Eindhoven University of Technology

Mathematics and practice in historical perspective

2WH60

The History of the Bell Curve

Author
A.P.O Lefevre

ID Number
1517600

September 27, 2020

Contents

Abstract

In this paper we discuss the following topic from the history of mathematics and its applications: the normal distribution. In order to understand the origin of this distribution we return to the time where mathematicians tried to relate the binomial distribution and the normal distribution, and to find a distribution for measurements' errors. Finally we consider one of the best known results involving the normal distribution: the central limit theorem. In this paper we will cover the above themes with both a mathematical and a historical perspective.

1 Introduction

We give in this section an introduction to the normal distribution with some definitions and properties. The normal distribution, or bell curve, denoted by $\mathcal{N}(\mu, \sigma^2)$, is a continuous two parameters distribution. Its probability density function is $f(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$. We say that a random variable X is normally distributed if $X \sim \mathcal{N}(\mu, \sigma^2)$ where $\mu = \mathbb{E}(X) \in \mathbb{R}$ is the mean of the distribution and $\sigma^2 = \mathrm{Var}(X) > 0$ is its variance. A special case of the normal distribution is the standard normal $\mathcal{N}(0, 1)$ where $\mu = 0$ and $\sigma^2 = 1$.

The normal distribution has very interesting properties. We have that μ is the expectation of the distribution, its median as well as its mode. In addition, the curve of normal distributions are symmetric and centered at μ. [13]

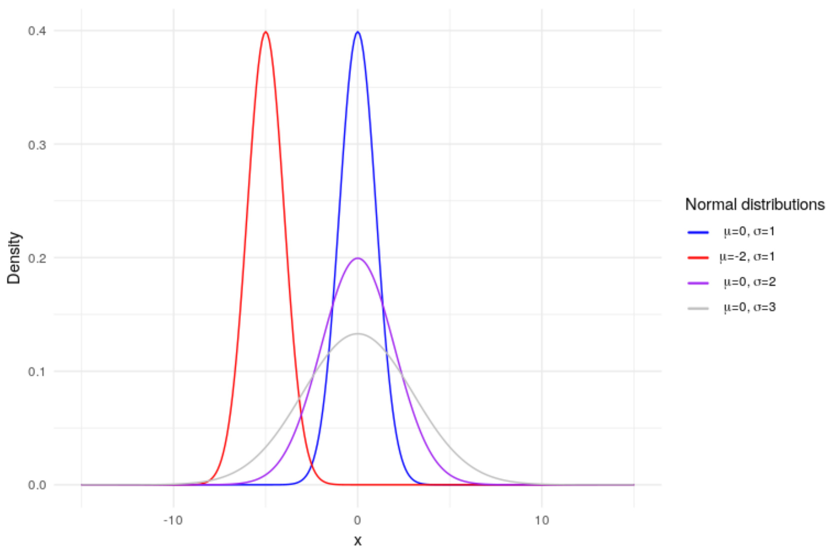

Figure 1: Probability density functions of normal distributions (generated using R)

The normal distribution is the most widely used in the field of statistics. Many researchers in social and natural sciences use the normality assumption in their works. This means that they assume their data are generated from a normal distribution. Throughout the decades, this distribution has been verified to be widespread in many physical and psychological phenomena by techniques such as hypothesis testing and QQ plots.[14]

The normal distribution is very present in nature and in the universe, it hasn't been created but communicated to mathematicians. The bell curve has two different origins, it first loomed when mathematicians tried to approximate the binomial distribution for gambling understanding purposes in the 17th and 18th centuries. It then appeared again when they tried to find a distribution for measurements' errors in the 18th and 19th centuries. In the next sections, we give more details about the birth of the normal distribution and its most important application, the Central Limit Theorem.

2 From binomial distribution to Normal distribution

In this section, we give the first origin of the normal distribution in the 17th and 18th centuries. We discuss the foundation of probability and then the binomial distribution along with its computational difficulties and finally its relation with the normal distribution.

2.1 Foundations of probability theory

The development of probability theory started in the 17th century with the questioning about the number of chances to win gambling games. For example the fair division of stakes when a game is interrupted before its conclusion. It was before the introduction of the word probability.

In 1654, Pierre de Fermat and Blaise Pascal, motivated by the Chevalier de Méré, exchanged several letters in which they shared their discoveries and solved probability problems. Pierre de Fermat was a French lawyer and mathematician from the 17th century. He is best known for his Fermat's principle for light propagation, but also for having studied various fields of mathematics such as analytic geometry, probability and optics. On the other side, still in the 17th century, Blaise Pascal was a French physicist, mathematician, writer and more. In mathematics, especially in probability theory, he is well known for the 'Pascal's triangle' and mathematical induction in dealing with 'his' triangle. He also computed 25 as the smallest number of rolls of 2 dice yielding a probability greater than 1/2 of achieving a double 6 [17]. We don't know much about the Chevalier de Méré's contribution since he did not take part in the main researches. He is the one who motivated gambling problem solving, but for personal purposes only. All of them talk for example about the number of chances before the word probability was introduced and about the value of a game which was 3 years later described as expectation by Christian Huygens in what is called the first book on probability: *De ratiociniis in ludo aleae (1657–60).* [17] and [4]

An example of such problems is the problem of points which was addressed by Chevlalier de Méré to Pascal around 1654:

> *'Two players who have equal chances of winning at each round of a game. The players contribute equally to a prize pot, and agree in advance that the first player to have won a certain number of rounds will collect the entire prize. Suppose that the game is interrupted by external circumstances before either player has achieved victory. How does one then divide the pot fairly?'[21]*

Let's now consider a game where player A still needs a points to win and respectively player B needs b points.

