The Mathematics Enthusiast

Volume 14
Number 1 *Numbers 1, 2, & 3*

Article 6

1-2017

# Studying "moments" of the Central Limit theorem

Benjamin A. Stark

# Studying "Moments" of the Central Limit Theorem

Benjamin A Stark[1]
University of Montana

**ABSTRACT**: The central limit theorem ranks high amongst the most important discoveries in the field of mathematics over the last three hundred years. This theorem provided a basis for approximation that turned the question of reaction into the art of prediction. This paper aims to map a course for the history and evolution of the famed theorem from its' initial origins in 1733, from Abraham de Moivre's inquiries to the most recent expressions of the theorem. The journey encompassing central limit theorem includes reformations of definition, relaxing of important associated conditions, and numerous types of rigorous proofs.

**Keywords**: Probability theory, Approximation, Normal Distribution, Independence, Random variables, Convergence, Binomial Distribution, Standard Normal Distribution, Method of Moments.

## 1. Introduction: A Basis For Approximation

Consider for a moment what it is you know about the field of probability and statistics. What does the study of probability theory and statistical methods really entail? Collect the image of this field in your mind, what do you see? When I do so, I see probability density functions, distributions, random variables, set theory, expectations, and many more swirling thoughts. At the root of all statistics and probability theory, there is a common goal, analyze the data at hand

[1] sbenjamin.stark@umconnect.umt.edu

*The Mathematics Enthusiast,* **ISSN 1551-3440, vol. 14, nos1, 2&3**, pp. 53 – 76

and predict certain outcomes. With this common goal in mind the question then becomes abstract yet clear, how do we carry out such processes of approximation? For hundreds of years mathematicians pondered the art of approximation. The developing answer to this gigantic range of approximation problems was the birth of the central limit theorem. Henk Tijms, highly esteemed author of "Understanding Probability", commented that the central limit theorem "is without a doubt the most important finding in the fields of probability theory and statistics." (Tijms 2012).

At the very core of the central limit theorem, there is a simple goal aimed to answer a difficult question. The most basic form of the result is as follows; when we have a large number of independent random variables, the central limit theorem helps calculate how probable a certain deviation is away from the sum of said random variables. Jarl Waldemar Lindeberg and Paul Lèvy presented the most common form of the central limit theorem in 1920(Degroot 2012). This form of the central limit theorem, henceforth called the Lindeberg- Lèvy central limit theorem, is the version taught to introductory statistics students worldwide. Their formulation of the theorem is as follows, "If a large random sample is taken from any distribution with mean $\mu$ and variance $\sigma^2$, regardless of whether this distribution is discrete or continuous, then the distribution of the sample mean will tend to a normal distribution with mean $\mu$ and variance $\sigma^2/n$." (DeGroot 2012). Let us break this result down into pieces, as each piece also has crucial conditions that need to be met. A random sample is simply a way of choosing samples from a population such that any sample is just as likely as another to be picked from the population. We need a random sample of random variables, variables that take on different values with different probabilities, to utilize the central limit theorem. The statement also mentions we need a large sample size; just how large this random sample must be will be discussed later. Lindeberg and

Lèvy's central limit theorem also includes information about the distribution of the sample mean, the sample mean of course being the sum of all the random variable values divided by the number of random variables taken in the sample. This is truly an astounding result! The sample mean of a distribution is commonly seen as some sort of constant, when in reality it is really more of a moving target, a random variable of its' own. The history surrounding the central limit theorem, including the development of the theorem, the criticism of some of the results, and the famous names that have laid their hands upon this idea, really results in an amazing journey through the history of mathematics. We start of course at the very roots, with a very simple idea.

**2. de Moivre Cracks the Code**

I propose a hypothetical scenario for consideration; let us say a man tosses a coin ten times. Of these ten tosses the man observes six heads and four tails. Is this claim believable? Let us run this experiment with the same man and the same coin again, however the man will now flip the coin ten thousand times. His tiresome task may take some time; however at the experiments' end the man records 5,227 heads and the rest tails. Would you believe that in ten thousand flips of a fair coin an individual can expect to see 5,227 heads?

If an answer escapes you, do not feel unintelligent or slow, questions of this very nature have baffled mathematicians for centuries. The earliest answers to such questions were brute and somewhat ineffective. Then in 1733, a breakthrough occurred. The French-born English mathematician Abraham de Moivre postulated the very first version of the central limit theorem in an attempt to approximate the distribution of the number of heads resulting from large numbers of fair coin tosses. This result was primitive compared to modern day methods, but what is most important is how de Moivre though to gauge the true probability of an event happening that was beyond the scope of mere experimentation. What de Moivre was essentially

doing was looking at a sequence of Bernoulli trials (trials that either have only two possible outcomes) and using a normal distribution to approximate the distribution of the number of heads resulting from successive tosses of a fair coin.

