The Central Limit Theorem (CLT) - Overview, Proof, Examples

Alexandre Acra

November 11, 2020

Abstract

In this paper, we state and prove the Central Limit Theorem. The approach we have taken is to assume little prior knowledge, and review the basics and main results of probability and random variables from first axioms and definitions. We construct the preparatory concepts necessary for our proof, such as moments and moment-generating functions, as these are central to the approach of the proof. We prove most – but not all – results that we state. We also give where possible practical insights, and describe some real-world applications of this very pervasive result.

1. INTRODUCTION

Many things that can be measured in Nature and Society appear to be distributed according to a "bell-shaped curve", formally known as the Normal or Gaussian distribution. This prompts us to investigate why this probability distribution appears so pervasively across so many domains.

We start by reviewing some of the basic definitions and results of probability and random variables. We then describe the standard normal distribution, which is central to the CLT. Before heading into the general proof, we illustrate a special case of the Bernoulli distribution converging to a standard normal.

We then approach the proof of the CLT by introducing moment generating functions (MGF) of random variables and reasoning on their convergence, and we will "accept" – but not prove – the ability to reason back on the convergence of the probability distributions themselves from the convergence of their MGFs.

We follow the proof with a few entertaining examples of the CLT, and we end the paper by making a final point on the various modes of convergence of random variables, in order to put the mode of convergence at play in the CLT in that broader context.

2. DEFINITIONS AND GENERAL RESULTS

In this section, we will recall a few definitions and results pertaining to probability spaces, events and independence, the axioms of probability, random variables, probability density or mass functions, cumulative distribution functions, expectations and variances of random variables and of combinations or functions of random variables. We will also quickly cover the concept of convolution of distributions, as the distribution of a sum of independent random variables.

2.1. PROBABILITY SPACES

When an experiment results in uncertain outcomes, probability theory quantifies the likelihood of occurrence of outcomes in the following sense: we want to measure the limit of the frequency of appearance of an outcome as a fraction of the total, if the experiment is repeated a large number of times tending to infinity. Probability theory builds on set theory by mapping experimental outcomes to sets, and by attributing to each set a non-negative "measure" $\mathbb{P}()$ that captures its

expected frequency of occurrence as a fraction of the total occurrences in an infinitely large numbers of runs of the experiment.

Sample Space. The sample space of an experiment is the set of all possible outcomes, and we can denote it by Ω. An event E in the sample space is a subset of the set of all possible outcomes. Since a set can be the union of other sets, or the intersection of other sets, or the complement of another set, we naturally want to define some algebraic rules governing the probability measures $\mathbb{P}(E)$ attributed to events such as E, when the events are made of intersections or unions or complements of other events.

Disjoint or Mutually Exclusive Events. As a definition, we say that two events are mutually exclusive or disjoint if the occurrence of one excludes the occurrence of the other in the same experimental run, i.e., E and F are mutually exclusive if $E \cap F = \emptyset$. We would intuitively want the probability of their simultaneous occurrence to be 0, i.e.:

$$E \cap F = \emptyset \implies \mathbb{P}(E \cap F) = 0$$

Also, if the sample space Ω captures the exhaustive set of all possible outcomes of a run of the experiment, then we would want the probability measure associated with Ω to correspond to "certainty", as – by definition – *some event* E is bound to be the result of each experimental run. We therefore expect to have:

$$\mathbb{P}(\Omega) = 1$$

2.2. AXIOMS OF PROBABILITY

The axioms of probability were defined by Kolmogorov, and there are a few variations to picking which rules are made axioms, so we choose the following:

- $0 \leq \mathbb{P}(E) \leq 1$, for any event E in Ω.
- $\mathbb{P}(\Omega) = 1$.
- For any sequence of countably many mutually exclusive events E_1, E_2, ..., we have:

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(E_i)$$

It can be shown that the following properties are consequences of the Kolmogorov axioms above:

- $\mathbb{P}(\Omega \setminus E) = 1 - \mathbb{P}(E)$.
- $\mathbb{P}(\emptyset) = 0$.
- $E \subseteq F \implies \mathbb{P}(E) \leq \mathbb{P}(F)$.
- $\mathbb{P}(E \cup F) = \mathbb{P}(E) + \mathbb{P}(F) - \mathbb{P}(E \cap F)$
- $\mathbb{P}(E \cup F) = \mathbb{P}(E) + \mathbb{P}(F)$ for disjoint events E and F.

2.3. INDEPENDENCE AND CONDITIONAL PROBABILITY

- **Independent Events.** We define two probability events E and F to be independent if the probability of their simultaneous occurrence during a run of the experiment is equal to the product of their individual probabilities of occurrence:

$$\mathbb{P}(E \cap F) = \mathbb{P}(E) \cdot \mathbb{P}(F)$$

- **Conditional Probability.** We define the conditional probability of an event E given another event F as:

$$\mathbb{P}(E|F) = \frac{\mathbb{P}(E \cap F)}{\mathbb{P}(F))}$$

- We note that if two events E and F are independent, then we have:

$$\mathbb{P}(E|F) = \frac{\mathbb{P}(E \cap F)}{\mathbb{P}(F))} = \frac{\mathbb{P}(E) \cdot \mathbb{P}(F)}{\mathbb{P}(F))} = \mathbb{P}(E)$$

  This is consistent with the intuition we might have of independence, i.e., the fact that event F has occurred does not change the probability of the event E occurring, and vice-versa.

2.4. RANDOM VARIABLES

- A random variable is a function X with domain Ω and co-domain $\mathbb{R}$, i.e., it assigns a real value to any event in the sample space of the probability space.

  For a simple example, in an experiment of flipping coins, a random variable might assign the value 1 to an outcome of Heads and 0 to an outcome of Tails. In this case, the range of X is the set $\{0, 1\} \subset \mathbb{R}$.

- **Discrete Random Variable.** A random variable is discrete if it takes on at most countably many real values.

  Some common examples of discrete random variables are the sum of the numbers from two die rolls, whose image is $\{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$, or the number of coin flips until the first Head appears, whose image is $\mathbb{N}$.

- **Continuous Random Variable.** A random variable is continuous if it takes on uncountably many real values.

  Examples of continuous random variables would be measurements of continuous physical quantities, such as heights of trees, or intensity of electric currents, or time duration between successive random events such as particle emissions by radioactive materials.

2.5. DISTRIBUTION FUNCTIONS

- **Probability Mass Function.** For a discrete random variable, the probability mass function (PMF) is a function that attributes to each value x that the random variable can take the probability that the random value takes that value x. This corresponds to the probability of the event in Ω whose occurrence results in the random variable X taking that value x. We note this as:

$$p_X(x) = \mathbb{P}(X = x) \geq 0.$$

  In the example of rolling two dice, with X being the random variable equal to their sum, we have the following:

$$p_X(4) = \mathbb{P}(X = 4) = \mathbb{P}((1, 3) \cup (2, 2) \cup (3, 1)) = \frac{1}{36} + \frac{1}{36} + \frac{1}{36} = \frac{1}{12}.$$

- **Probability Density Function.** For a continuous random variable, the probability density function (PDF) is a function $f_X(x)$ that attributes – via integration – to each "measurable" subset $B \subseteq \mathbb{R}$ the probability that the random variable X takes values that fall within that subset B. We note this as:

$$\mathbb{P}(X \in B) = \int_B f_X(x)dx \geq 0$$

  While the concepts of measures and measurable subsets are left out as the topics of a more advanced paper describing Lebesgue integration on possibly non-continuous functions, we will

stick in this paper with the simple notion of $f_X(x)$ being a Riemann-integrable function, and the sets B over which we seek probabilities of X falling being intervals or unions of intervals of $\mathbb{R}$.

- **Cumulative Distribution Function.** A cumulative distribution function (CDF) of a random variable X is defined as:

$$F_X(x) = \mathbb{P}(X \leq x)$$

In the case of a continuous random variable, we have:

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

In this continuous case, we see that the probability density function is the derivative of the cumulative distribution function.

The CDF of a random variable is the function which – measured at specific values – gives statistical insights into the shape of the distribution via quantiles.

- **Quantiles.** A quantile of a CDF distribution $F_X(x)$ is the value of x at which the probability of the random variable being less than or equal to a specified quantity is achieved. These are sometimes expressed as percentiles, when the desired probability is expressed as a percentage. Notable examples of quantiles are:

  – **Median:** The median is the point m such that $\mathbb{P}(X \leq m) = 0.5$.

  – **First Quartile:** The first quartile is the point q_1 such that $\mathbb{P}(X \leq q_1) = 0.25$.

  – **Third Quartile:** The third quartile is the point q_3 such that $\mathbb{P}(X \leq q_3) = 0.75$.

  – **Inter-Quartile Range:** The inter-quartile range is the difference between the third and the first quartiles, i.e., $IQR = q_3 - q_1$. It is one of the measures of spread of the distribution of a random variable.

  – **Ninth Decile:** The ninth decile is the point P_{90} such that $\mathbb{P}(X \leq P_{90}) = 0.90$.

  – **Ninety-Ninth Percentile:** The ninety-ninth percentile is the point P_{99} such that $\mathbb{P}(X \leq P_{99}) = 0.99$.

- **PMF or PDF of Sums of Independent Random Variables - Convolution.** We examine the PMF or PDF $p_Z(z)$ of the sum $Z = X + Y$ of two independent random variables X and Y, as a function of the PMFs or PDFs $p_X(x)$ and $p_Y(y)$ of the individual random variables X and Y. For a given z we have:

$$p_Z(z) = \mathbb{P}(Z = z) = \mathbb{P}(X + Y = z)$$

  – Discrete Case:

$$\mathbb{P}(X + Y = z) = \sum_{n=\infty}^{\infty} \mathbb{P}(X = n, Y = z - n)$$

By independence of X and Y, we have:

$$\mathbb{P}(X = n, Y = z - n) = \mathbb{P}(X = n) \cdot \mathbb{P}(Y = z - n) = p_X(n) \cdot p_Y(z - n),$$

so that:

$$p_Z(z) = \sum_{n=-\infty}^{\infty} p_X(n) \cdot p_Y(z-n)$$

The sum above is defined as the *convolution* of the two PMFs p_X and p_Y, and is denoted as:

$$p_Z(z) = \sum_{n=-\infty}^{\infty} p_X(n) \cdot p_Y(z-n) = (p_X * p_Y)(z).$$

A simple illustrative case of convolution of distributions is calculating the PMF of $Z = X+Y$ where X and Y are the readings of two six-sided dice thrown independently of each other. We have:

$$p_X(x) = \frac{1}{6} \text{ for } x \in \{1,\ 2,\ 3,\ 4,\ 5,\ 6\}$$

$$p_Y(y) = \frac{1}{6} \text{ for } y \in \{1,\ 2,\ 3,\ 4,\ 5,\ 6\}$$

$$\begin{aligned} p_Z(7) &= \sum_{i=1}^{6} p_X(i) p_Y(7-i) \\ &= p_X(1)p_Y(6) + p_X(2)p_Y(5) + p_X(3)p_Y(4) + p_X(4)p_Y(3) + p_X(5)p_Y(2) + p_X(6)p_Y(1) \\ &= \frac{1}{6} \cdot \frac{1}{6} \cdot 6 \\ &= \frac{1}{6}. \end{aligned}$$

