

Chapter 2

The Central Limit Theorem from Laplace to Cauchy: Changes in Stochastic Objectives and in Analytical Methods

In 1812, Pierre-Simon de Laplace (1749–1827) published the first edition of his *Théorie analytique des probabilités* (henceforth simply abbreviated by TAP).¹ With its typical problems, stochastic models, and analytic methods this book would considerably influence probability theory and mathematical statistics right until the beginning of the 20th century.

Until Laplace and his successors, classical probability consisted mainly in the sum of its applications to physical, social, and moral problems. However, as Laplace already pointed out in the concise preface to the first edition of his TAP, probability was also important for mathematics in a narrower sense. In many problems referring to stochastic models depending on a large number of trials, probabilities could only be expressed by formulae too complicated for direct numerical evaluation. Thus, for a reasonable application of many of the results of probability calculus, particular considerations were needed to obtain useful approximations of the “formulae of large numbers.” In the aforementioned preface, Laplace called these problems “the most delicate, the most difficult, and the most useful” of the entire theory. He expressed his hope that discussion of these problems would catch the attention of other “geometers.” Therefore, in addition to the qualitative feature of applicability, which was characteristic for classical probability theory, a new, purely mathematical aspect emerged: the relevance of specific analytical methods of probability theory.

Laplace had been intensely dealing with the “delicate problems” of probability just described from the very beginning of his scientific career. In his 1781 “Mémoire sur les probabilités,” one can already find “in nuce” almost all of the problems of TAP, which can be roughly divided into two categories: “sums of random variables”

¹ For a description of the origin and the major contents of this book, see [Stigler 2005; Sheynin 2005b, 99–110]. An English translation by Richard Pulskamp of the second, probabilistic, part of the TAP is available at

<http://www.cs.xu.edu/math/Sources/Laplace/index.html>.

and “inverse probabilities.”² The first category includes, for example, the a priori probabilities of profit and loss in certain games of chance, or of the arithmetic mean of observations being subject to random errors; the latter for instance deals with the a posteriori probabilities that the ratio of the chances of a boy’s and a girl’s birth is within a given interval centered around the ratio of the corresponding observed values. By 1774, Laplace had already developed useful approximation methods for those a posteriori probabilities depending on a large number of observations. He did not succeed in adapting this method to a priori probabilities until 1810, however. Only then, with a “tricky” modification of the method of generating functions, did he achieve any usable results on approximations of probabilities of sums of independent random variables, which, from the modern point of view, are subsumed under the rubric of the “central limit theorem.” It was the CLT which considerably shaped the contents and methods of the *TAP* and significantly influenced the development of probability and error theory during the 19th century.

As we have already seen (Sect. 1.4), the history of the CLT, as far as the contributions of Laplace and his successors are concerned, has already been studied in fair detail. Therefore, a main focus in the present section will be on those questions which still seem to be open: Which changes in the probabilistic and analytical context of the CLT occurred between ca. 1810 and 1850; how did these changes come about, and how have these changes influenced analytical style and methods in the treatment of this theorem?

2.1 Laplace’s Central “Limit” Theorem

As already noticed, Laplace’s CLT was the result of an almost forty years’ effort. In the following, we will describe the historical development of Laplace’s treatment of sums of independent random variables, his methods for finding appropriate approximation formulae, and the major applications of his finally achieved CLT.

² Inverse probabilities are conditional probabilities $P(H|B)$ for certain “hypothetic” causes H which may have entailed the observed results B . ($P(H|B)$ is considered as “inverse” to $P(B|H)$.) The probabilities $P(H|B)$ can be interpreted as if they quantify conclusions from an observation B to its causation H “a posteriori.” If there are n possible causes H_j ($j = 1, \dots, n$), and if the $P(H_j)$ are known, then, by virtue of Bayes’s formula:

$$P(H_k|B) = \frac{P(B|H_k)P(H_k)}{\sum_{j=1}^n P(B|H_j)P(H_j)}, \quad k = 1, \dots, n.$$

Since the probabilities $P(H_j)$ are unknown in most cases, one is often forced to the “subjective” assumption of the H_j being equiprobable. If, conversely, a certain probability distribution is—more or less arbitrarily—presupposed, then any probabilities derived therefrom can be interpreted as “a priori probabilities.”

2.1.1 Sums of Independent Random Variables

Sums of independent random variables had played an important role in Laplace's probabilistic work from the very beginning.³ In this context, the problem of calculating the probability distribution of the sum of angles of inclination, which were assumed to be determined randomly, as well as the related problem of calculating the probabilities of the deviations between the arithmetic mean of data which were afflicted by observational errors and the underlying "true value," became especially important. In one of his first published papers, Laplace [1776] had already set out to determine the probability that the sum of the angles of inclination of comet orbits (or the arithmetic mean of these angles respectively) is within given limits. He assumed that all angles, which had to be measured against the ecliptic, were distributed randomly according to a uniform distribution between 0° and 90° (and also tacitly presupposed that all angles were stochastically independent). Laplace succeeded in calculating these probabilities for an arbitrary number of comets via induction (with a minor mistake which was subsequently corrected in [Laplace 1781]). In this 1781 paper, Laplace even introduced a general—however very intricate—method, based on convolutions of density functions, in order to exactly determine the probability that a sum of independent random variables ("quantités variables," as Laplace put it) was within given limits.⁴ In the most simple case, each of the n variables had the same rectangular distribution between 0 and h . For the probability P that the sum of those variables was between a and b with $0 \leq a < b \leq nh$, Laplace obtained (in modern notation)

$$P = \frac{1}{h^n n!} \left(\sum_{i=0}^N \binom{n}{i} (-1)^i (b - ih)^n - \sum_{i=0}^M \binom{n}{i} (-1)^i (a - ih)^n \right), \quad (2.1)$$

where $N = \min(n, \lfloor \frac{b}{h} \rfloor)$ and $M = \min(n, \lfloor \frac{a}{h} \rfloor)$. Formulae of this kind were too complicated for a direct numerical evaluation if the number of random variables exceeded a relatively small value. The arithmetic mean of the actual observed angles of inclination of the then known 63 comets was $46^\circ 16'$. Through the use of (2.1) alone, Laplace was unable to address the hypothesis that the comets' planes of motion resulted at "random." At this stage of his mathematical work, however, Laplace could not develop usable approximations.

³ For a comprehensive biography also dealing with Laplace's probabilistic work, see [Gillispie 1997]. Detailed discussions of Laplace's contributions to probability and statistics can be found in [Sheynin 1976; 1977; 2005b; Stigler 1986; Hald 1998]. The web site already referred to in footnote 1 contains English translations of most works in probability theory by Laplace.

⁴ See [Sheynin 1973, 219 f.] and [Hald 1998, 56–60] for descriptions of this method.

2.1.2 Laplace's Method of Approximating Integrals, and "Algebraic Analysis"

Beginning with his "Mémoire sur la probabilité des causes" [1774], Laplace developed techniques for approximating integrals depending on a "great number," such as, for example, the Gamma function $\Gamma(s+1) = \int_0^\infty e^{-x} x^s dx$ with the "great number" s . The basic idea of this "Laplacian method of approximation" is as follows: Let the integrand $f(x)$ depend on a very large parameter such that the function f has a single, very sharp peak, with the consequence that appreciable contributions to the entire integral result only from a small interval around this maximum. Then it can be expected that the function f is asymptotically equal to a function of the form $f(a)e^{-\alpha(x-a)^{2k} \pm \dots}$ ($\alpha > 0$) if f attains its maximum at $x = a$. Based on this idea, the Laplacian method consists of appropriate series expansions around the abscissa of the maximum. In the case of the Gamma function, Laplace started with

$$\Gamma(s+1) = \int_0^\infty e^{-x} x^s dx = \int_{-s}^\infty e^{(-z+s)} (z+s)^s dz.$$

The maximum $M = e^{-s}s^s$ of the integrand is attained at $x = s$, or equivalently $z = 0$. Laplace [1785, 258 f.; 1812/20/86, 128–131] set

$$e^{-s} e^{-z} (z+s)^s = M e^{-t^2(z)}$$

and expanded $t^2 = -\log(e^{-z}(1+z/s)^s)$ into a series of powers of z . Conversely, he also expanded z into a series of powers of t , and obtained the following expansion after transforming the variable of integration from z to t :

$$\begin{aligned} \Gamma(s+1) &= M \int_{-\infty}^\infty e^{-t^2} \sqrt{2s} \left(1 + \frac{4t}{3\sqrt{2s}} + \frac{t^2}{6s} + \dots \right) dt \\ &= s^{s+1/2} e^{-s} \sqrt{2\pi} \left(1 + \frac{1}{12s} + \frac{1}{288s^2} + \dots \right). \end{aligned} \quad (2.2)$$

For many probabilistic formulae, Laplace's method of approximation worked extremely well. For the problem of sums of (independent) random variables, however, it was only at a rather late stage of his mathematical work that Laplace developed techniques based on which suitable approximations could be deduced.

In the above-mentioned article of 1774, Laplace treated approximation problems in an analytical style closely related to that of Euler. Laplace discussed the behavior of the peak with an "infinitely large" parameter, carefully considering "infinitely" large or small quantities. In his later work, however, he abandoned the "Eulerian" style of calculating with infinite quantities of different gradations and, influenced by Lagrange's algebraic analysis, developed a special algebraic-algorithmic style dealing primarily with formal series expansions, as we have just seen in connection with the Gamma function. Laplace's deduction of the CLT was likewise written in this style.

2.1.3 The Emergence of Characteristic Functions and the Deduction of Approximating Normal Distributions

Laplace for the first time exemplified his approach to the CLT in the "Mémoire sur les approximations des formules qui sont fonctions des très grands nombres et sur leur application aux probabilités" [1810a]. Crucial for this success in approximating distributions of sums of independent random variables by normal distributions was his modification of generating functions. Let me demonstrate the essentials of his approach to the CLT⁵ in the special case of identically distributed random variables X_1, \dots, X_n , which have zero means and which take the values $\frac{k}{m}$ (m a given natural number, $k = -m, -m+1, \dots, m-1, m$) with the respective probabilities p_k .⁶ For the calculation of the probability P_j that $\sum_{l=1}^n X_l$ has the value $\frac{j}{m}$ ($-nm \leq j \leq nm$), Laplace made use of the generating function $T(t) = \sum_{k=-m}^m p_k t^k$. Due to the mutual independence of the X_l 's—which was usually only tacitly presupposed by Laplace— P_j is equal to the coefficient of t^j in $[T(t)]^n$ after carrying out the multiplication. The direct execution of this method—its general principle going back to de Moivre, see [Seal 1949]—leads at best to very complicated algebraic terms for P_j . Laplace, however, introduced the trick of substituting the variable t by e^{ix} ($i = \sqrt{-1}$). Thus, he introduced the (now so-called) characteristic functions in a special case.

From

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} e^{-itx} e^{isx} dx = \delta_{ts} \quad (t, s \in \mathbb{Z}) \quad (2.3)$$

it follows that

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

The last integral above was at least formally accessible to Laplace's method of approximation. There was, however, a certain modification necessary, as Laplace did not consider an expansion of the whole integrand around its maximum at $x = 0$, but only of the power with exponent n (equal to the characteristic function). By expanding e^{ikx} into powers of x one gets

⁵ The most important sources for Laplace's treatment of the CLT are [Laplace 1810a; 1811], and the fourth chapter of the *TAP*.

⁶ The following explanation differs, as far as terminology and further details are concerned, from Laplace's exposition. Unlike Laplace, we only consider, for the sake of simplicity, random variables with values within the interval $[-1; 1]$. For paraphrases in Laplace's original style see [Sheynin 1977, 10–16] and [Fischer 2000, 29–33]. Hald [1998, 303–317] gives a thorough account on Laplace's analytical approach to the CLT.

$$\begin{aligned}
P(j) &= \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{-ijx} \left[\sum_{k=-m}^m p_k e^{ikx} \right]^n dx \\
&= \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{-ijx} \left[\sum_{k=-m}^m p_k \left(1 + ikx - \frac{k^2 x^2}{2} - \frac{ik^3 x^3}{6} + \dots \right) \right]^n dx.
\end{aligned}$$

Taking into consideration that $\sum_{k=-m}^m p_k k = 0$, and with the substitution $m^2 \sigma^2 = \sum_{k=-m}^m p_k k^2$, we obtain

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

where A is a constant depending on $\sum_{k=-m}^m p_k k^3$. The formal expansion of

$$\log \left[1 - \frac{m^2 \sigma^2 x^2}{2} - iAx^3 + \dots \right]^n =: \log z$$

into a series of powers of x leads to

$$\log z = -\frac{m^2 \sigma^2 n x^2}{2} - iAnx^3 + \dots,$$

and therefrom to

$$z = e^{-\frac{m^2 \sigma^2 n x^2}{2} - iAnx^3 + \dots} = e^{-\frac{m^2 \sigma^2 n x^2}{2}} (1 - iAnx^3 + \dots).$$

After transforming the variable of integration according to $x = \frac{y}{\sqrt{n}}$, the result is

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

For an approximation with a “very large” n we ignore, like Laplace, all series terms with a power of \sqrt{n} in the denominator, and at the same time, set the limits of integration equal to $\pm\infty$. In this way we get

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

where the last integral is, as Laplace showed in different ways, equal to

$$\frac{1}{m\sigma\sqrt{2\pi n}} e^{-\frac{j^2}{2m^2\sigma^2 n}}. \quad (2.4)$$

Summing up (2.4) for $\frac{j}{m} \in [r_1\sqrt{n}; r_2\sqrt{n}]$, which can be approximated by integration ($dx \approx \frac{1}{\sqrt{n}}$), leads to the result

$$\begin{aligned}
P(r_1\sqrt{n} \leq \sum X_l \leq r_2\sqrt{n}) &\approx \sum_{j \in [mr_1\sqrt{n}; mr_2\sqrt{n}]} \frac{1}{m\sigma\sqrt{2\pi n}} e^{-\frac{j^2}{2m^2\sigma^2 n}} \\
&\approx \int_{mr_1}^{mr_2} \frac{1}{m\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2m^2\sigma^2}} dx = \int_{r_1}^{r_2} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}} dx,
\end{aligned}$$

which corresponds to the integral form of the CLT. With only one exception (see Sect. 2.1.5.3) Laplace dealt with independent identically distributed and bounded random variables with densities.⁷ To this aim he at first considered the range of values of those random variables discrete (as described above), and then he assumed m "infinitely large."

Nowhere in his work did Laplace state a general theorem which would have corresponded to the CLT in today's sense. He only treated particular problems concerning the approximation of probabilities of sums or linear combinations of a great number of random variables (in many cases errors of observation, see Sect. 2.1.5.2) by methods which in principle corresponded to the procedure described above. In modern notation, Laplace's most general version of the CLT [Laplace 1812/20/86, 335–338] was as follows: Let $\epsilon_1, \dots, \epsilon_n$ be a large number of independent errors of observation, each having the same density with mean μ and variance σ^2 . If $\lambda_1, \dots, \lambda_n$ are constant multipliers and $a > 0$, then

$$P\left(\left|\sum_{j=1}^n \lambda_j(\epsilon_j - \mu)\right| \leq a \sqrt{\sum_{j=1}^n \lambda_j^2}\right) \approx \frac{2}{\sigma\sqrt{2\pi}} \int_0^a e^{-\frac{x^2}{2\sigma^2}} dx. \quad (2.5)$$

The special case of a CLT for the binomial distribution Laplace [1812/20/86, 280–284] on the basis of Stirling's formula treated in a particular section of his *TAP* by methods which are in principle due to de Moivre and still employed in modern textbooks.

2.1.4 The "Rigor" of Laplace's Analysis

From Laplace's point of view, approximating an analytical expression depending on a great number n meant transforming it into a series expansion with terms whose order of magnitude decreased sufficiently fast with increasing n . The greater the number of calculated terms and the faster these terms decrease, the better the approximation. Laplace did not determine absolute or relative errors of approximations, but instead put his trust, according to the leitmotif of algebraic analysis, in the power of series expansions.

In the case of Laplace's CLT, the series terms seem to decrease with ascending powers of $\frac{1}{\sqrt{n}}$ (or even of $\frac{1}{n}$ if the individual random variables have a symmetric distribution). Apparently, it was Laplace's point of view to trust in the quality

⁷ Laplace [1810a, 326 f.; 1812/20/86, 313 f.] hinted, though in a quite vague manner only, also at the possibility of analogous considerations concerning unbounded random variables.

of his approximations already because of those decreasing series terms. In the *Essai philosophique sur les probabilités*, whose first edition appeared in 1814 and served as a “popular” introduction to the *Théorie analytique*, Laplace [1814/20/86, XXXIX] wrote of his approximations:

(...) the series converges the faster the more complicated the formula is, such that the procedure is more precise the more it becomes necessary.