Recall the question at the beginning of the section; is it to be expected to get 5,227 heads in ten thousand successive tosses of a fair coin? The short answer is no, and the explanation is given by the central limit theorem. To prove this result to be either valid or unbelievable, de Moivre's idea was to look at the probability of more than 5,227 heads being flipped in ten thousand tosses, treating each flip as a Bernoulli trial denoted $X_i$. The whole point of a Bernoulli trial is to have a "yes or no" outcome where $X_i$ is 1 if the result is a "yes" and $X_i$ is a 0 if the result is a "no". Denoted this way, probabilities become much more simple as they are reduced to a binary response. Denoted this way, one could easily find a standard deviation for a result of 5,227 heads or more away from the expected value of heads in ten thousand flips, which is 5,000. Let us now see this analytically; we must calculate the probability of 5,227 or more heads appearing in 10,000 tosses of a fair coin (note that we do not seek the exact probability of 5,227 heads occurring in 10,000 tosses, an important distinction!). To carry out this computation we need two components, the expected value of our random variables and the standard deviation of our random variables. To validate the usage of the central limit theorem, we must view each toss of the coin as a random variable with two possible outcomes. Let the number of times the coin lands heads be denoted by the sum: $X_1, X_2, X_3, \dots, X_{10,000}$. Denoted in this fashion each flip of the coin becomes a Bernoulli random variable (as described earlier) and we can think of each $X_i$ as:

$$X_i = \begin{cases} 1 \; if \; the \; ith \; toss \; is \; a \; head \\ 0 \; if \; the \; ith \; toss \; is \; a \; tail \end{cases}$$

From pure intuition, we know that we expect to see 5,000 flips of the fair coin result in a head, as a head will be flipped exactly 50 percent of the time. We will also need to find the standard deviation for this computation, which can be found in general for a Bernoulli random variable as the square root of the probability of success multiplied by one minus the probability of success (Degroot 2012), we then take that quantity and multiply it by our 10,000 trials, in our case it is computed as: $\sqrt{10000 * \left(\frac{1}{2} * \left(1 - \frac{1}{2}\right)\right)} = 50$, with $\frac{1}{2}$ as our probability of success. We can then "standardize" our result of 5,227 heads obtained to get a distance from our observed value to our expected value as follows:

$$\frac{Observed\ Value - Expected\ Value}{Standard\ Deviation} = \frac{5{,}227 - 5{,}000}{50} = 4.54$$

The computed distance from the expected value in this particular scenario is 4.54 standard deviations away from the mean. This is an extremely unlikely outcome! To put an approximate percentage on this event occurring, the chance of this result or a more extreme outcome happening is approximately 1 in 3.5 million(Tijms 2012). As a result, such a claim would be largely rejected or regarded as a miracle of chance.

De Moivre's early version of the central limit theorem wasn't purely experimental; In fact de Moivre was able to show that by point probability (the probability of one particular occurrence of an event) the binomial distribution converges to a normal probability. His proof required only fundamental analysis and algebra manipulation, no probability theory was required, largely due to the fact that probability theory was just being born itself! In one of de Moivre's masterpieces, "The Doctrine of Chance", he displays a long, tedious proof that results in the following statement:

$$\binom{n}{k} p^k q^{n-k} \approx \frac{1}{\sqrt{2\pi npq}} e^{\frac{-(k-np)^2}{2npq}}$$

Where the left hand side of the equation above is the binomial distribution for a certain probability $p$, such that $0 \leq p \leq 1$ and $q = 1 - p$. For those familiar with distribution theory and probability theory, the right hand side of this equation is the integrand of the normal distribution, or a certain form of it at the least. The experimental findings of Abraham De Moivre would largely hinge off this result and he would go on to document many papers on his findings. However, as many mathematicians in history have unfortunately uncovered, some discoveries are not fully appreciated until far after their unveiling. Such is the case with Abraham de Moivre's original postulate of the central limit theorem.

**3. "Insignificant findings"**

At the time of its conception, de Moivre's discovery of approximations of normality and the beginning foundations of the central limit theorem were severely underappreciated. Of course this is understandable, at a time when approximation methods were considered to be mere fantasies the idea that you can simply "guess" an accurate result based of some kind of distribution was far-fetched. Given the lack of enthusiasm behind de Moivre's discoveries of normality at the time the results were largely ignored. It was not until 1812 that the famous French mathematician Pierre-Simon Laplace "rescued it from obscurity" (Tijms 2012) in his famous work *Theorié Analytique des Probabilités*. Laplace would expand a great deal on de Moivre's theory of approximating with a normal distribution. Laplace would broaden the prior results to also show that the binomial distribution (the sum of many Bernoulli trials) could also be well approximated by a normal distribution. This result is crucial, as it was the first step to showing that the normal distribution comparison can be applied to more than just one distribution. This idea would pave the way for Pierre-Simon Laplace to create a more formal statement on approximating the binomial distribution with the normal distribution. Realizing the

