

# Distributions, Central Limit Theorem, and Estimation in ML

## Random Variables: Discrete vs Continuous

A **random variable** is a function that assigns a numerical value to each outcome of a random experiment.

### What is a Random Variable?

Instead of thinking about abstract outcomes like "heads" or "tails," a random variable converts them to numbers.

#### Example:

Coin flip experiment:

- Outcome: Heads
- Random variable X:  $X = 1$  (assign 1 to heads)
- Or:  $X = 0$  (assign 0 to tails)

#### Example: Rolling a die

Outcome: Rolling a 4

Random variable X = value shown = 4

**Why use random variables?** They let us apply mathematical operations (addition, multiplication, calculus) to probabilistic events.

### Discrete Random Variables

A **discrete random variable** takes on a **finite or countably infinite** set of values.

#### Characteristics:

- Specific, separated values (gaps between possible values)
- Can be listed or counted
- Probability described by **probability mass function (PMF)**

#### Examples:

- Number of heads in 10 coin flips:  $X \in \{0, 1, 2, 3, \dots, 10\}$
- Number of emails received per hour:  $X \in \{0, 1, 2, 3, \dots\}$
- Test score (if graded on points):  $X \in \{0, 1, 2, \dots, 100\}$

- Product defects in a batch:  $X \in \{0, 1, 2, 3, \dots\}$

### Probability Mass Function (PMF):

$P(X = k)$  = probability that  $X$  equals exactly  $k$

Properties:

- $0 \leq P(X = k) \leq 1$
- $\sum P(X = k) = 1$  (probabilities sum to 1)

### Example: Fair die

$$P(X = 1) = 1/6$$

$$P(X = 2) = 1/6$$

...

$$P(X = 6) = 1/6$$

$$\text{Sum} = 6 \times (1/6) = 1 \checkmark$$

### Continuous Random Variables

A **continuous random variable** can take on **any value in an interval** (uncountably infinite values).

### Characteristics:

- Any real number in a range
- Cannot list all possible values
- Probability described by **probability density function (PDF)**

### Examples:

- Temperature tomorrow:  $X \in [-50^\circ\text{C}, 50^\circ\text{C}]$
- Height of a person:  $X \in [0\text{cm}, 300\text{cm}]$
- Stock price:  $X \in [0, \infty)$
- Response time of a website:  $X \in [0, \infty)$

### Probability Density Function (PDF):

$f(x)$  = probability density at value  $x$

Properties:

- $f(x) \geq 0$  (always non-negative)
- $\int f(x) dx = 1$  (total area under curve is 1)
- $P(a \leq X \leq b) = \int [a \text{ to } b] f(x) dx$  (probability is area under curve)

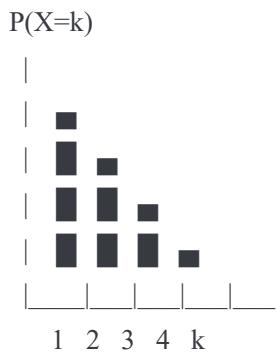
**Key difference:** For continuous variables,  $P(X = \text{exact value}) = 0$  (infinitesimal).

- $P(\text{temperature} = 25.000000^\circ\text{C exactly}) = 0$
- But  $P(24.5 \leq \text{temperature} \leq 25.5) > 0$

We can only talk about probability in intervals or ranges.

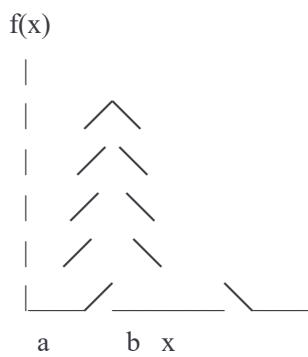
## Visual Comparison

### Discrete (PMF):



Probability is at specific points (bars).

### Continuous (PDF):



$P(a \leq X \leq b) = \text{area under curve}$

Probability is the area under the smooth curve.

