Central Limit Theorem Proof Explained in Plain English

What We're Trying to Prove

Imagine you have a bunch of random measurements—maybe people's heights, test scores, or coin flip results. Each measurement comes from the same underlying process, so they all have the same average (mean) and the same amount of spread (variance).

Now, if you take the average of many such measurements, something remarkable happens: **the distribution of this average becomes more and more like a bell curve (Gaussian distribution)**, no matter what the original distribution looked like!

The Central Limit Theorem makes this precise: if you properly adjust (standardize) your average by subtracting the true mean and dividing by the appropriate amount of spread, then as you take more and more measurements, this standardized average follows a standard normal distribution.

The Mathematical Setup

Let's say we have random variables $X_1, X_2, X_3, \ldots$ that are:

- **Independent**: Each measurement doesn't affect the others
- **Identically distributed**: They all come from the same underlying process
- **Finite mean μ**: They have a well-defined average
- **Finite variance σ^2**: They don't have infinite spread

We want to study the standardized sum:

$$Z_n = \frac{(X_1 + X_2 + \cdots + X_n) - n\mu}{\sigma\sqrt{n}}$$

This formula does two things:

1. **Centers** the sum by subtracting the expected total $(n\mu)$
2. **Scales** it by dividing by $\sigma\sqrt{n}$, which is the standard deviation of the sum

Why Use Characteristic Functions?

A **characteristic function** is like a "fingerprint" of a probability distribution. Every distribution has a unique characteristic function, and if two distributions have the same characteristic function, they're the same distribution.

The key insight is this: **instead of trying to prove that the distributions become Gaussian directly, we can prove that their "fingerprints" (characteristic functions) become the fingerprint of a Gaussian distribution**.

This is much easier because characteristic functions turn complicated operations (like adding random variables) into simple multiplication.

Step-by-Step Explanation

Step 1: Simplify by Centering

First, let's make our lives easier by working with centered variables. Define:

$$Y_i = X_i - \mu$$

Now each Y_i has mean 0 and variance σ^2. Our standardized sum becomes:

$$Z_n = \frac{Y_1 + Y_2 + \cdots + Y_n}{\sigma\sqrt{n}}$$

Step 2: Write Down the Characteristic Function

The characteristic function of Z_n is:

$$\varphi_{Z_n}(t) = \mathbb{E}\left[e^{itZ_n}\right] = \mathbb{E}\left[\exp\left(it \cdot \frac{Y_1 + \cdots + Y_n}{\sigma\sqrt{n}}\right)\right]$$

Because the Y_i are independent, we can separate this expectation into a product:

$$\varphi_{Z_n}(t) = \left[\mathbb{E}\left[\exp\left(\frac{itY_1}{\sigma\sqrt{n}}\right)\right]\right]^n = \left[\varphi_{Y_1}\left(\frac{t}{\sigma\sqrt{n}}\right)\right]^n$$

This is the magic of independence—adding independent random variables corresponds to multiplying their characteristic functions.

Step 3: Understand What Happens for Large n

Here's the crucial observation: as n gets very large, the argument $\frac{t}{\sigma\sqrt{n}}$ becomes very small (close to 0). So we only need to understand how the characteristic function of Y_1 behaves **near zero**.

Step 4: Expand Near Zero Using Taylor Series

Think of the characteristic function $\varphi_{Y_1}(s) = \mathbb{E}[e^{isY_1}]$ as a smooth curve. Near $s = 0$, we can approximate it using its Taylor series (like approximating a curve with a polynomial near a point).

Using Euler's formula $e^{ix} = \cos x + i \sin x$ and taking expectations:

$$\varphi_{Y_1}(s) = \mathbb{E}[\cos(sY_1)] + i\mathbb{E}[\sin(sY_1)]$$

For small s, we can use the approximations:

- $\cos(sY_1) \approx 1 - \frac{(sY_1)^2}{2}$
- $\sin(sY_1) \approx sY_1 - \frac{(sY_1)^3}{6}$

Taking expectations and remembering that $\mathbb{E}[Y_1] = 0$ and $\mathbb{E}[Y_1^2] = \sigma^2$:

$$\varphi_{Y_1}(s) \approx 1 - \frac{s^2\sigma^2}{2} + \text{(higher order terms)}$$

The key point is that **the behavior near zero only depends on the first two moments** (mean and variance) of the distribution. All the details about the original shape of the distribution are hidden in the "higher order terms" that become negligible.