gravity of his findings, Laplace formed a set of rules to follow when trying to approximate a random distribution by way of a normal curve. (Chernoff and Sriraman 2014). Essentially a random variable can be well approximated by a normal distribution if a large enough sample of the random variable is observed. This statement is key to the central limit theorem; the entire result is based off the fact that we have a significantly large number of trials to work with. Intuitively this condition should make a fair amount of sense. Statisticians can get a better idea of how a distribution behaves if there exists and abundant amount of observations from the distribution. With too few observations, the result breaks down completely. To illustrate this point, because it is rather important, imagine flipping a fair coin 3 times, and getting exactly 3 heads from successive flips. Based off this experiment, the naïve statistician would denote the probability of obtaining heads on any particular flip of the fair coin to be 1, as in every flip would result in a head being obtained. However the realist would know this result to be completely false! As you continuously flip the coin again and again and again you would expect the coin to turn up heads roughly 50 percent of the time, in other words you would expect the probability of obtaining a head on any particular flip of the coin to be .5. Ultimately, Laplace's form of the central limit theorem is shown below. For it's time, it was one of the most outstanding results in probability theory.

**4. Laplace's Central Limit Theorem via Generating Functions**

Here we will look at the Laplacian approach to the central limit theorem using generating functions and rigorous algebra. This proof has some complexity to it; the following explanation comes from Hans Fischer's, "History of the Central Limit Theorem". The set up to the problem is important, suppose we have the random variables $X_1, X_2, X_3 \ldots X_n$, which each have means of zero. We also assume that each $X_i$ can take on a value of

$\frac{k}{m}$ $(m\ is\ a\ natural\ number, k = -m, -m+1, \ldots, m-1, m.)$, meaning each $X_i$ can take on a value somewhere in between negative one and one. Now we assume that each $X_k$ is paired with some probability of occurrence $p_k$, we then want to calculate the probability that the sum of these random variables is in between some two numbers, mathematically represented as follows:

$$P_j = \Pr\left(\sum_{l=1}^{n} X_l = \frac{j}{m}\right), \{j: (-nm < j < nm)\}$$

Where $\frac{j}{m}$ is some arbitrary value. Laplace made use of the generating function,

$$T(t) = \sum_{k=-m}^{m} p_k\, t^k$$

Where $P_j$ is equal to the coefficient of $\mathrm{t}^j$ in $[T(t)]^n$. Now a generating function is defined to be a formal (meaning algebraically defined instead of analytically) power series whose coefficients give the sequence (Miller 2009). This in essence, was the general method that de Moivre employed, however where Laplace improved on this method is to substitute in the value of $e^{ix}$ for $t$ (where $i = \sqrt{-1}$). This method is really a sort of introduction to what is modern day called "characteristic functions". Now the prior form of the integral approximation for the probability mentioned earlier became far more approachable as follows, the original form:

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} e^{itx} e^{isx}\, dx = \delta_{st} \quad (s, t \in \mathrm{Z})$$

It then follows from the generating function and Laplace's substitution of $t$ that:

$$P(j) = \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{ijx}\left[\sum_{k=-m}^{m} p_k\, e^{ikx}\right]^n dx$$

This expression is now an operable one; we can expand $e$ into its power summation form (with respect to x) within the integral:

$$P(j) = \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{ijx}\left[\sum_{k=-m}^{m} p_k\left(1 + ikx - \frac{k^2x^2}{2} + \frac{ik^3x^3}{3!} + -\cdots\right)\right]^n dx$$

Now, due to the fact this is a generating function, we can take into consideration that $\sum_{k=-m}^{m} p_k\, k = 0$ and using the substitution $m^2\sigma^2 = \sum_{k=-m}^{m} p_k\, k^2$, we then get:

$$P(j) = \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{ijx}\left[\sum_{k=-m}^{m} p_k\left(1 - \frac{m^2\sigma^2x^2}{2} - iAx^3 - \cdots\right)\right]^n dx$$

Where $A$ is treated as some constant value depending on $\sum_{k=-m}^{m} p_k\, k^3$. We then focus on the term inside the sum, we define $log(z)$ as:

$$log(z) = \log(1 - \frac{m^2\sigma^2x^2}{2} - iAx^3 - \cdots)$$

As a power series of $x$ this quantity is then:

$$log(z) = -\frac{m^2\sigma^2nx^2}{2} - iAnx^3 - \cdots$$

Now we use the base $e$ to transform the above statement into:

$$z = e^{-\frac{m^2\sigma^2nx^2}{2} - iAnx^3 - \cdots} = \; e^{-\frac{m^2\sigma^2nx^2}{2}}(1 - iAnx - \cdots)$$

