Random Variables, the Law of Large Numbers, and the Central Limit Theorem.
Introduction to Machine learning
Week 4

Olivier JAYLET

September 24, 2025

Contents

1 Introduction

We all understand intuitively that the average of many measurements of the same unknown quantity tends to give a better estimate than a single measurement. Intuitively, this is because the random error of each measurement cancels out in the average. In these notes we will make this intuition precise in two ways: the law of large numbers (LoLN) and the central limit theorem (CLT). This document provides a summarized explanation of these topics for a necessary comprehension.

2 Random Variables

2.1 Definition

A **random variable** is a function that maps the outcomes of a random experiment to a numerical value. Random variables are typically denoted by capital letters, such as X, Y, or Z.

(i) **Discrete Random Variable:** A random variable that can take on a finite or countably infinite number of distinct values. For example, the number of heads in a sequence of three coin flips can only be 0, 1, 2, or 3.

(ii) **Continuous Random Variable:** A random variable that can take on any value within a given interval. For example, the height of a person or the time it takes for a reaction to occur.

Why Are Random Variables Important in Machine Learning?

(i) Random variables provide a framework to describe and analyze uncertainty in data and predictions. In machine learning, a model might predict not just a single value but a probability distribution over possible outcomes.

(ii) They allow us to define probability distributions (e.g., binomial for discrete RVs, normal for continuous RVs), which are essential for statistical inference and machine learning algorithms.

(iii) In fields like finance, engineering, and healthcare, random variables help in risk assessment, forecasting, and optimizing decisions under uncertainty.

When are random variables used?

(i) **In ML**: To model features, targets, and errors (e.g., in regression, classification, and clustering).

(ii) **Statistics**: To perform hypothesis testing, confidence interval estimation, and Bayesian inference.

(iii) **Simulation** : To generate synthetic data for testing models or understanding complex systems (e.g., Monte Carlo simulations).

(iv) **Real-World application** : From predicting weather patterns to optimizing supply chains, random variables are everywhere uncertainty exists.

2.2 Probability Distributions

The behavior of a random variable is described by its probability distribution, which specifies the probability of the variable taking on each of its possible values.

2.2.1 Discrete Probability Distribution

For a discrete random variable X, its distribution is defined by the **Probability Mass Function (PMF)**, denoted by $P(X = x)$. The PMF must satisfy two properties:

1. $0 \leq P(X = x) \leq 1$ for all values of x.
2. $\sum_x P(X = x) = 1$.

The **Cumulative Distribution Function (CDF)** for a discrete variable is:

$$F(x) = P(X \leq x) = \sum_{t \leq x} P(X = t).$$

2.2.2 Continuous Probability Distribution

For a continuous random variable X, its distribution is defined by the **Probability Density Function (PDF)**, denoted by $f(x)$. The PDF must satisfy:

1. $f(x) \geq 0$ for all x.
2. $\int_{-\infty}^{\infty} f(x)\,dx = 1$.

For continuous variables, the probability of a specific value is zero, so we consider probabilities over intervals: $P(a < X < b) = \int_a^b f(x)\,dx$.

The **CDF** for a continuous variable is:

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)\,dt.$$

2.3 Use in ML

The CDF is useful in machine learning for reshaping data, figuring out percentiles, or creating synthetic data for testing and training. When you really get how these distributions work, your ML models can do more than just guess a single number, they can give you predictions that include a sense of how confident (or uncertain) they are. That makes your models more reliable and practical.

2.4 Expected Value and Variance

Two key parameters describe the central tendency and dispersion of a random variable's distribution.

(i) **Expected Value (Mean):** The long-run average of the random variable. It is denoted by $E[X]$ or μ. The expected value (or mean) of a random variable is the "average" value you would expect to observe if you repeated an experiment infinitely many times. It represents the central tendency of the distribution.

- Discrete: $E[X] = \sum_x xP(X = x)$.

- Continuous: $E[X] = \int_{-\infty}^{\infty} x f(x)\, dx$.

Interpretation: The expected value is not necessarily a value that X can actually take (e.g., the expected number of heads in two coin flips is 1, even though you can only get 0, 1, or 2 heads). It's a theoretical average.

(ii) **Variance:** A measure of the spread of the distribution around the mean.
It is denoted by $Var(X)$ or σ^2. $Var(X) = E[(X-\mu)^2] = E[X^2] - (E[X])^2$.
The **Standard Deviation** is the square root of the variance, $\sigma = \sqrt{Var(X)}$.
A high variance means the values are widely dispersed; a low variance means they are clustered close to the mean.

(iii) **Skewness:** A measure of the asymmetry of the distribution around the mean.
It is denoted by γ_1 or $\text{Skew}(X)$.

$$\text{Skew}(X) = E\left[\left(\frac{X-\mu}{\sigma}\right)^3\right]$$

where μ is the mean and σ is the standard deviation.
A positive skewness indicates a longer or heavier right tail, while a negative skewness indicates a longer or heavier left tail. A skewness of zero suggests a symmetric distribution.

(iv) **Kurtosis:** A measure of the "tailedness" and "peakedness" of the distribution relative to a normal distribution.
It is denoted by γ_2 or $\text{Kurt}(X)$.

$$\text{Kurt}(X) = E\left[\left(\frac{X-\mu}{\sigma}\right)^4\right]$$

Excess kurtosis is often reported as $\text{Kurt}(X) - 3$, where 3 is the kurtosis of a normal distribution.
A high kurtosis indicates heavier tails and a sharper peak, while a low kurtosis indicates lighter tails and a flatter peak. A kurtosis of 3 (or excess kurtosis of 0) matches the shape of a normal distribution.