## Cumulative Distribution Function (CDF)

For both discrete and continuous, the **CDF** gives cumulative probability:

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

### Properties:

- $0 \leq F(x) \leq 1$
- Monotonically increasing (never decreases)
- $F(-\infty) = 0, F(\infty) = 1$

### Example: Fair die

$$F(1) = P(X \leq 1) = 1/6$$

$$F(2) = P(X \leq 2) = 2/6$$

$$F(3) = P(X \leq 3) = 3/6 = 1/2$$

...

$$F(6) = P(X \leq 6) = 1$$

## Applications in ML

**Discrete RVs:** Classification (probability of each class), counting events (defects, clicks)

**Continuous RVs:** Regression (predicting real values), measurements, time, distance

Most real-world ML problems involve continuous random variables, but many also model discrete outcomes (categorical classification).

---

## Common Distributions: Bernoulli, Binomial, Normal, Poisson

Understanding these distributions is essential—they model most real-world data.

### Bernoulli Distribution

The **Bernoulli distribution** models a single trial with two possible outcomes (success/failure, yes/no).

**Parameter:**  $p$  = probability of success ( $0 \leq p \leq 1$ )

### PMF:

$$P(X = 1) = p \quad (\text{success})$$

$$P(X = 0) = 1 - p \quad (\text{failure})$$

## Mean and Variance:

$$E[X] = p$$

$$\text{Var}(X) = p(1 - p)$$

## Examples:

- Coin flip:  $p = 0.5$
- Email is spam:  $p = 0.3$
- Customer converts:  $p = 0.02$
- Loan defaults:  $p = 0.05$

**In ML:** Binary classification output, click-through rates, success indicators.

## Binomial Distribution

The **Binomial distribution** models multiple independent Bernoulli trials.

## Parameters:

- $n$  = number of trials
- $p$  = probability of success in each trial

## PMF:

$$P(X = k) = C(n,k) \times p^k \times (1-p)^{n-k}$$

Where  $C(n,k) = n! / (k! \times (n-k)!)$  is "n choose k"

**Intuition:** Of  $n$  trials, exactly  $k$  succeed (with probability  $p$ ), the rest fail (with probability  $1-p$ ). The  $C(n,k)$  term counts how many ways this can happen.

## Mean and Variance:

$$E[X] = n \times p$$

$$\text{Var}(X) = n \times p \times (1 - p)$$

**Example: 5 coin flips, how many heads?**

$n = 5, p = 0.5$

$$\begin{aligned}P(X = 3) &= C(5,3) \times (0.5)^3 \times (0.5)^2 \\&= 10 \times 0.125 \times 0.25 \\&= 0.3125 \text{ (31.25% chance of 3 heads)}\end{aligned}$$

$$E[X] = 5 \times 0.5 = 2.5 \text{ (expect 2.5 heads on average)}$$

$$\text{Var}(X) = 5 \times 0.5 \times 0.5 = 1.25$$

### Example: Ad clicks

$n = 1000$  ad impressions,  $p = 0.02$  click-through rate

$$E[\text{clicks}] = 1000 \times 0.02 = 20 \text{ clicks}$$

$$\text{Var}(\text{clicks}) = 1000 \times 0.02 \times 0.98 = 19.6$$

**In ML:** Number of successes in fixed trials, conversion counts, defect counts in batches.

### Normal Distribution

The **Normal (Gaussian) distribution** is the most important distribution in statistics and ML.

#### Parameters:

- $\mu$  (mu) = mean (center of distribution)
- $\sigma$  (sigma) = standard deviation (spread)

#### PDF:

$$f(x) = (1 / (\sigma\sqrt{2\pi})) \times \exp(-(x - \mu)^2 / (2\sigma^2))$$

This intimidating formula creates the famous **bell curve**.

