CENTRAL LIMIT THEOREM

JUSTIN CHEONG

Abstract

In this paper, we will prove a few different versions of the Central Limit Theorem: first, a weak one, then the Lindeberg and Liapunov Central Limit Theorems. Additionally, we will show the Central Limit Theorem in action and explain why it is one of the most useful theorems in all of statistics.

1. Introduction

In many situations, we would like to find the average value of an event, but as the event has randomness, there must also some randomness in the mean. We know that with more samples taken to compute the mean, the more accurate it is; however, how do we know how many samples to use such that we have confidence in knowing the value we get is within a certain tolerance range? This question can be answered with the help of Central Limit Theorem, which states that under certain conditions, the probability distribution of a sum approaches a normal distribution as you sum more and more random variables. On top of that, it allows statisticians to assume a huge portion of sums are nearly normal, giving them access to tools they can use for such distributions. There are also many versions of the Central Limit Theorem, each having different prerequisites but having the same result. Versions included in this paper include 3.1, where random variables are independent and identically distributed, and 5.1 and 6.1, where random variables are independent and satisfy certain criteria but do not need to be identically distributed. The main results of this paper are showing various proofs for different versions of the Central Limit Theorem and also showing results and applications of it. First, we will go through some definitions, which you should check out before reading the paper. Next, we'll go over the weaker Central Limit Theorem described here. In the following section is a proof of that central limit theorem using Pafnuty Chebyshev's method of moment convergence. Then, we'll show and prove the Lindeberg Central Limit Theorem which only requires independent random variables that satisfy the Lindeberg Condition. Similarly, there exists a Liapunov Central Limit Theorem which only requires independent random variables that satisfy the Liapunov Condition; we'll prove this one too. Then, we'll show why the Central Limit Theorem is such a central part of statistics. Finally, a few further questions if you are interested in learning more about the Central Limit Theorem's applications.

Acknowledgements

The author would like to thank Simon Rubinstein-Salzedo, Paulina Paiz, and Kok-Wui Cheong for helpful conversations.

Date: July 16, 2024.

2. Background

2.1. Random Variables.

Definition 2.1. Random Variable
A random variable X is a real value outcome from an event.
For example, the outcome of rolling a fair 6-sided die. In this case, X could take on any of $1, 2, 3, 4, 5, 6$ with equal probability. There are two types of random variables: continuous random variables and discrete random variables. Continuous random variables are assigned a value from a continuous range, while discrete random variables are assigned a value from a discrete range, which may or may not be infinitely large. The dice roll above is an example of a discrete random variable, while a dart's distance from the bulls-eye after throwing it at a target is a continuous random variable.

Definition 2.2. Standard Normal Random Variable
A standard normal variable Z is a special type of random variable whose distribution follows the standard normal distribution, which will be discussed later.

2.2. Other Statistics definitions.

Definition 2.3. Probability Distribution Function
A probability distribution function is a function that describes the probabilities of the possible values for a random variable.

One thing to note is that for continuous random variables, you should look at the area under the distribution rather than the height of the distribution to find the probability of getting a value.

Notation. Probability $P(A)$ and $\Pr(A)$ denote the probability that an event A occurs.

Definition 2.4. Independent Variable
An independent variable is unaffected by the outcomes of the other variables in consideration.
For example, if you flip a fair coin twice, the second outcome isn't affected by the first outcome, so coinflips are independent variables.

Definition 2.5. Dependent Variable
On the other hand, a dependent variable is affected by the outcomes of other variables.

Definition 2.6. Identically Distributed Variables
Identically distributed variables have the same probability distribution, meaning they are essentially the same variable.
For example, all fair coinflips are identically distributed, since they all have a 50/50 chance for heads or tails.

Definition 2.7. i.i.d.
Independent and identically distributed is often shortened to i.i.d.

Definition 2.8. Expected Value
The expected value of a random variable is the average value you will get when running an experiment of a random variable. This can be calculated by summing up the product of each outcome's value and its probability.
For continuous random variables, we use an integral:

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x \Pr(x) dx.$$

As for discrete random variables, we just use a normal sum, but it still can be an infinite sum if there are uncountably many possible values for the random variable.

$$\mathbb{E}[X] = \sum_i x_i \Pr(x_i)$$

The expected value of a random variable X is often denoted with $\mathbb{E}[X]$. A property called the linearity of expectations states that for two random variables X and Y, $\mathbb{E}[X+Y] = \mathbb{E}[Y] + \mathbb{E}[X]$ no matter whether X and Y are independent. This stems from the definition of a random variable, mapping a real value outcome from an event.

A similar property, the product rule of expectation, states that $\mathbb{E}[X \cdot Y] = \mathbb{E}[X] \cdot \mathbb{E}[Y]$ if X and Y are independent. This can be done through manipulating some integrals for continuous random variables, and manipulating sums for discrete random variables.

There are two types of means we will be looking at:

Definition 2.9. Population Mean
A population mean, often denoted by μ, is another way to say the expected value of a random variable.

Definition 2.10. Sample Mean
A sample mean, often denoted by $\bar{x}$ is intended to be an estimate of the population mean, and is calculated by taking samples of the random variable and averaging them. As the number of times you run the experiment approaches infinity, the sample mean will approach the population mean, because when you do the experiment, your distribution of outcomes will approach the probability distribution, and hence the sample mean will approach the population mean.

We will be mostly looking at the population mean for the purposes of this paper.

Definition 2.11. Variance

Variance provides a way to characterize a distribution or random variable's spread. For a random variable or probability distribution, the variance is

$$\mathrm{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2].$$

To find the sample variance $\mathrm{Var}(X)$ or σ^2 of a distribution of data, for each data point X_n, take the square of the difference between itself and the mean.

Finally, average all those values found. This is represented mathematically by

$$\sigma^2 = \sum_{k=1}^{n} \frac{(X_n - \mu)^2}{n}.$$

Another useful form of variance can be derived from the definition:

$$\begin{aligned}\mathrm{Var}(X) &= \mathbb{E}[(X - \mathbb{E}[X])^2] \\ &= \mathbb{E}[X^2 - 2X\mathbb{E}[X] + (\mathbb{E}[X])^2] \\ &= \mathbb{E}[X^2] - 2\mathbb{E}[X]\mathbb{E}[X] + (\mathbb{E}[X])^2 \text{ by linearity of expectations} \\ &= \mathbb{E}[X^2] - (\mathbb{E}[X])^2\end{aligned}$$

One thing to note is that when making a random variable by summing independent random variables, for example turning four coinflips into 1 random variable, the variance of the final random variable is the sum of the variances of the original random variables.

Proof. Given two independent random variables X and Y with means 0,

$$\begin{aligned}\mathrm{Var}(X + Y) &= \mathbb{E}[(X + Y)^2] \\ &= \mathbb{E}[X^2 + 2XY + Y^2] \\ &= 2\mathbb{E}[x]\mathbb{E}[Y] + \mathbb{E}[X^2] + \mathbb{E}[Y^2] \\ &= \mathrm{Var}(X) + \mathrm{Var}(Y).\end{aligned}$$

If random variables were to not have mean 0, this would still apply as changing the mean doesn't change the variance. Another interesting fact is that since when finding the variance of $\mathrm{Var}(X - Y)$, the only change in the expected value is the sign of $2\mathbb{E}[X]]\mathbb{E}[Y]$, which is 0 anyways; therefore,

$$\mathrm{Var}(X + Y) = \mathrm{Var}(X - Y) = \mathrm{Var}(X) + \mathrm{Var}(Y).$$

■

Definition 2.12. Standard Deviation
The standard deviation of a distribution, normally denoted by σ, is the square root of the variance.

