]$ and $\mathrm{Var}[X]$ decrease, making the distribution more concentrated around its mean.

Remark 1. *The metrics of expectation and variance serve as the conceptual backbone for the subsequent discussion on the Central Limit Theorem. Their mathematical properties will play a pivotal role in comprehending why averages of random variables tend to follow a normal distribution.*

The Central Limit Theorem

Imagine a gambler tossing dice at a casino. The outcomes of individual dice rolls are random, but what about the average outcome over a series of rolls? Interestingly, this average tends to behave in a very predictable manner as the number of dice increases, thanks to a fundamental result in probability theory known as the Central Limit Theorem (CLT).

Dice and Averages: An Illustrative Example

In previous discussions, we delved into how the average outcome of multiple dice rolls doesn't follow a uniform distribution. For example, when you roll 10 dice, the sum of 35 is much more likely to occur than a sum of 10. We'll now extend this experiment to understand how these averages behave as the number of dice rolls increases.

Experiment 1 (Distribution of Dice Roll Averages)**.** ***Objective:*** *To investigate how the distribution of the averages of dice rolls changes as the number of dice rolls increases.*

Procedure:

- *Set* $N = 10,000$ *trials for the experiment.*

- *Let* $M = \{1, 2, 5, 10, 100, 1000\}$ *represent different dice counts.*
- *For each* $m \in M$*:*
  1. *Initialize an array for storing averages.*
  2. *For* $i = 1, \ldots, N$*:*
     - *Roll* m *dice.*
     - *Compute and store the average.*
- *Plot the distributions for each* m*.*

Observation: *The distribution narrows and assumes a symmetric shape as* m *increases.*