– Continuous Case:

We have a similar concept of convolution defined as follows for two functions f and g that are integrable over $\mathbb{R}$:

$$(f * g)(x) = \int_{-\infty}^{\infty} f(t) \cdot g(x-t)dt = \int_{-\infty}^{\infty} f(x-t) \cdot g(t)dt$$

And there is a similar result that we will not prove here, that the PDF $P_Z(z)$ of $Z = X+Y$, with X and Y continuous independent random variables, is given by the convolution of the PDFs p_X and p_Y, respectively of X and Y:

$$f_Z(z) = (f_X * f_Y)(z) = \int_{-\infty}^{\infty} f_X(t) \cdot f_Y(z-t)dt$$

2.6. EXPECTATION, VARIANCE, MOMENTS

- **Expected Value or Mean.** We define the Expected Value or Mean $\mathbb{E}(X)$ of a random variable X – also denoted as μ_X – as the weighted sum of the values that X can take, with the weights representing the respective probabilities that X can take the given values. In the continuous case, the definition is a little more delicate since positive probabilities are only assigned to sets of positive measure, i.e., to intervals of non-zero length. So the expected value is defined as the integral of the product of the variable x by the PDF value $p_X(x)$ at x. Specifically, we have:

  – Discrete Case:

$$\mathbb{E}(X) = \mu_X = \sum_{x=-\infty}^{\infty} x \cdot p_X(x) = \sum_{x:p_X(x)>0} x \cdot p_X(x)$$

– Continuous Case:

$$\mathbb{E}(X) = \mu_X = \int_{-\infty}^{\infty} x \cdot f_X(x) dx$$

- **"Distributive" Property of Expectation.** It is straightforward to show from the definitions that the expected value satisfies the following property. If a and b are real constants, then $aX + b$ is a random variable that takes value $ax + b$ whenever X takes value x, and we have:

$$\mathbb{E}(aX + b) = \mu_{aX+b} = a\mathbb{E}(X) + b = a\mu_X + b$$

We therefore define the following:

- **Variance.** The variance of a random variable is the expected value of the squared difference of the random variable and its mean.

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

- **Alternative Expression of Variance.** We have:

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

- **Effect on Variance of Translation and Scaling.** It is easy to see that for a and b real constants we have:

$$\begin{aligned}
\mathrm{Var}(aX + b) &= \mathbb{E}[(aX + b - a\mu_X - b)^2] \\
&= \mathbb{E}[(aX - a\mu_X)^2] \\
&= \mathbb{E}[a^2(X - \mu_X)^2] \\
&= a^2\mathbb{E}[(X - \mu_X)^2] \\
&= a^2 \,\mathrm{Var}(X).
\end{aligned}$$

The variance is therefore indifferent to a uniform translation of the random variable by a constant b, and when the random variable is scaled by a constant a, the variance is scaled by a^2, i.e., by the square of the scaling constant.

- **Standard Deviation.** We define the standard deviation σ_X of a random variable X as the square root of its variance.

$$\sigma_X = \sqrt{\mathrm{Var}(X)}$$

We then have:

$$\sigma_{aX+b} = \sqrt{\mathrm{Var}(aX + b)} = |a| \cdot \sigma_X$$

We note that, unlike the variance, the standard deviation and the mean are measured in the same units as those of the random variable itself. When they are well defined (i.e., finite), the variance and the standard deviation are key measures of dispersion of the distribution around the mean.

- **Expected Value of a Function of a Random Variable.** We have seen that the variance is the expected value of a particular transformation of the random variable (namely the square of the difference between the random variable and its mean), and we can generalize

this concept to expected values of any function g of a random variable as follows:

– Discrete Case: if X takes values in $\{x_i\}$, $i \in \mathbb{N}$, then

$$\mathbb{E}[g(x)] = \sum_i g(x_i) p_X(x_i)$$

Proof. We group together all the values x_i that share a common value $g(x_i)$, and we let g_j be the discrete set of values in the image of the function g when its domain is the support of the random variable X, $\{x_1, \ldots, x_i, \ldots\}$. We therefore have:

$$\begin{aligned}\sum_i g(x_i) p_X(x_i) &= \sum_j g_j \sum_{i:g(x_i)=g_j} p_X(x_i) \\ &= \sum_j g_j \mathbb{P}(X = g_j) \\ &= \mathbb{E}[g(X)]\end{aligned}$$

□

– Continuous Case:

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

- **Moments of a Random Variable.** Moments are an important building block towards our upcoming proof of the Central Limit Theorem. First off, some of the regularity conditions that determine applicability of the CLT are expressed as conditions on some of the moments. Futhermore, we will be using in our proof a concept of Moment Generating Function, whose name refers to the moments that we will define and explore here.

  We define the k^{th} moment, with $k \in \mathbb{N}$, of a random variable X as follows:

  $$M_k(X) = \mathbb{E}[X^k]$$

  With this definition, we see that:

  – The expected value of a random variable is also its first moment.

  $$\mathbb{E}[X] = \mu_X = M_1(X)$$

  – The variance of a random variable relates to its first two moments as follows:

  $$\mathrm{Var}(X) = \mathbb{E}(X^2) - \mu_X^2 = M_2(X) - M_1^2(X)$$

- **Central Moments of a Random Variable.** We define the k^{th} central moment, with $k \in \mathbb{N}$, of a random variable X as:

  $$\mu_k(X) = \mathbb{E}[(X - \mathbb{E}[X])^k]$$

  We see in particular that:

  $$\mathrm{Var}(X) = \mu_2(X).$$

- **Standardized Moments of a Random Variable.** We define the k^{th} standardized moment, with $k \in \mathbb{N}$, of a random variable X with mean μ_X and standard deviation σ_X as:

  $$\tilde{\mu_k}(X) = \mathbb{E}\left[\left(\frac{X - \mu_X}{\sigma_X}\right)^k\right]$$

- **Sknewness of a Random Variable.** The sknewness of a random variable is the third standardized moment of the random variable, i.e.:

$$\tilde{\mu_3}(X) = \mathbb{E}\left[\left(\frac{X - \mu_X}{\sigma_X}\right)^3\right] = \frac{\mu_3(X)}{\sigma_X^3}$$

The skewness is significant because when a distribution is unimodal, i.e., when it has a single peak in the shape of the PDF, then the skewness measures the asymmetry in the shape of the distribution about the mode of the random variable, with a skewness of 0 indicating a perfectly symmetric distribution about the mode, a positive (or right) skew indicating a right tail that tapers off more slowly than the left tail, and a negative (or left) skew indicating the opposite.

When a distribution is symmetric, then the mean is equal to the median (50^{th} percentile), and the skewness is 0. In addition, if the distribution is also unimodal as well as symmetric, then the mean, the mode, and the median are all equal. We will see later on that the normal distribution satisfies the conditions of unimodality and symmetry, and has skewness of 0.

- **Kurtosis of a Random Variable.** The kurtosis of a random variable is the fourth standardized moment of the random variable, i.e.:

$$\tilde{\mu_4}(X) = \mathbb{E}\left[\left(\frac{X - \mu_X}{\sigma_X}\right)^4\right] = \frac{\mu_4(X)}{\sigma_X^4}$$

The kurtosis is significant because it measures the extent to which a distribution has probability still concentrated at the outlying ends of the support (i.e., values with large absolute values), versus having more of the probability concentrated in the central part of the support of the random variable.

We will see in the section on the normal distribution that its kurtosis (equal to 3) is used as a standard against which to measure other distributions. Distributions with kurtosis higher than 3 have "heavier tails" than the normal distribution, i.e., they tend to have more mass or density at the outlier values. Conversely, distributions with kurtosis less than 3 have less mass or density at the outlier values (with large absolute values).

2.7. LINEAR COMBINATIONS AND PRODUCTS OF RANDOM VARIABLES

We examine a few key results relative to the expected value or the variance of linear combinations or of products of random variables.

- **Expected Value of a Linear Combination.** We have the following result for the linear combination of two random variables X and Y with α and β being two real constants, and we prove it for the discrete case although it is also true in the continuous case.

$$\mathbb{E}(\alpha X + \beta Y) = \alpha\mathbb{E}(X) + \beta\mathbb{E}(Y)$$

Proof. Let $\{x_1, \ldots, x_i, \ldots\}$ be the support of X and $\{y_1, \ldots, y_j, \ldots\}$ be the support of Y. We look at the expected value of the random variable $\alpha X + \beta Y$ as the expected value of the bi-variate function $g(x, y) = \alpha x + \beta y$ applied to the joint random variable (X, Y) whose

support is the Cartesian product of the supports of X and of Y, and we have:

$$\begin{aligned}
\mathbb{E}(\alpha X + \beta Y) &= \sum_{x_i,\, y_j} (\alpha x_i + \beta y_j)\mathbb{P}(X = x_i,\, Y = y_j) \\
&= \sum_{x_i} \sum_{y_j} (\alpha x_i + \beta y_j)\mathbb{P}(X = x_i,\, Y = y_j) \\
&= \sum_{x_i} \sum_{y_j} \alpha x_i \mathbb{P}(X = x_i,\, Y = y_j) + \sum_{x_i} \sum_{y_j} \beta y_j \mathbb{P}(X = x_i,\, Y = y_j) \\
&= \alpha \sum_{x_i} x_i \sum_{y_j} \mathbb{P}(X = x_i,\, Y = y_j) + \beta \sum_{y_j} y_j \sum_{x_i} \mathbb{P}(X = x_i,\, Y = y_j) \\
&= \alpha \sum_{x_i} x_i \mathbb{P}(X = x_i) + \beta \sum_{y_j} y_j \mathbb{P}(Y = y_j) \\
&= \alpha \mathbb{E}(X) + \beta \mathbb{E}(Y)
\end{aligned}$$

□

- **Expected Value of Sample Mean.** We frequently encounter experimental situations in which we collect a sample of n random variables $X_1, X_2, \ldots, X_n$ and we are interested in various statistical measures related to the sample. We define the sample mean as:

$$\bar{X} = \frac{X_1 + \cdots + X_n}{n}$$

Our previous result on linear combinations of random variables lets us derive the following result:

$$\mathbb{E}(\bar{X}) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}(X_i)$$

In the special case where all the random variables X_i have a common expected value μ_X, we have the following result that states that the average of the sample mean is equal to the common average of the random variables in the sample:

$$\mathbb{E}(\bar{X}) = \frac{1}{n} \sum_{i=1}^{n} \mu_X = \mu_X$$

- **Expected Value of Product of Independent Random Variables.** We have the following result when X and Y are independent random variables, and we prove it for the discrete case:

$$\mathbb{E}(X \cdot Y) = \mathbb{E}(X) \cdot \mathbb{E}(Y)$$

Proof.

$$\begin{aligned}
\mathbb{E}(X \cdot Y) &= \sum_{x_i} \sum_{y_j} x_i y_j \mathbb{P}(X = x_i,\, Y = y_j) \\
&= \sum_{x_i} \sum_{y_j} x_i y_j \mathbb{P}(X = x_i) \cdot \mathbb{P}Y = y_j) \\
&= \left(\sum_{x_i} x_i \mathbb{P}(X = x_i) \right) \cdot \left(\sum_{y_j} y_j \mathbb{P}(Y = y_j) \right) \\
&= \mathbb{E}(X) \cdot \mathbb{E}(Y)
\end{aligned}$$

□

This lets us prove the following important result.