Definition 2.13. Normal Distribution
A bell curve, or a normal distribution, is a special type of curve that holds many useful properties in statistics. Some properties include a standardized area distribution (68.2% of area lies within 1 standard deviation, 95.4% within 2, 99.7% within 3, etc.), symmetry about the mean, and equality between the mean, median, and mode. All normal distributions have area 1 under the curve and follow the equation

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

for their probability distribution function, where μ is the mean and σ is the standard deviation.

One special type of normal distribution, the standard normal distribution, has standard deviation 1 and mean 0. Here's a visual representation of it:

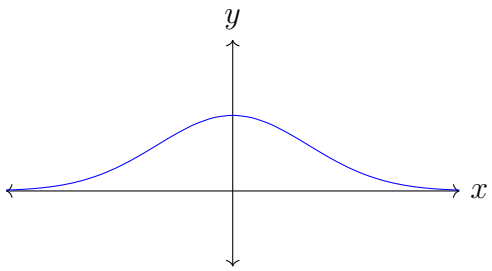


Figure 1. The standard normal distribution

Other normal distributions have this same general shape, but they can have different means and standard deviations. However, since the area must stay 1, if you increase the standard deviation, the curve will compress vertically, while if you decrease the standard deviation, the curve will expand vertically.

Normal distributions with mean μ and variance σ^2 can be represented with

$$\mathcal{N}(\mu, \sigma^2).$$

Notation. $\rightarrow$ in a distribution context Given two distributions Y_n and Y, if $Y_n \xrightarrow[n\to\infty]{} Y$, then for each interval $[a, b]$,

$$P[a \leq Y_n \leq b] \xrightarrow[n\to\infty]{} P[a \leq Y \leq b].$$

Definition 2.14. $\phi(a)$ function
$\phi(a)$ gives the area under a standard normal distribution from $-\infty$ to a. Specifically,

$$\phi(a) = \int_{-\infty}^{a} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}} dx$$

Definition 2.15. Moment
The kth moment of a distribution can be calculated with $\mathbb{E}[X^k]$. The kth central moment of a distribution is $\mathbb{E}[(X-\mu)^k]$. Moments are a way of characterizing a distribution, to the point where if all moments of two distributions are equal, then the distributions are identical. This is because moments capture the movement or tendency of data points away from a reference point. For example, when taking central moments, the first moment is expected value, the second is variance, the third is skewness, and the fourth is kurtosis. This fact will be useful in proving the Central Limit Theorem.

Definition 2.16. Moment Generating Function
A moment generating function, or MGF, provides another way to characterize distribution of a random variable. An MGF is commonly denoted with $M_X(t)$, and $M_X(t) = \mathbb{E}[e^{tX}]$. By expanding the Taylor series of e^{tX}, we can see why

it's called the moment generating function:

$$M_X(t) = \sum_{n=0}^{\infty} \frac{\mathbb{E}[X^n]}{n!} t^n$$

2.3. **Other definitions.**

Definition 2.17. Indicator Function
An indicator function, I_A, is 1 if A is true and 0 if A is false. Sometimes, this is represented with $I_A(x)$, which is 1 if $x \in A$ and 0 if $x \notin A$.

Definition 2.18. Double Factorial
Double factorials, $n!!$, represent the product of all positive integers less than or equal to n that have the same parity (oddness or evenness) as n. Formally,

$$n!! = \begin{cases} (n) \cdot (n-2) \cdot (n-4) \cdot \ldots \cdot (1) \text{ if } n \text{ is odd,} \\ (n) \cdot (n-2) \cdot (n-4) \cdot \ldots \cdot (2) \text{ if } n \text{ is even.} \end{cases}$$

For example, $4!! = 4 \cdot 2 = 8$, and $5!! = 5 \cdot 3 \cdot 1$.

3. The Weaker Central Limit Theorem

The central limit theorem says that if you sum independent random variables and normalize them accordingly, as you sum more and more random variables, the distribution of sums will converge to normal distribution. However, we'll start off with a weaker version of the CLT.

Theorem 3.1. *Given independent, identically distributed (i.i.d) random variables $X_1, X_2, \ldots, X_n$ with mean 0 and variance 1, as $n \to \infty$,*

$$\frac{X_1 + X_2 + \ldots + X_n}{\sqrt{n}} \xrightarrow[n\to\infty]{} \mathcal{N}(0,1).$$

In other words, the distribution of sums, when normalized through dividing by $\sqrt{n}$, will approach a standard normal distribution as n approaches infinity. We can extend this to say that the distribution of sums without normalization will approach a normal distribution as $n \to \infty$. Note that if our random variables don't have mean 0, like dice rolls, we can still apply the CLT by normalizing them through subtracting the mean from them. Additionally, although our proof will assume a variance of 1 for simplicity, the proof may be adapted to other variances, although it will get messier.

Instead of jumping straight into the proof, let's look at the fair dice roll example in action. As dice are independent, identically distributed random variables, and can be normalized to mean 0 and variance 1, the central limit theorem should apply. Below are some probability distributions for different values of n (number of dice rolled).

What's extraordinary is that this is one of the most basic applications of the central limit theorem. Because even the simplified version of the theorem applies to independent and identically distributed variables, we may apply it to unfair dice too.

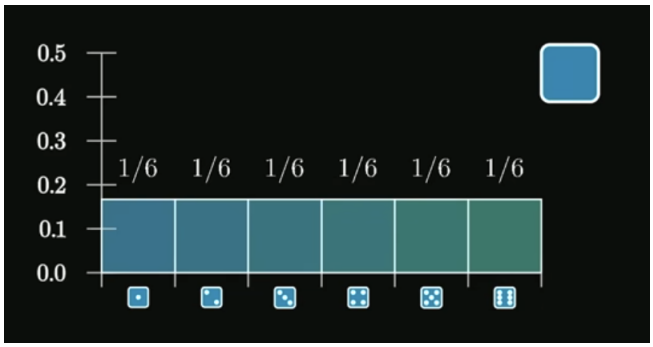


Figure 2. The fair die (n=1)