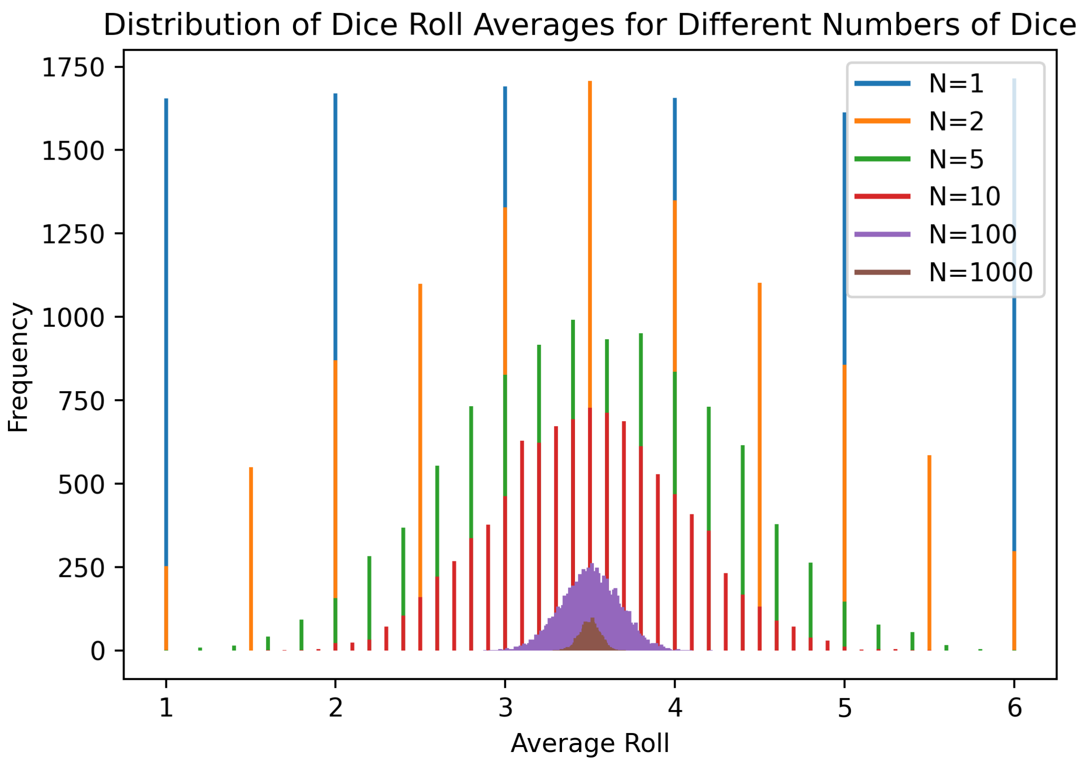

Figure 2: Frequency plots showing the sampling distribution of the mean from rolling a die. The histograms were created for different number of dices, ranging from 1 to 1000, and was done 10000 times for each.

The Central Limit Theorem (CLT) mathematically formalizes the phenomenon we observed in the dice experiment.

Theorem 1 (Central Limit Theorem)**.** *Let* $X_1, X_2, \ldots, X_n$ *be i.i.d. random variables with expected value* $E[X] = \mu$ *and variance* $Var[X] = \sigma^2$*. The standard deviation is* $\sigma = \sqrt{\sigma^2}$*. Define the sample mean* $\overline{X}$ *as*

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Then, as n *approaches infinity, the distribution of the sample mean* $\overline{X}$ *approaches a normal distribution with mean* μ *and variance* $\frac{\sigma^2}{n}$*, formally*

$$\overline{X} \xrightarrow{d} \mathcal{N}(\mu, \frac{\sigma^2}{n})$$

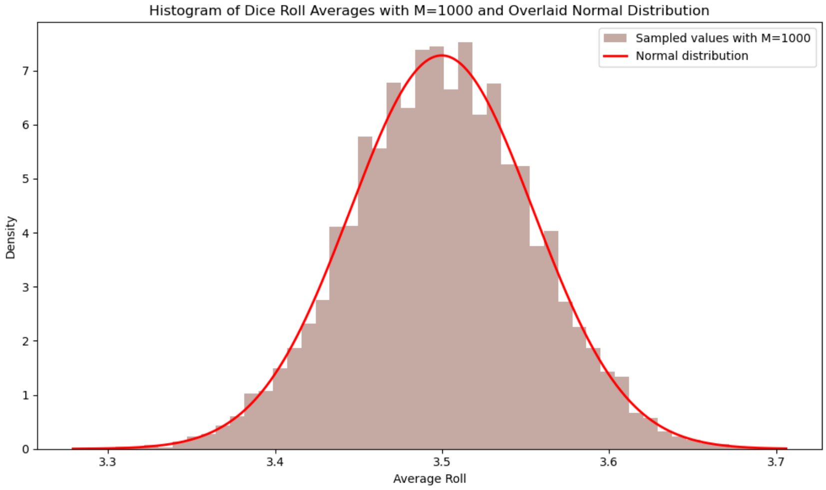

Figure 3: Histogram of Dice Roll Averages with $M = 1$ and Overlaid Normal Distribution

where $\xrightarrow{d}$ *denotes convergence in distribution. The PDF of* $\overline{X}$ *will then be:*

$$P(\overline{X} = x) \approx \frac{1}{\sqrt{2\pi \frac{\sigma^2}{n}}} e^{-\frac{(x-\mu)^2}{2\frac{\sigma^2}{n}}}$$

This is equivalent to $Z \sim N(0, 1)$*, where* Z *is defined as*

$$Z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

and its PDF is:

$$P(Z = z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

Experiment 2 (Conformance to Normal Distribution)**.** ***Objective:*** *To verify that the average of a large number of dice rolls approximates a normal distribution, in line with the Central Limit Theorem.*

Procedure:

1. *Use the averages obtained from the previous experiment where* $M = 1000$*.*
2. *Calculate* μ *and* σ *of these averages.*
3. *Plot a histogram of the averages and overlay it with a normal distribution curve parameterized by* μ *and* σ*.*