However, some authors did, if rather rarely, object to Laplace’s specific approach to approximations. A first hint came from Adrien Marie Legendre as early as 1811. In his *Exercices du calcul intégral* [1811, 290 f.] he discussed the approximation formula

$$n! \approx \frac{\sqrt{2\pi}n^{n+1/2}}{\exp(n)} \exp(E(n)), \quad E(n) = \sum_{k=1}^m \frac{B_{2k}}{2k(2k-1)n^{2k-1}}, \quad (2.6)$$

which can (with slight modifications) be traced back to de Moivre and Stirling around 1730 (see [Schneider 1968, 266–276]). The B_{2k} are the (Jakob) Bernoullian numbers; Leonhard Euler had already shown in 1739 that, from a certain index, these numbers grow faster than any geometric sequence [Schneider 1968, 276]. But only Legendre clearly addressed the divergence of the series $E(s)$ and the resulting difficulties for its analytical treatment. Laplace’s series (2.2) was, as apparent from its first terms, equivalent to (2.6). (An exact proof for the equality of both series expansions, however, was not given during the 19th century.) From Legendre’s description [1811, 343–348] of Laplace’s account it became therefore plausible that the Laplacian method of approximation could lead in the general case to (in Legendre’s own words) “semi-convergent expansions” only. Thus, for critical mathematicians, Laplace’s treatment of the CLT became suspicious as well. How could it be justified neglecting series terms of “higher order,” if the series was possibly divergent?

In 1844, Robert Leslie Ellis tried to discuss Laplace’s reasoning regarding the CLT in a modified form (see [Hald 1998, 333–335]). He also explicitly analyzed the example of mutually independent random variables with the common density function $f(x) = \frac{1}{2}e^{-|x|}$. Referring to his—only quite formal—manipulations with series expansions in treating this particular case, he wrote at the end of his explanations [1844, 215]:

But some doubt may perhaps remain, whether such an approximation to the *form* of the function P [the probability to be approximated], if such an expression may be used, is also an approximation to its numerical value (...)

A similar assessment of Laplace’s series expansions was given by Cauchy in [1853g’] (see Sect. 2.5.6).

In 1856 Anton Meyer⁸ submitted a proof of the CLT in the special case of two-valued random variables to the Academy in Brussels. Meyer’s proof was not based

⁸ Meyer was the author of a rather influential treatise of probability and error theory [Meyer 1874], which was also translated into German [Meyer 1874/79] and constitutes an important source for the state of the art at the beginning of the last quarter of the 19th century.

on the usual procedure which can be traced back to de Moivre, and which had also been elaborated in Laplace's *Théorie analytique*. He instead used Laplace's modification of generating functions. There exists a brief report by Jean Baptiste Brasseur on Meyer's article (which itself seems to have been lost). Brasseur [1856] hoped that Meyer's method would lead to a more exact discussion of the neglect of the "terms of higher order of smallness." Meyer's paper was accepted for publication, however on condition that a better examination of the "convergence of the series" be made. The publication failed, Meyer died the following year.

2.1.5 The Central Limit Theorem as a Tool of Good Sense

The examples of Ellis, Cauchy, and Meyer show that, in the middle of the 19th century, Laplace's methods of deducing approximative normal distributions for sums of random variables were considered to be unrigorous by some authors. Such criticism was quite rare, but this was in part due to the status of probability theory within mathematics during the 19th century. As Lorraine Daston [1988] explained, probability theory, at least until the middle of the 19th century, was not a discipline of mathematics in a narrower sense, but rather part of a "mathesis mixta." The value of probabilistic research was determined less by internal mathematical criteria, but rather by the quality of its application to "real" situations. Laplace's CLT met the latter point in an excellent manner. The results of all applications of this theorem matched with "good sense" and thus confirmed Laplace's well-known saying [1814/20/86, CLIII] that

Basically, probability is only good sense reduced to a calculus.

We shall test this claim with three prominent applications of CLT: the comet problem (already mentioned above), the problem of foundation of the method of least squares, and the problem of risk in games of chance.

2.1.5.1 The Comet Problem

In 1810, Laplace could base his examinations of the "randomness" of the orbits of comets on the observation of 97 comets. Under the hypothesis of a uniform distribution for the angles of inclination between 0 G and 100 G (centesimal degrees, corresponding to 0° and 90°) and with aid of the CLT, he calculated the probability that the arithmetic mean of all angles falls within a certain interval around "50 G." The mean of the observed values was 51.87663 G, and thus Laplace considered the interval $[50\text{ G} - 1.87663\text{ G}; 50\text{ G} + 1.87663\text{ G}]$. The probability of this interval was only around 0.5. Therefore, there was a considerable probability that, presupposing a uniform distribution, the mean of all angles deviated from 50 G even more than the observed mean. Laplace [1810a, 316] followed that there did not exist any "primitive cause" which affected the specific positions of comet orbits. Thus, Laplace, by

using probabilistic methods, succeeded in confirming the prior assertion of Achille Pierre de Séjour (stated in *Essai sur les comètes* 1775) which he had already referred to in his first pertinent contribution [Laplace 1776, 280].

In contrast, an analogous calculation regarding the 10 planets (and planetoids) known at that time, which could be carried out with the “exact” formula (2.1) of 1776/1781, showed that the position of their orbits depended on a common “cause” [Laplace 1810a, 307 f.]. Such considerations were important regarding the currently so-called Kant–Laplace nebular-hypothesis. Stigler [1986, 137 f.] and Hald [1998, 303–306], both referring to the first, although very specific and purely algebraic, applications of the tricky substitution $t^x = e^{x\varpi\sqrt{-1}}$ in generating functions discussed by Laplace in [1785, 267–270], maintain that Laplace had already discovered “his” CLT by the 1780s. However, the relevance of this theorem for astronomical issues, intensively studied by Laplace between 1785 and 1810, was likely to have led to the publication of pertinent results as soon as possible. Thus, Laplace presumably did not develop his method for deriving approximate normal distributions for sums of independent random variables much earlier than around 1810.

The problem whether orbits of comets and planets depended on “primitive causes” was only one of several opportunities when Laplace searched for “regular causes” in nature. Other examples, treated similarly as the comets and planets issue, such as the daily changes of air pressure between mornings and evenings, or the slight deviations to the east during the free fall of bodies, can be found in the fifth chapter of Laplace’s *TAP*.⁹

2.1.5.2 The Foundation of the Method of Least Squares

The most prominent application of the method of least squares¹⁰ during the 19th century was as follows:

Let d_i ($i = 1, \dots, s$) be observed values, a_{ij} ($j = 1, \dots, t, t < s$) given coefficients, and ξ_j “elements” to be determined such that

$$d_i + \epsilon_i = \sum_{j=1}^t a_{ij}\xi_j \quad (i = 1, \dots, s), \quad (2.7)$$

where the ϵ_i are unknown, mutually independent errors of observation. Laplace named the equations (2.7) “equations of condition” (“equations de condition”). The problem was to estimate the ξ_j as precisely as possible after observing the d_i . According to the method of least squares, first published by Legendre in 1805, estimators x_j for the ξ_j can be obtained by virtue of the principle

⁹ For a survey of the pertinent work of Laplace see [Hald 1998, 431–443].

¹⁰ There exists a good deal of historical literature on the method of least squares. For detailed discussions of the error theoretic development during the 18th and 19th centuries see [Stigler 1986; Hald 1998; Farebrother 1999]. The most important original sources can be found (mainly in German translation) in [Schneider 1988].

$$\sum_{i=1}^s \left(d_i - \sum_{j=1}^t a_{ij} x_j \right)^2 = \min, \quad (2.8)$$

from which the t equations

$$\sum_{i=1}^s \sum_{j=1}^t a_{ik} a_{ij} x_j = \sum_{i=1}^s a_{ik} d_i \quad (k = 1, \dots, t)$$

follow. Thus, the method of least squares belongs to those methods which combine the equations of condition after setting $\epsilon_i = 0$ linearly to a new system of t equations in t unknowns. In modern matrix-notation, this means: Given the system of equations of condition

$$\mathbf{d} + \boldsymbol{\epsilon} = A\boldsymbol{\xi}$$

for the vector of unknown elements $\boldsymbol{\xi} = (\xi_1, \dots, \xi_t)^T$ with

$$A = (a_{ij}) \in \mathbb{R}^{s,t} \ (s > t), \ \mathbf{d} = (d_1, \dots, d_s)^T, \ \boldsymbol{\epsilon} = (\epsilon_1, \dots, \epsilon_s)^T,$$

the goal is to find a system of "multipliers" $B \in \mathbb{R}^{t,s}$ such that the vector of solutions \mathbf{x} of the equation system

$$B\mathbf{d} = B A \mathbf{x}$$

is in a certain sense "optimal" with regard to the "true" $\boldsymbol{\xi}$. Choosing $B = A^T$, one gets exactly the same values for the coordinates of \mathbf{x} which result from the condition (2.8), that is, from the method of least squares.

In the special case of "direct observations" of one single element ξ , that means, in the case where the equations of condition have the particular form

$$d_i + \epsilon_i = \xi \quad (i = 1, \dots, s),$$

the method of least squares yields the arithmetic mean $x = \sum_{i=1}^s d_i / s$ as an estimator for ξ . This property rather frequently played an important role in foundational discussions on least squares during the 19th century.

Legendre [1805] had only given an intuitive justification of least squares, which did not use any probabilistic arguments. In 1809 Carl Friedrich Gauss succeeded in showing that the least squares estimators x_j according to (2.8) are equal to the estimators meeting the condition of being "most probable," a condition which is now called the "maximum-likelihood-principle" (see Sect. 3.1). For this justification of giving preference to the method of least squares, Gauss presupposed that the errors of observation were identically normally distributed (with expectation 0).

The joint occurrence of normal distributions in Gauss's argument and in Laplace's CLT possibly motivated the latter to give a new foundation of least squares in the case of a large number of equations of condition (see [Stigler 1986, 143] for a discussion of this "Gauss-Laplace Synthesis"). Laplace [1811] showed that the method of least squares was "optimal" according to certain criteria, which suggested to him calling this method later, in the TAP, the "most advantageous."

If one takes the recapitulating description in the *Essai philosophique* (the introduction to the *TAP*) as a standard, the “most advantageous” method was, according to Laplace [1814/20/86, LXII], the method in which “one and the same error of the results is less probable than with any other procedure.” A sensible translation of this sentence into modern mathematical language is: the estimator x' for a true value ξ according to the “most advantageous” method has, in comparison with all estimators x'' obtained by competing methods, the following property:

$$P(|\xi - x'| \geq a) < P(|\xi - x''| \geq a) \text{ for all } a > 0. \quad (2.9)$$

Laplace (e.g., [1812/20/86, 348]) claimed to have proven that the method of least squares would be, in this sense, the “most advantageous,” at least among those methods which combine a large number of observational equations linearly into a set of equations with (if possible) a uniquely determined system of solutions.

In his foundation of the method of least squares, Laplace [1811, 387–398; 1812/20/86, 318–327] treated first the simplest case of equations of condition with a single element ξ :

$$a_1\xi = d_1 + \epsilon_1, \dots, a_s\xi = d_s + \epsilon_s$$

(a_i given coefficients, d_i observations, ϵ_i mutually independent errors with zero means). Laplace estimated ξ in the form

$$x = \frac{\sum_{i=1}^s b_i d_i}{\sum_{i=1}^s b_i a_i},$$

b_1, \dots, b_s being indeterminate constants at first. The difference between the true value ξ and the estimator x became therefore

$$\xi - x = \frac{\sum_{i=1}^s b_i \epsilon_i}{\sum_{i=1}^s b_i a_i}. \quad (2.10)$$

In order to determine the “most advantageous” multipliers b_i , Laplace tried to calculate the probability law for linear forms $\sum_{i=1}^s b_i \epsilon_i$, s being a great number. For each error he assumed the same symmetric density function which vanished beyond a bounded interval. In his work of 1810 Laplace had already deduced an approximating normal distribution for the sum of a large number of identically distributed errors, a result which at first served only for a rather theoretical discussion of arithmetic means. Now, Laplace used an analogous analytical approach to the linear combination, with the following result (represented in modern notation):

$$P(-r\sqrt{s} \leq \sum b_i \epsilon_i \leq r\sqrt{s}) \approx \frac{\sqrt{2s}}{\sigma \sqrt{\pi \sum b_i^2}} \int_0^r e^{-\frac{su^2}{2\sigma^2 \sum b_i^2}} du,$$

σ^2 being the variance common to all errors. Setting $r\sqrt{s} = c\sigma\sqrt{2\sum b_i^2}$, Laplace for $\xi - x$ according to (2.10) deduced:¹¹

$$P\left(\left|\frac{\sum b_i \epsilon_i}{\sum a_i b_i}\right| \leq \frac{c\sigma\sqrt{2\sum b_i^2}}{|\sum a_i b_i|}\right) = P\left(|\xi - x| \leq \frac{c\sigma\sqrt{2\sum b_i^2}}{|\sum a_i b_i|}\right) \\ \approx \frac{2}{\sqrt{\pi}} \int_0^c e^{-t^2} dt. \quad (2.11)$$

Laplace now proceeded, without giving any explanations, as if the approximation (2.11) was, presupposing a large number s , even exact. This was one of the crucial points of his foundation of least squares. As we will see below, Cauchy's criticism of exactly this point would later become a major motivation for his own "rigorous proof" of the CLT. Also Gauss, at several places of his work, critically pointed out that, strictly speaking, Laplace's argumentation was only valid for the unrealistic situation of an "infinitely large" number of observations.¹²

On the basis of the assumption of an exact normal distribution, Laplace required that one choose the multipliers b_i according to the condition that for any probability level (depending only on c) the "limits of error" $\pm \frac{c\sigma\sqrt{2\sum b_i^2}}{|\sum a_i b_i|}$ should be minimal. Because the modulus of these limits is minimal if and only if $b_i = ka_i$, with constant $k \neq 0$, this condition in fact leads to the least squares estimator $x = \frac{\sum a_i d_i}{\sum a_i^2}$. The criterion of "minimal limits" is equivalent to condition (2.9), which was discussed only in Laplace's *Essai philosophique*.

Laplace [1811, 401–409; 1812/20/86, 327–332] also tried to apply his reasoning to the simultaneous treatment of more than one element. To achieve this, he developed a rudimentary form of the multidimensional CLT, from which he, however, passed on to a one-dimensional consideration. A truly complete multidimensional solution of this problem, by an explicit consideration of confidence ellipsoids, was only reached by Bienaymé [1852]. Presupposing mutually independent errors of observation $\epsilon_1, \dots, \epsilon_n$, each having the same density f with mean 0, Bienaymé by further developing Laplace's techniques derived a series expansion for the density $p(\boldsymbol{\tau})$ of the multi-dimensional linear combination $\boldsymbol{\Delta} := \sum_{i=1}^s \boldsymbol{\alpha}_i \epsilon_i$ with fixed $\boldsymbol{\alpha}_i \in \mathbb{R}^t$ ($t \leq n$). His result was equivalent to

$$p(\boldsymbol{\tau}) = \frac{1}{(2\pi)^{\frac{t}{2}} \sigma^t N} \exp\left(-\frac{1}{2\sigma^2} \sum_{j,k=1}^t a_{jk} \tau_j \tau_k\right) (1 - R(\boldsymbol{\tau})),$$

¹¹ In order to deduce the following approximation, Laplace in his *TAP* would have been able to apply equation (2.5), which was even derived for errors with an asymmetrical density, if he had set $\mu = 0$, $a = c\sigma\sqrt{2}$, and $\lambda_i = b_i / \sum a_i b_i$ there. As the *TAP* was largely a compilation of earlier work, he simply copied the argumentation from his 1811 paper, which was based on symmetric errors. And only in the subsequent section of the *TAP* did he establish the relation (2.5), however without any comment on its possible use for discussing least squares.

¹² See, for example, [Gauss 1821, 99; 1823, 18].

where σ^2 denotes the variance common to all errors, (a_{jk}) the inverse matrix of $(A_{i\ell}) \in \mathbb{R}^{t,t}$ with $A_{i\ell} = \sum_{r=1}^s \alpha_{r,i} \alpha_{r,\ell}$,¹³ N^2 the determinant of the matrix $(A_{i\ell})$, and $R(\tau)$ an infinite series of terms, each depending on moments of f ¹⁴ and tending to 0 as $n \rightarrow \infty$ (see [Heyde & Seneta 1977, 66–71; Hald 1998, 501–504]).

By around 1810, several methods of dealing with observational data were available, but the method of least squares was apparently the most useful in the general case. Thus, it was reasonable to champion least squares even without a probabilistic discussion. Yet the CLT “proved” that, at least under “natural” assumptions, this method was superior to other procedures. From Laplace’s point of view, his asymptotic discussion of least squares completely confirmed the established opinion of astronomers and geodesists. Thus, on the one hand, his CLT was a tool of good sense, and its rigor was not to be scrutinized. On the other hand, it became plausible that, in the time after Laplace, critical discussions of the superiority of least squares also questioned the validity of the applied normal approximations, and thus of the CLT itself.