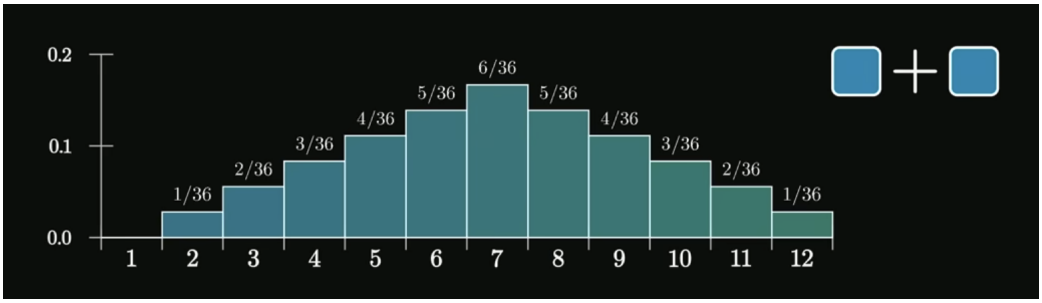


Figure 3. n=2

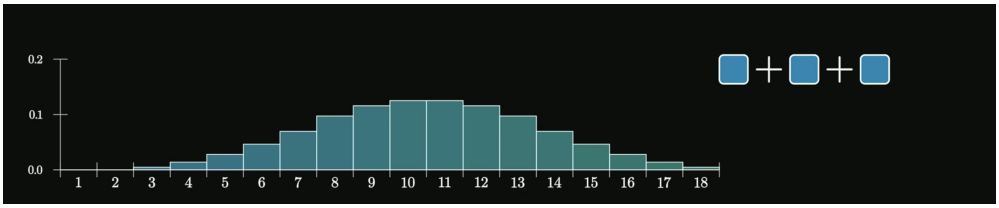


Figure 4. n=3

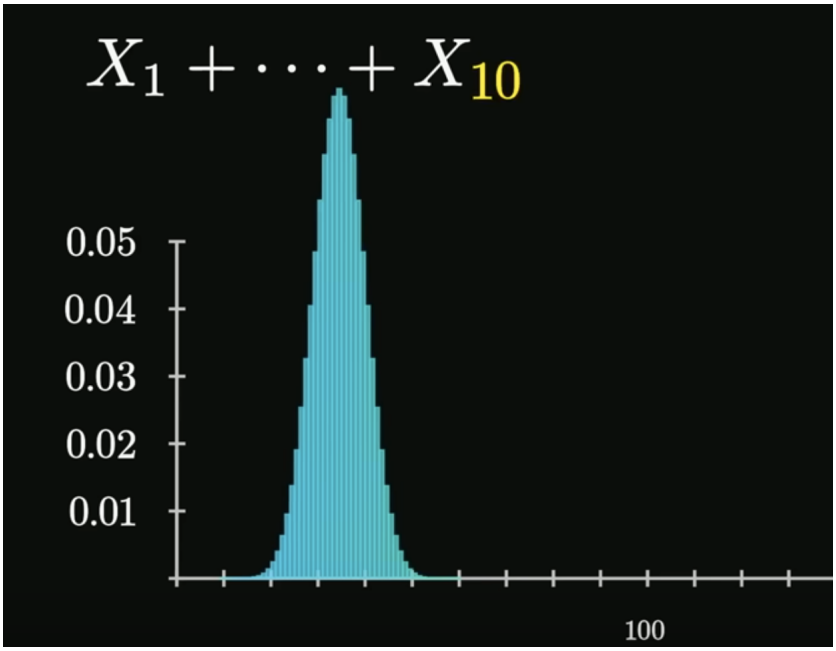


Figure 5. n=10

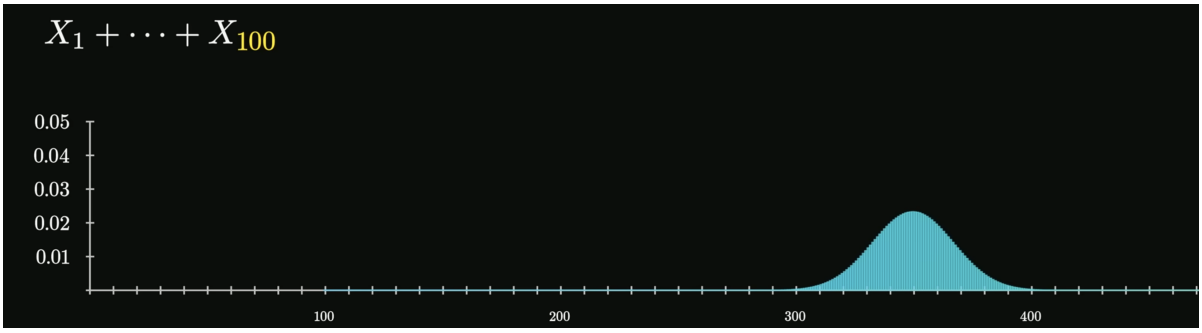


Figure 6. n=100

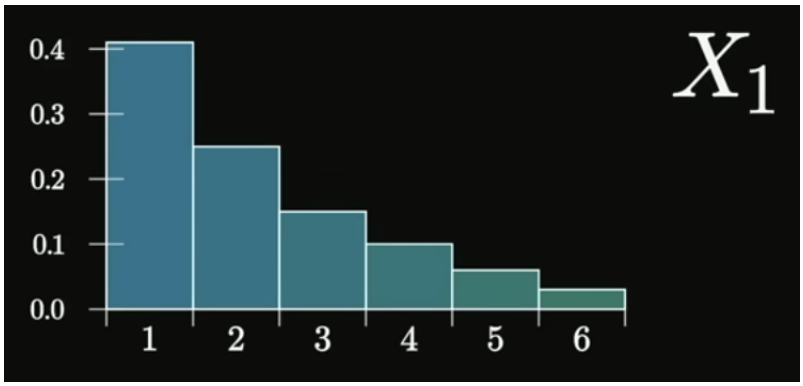


Figure 7. unfair die

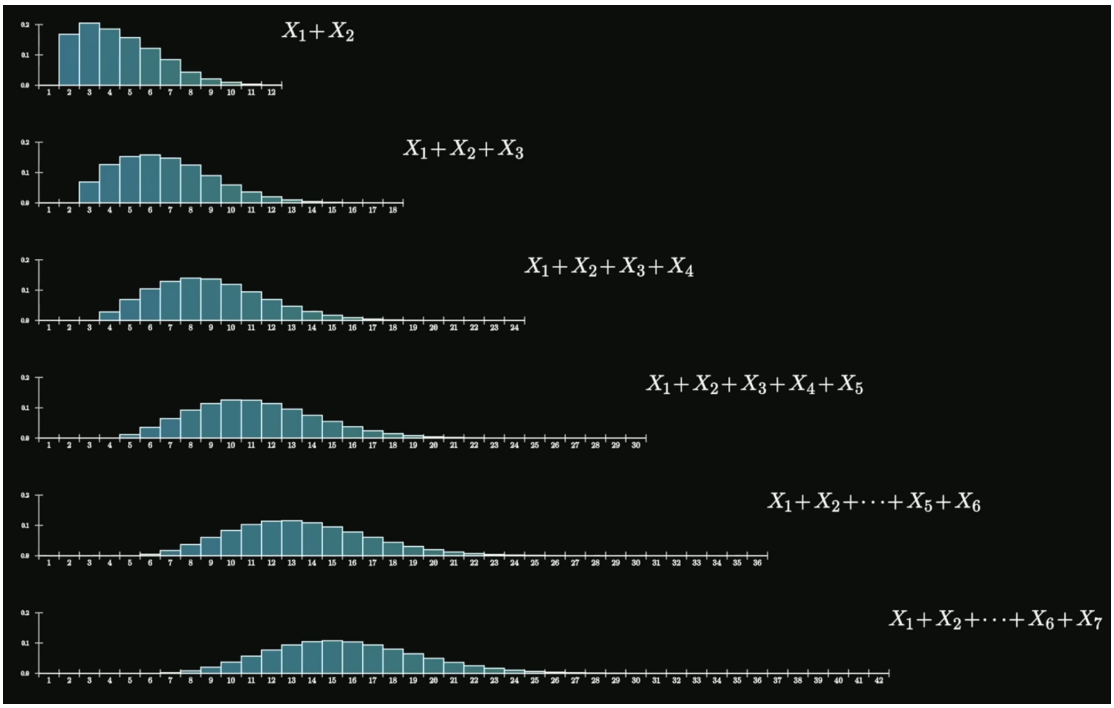


Figure 8. Convergence with unfair dice

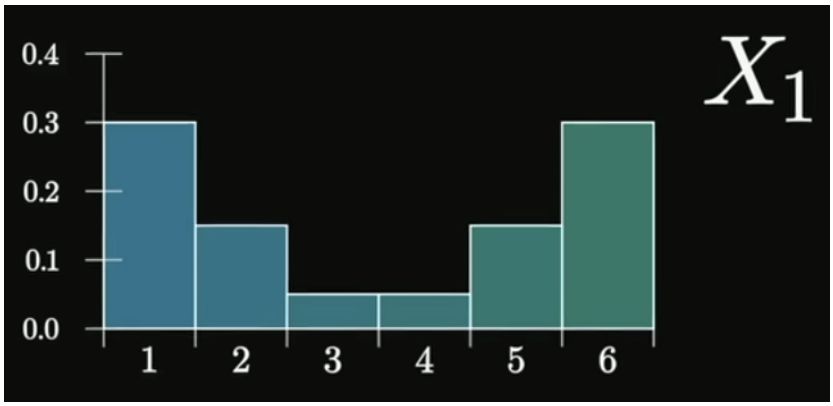


Figure 9. A different unfair die

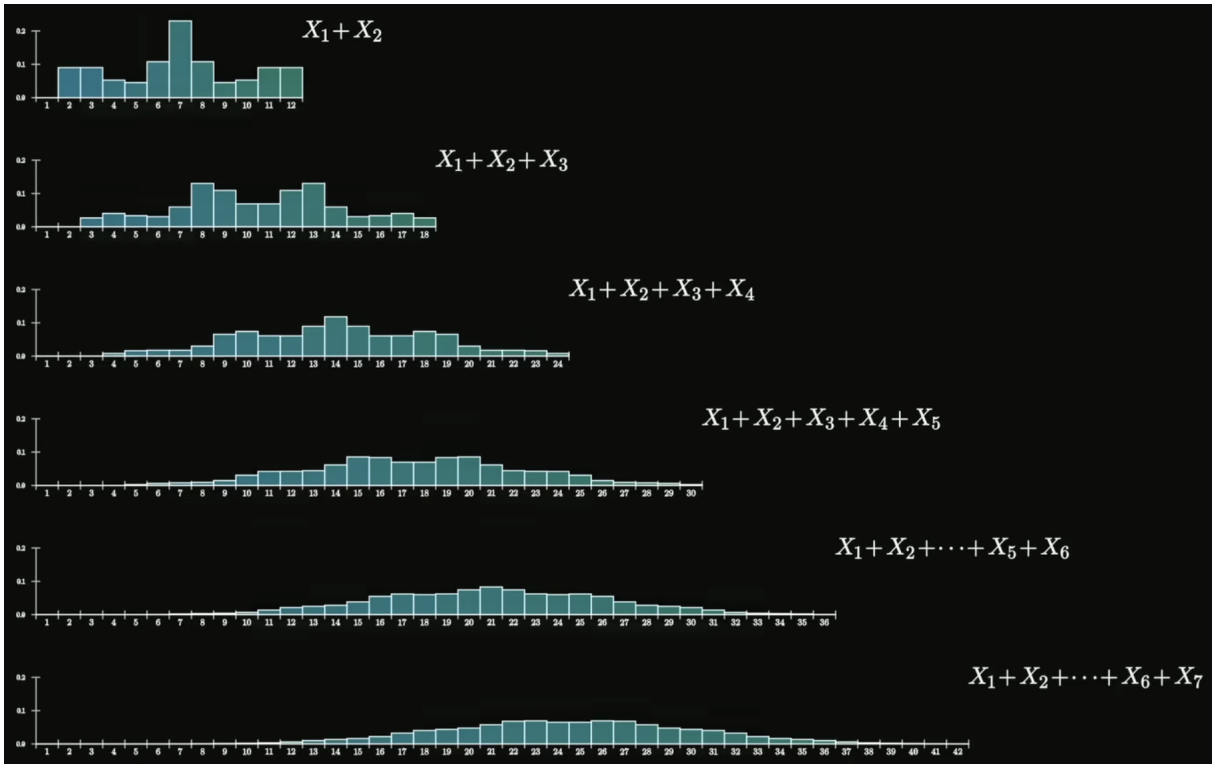


Figure 10. Convergence with different unfair dice

Further applications include sums of independent and not identically distributed variables, as if certain conditions are satisfied, for example Liapunov's condition which shall be explained later, the central limit theorem still applies. Unfortunately, I couldn't find any images of this, but to start, we will prove our weakened version of the central limit theorem.

4. Proof of the Weaker Central Limit Thorem

[Fil10]

Our strategy for proving this weaker central limit theorem will be to show that the moments of the standard normal variable and the moments of a normalized sum converge as $n \to \infty$. First, we'll find the moments of the standard normal variable.

4.1. **Finding the moments of the standard normal variable.**

Lemma 4.1.

$$for\ all\ k \in \mathbb{Z}, \mathbb{E}[Z^k] = \begin{cases} (k-1)(k-3)\dots(2)(0) = 0 & if\ k\ is\ odd, \\ (k-1)(k-3)\dots(1)(1) = (k-1)!! & if\ k\ is\ even. \end{cases}$$

Proof. The kth moment of a standard normal variable is