Fermat found a solution which was improved independently by Pascal. He stated that the game will be won in no more than $a + b - 1$ rounds, yielding 2^{a+b-1} versions of the game. The game could be won in less than $a + b - 1$ steps but we imagine that players can continue the game without purpose. Fermat lists all possible outcomes of the game and compute the odds (probability) for each player to win. He divides the stakes accordingly. [11]

Pascal reasons recursively, he doesn't list the outcomes of the entire remaining game but imagine that the game can be played another round before being interrupted. We assume that we know how to split the stakes after this round which could have two possible outcomes. Since players have equal chances of winning a round, we split the difference between the two future divisions evenly.
He finds using the Pascal's triangle [12] that A's stake is proportional to:

$$\frac{\sum_{k=0}^{a-1} \binom{a+b-1}{k}}{2^{a+b-1}}$$

Gambling problems involved combinatorics and repetition of experiments with some winning probability. This lead to the development of the Bernoulli experiments half a century later.

2.2 Binomial distribution and Bernoulli trials

A Bernoulli experiment is a sequence of n random variables $X_1, ..., X_n$ whose values are either 0 for failure or 1 for success. We say that each X_i follow a Bernoulli distribution $Ber(p)$ (with probability of success p) if $P(X = 1) = p$ and $P(X = 0) = 1-p$. In 1713, in his book *Ars conjectandi* published posthumously, Jakob Bernoulli, a Swiss mathematician, solved the problem of computing the probability of r successes in a Bernoulli experiment. He determined that the probability of r such outcomes in n repetitions is equal to the r-th term in the expansion of the binomial expression $(p + q)^n$ (Hence the name *binomial distribution*), where $q = 1 - p$. The binomial expression could be itself the subject of another historical report. But we decided to mention Isaac Newton who, in 1665, generalized the binomial theorem as follow :

$$(p + q)^r = \sum_{k=0}^{n} \binom{n}{k} p^{r-k} q^k$$

There is a great history behind the binomial distribution that we won't cover in this paper as it is not the topic. From the above equation, Jakob Bernoulli derived the binomial distribution $Bin(n, p)$ [15]:

Let S be the number of successes in a Bernoulli experiment of size n where the probability of success is p, we have that,

$$P(i \leq S \leq j) = \sum_{k=i}^{j} \binom{n}{k} p^k (1 - p)^{n-k}$$

2.3 Large computations and the duration of play problem

Bernoulli experiments and the binomial distribution raised new solutions to old problems but also new computational difficulties. An application of these experiments is the duration of play problem that could be stated as follows:

Let A and B be two players each having a counter a and b. In each game, A has probability p of winning while B has a probability of $q = 1-p$. The winner of a game gets a counter from the loser. The game is repeated until a player is ruined and has no more counters. What is the probability that the play ends in n rounds?

To understand this problem, let $u_n(x, a, b)$ with $-a \leq x \leq b$ be the probability that neither A nor B lost the game in n rounds and that B gave x counters to A. B and A are ruined at the n-th game with probabilities respectively $r_n^B(a, b) = pu_n(b - 1, a, b)$ and $r_n^A(a, b) = qu_n(a - 1, a, b)$. The probability that the game end in n rounds is $d_n = r_n^A(a, b) + r_n^B(a, b)$. Finally, the probability that the game lasts at most n rounds is $D_n(a, b) = \sum_{i=0}^{n} d_n = \sum_{i=0}^{n} r_n^B(a, b) + r_n^A(a, b) = R_n^A + R_n^B$. [5]

In 1710, Montmort, a French mathematician, sent a letter to Jakob Bernoulli saying that he has found a general solution for players of equal skill and with the same number of counters. He doesn't directly disclose his formula but gives numerical experiments. We won't mention Montmort's history but only some of his results since he is not a main actor of the normal distribution.
He treats the problem of finding the *median duration of the game*, n for which $D_n(b, b) = \frac{1}{2}$ when $p = \frac{1}{2}$. For example $D_{26}(6, 6) = 0.495 \leq \frac{1}{2} \leq D_{28}(6, 6) = 0.529$. He also believed that $D_{122}(12, 12) \leq \frac{1}{2} \leq D_{124}(12, 12)$. He asks Jakob Bernoulli to go pass over the problem to Nicolas Bernoulli, who was 23 years old at this time, believing that he was capable of solving such problems. In 1711, Nicolas Bernoulli sent the complete solution to Montmort but gave no proof about his formula.

Setting $c = a + b$

$$\begin{aligned} R_n^B(a, b) = p^b \sum_{k=0}^{\infty} (qp)^{kc} \sum_i \binom{n}{i} (p^{n-b-2kc-i} q^i + q^{n-b-2kc-i} p^i) - \\ p^b \sum_{k=0}^{\infty} (qp)^{a+kc} \sum_i \binom{n}{i} (p^{n-b-2kc-2a-i} q^i + q^{n-b-2kc-2a-i} p^i) \end{aligned} \tag{1}$$

where $0 \leq 2i \leq n - b - 2kc$ in the first sum and $0 \leq 2i \leq n - b - 2kc - 2a$ in the second.