- **Variance of Linear Combination of Independent Random Variables.** We state and prove for the discrete case the following general result, with the key step in the proof being the use of the result that the expected value of the product of independent random variables is equal to the product of their respective expected values.

$$\mathrm{Var}(\alpha X + \beta Y) = \alpha^2 \,\mathrm{Var}(X) + \beta^2 \,\mathrm{Var}(Y)$$

Proof. We let $\mu_X = \mathbb{E}(X)$ and $\mu_Y = \mathbb{E}(Y)$, an we have:

$$\begin{aligned}
\mathrm{Var}(\alpha X + \beta Y) &= \mathbb{E}[(\alpha X + \beta Y)^2] - (\alpha\mu_X + \beta\mu_Y)^2 \\
&= \mathbb{E}[\alpha^2 X^2 + 2\alpha\beta XY + \beta^2 Y] - (\alpha^2\mu_X^2 + 2\alpha\beta\mu_X\mu_Y + \beta^2\mu_Y^2) \\
&= \alpha^2\mathbb{E}(X^2) + 2\alpha\beta\mathbb{E}(XY) + \beta^2\mathbb{E}(Y^2) - \alpha^2\mu_X^2 - 2\alpha\beta\mu_X\mu_Y - \beta^2\mu_Y^2 \\
&= \alpha^2\mathbb{E}(X^2) + 2\alpha\beta\mathbb{E}(X)\mathbb{E}(Y) + \beta^2\mathbb{E}(Y^2) - \alpha^2\mu_X^2 - 2\alpha\beta\mu_X\mu_Y - \beta^2\mu_Y^2 \\
&= \alpha^2\mathbb{E}(X^2) + 2\alpha\beta\mu_X\mu_Y + \beta^2\mathbb{E}(Y^2) - \alpha^2\mu_X^2 - 2\alpha\beta\mu_X\mu_Y - \beta^2\mu_Y^2 \\
&= \alpha^2[\mathbb{E}(X^2) - \mu_X^2] + \beta^2[\mathbb{E}(Y^2) - \mu_Y^2] \\
&= \alpha^2 \,\mathrm{Var}(X) + \beta^2 \,\mathrm{Var}(Y)
\end{aligned}$$

□

2.8. INDEPENDENT, IDENTICALLY DISTRIBUTED RANDOM VARIABLES – i.i.d.

As mentioned during the definition of the sample mean above, it is a frequent experimental setup to be gathering a sample of n random variables $X_1, \ldots, X_n$. It is also common for the random variables to all be drawn from the same probability distribution with mean μ_X and variance $Var(X)$ and standard deviation σ_X, and in a manner that makes their draws independent of each other. We refer to this situation as having a sequence $X_1, \ldots, X_n$ of *independent, identically distributed (i.i.d.)* random variables.

Based on the results we have stated and proven in *Section* 2.7 above, we have the following for a sequence of *i.i.d.* random variables $X_1, \ldots, X_n$:

$$\mathbb{E}(X_1) = \cdots = \mathbb{E}(X_n) = \mu_X$$

$$\mathbb{E}(\bar{X}) = \mathbb{E}\left(\frac{X_1 + \cdots + X_n}{n}\right) = \mu_X$$

$$\mathbb{E}(X_i X_j)_{i \neq j} = 0$$

$$\mathbb{E}(X_1^2) = \cdots = \mathbb{E}(X_n^2) = Var(X) + \mu_X^2$$

$$\mathrm{Var}(\bar{X}) = \mathrm{Var}\left(\frac{X_1 + \cdots + X_n}{n}\right) = \frac{1}{n^2}\,\mathrm{Var}(X_1) + \cdots + \frac{1}{n^2}\,\mathrm{Var}(X_n) = n \cdot \frac{1}{n^2}\,\mathrm{Var}(X) = \frac{\mathrm{Var}(X)}{n}$$

We can also state the last result as:

$$\sigma_{\bar{X}}^2 = \frac{\sigma_X^2}{n}$$

$$\sigma_{\bar{X}} = \frac{\sigma_X}{\sqrt{n}}$$

The results above are significant and commonly used. They tell us in particular that the expected value of the sample mean of an *i.i.d.* sample is equal to the common mean, but that the standard deviation of that sample mean gets asymptotically close to 0 as the sample size n gets large.

3. NORMAL $\mathcal{N}(\mu, \sigma^2)$ AND STANDARD NORMAL $\mathcal{N}(0, 1)$

3.1. NORMALLY DISTRIBUTED RANDOM VARIABLE

We now describe the normal distribution for a continuous random variable, and we examine a few of its important properties in order to be come familiar with the "destination object" of the Central Limit Theorem. We say that a random variable X is distributed according to a normal distribution with mean μ and standard deviation σ, and we note this as $X \sim \mathcal{N}(\mu, \sigma^2)$, if its probability density function is:

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

and its cumulative distribution function is:

$$\mathbb{P}(X \leq x) = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt.$$

It can be shown that the PDF of the normal distribution integrates to 1 when taken over $(-\infty, \infty)$, and we sketch the proof here, which relies on *Fubini's Theorem* for integration of non-negative functions, and converting from integration in Cartesian coordinates to integration in polar coordinates:

$$\begin{aligned}
\int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt &= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{x^2}{2\sigma^2}} dx \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{u^2}{2}} du \\
&= \sqrt{\left(\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{u^2}{2}} du\right) \cdot \left(\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{v^2}{2}} dv\right)} \\
&= \frac{1}{\sqrt{2\pi}} \sqrt{\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-\frac{(u^2+v^2)}{2}} du dv} \\
&= \frac{1}{\sqrt{2\pi}} \sqrt{\int_{0}^{\infty} \int_{0}^{2\pi} e^{-\frac{r^2}{2}} r \cdot dr \cdot d\theta} \\
&= \frac{1}{\sqrt{2\pi}} \sqrt{2\pi \int_{0}^{\infty} r e^{-\frac{r^2}{2}} dr} \\
&= \sqrt{-e^{-\frac{r^2}{2}}\big|_0^{\infty}} \\
&= 1.
\end{aligned}$$

We will not show the proof here, but it can be shown using integration by parts that when $X \sim \mathcal{N}(\mu, \sigma^2)$, then it is the case that:

$$\begin{aligned}
\mathbb{E}(X) &= \mu \\
\text{Var}(X) &= \sigma^2 \\
\tilde{\mu_3}(X) &= 0 \\
\tilde{\mu_4}(X) &= 3
\end{aligned}$$

3.2. SHAPE OF THE NORMAL PDF

The normal distribution is unimodal and symmetric, and it has a measure of skewness of 0. In fact, given that the function is even around its mean, all of its odd standardized moments are equal to 0.

In addition, the kurtosis (standardized fourth moment) of any normally distributed random variable is equal to 3. This number is used as a benchmark against which other distributions are

compared for insights into whether they are "heavier tailed" (kurtosis greater than 3) or "thinner tailed" (kurtosis less than 3) than a normal distribution.

3.3. STANDARD NORMAL

The standard normal distribution $\mathcal{N}(0,\,1)$ is the special case of the normal distribution when the parameters are $\mu_X = 0$, and $\sigma_X^2 = 1$. Its PDF is therefore:

$$f_{\mathcal{N}(0,\,1)}(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

Some of the well-known statistics of the standard normal distribution are the probabilities of the random variable falling within one, two, or three units from its zero mean, and these are approximately 0.68, 0.95, and 0.999, respectively. The same probabilities apply for any normally distributed random variable falling within one, two, or three standard deviations from its mean.

3.4. MOMENTS OF THE STANDARD NORMAL

The k^{th} moment of a standard normal is given by:

$$\begin{aligned}
M_k &= \mathbb{E}[X^k] \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^k e^{-\frac{x^2}{2}} \\
&= \frac{1}{\sqrt{2\pi}} \frac{x^{k+1}}{k+1} e^{-\frac{x^2}{2}} \Big|_{-\infty}^{\infty} - \frac{1}{\sqrt{2\pi}} \int_{\infty}^{\infty} \frac{x^{k+1}}{k+1} \left(-x e^{-\frac{x^2}{2}}\right) dx \\
&= \frac{1}{\sqrt{2\pi}} \frac{1}{k+1} \int_{-\infty}^{\infty} x^{k+2} e^{-\frac{x^2}{2}} dx \\
&= \frac{1}{k+1} \cdot \mathbb{E}[X^{k+2}] \\
&= \frac{M_{k+2}}{k+1}
\end{aligned}$$

So we have a recurrence relation:

$$M_{k+2} = (k+1) M_k$$

With $M_0 = 1$, $M_1 = 0$, and $M_2 = 1$, it is easy to see by induction that we also have for all $k \in \mathbb{N}$:

$$M_{2k+1} = 0,$$

$$M_{2k+2} = (2k+1) M_{2k} = (2k+1)(2k-1)\ldots 1 = \frac{(2k+1)!}{2^k k!}$$

Indeed, we expect all the odd moments to be equal to 0 because we are integrating an odd function over $\mathbb{R}$.
We also retrieve from the formula for even moments the value of the kurtosis for the standard normal:

$$M_4 = \mu_4 = \frac{3!}{2^1 \cdot 1!} = 3.$$

3.5. THE SUM OF INDEPENDENT NORMAL r.v.'s IS A NORMAL r.v.

A significant result is that if X and Y are normally distributed with $X \sim \mathcal{N}(\mu_X,\, \sigma_X^2)$ and $Y \sim \mathcal{N}(\mu_Y,\, \sigma_Y^2)$, then $X + Y$ is also normally distributed with $X + Y \sim \mathcal{N}(\mu_X + \mu_Y,\, \sigma_X^2 + \sigma_Y^2)$. We sketch the proof using the concept of convolution that we described in *Section* 2 above.

Proof. We know that $X+Y$ will have mean $\mathbb{E}(X+Y) = \mu_X + \mu_Y$ and variance $Var(X+Y) = \sigma_X^2 + \sigma_Y^2$, by linearity of the expectation operator, on the one hand, and by independence of the variables X and Y, on the other hand.