$$\begin{aligned} \mathbb{E}[Z^k] &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^k e^{-\frac{1}{2}x^2} dx \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^{k-1} (x e^{-\frac{1}{2}x^2}) dx. \end{aligned}$$

Using integration by parts, this can be simplified to

$$\begin{aligned} \mathbb{E}[Z^k] &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} (k-1) x^{k-2} e^{-\frac{1}{2}x^2} dx \\ &= (k-1) \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^{k-2} e^{-\frac{1}{2}x^2} dx \\ &= (k-1)\mathbb{E}[Z^{k-2}]. \end{aligned}$$

Therefore,

$$\mathbb{E}[Z^k] = \begin{cases} (k-1)(k-3)\dots(2)\mathbb{E}[Z^1] & \text{if } k \text{ is odd}, \\ (k-1)(k-3)\dots(1)\mathbb{E}[Z^0] & \text{if } k \text{ is even}. \end{cases}$$

We know that $\mathbb{E}[Z^1] = 0$ since Z has mean 0, and that $\mathbb{E}[Z^0] = \mathbb{E}[1] = 1$, so

$$\mathbb{E}[Z^k] = \begin{cases} (k-1)(k-3)\dots(2)(0) = 0 & \text{if } k \text{ is odd}, \\ (k-1)(k-3)\dots(1)(1) = (k-1)!! & \text{if } k \text{ is even}. \end{cases}$$

Another explanation for why the kth moment is 0 if k is odd is because the probability distribution function of Z is symmetric about $x = 0$, so the probability distribution of Z^k will be an even function if k is odd, and therefore will have mean 0. When k is even, Z^k will only have positive values, hence why $\mathbb{E}[Z^k]$ is positive. ∎

4.2. **Finding the moments of the sum of random variables.** The second and tougher part of our proof is to show that as $n \to \infty$, the moments of a normalized sum converge to the moments we just found.