Insight: *The histogram closely matches the overlaid normal distribution, substantiating the Central Limit Theorem and reinforcing its applicability to gambling scenarios.*

Further Considerations: *Given that* $\mu = 3.5$ *and* $\sigma^2 = 2.92$ *for this dice game, one can transform the sample means into Z-scores using*

$$Z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Plotting these Z-scores should result in a distribution that approximates a standard normal distribution $P(Z = z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$*.*

Monte Carlo Integration

Monte Carlo methods are a broad class of algorithms that rely on random sampling to obtain numerical results. One of the most classic applications of these methods is *Monte Carlo integration*.

Imagine you wish to find the area under a curve defined by a function $f(x)$ over an interval $[a, b]$. Traditional numerical integration methods, like the trapezoidal rule or Simpson's rule, divide the interval into small sub-intervals and approximate the area under the curve using geometric shapes. In contrast, Monte Carlo integration estimates this integral by taking random samples.

The process is as follows:

1. Define a rectangle that contains the region you wish to integrate. This rectangle should span from a to b in the x-direction and from 0 to a value M which is greater than the maximum value of $f(x)$ on the interval in the y-direction.

2. Generate a large number of random points inside this rectangle.

3. Count the number of points that fall below the curve $f(x)$.

4. The ratio of the number of points below the curve to the total number of points, multiplied by the area of the rectangle, gives an estimate of the integral.

Mathematically, the estimate I for the integral of $f(x)$ over $[a, b]$ is:

$$I = \left(\frac{\text{Number of points below } f(x)}{\text{Total number of points}}\right) \times \text{Area of rectangle}$$

Law of Large Numbers

The expectation of a random variable offers a theoretical average or "center of mass" for its distribution. A natural question arises: if we sample repeatedly from this distribution, how closely does the sample average approximate this theoretical expectation? The Law of Large Numbers provides an answer, guaranteeing that, under specific conditions, the sample average converges to the theoretical expectation as the number of samples increases. This law underpins the intuitive notion that as we collect more independent observations, their average tends towards the true mean of the distribution.

Theorem 2 (Law of Large Numbers)**.** *Let $X_1, X_2, \dots, X_n$ be a sequence of independent and identically distributed (i.i.d.) random variables with a finite expectation denoted by $\mathbb{E}[X_i]$. Then, as n approaches infinity, the sample average converges to the expected value:*

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow[n\to\infty]{} \mathbb{E}[X_i].$$

In practical terms, the average outcome from a large number of trials will approximate the expected value, and this approximation improves with more trials.

Fundamental Theorem of Monte Carlo Integration

Building on the Law of Large Numbers, Monte Carlo Integration approximates integrals by averaging function values at randomly chosen points.

Theorem 3 (Monte Carlo Integration). *Consider a real-valued function $f(x)$ defined over a domain D. The Monte Carlo estimate for the integral $\int_D f(x)\,dx$ is:*

$$\frac{1}{N}\sum_{i=1}^{N} f(x_i),$$

where x_i are random samples drawn uniformly from D. As N approaches infinity, and under certain conditions, this estimate converges to the true value of the integral, thanks to the Law of Large Numbers.

Monte Carlo Integration can be approached using indicator random variables, especially useful when the domain of integration, D, is complex or irregularly shaped. This method leverages random sampling and probability to provide an estimate for the integral.

The methodology is:

1. **Random Sampling**: Draw a random point (u_1, u_2) uniformly from a larger domain R that encompasses D.

2. **Indicator Variable**: Define a binary random variable I as:

$$I = \begin{cases} 1 & \text{if } u_1 \leq f(u_2) \text{ and } (u_1, u_2) \in D \\ 0 & \text{otherwise} \end{cases}$$

   This variable I is 1 if the point (u_1, u_2) lies below the curve of f within D, and 0 otherwise.

3. **Compute the Proportion**: After drawing N random points, compute the proportion $\hat{p}$ of points for which $I = 1$. This proportion estimates the ratio of the area under f in D to the area of R.