He also corrected Montmort on his estimation and finds $D_{108}(12, 12) = 0.499 \leq \frac{1}{2} \leq D_{110}(12, 12) = 0.507$. We decided to add equation 1 to this report to show you the computation difficulties raised by the binomial distribution. In fact theses difficulties are the main reasons why mathematicians such as de Moivre decided to work on the binomial distribution. Later in 1969, Takacs derives a possible proof for Bernoulli's solution which we will not mention here because it is out of scope of this paper. If it caught your attention you can find more information on the topic in [5].

At this point in history, computing binomial distributions could turn out to be really expensive for large values of n. For example Montmort and Nicolas Bernoulli calculated sums of binomial coefficients for n greater than 100 to find the *median duration of the game*.

2.4 De Moivre's approximation to the Binomial distribution

Because of these computational difficulties, the Bernoullis already tried to find an approximation of the formula:

$$\begin{aligned} P_d &= P(|x - np| \leq d) \\ &= \sum_{|x-np| \leq d} \binom{n}{x} p^x q^{n-x} \end{aligned} \tag{2}$$

They managed to find a lower bound for this value, which we refer to as the Bernoulli Theorem [15]. Between 1713 and 1733, de Moivre put aside the analysis of P_d and tried to find an approximation to the binomial distribution.

Figure 2: From left to right: Nicolaus and Jakob Bernoulli, Abraham de Moivre

De Moivre started by getting an approximation for $n!$ and found that, for some constant C,

$$n! \sim C n^{n+\frac{1}{2}} e^{-n}$$

Stirling completed his work by finding $C = \sqrt{2\pi}$ and stated the so-called Stirling formula [6]:

$$n! \sim \sqrt{2\pi n} \left(\frac{n}{e}\right)^n$$

In 1733, de Moivre finally comes up with a result and states without proof (as he was used to do):

$$b(np + d, n, p) \sim \frac{1}{\sqrt{2\pi npq}} e^{-\frac{d^2}{2npq}} \quad d = O(\sqrt{n}) \tag{3}$$

We show below the proof of de Moivre's derivation since it is the premises of the normal distribution.

Proof: From the previous results of his papers, we can imagine what his proof may have been.

From Stirling's formula, we have,

$$b(np, n, p) \sim \frac{1}{\sqrt{2\pi npq}}$$

Using that

$$\frac{b(np, n, p)}{b(np + d, n, p)} = (1 + \frac{d}{np}) \prod_{i=1}^{d-1} \frac{1 + i/np}{1 - i/np}$$

and by Newton's infnite series for $\ln \frac{1+x}{1-x}$

$$\ln \frac{1+i/np}{1-i/np} \sim \frac{i}{npq}$$

we have,

$$\ln \frac{b(np)}{b(np+d)} \sim \frac{d^2}{2npq}$$

which completes the proof. [8]

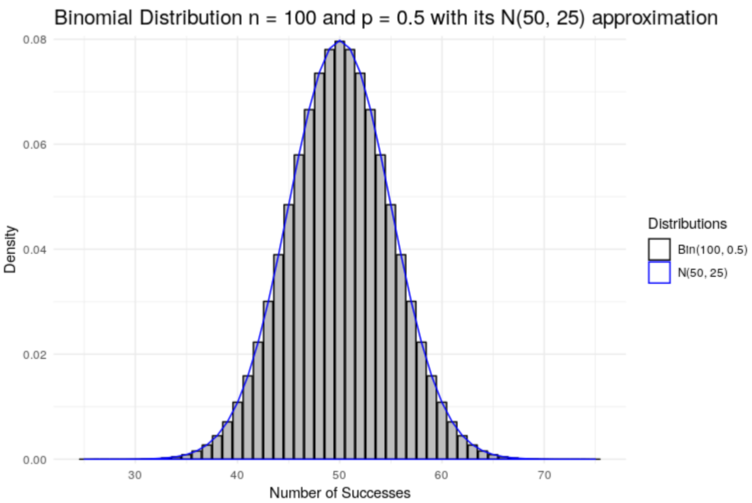

Figure 3: $Bin(n = 100, p = 0.5)$ with its normal approximation $\mathcal{N}(np = 50, npq = 25)$(plot with R)

We have explained so far how the normal distribution was derived for the first time. The main objective was to find an approximation for the binomial distribution which was used to compute the number of successes in an experiment. It was at this time used in many applications such as gambling games and statistics.

De Moivre's work was published in a paper *Approximatio ad Summam Terminorum Binomii* $(a+b)^n$ *in Seriem expansi* in 1733 for private circulation. It had remained concealed from the mathematician's community until 1756, when it was published in his masterpiece *The Doctrine of Chances*, one of the first books on probability theory, after Christian Huygens. Only 6 copies of the *Approximatio* have ever been found. But the first book about games of chance, *Liber de ludo aleae*, was written in 1550 by Girolamo Cardano [17] (even if it was not published before 1663). [6]

3 From Measurement's Errors to the Normal Distribution

We discuss in this section the second origin of the normal distribution: errors in measurements. It started in the 16th century with some observations and intuitions from some astronomers. But most of the theoretical work was done during the 18th and 19th centuries.