Lemma 4.2. *As $n \to \infty$, for all $k \in \mathbb{Z}$,*

$$\mathbb{E}\left[\left(\frac{X_1 + X_2 + \dots + X_n}{\sqrt{n}}\right)^k\right] \to \begin{cases} 0 & if\ k\ is\ odd. \\ (k-1)!! & if\ k\ is\ even. \end{cases}$$

Proof. The kth moment of our sum is

$$\mathbb{E}[(X_1 + X_2 + \ldots + X_n)^k]$$

This is a lot harder to break down, so let's start with $k = 1$ and see if we can find a general pattern. Because of the linearity of expectations,

$$\begin{aligned}\mathbb{E}[X_1 + X_2 + \ldots + X_n] &= \mathbb{E}[X_1] + \mathbb{E}[X_2] + \ldots + \mathbb{E}[X_n] \\ &= 0 \text{ because our variables have mean 0.}\end{aligned}$$

When expanding for $k = 2$, we get

$$\begin{aligned}\mathbb{E}[(X_1 + X_2 + \ldots + X_n)^2] &= \mathbb{E}[\sum_{i=1}^{n}\sum_{j=1}^{n} X_i X_j] \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} \mathbb{E}[X_i X_j]\end{aligned}$$

Notice how there are two types of terms: ones in the form of $\mathbb{E}[X_i^2]$, and ones in the form of $\mathbb{E}[X_i X_j]$, where $i \neq j$. There are n of the $\mathbb{E}[X_i^2]$ terms, and $n(n-1)$ of the $\mathbb{E}[X_i X_j]$ terms. We can use the product rule of expectations on the $\mathbb{E}[X_i X_j]$ terms to get

$$\begin{aligned}\mathbb{E}[X_i X_j] &= \mathbb{E}[X_i]\mathbb{E}[X_j] \\ &= 0.\end{aligned}$$

So we only need to worry about the terms in the form of $\mathbb{E}[X_i^2]$. Our other definition of variance states that $\text{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$, so in our case,

$$\begin{aligned}1 &= \mathbb{E}[X^2] - 0 \\ \mathbb{E}[X^2] &= 1.\end{aligned}$$

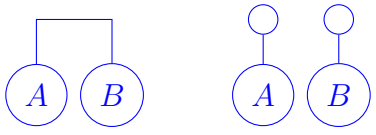
Therefore, since there are n of these terms,

$$\mathbb{E}[(X_1 + X_2 + \ldots + X_n)^2] = n \cdot 1 + n(n-1) \cdot 0 = n$$

However, to normalize this, we divide both sides by n:

$$\begin{aligned}\mathbb{E}[(X_1 + X_2 + \ldots + X_n)^2] &= n \\ \frac{\mathbb{E}[(X_1 + X_2 + \ldots + X_n)^2]}{n} &= 1 \\ \mathbb{E}\left[\left(\frac{X_1 + X_2 + \ldots + X_n}{\sqrt{n}}\right)^2\right] &= 1.\end{aligned}$$

So the second moment of $\frac{X_1+X_2+\ldots+X_n}{\sqrt{n}}$ is 1.
In fact, we can use a model to represent this relationship, which will be helpful later:



In the first diagram, A and B are connected, representing how the $\mathbb{E}[X_i^2]$ terms have two of the same random variable, and the diagram with the disconnected A and B represents how the $\mathbb{E}[X_iX_j]$ terms have two singletons (variables appearing only once in the term).

For $k = 3$, things get even more complex. Now there are 3 types of terms: ones in the form of $\mathbb{E}[X_i^3]$, ones in the form of $\mathbb{E}[X_i^2X_j]$, and ones in the form of $\mathbb{E}[X_iX_jX_l]$, with $i \neq j \neq l$. Again,

$$\mathbb{E}[X_iX_jX_h] = \mathbb{E}[X_i] \cdot \mathbb{E}[X_j] \cdot \mathbb{E}[X_l] = 0,$$

so we don't need to worry about those terms. In fact, whenever there is a singleton, we can ignore that term since its expectation is 0, so we will ignore terms in the form of $\mathbb{E}[X_i^2X_j]$ too. Since there are n terms in the form of $\mathbb{E}[X_i^3]$,

$$\mathbb{E}[(X_1 + X_2 + \ldots + X_n)^3] = \sum_{i=1}^{n} \mathbb{E}[X_i^3] = n\mathbb{E}[X_i^3],$$

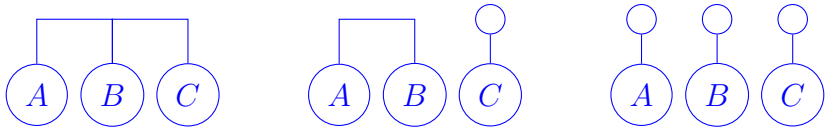
since our variables are identically distributed. Therefore,

$$\mathbb{E}\left[\left(\frac{X_1 + X_2 + \ldots + X_n}{\sqrt{n}}\right)^3\right] = \frac{1}{\sqrt{n}}\mathbb{E}[X_i^3].$$

However, since we want to look at what happens when $n \to \infty$, we look for its limit as $n \to \infty$. Because $\mathbb{E}[X_i^3]$ is constant,

$$\lim_{n\to\infty} \frac{1}{\sqrt{n}}\mathbb{E}[X_i^3] = 0,$$

and so the third moment is 0. Once again, another diagram can be created:



In this case, the first diagram represents $\mathbb{E}[X_i^3]$, the second $\mathbb{E}[X_i^2X_j]$, and the third $\mathbb{E}[X_iX_jX_l]$. For $k = 4$, there are two types of terms without singletons: $\mathbb{E}[X_i^2X_j^2]$ and $\mathbb{E}[X_i^4]$. There are n $\mathbb{E}[X_i^4]$ terms, and $3n(n-1)$ $\mathbb{E}[X_i^2X_j^2]$ terms, with a factor of 3 because there are 3 ways to pick 2 pairs in a group of 4. Therefore,

$$\mathbb{E}[(X_1 + X_2 + \ldots + X_n)^4] = n\mathbb{E}[X_i^4] + 3n(n-1),$$

$$\mathbb{E}\left[\left(\frac{X_1 + X_2 + \ldots + X_n}{\sqrt{n}}\right)^4\right] = \frac{\mathbb{E}[X_i^4]}{n} + \frac{3n(n-1)}{n^2},$$

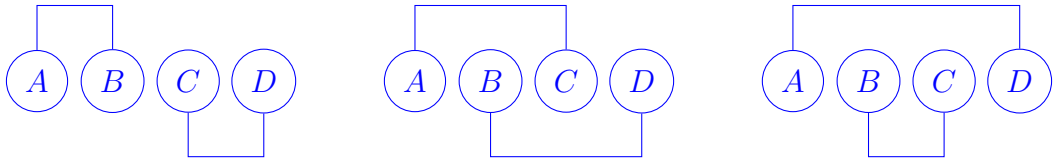
and

$$\begin{aligned}\lim_{n\to\infty} \mathbb{E}\left[\left(\frac{X_1 + X_2 + \ldots + X_n}{\sqrt{n}}\right)^4\right] &= \lim_{n\to\infty} \frac{\mathbb{E}[X_i^4]}{n} + \lim_{n\to\infty} \frac{3n(n-1)}{n^2} \\ &= 0 + 3 = 3.\end{aligned}$$

Notice how for all k values, only terms with exclusively squares are counted in the limit as $n \to \infty$. Other terms vanished as $n \to \infty$, or were 0 from the start. This is not a coincidence; as stated earlier, terms with singletons are 0 automatically, and terms with 3 or more of a

variable end up having less than $\frac{k}{2}$ unique variables, leading to where the coefficient for it has n with degree less than $\frac{k}{2}$ (since each unique variable is a new factor). Meaning, when we normalize by dividing by $n^{k/2}$ on both sides, these terms will have negative exponents of n, so as $n \to \infty$, these terms will approach 0.