Step 5: Plug in the Small Argument

Now substitute $s = \frac{t}{\sigma\sqrt{n}}$:

$$\varphi_{Y_1}\left(\frac{t}{\sigma\sqrt{n}}\right) \approx 1 - \frac{1}{2}\left(\frac{t}{\sigma\sqrt{n}}\right)^2 \sigma^2 = 1 - \frac{t^2}{2n}$$

Notice how beautiful this is: the σ^2 cancels out, leaving us with something that doesn't depend on the original variance at all!

Step 6: Raise to the nth Power

Now remember we need to raise this to the nth power:

$$\varphi_{Z_n}(t) \approx \left(1 - \frac{t^2}{2n}\right)^n$$

Step 7: Take the Limit

Here's where calculus comes in. There's a famous limit from calculus:

$$\lim_{n\to\infty}\left(1 + \frac{a}{n}\right)^n = e^a$$

In our case, $a = -\frac{t^2}{2}$, so:

$$\lim_{n\to\infty}\varphi_{Z_n}(t) = e^{-t^2/2}$$

Step 8: Recognize the Gaussian Fingerprint

The function $e^{-t^2/2}$ is exactly the characteristic function of the **standard normal distribution**! This is the "fingerprint" we were looking for.

Why This Proof is So Powerful

This proof reveals something profound: **the Gaussian distribution is universal because it's what you get when you only care about the first two moments (mean and variance) of a distribution**.

When you add up many independent random variables:

- The details of their individual distributions get "washed out"
- Only their means and variances matter for the limiting behavior
- The mathematical structure forces the result to be Gaussian

It's like mixing many different colors of paint—if you mix enough different colors together, you always get a muddy brown, regardless of what specific colors you started with. Similarly, if you add enough independent random effects together, you always get something that looks Gaussian.

The Big Picture

The characteristic function approach works because it transforms a difficult problem about distributions into an easier problem about functions. By focusing on the behavior near zero (which corresponds to the large-scale behavior of the sum), we can ignore all the complicated details of the original distribution and see the universal Gaussian pattern emerge.

This is why the Central Limit Theorem appears everywhere in nature, science, and engineering—any phenomenon that results from the accumulation of many small, independent influences will naturally follow a bell curve, and this proof shows us exactly why that happens mathematically.

1. What is the Binomial Distribution?

Imagine you're flipping a (possibly biased) coin n times. Each flip is a **Bernoulli trial**:

- It has two outcomes: **success** (e.g., heads) with probability p, and **failure** (tails) with probability $1 - p$.
- All flips are **independent**: the result of one doesn't affect the others.

Let X be the total number of successes in those n flips. Then X follows a **binomial distribution**, written as:

$$X \sim \text{Binomial}(n, p).$$

Its probability mass function is:

$$\mathbb{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad \text{for } k = 0, 1, \ldots, n.$$

This tells us how likely it is to get exactly k heads in n flips.

2. What Does "Converges to the Gaussian" Mean?

As n gets very large, the shape of the binomial distribution starts to look more and more like a **bell curve**—the famous **Gaussian (normal) distribution**.

But we must be careful: the binomial is **discrete** (only defined at integer values), while the Gaussian is **continuous**. So we don't mean they become *identical*, but rather that **after appropriate centering and scaling**, the binomial's shape becomes indistinguishable from a normal curve.

This is where the **Central Limit Theorem (CLT)** comes in.

3. The Central Limit Theorem (CLT): The Big Idea

The CLT is one of the most powerful results in probability. In simple terms, it says:

> If you add up a large number of independent, identically distributed (i.i.d.) random variables—each with finite mean and variance—then the sum (properly normalized) will look approximately like a normal distribution, no matter what the original distribution was!

More formally, suppose $Y_1, Y_2, \ldots, Y_n$ are i.i.d. random variables with:

$$\mathbb{E}[Y_i] = \mu, \quad \mathrm{Var}(Y_i) = \sigma^2 < \infty.$$

Define their sum: $S_n = Y_1 + Y_2 + \cdots + Y_n$.

Then the **standardized version** of this sum:

$$Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}}$$

converges in distribution to a **standard normal random variable** as $n \to \infty$. That is:

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

This means the cumulative distribution function (CDF) of Z_n gets closer and closer to the CDF of a standard normal.