2.1.5.3 Benefits from Games of Chance

As a general rule, Laplace considered independent identically distributed random variables with densities. A rare exception from this rule can be found in his discussion of the “benefits depending on the probability of future events” (chapter IX of *TAP*). Laplace [1812/20/86, 428–432] dealt with a particular sequence of games with only two outcomes for each single game: “gain” and “loss.” He assumed that the respective probabilities of gain and loss were possibly different from game to game. According to these assumptions, Laplace based his analysis on a large number s of single games (tacitly considered as being independent) with results X_1, \dots, X_s , where each X_i could take the values v_i (gain) and $-\mu_i$ (loss) with probabilities q_i and $1 - q_i$, respectively. Proceeding in a way analogous to his treatment of sums of observational errors, he achieved the result that

$$P \left(\left| \sum X_i - \sum (q_i v_i - (1 - q_i) \mu_i) \right| \leq r \sqrt{2 \sum q_i (1 - q_i) (v_i + \mu_i)^2} \right) \\ \approx \frac{2}{\sqrt{\pi}} \int_0^r e^{-t^2} dt.$$

Laplace argued that $\sum (q_i v_i - (1 - q_i) \mu_i)$ was of the order of magnitude s if each summand was “a little” greater than 0, whereas $r \sqrt{2 \sum q_i (1 - q_i) (v_i + \mu_i)^2}$ was of the order \sqrt{s} only. Therefore, for arbitrarily large $r > 0$ and sufficiently large s , even

$$\sum (q_i v_i - (1 - q_i) \mu_i) - r \sqrt{2 \sum q_i (1 - q_i) (v_i + \mu_i)^2}$$

¹³ $\alpha_{r,i}$ designating the i -th coordinate of the vector α_r .

¹⁴ Bienaymé explicitly calculated those terms which depend on moments up to the 4th order.

became greater than 0.¹⁵ Laplace followed that an “infinitely large and certain” total gain would be accumulated if only $q_i v_i - (1 - q_i)\mu_i > 0$ for all $1 \leq i \leq s$. By this application of the CLT, Laplace provided the basis for a theory of risk, which in turn would even play an important role in the history of the CLT during the 1920s (see Sect. 5.2.8.1).

2.2 Poisson's Modifications

Among all contributions of the 19th century in connection with Laplace's CLT aiming at a more comprehensible presentation or at modifications of the Laplacian methods according to contemporary analytical standards, the two approaches [1824; 1829] by Siméon Denis Poisson (1781–1840) had a special influence on the contributions of later authors. Poisson shared Laplace's view on the status of probability theory in the classical sense.¹⁶ Concerning moral problems, however, Poisson generalized Laplace's stochastic models to a considerable extent, and he did not share Laplace's cautious attitude toward these issues. Poisson's idea of all processes in the physical and moral world being governed by distinct mathematical laws is in line with his attempts toward a more exact mathematical analysis. Accordingly, the consequences for CLT were twofold: Firstly, Poisson formulated and proved this theorem generally for “choses,” thus creating an early concept of random variables, and secondly, he tried to discuss the validity of this theorem, mainly through counterexamples.

2.2.1 Poisson's Concept of Random Variable

In the first [1824] of the above-mentioned articles, Poisson treated sums and linear combinations of observational errors with different (not necessarily symmetrical) distributions, followed by a discussion of the Laplacian foundation of least squares. In the second article of 1829, he took up the issue from a far more general point of view. There, Poisson investigated asymptotic behavior of the distribution of a sum of functions (!) of the values of a “thing” (“chose”), where in several independent experiments these values were obtained with possibly different probabilities. The additional complication of considering a “function” essentially served to cover both sums of random values and of powers of these values within the same theorem. From today's point of view, all these quantities would plainly be described as random variables. Thus, Poisson's concept of the values of a “thing” was directed primarily

¹⁵ Apparently, Laplace tacitly assumed the existence of positive constants a, b such that $q_i v_i - (1 - q_i)\mu_i > a$ and $(v_i + \mu_i)^2 < b$ for all i .

¹⁶ Poisson's work in probability is well described in [Sheynin 1978; Bru 1981; Hald 1998; Sheynin 2005b].

toward the most important applications, and was still far away from the modern conception of abstract “random variable,” as explained by Kolmogorov [1933a].¹⁷

2.2.2 Poisson’s Representation of the Probabilities of Sums

In his discussion of sums of independent random variables, Poisson normally assumed that each variable X_n took values within the interval $[a; b]$ ($-a$ and b could be even infinitely large) and had a density function f_n , which was introduced by $f_n(x) = F'_n(x)$, where $F_n(x) = P(X_n \leq x)$. In a manner similar to Laplace’s approach, Poisson started his analysis with discrete random variables. Unlike Laplace, however, he did not consider probabilities of single discrete values but immediately calculated, partly through combinatorial considerations, the probability that the sum $S_s = X_1 + \dots + X_s$ would be within certain limits. Through the strict use of infinitesimal quantities in the transition from discrete to continuous random variables, he [1829, 5; 1824, 275; 286] established the formula

$$P(c - \varepsilon \leq S_s \leq c + \varepsilon) = \frac{1}{\pi} \int_{-\infty}^{\infty} \left(\prod_{n=1}^s \int_a^b f_n(x) e^{\alpha x \sqrt{-1}} dx \right) e^{-\alpha c \sqrt{-1}} \sin(\varepsilon \alpha) \frac{d\alpha}{\alpha}. \quad (2.12)$$

The justification of this formula was incomplete, even from a contemporary point of view. But Poisson [1824, 276] examined the validity of (2.12) in the special case $s = 1$. By interchanging the order of integration he concluded from (2.12) that

$$P(c - \varepsilon \leq X_1 \leq c + \varepsilon) = \frac{1}{\pi} \int_a^b \int_{-\infty}^{\infty} \left(e^{(x-c)\alpha \sqrt{-1}} \sin(\varepsilon \alpha) \frac{d\alpha}{\alpha} \right) f_1(x) dx. \quad (2.13)$$

By virtue of the addition theorems for sine and cosine, and the well-known formula¹⁸

$$\int_0^{\infty} \frac{\sin(kx)}{x} dx = \frac{\pi}{2} \quad (k > 0),$$

he showed that

$$\int_{-\infty}^{\infty} e^{(x-c)\alpha \sqrt{-1}} \sin(\varepsilon \alpha) \frac{d\alpha}{\alpha} = \begin{cases} \pi & \text{for } x \in]c - \varepsilon; c + \varepsilon[\\ 0 & \text{for } x \notin [c - \varepsilon; c + \varepsilon]. \end{cases} \quad (2.14)$$

¹⁷ Poisson’s approach to random variables was taken up and further developed soon afterwards by Carl Friedrich Hauber [1830], in his “Theorie der mittleren Werthe” (“Theory of Mean Values”), in an interesting attempt to develop a concept of far-reaching generality for random variables, which were named “unbestimmte Größen” (“indetermined quantities”). Many properties of expectations and variances of sums or products of independent random variables which today belong to the standards of each elementary theory of random variables, were explicitly stated and proven for the first time by Hauber.

¹⁸ For a history of this formula, which can be essentially traced back to Euler and still plays an important role in several branches of analysis, see [Fischer 2007].

The required result

$$P(c - \varepsilon \leq X_1 \leq c + \varepsilon) = \int_{c-\varepsilon}^{c+\varepsilon} f_1(x) dx$$

followed immediately from (2.13) and (2.14). In turn, it must have been within Poisson's scope to establish (2.12) by means of (2.14), even in the general case of arbitrary s . But only Dirichlet and Cauchy, as we will see below, directly used the jump function in (2.14) for elegant derivations of formulae equivalent to (2.12) for the probabilities of sums. Dirichlet at least was most probably motivated by Poisson's discussion of (2.13) and (2.14).

Dealing with the general case, Poisson set

$$\int_a^b f_n(x) \cos(\alpha x) dx =: \rho_n \cos \varphi_n, \quad \int_a^b f_n(x) \sin(\alpha x) dx =: \rho_n \sin \varphi_n, \quad (2.15)$$

and

$$R := \rho_1 \cdots \rho_s, \quad \psi := \varphi_1 + \cdots + \varphi_s. \quad (2.16)$$

Using $R(-\alpha) = R(\alpha)$ and $\psi(-\alpha) = -\psi(\alpha)$, he concluded from (2.12):

$$P(c - \varepsilon \leq S_s \leq c + \varepsilon) = \frac{2}{\pi} \int_0^\infty R \cos(\psi - c\alpha) \sin(\varepsilon\alpha) \frac{d\alpha}{\alpha}. \quad (2.17)$$

In his article of 1824, Poisson dealt with the case of an “infinitely large” s by calculating with infinitely small and infinitely large quantities. In his article of 1829, however, series expansions constituted the analytical background for an approximation with “large” (but not infinite) s . Afterwards, Poisson apparently preferred the second version (described in detail by Hald [1998, 317–327]), which was also adopted into his major probabilistic work, the *Recherches sur la probabilité des jugements en matière criminelle et en matière civile* [1837].

2.2.3 The Role of the Central Limit Theorem in Poisson's Work

As we will see in the following, Poisson's work on the CLT was based on Laplace's ideas on the one hand; on the other hand, however, Poisson's discussion of new analytical aspects paved the way toward a more rigorous treatment of the CLT.

2.2.3.1 Poisson's Version of the Central Limit Theorem

Poisson's results concerning the CLT can be summarized in modern terminology essentially as follows:

Let X_1, \dots, X_s be a great number of independent random variables with density functions which decrease sufficiently fast (Poisson did not specify exactly how fast)

as their arguments tend to $\pm\infty$. It is supposed that for the absolute values $\rho_n(\alpha)$ of the characteristic functions of X_n (see (2.15)) there exists a function $r(\alpha)$ independent of n with $0 \leq r(\alpha) < 1$ for all $\alpha \neq 0$ such that

$$\rho_n(\alpha) \leq r(\alpha). \quad (2.18)$$

Then, for arbitrary γ, γ' ,

$$P\left(\gamma \leq \frac{\sum_{n=1}^s (X_n - EX_n)}{\sqrt{2 \sum_{n=1}^s \text{Var} X_n}} \leq \gamma'\right) \approx \frac{1}{\sqrt{\pi}} \int_{\gamma}^{\gamma'} e^{-u^2} du, \quad (2.19)$$

where the approximation becomes all the better the larger s is, and the difference between the left and the right side becomes “infinitely small” with “infinite” s . Strictly speaking, Poisson’s analysis could be used for arbitrary γ, γ' , though he explicitly expressed end results in the sense of (2.19) only for the special case $\gamma = -\gamma' < 0$.

Poisson was convinced that this CLT was also valid for discrete random variables. In this case one could, according to Poisson [1837, 274 f.], assume that the values c_1, \dots, c_v of a random variable of this kind were subject to the respective probabilities $\gamma_1, \dots, \gamma_v$ which were represented by $\gamma_i = \int_{c_i-\delta}^{c_i+\delta} f(z) dz$ with an “infinitely small” quantity δ and a “discontinuous” density function f .¹⁹

As with Laplace, the CLT for Poisson was an important tool of classical probability, but not an autonomous mathematical theorem. Unlike Laplace, however, Poisson pointed out essential presuppositions “en passant,” such as the above-mentioned condition (2.18) for characteristic functions, and he discussed counterexamples to an overall validity of asymptotic normal distributions for sums. The most prominent of these counterexamples [Poisson 1824, 278] concerns the sum of identically distributed random variables with the probability density

$$f(x) = \frac{1}{\pi(1+x^2)},$$

for which the direct evaluation of (2.12) shows that

$$P(c - \varepsilon \leq \sum X_n \leq c + \varepsilon) = \frac{1}{\pi} \arctan\left(\frac{2\varepsilon s}{s^2 + c^2 - \varepsilon^2}\right).$$

Therefore in this case, even for large s , an approximate normal distribution can not be reached. Poisson [1824, 280] pointed out, however, that such cases of very slowly decreasing densities would not occur in practice, because all errors of observation were uniformly bounded in reality. Random variables with the density function f would later play an important role in Cauchy’s critical discussion of least squares (see Sect. 2.5.2). In fact, such random variables are now called “Cauchy-distributed.”

¹⁹ Poisson at this place used the adjective “discontinuous” in the traditional sense, as being inaccessible to a representation through a uniform algebraic term.

The significance of his condition for characteristic functions (2.18) Poisson [1824, 289–291] illustrated by two similar examples, where neither the assertion of the CLT was true nor this condition was met: He considered linear combinations $\sum_{n=1}^s \gamma_n \epsilon_n$ of identically distributed errors obeying the law

$$f(x) = e^{-2|x|}.$$

Using the formula (2.12) he showed that, for an “infinitely large” s ,

$$P(-c \leq \sum \gamma_n \epsilon_n \leq c) = \frac{1 - e^{-2c}}{1 + e^{2c}} \quad \text{if} \quad \gamma_n = \frac{1}{n},$$

and

$$P(-c \leq \sum \gamma_n \epsilon_n \leq c) = 1 - \frac{4}{\pi} \arctan(e^{-2c}) \quad \text{if} \quad \gamma_n = \frac{1}{2n-1}.$$

According to Poisson, in the first example we have

$$\rho_1(\alpha) \cdots \rho_s(\alpha) = \frac{1}{(1 + \frac{\alpha^2}{4})(1 + \frac{\alpha^2}{4 \cdot 4}) \cdots (1 + \frac{\alpha^2}{4s^2})} \rightarrow \frac{\pi\alpha}{e^{\frac{1}{2}\pi\alpha} - e^{-\frac{1}{2}\pi\alpha}},$$

whereas in the second

$$\rho_1(\alpha) \cdots \rho_s(\alpha) = \frac{1}{(1 + \frac{\alpha^2}{4})(1 + \frac{\alpha^2}{4 \cdot 9}) \cdots (1 + \frac{\alpha^2}{4(2s-1)^2})} \rightarrow \frac{2}{e^{\frac{\pi\alpha}{4}} + e^{-\frac{\pi\alpha}{4}}}.$$

2.2.3.2 Poisson's Law of Large Numbers

Regarding error theory, Poisson hardly made any modifications to the Laplacian discussion of least squares based on the CLT. Yet the discussion of (in modern terms) stochastic convergence of mean values and relative frequencies, respectively, which did not play a too dominant role in Laplace's work, became vital for Poisson and his major probabilistic work, the *Recherches*. Like Laplace, Poisson based such considerations on the CLT.

The approximate stability of arithmetic means or relative frequencies, quite often observed within different sequences of random experiments of the same kind, was so important for Poisson's probabilistic approach that he coined the term “law of large numbers” for this fact. In the introduction of his *Recherches*, he characterized this law as follows:

The phenomena of any kind are subject to a general law, which one can call the *Law of Large Numbers*. It consists in the fact, that, if one observes very large numbers of phenomena of the same kind depending on constant or irregularly changeable causes, however not progressively changeable, but one moment in the one sense, the other moment in the other sense; one finds ratios of these numbers which are almost constant [Poisson 1837, 7].

It must be emphasized that Poisson's interpretation of “law of large numbers” is different from the modern definition of this term.

For a “proof” of his law of large numbers, Poisson [1837, 139–143, 277 f.] introduced a special two-stage model of causation for the occurrence of an event (or, more generally, for the occurrence of a special value of a “chose”), and he established two auxiliary theorems on stochastic convergence: the first concerning the arithmetic means of non-identically distributed random variables, the second concerning the relative frequencies of an event which generally does not occur with constant probability. He based these theorems, which are equivalent to the *now* so-called “laws of large numbers,” on his general CLT (for comprehensive historical accounts see [Bru 1981, 69–75] and [Hald 1998, 577–580]). A distinct deviation of the relative frequencies with which a certain event had occurred in different sequences of observations respectively, possibly gave rise to the assumption that these sequences originated from different systems of causation. In the third part of his *Recherches*, Poisson gave a probabilistic discussion of the significance of such hypotheses in the context of conviction rates, and he essentially used the CLT for calculating the respective probabilities (see [Stigler 1986, 186–194] for a detailed discussion).

Poisson’s law of large numbers (in its original form) was heavily criticized during the 19th century. Among these discussions, two crucial points became subject of debates: the practical meaning of Poisson’s causation system was scrutinized (mainly by Bienaymé, see [Stigler 1986, 185; Heyde & Seneta 1977, 46–49]), and the analytical rigor of the deduction of the “auxiliary” CLT was questioned. Chebyshev [1846, 17] criticized that Poisson’s analysis was only “approximative,” and did not provide exact “error limits.” In this way he showed a—still rather vague—unease with Poisson’s analytical approach. One can, however, interpret Chebyshev’s criticism as an indication of the shift from “classical” probability, chiefly determined by its applications, toward a “new mathematical” probability. Perhaps, Chebyshev’s objections resulted from Poisson’s (as well as Laplace’s) procedure of neglecting “higher” series terms without giving any justification for that. Yet, if this was the case, Chebyshev did possibly not realize that Poisson had given an—at least indirect—justification of this procedure with his first, infinitistic approach.