As for the terms with exclusively squares, like $\mathbb{E}[X_i^2 X_j^2]$, n has an exponent of $\frac{k}{2}$, so when normalizing, the degree of n will be the same on the numerator and denominator. So when taking the limit as $n \to \infty$, the term will approach its coefficient. As there is no term with exclusively squares for odd k, those moments will approach 0 as $n \to \infty$. However, what about the moments for even k?



These represent the number of pair partitions in a group of 4, but also the coefficient of 3 we saw earlier when finding the 4th moment. In fact, the coefficient of the term with only square exponents is always the number of pair partitions in a group of k. This is because we need to find $k/2$ pairs, $X_{i_1}^2, X_{i_2}^2, \ldots, X_{i_{\frac{k}{2}}}^2$, so we're just splitting the group of k into pairs.

We're nearly done; all we need to do now is find the number of pair partitions in a group of k elements, $\{1, 2, 3, \ldots, k\}$ where k is even. Since we have $k-1$ options to pair a number with the element k, and $k-3$ options to pair a number with the new largest element, and so on, we know that the number of pair partitions in a group of k is

$$(k-1)(k-3)\ldots(1) = (k-1)!!.$$

Therefore, as $n \to \infty$,

$$\mathbb{E}\left[\left(\frac{X_1 + X_2 + \ldots + X_n}{\sqrt{n}}\right)^k\right] \to \begin{cases} 0 & \text{if } k \text{ is odd.} \\ (k-1)!! & \text{if } k \text{ is even.} \end{cases}$$

■

Finally, we'll combine these lemmas to prove our theorem.

Proof. Since for all $k \in \mathbb{Z}$, as $n \to \infty$, 4.2

$$\mathbb{E}\left[\left(\frac{X_1 + X_2 + \ldots + X_n}{\sqrt{n}}\right)^k\right] \to \begin{cases} 0 & \text{if } k \text{ is odd} \\ (k-1)!! & \text{if } k \text{ is even,} \end{cases}$$

and the moments of the standard normal variable are 4.1

$$\mathbb{E}[Z^k] = \begin{cases} 0 & \text{if } k \text{ is odd} \\ (k-1)!! & \text{if } k \text{ is even,} \end{cases}$$

all moments of the two random variables approach each other as $n \to \infty$, meaning as $n \to \infty$,

$$\frac{X_1 + X_2 + \ldots + X_n}{\sqrt{n}} \to \mathcal{N}(0, 1).$$

and so $X_1 + X_2 + \ldots + X_n$ approaches a normal distribution as $n \to \infty$. ■

5. Lindeberg's Central Limit Theorem

As it turns out, there are many other different forms of the Central Limit Theorem [Sch11]. The proofs for some other versions follow from the proof for Lindeberg's Central Limit Theorem, so we will start by looking at that one.

Theorem 5.1. *Given n independent random variables with means 0 denoted $X_{n_1}, X_{n_2}, \ldots, X_{n_n}$ and respective variances $\sigma^2_{n_1}, \sigma^2_{n_2}, \ldots, \sigma^2_{n_n}$, denote the sum of the random variables, $X_{n_1} + X_{n_2} + \ldots + X_{n_n}$, with S_n, its variance with τ_n^2.*

If the Lindeberg condition $L_n(\epsilon)$ is satisfied, or in other words, if for every $\epsilon > 0$,

$$L_n(\epsilon) = \frac{1}{\tau_n^2} \sum_{i=1}^{n} \mathbb{E}[X^2_{n_i} I_{(X_{n_i} | \geq \epsilon \tau_n)}] \to 0 \text{ as } n \to \infty.$$

then for every real a,

$$P\left(\frac{S_n}{\tau_n} \leq a\right) - \phi(a) \to 0 \text{ as } n \to \infty.$$

Essentially, the theorem's saying that if the Lindeberg condition is satisfied, as $n \to \infty$, then the probability that the sum of n random variables, scaled to a standard normal distribution by dividing by the standard deviation, is less than a real number a is equal to the probability that a standard normal variable is less than a. Best put, as you sum more and more random variables, the distribution of sums will match a normal curve. Note that this time, the random variables don't need to be identically distributed. On the other hand, the Lindeberg condition is the sum of the expected value of each random variable squared only if it is more than ϵ standard deviations away from the mean, all divded by the variance of the sum. Something helpful to keep in mind is that $\frac{S_n}{\tau_n}$ is just the sum divided by the standard deviation, so it's the distribution normalized to have standard deviation 1. Note that because S_n is the sum of random variables, its variance is the sum of the variances of those random variables, meaning $\tau_n^2 = \sigma^2_{n_1} + \sigma^2_{n_2} + \ldots + \sigma^2_{n_n}$.

First, we'll start with a bold assumption that is proved in the cited source's appendix. Let a smooth function f mean that f is bounded and has 3 bounded, continuous derivatives. For certain functions f, including all smooth f,

$$\mathbb{E}[f(S_n/\tau_n)] - \mathbb{E}[f(Z)] \xrightarrow[n\to\infty]{} 0.$$

Let $f_a(x) = I_{(-\infty,a]}(x)$. If the convergence above would also work for the function $f_a(x)$, then we would obtain

$$\mathbb{E}[f_a(S_n/\tau_n)] - \mathbb{E}[f_a(Z)] = P\left(\frac{S_n}{\tau_n} \leq a\right) - \phi(a) \xrightarrow[n\to\infty]{} 0,$$

which would mean our proof would be quite short. However, we'll have to work to show that the convergence is true for $f_a(x)$.

Our strategy will consist of sandwiching $f_a(x)$ between two smooth functions, but to do that, we'll start with sandwiching a smooth function between two $f_a(x)$: $f_a(x)$ and $f_{a+\delta}(x)$, where $\delta > 0$. This would mean our smooth function $f(x)$ must satisfy

$$f_a(x) \leq f(x) \leq f_{a+\delta}(x) \text{ for all } x \in \mathbb{R}.$$

More specifically, $f(x) = 1$ for $x \leq a$, $f(x) = 0$ for $x > a + \delta$, and $0 \leq f(x) \leq 1$ for $a < x \leq a + \delta$, along with having 3 bounded, continuous derivatives. Although the exact function doesn't matter, we would like to show that such a smooth function always exists.