We then return to our integration, we transform the variable of integration as $x = \frac{y}{\sqrt{n}}$ the resulting integral is a bit nasty,

$$P(j) = \frac{1}{2\pi\sqrt{n}}\int_{-\pi\sqrt{n}}^{\pi\sqrt{n}} e^{ij\frac{y}{\sqrt{n}}}\left(e^{-\frac{m^2\sigma^2ny^2}{2}}\left(1 - \frac{iAy^3}{\sqrt{n}} - \cdots\right)\right)dy$$

This calculation is clearly very advanced; it takes a very high level of mathematical knowledge to interpret these results. However we have now reached the point where intuition can start to play a part in the solution to our problem. We notice that consecutive terms in the inner sum of the integral have denominators of square root of n, as do the limits of integration. This result was

one of Laplace's great improvements on De Moivre's work. The assumption is now that as n gets arbitrary large, these before mentioned terms become almost negligible to the solution, and the limits of integration now stretch to the entire real line thus the integral can be simplified even further. However now, because we make assumptions of n infinitely large, our exact formula becomes an approximation.

$$P(j) \approx \frac{1}{2\pi\sqrt{n}} \int_{-\infty}^{\infty} e^{ij\frac{y}{\sqrt{n}}} e^{-\frac{m^2\sigma^2y^2}{2}} dy$$

This last integral is shown by Laplace in to be equal to

$$\frac{1}{m\sigma\sqrt{2\pi n}} e^{\frac{-j^2}{2m^2\sigma^2 n}}$$

We finally reach the result, which would become Laplace's closed form central limit theorem, summing the statement above over all $j \in (mr_1\sqrt{n}, mr_2\sqrt{n})$:

$$\Pr\left(r_1\sqrt{n} \leq \sum X_l \leq r_2\sqrt{n}\right) \approx \int_{mr_1}^{mr_2} \frac{1}{m\sigma\sqrt{2\pi}} e^{\frac{-x^2}{2\sigma^2m^2}} dx \overset{m=1}{\Longrightarrow} \int_{r_1}^{r_2} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-x^2}{2\sigma^2}} dx$$

This restriction of $m = 1$ lets us look at the values $\frac{j}{m} : -1 \leq \frac{j}{m} \leq 1$. Note the final form of this expression; to statisticians it should look fairly familiar. It is the form of a normal distribution with a mean of zero and some finite variance $\sigma^2$. This shows that Laplace had a method for proving that the sum of a sequence of independent and identically distributed random variables is well approximated by a normal distribution under certain conditions.

To mathematical enthusiasts this result is substantial, especially in the mathematical realm of statistics. After all, one of the ultimate goals of statistics in prediction of future events, and Laplace's expansion of de Moivre's ideas took mathematical science one-step closer to predicting outcomes more accurately than ever before. However this result was over-shadowed

in Laplace's time, largely because the mathematics to prove such results was still highly contested. There were no prior formal proofs of these statements; only analytical results based off experiments were presented in the papers of de Moivre and Laplace. In fact many prominent mathematicians formed small back and forth debates on the validity of Laplace's proof. This fact contributed to these findings being buried and squabbled in history for almost a hundred years. It wasn't until 1901, 89 years after Laplace's papers on normal approximation, that the Russian Mathematician Aleksandr Lyapunov gave the central limit theorem it's first rigorous proof. His result is really quite exquisite. However before we reach the golden age of Russian probability theorists, we will spend some time discovering aspects of analysis that were extremely vital to the formation of a central limit theorem. It should be noted that Laplace's technique of using generating functions (shown earlier) did incorporate some analysis concepts, however because it his proof was more centered towards probability rather than randomness and error approximation I mentioned his techniques first.

**5. Mathematical Analysis Gives Structure to Central Limit Theorem**

As mentioned before, the early formations of central limit theorem were predominantly based off of experimental outcomes and a frequentist approach. Data collected by natural scientists from all disciplines were quickly becoming somewhat of a playground for mathematicians such as de Moivre, Laplace, and even Carl Friedrich Gauss the so-called "Prince of Mathematics" saw the usefulness of a normal approximation for large samples of data. This led early probability theorists to somewhat of a plateau. Being that all of this theory was purely based off of hypothesis and data, the question remained, was there a way to mathematically prove that the sums of random variables converge to a normal distribution? The central limit theorem really gained its structural integrity alongside the birth of modern day mathematical

analysis. In particular, the combined studies of finite algebra and epsilon-infinitesimal considerations. Combined, these concepts would lay out a blueprint of sorts for central limit theorem to be thoroughly constructed (Fischer, 2011, pg. 10). These concepts are well above that of introductory statistician, to make them a little clearer, we must consider again what it is we are working with. At it's most basic roots, we are considering a sum of independent random variables from any distribution, and determining the long run behavior of such a distribution. Now what comes to mind when mathematicians talk of "long-run" behavior? Anyone who took introductory calculus would know the answer, limits and infinitesimal analysis. We must remember that the encompassing theory around central limit theorem is based off of inequalities, upper and lower bounds on possible deviations from the expected value of distributions. Thus it would make sense to consider sums at their largest extremes and make comments on their behavior rather than there value at certain stages. So early probability theorists needed this study of the infinite extremes, both infinitely small and large, to solidify this statement for large enough sample sizes, we will see many examples of this throughout the remainder of this inquiry.