2.2.4 Poisson’s Infinitistic Approach

Poisson’s discussions of 1824 and 1829 on the CLT were essentially equivalent. The first account, however, clarified the fundamentals of Laplace’s method of approximations as applied to the CLT much more directly, and, as we will see below, paved the way for a more “rigorous” treatment of asymptotic normal distributions for sums of independent random variables. For a discussion of the essentials of Poisson’s “first” approach it is sufficient to confine the description to the special case of identically distributed random variables with a density f_1 vanishing beyond the finite interval $[a; b]$.

From (2.15), (2.16), (2.17) one gets with

$$\rho := \rho_1 = \sqrt{\left(\int_a^b f_1(x) \cos(\alpha x) dx\right)^2 + \left(\int_a^b f_1(x) \sin(\alpha x) dx\right)^2},$$

and $\varphi := \varphi_1$:

$$P(c - \varepsilon \leq S_s \leq c + \varepsilon) = \frac{2}{\pi} \int_0^\infty \rho^s \cos(s\varphi - c\alpha) \sin(\varepsilon\alpha) \frac{d\alpha}{\alpha}. \quad (2.20)$$

Poisson [1824, 279–281] carefully justified that $0 < \rho < 1$ if $\alpha \neq 0$. The following excerpt illustrates his notion of “infinite” quantities and his handling of these quantities in connection with an asymptotic representation of $P(c - \varepsilon \leq S_s \leq c + \varepsilon)$:

We want to consider the number s infinitely large, such that the following formulae are rigorously true at this limit, and the more approximated, the larger s is. Now, from the quantity ρ being less than 1 if the variable α is not $= 0$ it follows that at the limit $s = \infty$ the power ρ^s attains finite values only for infinitely small values of this variable, and becomes infinitely small if α has a finite value [Poisson 1824, 280].

Poisson expressed in this text the contemporary view of the meaning of “approximation”: Approximation formulae had to be “rigorously true” at the “limit.” Moreover, he considered, as can be inferred from his phrasing “limit,” an “infinite” quantity not as actually infinite. On the other hand, he treated infinitely small quantities as belonging to the common system of numbers.²⁰ This ambivalence in the attitude toward the infinite is typical for the “infinitesimal” period in the first half of the 19th century, which led away from the priority of algebraic analysis.

On the basis of the above-cited comment and on account of $\cos(\alpha x) \approx 1 - \frac{\alpha^2 x^2}{2}$ and $\sin(\alpha x) \approx \alpha x$ for “infinitely small” α , Poisson could deduce—at least for a finite interval $[a; b]$:

$$\rho^s \approx \begin{cases} (1 - h^2 \alpha^2)^s & \text{for an “infinitely small” } \alpha \\ 0 & \text{otherwise,} \end{cases}$$

$$\text{where } h^2 := \frac{1}{2} \left(\int_a^b x^2 f_1(x) dx - \left(\int_a^b x f_1(x) dx \right)^2 \right).$$

The sign \approx (not explicitly used by Poisson) is used here to indicate an “infinitely close” position of one value to another. Poisson [1824, 281] set $\alpha =: y/\sqrt{s}$, “where the new variable y can attain finite values.” Taking into account that $\rho \approx 1$ and $\int_a^b f_1(x) \sin(\alpha x) dx \approx \int_a^b f_1(x) \alpha x dx$ for $\alpha \approx 0$, he concluded that $\sin \varphi \approx k\alpha$ for $\alpha \approx 0$, where k is the expectation of the random variables. As a result of $\sin \varphi \approx \varphi$ for $\sin \varphi \approx 0$ it followed that $\varphi \approx k\alpha$ for $\alpha \approx 0$.

In this way Poisson obtained for “infinitely large” s on account of $(1 - \frac{h^2 y^2}{s})^s \approx e^{-h^2 y^2}$:

$$P(c - \varepsilon \leq S_s \leq c + \varepsilon) \approx \frac{2}{\pi} \int_0^\infty e^{-h^2 y^2} \cos[(ks - c) \frac{y}{\sqrt{s}}] \sin \frac{\varepsilon y}{\sqrt{s}} \frac{dy}{y}. \quad (2.21)$$

²⁰ For Poisson's general preference to infinitesimals see [Schubring 2005, 455 f.].

For Poisson's inference from (2.20) to (2.21) further explanations would have been necessary. The only comment which Poisson gave in this context was in relation to (2.21):

Strictly speaking, one is allowed to attribute to the variable y only finite values; because of the exponential factor $e^{-h^2 y^2}$, however, one can expand the respective integral into the infinite, without a considerable error [Poisson 1824, 282].

From a rigorous point of view, one can deduce from (2.20) only that, for an arbitrarily large but finite Y ,

$$P(c - \varepsilon \leq S_s \leq c + \varepsilon) \approx \frac{2}{\pi} \int_0^Y e^{-h^2 y^2} \cos[(ks - c) \frac{y}{\sqrt{s}}] \sin \frac{\varepsilon y}{\sqrt{s}} \frac{dy}{y} + \frac{2}{\pi} \int_{\frac{Y}{\sqrt{s}}}^{\infty} \rho^s \cos(s\varphi - c\alpha) \sin(\varepsilon\alpha) \frac{d\alpha}{\alpha}.$$

Apparently, for Poisson it was a matter of course, which did not need any special justification, that

$$\int_{\frac{Y}{\sqrt{s}}}^{\infty} \rho^s \cos(s\varphi - c\alpha) \sin(\varepsilon\alpha) \frac{d\alpha}{\alpha} \approx 0$$

for an “infinitely large” s .

From (2.21) one could infer, with the aid of the relation

$$\frac{1}{y} \cos[(ks - c) \frac{y}{\sqrt{s}}] \sin \frac{\varepsilon y}{\sqrt{s}} = \frac{1}{\pi \sqrt{s}} \int_{-\varepsilon}^{\varepsilon} \cos[(ks - c + z) \frac{y}{\sqrt{s}}] dz,$$

and consequently

$$\begin{aligned} \frac{2}{\pi} \int_0^{\infty} e^{-h^2 y^2} \cos[(ks - c) \frac{y}{\sqrt{s}}] \sin \frac{\varepsilon y}{\sqrt{s}} \frac{dy}{y} \\ = \frac{1}{\pi \sqrt{s}} \int_{-\varepsilon}^{\varepsilon} \left(\int_0^{\infty} e^{-h^2 y^2} \cos[(ks - c + z) \frac{y}{\sqrt{s}}] dy \right) dz, \end{aligned}$$

that

$$P(c - \varepsilon \leq S_s \leq c + \varepsilon) \approx \frac{1}{2h\sqrt{\pi s}} \int_{-\varepsilon}^{\varepsilon} e^{-\frac{(ks - c + z)^2}{4h^2 s}} dz. \quad (2.22)$$

Setting $c = ks$ and $\varepsilon = 2hr\sqrt{s}$ in (2.22), Poisson finally obtained the result:

$$P(ks - 2hr\sqrt{s} \leq S_s \leq ks + 2hr\sqrt{s}) \approx \frac{2}{\sqrt{\pi}} \int_0^r e^{-t^2} dt.$$

Poisson's discussion of sums of non-identically distributed random variables followed the model just described for identically distributed random variables. For the validity of his deductions in the general case, Poisson made a further condition explicit which was equivalent to (2.18).

2.2.5 Approximation by Series Expansions

The essential goal of Poisson's paper from 1829 on the CLT was to approximate the probabilities of a sum of a large number of random variables X_n , whose densities f_n vanish beyond a finite interval $[a; b]$, by a series expansion rather than to derive a "limiting formula." By the argument that the product $\rho_1(\alpha) \cdots \rho_s(\alpha)$, where

$$\rho_n = \sqrt{\left(\int_a^b f_n(x) \cos(\alpha x) dx\right)^2 + \left(\int_a^b f_n(x) \sin(\alpha x) dx\right)^2},$$

attained values significantly different from zero for very small α only, Poisson justified that it was possible to cut off that series expansion after its first terms.

With the abbreviations $\int_a^b x f_n(x) dx =: k_n$, $\int_a^b x^2 f_n(x) dx =: k'_n$, \dots , and the designations (2.15), Poisson [1829, 8 f.] derived the series expansions

$$\begin{aligned}\rho_n \cos \varphi_n &= 1 - \frac{\alpha^2}{2} k'_n + \frac{\alpha^4}{4!} k_n''' - \dots, \\ \rho_n \sin \varphi_n &= \alpha k_n - \frac{\alpha^3}{3} k_n'' + \dots.\end{aligned}$$

Because of $|k_n| < |b| + |a|$, $|k'_n| < (|b| + |a|)^2, \dots$ these series are convergent. Poisson [1829, 8] expressed the opinion that this convergence guaranteed the respective left sides being actually represented by the series expansions on the right. In this way, Laplace's formal calculations according to his method of approximation, which in many cases led to divergent series, were substituted by an explicit discussion of convergence.

By use of the series expansions for $\rho_n \cos \varphi_n$ and $\rho_n \sin \varphi_n$, series expansions for R , ψ (see (2.16)), and for $\cos(\psi - c\alpha)$ in powers of α were accomplished such that, because of (2.17),

$$\begin{aligned}P(c - \varepsilon \leq S_s \leq c + \varepsilon) &= \frac{2}{\pi} \int_0^\infty e^{-\alpha^2 h s} (1 + \alpha^4 l s + \dots) \times \\ &\times (\cos[(k s - c)\alpha] + \alpha^3 g s \sin[(k s - c)\alpha] + \dots) \sin(\varepsilon \alpha) \frac{d\alpha}{\alpha}.\end{aligned}$$

In this formula k, h, g, l denote quantities depending on the moments of the single random variables; the absolute values of these quantities have upper bounds independent of s , as Poisson proved. In particular, $k = \frac{\Sigma E X_n}{s}$ and $h = \frac{\Sigma \text{Var} X_n}{2s}$ ensued. On the basis of these considerations Poisson apparently believed to have given an additional justification for the neglect of those terms which are, after having carried out the substitution $\alpha =: \beta / \sqrt{s}$, divided by a power of s larger than $s^{1/2}$. In this way, the approximation

$$\begin{aligned}
P(c - \varepsilon \leq S_s \leq c + \varepsilon) &\approx \frac{2}{\pi} \int_0^\infty e^{-\beta^2 h} \cos \frac{(ks - c)\beta}{\sqrt{s}} \sin \frac{\varepsilon\beta}{\sqrt{s}} \frac{d\beta}{\beta} + \\
&+ \frac{2g}{\pi\sqrt{s}} \int_0^\infty e^{-\beta^2 h} \sin \frac{(ks - c)\beta}{\sqrt{s}} \sin \frac{\varepsilon\beta}{\sqrt{s}} \beta^2 d\beta \quad (2.23)
\end{aligned}$$

was reached [Poisson 1829, 9].

In his 1929 paper, Poisson's further proceeding was rather complicated. A considerably simplified approach was given in his book [1837, 270 f.]: Poisson in (2.23) set $c = ks$ and $\varepsilon = 2\gamma\sqrt{hs}$, with the result

$$P(ks - 2\gamma\sqrt{hs} \leq S_s \leq ks + 2\gamma\sqrt{hs}) \approx \frac{2}{\pi} \int_0^\infty e^{-\beta^2 h} \sin(2\beta\gamma\sqrt{h}) \frac{d\beta}{\beta}. \quad (2.24)$$

Essentially making use of

$$\int_{-\infty}^\infty e^{-x^2} \cos(\alpha x) dx = \sqrt{\pi} e^{-\frac{\alpha^2}{4}},$$

Poisson showed that the integral in (2.24) was equal to

$$\frac{2}{\sqrt{\pi}} \int_0^\gamma e^{-u^2} du.$$

Poisson's preference for the just-described approach to the CLT by means of explicit series expansions might have been mainly caused by the fact that this method gave additional correction terms of the order $s^{-1/2}$ and less for "large" (but not infinite) s , and therefore was considered to be more general than the "simple" approximation by the normal distribution only. For the subsequent development of the CLT, Poisson's "infinitistic" approach seems to have been more influential, however.

2.3 The Central Limit Theorem After Poisson

During the time after Poisson, two crucial changes occurred in the development of probability theory. Firstly, probability eventually lost one of its major branches, the application to moral sciences. Secondly, the movement toward a purely mathematical view of stochastics, which in a certain sense had already begun with Laplace, gained momentum. The development of the CLT was connected with both fields, as we will see in the cases of Cauchy's and Dirichlet's contributions.

2.3.1 Toward a New Conception of Mathematics

Both Cauchy and Dirichlet are seen as representatives of a new mathematical conception emerging after 1800 which was generally accepted during the last third

of the 19th century. The essentials of this new point of view can be summarized as follows: A separation of mathematics from its ontological relation to the physical and moral world was beginning to form, as stated by Kline [1972, 619 f.]. In [Laugwitz 1999, 187–191] this development is described as a transition from the consideration of the “contents” to the discussion of the “scope” of “concepts.” The role of counterexamples in this context changed from irrelevant “curiosities” toward boundary posts indicating the limits of the specific concepts. Poisson, for example, still understood his examples of nonconvergence to the normal distribution in the sense of singular exceptions, which do not occur “in practice.” Without external criteria, such as applicability, however, mathematics experienced an increased need to reflect on its internal logical consistency, as pointed out by Mehrtens [1990]. In this sense, Poisson’s main counterexample would become especially important for Cauchy’s critique of the method of least squares.

The framework of the growing abstraction of mathematics during the 19th century can only be roughly described in this exposition. An excellent survey is given by Schneider [1981a]. There were changes in the employment of mathematicians (from 18th-century academies to universities), which helped to promote pure mathematics.²¹ The computational potentialities of analysis seemed to become gradually exhausted, so a turn to the discussion of analytical fundamentals or even to other, temporarily neglected disciplines, such as synthetic geometry, became plausible. The intellectual background was perhaps even more decisive. After the political upheavals due to the French Revolution, the confidence of the Enlightenment in a common standard of rationality began to vanish. The commonly accepted unity of mathematics and good sense began to drift apart (this process is exactly described by Daston [1988, 370–386], for the field of probability theory). The growing re-examination of basic definitions after 1800 can be considered as a reaction to the decline of the idea of self-evident “natural” standards.

The resulting changes toward “mathematical rigor” are not to be confused with changes in analytical style and methods. As several authors have pointed out since Lakatos [1966] and Spalt [1981], analytic reasoning during the first half of the 19th century using the language of infinitesimals was not fundamentally less rigorous than the application of epsilon methods.²² The decline of algebraic analysis, however, was closely connected to the new standards of mathematical rigor. This was also an essential point in the history of the CLT.

Certainly, the changes described above did not happen overnight. Cauchy and Dirichlet still worked a good deal in the tradition of problem solving of the 18th century. In the case of the CLT, however, the “new mathematics” can clearly be seen in the contributions of both authors.

²¹ For more material on this topic see [Mehrtens, Bos, & Schneider 1981; Schubring 2005, Chapt. VII].

²² Especially regarding Cauchy’s work, the discussion is still quite controversial, see Sect. 1.3.

2.3.2 *Changes in the Status of Probability Theory*

Several subjects of classical probability were heavily attacked after Laplace's death. His personal authority, however, remained unharmed. This criticism was mainly directed toward applications of probability to human decisions, for example at court trials. Especially Poisson's work in this field caused a broad disapproval of the claim of classical probability for universal applicability, at least in France.²³ Daston [1988, 384] has pointed out that, as a consequence, a shift from the focus on the individual man toward the probability of mass phenomena occurred. Naturally, the CLT was also an excellent tool for the latter field. A further consequence was that a more critical awareness replaced the "natural" and often only tacit presuppositions of classical probability also in "unsuspicious" applications, such as error theory. In this way, error theory became the discipline of probability being subject to the most far-reaching mathematization. Some sources showed a rather abstract view of error theory and gave rise to demanding analytical discussions. This development was responsible for Cauchy's "rigorous" proof of the CLT during his dispute with Bienaymé over the priority of the method of least squares, as we will see below.

At several occasions during his work, Laplace had already pointed out the extreme relevance of his analytic methods of probability theory, especially his methods for approximating integrals depending on large numbers. Thus, from the analytical point of view, statements now interpreted as probabilistic limit theorems became appendages of the theory of definite integrals. Based on this idea, Dirichlet rather frequently gave courses on probability theory during the 1830s and 1840s, in which he directly referred to Laplacian methods, however with considerable modifications toward a "new" analytical rigor, from which his "rigorous" proof of the CLT (discussed in detail below) resulted. In this context, the CLT reached a quality different from the framework of classical probability theory. It was no longer only a tool for applications beyond mathematics, but also became a subject within (pure) mathematics, albeit with a mainly auxiliary character (serving as an illustration of the theory of definite integrals).