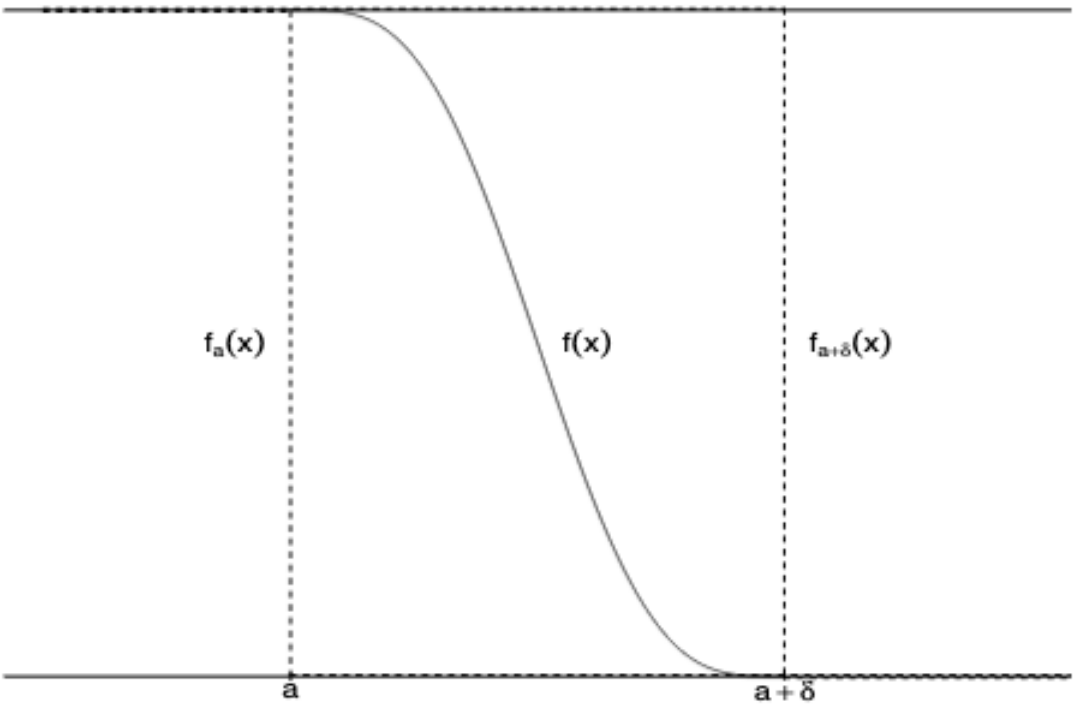


Figure 11. Sandwiching the smooth function

Lemma 5.2. *A smooth function $f(x)$ always exists such that*

$$f_a(x) \leq f(x) \leq f_{a+\delta}(x) \text{ for all } x \in \mathbb{R}.$$

Proof.

$$g(x) = 1 - 140\left(\frac{1}{4}x^4 - \frac{3}{5}x^5 + \frac{1}{2}x^6 - \frac{1}{7}x^7\right)$$

is a smooth function that has $g(0) = 1, g(1) = 0, g'(0) = g'(1) = 0, g''(0) = g''(1) = 0,$ and $g'''(0) = g'''(1) = 0$. This means if we have $a = 0$, $\delta = 1$, and

$$f(x) = \begin{cases} 1 & x \leq a \\ g(x) & a \leq x \leq a + \delta \\ 0 & x > a + \delta, \end{cases}$$

$g(x)$ and its first three derivatives will connect with the other segments, meaning $f(x)$ is one of the smooth functions we are looking for. To generalize for any a and δ,

$$g(x) = \delta\left(1 - 140\left(\frac{1}{4}(x-a)^4 - \frac{3}{5}(x-a)^5 + \frac{1}{2}(x-a)^6 - \frac{1}{7}(x-a)^7\right)\right),$$

and like before,

$$f(x) = \begin{cases} 1 & x \leq a \\ g(x) & a \leq x \leq a + \delta \\ 0 & x > a + \delta. \end{cases}$$

Therefore, a smooth function that satisfies

$$f_a(x) \leq f(x) \leq f_{a+\delta}(x) \text{ for all } x \in \mathbb{R}$$

exists for all a and δ. ■

Now, we can get started on our proof.

Proof. Following 5.2,

$$f(x) \leq f_{a+\delta}(x),$$

meaning

$$f(x+\delta) \leq f_{a+\delta}(x+\delta).$$

Additionally, $f_a(x) = f_{a+\delta}(x+\delta)$, so by combining these, we obtain

$$f(x+\delta) \leq f_a(x) \leq f(x),$$

meaning we have sandwiched $f_a(x)$ between 2 smooth functions.

Because this inequality holds for all x, we also know that

$$\mathbb{E}[f(x+\delta)] \leq \mathbb{E}[f_a(x)] \leq \mathbb{E}[f(x)].$$

When applying this inequality for both $\frac{S_n}{\tau_n}$ and Z, we get the following inequalities:

$$\mathbb{E}[f(\frac{S_n}{\tau_n}+\delta)] \leq \mathbb{E}[f_a(\frac{S_n}{\tau_n})] \leq \mathbb{E}[f(\frac{S_n}{\tau_n})]$$

$$\mathbb{E}[f(Z+\delta)] \leq \mathbb{E}[f_a(Z)] \leq \mathbb{E}[f(Z)]$$

However, recall that f is a smooth function and we can apply the convergence we showed earlier:

$$\begin{array}{ccccc} \mathbb{E}[f(\frac{S_n}{\tau_n}+\delta)] & \leq & \mathbb{E}[f_a(\frac{S_n}{\tau_n})] & \leq & \mathbb{E}[f(\frac{S_n}{\tau_n})] \\ \wr\!\wr & & & & \wr\!\wr \\ \mathbb{E}[f(Z+\delta)] & \leq & \mathbb{E}[f_a(Z)] & \leq & \mathbb{E}[f(Z)], \end{array}$$

where $\wr\!\wr$ represents that the two terms above and below it will converge as $n \to \infty$.

Since $f(Z+\delta) - f(Z) = 0$ for $Z \notin [a-\delta, a+\delta]$ and $|f(Z+\delta) - f(Z)| \leq 1$ otherwise, we have

$$\begin{aligned} |\mathbb{E}[f(Z+\delta)] - \mathbb{E}[f(Z)]| &= |\mathbb{E}[(f(Z+\delta) - f(Z))I_{[a-\delta \leq Z \leq a+\delta]}]| \\ &\leq |\mathbb{E}[I_{[a-\delta \leq Z \leq a+\delta]}]| \\ &= P(a-\delta \leq Z \leq a+\delta) \\ &= \phi(a+\delta) - \phi(a-\delta) \\ &\leq \phi(\delta) - \phi(-\delta) \\ &\leq 2\delta\phi(0) \\ &= \frac{2\delta}{\sqrt{2\pi}}. \end{aligned}$$

There are some pretty confusing steps here; the first step comes from how $f(Z+\delta) - f(Z) = 0$ for $Z \notin [a-\delta, a+\delta]$, so we can just multiply by 0 and not change anything anyways in