As a result, mathematicians who contributed to the development of the central limit theorem also happened to be extraordinary wielders of analysis knowledge. As we saw earlier, Laplace used some forms of analysis of the extremes in his generating function proof of the central limit theorem. The likes of famous probability theorists such as Pafnuty Chebyshev, Andrei Markov, and Lyapunov all follow this analysis-oriented approach to the central limit theorem, and as we will see in far more detail. Going forward in this paper it is important to keep in mind this philosophy of analysis-based interpretation, as it will take our prior talks of approximation and examine them in the limit realm. This conversion of thought leads to

approximations transforming into more exact values. This transition certainly marks the ending of classical probability theory and reveals the modern approach to probabilistic thinking, and is vital to our conversation on the central limit theorem.

It is very worth noting that before the central limit theorem made its way into the realm of probability theory, some of the greatest mathematical minds in history would analyze the theorem as it relates to error bounds outside the concept of random variables. Between the years of 1812 and 1887, notable mathematicians such as Augustin Cauchy, Peter Dirichlet, and Siméon Poisson would look at the central limit theorem. However these famous names were looking at the topic from somewhat of a broad view, merely as a means of error approximation. We will take a look at some interesting methods executed by Poisson; his major contribution to the central limit theorem was the reinforcement of the earlier works of Laplace.

Siméon Poisson had an interesting view of mathematics and nature that drove his research behind the central limit theorem. Poisson's mentality was that mathematical laws govern all processes in the physical and moral world; this frame of mind drove him to derive more exact mathematical analysis (Fischer 2012). Poisson's central limit theorem proofs were quite fascinating. Like many mathematicians do Poisson used multiple different proofs including series expansion, an "infinitistic" approach, and a definition driven analysis. Like Laplace, Poisson was still oriented on using analysis to prove the limit theorem, which would make a fair amount of sense given the rigor of the problem. However what is interesting about Poisson's version of the central limit theorem is the growing utilization of probability theory concepts in his work. Poisson's version of the central limit theorem is as follows (Fischer 2012, pg. 33):

*Let $X_1, X_2, X_3, \dots, X_n$ be a great number of independent random variables with density functions which decrease sufficiently fast as their arguments tend to $\pm\infty$. It is supposed*

*that for the absolute values of* $p_n(\alpha)$ *of the characteristic functions of* $X_n$ (the sample mean) *there exists a function* $r(\alpha)$ *independent of n with* $0 \leq r(\alpha) \leq 1$ *for all* $\alpha \neq 0$ *such that:* $p_n(\alpha) \leq r(\alpha)$. *Then for arbitrary* $\gamma, \gamma'$,

$$\Pr\left(\gamma \leq \frac{\sum_{n=1}^{s}(X_n - E[X_n])}{\sqrt{2\sum_{n=1}^{s} Var[X_n\}}} \leq \gamma'\right) \approx \frac{1}{\sqrt{\pi}} \int_{\gamma}^{\gamma'} e^{-u^2}\, du$$

Poisson's version of the central limit theorem embodies the objective of this paper, to map a course throughout the history of the central limit theorem. To me, Poisson's version of the central limit theorem was a major transitional step out of the classical, Laplacian probability theory and into a more conditional and notation oriented statement. The Poisson central limit theorem signals are large shift in the way mathematicians were thinking about limit theorems, especially in regards to probability theory.

**6. The Russian Influence**

We now move into a more specified talk of the central limit theorem, with the intent of describing its earliest structured proofs in the realm of probability theory and statistics. This takes us to early 20th century Russia, were the works of Pafnuty Chebyshev and Aleksandr Lyapunov provided the structure needed to prove central limit theorem. We first look at the works of Chebyshev, his labor on the central limit theorem would ultimately be labeled as "incomplete", however his ideas and techniques would open the door for other mathematicians including his most famous student, Andrei Markov, to finish and expand the proof. In 1887, Chebyshev published a paper detailing the first attempt at a proof of the central limit theorem utilizing a technique known as the method of moments.