#### Mean and Variance:

$$E[X] = \mu$$

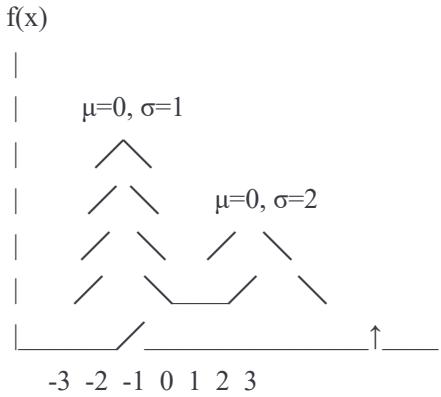
$$\text{Var}(X) = \sigma^2$$

#### Properties:

- Symmetric around  $\mu$  (mean = median = mode)
- 68% of data within  $\mu \pm \sigma$

- 95% of data within  $\mu \pm 2\sigma$
- 99.7% of data within  $\mu \pm 3\sigma$

### Visualizing normal distributions:



Narrower curve: smaller  $\sigma$  (more concentrated)

Wider curve: larger  $\sigma$  (more spread out)

### Standard Normal Distribution (Z):

$Z \sim N(0, 1)$  [mean = 0, std dev = 1]

Any normal variable can be standardized:

$$Z = (X - \mu) / \sigma$$

### Example: Heights

Adult male heights:  $X \sim N(175 \text{ cm}, 8 \text{ cm})$

$$P(X < 183) = P(Z < (183-175)/8) = P(Z < 1) \approx 0.84 \text{ (84%)}$$

$$P(165 < X < 185) = P(-1.25 < Z < 1.25) \approx 0.79 \text{ (79%)}$$

### Why Normal Distribution?

- **Central Limit Theorem** (see next section): averages of many independent variables approach normal
- Many natural phenomena follow normal distribution (heights, test scores, measurement errors)
- Mathematically convenient (easy to work with)
- Many ML algorithms assume normality (linear regression, Gaussian processes)

### In ML:

- Modeling measurement errors
- Prior distributions in Bayesian methods
- Assuming normally distributed features improves some algorithms
- Regularization often assumes normally distributed weights

## Poisson Distribution

The **Poisson distribution** models the number of events occurring in a fixed time/space interval.

**Parameter:**  $\lambda$  (lambda) = average rate of events

**PMF:**

$$P(X = k) = (e^{-\lambda} \times \lambda^k) / k!$$

**Mean and Variance:**

$$E[X] = \lambda$$

$$\text{Var}(X) = \lambda$$

Unique property: **mean equals variance** (unusual!).

**Example: Customer arrivals**

$$\lambda = 10 \text{ customers per hour}$$

$$P(\text{exactly 5 arrive}) = (e^{-10} \times 10^5) / 5! \approx 0.038 \text{ (3.8%)}$$

$$P(\text{more than 15 arrive}) = ?$$

$$E[X] = 10 \text{ (expect 10 on average)}$$

$$\text{Var}(X) = 10 \text{ (std dev } \approx 3.16)$$

**Examples:**

- Emails received per hour
- Website traffic per minute
- Accident rates per month
- Defects per meter of material
- Calls to a help desk

## When to use Poisson:

- Events occur independently
- Events happen at constant average rate
- Time/space intervals are independent

In ML: Modeling count data, anomaly detection (unusual spike in counts), forecasting event frequencies.

## Distribution Selection Guide

Distribution	Situation	Parameter(s)
Bernoulli	Single yes/no event	p (success rate)
Binomial	Fixed number of trials, count successes	n, p
Normal	Measurement, aggregated data, errors	$\mu, \sigma$
Poisson	Count of events in fixed interval	$\lambda$ (rate)

## Central Limit Theorem (CLT) and Why It Matters in ML

The **Central Limit Theorem** is arguably the most important theorem in statistics and ML.

### The Theorem Statement

If you take repeated samples of size  $n$  from ANY distribution and compute the sample mean, those means will form a **Normal distribution**, regardless of the original distribution's shape.

Mathematically:

If  $X_1, X_2, \dots, X_n$  are independent samples from any distribution with mean  $\mu$  and variance  $\sigma^2$

Then the sample mean  $\bar{X} = (X_1 + X_2 + \dots + X_n) / n$

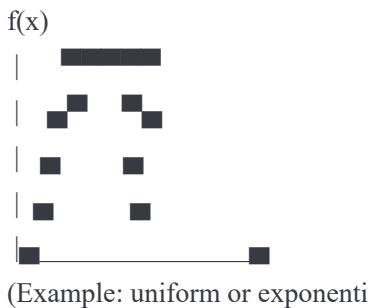
Approaches  $N(\mu, \sigma^2/n)$  as  $n \rightarrow \infty$