As calculations are less tedious, we first show that the sum of two standard normal random variables is also a normal random variable, with mean 0 and variance 2:

$$\begin{aligned}\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-t)^2}{2}} dt &= \frac{1}{\sqrt{2\pi}} \cdot \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{(2t^2-2tx+x^2)}{2}} \\ &= \frac{1}{\sqrt{2\pi}} \cdot \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{\left[2\left(t-\frac{x}{2}\right)^2 - 2\frac{x^2}{4} + x^2\right]}{2}} dt \\ &= \left(\frac{e^{-\frac{x^2}{4}}}{\sqrt{2\pi}}\right) \cdot \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}\left[\sqrt{2}\left(t-\frac{x}{2}\right)\right]^2} dt\end{aligned}$$

We perform the change of variables $u = \sqrt{2}\left(t - \frac{x}{2}\right)$ in the integral, which implies that $dt = \frac{1}{\sqrt{2}} du$, and the integration bounds are still from $-\infty$ to ∞, so our integral becomes:

$$\left(\frac{1}{\sqrt{2}\cdot\sqrt{2\pi}}\right) e^{-\frac{1}{2}\left(\frac{x}{\sqrt{2}}\right)^2} \cdot \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}u^2} du = \frac{1}{\sqrt{2}\cdot\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x}{\sqrt{2}}\right)^2}$$

where the next to last expression simplifies because the right-hand operand of the product is simply the integral over $\mathbb{R}$ of the PDF of the standard normal distribution, which is of course equal to 1.

We recognize in the final expression the PDF of $\mathcal{N}(0,\, 2)$, which shows that the sum of two independent, and standard normally distributed random variables is also a normally distributed random variable with mean 0 and with standard deviation $\sqrt{2}$, i.e., with variance 2. □

This result can be generalized (by affine variable transformation of scaling and shifting) to the sum of two independent, non-standard normally distributed random variables being a normally distributed random variable with mean equal to the sum of the two means, and variance equal to the sum of the two variances.

The result can also be generalized by induction to a sum of any finite number of independent, normally distributed random variables being normal with mean equal to the sum of the means and variance equal to the sum of the variances. Similarly, the result generalizes to any linear combination of independent, normally distributed random variables, whose distribution is also normal, and whose mean and variance are derived by following the rules of additivity and scaling for means and variances that we saw in *Sections* $2.7 - 2.8$.

4. CENTRAL LIMIT THEOREM (CLT).

4.1. CONTEXT

We have seen in the previous section that a sum – or in fact any linear combination – of independent, normally distributed random variables, is also normally distributed. In particular, we have the case of a sample mean when dealing with *i.i.d.* random variables. If $X_1, \ldots, X_n$ are *i.i.d.* with $X_i \sim \mathcal{N}(\mu_X, \sigma_X^2)$ for all $i : 1 \leq i \leq n$, and we let $S_n = X_1 + \cdots + X_n$, then the sample mean is distributed exactly as follows:

$$\overline{X} = \frac{X_1 + \cdots + X_n}{n} = \frac{S_n}{n} \sim \mathcal{N}(\mu_X,\, \frac{\sigma_X^2}{n})$$

The Central Limit Theorem constitutes an asymptotic generalization of the result above to a sample mean of *i.i.d.* random variables that *may not* be normally distributed, but whose sample mean will have a distribution that approaches a normal distribution as n gets large. Specifically, we have the following statement.

4.2. STATEMENT OF THE CENTRAL LIMIT THEOREM

Let $X_1, \ldots, X_n$ be n independent and identically distributed random variables, all distributed according to a common distribution with mean μ_X and with finite standard deviation σ_X, and let $S_n = X_1 + \cdots + X_n$ and $\overline{X} = \frac{S_n}{n}$. Then, the cumulative distribution function (CDF) of their *standardized sum*

$$Z_n = \frac{X_1 + \cdots + X_n - n\mu_X}{\sigma_X \sqrt{n}} = \frac{S_n - n\mu_X}{\sigma_X \sqrt{n}} = \frac{\overline{X} - \mu_X}{\frac{\sigma_X}{\sqrt{n}}}$$

approaches a standard normal distribution as $n \to \infty$. This is formally stated as:

$$\mathbb{P}\left(Z_n = \frac{X_1 + \cdots + X_n - n\mu_X}{\sigma_X \sqrt{n}} \leq x\right) \to \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{x^2}{2}} dx \text{ as } n \to \infty.$$

We note that the CLT is a statement about the convergence of a sequence of functions to a limit function, with these functions being CDFs of random variables. As such it is an example of *convergence in distribution*, which in general states the convergence of the CDFs of a sequence of random variables to a limit CDF.

Under stronger conditions of the common distribution of the X_i having finite third absolute moment (i.e., if $\mathbb{E}(|X_i|^3 < \infty)$, then the probability density function (or probability mass function for the discrete case) of the standardized sum also converges to the probability density function of the standard normal $\mathcal{N}(0, 1)$.

4.3. HISTORY

Historically, it is Abaraham De Moivre who, in 1738 first stated the result and proved it for the special case of Bernoulli distributed random variables in his *The Doctrine of Chances*, and with the parameter $p = \frac{1}{2}$ making the distribution symmetric (equal odds of success and failure of the binary experiment). Laplace subsequently generalized the theorem to $p \neq \frac{1}{2}$ and his result was published in 1812 in *Théorie Analytique des Probabilités*. He also stated the more general (non-Bernoulli) form of the CLT, but didn't give a rigorous proof, and Gauss further improved on Laplace's results.

The first rigorous proof of the CLT was published by Alexander Liapunov in $1901 - 1902$, and the theorem is occasionally referred to as *Liapunov's Theorem*. The designation of *Central Limit Theorem* is due to George Pólya who used it in a paper he published in 1920.

4.4. REMARK

The CLT is a remarkable result in its generality, as it applies to discrete as well as continuous random variables, and almost regardless of the shape of the individual random variables' common distribution, as long as the mean and variance are finite (for the CDF convergence result), and additionally the third absolute moment is finite (for the PDF convergence result).

It is important to note that the main condition that causes the CLT to fail to apply is when the individual random variables' common distribution has too high kurtosis ("too heavy-tailed")

which in practice translates to a variance not being finite, or sometimes even a mean being undefined. This is the case for the Cauchy distribution, for instance, whose PDF and CDF are given by:

$$f_{X,\,x_0,\,\gamma}(x) = \frac{1}{\pi\gamma}\frac{\gamma^2}{(x-x_0)^2+\gamma^2}$$

$$F_{X,\,x_0,\,\gamma}(x) = \frac{1}{\pi}\arctan\left(\frac{x-x_0}{\gamma}\right)+\frac{1}{2}$$

A Cauchy-distributed random variable has undefined mean and undefined variance. It also has undefined sknewness and undefined kurtosis! The paramater x_0 is called a location parameter and it specifies both the median and the mode of the distribution. The parameter γ is the scale parameter, which specifies the half-width at half-maximum from the peak.

4.5. SPEED OF CONVERGENCE

Practice of Statistics. In applied statistics, the question of "how large does n need to be for good approximations" is a very important one to decide on minimum sample sizes needed to achieve results at desired tolerance levels for the error in the approximation of the distribution.

In addition to the heavy tailed considerations mentioned above, another one of the main hurdles to rapid convergence towards the standard normal asymptotic distribution, is the skewness of the original common distribution of the X_i. When the individual random variables have a distribution that is fairly symmetric around a point (where the mean, median, and mode would all be concentrated), then convergence happens rather quickly.

If, however, the original distribution has a high amount of sknewness, then "the memory of that sknewness" lasts through quite large sample sizes even when the averaging from the sample mean is taking place. The distribution of the sample mean in these cases can continue to exhibit asymmetry in its tails even with large sample sizes. Additionally, lack of unimodality (i.e., multiple peaks) in the original distribution can be another source of slow convergence towards the target standard normal distribution.

Theoretical Bounds on Speed of Convergence. More formally than the heuristics and qualitative observations above, there are formal results for bounds on the rate of convergence of the CDF of the standardized sample mean towards the CDF of the standard normal distribution, under additional conditions. With no added conditions besides those of the CLT, convergence of the CDFs is not uniform, i.e., it occurs more slowly for the tails of the distribution than for its center.

Berry-Esseen Theorem. The best known result for rate of convergence, with additional conditions to those of the CLT but still fairly broad applicability, is the following:

If $X_1, \ldots, X_n$ are *i.i.d.* with $\mathbb{E}[X_i] = 0$, with $\mathbb{E}[X_i^2] = \sigma_X^2 < \infty$, and with $\mathbb{E}(|X_i|^3) = \rho_X < \infty$, and if we designate by Z_n the standardized sum of the X_i, i.e.:

$$Z_n = \frac{X_1+\cdots+X_n-n\mu_X}{\sigma_X\sqrt{n}} = \frac{X_1+\cdots+X_n}{\sigma_X\sqrt{n}},$$

then, with $F_{Z_n}(x)$ being the CDF of Z_n, and $F_{\mathcal{N}(0,\,1)}(x)$ being the CDF of the standard normal $\mathcal{N}(0,\,1)$, we have:

$$|F_{Z_n}(x) - F_{\mathcal{N}(0,\,1)}(x)| \leq \frac{C\rho_X}{\sigma_X^3\sqrt{n}},\ \text{with } 0.4 \leq C \leq 0.48.$$

4.6. BERNOULLI CASE - DE MOIVRE-LAPLACE THEOREM

The De Moivre Laplace Theorem is simply the application of the Central Limit Theorem to the specific case of the sample mean of a set of Bernoulli distributed, *i.i.d.* random variables, with $P_{n,\,p}X(k) = p^k(1-p)^{n-k}$, and $0 \leq k \leq n$.

4.7. DISCRETE CASE - DE MOIVRE-LAPLACE 1/2 CORRECTION

In the case of discrete random variables with discrete PMFs and CDFs, the standard normal CDF provides a good approximation of the sample sum's CDF for large n, but the standard normal PDF can be off from the sample sum's PMF. This is because the sum S_n only takes integer values, so $\mathbb{P}(S_n \leq k) = \mathbb{P}(S_n < k+1)$, but this does not translate to the equivalent inequalities for the normal PDF, which is of course a continuous PDF that puts zero probability on any specific point on the real line.