3.1 Average and errors in measurements

In the 16th century, the Danish astronomer Tycho Brahe supported by the king was the leading observational astronomer. He collected during his 25 years his observations in a catalog of 1000 stars. In most cases, his errors didn't exceed 61 cm. Tycho knew that his observations were affected by systematic errors such as refraction on lens. He was the first to introduce the arithmetic mean of observations in order to minimize random errors. [7]

Galileo (1564-1642) was the first scientist to reckon that errors were worthy of scientific treatment and concern. We will not give details about his personal achievements since the normal distribution was not his main focus and since he is already well know as the father of modern science for his astonishing discoveries. He states in an informal discussion about errors:

1. There is only one number which gives the distance of the star from the center of the earth, the true distance.
2. All observations are encumbered with errors, due to the observer, the instruments, and the other observational conditions.
3. The observations are distributed symmetrically about the true value; that is the errors are distributed symmetrically about zero.
4. Small errors occur more frequently than large errors.
5. The calculated distance is a function of the direct angular observations such that small adjustments of the observations may result in a large adjustment of the distance.

These properties became the main tool for error treatment and scientists began to use the median and the average of observations in order to capture a more accurate value of their observations. [7]

3.2 Simpson's probability distribution of errors

Thomas Simpson was a British mathematician very interested in astrology and in de Moivre's researches. His work on measurements' errors was read to the Royal Society in 1755. It took the form of a letter to the Earl of Macclesfield *'On the Advantage of Taking the Mean of a Number of Observations in Practical Astronomy'*. His focus was on the mean error instead of the mean observation, which made his studies different from astronomers of his time.

Simpson was the first to give a formal definition of error distribution. We can find a lot of similarities between Simpson's definition of error distribution and the normal distribution, this is why we decided to present his work more in details. He supposed that errors' sizes were in the range:

$$-v, -v+1, \ldots, -1, 0, 1, \ldots, v-1, v$$

And each value has probability proportional to:

$$r^v, r^{v-1}, \ldots, r^1, r^0, r^1, \ldots, r^{v-1}, r^v \quad r > 0$$

or

$$r^v, 2r^{v-1}, \ldots, vr^1, (v+1)r^0, vr^1, \ldots, 2r^{v-1}, r^v \quad r > 0$$

Simpson's interest was mainly on the case where $r = 1$.[1]

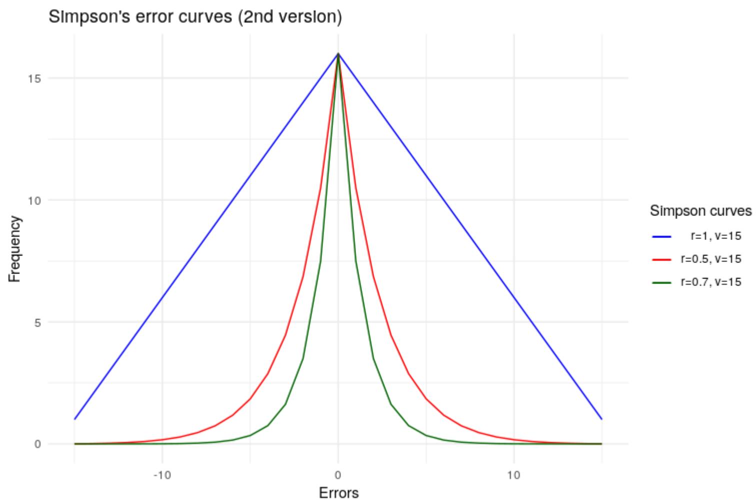

Figure 4: Simpson's error curves (generated using R)

3.3 Laplace's error curves

Pierre-Simon de Laplace was a french polymath of the 18th century. He studied many fields such as mathematics, statistics, physics, astronomy, and philosophy. He is well know for Laplace's equations and Laplace transform, but he also contributed in the development of the normal distribution and its application. Twenty years after Simpson, Laplace started thinking of an error curve $\phi(x)$ in 1772. He listed 3 conditions that the curve of $\phi(x)$ must respect:

1. The curve should be symmetrical around a mean V because it is as probable that errors deviate to the left or the right.
2. The curve must decrease towards the axis since the probability that the observation differs from the true value by an infinite distance is 0.
3. The area under the curve must be 1 because we will surely fall on one point of the curve.

You can easily observe that theses 3 conditions are also 3 main characteristic of the normal distribution we know today. This is why we decided to focus on Laplace's researches.