This method, first attributed to Irenée Bienaymé in 1853 for his work on the weak law of large numbers, is a way of proving convergence in distribution by showing convergence

in a sequence of moments (Heyde 2001 pg.134). The method of moments is normally introduced to more experienced statistics students, in our settings it is important we know about this technique to gain an understanding of the way early probability theorists thought about the central limit theorem. Let us now break it down step by step, a moment is a certain form of the expected value of a random variable denoted $E[x^k], k = 1,2,3$ ... For example, $E[x^1]$ is the "first" moment of the random variable x, or simply the expected value of x. Formally, a moment is defined mathematically as follows:

$$E[x^k] = \int_{-\infty}^{\infty} x^k f(x)dx \ \text{, k} = 0,1,2, \dots, \text{n}$$

.

In the simplest way, the moments of a random variable describe how you'd expect certain powers of the function to behave, these values give statisticians invaluable information about sequences of random variables. Considering that the central limit theorem deals with a sequence of random variables, this appears to be the proper way of thinking. The method of moments requires the condition that the distribution of X is completely characterized by it's moments, meaning that it's sequence of moments ($E[x^1], E[x^2], E[x^3]$, ...) is unique to that distribution alone. At the heart of the technique is this, if the distribution has unique moments (as stated above) then if the limit as the number of observations goes to infinity of the sequence of moments converges to a single moment, that sequence of random variables converges to a distribution. This statement can be seen mathematically as follows:

$$\lim_{n\to\infty} E[X_n^k] = E[X^k]$$

This result is still today considered to be spectacular. Using this method, Chebyshev was able to show that the moments of a sequence of random variables converges to that of a moment of a

normally distributed random variable. The result is sufficient to say that any sequence of random variables, regardless of distribution parameters, converges to a normally distributed parameter.

**7.Chebyshev's Central Limit Proof via Method of Moments**

Chebyshev's method of moments approach to the central limit theory was revolution for its time, it differed vastly from the sort of earlier computations relied on by Laplace and de Moivre in the sense that it relied less on clever transformations and tedious algebra and more on definitions and theory. This transition is extremely important, as it signifies the beginning of looking at the central limit theorem from a new perspective. Before Chebyshev used the method of moments to prove central limit theorem, he outlined the concept in very basic terms. The following explanation of his method is from William J. Adams author of "The Life and Times of the Central Limit Theorem". In a paper presented to the Congress of the French association for the Advancement of Science at Lyons in August 1873, Chebyshev presented the following theorem:

*If* $f(x)$ *and* $g(x)$ *are both positive in the interval* $[a, \mathrm{b}]$ *and the functions have the same moments i.e.*

$$\int_a^b f(x)dx = \int_a^b g(x)dx \ ,$$

$$\int_a^b xf(x)dx = \int_a^b xg(x)dx \ ,$$

$$\int_a^b x^{2m-1}f(x)dx = \int_a^b x^{2m-1}g(x)dx \ ,$$

*Then for some v in* $[a, \mathrm{b}]$

$$\left| \int_a^v g(x)dx - \int_a^v f(x)dx \right| \leq \frac{1}{a_2\psi_1^2(v) + \cdots + a_m\psi_{m-1}^2(v)}$$

*where* $\psi_i^2(v), i = 1,2,3 \ldots, m-1$ *are denominators of convergent of the expansion of* $\int_a^b \frac{f(x)}{z-x}x$ *into a continued fraction of the form*

$$\cfrac{1}{a_1 z + b_1 - \cfrac{1}{a_2 z + b_2 - \cfrac{1}{a_3 z + b_3 - \ddots}}}$$

To put this idea in the simplest of terms, the difference between two functions that share the same moments grows arbitrarily small in the continued fraction expansion. With this theorem in mind, Chebyshev would apply his findings fourteen years later in an attempt to prove central limit theorem. Following the same template as Chebyshev's original theorem, we take $a = -\infty, b = \infty$ and work with the function:

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

This function is of a normal form, thus the proof will aim to relate the moments of $g(x)$ to the moments of $f(x)$ by showing they become arbitrarily close to each other for infinitely large moments.

If $g(x) > 0 \; and \; f(x) = \frac{q}{\sqrt{2\pi}} e^{\frac{-q^2x^2}{2}}$ have the same $2m-1$ moments, that is:

$$\int_{-\infty}^{\infty} g(x)dx = 1\,, \int_{-\infty}^{\infty} xg(x)dx = 0\,, \int_{-\infty}^{\infty} x^2 g(x)dx = \frac{1}{q^2}\,, \cdots,$$

$$\int_{-\infty}^{\infty} x^{2m-2} g(x)dx = \frac{1 * 3 \cdots * (2m-3)}{q^{2m-2}}\,, \int_{-\infty}^{\infty} x^{2m-1} g(x)dx = 0$$

Then for all values of $v$:

$$\left| \int_a^v g(x)dx - \frac{q}{\sqrt{2\pi}} \int_a^v e^{\frac{-q^2x^2}{2}} dx \right| \leq \frac{3\sqrt{3}(m^2 - 2m + 3)^{\frac{3}{2}}(q^2v^2 + 1)^3}{2(m-3)^3\sqrt{m-1}}$$