2.3.3 *The Rigorization of Laplace's Idea of Approximation*

As we have seen in the discussion of Poisson's deduction of an approximate normal distribution for sums of independent random variables, the following basic idea (for the sake of simplicity described only for identically distributed random variables with symmetric density function f on $[-a; a]$) was pursued: The probability P that a sum of s random variables of this kind has values within $[b - c; b + c]$ is (cf. formula (2.12)):

²³ There is also a German example: Jakob Fries's *Versuch einer Kritik der Prinzipien der Wahrscheinlichkeitsrechnung* [1842], which was based on Kant's philosophy, and met with Gauss's approval; see [Fischer 2004].

$$P = \frac{2}{\pi} \int_0^\infty \left(\int_{-a}^a f(x) \cos(\alpha x) dx \right)^s \cos(b\alpha) \sin(c\alpha) \frac{d\alpha}{\alpha}.$$

As expressed in the infinitesimal style of the first half of the 19th century, the power

$$\left(\int_{-a}^a f(x) \cos(\alpha x) dx \right)^s$$

with the “infinitely large” exponent s attains values which differ essentially from 0 only for “infinitely small” α . The whole integrand is, as a function of α , similar to a bell-shaped function, whose maximum peak becomes sharper and sharper as s increases. This circumstance gives rise to the conjecture that for “infinitely large” s the whole range $]-\infty; \infty[$ of the integral with respect to α can, with only an “infinitely small” error, be reduced to an “infinitely small” neighborhood of $\alpha = 0$. It was exactly the latter point which was used by Poisson (and many of his imitators) without any detailed justification. But, why should it be impossible for the value of the integral of an “infinitely” small function to be considerably large if the domain of integration itself is unbounded? This unsolved problem corresponded, in the end, to the unjustified neglect of higher terms in the approach via series expansions, and was most probably responsible for the already described unease (see Sect. 2.1.4) associated with Laplace’s deduction of the CLT.

A more exact analysis of the CLT, which explicitly referred to the basic idea of the Laplacian method of approximation, had to show that for r in a specified range of “infinite smallness” the integral

$$\int_r^\infty \left(\int_{-a}^a f(x) \cos(\alpha x) dx \right)^s \cos(b\alpha) \sin(c\alpha) \frac{d\alpha}{\alpha}$$

would in fact become “infinitely small” for “infinitely large” values of s . As we have seen, Poisson’s analysis had already shown that r had to be of an order around $1/\sqrt{s}$. Corresponding considerations were to be applied in the general case of non-identically and non-symmetrically distributed random variables.

Similar ideas led to Cauchy’s sketch of the rigorous proof of a (if still rather specific) CLT in 1853, and also to Lyapunov’s epochal proof of a very general form of the theorem in 1900/01. Cauchy had already begun in the 1820s to discuss “functions of great numbers,” such that his work of 1853 was not only connected with error theory but was also produced in the broader context of his analytical studies. Dirichlet had, independent of Cauchy and actually even before him, also advanced similar ideas. He did not, however, publish his results, but only presented them in his lecture course of 1846.

2.4 Dirichlet's Proof of the Central Limit Theorem

Peter Gustav Lejeune Dirichlet (1805–1859) is renowned for his pioneering contributions to mathematical physics and number theory. In the field of probability theory, however, one can find only a few brief notices in Dirichlet's collected *Werke*. Actually, during his Berlin period (1828–1855), he quite frequently gave courses on probability and error theory presenting new and original ideas, as we can see from unpublished lecture notes (see [Fischer 1994]). In these lecture courses, Dirichlet's main concern was not the treatment of probabilistic fundamentals or applications, but rather the discussion of demanding analytical problems. He considered these problems as applications of the theory of definite integrals, and therefore plainly named several of the pertinent courses “Anwendungen der Integralrechnung” (“applications of integral calculus”). In one of these “Anwendungen”—dedicated to foundational issues of least squares that served as a 1-hr appendage to a 4-hr course on definite integrals in 1846²⁴—one can find a very notable and innovative approach to a proof of the CLT.

Dirichlet's analytical style varied between an almost “epsilonic” presentation, as used in his publications, and a rather intuitive handling of problems, quite often connected with infinitistic methods. Evidence of this can be found in his lectures or unpublished drafts (see [Fischer 1994]). The style of Dirichlet's contribution to the CLT [1846] seems mainly of the second kind; yet, as we will see, all essential steps (only sketched out in the original source) can be taken using finitistic considerations which were within Dirichlet's scope.

2.4.1 Dirichlet's Modification of the Laplacian Method of Approximation

Dirichlet's main probabilistic interests lay in problems of approximating “functions of large numbers.” Thus, he actually satisfied Laplace's hope that such questions would interest the “geometers” (see the introductory part of the present chapter). At the same time, one can see in Dirichlet's activities a shift from the typical objects of classical probability, concentrating on practical applicability, toward the discussion of the respective analytical methods.

In the 1830s, Dirichlet presented (e.g., [1838, 67 f.]) Laplace's original deduction of Stirling's formula in his lectures. He succeeded at least in deducing the law of Laplace's series (2.2), which Cauchy [1844, 68] would still consider to be unknown. As we have seen in Sect. 2.1.2, Laplace had set

$$\Gamma(s + 1) = M \int_{-s}^{\infty} e^{-z} (1 + z/s)^s dz = M \int_{-\infty}^{\infty} e^{-t^2} \frac{dz}{dt} dt,$$

²⁴ The corresponding lecture notes, written by an unknown author, are undated. From all we know about Dirichlet's teaching activities in probability theory, it seems evident, however, that the lecture notes pertain to Dirichlet's course in summer semester 1846 [Fischer 1994, 56, 60].

where z is a power series in t and $M = e^{-s}s^s$. Dirichlet differentiated the equality

$$e^{-z}(1 + z/s)^s = e^{-t^2}$$

by t to obtain

$$z \frac{dz}{dt} = 2t(s + z).$$

By employing the formula $z = k_1 t + k_2 t^2 + \dots$ with unknown coefficients k_i ($z = 0$ if and only if $t = 0$) in the latter equation and by comparing the coefficients of powers of t , Dirichlet determined the first terms of $\sum_{n \geq 1} k_n t^n$. In essence, he developed the recursion formula

$$k_1 = \sqrt{2s}, \quad k_n = \frac{2k_{n-1}}{(n+1)k_1} - \frac{1}{2k_1} \sum_{i=2}^{n-1} k_i k_{n+1-i} \quad (n \geq 2).$$

From this, the series expansion

$$\Gamma(s+1) = s^{s+1/2} e^{-s} \sqrt{2\pi} \left(1 + \sum_{n \geq 1} \frac{1 \cdot 3 \cdot 5 \cdots (2n+1) a_{2n+1}}{s^n} \right),$$

where

$$a_i = 2^{1-i} (\sqrt{2s})^{i-2} k_i$$

follows. (Dirichlet, however, made explicit only the first terms of the latter series expansion, which can also be deduced by different “modern” methods, see [Copson 1965, 53–57; Fischer 2006].)

In the 1840s, Dirichlet's interest in Stirling's formula no longer aimed at formal series expansions, but at a modification of the basic procedure concerning the Laplacian method of approximation, in exactly the sense which was described in Sect. 2.3.3 for the case of the CLT. Dirichlet [1841/42, 56–61] split the entire integral

$$\int_{-n}^{\infty} e^{-z} \left(1 + \frac{z}{n}\right)^n dz = \int_{-n}^{\infty} y dz = \Gamma(n+1) e^n n^{-n}$$

into the sum

$$\int_{-n}^{-n^m} y dz + \int_{-n^m}^{n^m} y dz + \int_{n^m}^{\infty} y dz = I_1 + I_2 + I_3,$$

where $\frac{1}{2} < m < \frac{2}{3}$. He set $y(z) = e^{-t^2(z)}$, and considering the convergent (!) series expansion of $\log y(z)$ around $z = 0$ (the abscissa of the maximum of y) he showed that I_1 and I_3 tend to 0 as n increases indefinitely, whereas

$$\frac{I_2}{\sqrt{2n}} \rightarrow \int_{-\infty}^{\infty} e^{-u^2} du = \sqrt{\pi}.$$

Thus, he obtained the expected result for $\Gamma(n+1)$ for “infinitely large” n (for more details see [Fischer 1994, 49 f.]).

2.4.2 The Application of the Discontinuity Factor

In order to adopt his reasoning from the case of Stirling's formula to the CLT, Dirichlet first needed an appropriate representation of the exact probabilities for sums or linear combinations of random variables. As one can see from the development of Dirichlet's ideas, as represented in his lectures of 1838 compared to his lectures of 1846, Poisson's discussion of the jump function (2.14) apparently led to Dirichlet's general method of calculating integrals over complicated domains with the aid of "discontinuity factors."

In his courses on Laplacian error theory as of 1838 and 1846, Dirichlet proposed the central problem of finding an approximate term for the probability P that the value of the linear combination $\alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_n x_n$ was within $-\lambda'$ and $+\lambda'$, where $\lambda' = \lambda \sqrt{n}$ and λ was a given positive constant. More precisely, x_1, x_2, \dots, x_n stood for independent observation errors (n being a large number), with expectations 0 and with (in general different) symmetric probability densities f_1, f_2, \dots, f_n , vanishing beyond the finite interval $[-a; a]$.

Initially, Dirichlet [1838, 142–144] repeated Poisson's "combinatorial" procedure for the deduction of a formula for the probability that a linear combination of errors is within a given interval (see Sect. 2.2.2). But then, he presented—unlike Poisson also for the general case of arbitrary n and arbitrary λ —the application of "his" discontinuity factor

$$\frac{2}{\pi} \int_0^\infty \frac{\sin \varphi}{\varphi} \cos(k\varphi) d\varphi = \begin{cases} 0 & \text{for } |k| > 1 \\ 1 & \text{for } -1 < k < 1, \end{cases} \quad (2.25)$$

which he deduced from

$$\int_0^\infty \frac{\sin(kt)}{t} dt = \frac{\pi}{2} \quad (k > 0)$$

using trigonometric addition theorems. To this aim he calculated the probability

$$P = \int_G f_1(x_1) \cdots f_n(x_n) dx_1 \cdots dx_n,$$

where

$$G = \{x \in \mathbb{R}^n \mid -\lambda' < \alpha_1 x_1 + \cdots + \alpha_n x_n < \lambda'\},$$

by use of the jump function (2.25), with the result

$$P = \frac{2}{\pi} \int_{[-a;a]^n} \int_0^\infty f_1(x_1) \cdots f_n(x_n) \frac{\sin \varphi}{\varphi} \cos[(\alpha_1 x_1 + \cdots + \alpha_n x_n) \frac{\varphi}{\lambda'}] d\varphi dx_1 \cdots dx_n. \quad (2.26)$$

Dirichlet [1839a;b;c] published three papers in which the jump function (2.25) was used for the calculation of specific multiple integrals that were important for

the determination of space volumes and for potential theory, but he did not mention any applications in probability theory. In his 1846 course, he totally ignored Poisson's combinatorial approach, and by applying his discontinuity function he deduced (2.26) through a consideration of the analogies between probabilities and space volumes.²⁵ From (2.26) Dirichlet [1846, 27] deduced

$$\begin{aligned} P &= \frac{2}{\pi} \int_0^\infty \frac{\sin(\lambda' \varphi)}{\varphi} \int_{-a}^a f_1(x_1) e^{\alpha_1 x_1 \varphi \sqrt{-1}} dx_1 \cdots \int_{-a}^a f_n(x_n) e^{\alpha_n x_n \varphi \sqrt{-1}} dx_n d\varphi \\ &= \frac{2}{\pi} \int_0^\infty \frac{\sin(\lambda \sqrt{n} \varphi)}{\varphi} \int_{-a}^a f_1(x_1) \cos(\alpha_1 x_1 \varphi) dx_1 \cdots \\ &\quad \cdots \int_{-a}^a f_n(x_n) \cos(\alpha_n x_n \varphi) dx_n d\varphi. \quad (2.27) \end{aligned}$$

The interchange of the order of integration was not discussed. In his paper [1839c], however, Dirichlet—without referring to probabilistic applications—pointed out the need for a proof of such interchanges. He suggested multiplying the integrands with factors such as $e^{-r\varphi}$. For $r > 0$ the absolute “convergence” of the modified integrals would be guaranteed (and, thus, the interchangeability of the order of integration). For both multiple integrals, the one before and the other after the interchange, one had finally to examine the limit $r \rightarrow 0$. Actually, this method is practical in the case of the probabilities of linear combinations of mutually independent random variables if one assumes for the densities of these variables certain—not very drastic—conditions, from which the absolute integrability of the function $\varphi \mapsto \frac{\sin \varphi}{\varphi} \int_{[-a;a]} f_1(x_1) \cdots f_n(x_n) \cos[(c_1 x_1 + \cdots + c_n x_n) \varphi] dx_1 \cdots dx_n$ over $[0; \infty[$ follows for fixed c_1, \dots, c_n . The hypotheses regarding the density functions, which Dirichlet supposed more or less tacitly, are in fact sufficient for this condition.²⁶

2.4.3 Dirichlet's Proof

Dirichlet's discussion of the asymptotic distribution of linear combinations of observational errors can be reconstructed in the sense of a rigorous proof of the CLT, even from today's point of view.

²⁵ Glaisher [1872a, 195; 1872b, 98] was perhaps the first—of course without being directly influenced by Dirichlet—to publish the use of Dirichlet's factor in exactly the same way as Dirichlet had presented it in his 1846 lecture course. [Cauchy 1853d], as it seems without knowledge of Dirichlet's prior contributions, had already given a very similar consideration, see Sect. 2.5.2.

²⁶ For an account of post-Weierstrassian era on the problem of interchanging the order of integration in applying Dirichlet's factor, see [David 1909].

2.4.3.1 Tacit Assumptions and Proposition

As described above, Dirichlet discussed linear combinations $\alpha_1 x_1 + \dots + \alpha_n x_n$ of random errors. The densities of these errors were not only considered to be symmetric and concentrated on a fixed interval, but also to be smooth (in the sense of the existence of continuous derivatives) and unimodal, as it appears from a picture in the lecture notes [1846, 21]. The latter assumption was, however, not absolutely necessary for Dirichlet's deductions. As we will see, Dirichlet tacitly presupposed that the sequence of the α_v had a positive lower bound (named α by me) and a positive upper bound (A), and that all variances of the random errors should be uniformly bounded away from 0 (by a positive lower bound to which I refer as k). Such tacit assumptions were natural within error theory. For a rigorous completion of Dirichlet's line of proof in the case of non-identically distributed observation errors, one has to additionally assume a certain uniformity in the shape of all the density functions, such as, for example, the existence of an upper bound C such that $|f'_v(x)| < C$ for all $x \in [-a; a]$ and all v . (From this condition one can already deduce the existence of the above-mentioned constant k .)

Expressed as a "modern" limit assertion, the main result of Dirichlet's lecture course on error theory in 1846 was

$$\left| P \left(-\lambda \sqrt{n} \leq \sum_{v=1}^n \alpha_v x_v \leq \lambda \sqrt{n} \right) - \frac{2}{\sqrt{\pi}} \int_0^{\lambda/r} e^{-s^2} ds \right| \rightarrow 0 \quad (n \rightarrow \infty),$$

where

$$r = 2 \sqrt{\frac{1}{n} \sum_{v=1}^n k_v \alpha_v^2}.$$

Even if the transcriber of the lecture notes did apparently not render all arguments entirely correctly, the basic ideas for a rigorous proof of this limit can be clearly discerned. At least in the special case of identically distributed errors a complete argumentation can be reached with such methods that Dirichlet himself used.²⁷

2.4.3.2 Dirichlet's Discussion of the Limit

Analogous to his derivation of Stirling's formula, Dirichlet split the integral (2.27) with respect to φ into three parts

$$\frac{2}{\pi} \int_0^\delta \dots d\varphi + \frac{2}{\pi} \int_\delta^\Delta \dots d\varphi + \frac{2}{\pi} \int_\Delta^\infty \dots d\varphi = p + q_1 + q_2,$$

where δ and Δ depend on n in such a way that

$$\delta \sqrt[4]{n} \rightarrow 0, \quad \delta \sqrt{n} \rightarrow \infty \quad (n \rightarrow \infty) \tag{2.28}$$

²⁷ For an edition of the original source see Appendix.

and

$$\Delta \propto n^\gamma \quad \text{with an arbitrary, but fixed } \gamma > 0. \quad (2.29)$$

Dirichlet represented the product $\Pi(\varphi)$ of the integrals

$$g_v(\varphi) := \int_{-a}^a f_v(x_v) \cos(\alpha_v x_v \varphi) dx_v$$

by

$$\Pi(\varphi) = e^{-\sum k_v \alpha_v^2 \varphi^2} e^{R(\varphi)}, \quad k_v := \frac{1}{2} \int_{-a}^a z^2 f_v(z) dz. \quad (2.30)$$

It was not explained in the lecture notes [Dirichlet 1846] that for general densities f_v this representation with real $R(\varphi)$ is only valid for sufficiently small φ , and therefore only in the first of the three integrals for small δ . Since $g_v(\varphi) > 0$ for $0 \leq \varphi \leq \frac{\pi}{2Aa}$, $R(\varphi)$ exists for at least all $\varphi \in [0; \frac{\pi}{2Aa}]$. Dirichlet perhaps supposed unimodal densities f which diminish sufficiently fast with growing absolute values of the argument; then the term $\int_{-a}^a f(x) \cos(\alpha \varphi x) dx$ is positive for all α and all φ . As we will see below, however, it actually suffices that (2.30) holds for a small interval of φ -values.