4. **Estimate the Integral**: The integral of f over D is approximately $\hat{p} \times |R|$.

The expectation of the indicator random variable I is given by:

$$\mathbb{E}[I] = P((u_1, u_2) \text{ is under } f \text{ and in } D)$$

This expectation is essentially the proportion of the area under f within D relative to R:

$$\mathbb{E}[I] = \frac{\text{Area under } f \text{ in } D}{|R|}$$

The Monte Carlo estimate for this expectation, after N trials, is:

$$\hat{p} = \frac{1}{N}\sum_{i=1}^{N} I_i$$

where I_i is the value of I for the i-th random sample. Thus, the Monte Carlo estimate for the integral of f over D becomes:

$$\hat{p} \times |R|$$

By the Law of Large Numbers, as N grows larger, $\hat{p}$ converges to $\mathbb{E}[I]$, making our integral estimate increasingly accurate.

Monte Carlo Integration can be applied to various problems, one of the most illustrative being the estimation of the mathematical constant π.

Example 4 (Estimation of π using Monte Carlo). *Consider a unit circle inscribed in a unit square. If we uniformly sample random points within this square, the probability that a point lies inside the circle is equal to the ratio of the area of the circle to the area of the square. Given that the area of the unit circle is π and the area of the unit square is 1, this ratio is $\frac{\pi}{4}$.*

Let's define an indicator random variable I:

$$I = \begin{cases} 1 & \textit{if the point is inside the unit circle} \\ 0 & \textit{otherwise} \end{cases}$$

After drawing N random points in the square, the proportion $\hat{p}$ of points for which $I = 1$ approximates the ratio of the area of the circle to the square. Therefore, an estimate of π is given by:

$$\pi \approx 4 \times \hat{p}$$

This method leverages the geometric interpretation of π and the probabilistic foundations of Monte Carlo to provide an estimate. As N grows larger, the estimate becomes more accurate due to the Law of Large Numbers.