The distribution of means is approximately **normal** with:

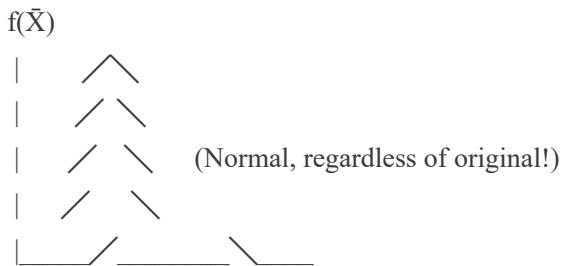
- Mean =  $\mu$  (same as original)
- Variance =  $\sigma^2/n$  (decreases as sample size increases!)
- Standard error =  $\sigma / \sqrt{n}$

## Visual Intuition

**Original distribution** (could be any shape):



**Distribution of sample means:**



**Effect of sample size:**

n=10: (wider, more uncertain)

n=100: (narrower, more certain)

n=1000: (very narrow, very certain)

As n increases, distribution of means gets narrower and more concentrated.

## Why This Matters: Three Key Reasons

**1. Generalization from Theory to Practice** Many ML algorithms assume normally distributed data. CLT justifies this:

- If you aggregate many independent factors, the result is approximately normal
- Most real measurements are aggregates of many small effects

**2. Confidence Intervals and Hypothesis Testing** Because sample means are normal, we can use normal distribution properties:

95% Confidence Interval for true mean  $\mu$ :

$$[\bar{X} - 1.96 \times SE, \bar{X} + 1.96 \times SE]$$

Where  $SE = \sigma / \sqrt{n}$  (standard error)

### Example: Average customer lifetime value

Sample of 100 customers:  $\bar{X} = \$500$

Sample std dev:  $\sigma = \$200$

$$SE = \$200 / \sqrt{100} = \$20$$

$$95\% CI = [500 - 1.96 \times 20, 500 + 1.96 \times 20]$$

$$= [\$460.80, \$539.20]$$

We're 95% confident true mean is between \$460.80 and \$539.20

### 3. Sample Size Calculation

CLT determines required sample size:

$$SE = \sigma / \sqrt{n}$$

To halve standard error (be twice as confident):

Need  $\sqrt{4} = 4$  times as many samples

If 100 samples give  $\pm \$20$ , need 400 samples for  $\pm \$10$ .

## CLT in Machine Learning Practice

### 1. Model Averaging

Final prediction = average of many model predictions

By CLT: prediction uncertainty decreases with more models

### 2. Confidence in Parameter Estimates

Model parameter  $w$  estimated from data

Distribution of  $w$  estimates across different samples approaches normal

Standard error tells us confidence in the estimate

### 3. Regularization Justification

Many ML algorithms find weights that minimize average error

By CLT: error distribution is normal

Regularization assumes normal distribution of errors

## 4. Batch Normalization in Neural Networks

Each mini-batch is a sample

Normalizing based on batch statistics relies on CLT reasoning

## 5. Bootstrap Confidence Intervals

Resample data many times, compute statistic for each sample

Distribution of statistics approaches normal by CLT

### Example: Website Conversion Rate

Your website converts visitors to customers. You want to estimate true conversion rate.

#### Data collection:

Sample 1: 100 visitors, 10 conversions  $\rightarrow \hat{p}_1 = 0.10$

Sample 2: 100 visitors, 15 conversions  $\rightarrow \hat{p}_2 = 0.15$

Sample 3: 100 visitors, 12 conversions  $\rightarrow \hat{p}_3 = 0.12$

...

Sample 100: 100 visitors, 11 conversions  $\rightarrow \hat{p}_{100} = 0.11$

#### Distribution of sample proportions:

By CLT, these proportions form a normal distribution!

Mean of proportions:  $(0.10 + 0.15 + 0.12 + \dots + 0.11) / 100 \approx 0.127$

Standard error:  $\sigma/\sqrt{n}$

95% CI:  $[0.127 - 1.96 \times \text{SE}, 0.127 + 1.96 \times \text{SE}]$

This is how we estimate true conversion rate with confidence.