The idea behind the $\frac{1}{2}$ correction, therefore, is to improve the quality of the approximation of the sample sum's PMF at a particular discrete value k through the use of the following "trick" of stating that:

$$\mathbb{P}(S_n = k) = \mathbb{P}\left(k - \frac{1}{2} \leq S_n \leq k + \frac{1}{2}\right) = \mathbb{P}\left(\frac{k - \frac{1}{2} - n\mu_X}{\sigma_X\sqrt{n}} \leq \frac{S_n - n\mu_X}{\sigma_X\sqrt{n}} \leq \frac{k + \frac{1}{2} - n\mu_X}{\sigma_X\sqrt{n}}\right)$$

for which we can get an estimate from the CDF of the continuous standard normal $\mathcal{N}(0,\,1)$:

$$\begin{aligned}\mathbb{P}(S_n = k) &= \mathbb{P}\left(\frac{k - \frac{1}{2} - n\mu_X}{\sigma_X\sqrt{n}} \leq \frac{S_n - n\mu_X}{\sigma_X\sqrt{n}} \leq \frac{k + \frac{1}{2} - n\mu_X}{\sigma_X\sqrt{n}}\right) \\ &\approx F_{\mathcal{N}(0,\,1)}\left(\frac{k + \frac{1}{2} - n\mu_X}{\sigma_X\sqrt{n}}\right) - F_{\mathcal{N}(0,\,1)}\left(\frac{k - \frac{1}{2} - n\mu_X}{\sigma_X\sqrt{n}}\right)\end{aligned}$$

This correction considerably improves the error on the approximation of individual probability masses for a discrete random variable's sample sum, from the use of the CDF of the standard normal distribution according to the CLT. This addresses an issue that is specific to discrete random variables and is not needed for continuous random variables and their sample sums.

4.8. SUMMARY

In summary, the scope of applicability of the CLT is quite broad as it applies whenever the distribution of the *i.i.d.* random variables has finite means and variances. It allows us to approximate the sample mean CDF which is possibly very complicated or intractable to calculate in closed form, by a CDF for which there are readily available precise calculations.

A slightly stronger condition on the third absolute moment being finite also lets us approximate the PDF of the sample mean, and – with the $\frac{1}{2}$ Correction – get good approximations for the PMF in the discrete case.

In all cases, symmetry (i.e., low absolute sknewness), unimodality, and absence of heavy tails (i.e., low to medium kurtosis) contribute to much faster convergence towards the asymptotic distribution $\mathcal{N}(0,\,1)$.

5. DISTRIBUTION-SPECIFIC PROOF – CLT FOR BERNOULLI

For the case of Bernoulli distributed *i.i.d.* random variables $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$, the sample sum S_n is distributed as a Binomial $\mathcal{B}(n, p)$, with mean np and with variance $np(1-p)$ and with PMF:

$$\mathbb{P}(S_n = x) = \binom{n}{x} p^x (1-p)^{n-x}, \text{ for } 0 \leq x \leq n.$$

We can sketch the proof of convergence of the PMF towards the PDF of a standard normal using the Stirling approximation of the factorial function, and some Taylor expansion but with no need for the mechanics of MGFs which we will use for the general case in the next section.

Proof. With the Stirling approximation:

$$n! \approx \frac{n^n}{e^n}\sqrt{2\pi n}$$

we have:

$$\mathbb{P}(S_n = x) \approx \frac{\sqrt{n}}{\sqrt{2\pi}\sqrt{x(n-x)}} \left(\frac{np}{x}\right)^x \left(\frac{n(1-p)}{n-x}\right)^{n-x} = \frac{1}{\sqrt{2\pi np(1-p)}} \left(\frac{x}{np}\right)^{-x-\frac{1}{2}} \left(\frac{n-x}{n(1-p)}\right)^{-(n-x)-\frac{1}{2}}$$

We now take natural logarithms on both sides, and we get:

$$\log \mathbb{P}(S_n = x) \approx -\frac{1}{2}\log[2\pi np(1-p)] - (x+\frac{1}{2})\log\left(\frac{x}{np}\right) - (n-x+\frac{1}{2})\log\left(\frac{n-x}{n(1-p)}\right)$$

We now let $Z_n = \frac{S_n - np}{\sqrt{np(1-p)}}$ and $z_n = \frac{x-np}{\sqrt{np(1-p)}}$, our approximation becomes:

$$\log \mathbb{P}(Z_n = z) \approx -\frac{1}{2}\log[2\pi np(1-p)] + L_n(z, p) + L_n(-z, p)$$

where:

$$L_n(z, p) = -(np + z\sqrt{np(1-p)} + \frac{1}{2})\ln\left(1 + z\sqrt{\frac{1-p}{np}}\right)$$

With n large, we can use a Taylor expansion of the logarithm and approximate $L_n(z, p)$ as follows:

$$L_n(z, p) = -(np + z\sqrt{np(1-p)} + \frac{1}{2})\left[z\sqrt{\frac{1-p}{np}} - \frac{z^2(1-p)}{2np} + \frac{z^3(1-p)^{\frac{3}{2}}}{3(np)^{\frac{3}{2}}} - \ldots\right]$$

After simplification, the expressions become:

$$L_n(z, p) = -z\sqrt{np(1-p)} - (1-p)\frac{z^2}{2} + O(\frac{1}{\sqrt{n}})$$

$$L_n(-z, p) = z\sqrt{np(1-p)} - p\frac{z^2}{2} + O(\frac{1}{\sqrt{n}})$$

So their sum becomes:

$$L_n(z, p) + L_n(-z, p) = -\frac{z^2}{2} + O(\frac{1}{\sqrt{n}})$$

and our logarithm of the approximated PMF becomes:

$$\log \mathbb{P}(Z_n = z) \approx -\frac{1}{2}\log[2\pi np(1-p)] - \frac{z^2}{2} + O(\frac{1}{\sqrt{n}})$$

After going back from the logarithm, our approximation becomes:

$$\mathbb{P}(Z_n = z) \approx \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz$$

where we have written $\frac{1}{\sqrt{np(1-p)}} = \Delta_n \approx dz$, which goes to 0 as $n \to \infty$.

Going back to S_n, we have shown that:

$$\mathbb{P}(S_n = x) \approx \frac{1}{\sqrt{2\pi}\sqrt{np(1-p)}} e^{-\frac{(x-np)^2}{2[np(1-p)]}}$$

and its standardized form $Z_n = \frac{S_n - np}{\sqrt{np(1-p)}}$ has a distribution that approaches the standard normal $\mathcal{N}(0, 1)$. □

6. GENERIC PROOF OF CLT — MOMENT-GENERATING FUNCTIONS

In this section, we first introduce and prove the Markov and Chebyshev inequalities, as these proofs are short yet informative, then derive the Weak Law of Large Numbers (WLLN).

We then introduce the concept of moment-generating functions (MGF) corresponding to random variables, and how they relate to sums and scalings of random variables. We then explore the MGF of the standard normal $\mathcal{N}(0, 1)$, then we show how the MGFs of unspecified — but sufficiently regularly distributed — independent random variables result in MGFs for their sample means that converge to the MGF of $\mathcal{N}(0, 1)$.

We will rely on a particularly important lemma which we will not prove, however: the result that if a sequence of random variables have their MGFs converging to a limit MGF, then their cumulative distribution functions also converge to the cumulative distribution function that corresponds to the limit MGF. Under certain additional conditions, the probability density functions of these random variables also converge to the probability density function that corresponds to the limit MGF.

This ability to "walk back" from convergence of MGFs to convergence of CDFs and PDFs is critical to "sealing the proof" of the CLT, but proving it relies on the theory of Fourier Transforms, which would be a subject for its own paper...

6.1. MARKOV'S INEQUALITY

Markov's Inequality is a stepping stone towards proving Chebyshev's Inequality, which is needed to prove the Weak Law of Large Numbers. That law in turn will be needed for the proof of the Central Limit Theorem.

For a random variable X that takes only non-negative values, and with finite mean $\mathbb{E}(X)$, and for any positive real constant a, the following inequality holds:

$$\frac{\mathbb{E}[X]}{a} \geq \mathbb{P}(X \geq a).$$

Proof. For any event E, the indicator random variable $\mathbb{1}_E$ of the event E is the random variable that takes value 1 when the event E occurs, and the value 0 otherwise. Using this notation, we consider the indicator random variable $\mathbb{1}_{\{X \geq a\}}$ and we have:

$$\mathbb{1}_{\{X \geq a\}} = 1 \text{ if } X \geq a$$

$$\mathbb{1}_{\{X \geq a\}} = 0 \text{ if } X < a$$

Multiplying the equalities by $a > 0$, we have:

$$a \cdot \mathbb{1}_{\{X \geq a\}} = a \text{ if } X \geq a$$

$$a \cdot \mathbb{1}_{\{X \geq a\}} = 0 \text{ if } X < a$$

These two equalities make it clear that:

$$a \cdot \mathbb{1}_{\{X \geq a\}} \leq X$$

Indeed, when $X \geq a$, the left-hand side is equal to a and – by assumption of this case – $X \geq a$, so we have $a \cdot \mathbb{1}_{\{X \geq a\}} = a \leq X$. And when $X < a$, the left-hand side is equal to 0, and the random variable X is given as always taking non-negative values, so $a \cdot \mathbb{1}_{\{X \geq a\}} = 0 \leq X$ in this case too.

We now take expectations on both sides of the inequality, which preserves the inequality:

$$\mathbb{E}[a\mathbb{1}_{\{X \geq a\}}] \leq \mathbb{E}[X]$$

The left-hand side random variable takes value a with probability equal to $\mathbb{P}(X \geq a)$ and it takes value 0 with probability equal to $\mathbb{P}(X < a)$, so its expected value is equal to:

$$\mathbb{E}[a\mathbb{1}_{\{X \geq a\}}] = a \cdot \mathbb{P}(X \geq a) + 0 \cdot \mathbb{P}(X < a) = a \cdot \mathbb{P}(X \geq a)$$

We therefore have, by replacement into our inequality with $\mathbb{E}[X]$:

$$a \cdot \mathbb{P}(X \geq a) \leq \mathbb{E}[X]$$

. Dividing boths sides by $a > 0$ preserves the inequality and we get:

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}[X]}{a}.$$

□

6.2. CHEBYSHEV'S INEQUALITY

Chebyshev's Inequality is important for the derivation of the Weak Law of Large Numbers. It consists of applying Markov's Inequality to the specific non-negative random variable $(X - \mathbb{E}[X])^2$ whose expected value is the variance $\mathrm{Var}(X)$ of the random variable.

Specifically, letting X be a random variable with finite mean and variance, we have:

$$\mathbb{P}\left(|X - \mu_X| \geq k\sigma_X\right) \leq \frac{1}{k^2}.$$

Proof. We apply Markov's Inequality to the non-negative random variable $(X - \mu_X)^2$, and using $a = k^2\sigma_X^2 > 0$ as the positive constant, and we have:

$$\mathbb{P}[(X - \mu_X)^2 \geq k^2\sigma_X^2] \leq \frac{\mathbb{E}[(X - \mu_X)^2]}{k^2\sigma_X^2} = \frac{\mathrm{Var}(X)}{k^2\sigma_X^2} = \frac{\sigma_X^2}{k^2\sigma_X^2} = \frac{1}{k^2}$$

We now note that:

$$\mathbb{P}[(X - \mu_X)^2 \geq k^2\sigma_X^2] = \mathbb{P}(|X - \mu_X| \geq k\sigma_X)$$

Our inequality therefore becomes:

$$\mathbb{P}(|X - \mu_X| \geq k\sigma_X) \leq \frac{1}{k^2}.$$

□

The significance of Chebyshev's Inequality is that it puts an upper bound on the probability of a random variable deviating from its mean, in a general way that applies to all random variables with finite mean and variance. When the deviation from the mean is measured in multiples of the standard deviation, the bound is the squared inverse of the multiple.