Earlier mathematicians agreed on these conditions but then faced the problem of choosing a function, which was usually done arbitrarily. As Laplace wrote in his 1774 memoir, *'But of an infinitely number of possible functions, which one is to be preferred?'*. For example, Simpson promoted uniform and triangular distribution upon mathematical convenience. Laplace aimed for an exact computation. [18]

Laplace elaborated a hypothesis he could accept. He rejected the possibility of constant decrease of $\phi(x)$ as x deviates from the true value. So he asked himself that given the decrease of $\phi(x)$ and $\frac{d\phi(x)}{dx}$, what could be the relation between the two rates of change? He states :

> *'Now, as we have no reason to suppose a different law for the ordinates than for their differences, it follows that we must, subject to the rules of probabilities, suppose the ratio of two infinitely small consecutive differences to be equal to that of the corresponding ordinates.*

We write in what follows the proof of Laplace's first error curve, i.e. how he derived its formula.[1]

Proof :

From his intuition on the relation between the rate of change of $\phi(x)$ and its differential $\frac{d\phi(x)}{dx}$ stated above. Laplace writes for positive x,

$$\frac{d\phi(x+dx)}{d\phi(x)} = \frac{\phi(x+dx)}{\phi(x)}$$

Therefore,

$$\frac{d\phi(x)}{dx} = -m\phi(x)$$

Which gives,

$$\phi(x) = \mathcal{C}e^{-mx}$$

Since the area of the whole curve must be 1 and the curve is symmetric, we have,

$$\int_0^\infty \mathcal{C}e^{-mx}dx = 1/2$$

$$\left[-\frac{1}{m}\mathcal{C}e^{-mx}\right]_0^\infty = \frac{1}{m}\mathcal{C} = \frac{1}{2}$$

We get $\mathcal{C} = \frac{m}{2}$ and therefore for $x \in \mathbb{R}$,

$$\phi(x) = \frac{m}{2}e^{-m|x|}$$

[1]quote and formula from [18]

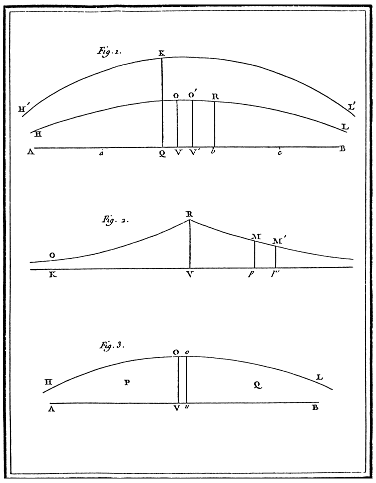

Figure 5: Diagrams accompanying Laplace's 1774 memoir on inverse probability, figure 2 represents the exponential density while Fig 1 and Fig 3 represents previous distributions (figure from [18])

Laplace was not satisfied with his first error curve. In 1777, he came out with another argument. In modern explanation, the idea could be explained as follows:
Imagine the unit interval is spitted into $n+1$ parts by choosing n uniformly random points in this interval. Let $\{d_i | 1 \leq i \leq n\}$ be the resulting spacings ordered in decreasing order, we have,

$$d_1 > d_2 > \cdots > d_{n+1} \qquad \sum_{i=1}^{n+1} d_i = 1$$

If we plot the points $(\frac{i}{n}, \mathbb{E}[d_i])$ and taking $n \to \infty$, Laplace found a curve proportional to $\log(\frac{a}{x})$ on $(0, a]$ By symmetry and taking the area under the curve equal to 1, he got

$$y = \frac{1}{2a} \log\left(\frac{a}{|x|}\right)$$

We don't give the complete proof of the formula in this paper since Laplace's work was immense. Printing his argument in 1979 consumed the better part of 20 pages [18].

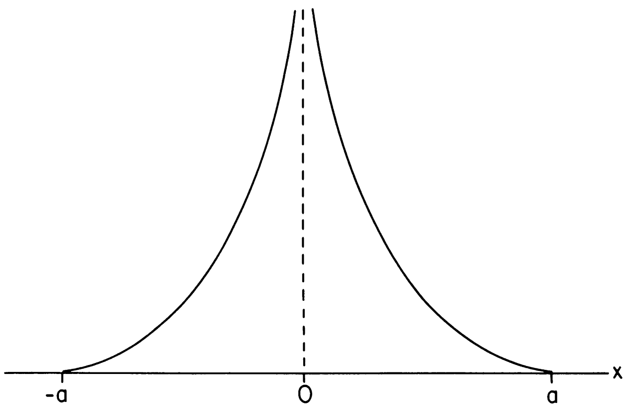

Figure 6: Laplace's 1777 error curve (figure from [18])

3.4 Gauss's proof

Laplace hoped to show that the arithmetic mean is the best estimate of the true value in observation's distribution. He used the absolute value of the deviation from the true value as error estimate $|x_i - \theta|$. He also used error minimization as estimation method and therefore gets the median of observations as minimizer.

Johann Carl Friedrich Gauss is a German mathematician, astronomer and physicist of the 18th-19th centuries. Sometimes referred to as the *Princeps mathematicorum* (Prince of Mathematics), he is well known for his work on the normal distribution, the maximum likelihood estimation and especially the method of least squares and more. In 1809, Gauss solves the problem of arithmetic mean by taking another distribution, his equation's unknown, and a new estimation method which he introduces, the maximum likelihood estimation. He writes:

> *'It has been customary to regard as an axiom the hypothesis that if any quantity has been determined by several direct observations, made under the same circumstances and with equal care, the arithmetic mean of the observed values gives the most probable value, if not rigorously, yet very nearly, so that it is always most safe to hold on to it.' []*

We give here Gauss's proof, the derivation of the normal distribution as it concludes one crucial step of the history of the normal distribution after 200 years of research.