### When CLT Fails

CLT assumes:

- **Independence:** Samples don't influence each other

- **Large enough n:** Rule of thumb:  $n \geq 30$  (or more if original distribution very skewed)
- **Finite variance:** Original distribution must have finite variance (some distributions don't)

With these conditions met, CLT is remarkably robust.

---

## Point Estimation of Parameters

When you have data, you need to estimate the distribution's parameters (like  $\mu$  and  $\sigma$  for normal).

### What is Point Estimation?

**Point estimation** is computing a single value (point estimate) as an estimate of an unknown parameter.

#### Examples:

True population mean  $\mu$  (unknown) → Estimate with sample mean  $\bar{x}$

True population variance  $\sigma^2$  (unknown) → Estimate with sample variance  $s^2$

True probability  $p$  (unknown) → Estimate with sample proportion  $\hat{p}$

### Common Point Estimators

#### Sample Mean (estimates population mean $\mu$ ):

$$\bar{x} = \sum x_i / n$$

Unbiased:  $E[\bar{x}] = \mu$  ✓

Efficient: Low variance, CLT applies

Most commonly used

#### Sample Variance (estimates population variance $\sigma^2$ ):

$$s^2 = \sum (x_i - \bar{x})^2 / (n - 1)$$

Unbiased:  $E[s^2] = \sigma^2$  ✓

Note: divide by  $(n-1)$ , not  $n$  [Bessel's correction]

#### Sample Proportion (estimates probability $p$ ):

$\hat{p} = (\text{number of successes}) / n$

Unbiased:  $E[\hat{p}] = p \checkmark$

Used for binary events

### Sample Median (estimates population median):

Sort data, take middle value

Robust to outliers (unlike mean)

## Properties of Good Estimators

### 1. Unbiasedness

$E[\text{estimator}] = \text{true parameter}$

On average, the estimate equals the truth (no systematic over/under estimation).

**Example:** Sample mean is unbiased for  $\mu$

$E[\bar{x}] = \mu$

### 2. Consistency

As  $n \rightarrow \infty$ , estimator  $\rightarrow$  true parameter (with probability 1)

With more data, estimate gets better.

### 3. Efficiency

Smaller variance is better

Among unbiased estimators, choose the one with lowest variance.

**Example:**

Sample mean:  $\text{Var}(\bar{x}) = \sigma^2/n$

Sample median:  $\text{Var}(\text{median}) \approx 1.25 \times \sigma^2/n$  (less efficient)

### 4. Robustness

Not sensitive to outliers or violations of assumptions

Sample median is robust; sample mean is not.

## Maximum Likelihood Estimation (MLE)

**Maximum Likelihood Estimation** finds parameters that make observed data most likely.

**Idea:** Given data, what parameter values would make this data most probable?

**Likelihood function:**

$$L(\theta | \text{data}) = P(\text{data} | \theta)$$

Probability of observing this data given parameter  $\theta$

**MLE:**

$$\theta = \operatorname{argmax} L(\theta | \text{data})$$

Find parameter that maximizes likelihood.

### Example: Coin flips

Observed: 7 heads in 10 flips. What's  $p$  (probability of heads)?

$$\text{Likelihood: } L(p | 7H, 3T) = p^7 \times (1-p)^3$$

$$L(0.5) = 0.5^7 \times 0.5^3 = 0.001$$

$$L(0.7) = 0.7^7 \times 0.3^3 = 0.0097$$

$$L(0.8) = 0.8^7 \times 0.2^3 = 0.0167$$

$$L(0.9) = 0.9^7 \times 0.1^3 = 0.000729$$

Maximum at  $p = 0.7$

MLE says  $p \approx 0.7$  (which equals  $7/10$ , the observed proportion!).