3 The Law of Large Numbers (LLN)

The LLN is a theorem that describes the result of performing the same experiment a large number of times.

3.1 Statement of the Theorem

Let $X_1, X_2, \ldots, X_n$ be a sequence of independent and identically distributed (i.i.d.) random variables with finite expected value $E[X_i] = \mu$. Let $\bar{X}_n$ be the sample mean:

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

The LLN states that as $n \to \infty$, the sample mean $\bar{X}_n$ converges to the population mean μ.

3.2 Types of LLN

(i) **Weak Law of Large Numbers (WLLN):** States that the sample mean converges in probability to the population mean.

$$\bar{X}_n \xrightarrow{p} \mu \quad \text{as } n \to \infty.$$

This means that for any small positive number $\epsilon > 0$:

$$\lim_{n \to \infty} P(|\bar{X}_n - \mu| > \epsilon) = 0.$$

The probability of the sample mean being far from the population mean approaches zero as n grows. The sample mean gets "close" to the population mean as n increases, but there's still a tiny chance it could be far off. This is called convergence in probability.

Conditions:

(a) The random variables $X1, X2, \ldots, Xn$ must be independent and identically distributed (i.i.d.).

(b) The expected value $E[X_i] = \mu$ must be finite (i.e., the mean exists and is not infinite).

(ii) **Strong Law of Large Numbers (SLLN):** States that the sample mean converges almost surely to the population mean. This is a more rigorous form of convergence, implying that the sample mean will almost certainly equal the population mean as $n \to \infty$.

$$\bar{X}_n \xrightarrow{a.s.} \mu \quad \text{as } n \to \infty.$$

i.e.

$$P\left(\lim_{n \to \infty} \bar{X}_n = \mu\right) = 1.$$

Conditions:

(a) The random variables $X1, X2, \ldots, Xn$ must be i.i.d. (just like the WLLN).

(b) The expected value $E[Xi] = \mu$ must be finite.

(c) Additional condition: The SLLN often requires that the random variables have a finite fourth moment (i.e., $E[X_i^4] < \infty$), though for many common distributions (e.g., normal, exponential), the SLLN holds as long as the mean is finite.

3.3 Practical significance

1. **Data Science and Statistics**
   When you calculate the average of a dataset (e.g., average income, average temperature), the LLN ensures that your sample average is a good estimate of the true population average, especially as your dataset grows.
   In political polling, the LLN ensures that the average preference for a candidate, calculated from a large random sample of voters, converges to the true population preference. This allows pollsters to predict election outcomes with increasing accuracy as the sample size grows.

2. **Machine Learning**
   In supervised learning, the LLN justifies why models trained on large datasets generalize better. For example, consider training a spam detection model using a dataset of labeled emails (spam or not spam).
   The model's performance is evaluated using the average classification error on a test set. According to the LLN, as the size of the test set grows, the average error rate observed on the test set will converge to the model's true expected error rate on the entire population of emails. This ensures that the model's performance metrics (e.g., accuracy, precision, recall) become more reliable and reflective of real-world performance as the test set size increases.

3. **Monte Carlo Simulations**
   Monte Carlo methods rely on the LLN to approximate complex integrals or probabilities. For example, if you're simulating the value of a financial option, the average of many random simulations will converge to the true expected value.

4. **A/B Testing**
   When comparing two versions of a product (e.g., a website layout), the LLN ensures that the average performance metrics (e.g., click-through rates) will converge to their true values as more users are included in the test.

4 The Central Limit Theorem (CLT)

While the LLN tells us that the sample mean will converge to the population mean, the CLT describes the shape of the distribution of the sample mean itself.

4.1 Central Limit Theorem (CLT): Statement and Interpretation

4.2 Statement of the Theorem

Let $X_1, X_2, \ldots, X_n$ be a sequence of i.i.d. random variables with finite expected value μ and finite variance σ^2.

As the sample size n becomes large, the distribution of the sample mean $\bar{X}_n$ will be approximately normal, regardless of the shape of the original population distribution.

More formally, the random variable $Z_n = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$ converges in distribution to the standard normal distribution $N(0,1)$ as $n \to \infty$.

$$Z_n \xrightarrow{d} N(0,1)$$

4.3 Key Properties of the Sample Mean Distribution

Based on the CLT, we can state the properties of the sampling distribution of the mean, denoted by $\bar{X}$:

1. **Mean:** The mean of the sampling distribution is equal to the population mean. $E[\bar{X}] = \mu$.

2. **Standard Deviation (Standard Error):** The standard deviation of the sampling distribution is equal to the population standard deviation divided by the square root of the sample size.
$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

This property demonstrates that as the sample size n increases, the standard error decreases, causing the distribution of sample means to become more tightly clustered around the population mean.

4.4 Intuition

1. Population Shape Doesn't Matter:
   - Even if the original population X_n is not normal (e.g., highly skewed, uniform, etc.), the distribution of $\bar{X}_n$ will approach a normal distribution due to the CLT.

2. Larger Sample Size = Better Approximation:
   - The approximation of $\bar{X}_n \sim \mathcal{N}(\mu, \sigma^2/n)$ improves as n increases.

4.5 Practical Significance

The CLT is the foundation for much of inferential statistics. It allows us to:

1. Construct **confidence intervals** for the population mean, even when the population's distribution is unknown.

2. Perform **hypothesis tests** on population means.

3. Assume normality for sample statistics when working with large datasets, simplifying statistical analysis.

In Machine learning, CLT is used to evaluate the performance of models. For example, when comparing the performance of different models, you can use the CLT to calculate the standard error of the estimate, which helps to determine if the difference in performance is statistically significant.