While the bound appears very generic in its form, the shape of the distribution still implicitly appears in the tightness of the bound. Indeed, if σ_X is very large, then for a given k, $k\sigma_X$ will also be very large, so the interval within which Chebyshev's Inequality concentrates the random variable with high probability is very wide, i.e., not that precise. Conversely, if σ_X is very small, then the inequality tells us that the probability of X being outside of a rather narrow interval is small, which is useful in practice.

While we can achieve better bounds for specific distributions, we show one example where Chebyshev's Inequality is sharp, i.e., it cannot be improved upon. The example is the following:

Let $k \geq 1$ be a real constant and let X be a discrete random variable distributed as follows:

$$X = -1 \text{ with probability } \frac{1}{2k^2},$$

$$X = 0 \text{ with probability } 1 - \frac{1}{k^2},$$

$$X = +1 \text{ with probability } \frac{1}{2k^2}.$$

We can calculate that for this distribution, the mean $mu_X = 0$ and the standard deviation $\sigma_X = \frac{1}{k}$. This implies the following equality:

$$\mathbb{P}(|X - \mu_X| \geq k\sigma_X) = \mathbb{P}(|X| \geq 1) = \frac{1}{2k^2} + \frac{1}{2k^2} = \frac{1}{k^2}.$$

6.3. WEAK LAW OF LARGE NUMBERS – WLLN

The Weak Law of Large Numbers (WLLN) is a very important result in the context of a sample of *i.i.d.* random variables with finite mean μ_X, and where we are interested in characterizing whether and how the sample mean $\overline{X} = \frac{X_1 + \cdots + X_n}{n}$ approaches μ_X as n gets asymptotically large. We have seen in *Section* 2.7 that $\mathbb{E}[\overline{X}] = \mu_X$. We are now interested in probabilities of the random variable $\overline{X}$ being at some distance or greater from its own mean μ_X.

Unlike the CLT which is an illustration of *convergence in distribution*, i.e., of convergence of a sequence of CDF functions towards a limit CDF function, the WLLN illustrates a different concept of convergence known as *convergence in probability*, and which is stronger than *convergence in distribution* in that it implies it.

We now state and prove the WLLN, which is sometimes referred to as *Khinchin's Law*.

For any positive real number ε,

$$\lim_{n \to \infty} \mathbb{P}\left(|\overline{X} - \mu_X| > \varepsilon\right) = 0.$$

Proof. We recall that:

$$\text{Var}(\overline{X}) = \frac{\sigma_X^2}{n}$$

which, equivalently, means:

$$\sigma_{\overline{X}} = \frac{\sigma_X}{\sqrt{n}}$$

We have:

$$\mathbb{P}\left(|\overline{X} - \mu_X| > \varepsilon\right) = \mathbb{P}\left(|\overline{X} - \mu_X| > \left(\frac{\varepsilon}{\sigma_{\overline{X}}}\right)\sigma_{\overline{X}}\right)$$

By Chebyshev's Inequality applied to $\overline{X}$, and with the positive constant $k = \frac{\varepsilon}{\sigma_{\overline{X}}}$, we have:

$$\mathbb{P}\left(|\overline{X} - \mu_X| > \varepsilon\right) \leq \frac{\sigma^2_{\overline{X}}}{\varepsilon^2} = \frac{\sigma^2_X}{\varepsilon^2 n} \to 0 \text{ as } n \to \infty.$$

□

The interpretation of the WLLN is that if we set ourselves a given tolerance band of $\pm\varepsilon$ around μ_X, then for any desired small positive number $\eta > 0$, there is a value N_η after which the sample size n being $\geq N_\eta$ makes the sample mean $\overline{X}$ have all of its probability except η, i.e., probability $1 - \eta$ of falling inside the band of width 2ε centered at μ_X, i.e., in $[\mu_X - \varepsilon, \mu_X + \varepsilon]$.

Said otherwise, having chosen our tolerance band ε on either side of μ_X, if we pick a sample size $n \geq N_\eta$, and generate a very large number of samples of that size $\geq N_\eta$ and measure their sample means, we will see a fraction $1 - \eta$ of these sample means falling inside the interval $[\mu_X - \varepsilon, \mu_X + \varepsilon]$. And we can make η as small as we want, and we'll find an N_η as a sample size that bunches a fraction $1 - \eta$ of the sample means in our desired interval.

6.4. MOMENT-GENERATING FUNCTIONS

The moment-generating function of a random variable will play a critical in the upcoming proof of the CLT, so we introduce it here and we examine its most important properties.

We recall that we defined in *Section* 2.6 the expected value of a function g of a random variable X, and we showed that this is a well-defined operator with:

$$\mathbb{E}[g(X)] = \sum_x p_X(x) \cdot g(x)$$

for a discrete r.v. with PMF $p_X(x)$,
and

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} f_X(x) \cdot g(x) dx$$

for a continuous r.v. with PDF $f_X(x)$.

We also recall that the k^{th} moment of a random variable X, with $k \in \mathbb{N}$, is defined as:

$$M_k(X) = \mathbb{E}[X^k]$$

6.4.1. DEFINITION OF MGF

We define the moment-generating function $\varphi_X(t)$ of a random variable X as the following function of a real variable $t \in \mathbb{R}$:

$$\varphi_X(t) = \mathbb{E}[e^{tX}]$$

It is therefore equal to:

$$\varphi_X(t) = \sum_x e^{tx} p_X(x) \text{ (discrete case).}$$

$$\varphi_X(t) = \int_{-\infty}^{\infty} e^{tx} f_X(x) dx \text{ (continuous case).}$$

6.4.2. GENERATION OF MOMENTS

The reason this function is called moment-generating is that we can retrieve all the moments of a random variable by successive derivatives with respect to the variable t of the function, and evaluating the derivatives at $t = 0$.

More specifically, we retrieve the k^{th} moment $M_k(X)$ of the random variable X as follows:

$$M_k(X) = \mathbb{E}[X^k] = \varphi_X^{(k)}(0) = \varphi_X^{(k)}(t)\big|_{t=0}$$

We see this by noting that:

$$\varphi'_X(t) = \frac{d}{dt}\varphi_X(t) = \mathbb{E}[Xe^{tX}]$$

and more generally that, for $k \in \mathbb{N}$:

$$\varphi_X^{(k)}(t) = \frac{d^k}{dt^k}\varphi_X(t) = \mathbb{E}[X^k e^{tX}]$$

Proof. We show this by induction, and we accept the fact – without proving it – that we can swap the expectation operator (over the distribution of X) with the differentiation operator (with respect to t):

– Base Case:

$$\varphi'(t) = \frac{d}{dt}\mathbb{E}[e^{tX}] = \mathbb{E}[\frac{d}{dt}e^{tX}] = \mathbb{E}[Xe^{tX}].$$

– Induction Step:

We assume that $\varphi^{(k)}(t) = \mathbb{E}[X^k e^{tX}]$, and we have:

$$\varphi^{(k+1)}(t) = \frac{d}{dt}\varphi^{(k)}(t) = \frac{d}{dt}\mathbb{E}[X^k e^{tX}] = \mathbb{E}[\frac{d}{dt}X^k e^{tX}] = \mathbb{E}[X^{k+1}e^{tX}].$$

□

We now see that when we evaluate these derivatives of the function $\varphi(t)$ with respect to t at the point $t = 0$, we get:

$$\varphi'(0) = \frac{d}{dt}\varphi(t)\big|_{t=0} = \mathbb{E}[Xe^{0\cdot X}] = \mathbb{E}[X].$$

$$\varphi^{(k)}(0) = \frac{d^k}{dt^k}\varphi(t)\big|_{t=0} = \mathbb{E}[X^k e^{0\cdot X}] = \mathbb{E}[X^k].$$

6.4.3. MGF OF SUMS AND AFFINE TRANSFORMS

Affine Transformation. If $\varphi_X(t)$ is the MGF of the random variable X, and if $Y = aX + b$ is an affine transformation of X, we calculate the MGF of Y as follows:

$$\varphi_Y(t) = \mathbb{E}[e^{t(aX+b)}] = \mathbb{E}[e^{bt}e^{(at)X}] = e^{bt}\varphi_X(at).$$

Sum of Independent Random Variables. Let X and Y be independent random variables, with MGFs $\varphi_X(t)$ and $\varphi_Y(t)$, respectively. We note that for a given value of the parameter t, the random variables e^{tX} and e^{tY} are independent too, and we have:

$$\varphi_{X+Y}(t) = \mathbb{E}[e^{t(X+Y)}] = \mathbb{E}[e^{tX} \cdot e^{tY}] = \mathbb{E}[e^{tX}] \cdot \mathbb{E}[e^{tY}] = \varphi_X(t) \cdot \varphi_Y(t).$$

6.4.4. MGF OF STANDARD NORMAL r.v. $\mathcal{N}(0,\,1)$

We now calculate the MGF of a standard normally distributed random variable, as this will also come into the proof of the CLT. We have:

$$\begin{aligned}
\varphi_{\mathcal{N}(0,\,1)}(t) &= \mathbb{E}[e^{t(X\sim\mathcal{N}(0,\,1))}] \\
&= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{tx}\cdot e^{-\frac{x^2}{2}}\,dx \\
&= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\frac{1}{2}(x^2-2xt)}\,dx \\
&= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\frac{1}{2}(x^2-2xt+t^2)}\cdot e^{\frac{t^2}{2}}\,dx \\
&= e^{\frac{t^2}{2}}\cdot\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\frac{1}{2}(x-t)^2}\,dx \\
&= e^{\frac{t^2}{2}}\cdot 1 \\
&= e^{\frac{t^2}{2}}
\end{aligned}$$

We have therefore established that the MGF of the standard normal random variable is:

$$\varphi_{\mathcal{N}(0,\,1)}(t) = e^{\frac{t^2}{2}}.$$

6.4.5. MGF OF NORMAL r.v. $\mathcal{N}(\mu_X,\,\sigma_X^2)$

If $X\sim\mathcal{N}(\mu,\,\sigma^2)$, then $Z=\frac{X-\mu}{\sigma}\sim\mathcal{N}(0,\,1)$ and $\varphi_Z(t)=e^{\frac{t^2}{2}}$.

We have:

$$Z=\frac{X-\mu}{\sigma}\iff X=\sigma Z+\mu$$

We apply the result from *Section* 6.4.3 for the MGF of an affine transformation of a random variable, and we have:

$$\begin{aligned}
\varphi_X(t) &= e^{\mu t}\varphi_Z(\sigma t) \\
&= e^{\mu t}e^{\frac{\sigma^2t^2}{2}} \\
&= e^{\frac{\sigma^2t^2+2\mu t}{2}} \\
&= e^{-\frac{\mu^2}{2\sigma^2}}e^{\frac{\left(\sigma t+\frac{\mu}{\sigma}\right)^2}{2}}
\end{aligned}$$

6.5. GENERAL PROOF OF THE CENTRAL LIMIT THEOREM

We now move to the final part of the proof of the CLT. We start with a Lemma which we will accept without proof, but whose result is critical to the CLT. We then state the CLT and show a result which – combined with the result of the Lemma – give a proof of the CLT.