## Practical Estimation

In practice with real data:

**Step 1:** Determine likely distribution (histogram, domain knowledge)

**Step 2:** Compute parameter estimates:

```

python

# If normal distribution
μ_hat = mean(data)
σ_hat = std(data)

# If binomial
n = total_trials
p_hat = successes / total_trials

# If Poisson
λ_hat = mean(data)

```

### Step 3: Compute confidence intervals (using CLT):

$$95\% \text{ CI} = [\text{estimate} - 1.96 \times \text{SE}, \text{estimate} + 1.96 \times \text{SE}]$$

### Example: Customer spending

Data: [50, 75, 100, 60, 90, 80, 110, 95, 85, 100] (10 customers)

$$\bar{x} = 845 / 10 = \$84.5$$

$$s^2 = \sum(x_i - \bar{x})^2 / 9 = 1592.5 / 9 \approx 176.9$$

$$s = \sqrt{176.9} \approx \$13.3$$

$$SE = s / \sqrt{n} = 13.3 / \sqrt{10} \approx \$4.2$$

$$\begin{aligned} 95\% \text{ CI} &= [84.5 - 1.96 \times 4.2, 84.5 + 1.96 \times 4.2] \\ &= [\$76.3, \$92.7] \end{aligned}$$

We're 95% confident true mean spending is between \$76.3 and \$92.7.

## Introduction to Monte Carlo Simulation

**Monte Carlo simulation** estimates quantities by randomly sampling and computing.

### Core Idea

Instead of solving a problem analytically (with formulas), simulate it with random numbers.

### Process:

1. Generate random samples from a distribution

2. Apply a function or rule to each sample
3. Aggregate results (average, count, etc.)
4. Use aggregate as estimate

## Why Monte Carlo?

Many problems are hard or impossible to solve analytically:

- Complex probability calculations
- High-dimensional integrals
- Non-linear systems
- Real-world simulations

Monte Carlo provides approximate answers.

## Simple Example: Estimating $\pi$

**Problem:** Estimate  $\pi$  using random sampling.

### Method:

1. Generate random points in a  $1 \times 1$  square
2. Check if each point is inside a quarter-circle (radius 1)
3. Ratio: (points in circle) / (total points)  $\approx (\pi/4) / 1 = \pi/4$
4. Multiply by 4 to get  $\pi$

Algorithm:

For i = 1 to N:

  x = random(0, 1)

  y = random(0, 1)

  if  $x^2 + y^2 \leq 1$ :

    count\_inside += 1

$$\pi_{\text{estimate}} = 4 \times \text{count\_inside} / N$$

## Results with different sample sizes:

N = 100:  $\pi_{\text{estimate}} \approx 3.16$   
N = 1,000:  $\pi_{\text{estimate}} \approx 3.14$   
N = 10,000:  $\pi_{\text{estimate}} \approx 3.1415$   
N = 100,000:  $\pi_{\text{estimate}} \approx 3.14159\dots$

With more samples, estimate approaches true value ( $\pi \approx 3.14159$ ).

## Computing Expected Values

**Problem:** Find  $E[g(X)]$  for complex function g.

**Analytical approach:**  $E[g(X)] = \int g(x) \times f(x) dx$  (hard integral!)

**Monte Carlo approach:**

1. Sample  $X_1, X_2, \dots, X_n$  from distribution of X
2. Compute  $g(X_1), g(X_2), \dots, g(X_n)$
3. Average:  $\hat{E}[g(X)] = \sum g(X_i) / n$

By law of large numbers, this average converges to true expected value.

## Example: Option pricing

Black-Scholes formula for option price is complex. Monte Carlo approach:

```
for i = 1 to N:  
    simulate stock price path to maturity  
    payoff = max(stock_price - strike, 0)  
    payoffs.append(payoff)  
  
option_price = average(payoffs) × discount_factor
```

This works even for complex, non-linear models.

## Monte Carlo in Machine Learning

### 1. Bayesian Inference

Sample from posterior distribution  
Use samples to estimate credible intervals

### 2. Uncertainty Quantification

Propagate input uncertainty through model

Estimate output distribution

Quantify prediction uncertainty

### 3. Model Evaluation

Resample from data distribution

Train model on each resample

Estimate distribution of performance metrics

### 4. Hyperparameter Tuning

Randomly sample hyperparameter combinations

Train and evaluate on each

Find best combination

### 5. Reinforcement Learning

Monte Carlo tree search: sample possible action sequences

Estimate value of each path

Choose highest-value action

## Variance Reduction Techniques

Monte Carlo estimates improve with more samples, but sampling is expensive.