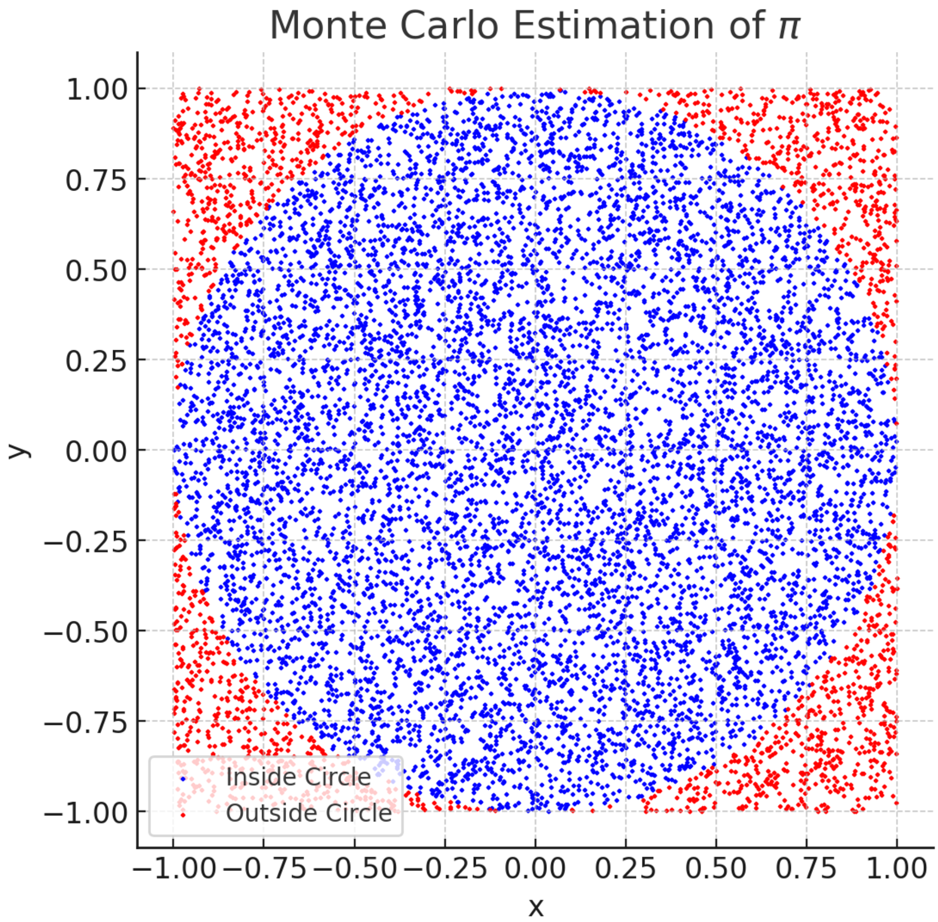

Figure 4: Monte Carlo Estimation of π. Blue dots represent random points inside the unit circle, while red dots represent points outside the circle but inside the unit square.

Example 5. *Consider the function $f(x) = x^2$ over the interval $[0, 1]$. The actual value of this integral is $\frac{1}{3}$. Using Monte Carlo integration, we can estimate this value.*

As seen in the Figure, the blue curve represents the function $f(x) = x^2$. The green dots represent the random points that fall below the curve, and the red dots represent the points

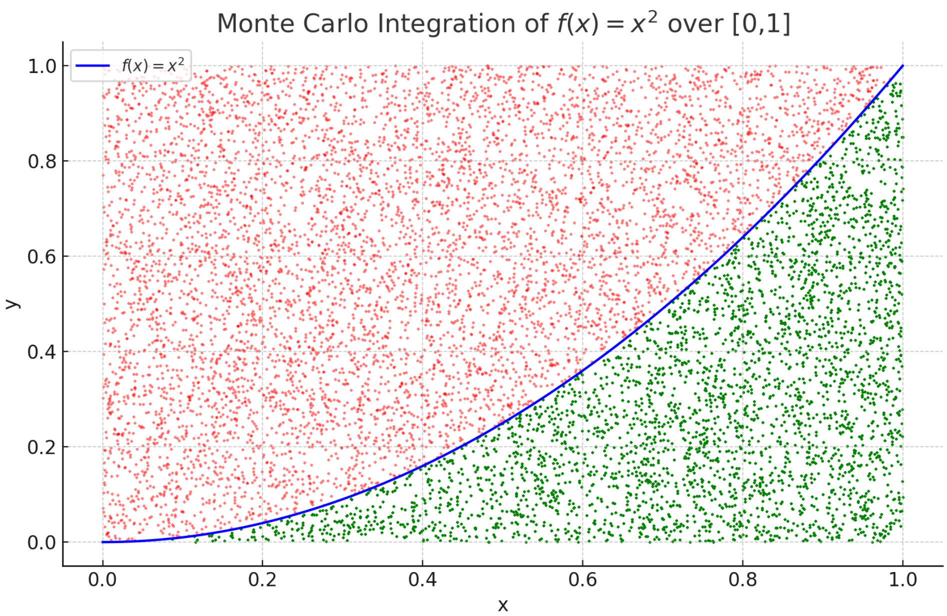

Figure 5: Monte Carlo Integration of $f(x) = x^2$ over [0,1]. Green dots represent points below the curve, and red dots represent points above the curve.

that fall above the curve. Through this method, we estimated the value of the integral to be approximately 0.3349*, which is close to the actual value of* $\frac{1}{3} \approx 0.3333$.

Generating continuos random variables

We begin our exploration of continuous random variables by examining how to generate them using specific probability density functions (PDFs). Let's first focus on the inverse transform sampling method, and then we will move on to rejection sampling.