In order to justify the asymptotic disappearance of

$$R(\varphi) = \sum_{v=1}^n \left(\log \left(\int_{-a}^a f_v(x_v) \cos(\alpha_v x_v \varphi) dx_v \right) + k_v \alpha_v^2 \varphi^2 \right)$$

in the first integral, Dirichlet expanded each logarithmic term into a power series of φ (in each case he explicitly took into account only the first nontrivial power of φ), and thus obtained for $0 \leq \varphi \leq \delta$ an estimate equivalent to the form

$$|R(\varphi)| < nL\delta^4 + nM\delta^6 + \dots. \quad (2.31)$$

L, M, \dots designate the absolute values of the largest coefficients of $\varphi^4, \varphi^6, \dots$ among all expansions of the individual logarithmic terms, and are therefore constants depending only on the functions f_v and the multipliers α_v . Dirichlet did not discuss the exact form of these constants. On the basis of (2.31) and (2.28) Dirichlet concluded that $R(\varphi)$ could be neglected in the first integral

$$p = \frac{2}{\pi} \int_0^\delta \frac{\sin(\lambda \sqrt{n} \varphi)}{\varphi} \Pi(\varphi) d\varphi = \frac{2}{\pi} \int_0^\delta \frac{\sin(\lambda \sqrt{n} \varphi)}{\varphi} e^{-\sum k_v \alpha_v^2 \varphi^2} e^{R(\varphi)} d\varphi$$

as $n \rightarrow \infty$. For a complete justification (see [Fischer 2000, sect. 2.3.1]) of Dirichlet's hints, one can show, with the aid of the elementary inequalities

$$\cos z \geq 1 - \frac{z^2}{2},$$

$$\cos z \leq 1 - \frac{z^2}{2} + \frac{z^4}{24},$$

$$\log(z) < z,$$

$$\log(1 - z) > -z - 2z^2 \quad (0 < z \leq \frac{1}{2}),$$

and by considering the above-mentioned “tacit presuppositions,” that

$$|R(\varphi)| < nL\delta^4 \quad (L = \frac{a^4}{2}A^4). \quad (2.32)$$

In the integral p , Dirichlet now made the substitution of variables $\psi = \sqrt{n}\varphi$. The upper bound $\delta\sqrt{n}$ of the domain of integration of the new integral became equal to ∞ as $n \rightarrow \infty$ because of (2.28). Thus, for a “large” number of observations the relation

$$p \approx \frac{2}{\pi} \int_0^\infty \frac{\sin(\lambda\psi)}{\psi} e^{-\psi^2 \frac{\sum k_v \alpha_v^2}{n}} d\psi$$

followed. This relation can be rigorously deduced from the inequalities (which were not explicitly stated by Dirichlet):

$$\left| \int_0^{\delta\sqrt{n}} \frac{\sin(\lambda\psi)}{\psi} e^{-\psi^2 \frac{\sum k_v \alpha_v^2}{n}} d\psi - \int_0^{\delta\sqrt{n}} \frac{\sin(\lambda\psi)}{\psi} e^{-\psi^2 \frac{\sum k_v \alpha_v^2}{n}} e^{R(\psi/\sqrt{n})} d\psi \right| \\ < \max(e^{nL\delta^4} - 1; 1 - e^{-nL\delta^4}) \lambda \int_0^\infty e^{-k\alpha^2\psi^2} d\psi =: \lambda C_1(n)$$

(based on (2.32)) and

$$\left| \int_{\delta\sqrt{n}}^\infty \frac{\sin(\lambda\psi)}{\psi} e^{-\psi^2 \frac{\sum k_v \alpha_v^2}{n}} d\psi \right| \leq \frac{1}{\delta\sqrt{n}} \int_0^\infty e^{-\psi^2 k\alpha^2} d\psi =: C_2(n).$$

From (2.28) one sees that the right sides of these inequalities tend to 0 as $n \rightarrow \infty$.

Finally, Dirichlet [1846, 30] concluded by use of well-known integral formulae that

$$p \approx \frac{2}{\sqrt{\pi}} \int_0^{\lambda/r} e^{-s^2} ds.$$

The justification that q_1 and q_2 tend to 0 as n increases, is only hinted at in the lecture notes [Dirichlet 1846, 30 f.], and seems to go as follows: $g_v(\varphi) = \int_{-a}^a f_v(x) \cos(\alpha_v \varphi x) dx$ is strictly monotonic decreasing in the interval—dependent on v (!)— $[0; \varepsilon_v]$, and thus $g_v(\varphi_0) > |g_v(\varphi)|$ for all $\varphi_0 \in [0; \varepsilon_v]$ and all $\varphi > \varphi_0$.²⁸ From this, Dirichlet concluded (loosely translated) that there must exist a $\delta_n > 0$ such that for all $\varphi_0 \in [0; \delta_n]$ and all $\varphi > \varphi_0$ also $\Pi(\varphi_0) > |\Pi(\varphi)|$ holds. Apparently, the possible dependence of the ε_v on v , and thus of the δ_n on n , was not taken into consideration, and it was supposed that, for a sufficiently large n , the δ according to (2.28) would be smaller than δ_n , and therefore

²⁸ We have $g'_v(\varphi) < 0$ in a neighborhood of $\varphi = 0$ and $|g_v(\varphi)| < 1$ for $\varphi > 0$. Moreover, $g_v(\varphi) \rightarrow 0$ for $\varphi \rightarrow \infty$, as one can deduce after partial integration, see below. Finally, the asserted behavior of g_v follows from its continuity with respect to all $\varphi \geq 0$.

$$|\Pi(\varphi)| < \Pi(\delta) \quad \forall \varphi > \delta \quad (2.33)$$

would apply. However, because δ_n might tend to 0 even faster than δ as $n \rightarrow \infty$, (2.33) only holds if a certain uniformity in the shape of the factors $g_v(\varphi)$ of $\Pi(\varphi)$ as functions of φ is presupposed. (Actually, this can be deduced from the “tacit assumptions,” though, as it seems, only by methods which were not known to Dirichlet, see [Fischer 2000, Sect. 2.3.1].) From (2.33) one gets for sufficiently large n

$$|q_1| < \int_{\delta}^{\Delta} \left| \frac{\sin(\lambda \sqrt{n} \varphi)}{\varphi} \Pi(\varphi) d\varphi \right| < \lambda \sqrt{n} \Delta \Pi(\delta).$$

By definition $\Pi(\delta) = e^{-\sum k_v \alpha_v^2 \delta^2} e^{R(\delta)}$ and therefore, using the “tacit assumptions”:

$$|q_1| < \lambda \sqrt{n} \Delta e^{-n k \alpha^2 \delta^2} e^{R(\delta)} =: \lambda C_3(n).$$

If one sets $\delta = n^{-\frac{1}{3}}$, as suggested by Dirichlet [1846, 29] as an example of a possible δ in accordance with (2.28), the right side tends to 0 as n increases.

In order to justify that

$$q_2 = \frac{2}{\pi} \int_{\Delta}^{\infty} \frac{\sin(\lambda \sqrt{n} \varphi)}{\varphi} \Pi(\varphi) d\varphi$$

can also be neglected for “infinite” n , Dirichlet used the relation

$$\int_{-a}^a \cos(\alpha_v \varphi x) f_v(x) dx = \frac{2 f_v(a) \sin(\alpha_v \varphi a)}{\alpha_v \varphi} - \int_{-a}^a \frac{\sin(a \varphi x)}{\alpha_v \varphi} f_v'(x) dx,$$

which can be derived by partial integration. (For the existence of continuous derivatives of the densities see the “tacit assumptions.”) From that, Dirichlet concluded that $|\Pi(\varphi)|$ must be smaller than $\left(\frac{c}{\varphi}\right)^n$ with a constant c independent of n , which is only true under the “tacit assumptions.” Dirichlet's reasoning can be completed as follows: From the estimate $|\Pi(\varphi)| < \left(\frac{c}{\varphi}\right)^n$ one gets

$$|q_2| < \int_{\Delta}^{\infty} \frac{1}{\varphi} \left(\frac{c}{\varphi}\right)^n d\varphi = \left(\frac{c}{\Delta}\right)^n \frac{1}{n} =: C_4(n).$$

From the hypothesis (2.29) on the growth of Δ , the latter term tends to 0 as $n \rightarrow \infty$.

On the basis of the inequalities stated above, we can reconstruct Dirichlet's result by the inequality

$$\left| P \left(-\lambda \sqrt{n} \leq \sum_{v=1}^n \alpha_v x_v \leq \lambda \sqrt{n} \right) - \frac{2}{\sqrt{\pi}} \int_0^{\lambda/\sqrt{n}} e^{-s^2} ds \right| \leq \lambda C_1(n) + C_2(n) + \lambda C_3(n) + C_4(n),$$

which is valid for sufficiently large n . Presupposing

$$\delta = n^{-1/2+\varepsilon}, \quad 0 < \varepsilon < \frac{1}{4},$$

the bounds C_1, C_2, C_3, C_4 have the respective asymptotic orders

$$C_1(n) = O(n^{-1+4\varepsilon}), \quad C_2(n) = O(n^{-\varepsilon}), \quad C_3(n) = o(n^{-\rho}), \quad C_4(n) = o(n^{-\rho}),$$

where ρ is an arbitrary positive constant. From this, we can see that Dirichlet's method gives an estimate for the error of approximation that is far from the optimal one as developed by modern methods. It was, however, not Dirichlet's intention at all to find a "very good" approximation error for the normal distribution. Apparently, he wanted to show that his modification of the Laplacian method of approximation could also be applied to the problem of probabilities of linear combinations of random errors. In this sense, the central CLT for Dirichlet served chiefly as an illustration of special analytical techniques and was less a problem which he treated in its own right.

2.5 Cauchy's Bound for the Error of Approximation

Augustin Louis Cauchy (1789–1857) provided fundamental contributions to a great number of mathematical subjects and essentially determined the development of mathematics during the 19th century. On probability theory in the narrow sense, Cauchy only published a few papers, in 1853, printed in the *Comptes rendus*, which referred to his dispute with Irénée Jules Bienaymé (1796–1878) over the Laplacian foundation of the method of least squares. In this scientific controversy, which occurred during the months of June, July, and August in the summer of 1853 at the Paris Academy, Bienaymé defended the Laplacian error theory, whose basic ideas were repeatedly criticized by Cauchy.²⁹ Cauchy's last article in a total of eight papers contains an interesting discussion of the approximate normal distribution of linear combinations of random errors. Basically, his line of analytical argumentation is similar to Dirichlet's and employs methods which are still being used in the modern treatment of the CLT. His (rather narrow) conditions are in essence the same as Dirichlet's.

2.5.1 The Cauchy–Bienaymé Dispute

From a historian's point of view, Cauchy's and Bienaymé's interest in treating stochastic problems in an almost purely mathematical manner, indicating a shift from classical toward mathematical probability, is especially important. However,

²⁹ For more details on this dispute see [Heyde & Seneta 1977] and [Fischer 2000, 76–97]. Bienaymé's contributions are, as listed in the Bibliography, [Bienaymé 1853a] to [Bienaymé 1853e], Cauchy's contributions are [Cauchy 1853a] to [Cauchy 1853h].

Cauchy's position of only accepting arguments within mathematics for a discussion of the error theoretic foundations (which became more and more adamant during the controversy), met with Bienaymé's opposition, who still demanded the critical "good sense" assessment of those problems.

The political and private connections of both opponents might have been especially important for the background of their scientific quarrel. As a consequence of the revolution of July 1830, which brought Louis-Philippe, the "king of the people," to power, Cauchy, being a supporter of the overthrown Charles X Bourbon, had to give up his positions in higher education and go into exile.³⁰ From 1833 to 1838 he was in charge of the education of Charles's eldest son in Prague. After the completion of his duties there, he went back to Paris and resumed work at the Academy. After the revolution of February 1848, which, for a brief period, reestablished the republic, he was able to return to teaching at the university. With the seizure of power by Napoleon III in 1851, Cauchy's official position remained unchanged. As a supporter of the house of Bourbon, however, he did not look on this political change especially enthused.

In 1820 Bienaymé³¹ set out on a brilliant career in government finance which remained entirely unscathed by the 1830 revolution. Whereas the revolution of 1848 had brought some advantages to Cauchy, Bienaymé had to resign from his positions. Consequently he delved into more scientific endeavors. Bienaymé, in contrast to Cauchy, sympathized with Napoleon III, and after his seizure of power regained a certain influence on the country's financial politics.

Apart from differences in their political views, Cauchy and Bienaymé seem to have had personal misgivings as well. As suggested by [Heyde & Seneta 1977, 13], these could have originated for one thing from different religious beliefs—Cauchy was a fanatic Catholic, and Bienaymé tended toward agnosticism. Further, Bienaymé cultivated a close friendship with Antoine Auguste Cournot,³² who was very influential in science back then, while Cournot and Cauchy were bitter enemies.

Bienaymé presented his essay on foundational problems of least squares (see Sect. 2.1.5.2) in 1852 at the Paris Academy. His good reception there contributed significantly to his election as an ordinary member of the Academy soon thereafter. It is only natural that Bienaymé would have been very interested in contributing to discussions on "his field," error calculus, at Academy conventions. He found a suitable opportunity when Cauchy once again presented his method of interpolation (introduced already in 1835); Cauchy suggested that this method be applied instead of least squares even in those cases which had not yet been taken into consideration when his procedure of interpolation was introduced.

In presenting his method in 1835, Cauchy began with the following problem: He assumed that a function $y(x)$ could be expanded into a convergent series of the form

³⁰ For biographical details on Cauchy see [Belhoste 1991].

³¹ For biographical details see [Heyde & Seneta 1977].

³² Regarding probability theory, Cournot became especially prominent by his elementary treatise [Cournot 1843], in which a clear distinction was made between the subjective and the objective notion of probability.

$$y(x) = au(x) + bv(x) + cw(x) + \dots$$

with given functions $u(x)$, $v(x)$, $w(x)$, \dots , but unknown coefficients a, b, c, \dots . Assigned to the given abscissae x_1, x_2, \dots, x_n were observed function values y_1, y_2, \dots, y_n , which were, however, subject to the observation errors $\epsilon_1, \epsilon_2, \dots, \epsilon_n$. Cauchy searched for a method of “interpolation” with which one could jointly 1) assess, with regard to the order of magnitude of the observational errors, how many series terms had to be calculated to obtain a sufficiently exact approximation of the true function value for each arbitrary x , and 2) calculate those series terms in an easy way. Cauchy [1835/37, 8–16] presented a procedure by which the coefficients a, b, c, \dots could be approximated by a method that allowed one to calculate the coefficients with a simple correction from the ones already determined, if the number of the coefficients was increased by 1. From the error theoretic point of view, Cauchy’s reasoning was based on the idea of minimizing the maximal possible error in each single stage of his procedure.

Cauchy’s method of interpolation can be considered as a procedure for determining compromise solutions $\bar{a}_1, \bar{a}_2, \dots$ of the overdetermined system

$$y_i = a_1 u_{i1} + a_2 u_{i2} + \dots + a_r u_{ir} + \dots \quad (i = 1, \dots, n)$$

with the given u_{ih} (according to the function values $u(x_i), v(x_i), \dots$) and y_i (the observations afflicted by errors), where, however, the number $r < n$ of the \bar{a}_i needed is not known at the beginning. Yet it was obvious that Cauchy’s procedure could also be applied to the case of overdetermined systems of linear equations with a fixed number r of variables.