Proof : Let $\phi(x-\theta)$ be the probability distribution of errors and assume that ϕ is differentiable. $x_1, ..., x_n$ are the independent observations.

From Laplace's principle of inverse probability, we have:

$$p(\theta|x_1, ..., x_n) = \frac{\phi(x_1-\theta)\dots\phi(x_n-\theta)}{\int_\theta \phi(x_1-\theta)\dots\phi(x_n-\theta)\,d\theta}$$

Gauss requires that the most probable value should be set to the arithmetic mean.
In other words,

$$\frac{\partial \ln(p(\theta|x_1, ..., x_n)}{\partial x} = 0 \quad \text{for } \theta = \overline{x} = \frac{x_1 + \cdots + x_n}{n}$$

We therefore need to find ϕ such that,

$$\frac{\phi'(x_1-\overline{x})}{\phi(x_1-\overline{x})} + \cdots + \frac{\phi'(x_n-\overline{x})}{\phi(x_n-\overline{x})} = 0$$

Since $x_1, \dots, x_n$ are arbitrary values, let's suppose in particular that $x_2 = x_3 = ... = x_1 - nN$ for some arbitrary real value N. We get $\overline{x} = x_1 - (n-1)N$ and,

$$\frac{\phi'((n-1)N)}{\phi((n-1)N)} + (n-1)\frac{\phi'(-N)}{\phi(-N)} = 0$$

By symmetry of the distribution, we have $\phi(-x) = \phi(x)$ and $\phi'(-x) = -\phi'(x)$ so $\frac{\phi'(-x)}{\phi(-x)} = -\frac{\phi'(x)}{\phi(x)}$, we rewrite the above result as,

$$\frac{\phi'((n-1)N)}{\phi((n-1)N)} = (n-1)\frac{\phi'(N)}{\phi(N)}$$

If we take $f(x) = \frac{\phi'(x)}{\phi(x)}$, we have with continuity of f and the above homogeneity condition that $f(x) = kx$ and therefore we get the differential equation,

$$\frac{\phi'(x)}{\phi(x)} = kx$$

Which is solved by integration,

$$\phi(x) = \mathcal{X}e^{\frac{-kx^2}{2}}$$

To have a maximum at $x = 0$, k must be negative, so we set $\frac{1}{2}k = -h^2$.
From Laplace integral result (1782) we have that,

$$\int_{-\infty}^{\infty} e^{-h^2x^2} = \frac{\sqrt{\pi}}{h}$$

With the unit area constraint, we finally get

$$\phi(x) = \frac{h}{\sqrt{\pi}}e^{-h^2x^2}$$

Gauss published his solution in *Theoria motus corporum coelestium in sectionibus conicis solem ambientium* (1809). [19].

At first, the normal distribution was given many names such as Gaussian law, law of errors, etc. The name *normal distribution* was popularized by Pearson in the 20th century who wanted to give credits to Laplace. This name comes from the normal equations involved in its application in the method of least squares. Later in the 20th century, Pearson introduced the standard deviation term σ in the distribution. In 1915 Fisher gave the normal distribution its final form that we know today by adding the location parameter μ. [16]

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

4 Central Limit Theorem

The Central Limit Theorem was officially proved by Laplace in 1812. It is certainly the most powerful property of the normal distribution. Its key idea is that, under some conditions (which differs depending on the version), the sum of independent and identically distributed random variables tends towards a normal distribution.

4.1 Genesis

The first version of the CLT was stated in 1733 by de Moivre in his approximation of the binomial distribution (cf. Section 2.4). He found that $Bin(n, p)$ approaches a normal distribution when n grows (cf. Definition 3). In other words, he found that the sum of n bernouilli trials is normally distributed with mean np and npq. We can rewrite his result as follows.

Let $X_i \sim Ber(p) \in \{0, 1\}$ for $i = 1, \ldots, n$, let $S = \sum_{i=1}^{N} X_i \sim Bin(n, p)$ then when n grows large,

$$P(S = s) = \binom{n}{s} p^s q^{n-s} \sim \frac{1}{\sqrt{2\pi npq}} e^{-\frac{(s-np)^2}{2npq}}$$

4.2 The Central Limit Theorem and Laplace's proof

In 1781, Laplace already derived the distribution of the arithmetic mean for i.i.d. random variables having a continuous density. But in practice, his focus was on densities for which calculations where too cumbersome and complicated which compromised his success. He got back to his works in 1810, and in 1812, Laplace pubishes in *Théorie analytique des probabilités* the first official version of the CLT. [9]

Let $X_1, \ldots, X_n$ be a sequence of i.i.d. random variables such that $\mathbb{E}[X_i] = \mu$ and $\text{Var}(X_i) = \sigma^2$. The support of X_i's distribution is finite and non-degenerate. We then have $0 < m < \sigma^2 < M < \infty$. Let $S = X_1 + \cdots + X_n$, the, S converges asymptotically to the normal distribution $\mathcal{N}(n\mu, n\sigma^2)$. Therefore,

$$z = \frac{S - n\mu}{\sigma\sqrt{n}} = \frac{(\overline{X} - \mu)\sqrt{n}}{\sigma} \sim \mathcal{N}(0, 1)$$

We give below the core idea of Laplace's proof for the Central Limit Theorem as it concludes the most important property of the normal distribution.