Rejection Sampling

Theorem 4 (Rejection Sampling with Uniform Distribution)**.** *Let $f(x)$ be a non-negative and bounded probability density function (pdf) defined on a domain $\mathcal{D}$ such that $0 \leq f(x) \leq M$ for all $x \in \mathcal{D}$, where M is the known maximum of f. Then, we can sample from $f(x)$ using the following rejection sampling method:*

1. *Generate $X \sim$ Uniform$(\mathcal{D})$ and $Y \sim U(0, M)$.*
2. *If $Y \leq f(X)$, accept X as a sample from f. Otherwise, reject X and go back to step 1.*

The accepted samples X are distributed according to $f(x)$.

Intuitively, the probability that a point (X, Y) falls below the curve $y = f(x)$ is proportional to the area under f. By accepting only points (X, Y) for which $Y \leq f(X)$, we ensure that the accepted samples are distributed according to $f(x)$. The rejection step ensures that we discard points that fall above $f(x)$, thus compensating for the use of the uniform distribution over $[0, M]$ instead of f itself.

Now, equipped with these tools, we can compute the acceptance probability for the rejection sampler that uses a uniform distribution. Given:

$$\begin{aligned}
P(U < f(Y)) &= \int P(U < f(Y)|Y = y)\frac{1}{b-a}dy \\
&= \int \frac{f(x)}{M}\frac{1}{b-a}dy \\
&= \frac{1}{M(b-a)}
\end{aligned}$$

Inverse transforming sampling

Theorem 5 (Inverse Transform Sampling)**.** *Let U be a uniform random variable in the interval $[0, 1]$, and let $F_X^{-1}(u)$ be the inverse function of $F_X(x)$, the CDF of X. Then $X = F_X^{-1}(U)$ will have the PDF $f_X(x)$.*

This theorem provides us a convenient way to generate random variables that follow any given distribution, as long as we can compute its inverse CDF $F_X^{-1}(u)$.

Example 6 (Exponential Distribution)**.** *The exponential distribution is commonly used to model waiting times between events in a Poisson process. Its PDF is given by:*

$$f(x; \lambda) = \begin{cases} \lambda e^{-\lambda x} & \text{for } x \geq 0, \\ 0 & \text{otherwise.} \end{cases}$$

The CDF is then:

$$F(x; \lambda) = \begin{cases} 1 - e^{-\lambda x} & \text{for } x \geq 0, \\ 0 & \text{otherwise.} \end{cases}$$

To generate a random variable X that follows an exponential distribution with rate λ, we can first generate a uniform random variable U in $[0, 1]$, and then use the inverse of the CDF:

$$X = F^{-1}(U) = -\frac{1}{\lambda}\ln(1 - U)$$

By using the inverse transform sampling method, we can simulate any continuous random variable, provided we can find the inverse of its CDF. This method is widely used in simulation studies, statistical modeling, and machine learning applications.

Theorem 6. *If $U \sim U(0, 1)$ and $F(\cdot)$ is a valid invertible cumulative distribution function (CDF), then*

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

Proof.

$$\begin{aligned}
P(X \leq x) &= P(F^{-1}(U) \leq x) \\
&= P(U \leq F(x)) \\
&= F_U(F(x)) \\
&= F(x).
\end{aligned}$$

□

This theorem provides a straightforward way to generate numbers from an arbitrary probability distribution by simulating uniform distriuted numbers and calculating the proper transformation.

Example 7. *The exponential distribution with rate parameter $\lambda > 0$ has the probability density function (pdf):*

$$p(x; \lambda) = \lambda \exp(-\lambda x) \quad \text{for } x \geq 0. \tag{5}$$

The cumulative distribution function (CDF) of the exponential distribution is:

$$F(x; \lambda) = 1 - \exp(-\lambda x) \quad \text{for } x \geq 0. \tag{6}$$

The inverse of the CDF is:

$$F^{-1}(y; \lambda) = -\frac{1}{\lambda} \log(1 - y) \quad \text{for } 0 < y < 1. \tag{7}$$

To generate a random sample from the exponential distribution using the inverse transform method, we first generate a random sample y from the uniform distribution on the interval (0, 1), and then apply the inverse CDF to y:

$$x = F^{-1}(y; \lambda) = -\frac{1}{\lambda} \log(1 - y). \tag{8}$$

This x is a random sample from the exponential distribution with rate parameter λ.