Around 1840, Cauchy began to show increased interest in astronomy and especially in perturbation theory. Belhoste [1991, 205 f.] sees a connection with Cauchy’s election to the “Bureau des Longitudes” in 1839, which had to be revoked because, being a royalist, Cauchy had refused to show any kind of allegiance to the “king of the people” Louis-Philippe. The works of astronomers Hervé Faye, Urbain Jean Joseph Leverrier (whose investigations in perturbation theory led to the discovery of Neptune in 1846), and Antoine François Yvon-Villarceau were influenced by Cauchy, and in turn stimulated some contributions by him. The problem of comparing observations and results obtained by perturbation theory kept Cauchy busy for most of the second half of 1847, when he issued a series of papers, and led him back to his own method of interpolation. Now, he [1847a] wanted to see this method also applied to overdetermined systems of linear equations with an a priori fixed number of unknowns. One can assume that this problem was being repeatedly discussed by the astronomy-prone members of the Academy. Cauchy [1847b] referred to a paper published by Villarceau in 1845 (this paper was not further specified) because approximation methods had apparently been used in it, analogous to his method of interpolation. Around 1849, Villarceau used Cauchy’s method in extensive calculations of approximations of various orbit parameters [Heyde & Seneta 1977, 74]. Cauchy [1853a, 36] quoted a remark made by Faye on the usefulness of his interpolation procedure (the corresponding paper of Faye’s

cannot be bibliographically determined). So, when declaring himself to be partial to the method of least squares and against the method of interpolation, Bienaymé met not only with opposition from Cauchy, but from a whole group of astronomers.

2.5.2 Cauchy's Exceptional Laws of Error

Cauchy's initial line of argument was to minimize the maximum possible errors of approximation. Thus, he used a typical interpolation justification, which practically did not touch probability at all. Bienaymé [1853a, 5; 10], on the other hand, criticized this lack of probabilistic argumentation: Errors of observation are subject to chance. Thus, in order to fit the parameters to the observations, those methods should be preferred that can be analyzed and justified by stochastic considerations. In this way, Bienaymé emphasized the universal claim of classical probability being responsible for all fields in which complete knowledge of causes and laws could not be obtained. In response to this criticism, Cauchy began his probabilistic research. According to Schneider [1987a, 200 f.], Cauchy did not disapprove probability completely, but was only willing to accept probabilistic results which could be justified within mathematics. For Cauchy, the usual reasoning of classical probability, based on the unity of good sense and mathematics, had become obsolete. In the case of error theory, Laplace had claimed that the method of least squares should be preferred "in any case." Now, Cauchy set out to ridicule this claim by using Laplace's (and Bienaymé's) own probabilistic methods, although from a strictly mathematical point of view.

Like Laplace, Cauchy considered the system of n "approximative" equations

$$a_j x + b_j y + \cdots + g_j v + h_j w = k_j \quad (j = 1, \dots, n)$$

with m "unknowns" x, y, \dots, v, w and n observed values k_1, k_2, \dots, k_n . Cauchy approximated the "unknown" x by $\bar{x} = \sum_{j=1}^n \lambda_j k_j$, where the multipliers $\lambda_1, \dots, \lambda_n$ had the additional property

$$\sum_{j=1}^n \lambda_j a_j = 1, \sum_{j=1}^n \lambda_j b_j = 0, \dots, \sum_{j=1}^n \lambda_j h_j = 0. \quad (2.34)$$

From the "exact" equations

$$a_j x + b_j y + \cdots + g_j v + h_j w = k_j - \epsilon_j \quad (j = 1, \dots, n),$$

where the ϵ_j represent the observational errors, it followed that

$$\sum_{j=1}^n \lambda_j a_j x + \sum_{j=1}^n \lambda_j b_j y + \cdots + \sum_{j=1}^n \lambda_j g_j v + \sum_{j=1}^n \lambda_j h_j w = \sum_{j=1}^n \lambda_j k_j - \sum_{j=1}^n \lambda_j \epsilon_j.$$

On account of (2.34) the estimate \bar{x} was distorted by the “error” $\bar{x} - x = \sum_{j=1}^n \lambda_j \epsilon_j$. Cauchy restricted his discussion to the determination of \bar{x} as being representative of all of the other variables. For the errors ϵ_j he presupposed a common symmetrical density $f(x)$, concentrated on the interval $[-\kappa; \kappa]$ with $\kappa \leq \infty$. For those densities Cauchy coined the term “indice de probabilité.” Taking up the Laplacian characterization of the “most advantageous value,” he demanded that

$$p = P(|x - \bar{x}| \leq \nu) = P\left(\left|\sum_{j=1}^n \lambda_j \epsilon_j\right| \leq \nu\right) = \max \quad (2.35)$$

for all $\nu > 0$.

In his discussion of this condition, Cauchy made systematic use of the (now so-called) characteristic function, which he named “fonction auxiliaire.” If $g(x)$ was the “indice de probabilité” of an error with values within $[\kappa_1; \kappa_2]$, then the “fonction auxiliaire” related to it was defined by³³

$$\varphi(x) = \int_{\kappa_1}^{\kappa_2} e^{-izx} g(z) dz \quad (i = \sqrt{-1}).$$

Repeating arguments of his proof [1818; 1827, note VI] of the Fourier inversion formula,³⁴ he [1853d; 1853e] showed that for symmetrical densities f , defined as above, and their characteristic functions

$$\varphi(x) = 2 \int_0^{\kappa} f(z) \cos(xz) dz$$

³³ The designation “indice de probabilité” is used, for example, in [Cauchy 1853f, 106], the designation “fonction auxiliaire” in [Cauchy 1853h, 125]. In a slightly different form compared with Cauchy’s use, in modern probability theory the characteristic function of a random variable X is defined by $Ee^{+iX\theta}$ instead of $Ee^{-iX\theta}$. For symmetrically distributed random variables with zero means (which case was predominantly considered by Cauchy) both terms coincide.

³⁴ Fourier, Poisson, and Cauchy around 1820 (more or less independently) published very similar versions of the inversion formula [Laugwitz 1990, 30–34]. An early form, which remained, however, unpublished, had been presented by Fourier already in 1807 [Grattan-Guinness & Ravetz 1972]. The complex version of the formula

$$f(x) = \lim_{c \rightarrow \infty} \frac{1}{2\pi} \int_{-c}^c \int_{-\infty}^{\infty} f(t) e^{iu(t-x)} dt du$$

for functions $f(x)$ continuous in x (precise properties of those functions were not explained for the time being) is essentially due to Cauchy. In the collection of Gauss’s private papers (“Nachlass”) an unpublished note (written presumably before 1813) on a complex version, with title “Schönes Theorem der Wahrscheinlichkeitsrechnung” (“beautiful theorem of probability calculus”) [Gauss 1900, 88 f.] was found as well. From Gauss’s remarks one can see that he derived his formulae on the basis of orthogonality relations like (2.3), by considering Fourier series with periods tending to infinity. One may suppose that Gauss was inspired to these observations by reading Laplace’s *TAP*. For surveys of the history of the Fourier inversion formula during the 19th century, see [Burckhardt 1914, 1085–1097] (up to ca. 1850) and [Pringsheim 1907] (for the time 1850–1900). An outline of the entire development up to ca. 1940 is given by Cooke [2005].

the equation

$$f(x) = \frac{1}{\pi} \int_0^\infty \varphi(z) \cos(xz) dz \quad (2.36)$$

holds. For the characteristic function Φ of the linear combination $\lambda_1 \epsilon_1 + \dots + \lambda_n \epsilon_n$, where each of the mutually independent errors $\epsilon_1, \dots, \epsilon_n$ obeys the law f , Cauchy derived:

$$\Phi(x) = \varphi(\lambda_1 x) \cdots \varphi(\lambda_n x). \quad (2.37)$$

From a modern point of view, Cauchy's proof [1853d, 86] for the latter identity was unnecessarily complicated, as it was not based on the now common conception of characteristic function as expectation. Instead, Cauchy used, in a rather intricate way, the jump function, very similar to Dirichlet's,

$$\frac{1}{2\pi} \int_{-\infty}^\infty \int_a^b e^{\theta(\tau-x)i} d\tau d\theta = \begin{cases} 1 & \text{for } x \in]a; b[\\ 0 & \text{for } x \notin [a; b], \end{cases}$$

which he derived by a (rather formal) use of the Fourier inversion formula.³⁵ Hinting at this jump function, Cauchy [1853e, 96] also stated

$$p = \frac{2}{\pi} \int_0^\infty \frac{\sin(\theta v)}{\theta} \Phi(\theta) d\theta, \quad (2.38)$$

where $\Phi(\theta)$ was defined as above (see (2.37)).

Cauchy [1853e, 98–101] gave a plausible justification that condition (2.35) is met if and only if $\kappa = \infty$ and

$$\varphi(x) = e^{-c|x|^N} \quad (2.39)$$

with positive constants c and N (see [Heyde & Seneta 1977, 82–85]). Cauchy's arguments for the “only if” were not sound.

From (2.37) to (2.39) it followed that

$$p = \frac{2}{\pi} \int_0^\infty e^{-c\theta^N \sum_{j=1}^n |\lambda_j|^N} \frac{\sin(\theta v)}{\theta} d\theta$$

[Cauchy 1853e, 102]. Independent of v , p is maximized if $\sum_{j=1}^n |\lambda_j|^N$, under the constraint (2.34), is minimized. This implies, as Cauchy [1853e, 102 f.] showed, in the case for which a single element x is to be determined ($b_j = \dots = g_j = h_j = 0$), that the condition

$$\lambda_j = \text{sign}(a_j) |a_j|^{\frac{1}{N-1}} \left(\sum_{r=1}^n |a_r|^{\frac{N}{N-1}} \right)^{-1}$$

³⁵ Jump functions played an important role in Cauchy's analytical work. As a means for integration he used this device not until 1853, however. See [Burckhardt 1914, 963; 1320–1324] for a general account on the use of jump functions during the first half of the 19th century.

holds for p being maximal. Only for the case $N = 2$ are the λ_j the least square multipliers. Cauchy did not observe that only for exponents N with $N \leq 2$ the function $\varphi(x)$ in (2.39) was the characteristic function of a probability distribution. On the contrary, he assigned to the case $N = \infty$ an essential importance. As Cauchy [1853e, 103 f.] argued, this case corresponded to his own method of interpolation with multipliers $\lambda_j = \pm 1$.

With the aid of the inversion formula (2.36) Cauchy was able to determine the specific law of error corresponding to the constants c and N in two special cases: For $N = 2$ one gets the Gaussian law of error, and for $N = 1$ one gets the density

$$f(x) = \frac{k}{\pi} \frac{1}{1 + k^2 x^2} \quad \left(k = \frac{1}{2\sqrt{c}} \right).$$

Poisson (see Sect. 2.2.3.1) had already shown that the sum of independent identically distributed random variables with this density does not satisfy the CLT.

The main result of the article [Cauchy 1853e] was the fact that laws of error which are different from the Gaussian error law can lead to systems of multipliers entirely different from the least squares multipliers if Laplace's criterion for the "most advantageous value" is taken as a basis. Thus, from a purely mathematical point of view such as Cauchy's, the method of least squares was not distinguished from other fitting methods, but was in principle only one possible method among many equivalent methods.

Naturally, Cauchy knew that observation errors are bounded. Laplace had shown that linear combinations of identically distributed bounded errors were normally distributed in the asymptotic sense, and, on this basis, one could expect that the method of least squares would produce fitting values rather close to the "optimal" possible fitting values (assuming a large number of observations). But, what assertion concerning the method of least squares had been actually proven by Laplace? As Schneider [1998] has pointed out, it was Laplace's style to avoid formulations that permitted a refutation of his arguments. Phrases like "Preference should (!) thus be given [to the method of least squares]," or, "if we have a very great number of observations," without a closer specification of "how great," could hardly be disproved. For a mathematical refutation of Laplace's assertions, Cauchy had to transform Laplace's application-oriented propositions into mathematical claims. But this thrust Cauchy into the dilemma that the presentation of some impractical counterexamples could hardly compromise Laplace's position, as Bienaymé immediately pointed out. Yet there was still Laplace's deduction of the approximate normal distribution, which no longer met the analytical standards of the mid-19th century, and did not produce an adequate and exact estimate of the deviation of the approximative distribution from the actual one. As we have seen (Sect. 2.1.5.2), the sore point in Laplace's argument was the assumption of a very substantial proximity (strictly speaking even of equality) of both distributions. Making precisely this point to the subject of discussion, Cauchy could argue that Laplace had not examined his approximations with sufficient scrutiny.

Cauchy [1853f] indeed gave a first discussion on the approximate normal distribution for linear combinations of errors, however without exactly discussing the quality of approximation. His account essentially endorsed Laplace's foundation of least squares. Still, Cauchy announced further critical examinations.

In [1853g] he actually presented several "candidates" for the failure of a sufficiently close proximity between approximate and exact distribution. One example referred to bounded errors, however with a density close to the above-mentioned "Cauchy-density" f_k . Another referred to cases in which large deviations concerning the order of magnitude among the least square multipliers λ_j occurred. From the point of view of common practice of observation and measurement, however, both examples seemed to be far-fetched, as Bienaymé would shortly point out.

2.5.3 Bienaymé's Arguments

Bienaymé's reply to Cauchy's arguments is mainly contained in the "Considérations à l'appui de la découverte de Laplace sur la loi de probabilité dans la méthode des moindres carrés" [Bienaymé 1853e]. This article consists of four parts: In the first, one can find a general defense of the principles of Laplacian probability theory. The second part contains a discussion of the importance of the "mean of the squares of the differences of the errors and their mean value" which, in modern terminology, is simply the variance of observational errors. Through this discussion, Bienaymé confirmed his preference for least squares. In the third part, Bienaymé deduced the inequality which is now named after him and Chebyshev, however not aiming at the (now common) discussion of a weak law of large numbers, but for the sake of giving an additional intuitive argument for the superiority of least squares in the case of a large number of observations. Finally, the practical irrelevance of Cauchy's exceptional laws was discussed by Bienaymé in the fourth part.

The first part of Bienaymé's considerations is well described by Heyde & Seneta [1977, 87 f.] and by Schneider [1987a, 208–210]. Here, Bienaymé pointed out the statistical importance of large samples, and, in the same context, the importance of Laplace's CLT. This exposition was connected with the refusal of small samples because of their insignificance, at least implicitly.

This refusal was discussed in more detail in the second part. Bienaymé gave a plausibility consideration in order to show that for linear combinations $\sum_{j=1}^n h_j \epsilon_j$ of independent identically distributed observational errors (each with only a finite number of possible values) the asymptotic relation

$$\sqrt[m]{E \left(\sum (h_j \epsilon_j - E h_j \epsilon_j) \right)^{2m}} \approx \text{const.} \sum E (h_j \epsilon_j - E h_j \epsilon_j)^2 \quad (2.40)$$

is valid for each natural m (the constant "const." depending on m). In this context, Bienaymé criticized Gauss's remark [Gauss 1823, 6 f.] that the variance was not distinguished as a precision measure from other central moments of even order. Gauss had made this statement in the context of an arbitrary number of observational

errors. Bienaymé, however, was apparently convinced that only the case of “large numbers” was worthy of consideration. Because in this case all central moments of even order of the deviation $\sum h_j \epsilon_j$ between the true and the estimated value could be reduced to the variance $\text{Var} \epsilon_1 \sum h_j^2$ by virtue of (2.40), he maintained that “nothing is simpler, than to recognize that one has to render the sum of the squares of the factors h_j a minimum” [Bienaymé 1853e, 319].

Bienaymé’s arguments in the second part were complemented by a discussion of Laplace’s criterion (2.9) for the “most advantageous value.” Bienaymé, applying a rather simple procedure (equivalent to the modern textbook proof of the Bienaymé–Chebyshev inequality), calculated the “form” of the probability of the deviation between the true and the estimated value in the case of identically distributed observational errors with zero means and the common variance σ^2 :

$$P\left(\left|\sum h_j \epsilon_j\right| \leq t\sqrt{2\sigma^2}\right) = 1 - \frac{\theta f}{2t^2} \sum h_j^2,$$

where θ and f are positive “constants” less than 1, depending on the error law and the factors h_j . From this estimation, Bienaymé plausibly argued that Laplace’s criterion is met if $\sum h_j^2$ becomes a minimum, which condition leads to the method of least squares. For a more exact discussion, however, Bienaymé, somewhat maliciously, referred to the article [Cauchy 1853f], in which a first reexamination of Laplace’s normal approximation was given (still without suitable limits for the approximation error).

For a discussion of Cauchy’s exceptional laws, Bienaymé confined himself to the examination of the now so-called “Cauchy distribution” with the density

$$f_k(\epsilon) = \frac{k}{\pi} \frac{1}{1 + k^2 \epsilon^2}.$$

This restriction was probably due to the fact that this density was the only one which could be given explicitly by an algebraic formula. However, it also seems that Bienaymé treated this density as representative of all exceptional laws. He argued first, with the aid of a table of $\int_{-a}^a f_1(x) dx$ for several values a , that, presupposing this error law, the probabilities of very large values were so high that no reasonable person would use a corresponding observation instrument. Second, Bienaymé advanced the argument that in the case of direct observations the probability of a certain deviation between the true value and the arithmetic mean would not depend on the number of observations, in contrast to all experiences of observational practice.³⁶ Bienaymé

³⁶ Bienaymé [1853e, 323] only noticed that it would be “very easy” to show this. The probably easiest way is the following: Let $y_j = x + \epsilon_j$ ($j = 1, \dots, n$), and let $\varphi(x) = e^{-|x|}$ be the characteristic function of each single error ϵ_j . Then the characteristic function $\Phi(z)$ of the difference $\frac{\sum \epsilon_j}{n}$ between the arithmetic mean $\frac{\sum y_j}{n}$ and the real value x is

$$\Phi(z) = \left(\varphi\left(\frac{z}{n}\right)\right)^n = \left(e^{-\frac{|z|}{n}}\right)^n = \varphi(z).$$

did not fail to remark that Poisson had already realized—in contrast to Cauchy, as it seemed—the practical irrelevance of the error laws f_k .