### Importance Sampling:

Sample from different (easier) distribution

Weight samples by likelihood ratio

Reduces variance without more samples

### Stratified Sampling:

Divide population into strata

Sample proportionally from each stratum

Ensures representative sampling

Reduces variance

### Control Variates:

Use a correlated variable with known expected value

Reduce variance by exploiting correlation

## Practical Monte Carlo Example: Risk Assessment

**Scenario:** Uncertain project completion time due to multiple tasks.

Task A: 5-7 days (uniform)

Task B: 8-12 days (uniform)

Task C: 4-6 days (uniform)

Total time = A + B + C

What's distribution of total time?

### Monte Carlo approach:

```
for i = 1 to 10,000:
```

```
    a = random(5, 7)
```

```
    b = random(8, 12)
```

```
    c = random(4, 6)
```

```
    total = a + b + c
```

```
    times.append(total)
```

Results:

Mean completion time:  $E[\text{total}] \approx 17.5$  days

5th percentile: 14.2 days (90% confident of finishing by then)

95th percentile: 20.8 days

From 10,000 simulations, we estimate the complete distribution without analytical formulas.

### Convergence and Accuracy

**Accuracy improves with  $\sqrt{N}$ :**

Error  $\propto 1/\sqrt{N}$

To get 10x more accuracy, need 100x more samples. This is the fundamental tradeoff.

### Standard error of estimate:

$$SE \approx \sigma / \sqrt{N}$$

$\sigma$  = std dev of function values

N = number of samples

This follows from CLT—the average of samples is normally distributed!

### Pseudocode: General Monte Carlo

```
function monte_carlo(N, distribution, function):
    samples = 0
    for i = 1 to N:
        x = random_sample_from(distribution)
        samples += function(x)

    return samples / N
```

Simple algorithm, but powerful.

---

## Putting It All Together: A Complete Example

### Scenario: Customer Lifetime Value (CLV) Estimation

You want to estimate customer lifetime value, which depends on:

- Annual spending:  $S \sim \text{Normal}(\mu=\$500, \sigma=\$100)$
- Retention rate:  $r \sim \text{Beta}(\alpha=2, \beta=1)$  [higher probability of staying]
- Customer lifetime:  $T = \text{geometric sum based on } r$
- Discount rate:  $d = 5\%$  per year

**Analytical approach:** Complex (involves infinite series)

**Monte Carlo approach:**

```

for i = 1 to 100,000:
    annual_spend = sample from Normal(500, 100)
    retention_rate = sample from Beta(2, 1)

    clv = 0
    for year in 1 to 20: # assume max 20 years
        probability_staying = retention_rate ^ year
        discounted_value = annual_spend / (1.05 ^ year)
        clv += probability_staying * discounted_value

    clv_estimates.append(clv)

```

Results:

Mean CLV:  $E[CLV] \approx \text{average}(clv\_estimates)$

Std Dev:  $\sigma \approx \text{std}(clv\_estimates)$

95% CI: [percentile(5), percentile(95)]

## Output:

Mean CLV: \$4,250

95% CI: [\$3,800, \$4,700]

Median: \$4,200

Std Dev: \$450

Monte Carlo lets us estimate CLV without closed-form formulas, handling complex uncertainty.

---

## Summary: From Theory to Simulation

**Random variables** convert outcomes to numbers, enabling mathematics.

**Common distributions** (Bernoulli, Binomial, Normal, Poisson) describe most real phenomena.

**Central Limit Theorem** explains why normal distributions are ubiquitous and justifies confidence intervals.

**Point estimation** finds best-guess parameters from data using sample statistics.

**Monte Carlo simulation** estimates probabilities and expected values through random sampling.

Together, these form the statistical foundation of machine learning:

- Understanding distributions helps choose appropriate models
- CLT justifies statistical inference and confidence intervals

- Point estimation fills in unknown parameters
- Monte Carlo simulates complex systems and estimates uncertainty

Master these concepts, and you understand the probabilistic foundations underlying all of modern data science and machine learning.