4. Connecting the Binomial to the CLT

Here's the key insight:

A binomial random variable is just the sum of n independent Bernoulli random variables.

Define Y_i as the outcome of the i-th coin flip:

$$Y_i = \begin{cases} 1 & \text{with probability } p \quad \text{(success)}, \\ 0 & \text{with probability } 1-p \quad \text{(failure)}. \end{cases}$$

Each $Y_i \sim \text{Bernoulli}(p)$, and they're independent.

Then the total number of successes is:

$$X = Y_1 + Y_2 + \cdots + Y_n,$$

so indeed $X \sim \text{Binomial}(n, p)$.

Now compute the mean and variance of each Y_i:

$$\mathbb{E}[Y_i] = p, \quad \text{Var}(Y_i) = p(1-p).$$

Therefore, for the sum X:

$$\mathbb{E}[X] = np, \quad \text{Var}(X) = np(1-p).$$

5. Apply the CLT to the Binomial

We now plug the binomial into the CLT framework.

Standardize X as the CLT prescribes:

$$Z_n = \frac{X - \mathbb{E}[X]}{\sqrt{\mathrm{Var}(X)}} = \frac{X - np}{\sqrt{np(1-p)}}.$$

According to the CLT, as $n \to \infty$, this standardized variable converges in distribution to a standard normal:

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

In words:

> If you take a binomial count, subtract its average (np), and divide by its standard deviation ($\sqrt{np(1-p)}$), then for large n, the result behaves like a standard normal random variable.

6. What Does This Mean Visually?

Imagine plotting the binomial probabilities $\mathbb{P}(X = k)$ for $k = 0, 1, \ldots, n$. For small n, the distribution may look lopsided or jagged—especially if p is far from 0.5.

But as n grows:

- The distribution becomes **smoother** (more possible values of k).
- Its shape becomes **symmetric and bell-shaped**, centered at np, with spread proportional to $\sqrt{n}$.

If you rescale the horizontal axis using

$$z = \frac{k - np}{\sqrt{np(1-p)}},$$

then the histogram of these rescaled points will hug the standard normal curve tighter and tighter as n increases.

This result is historically known as the **de Moivre–Laplace theorem**, a special case of the CLT for binomial distributions.

7. A Note on "Convergence in Distribution"

When we write $Z_n \xrightarrow{d} \mathcal{N}(0,1)$, we mean that for any real number z,

$$\mathbb{P}(Z_n \le z) \longrightarrow \Phi(z) \quad \text{as } n \to \infty,$$

where $\Phi(z)$ is the CDF of the standard normal distribution.

This does *not* mean that individual point probabilities $\mathbb{P}(X = k)$ become equal to the normal density—but it *does* mean that **probabilities over intervals** (e.g., $\mathbb{P}(a \le X \le b)$) can be well-approximated by the area under the normal curve after standardizing.

Hence, in practice, we use the **normal approximation to the binomial**:

$$\mathbb{P}(a \le X \le b) \approx \Phi\left(\frac{b + 0.5 - np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{a - 0.5 - np}{\sqrt{np(1-p)}}\right),$$

where the ± 0.5 is the **continuity correction**, accounting for the fact that we're approximating a discrete distribution with a continuous one.

8. Why Does This Happen? Intuition

Think of each coin flip as a tiny random "push"—sometimes +1 (success), sometimes 0 (failure). When you add up thousands of these tiny, independent pushes, the randomness tends to **average out**, and the total fluctuation around the mean becomes **smooth and symmetric**.

Extreme outcomes (like all heads or all tails) become astronomically unlikely, while outcomes near the average dominate. The mathematics of summing independent noise naturally leads to the bell shape—this is the essence of the CLT.

Since the binomial is just a sum of simple yes/no trials, it's a perfect candidate for this averaging effect.

9. Summary in One Sentence

Because a binomial random variable is the sum of many independent, identical Bernoulli trials, the Central Limit Theorem guarantees that—when properly centered and scaled—it converges in distribution to a Gaussian (normal) random variable as the number of trials goes to infinity.

Final Thought

This convergence is not just theoretical—it's **practically useful**. It lets statisticians approximate binomial probabilities (which can be hard to compute for large n) using the well-understood normal distribution. And it's a beautiful example of how **order emerges from randomness** when you look at large systems.