6.5.1. LEMMA "ON GOING BACK FROM MGF TO CDF"

Let $X_1,\,\ldots,\,X_n,\,\ldots$ be a sequence of random variables having CDFs $F_{X_n}(x)$ and MGFs $\varphi_{X_n}(t)$, and let Z be a random variable having CDF $F_Z(x)$ and MGF $\varphi_Z(t)$. Then:

If $\lim_{n\to\infty} \varphi_{X_n}(t) = \varphi_Z(t)$ for all t, then $\lim_{n\to\infty} F_{X_n}(x) \to F_Z(x)$ for all x at which $F_Z(x)$ is continuous.

The importance of this lemma is that it allows us to derive convergence of Cumulative Distribution Functions from convergence of Moment Generating Functions. Notably for the case where Z is normally distributed, its CDF is continuous everywhere, which means that the CDFs of the X_n will converge to the CDF of Z at all points.

6.5.2. THE THEOREM AND ITS PROOF

Let $X_1, \ldots, X_n, \ldots$ be a sequence of *i.i.d.* random variables each having mean μ and variance σ^2. Then the CDF of $\frac{X_1+\cdots+X_n-n\mu}{\sigma\sqrt{n}} = \frac{\frac{X_1+\cdots+X_n}{n}-\mu}{\frac{\sigma}{\sqrt{n}}}$ tends to the CDF of the standard normal as $n \to \infty$. That is, $\forall x \in \mathbb{R}$,

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} \leq x\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}}\, dt.$$

Proof. In order to simplify notation, we show the proof for the case where $\mu = 0$ and $\sigma = 1$, and we know how to generalize since we know how to handle affine transformations of random variables. With this simplifying assumption, we have:

$$\frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} = \frac{X_1 + \cdots + X_n}{\sqrt{n}} = \sum_{i=1}^{n} \frac{X_i}{\sqrt{n}}$$

We now calculate the MGF of this random variable, using the rules for MGFs of scalings of random variables, and of sums of independent random variables from *Section* 6.4.3 above.

Let us designate by $\varphi_{X_i}(t) = \varphi_X(t)$ the common MGF of the *i.i.d.* random variables $X_1, \ldots, X_n$. We have:

$$\begin{aligned}
\varphi_{\left(\sum_{i=1}^{n} \frac{X_i}{\sqrt{n}}\right)}(t) &= \prod_{i=1}^{n} \varphi_{\frac{X_i}{\sqrt{n}}}(t) && (1)\\
&= \prod_{i=1}^{n} \varphi_{X_i}\left(\frac{t}{\sqrt{n}}\right) && (2)\\
&= \prod_{i=1}^{n} \varphi_X\left(\frac{t}{\sqrt{n}}\right) && (3)\\
&= \left[\varphi_X\left(\frac{t}{\sqrt{n}}\right)\right]^n && (4)
\end{aligned}$$

where (1) above is derived from calculating the MGF of a sum of independent random variables, (2) is because the X_i are identically distributed and have a common MGF $\varphi_X(t)$, and (3) is derived from calculating the MGF of an affine transformation (here a simple scaling) of a random variable.

Now we proceed to prove that the MGF above tends to $e^{\frac{t^2}{2}}$, the MGF of the standard normal, as $n \to \infty$. To do so, it will be easier to first prove that the logarithm of the MGF converges to the exponent $\frac{t^2}{2}$, then raise the result as an exponent of e, so we start by letting:

$$\Lambda(t) = \log \varphi_X(t)$$

In the final step of the proof, we will be using l'Hôpital's rule for limits of ratios that look undefined (e.g., $\frac{\infty}{\infty}$ or $\frac{0}{0}$), so we start by calculating the quantities $\Lambda(0)$, $\Lambda'(0)$, $\Lambda''(0)$:

$$\Lambda(0) = \log \varphi_X(0) = \log \mathbb{E}[e^{0 \cdot X_i}] = \log \mathbb{E}[1] = \log 1 = 0$$

$$\Lambda'(0) = \frac{d}{dt}\log\varphi_X(t)\big|_{t=0} = \frac{\varphi'_X(0)}{\varphi_X(0)} = \varphi'_X(0) = \mathbb{E}[X] = \mu = 0$$

$$\Lambda''(0) = \frac{d^2}{dt^2}\log\varphi_X(t)\big|_{t=0} = \frac{\Lambda''(0)\Lambda(0) - [\Lambda'(0)]^2}{[\Lambda(0)]^2} = \frac{\mathbb{E}[X^2] - (\mathbb{E}[X])^2}{1^2} = \sigma^2 - \mu^2 = 1 - 0 = 1$$

We have therefore established that:

$$\Lambda(0) = 0,$$

$$\Lambda'(0) = \frac{d}{dt}\Lambda(t)\big|_{t=0} = \mu = 0,$$

$$\Lambda''(0) = \frac{d^2}{dt^2}\Lambda(t)\big|_{t=0} = \sigma^2 - \mu^2 = 1.$$

Finally, we prove that $\lim_{n\to\infty}\left[\varphi_X(\frac{t}{\sqrt{n}})\right]^n = e^{\frac{t^2}{2}}$ by first taking logarithms on both sides and proving that $\lim_{n\to\infty}\log\{\left[\varphi_X(\frac{t}{\sqrt{n}}\right]^n\} = \frac{t^2}{2}$, i.e., that $\lim_{n\to\infty} n\Lambda\left(\frac{t}{\sqrt{n}}\right) = \frac{t^2}{2}$.

We note that in the calculations below, we will be using l'Hôpital's rule for asymptotics when $n \to \infty$, so we will be replacing ratios by ratios of their derivatives *with respect to* n, which we will express as derivatives of composite functions that will exhibit the derivatives *with respect to* t which we calculated above.

We first observe that $\Lambda\left(\frac{t}{\sqrt{n}}\right) = \log\mathbb{E}\left(e^{\frac{tX}{\sqrt{n}}}\right)$ and:

$$\lim_{n\to\infty}\frac{tX}{\sqrt{n}} = 0 \implies \lim_{n\to\infty} e^{\frac{tX}{\sqrt{n}}} = 1 \implies \lim_{n\to\infty}\mathbb{E}\left(e^{\frac{tX}{\sqrt{n}}}\right) = \mathbb{E}[\lim_{n\to\infty} e^{\frac{tX}{\sqrt{n}}}] = 1 \implies \lim_{n\to\infty}\ln\mathbb{E}\left(e^{\frac{tX}{\sqrt{n}}}\right) = 0$$

This shows that:

$$\lim_{n\to\infty}\Lambda\left(\frac{t}{\sqrt{n}}\right) = 0$$

We can similarly show that:

$$\lim_{n\to\infty}\Lambda'\left(\frac{t}{\sqrt{n}}\right) = 0$$

We will use these two results in the calculation of the limit below, where we will replace ratios of quantities tending to 0 as $n \to \infty$ with ratios of their derivatives with respect to n:

$$\lim_{n\to\infty} n\Lambda\left(\frac{t}{\sqrt{n}}\right) = \lim_{n\to\infty} \frac{\Lambda\left(\frac{t}{\sqrt{n}}\right)}{\frac{1}{n}} \tag{5}$$

$$= \lim_{n\to\infty} \frac{\frac{d}{dn}\Lambda\left(\frac{t}{\sqrt{n}}\right)}{\frac{d}{dn}\left(\frac{1}{n}\right)} \tag{6}$$

$$= \lim_{n\to\infty} \frac{\frac{-t}{2}n^{-\frac{3}{2}}\Lambda'\left(\frac{t}{\sqrt{n}}\right)}{\frac{-1}{n^2}} \tag{7}$$

$$= \lim_{n\to\infty} \frac{t\Lambda'\left(\frac{t}{\sqrt{n}}\right)}{\frac{2}{\sqrt{n}}} \tag{8}$$

$$= t\lim_{n\to\infty} \frac{\frac{d}{dn}\Lambda'\left(\frac{t}{\sqrt{n}}\right)}{\frac{d}{dn}\left(\frac{2}{\sqrt{n}}\right)} \tag{9}$$

$$= t\lim_{n\to\infty} \frac{t\left(-\frac{1}{2}\right)n^{-\frac{3}{2}}\Lambda''\left(\frac{t}{\sqrt{n}}\right)}{(-1)n^{-\frac{3}{2}}} \tag{10}$$

$$= \frac{t^2}{2}\lim_{n\to\infty} \Lambda''\left(\frac{t}{\sqrt{n}}\right) \tag{11}$$

$$= \frac{t^2}{2}\Lambda''(0) \tag{12}$$

$$= \frac{t^2}{2}\cdot 1 \tag{13}$$

$$= \frac{t^2}{2} \tag{14}$$

In the sequence above, step (6) is using l'Hôpital after we've shown that numerator and denominator both have a limit of 0 as $n \to \infty$, step (7) takes the derivative with respect to n by expressing it for the numerator via the derivative with respect to t, step (8) is a regrouping of terms, step(9) is again using l'Hôpital after we've shown that numerator and denominator both have a limit of 0 as $n \to \infty$, step (10) is similar to step (7), step (12) is simply recognizing that $\lim_{n\to\infty} \frac{t}{\sqrt{n}} = 0$, and step (13) recalls the previously calculated value of $\Lambda''(0)$.

We have therefore shown that:

$$\frac{t^2}{2} = \lim_{n\to\infty} n\Lambda\left(\frac{t}{\sqrt{n}}\right) = \lim_{n\to\infty} n\log\varphi_X\left(\frac{t}{\sqrt{n}}\right) = \lim_{n\to\infty} \log\left[\varphi_X\left(\frac{t}{\sqrt{n}}\right)\right]^n = \lim_{n\to\infty} \log\left[\varphi_{\left(\sum_{i=1}^n \frac{X_i}{\sqrt{n}}\right)}(t)\right]$$

By continuity of the exponential function, the limit of the exponential of the right hand side is equal to the exponential of the limit, and we have:

$$\lim_{n\to\infty}\left[\varphi_{\left(\sum_{i=1}^n \frac{X_i}{\sqrt{n}}\right)}(t)\right] = e^{\frac{t^2}{2}}$$

□

We have shown that the MGF of the normalized sum of a sequence of *i.i.d.* random variables converges to the MGF of a standard normal as $n \to \infty$. The proof above showed the result for the case $\mu = 0$, $\sigma^2 = 1$, and using the results for the MGF of affine transformations of random variables generalizes the result to the case of arbitrary $\mathbb{E}[X] = \mu$ and $\mathrm{Var}[X] = \sigma^2$.