Notice that this upper bound on the difference of the functions is going to zero as $m \to \infty$. In this case we have defined $\psi_z(n) = e^{q^2z^2} * \frac{d^n}{dz^n}\left[e^{\frac{-q^2z^2}{2}}\right]$ with $n = 0,1,\cdots,m-1$ as denominators of the continued fraction expansion of:

$$\frac{q}{\sqrt{2\pi}} \int_a^b \frac{e^{\frac{-q^2x^2}{2}}}{z-x} dx$$

This then shows that the difference between the two functions with the same moments grows arbitrarily small as $m \to \infty$. Thus we have convergence in the sequence of moments of both functions, they converge to the same sequence of moments. In this fashion Chebyshev then formalized his central limit theorem as follows:

*If the mathematical expectations of the variables $U_1, U_2, \cdots, U_n$ are all zero, and if the mathematical expectation of all their powers* (their moments) *are less than some finite bound, then the probability that the sum $U_1 + U_2 + U_3 + \cdots + U_n$ of n of these variables divided by the square root of two times the sum of the mathematical expectations of their squares will be between two arbitrary limits, a and b, approaches:*

$$\frac{1}{\sqrt{2\pi}} \int_a^b e^{-x^2} dx$$

*as n approaches infinity* (Williams, 2009, pg. 47).

Utilizing the earlier explanation of the method of moments, it is then clear to see how Chebyshev constructed his proof. First we define $X_n$ as follows:

$$X_n = \frac{U_1 + U_2 + U_3 + \cdots + U_n}{\sqrt{n}}$$

We then define the function $g_n(x)$:

$$Pr(X_n < x) = \int_{-\infty}^{x} g_n(x)dx$$

Then the moments of $g_n(x)$ converge to the moments of $f(x) = \frac{q}{\sqrt{2\pi}} e^{\frac{-q^2x^2}{2}}$ by the reasoning of the method of moments (shown above). Mathematically speaking, we can then say:

$$\lim_{n\to\infty} \int_{-\infty}^{\infty} x^k g_n(x)dx = \frac{q}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^k e^{\frac{-q^2x^2}{2}} dx = \begin{cases} \frac{1 * 3 * 5 * \cdots * (k-1)}{q^k} & , k \text{ is even} \\ 0 & , k \text{ is odd} \end{cases}$$

This, in essence, is the infrastructure that every single attempted proof of the central limit theorem passed Chebyshev's time would follow. Unfortunately, this result was deemed incomplete, as explained by Chebyshev's famous pupil Andrei Markov in a letter to a colleague, "Unfortunately its (Chebyshev's central limit theorem proof) significance is obscured because of two factors: 1) the complexity of the derivation, 2) the insufficient rigor of the reasoning. The theorem, which Chebyshev proved in the aforementioned memoir, had long been considered true, but it was established by means of extremely loose reasoning. I (Markov) do not say proven since I do not recognize loose proofs unless I perceive opportunities to make them rigorous." Andrei Markov's letters to his colleague A.V. Vasilev, a professor of mathematics at Kazan University, would give several counter examples that show when certain conditions of Chebyshev's proof were not satisfied, the result proved to be incorrect. This "loose" proof structure would influence Andrei Markov to improve the statement given by Chebyshev drastically.

Markov's revision of the central limit theorem proposed by Chebyshev is far more rigorous. From a statistician's point of view, Markov's version of the central limit theorem is an

absolute work of beauty; as such it will be presented it in its entirety. In one letter to his colleague at Kazan University, Markov gives the following form of the central limit theorem:

*Let* $U_1, U_2, \ldots$ *denote random variables and let* $E[U_n^v]$ *denote the expected value of* $U_n^v$. *If*

*(1) The expected value* $E[U_n^v]$ *of* $U_n$ *is 0, n=1,2,…,*

*(2) The kth moments* $E[U_1^k], \ldots, E[U_n^k]$, *of* $U_1 \ , \ldots U_n$ *are bounded*

*(3)* $U_1 \ , \ldots U_n$ *is a sequence of independent random variables*

*(4)* $\lim_{n\to\infty} \frac{E[U_1^2]+\cdots+E[U_n^2]}{n} \neq 0$

*then as n* $\to \infty$

$$P\left[ s < \frac{U_1 + U_2 + \cdots + U_n}{\sqrt{2(E[U_1^2] + \cdots + E[U_n^2])}} < t \right] \to \frac{1}{\sqrt{\pi}} \int_s^t e^{-x^2} dx.$$

It is plain to see that Markov's version of the central limit theorem is much more structured than that of his mentor. The conditions set by Markov provided the rigidity to Chebyshev's central limit theorem, and accounted for certain conditions that would have invalidated the predicated theorem (specifically condition 4 of Markov's central limit theorem). It is also very clear to interpret exactly what Markov was trying to say with this expression. Essentially, given a sum of random variables under certain conditions, as the number of said random variables increases the probability that these values are between two finite values (here s and t) converges in distribution to the standard normal density evaluated between said finite values. It is my hope to convey how vastly this result improved from the earlier versions of the central limit theorem presented by de Moivre and Laplace. Markov's central limit theorem marks the first truly accurate version based on probability theory and closed form expressions.