The efficiency of this method is deterined by the aceptance probability. To calculate this probability as well as to provide a rigorous proof of the theoretical support of the rejection sampling method, it is convenient to define basic properties of probability.

Theorem 7. *Realizations of X can be generated by simulation $U \sim U(0, 1)$, and then*

$$X = min\{x | F_X(x) \geq U\}$$

Proof.

$$\begin{aligned} P(X = x) &= P_{(U)}(U \in (P_X(X \leq x - 1), P_X(X \leq x)]) \\ &= P_X(X \leq x - 1) - P_X(X \leq x) \\ &= P_X(X = x). \end{aligned}$$

□

To exemplify the previous theorem let's study the following experiment.

Experiment 3 (Poisson Sampling Experiment)**.** *Given the cumulative distribution function (CDF) of the Poisson distribution,*

$$F_X(k; \lambda) = \sum_{i=0}^{k} \frac{e^{-\lambda} \lambda^i}{i!},$$

we wanted to generate random samples using the inverse transform sampling method. The procedure was as follows:

1. *Generate a random number U from a uniform distribution $U(0, 1)$.*
2. *Compute the value X such that:*

   $$X = \min\{x | F_X(x; \lambda) \geq U\}.$$

3. *Repeat the above steps 10,000 times to generate a sufficient number of samples.*

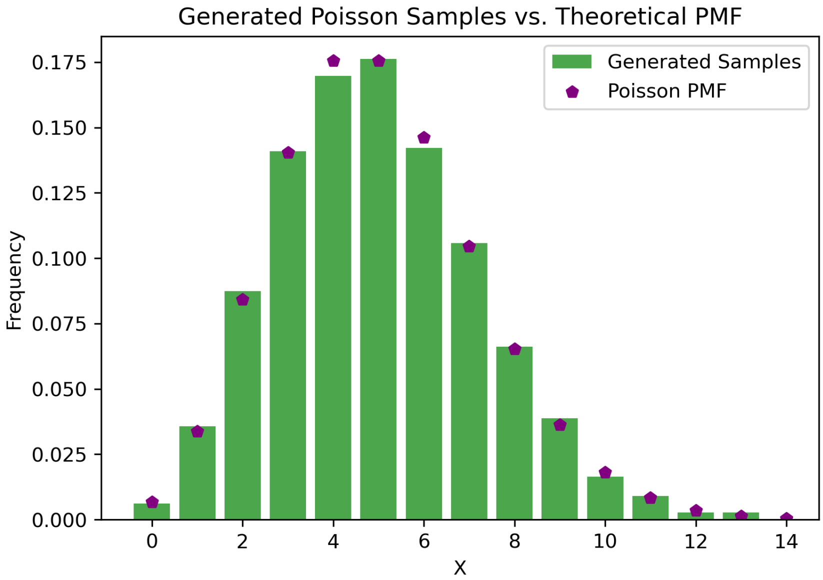

Figure 6: The result is visualized in a histogram where the blue bars represent the generated samples, and the green stems and markers represent the Poisson PMF. The generated samples closely followed the theoretical Poisson distribution, validating the sampling method.

4. *Plot a histogram of the generated samples to visualize their distribution.*

5. *On top of the histogram, overlay the probability mass function (PMF) of the Poisson distribution to compare the generated samples with the theoretical distribution:*

$$P(X = k; \lambda) = \frac{e^{-\lambda}\lambda^k}{k!}.$$

This experiment, shows very well the relationship between the frequency of an event to occur and it's probability.