<u>*Proof :*</u> Laplace defines the characteristic function,

$$\psi(t) = \mathbb{E}[e^{iXt}] = \mathbb{E}\left[\sum_{r=0}^{\infty} \frac{(it)^r X^r}{r!}\right] = 1 + it\mathbb{E}[x] - \frac{t^2\mathbb{E}[x^2]}{2!} + ...$$

He takes the logarithm and gets what was later called the cumulant-generating function,

$$\ln \psi(t) = it\mathbb{E}[X] - \frac{\sigma^2 t^2}{2!} + \ldots \quad \sigma^2 = \mathbb{E}[X^2] - \mathbb{E}[X]^2$$

Since the characteristic function of S is $\psi_s(t) = \mathbb{E}[e^{i(X_1+\cdots+X_n)t}] = \mathbb{E}[e^{iX_1t}]\ldots\mathbb{E}[e^{iX_nt}] = \psi^n(t)$, we have $\ln\psi_s(t) = n\ln\psi(t)$.
With $\psi(t) = \sum_k P(X=k)e^{ikt}$ (finite support), and using,

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} e^{ikt}dt = \begin{cases} 0 & k=1 \\ 1 & k=0 \end{cases}$$

He gets the inversion formula,

$$P(X=k) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \psi(t)e^{-ikt}\,dt$$

$$P(S=s) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \psi^n(t)e^{-ist}\,dt = \frac{1}{2\pi}\int_{-\pi}^{\pi}\left[\sum_k P(X=k)e^{ikt}\right]^n e^{-ist}\,dt$$

Which with more development leads him to the final result,

$$P(S=n\mathbb{E}[X]+s) = \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{(-ist-\frac{1}{2}n\sigma^2t^2+\ldots)}\,dt$$

Neglecting higher order terms he gets,

$$P(S=n\mathbb{E}[X]+s) \sim \frac{1}{\sqrt{2\pi n}\sigma}e^{-\frac{s^2}{2n\sigma^2}}$$

More details on the proof can be found in [2] Laplace later noticed that the CLT holds for infinite support as long as σ^2 is bounded away from infinity and zero. His intuitions were later proved right.[9]

The Central Limit Theorem confirms the ubiquity of the normal distribution in nature. As Sir Francis Galton said [3]:

> *'I know of scarcely anything so apt to impress the imagination as the wonderful form of cosmic order expressed by the 'Law of Frequency of Error'. The law would have been personified by the Greeks and deified, if they had known of it. It reigns with serenity and in complete self-effacement, amidst the wildest confusion. The huger the mob, and the greater the apparent anarchy, the more perfect is its sway. It is the supreme law of Unreason. Whenever a large sample of chaotic elements are taken in hand and marshalled in the order of their magnitude, an unsuspected and most beautiful form of regularity proves to have been latent all along.'*

This very powerful theorem was later studied and refined by many mathematicians throughout the centuries notably by Poisson (1837), Cauchy (1853) and the russian mathematician Lyapunov (1901).[2]

5 Conclusion

As we have seen in this paper, the normal distribution bears a great history. It started with the development of probability theory in the 17th century with Pascal and Fermat solving gambling problems, notably the problem of points raised by the Chevalier de Méré. Jakob Bernouilli made gambling games easier to analyze by creating the binomial distribution in 1713. However, computing this distribution for large parameters was very difficult. In 1733, de Moivre found an approximation for the binomial distribution. This was the first time the normal distribution was ever derived.

On the other hand, astronomers like Brahe and Galileo were confused about how to treat measurements errors and the search for an error distribution started. Simpson created one of these curves in 1755 but didn't convince Laplace who sought for an exact computation. He found a proof for the distribution of errors but didn't manage to get the arithmetic mean as best estimator for measurements' location. Instead, the error minimizer was the median. In 1809, Gauss gave a proof for the distribution of errors having the arithmetic mean as best estimator by introducing a new distribution (the normal distribution) along with a new estimation method, the maximum likelihood estimation. Later in 1812, Laplace derived the Central Limit Theorem which showed the importance and omnipresence of the normal distribution in nature. It approximate the distribution of the sum of any i.i.d. random variables with fixed mean and variance to a normal distribution.

The bell curve has been widely used throughout the 19th and 20th century in social and natural sciences. For example, in 1846, Quetelet gave the first application of the normal distribution in social sciences. He showed that the chest measurements of Scottish soldiers were normally distributed and created the concept of 'the Averaged Man'. It is a Man who is characterized by the mean of all Man's measurable characteristics. It was elevated to the standard of perfection and beauty.[20].

The normal distribution also found its application more recently in finance for example. After the crisis of 1987, statisticians and risk managers created the concept of Value at Risk (VaR). It models the losses and profits of a portfolio with a normal distribution. The analysis of quantiles allows the trader to know how much money is at risk with a given probability [10].