**8. Modern Transitions**

The central limit theorem carves a very fluid stream of mathematical reason throughout the ages. Beginning with de Moivre's earliest conceptions of approximation and essentially ending with the Russian forming of certain probability theory notations. If you have noticed the continuum characteristics of content in this inquiry, you may have seen a very drastic transition of thinking, a transition that is rather frequently seen in studying the history of certain mathematical concepts. For example, consider de Moivre's idea of the central limit theorem, It was rather vague and lacking justification. Though revolutionary for its time, de Moivre's theories on normality were proved with concrete examples (such as coin flips seen earlier) and loose, approximation-based, mathematical analysis. We then moved to the likes of Pierre-Simon Laplace, who used clever substitutions and generating functions to give a backbone to this idea of approximating random variables with mathematical analysis. In the late $18^{th}$ century several Russian mathematicians gave birth to the concept modern probability theory. Famous men such as Pafnuty Chebyshev, Andrei Markov, and Aleksandr Lyapunov set more rigorous conditions for the random variables themselves, incorporated new notations to answer old problems, and solidified a new way of looking at probability (via moment generating functions) to prove numerous results based off concepts much more simple to understand than complex analysis techniques.

Studying the history of well-known theorems in mathematics divulges certain patterns, certain paths of thought. These are patterns that at first may seem very obvious but raise certain questions after some investigation. For example, in regards to the central limit theorem, the re-occurring question that has echoed through out every proof structure is, "how can this idea become more concrete?" What I find most fascinating about this particular question is that the

several answers offered over the ages have always reflected the mathematical ideals of the era it comes from. Whereas many theorems share this same characteristic of taking on the proofing techniques of the current mathematical acumen, the grand question of the central limit theorem offers a path through the history of math that has been traveled by the greatest mathematicians from numerous branches of the mathematical world. Through the course of this inquiry, I have not touched on the most obvious question surrounding central limit theorem, namely why is it called "central limit theorem". I feel it would be appropriate to conclude with the answer to this question. In 1920, Hungarian mathematician George Pólya referred to this particular theorem in a paper and called it "central" due to its importance in probability theory. The paper was written in German, the title translates into English as, "On the Central Limit Theorem of Calculus and Probability and the Problem of Moments." Pólya's title does embody the character of the theorem accurately, as the central limit theorem is by its very nature at the center of all approximation, and thus it is at the center of every mathematical question regarding "How correct are we?"

**Acknowledgement**

The author thanks Prof. Sriraman for encouraging the line of inquiry pursued in this paper in M429: History of Mathematics, and for finding the appropriate sources.

**9. References**

Adams, W. (2009). *The Life and Times of the Central Limit Theorem (2nd edition).* American Mathematical Society.

Chernoff, E and Sriraman, B. (2014). *Probabilistic thinking: presenting plural perspectives.* Springer Science and Business.

DeGroot, M and Schervish, M. (2010). *Probability and Statistics (4th edition)* (pp. 360-371). Carnegie Mellon University. Addison-Wesley.

Fischer, H. (2011). *A History of the Central Limit Theorem.* Springer Science and Business.

Hald, A. (1998). *A History of Mathematical Statistics from 1750 to 1930*. Wiley Publications.

Heyde, C and Seneta, E. (2001). *Statisticians of the Centuries.* Springer Science.

Miller, S. (2009). *From Generating Functions to the Central Limit Theorem.* Accessed at https://web.williams.edu/Mathematics/sjmiller/public_html/342/handouts/GFtoCLT02.pdf.

Porter, T. (1986). *The Rise of Statistical Thinking 1820-1900.* Princeton University Press.

Pólya, George (1920), "Über den zentralen Grenzwertsatz der Wahrscheinlichkeitsrechnung und das Momentenproblem", *Mathematische Zeitschrift*.

Stigler, S.(1986). *The History of Statistics:The Measurement of Uncertainty before 1900.* Harvard University Press.

Tijms, H. (2012). *Understanding Probability (3rd Edition)* (pp. 162-167). Cambridge University Press.

Walker, H. (1985). *De Moivre on the law of normal probability*. Columbia University. Retrieved March 20, 2016 from http://www.york.ac.uk/depts/maths/histstat/demoivre.pdf.

Watkins, T. *Illustration of the Central Limit Theorem.* San Jose State University. Retrieved March 18, 2016 from http://www.sjsu.edu/faculty/watkins/randovar.htm.