We now invoke the *Lemma "On Going Back from MGF to CDF"*, and we can state that

the CDF of the standardized sum of *i.i.d.* random variables with finite mean and variance converges as $n \to \infty$ to the CDF $F_{\mathcal{N}(0,\,1)}(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}\, dt$ of the standard normal $\mathcal{N}(0,\,1)$.

With the extra condition of finite absolute third moment of the *i.i.d.* random variables, we can also state (but have not proven) that the PDF of the standardized sum of *i.i.d.* random variables converges as $n \to \infty$ to the PDF $f_{\mathcal{N}(0,\,1)}(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ of the standard normal $\mathcal{N}(0,\,1)$.

7. FUN EXAMPLES OF THE CLT

We look at two tangible, entertaining, examples of the CLT in action.

Galton Board. The Galton Board, also known as the Quincunx or "Bean Machine" is a game that illustrates the CLT in action (specifically, the De Moivre-Laplace Theorem) by means of the convergence of the sum of Binomially distributed random variables to a Normal distribution.

The game has a vertical board with a lattice of pegs inserted in a pattern of interleaved rows, in such a way that an object dropped from above and hitting a peg has a given chance of being deflected to the left or to the right of the peg. Counting from the top, the first row needs to have one peg, the second row needs two pegs disposed below and on either side of the top peg, the third row needs three pegs, and so on until a last row of n pegs at the bottom.

Little beads or beans are dropped down a vertical chute located above the top row's peg, and at each peg they hit, they get deflected left or right with a pre-set probability p, where p is typically equal to $\frac{1}{2}$, although one can also have asymmetric pegs that bias the deflection to one side, the way a biased coin would.

At the end of their drop, the beans gather at the bottom in one of $n+1$ narrow bins. The left-most bin #0 gathers the pile of balls that have bounced 0 times to the right of the n pegs they have hit on their way down (i.e., they have had n left-side bounces), bin #1 gathers the balls that have bounced once to the right and $n-1$ times to the left, and – for $0 \leq i \leq n$ – bin #i gathers the balls that have experienced i right-side deflections and $n-i$ left-side deflections.

The outcome (bin number where it landed) for each bean follows a Binomial $\mathcal{B}(n\,p)$ distribution, and when we consider the total of a large number of beans, we are in effect looking at the distribution of a sum of Binomials.

Repeating the large-count drop several times exhibits a large majority of the times a bell shape for the vertical pilings of beans, characteristic of a normal distribution!

Depressed Stair Steps. Old and highly visited monuments such as centennial temples or museums, especially those with narrow stairs or entrance stepping stones, show a smooth pattern of wear where the central part of the stone is at a vertical depression from the sides of the stone.

It turns out that a careful look at the shape of the stone reveals an inverted (upside down) bell shape characteristic of the normal distribution, showing the CLT in action!

If a person stepping on the stone has a few grains of sand attached to their soles, then the grind from the sand grains attached to a shoe sole causes some wear of the stone at the point where the footstep landed.

We can model each visitor's footstep as following a uniform distribution over the width of the stepping stone, so that if each footstep erodes its landing spot by a few microns, the total erosion at a given

point of the stone is the sum of the effects from all the footsteps that have landed at that spot after centuries of visitors stepping in and out. In particular, we expect the erosion to be proportional to the footsteps that have landed there (we could make the model more complex by modeling another independent variable for the amount of sand on a person's sole, but we won't...).

We overlay a one-dimension coordinate system on the lateral dimension of the stone, with the midpoint at 0 and the ends at $\pm L$, where $2L$ is the lateral dimension of the stone. The uniform distribution over that interval $[-L, L]$ has PDF $f(x) = \frac{1}{2L}$, mean $\mu = 0$, and variance $\sigma^2 = \int_{-L}^{L} \frac{1}{2L} x^2 dx = \frac{L^2}{3}$.

By the CLT, we expect the standardized sum:

$$\frac{X_1 + \cdots + X_n - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{X_1 + \cdots + X_n}{\frac{L}{\sqrt{3n}}}$$

to converge in distribution towards a standard normal $\mathcal{N}(0, 1)$. The sum $X_1 + \cdots + X_n$ would therefore follow a normal distribution $\mathcal{N}(0, \frac{L^2}{3n})$, i.e., with standard deviation $\sigma = \frac{L}{\sqrt{3n}}$.

The height of a normal distribution at its mode is given by $\frac{1}{\sqrt{2\pi}\sigma}$, which in our case, would be:

$$H = \frac{1}{\sqrt{2\pi}\frac{L}{\sqrt{3n}}} = \frac{\sqrt{3n}}{\sqrt{2\pi}L}$$

We see that the maximal erosion is proportional to the square root of the number of visitors over the centuries. It is also inversely proportional to the lateral width of the stone, which makes sense as the narrower the stone, the more concentrated the steps will laterally be, and the higher the "hit rate" on the central part of the stone.

8. A FINAL POINT ON CONVERGENCE OF RANDOM VARIABLES

We saw different modes of convergence of random variables through two examples in this paper.

Convergence in Distribution. The Central Limit Theorem is an example of convergence in distribution, which is in fact a weak form of convergence. We have seen that this form of convergence makes a statement about the CDFs of a sequence of random variables converging to a target CDF.

Notably, convergence in distribution does *not* guarantee uniform convergence over the entire domain of the reals, but it does state convergence at each point on the real line at which the target (limit) CDF is continuous. This leaves room for convergence to happen at different rates at different points on the real line, and – in the case of the CLT – it is indeed the case that the tails may converge more slowly than the center part of the distributions.

Convergence in distribution of the sequence X_n to a target distribution associated with random variable X_∞ can be denoted as:

$$X_n \xrightarrow{d} X_\infty.$$

Convergence in Probability. The Weak Law of Large Numbers (WLLN) is an example of convergence in probability which is a stronger form of convergence than convergence in distribution.

Convergence in probability conveys that for any tolerated difference ε that we're willing to consider, the probability that the absolute difference between a random variable X_n in the sequence and the target random variable X_∞ exceeds ε tends to 0 as $n \to \infty$.

This means that we can pick an arbitrarily small tail probability $\eta > 0$, and there will be a value

of n beyond which all random variables in the sequence differ from the target random variable by less than ε, with probability at least $1 - \eta$. Formally, it implies:

$$\forall \varepsilon > 0 : \forall \eta > 0, \exists\, N_\eta \in \mathbb{N} \text{ such that } n \geq N_\eta \implies \mathbb{P}(|X_n - X_\infty| < \varepsilon) > 1 - \eta.$$

This form of convergence is represented as:

$$X_n \xrightarrow{p} X_\infty.$$

If a sequence of random variables converges to a limit random variable in probability, then it will also converge to that limit in distribution, but the converse is not necessarily true. The converse is true, however, if the limit is a constant, i.e., a degenerate random variable concentrated on one value (which it is not – in the case of the CLT for example – because the target is a standard normal).

Convergence Almost Surely. There is an even stronger form of convergence known as convergence almost surely, or also convergence with probability 1. A sequence of random variables $X_1, \ldots, X_n$ converges to a limit random variable X_∞ almost surely if the set of events $\omega \in \Omega$ over which the convergence is not true has probability 0. This form of convergence can be stated as:

$$\mathbb{P}\left(\lim_{n\to\infty} X_n = X_\infty\right) = 1.$$

or also:

$$\mathbb{P}\left(\omega \in \Omega : \lim_{n\to\infty} X_n(\omega) = X_\infty(\omega)\right) = 1.$$

This form of convergence is denoted as:

$$X_n \xrightarrow{a.s.} X_\infty.$$

This form of convergence looks at random variables as functions from the event space Ω to $\mathbb{R}$, and considers the events for which the sequence $X_n(\omega) \to X_\infty(\omega)$ versus those where convergence does not occur. Almost sure convergence says that the events for which convergence does not occur have probability 0.

An example of almost sure convergence of random variables would be a player who earns one dollar each time she flips a coin and gets a Head, and stops receiving money at the first flip that results in Tails. Let X_in be the random variable of the amount earned by the player at the in^{th} flip of the coin. Then, this random variable of amount earned per flip converges almost surely to the constant 0 (random variable equal to 0 with certainty), because the probability of a Tail never happening has probability 0. Yet, for any finite $N \in \mathbb{RN}$, there is a positive (non-zero) probability that all the X_n, with $1 \leq n \leq N$ are equal to 1, from the non-zero probability of having a run of N Heads.

Strong Law of Large Numbers – SLLN. For completion's sake, we mention here that there is a Strong Law of Large Numbers which states that the sample mean of a sequence of $i.i.d.$ random variables converges almost surely, i.e., with probability 1, to the common expected value μ of the random variables in the sequence.

ACKNOWLEDGEMENTS

I would like to express my gratitude to Sawyer Robertson for his guidance and suggestions that have improved the structure and readability of this paper. I would also like to sincerely thank Simon Rubinstein-Salzedo for the inspiration that he instills in me to pursue and research mathematical questions, even (or especially) when they seem difficult and challenging. Lastly, I would like to acknowledge and thank my peers in the Euler Circle: I have benefited from collaborating with them, and learned from their questions and comments throughout the class.

REFERENCES

References

[1] Subhash Bagui and K.L. Mehra, *Convergence of binomial to normal: Multiple proofs*, (2017), https://doi.org/10.12988/imf.2017.7118.

[2] Steven Dunbar, *Topics in probability theory and stochastic processes*, https://www.math.unl.edu/~sdunbar1.

[3] Matt E, *Cool examples of the central limit theorem in action*, math.stackexchange, https://math.stackexchange.com/questions/38825/cool-examples-of-the-central-limit-theorem-in-action.

[4] Yuval Filmus, *Two proofs of the central limit theorem*, https://www.cs.toronto.edu/~yuvalf/CLT.pdf.

[5] Charles M. Grinstead and J. Laurie Snell, *Introduction to probability*, 2nd rev. ed. ed., American Mathematical Society, Providence, Rhode Island, 1997.

[6] Krishna Jagannathan, *Ee5110: Probability foundations for electrical engineers*, (2015), https://nptel.ac.in/content/storage2/courses/108106083/lecture28_Convergence.pdf.

[7] Vlad Krokhmal, *Introductory probability and the central limit theorem*, http://www.math.uchicago.edu/~may/VIGRE/VIGRE2011/REUPapers/Krokhmal.pdf.

[8] Hossein Pishro-Nik, *Introduction to probability, statistics, and random processes*, Kappa Resources, LLC, 2014, https://www.probabilitycourse.com/chapter7/7_2_0_convergence_of_random_variables.php.

[9] John Tsitsiklis, *Probabilistic systems analysis and applied probability*, (2010), https://ocw.mit.edu.

[10] Larry Wasserman, *All of statistics: A concise course in statistical inference*, Springer, 2004, http://www.stat.cmu.edu/ larry/=stat325.01/chapter5.pdf.