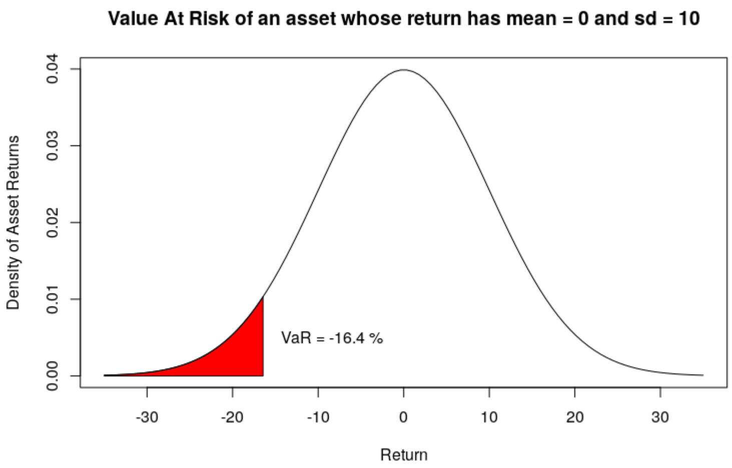

Figure 7: 5% VaR of an asset, we risk of losing 16.4% with probability 0.05 (generated using R)

The normal distribution stays nowadays the best distribution to model errors or deviations from a natural average.

References

[1] Hans Fischer. *History of the Central Limit Theorem: from Classical to Modern Probability Theory*, chapter 2, pages 17–75. Springer, New-York, New York, 2011. Chapter: Central Limit Theorem from Laplace to Cauchy: Changes in Stochastic Objectives and in Analytical Methods.

[2] Hans Fischer. *History of the Central Limit Theorem: from Classical to Modern Probability Theory.* Springer, New-York, New York, 2011.

[3] Sir Francis Galton. *Natural Inheritance*, chapter 5, page 66. Francis Galton, 1889. Chapter: Normal Variability.

[4] Anders Hald. *A History of Probability and Statistics and Their Applications before 1750*, chapter 5, pages 42–65. Wiley-Interscience, Hoboken, New Jersey, 2003. Chapter: The Foundation of Probability Theory by Pascal and Fermat in 1654.

[5] Anders Hald. *A History of Probability and Statistics and Their Applications before 1750*, chapter 20, pages 347–375. Wiley-Interscience, Hoboken, New Jersey, 2003. Chapter: The problem of the duration of play, 1708-1718.

[6] Anders Hald. *A History of Probability and Statistics and Their Applications before 1750*, chapter 24, pages 468–508. Wiley-Interscience, Hoboken, New Jersey, 2003. Chapter: De Moivre's Normal Approximation to the Binomial Distribution 1733.

[7] Anders Hald. *A History of Probability and Statistics and Their Applications before 1750*, chapter 10, pages 144–170. Wiley-Interscience, Hoboken, New Jersey, 2003. Chapter: Mathematical Models and Statistical Methods in Astronomy from Hipparchus to Kepler and Galileo.

[8] Anders Hald. *A History of Parametric Statistical Inference from Bernoulli to Fisher, 1713-1935*, chapter 3, pages 17–25. Springer, New-York, New York, 2007. Chapter: De Moivre's Normal Approximation to the Binomial,1733, and its Generalization.

[9] Anders Hald. *A History of Parametric Statistical Inference from Bernoulli to Fisher, 1713-1935*, chapter 12, pages 83–93. Springer, New-York, New York, 2007. Chapter: Laplace Central Limit Theorem and Linear Minimum Variance Estimation.

[10] Philippe Jorion. *Value at Risk: the New Benchmark for Managing Financial Risk.* McGraw - Hill Education, New-York New York, 3 edition, 2006.

[11] Pascal. Letter to fermat. *Games, Gods, and Gambling*, 1962. The letter is quoted in the book from F. N. David.

[12] Blaise Pascal. Traité du triangle arithmétique. 1965. Archived at the Wayback Machine at the Cambridge University Library (in French) with short English summary.

[13] Jagdish Patel and Campbell Read. *Handbook of the normal distribution*, chapter 2, pages 18–19. Marcel Dekker, inc., New-York, New York, 1982.

[14] Neil Salkind. *Encyclopedia of Research Design*, volume 1, pages 932–934. SAGE Publications, Thousand Oaks, California, 2010. Chapter: Normality Assumption.

[15] Neil Salkind. *Encyclopedia of Research Design*, volume 1, pages 78–81. SAGE Publications, Thousand Oaks, California, 2010. Chapter: Bernoulli Distribution.

[16] Bruce D. [Editor] Spencer. *Statistics And Public Policy.* Oxford University Press, Oxford, UK, 1997.

[17] Hans Sterk. Probability. *2WH60 - Week 6*, page 1, 2020.

[18] Stephen Stigler. *The History of Statistics : The Measurement of Uncertainty before 1900.*, chapter 3, pages 99–139. President and Fellows of Harvard College, Cambridge, Massachusetts, 1986. Chapter: Inverse Probability.

[19] Stephen Stigler. *The History of Statistics : The Measurement of Uncertainty before 1900.*, chapter 4, pages 139–159. President and Fellows of Harvard College, Cambridge, Massachusetts, 1986. Chapter: The Gauss-Laplace Synthesis.

[20] Stephen Stigler. *The History of Statistics : The Measurement of Uncertainty before 1900.*, chapter 5, pages 161–221. President and Fellows of Harvard College, Cambridge, Massachusetts, 1986. Chapter: Quetelet's two attempts.

[21] Wikimedia Foundation Wikipedia. Problem of points, last update 2020-05-13.