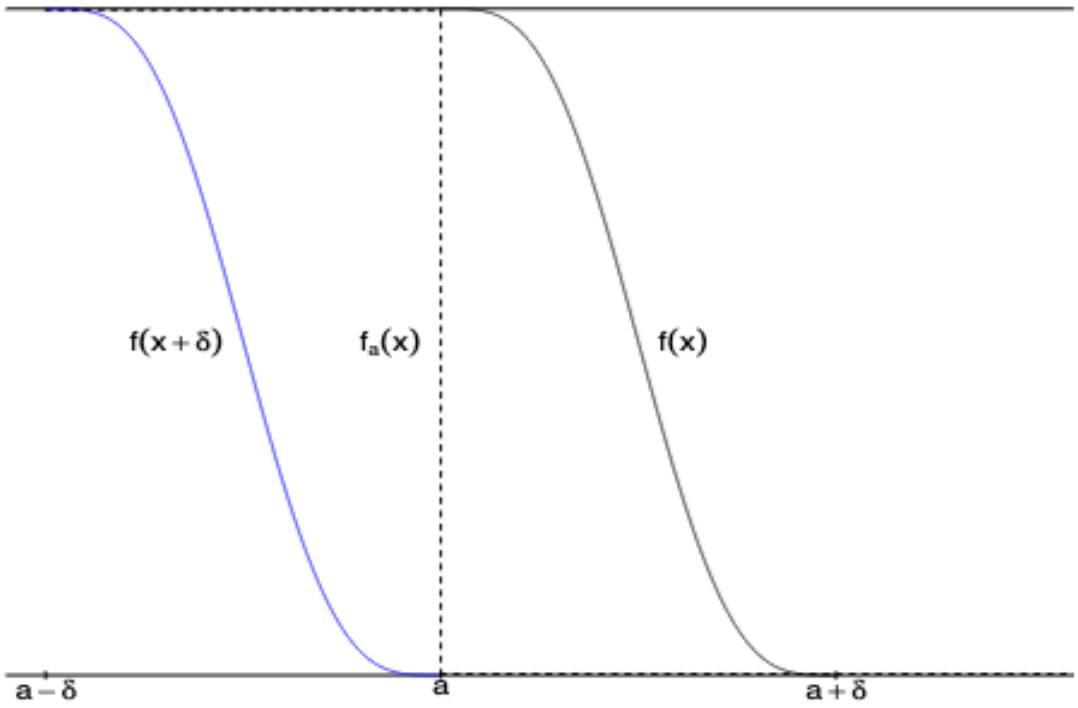


Figure 12. Sandwiching $f_a(x)$

that range. The second because $f(Z+\delta) - f(Z) \leq 1$ for all Z. Recall that ϵ is part of the Lindeberg condition; for all $\epsilon > 0$,

$$L_n(\epsilon) = \frac{1}{\tau_n^2} \sum_{i=1}^{n} \mathbb{E}[X_{n_i}^2 I_{(X_{n_i} | \geq \epsilon \tau_n)}] \to 0 \text{ as } n \to \infty.$$

If we were to take $\delta = \epsilon \frac{\sqrt{2\pi}}{6}$, our bound would become $\frac{\epsilon}{3}$, no matter what f is. From the above diagram, we get

$$\left| \mathbb{E}\left[f_a\left(\frac{S_n}{\tau_n} \right) \right] - \mathbb{E}[f_a(Z)] \right| \leq \frac{\epsilon}{3} + \frac{\epsilon}{3} + \frac{\epsilon}{3} = \epsilon.$$

Since we can do this for all $\epsilon > 0$, we have shown that

$$\left| \mathbb{E}\left[f_a\left(\frac{S_n}{\tau_n} \right) \right] - \mathbb{E}[f_a(Z)] \right| \xrightarrow[n\to\infty]{} 0,$$

meaning

$$P\left(\frac{S_n}{\tau_n} \leq a \right) - \phi(a) \xrightarrow[n\to\infty]{} 0.$$

■

6. Liapunov's Central Limit Theorem

Following the Lindeberg Central Limit Theorem, we can establish another version of the central limit theorem: Liapunov's Central Limit Theorem.

Theorem 6.1. *Let $Y_1, Y_2, \ldots, Y_n$ be independent random variables with means $\mu_1, \mu_2, \ldots, \mu_n$, variances $\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2$ and finite absolute central moments, meaning $\mathbb{E}[|Y_i - \mu_i|^3] < \infty$. If Liapunov's condition,*

$$l_n = \frac{1}{\tau^3} \sum_{i=1}^{n} \mathbb{E}[|Y_i - \mu_i|^3] \to 0 \text{ as } n \to \infty,$$

then

$$P\left(\frac{\sum Y_i - \sum \mu_i}{\sqrt{\sum \sigma_i^2}} \leq a \right) - \phi(a) \to 0 \text{ as } n \to \infty.$$

This means that if the sum of absolute central moments divided by the standard deviation cubed approaches 0 as $n \to \infty$, then the normalized sum will approach the standard normal distribution as $n \to \infty$. The sum is normalized by subtracting all the means, giving it mean 0, and also by dividing by the standard deviation of the sum to give the normalized sum a standard deviation of 1.

Proof. To show that Liapunov's Central Limit Theorem is true, we will show that Liapunov's condition implies Lindeberg's condition. Let $X_i = Y_i - \mu_i$, which means it's just Y_i with mean 0 and the same variance. If $l_n \xrightarrow[n\to\infty]{} 0$, then since

$$\begin{aligned} l_n &= \frac{1}{\tau_n^3} \sum_{i}^{n} \mathbb{E}[|X_i|^3] \\ &\geq \frac{1}{\tau_n^3} \sum_{i}^{n} \mathbb{E}[|X_i|^3 I_{[|X_i| \geq \epsilon \tau_n]}] \\ &\geq \frac{\epsilon \tau_n}{\tau_n^3} \sum_{i}^{n} \mathbb{E}[|X_i|^2 I_{[|X_i| \geq \epsilon \tau_n]}] \\ &= \epsilon L_n(\epsilon) \end{aligned}$$

and $\epsilon > 0$, $L_n(\epsilon) \to 0$ as $n \to \infty$, meaning the Lindeberg condition is satisfied if the Liapunov condition is satisfied. Therefore,

$$P\left(\frac{\sum Y_i - \sum \mu_i}{\sqrt{\sum \sigma_i^2}} \leq a \right) - \phi(a) \to 0 \text{ as } n \to \infty.$$

■

7. Why Does the Central Limit Theorem Matter?

The Central Limit Theorem is important in statistics because it opens the door for statisticians to assume distributions are essentially normal. This allows the use of tools only available for normal distributions; for example, you may calculate area within a certain number of standard deviations as mentioned in our definition for normal distributions. Consider the problem of means in the introduction, where you want to find the probability the mean of a sum lies within a certain range. Given you have a sufficient amount of random variables, you could assume the distribution of sums is normal, and by extension the distribution of means. Therefore, you could estimate the probability that the mean lies within a

certain range given the standard deviation of the distribution. Another use case is reverse engineering the mean and variance of a random variable with an unknown distribution.

8. Further Questions

Although we have studied the behavior of the sum as $n \to \infty$, you might be wondering what happens when we study smaller n. Well, one area where you might be interested in is how fast the sum converges to a normal distribution. Unlike the theorem itself, this heavily depends on the random variable you choose; for example, 10 was an unfair die that looked like it approached a normal distribution much slower than 8, at least for the first few values of n. Additionally, you could research what approaching normal means; how can you tell one sum converges slower than another? Lots of stuff in this area is estimation, so perhaps you could be the one to make better estimations or even precise answers.

References

[Fil10] Yuval Filmus. Two proofs of the central limit theorem. *Recuperado de http://www. cs. toronto. edu/yuvalf/CLT. pdf*, 2010.

[Sch11] FW Scholz. Central limit theorems and proofs. *Lecture Notes, Univ. Washington*, 2011.