Bienaymé [1853e, 324] also discussed Cauchy's example of multipliers which considerably deviate in their respective orders of magnitude. He emphasized that such cases were far from any “well-planned and careful” application of the method of least squares.

Bienaymé's comments constitute a mix of purely mathematical arguments and reflection upon these arguments within the framework of the practice of observation. If Cauchy tried to transpose Laplace's statements into purely mathematical claims, then Bienaymé conversely tried to transform Cauchy's mathematical considerations into concrete situations of observation. In doing this, both mathematicians executed a separation of mathematics and its applications which had remained foreign to Laplace's classical probability. Bienaymé, however, did not share Cauchy's attitude of attributing the same value to any stochastic model which could be mathematically derived, but instead insisted on an assessment of any implication by “good sense.”

2.5.4 Cauchy's Version of the Central Limit Theorem

In his last contribution to the scientific discussion with Bienaymé on least squares, Cauchy [1853h] established explicit upper bounds for the error of a normal approximation to the distribution of a linear combination $\sum_{j=1}^n \lambda_j \epsilon_j$ of identically distributed independent errors ϵ_j with a symmetric density f vanishing for arguments beyond the compact interval $[-\kappa; \kappa]$. He additionally required that the λ_j should have “the order of magnitude” of $\frac{1}{n}$ or less, and that $\sum \lambda_j^2 =: \Lambda$ should be of the order $\frac{1}{n}$. For a precise formulation of the first requirement, we have to assume that there exist positive constants α and β independent of n , such that for all $j = 1, \dots, n$ there is a $\gamma(j) \geq 1$ with

$$\alpha \leq n^{\gamma(j)} |\lambda_j| \leq \beta. \quad (2.41)$$

Cauchy [1853h] only gave a sketch of proof (some details are discussed in the next section) that, for $v > 0$ with the notation $c := \int_0^\kappa x^2 f(x) dx$,

$$\left| P \left(-v \leq \sum_{j=1}^n \lambda_j \epsilon_j \leq v \right) - \frac{2}{\sqrt{\pi}} \int_0^{\frac{v}{2\sqrt{c\Lambda}}} e^{-\theta^2} d\theta \right| \leq C_1(n) + C_2(n, v) + C_3(n), \quad (2.42)$$

but at least he made the formulae for the bounds C_1 , C_2 , and C_3 explicit, which are valid for sufficiently large n and which tend to 0 as n increases.

In his doctoral thesis, Ivan Vladislavovich Sleshinskii [1892] gave detailed deductions for the constants C_1 , C_2 , and C_3 , and he corrected apparent misprints in Cauchy's formulae (see [Heyde & Seneta 1977, 94–96] and [Seneta 1984, 48–50]), with the following result: Let $\Theta = n^{\frac{1}{2}+\delta}$ ($0 < \delta < \frac{1}{4}$); then

$$C_1 = \frac{1}{\pi \mathcal{N}} e^{-\mathcal{N}}, \text{ with } \mathcal{N} = \frac{1}{2} \frac{r \Lambda \Theta^2}{1 + r \lambda^2 \Theta^2}, \quad (2.43)$$

$$C_2(n, v) = \frac{2h\sqrt{3}}{\pi} \log \left(\frac{\Theta v}{\sqrt{3}} + \sqrt{1 + \frac{\Theta^2 v^2}{3}} \right), \quad (2.44)$$

where

$$\lambda := \max(|\lambda_1|, \dots, |\lambda_n|); \quad h := \max \left(e^{\frac{1}{4} c \Lambda \lambda^2 \Theta^4 \kappa^2} - 1, 1 - e^{-\frac{c^2 \Lambda \lambda^2 \Theta^4}{1 - c \lambda^2 \Theta^2}} \right),$$

and

$$C_3(n) = \frac{e^{-c \Lambda \Theta^2}}{\pi c \Lambda \Theta^2}. \quad (2.45)$$

There are minor differences with regard to C_2 and C_3 between Cauchy's original formulae and Sleshinskii's.

The quantity C_2 has the minor flaw that it is not independent of v ; it grows for fixed n together with v . However, presupposing (2.41), and considering that $|\epsilon_j| \leq \kappa$, one can deduce

$$P(|\sum \lambda_j \epsilon_j| \leq v) = 1 \quad \text{if } v \geq \beta \kappa. \quad (2.46)$$

Since C_2 is monotonically increasing as a function of v , one gets

$$\left| P \left(-v \leq \sum_{j=1}^n \lambda_j \epsilon_j \leq v \right) - \frac{2}{\sqrt{\pi}} \int_0^{\frac{v}{2\sqrt{c\Lambda}}} e^{-\theta^2} d\theta \right| \leq C_1(n) + C_2(n, \beta \kappa) + C_3(n)$$

for $v \leq \beta \kappa$. On the other hand, for $v > \beta \kappa$, (2.46) yields

$$\left| P \left(-v \leq \sum_{j=1}^n \lambda_j \epsilon_j \leq v \right) - \frac{2}{\sqrt{\pi}} \int_0^{\frac{v}{2\sqrt{c\Lambda}}} e^{-\theta^2} d\theta \right| \leq \frac{2}{\sqrt{\pi}} \int_{\frac{\beta \kappa}{2\sqrt{c\Lambda}}}^{\infty} e^{-\theta^2} d\theta =: C_4(n).$$

Now, because Λ must be of the order of magnitude $\frac{1}{n}$, $C_4(n)$ tends to 0 independent of v . Altogether, it follows that for any $v \in \mathbb{R}^+$,

$$\left| P \left(-v \leq \sum_{j=1}^n \lambda_j \epsilon_j \leq v \right) - \frac{2}{\sqrt{\pi}} \int_0^{\frac{v}{2\sqrt{c\Lambda}}} e^{-\theta^2} d\theta \right| \leq C_1(n) + C_2(n, \beta \kappa) + C_3(n) + C_4(n),$$

where the right side tends to 0 independent of v .

Apparently, the convergence of C_2 to 0 as $n \rightarrow \infty$ (v fixed) is the slowest among all of the "constants," because $C_2 = O\left(\frac{\log n}{n^{1-4\delta}}\right)$. Thus, the order of magnitude of Cauchy's upper bounds was rather close to the optimal asymptotic order, which is, in the case at hand, and according to Harald Cramér [1928], equal to $O(\frac{1}{n})$.

From today's point of view Cauchy's account can be interpreted as the more or less rigorous proof of the finite version of a CLT for linear combinations of independent identically distributed random variables. In fact, a "modern" CLT can be inferred from Cauchy's version by considering a sequence of independent random variables X_j , distributed like Cauchy's observational errors, and by setting $\lambda_j = \frac{1}{n}$, $v = \frac{a}{\sqrt{n}}$ ($a > 0$), $c = \frac{1}{2}\text{Var}X_1$. Then, by virtue of (2.42),

$$\left| P\left(-a\sqrt{n} \leq \sum_{j=1}^n X_j \leq a\sqrt{n}\right) - \frac{2}{\sqrt{\pi}} \int_0^{\frac{a}{2\sqrt{c}}} e^{-x^2} dx \right| \leq C_1(n) + C_2(n, \frac{a}{\sqrt{n}}) + C_3(n) \rightarrow 0 \quad (n \rightarrow \infty).$$

Though Sleshinskii gave more precise explanations in comparison to Cauchy, he did not substantially go beyond the latter's ideas, and, in particular, he did not succeed in weakening Cauchy's still rather restrictive assumptions. Like Cauchy, Sleshinskii was primarily interested in solving an—although quite abstract—problem of error theory. Therefore, we may actually follow Freudenthal [1970–76, 142] in championing Cauchy for the "first rigorous proof" of the CLT, we must not forget, however, that his goals were quite different from those of modern probability theory.

2.5.5 Cauchy's Idea of Proof

There was a rule that only brief articles were accepted for publication in the *Comptes rendus*, and thus, Cauchy [1853h] had to restrict his presentation to a description of the major steps of his reasoning. The basic ideas, however, can be clearly discerned from his account. In particular, the deduction of (2.42) was based on Cauchy's use of characteristic functions and his modification of the Laplacian method of approximation, which he had already dealt with in several articles published in the 1840s [Cauchy 1844; 1845; 1849]. In [1849, 138–140], for example, Cauchy discussed the asymptotic behavior (as $n \rightarrow \infty$) of the integral

$$S = \frac{1}{2\pi} \int_{-\pi}^{\pi} \left(Z(r_0 e^{(p_0+\varphi)\sqrt{-1}}) \right)^n d\varphi,$$

where $Z(z)$ is an analytic function whose derivative has the property $Z'(r_0 e^{p_0\sqrt{-1}}) = 0$, and whose modulus $|Z(r_0 e^{(p_0+\varphi)\sqrt{-1}})|$ attains its maximum at $\varphi = 0$. Cauchy split the entire integral S into a "major" part with a domain of integration close to the maximum and, in comparison with this part, very small remaining parts which vanish as $n \rightarrow \infty$. Thus, he developed asymptotic methods similar to those of Dirichlet, who, on the other hand, had not published his contributions.

Proving his version of the CLT, Cauchy [1853h, 125 f.] first summarized the most important properties of the "fonction auxiliaire" $\varphi(\theta) = 2 \int_0^K f(\epsilon) \cos(\theta\epsilon) d\epsilon$

for the error law $f(\epsilon)$, and in this context he repeated the fundamental relation

$$P\left(-v \leq \sum_{j=1}^n \lambda_j \epsilon_j \leq v\right) = \frac{2}{\pi} \int_0^\infty \Phi(\theta) \frac{\sin(\theta v)}{\theta} d\theta, \quad (2.47)$$

$$\Phi(\theta) = \varphi(\lambda_1 \theta) \cdots \varphi(\lambda_n \theta).$$

Basically resuming his approach of [1853f], Cauchy from $\varphi(0) = 1$ and $|\varphi(\theta)| < 1$ for $\theta > 0$ concluded that

$$[\varphi(\theta)]^2 = \frac{1}{1 + \rho(\theta)\theta^2}$$

($\rho(\theta) > 0$ for $\theta > 0$). He briefly justified that $\rho(\theta)$ has a positive lower bound r , such that

$$[\varphi(\theta)]^2 \leq \frac{1}{1 + r\theta^2}. \quad (2.48)$$

For this justification he needed the estimate “ $\rho(\infty) \geq \left[\frac{1}{2f(\kappa)}\right]^2$,” which was, as we can see from a similar consideration in [1853f, 107], most likely obtained by partial integration under the tacit presupposition that f possessed a continuous derivative.³⁷

Finally, he [1853h, 126] referred to a consideration in [Cauchy 1853f, 107 f.] (based on the mean value theorem of differential calculus as applied to $\sin(z)$ and $\log(1 - z)$) that for sufficiently small $\theta > 0$:

$$\varphi(\theta) = 1 - \int_0^\kappa \left(2 \sin \frac{\theta \epsilon}{2}\right)^2 f(\epsilon) d\epsilon = e^{-\sigma \theta^2}$$

with

$$1 - \left(\frac{\theta \kappa}{2}\right)^2 < \frac{\sigma}{c} < \frac{1}{1 - c\theta^2} \quad (c = \int_0^\kappa x^2 f(x) dx).$$

By virtue of these estimates, Cauchy's further proceeding [1853h, 127–129] corresponded to his above-mentioned modification of the Laplacian method of approximation as applied to the integral (2.47). The integrand $\Phi(\theta) \frac{\sin(\theta v)}{\theta}$ of this integral attains its absolute maximum at $\theta = 0$. For Θ of an “order greater than \sqrt{n} but smaller than $n^{\frac{3}{4}}$ ” (e.g., $\Theta = n^{\frac{1}{2} + \delta}$, $0 < \delta < \frac{1}{4}$), and for sufficiently large n , Cauchy established the following inequalities for the grade of accuracy regarding the approximation of the integrand by a bell-shaped function in the neighborhood of $\theta = 0$:

$$\left| \frac{2}{\pi} \int_0^\Theta \Phi(\theta) \frac{\sin(\theta v)}{\theta} d\theta - \frac{2}{\pi} \int_0^\Theta e^{-c \Lambda \theta^2} \frac{\sin(\theta v)}{\theta} d\theta \right| < C_2(n, v),$$

³⁷ By partial integration one gets $\varphi(\theta) = 2 \frac{f(\kappa) \sin(\theta \kappa) - \int_0^\kappa f'(x) \sin(\theta x) dx}{\theta}$. If we set $\eta(\theta) := |\int_0^\kappa f'(x) \sin(\theta x) dx|$ the inequality $\rho(\theta) > \frac{1}{4(f(\kappa) + \eta(\theta))^2} - \frac{1}{\theta^2}$ ensues. By taking into account the relation $\lim_{\theta \rightarrow \infty} \eta(\theta) = 0$ (which for Cauchy most probably was a matter of course) the asserted estimate can be followed.

and

$$\left| \frac{2}{\pi} \int_0^{\frac{\nu}{2\sqrt{c\Lambda}}} e^{-\theta^2} d\theta - \frac{2}{\pi} \int_0^{\Theta} e^{-c\Lambda\theta^2} \frac{\sin(\theta\nu)}{\theta} d\theta \right| < C_3(n).$$

In order to estimate the “tail,” Cauchy derived

$$\left| \frac{2}{\pi} \int_{\Theta}^{\infty} \Phi(\theta) \frac{\sin(\theta\nu)}{\theta} d\theta \right| < C_1(n).$$

The constants C_1, C_2, C_3 are already quoted in (2.43), (2.44), (2.45), respectively.

2.5.6 The End of the Controversy

Cauchy [1853h, 130] wrote that for “very large values” of n (the total number of errors) there would be “une grande approximation” between exact and approximate probability. He stated:

The various formulae that we have just written down also permit us to assess, by reducing them to their true significance, the advantages of the employment of the one or the other system of factors, and consequently of the one or the other method.

Cauchy’s “formulae,” in particular those concerning the upper bounds C_1, C_2, C_3 , were indeed appropriate, at least in cases of “large numbers” of observations, for confirming the closeness of the actual distribution of a linear combination of errors to the corresponding normal distribution, and therefore for confirming the superiority of the method of least squares. One could rarely use them for a rejection of least squares, however.

At the end of his article, Cauchy announced that he would return to the issue, but he did never resume his probabilistic studies. There does not exist any explicit evidence as to why he did not continue his discussion of the method of least squares. Beginning with Sleshinskii [1892], the common opinion has been established that Cauchy had come so close to Laplace’s (and Bienaymé’s) position with his asymptotic result that a continuation of the dispute did not appear advisable (see [Heyde & Seneta 1977, 96; Stigler 1974/1999]).

A closer examination, however, shows that Cauchy’s result was not even properly suited—at least from the practical point of view of error theory—for a really sound justification of the Laplacian approach. As we have seen above, Cauchy’s bounds for the difference of the actual and the normal probability distribution were quite appropriate in an asymptotic sense. In many cases of practical importance, however, his bounds were scarcely usable.

In the case of direct observations, for example, the equations of condition are

$$x = k_j - \epsilon_j \quad (j = 1, \dots, n).$$

The least square multipliers are identical with $\frac{1}{n}$. If the errors obey a uniform density within the interval $[-1; 1]$, then for $\Theta \geq \sqrt{n}$ the constant r in (2.48) (which is only

important for $x \geq \Theta$) can be assumed to be $r = 0.9$ if $n \geq 10$.³⁸ According to [Sleshinskii 1892, 255], Cauchy's estimates can be applied if

$$n > \max \left(8; \frac{4\beta^2}{\alpha^2}; \frac{8\kappa^2\beta^2}{r\alpha^2} \right)$$

and

$$\frac{2\sqrt{2n}}{\alpha\sqrt{r}} < \Theta < \frac{n}{\kappa\beta},$$

where $[-\kappa; \kappa]$ is the support of the error density f , and α, β are according to (2.41). In our case we can choose $\alpha = \beta = 1$, and the first of the latter conditions is satisfied for all $n \geq 9$. For $n = 10$, $v = 0.1$, and $r = 0.9$ the sum $C_1 + C_2 + C_3$ (dependent on a Θ which has still to meet Sleshinskii's second condition) is at its minimum (for $\Theta \approx 9.43$) approximately equal to 0.288. In the case at hand, the probability $P(-v \leq \sum \lambda_j \epsilon_j \leq v)$ with $\lambda_j = \frac{1}{n}$ can be directly calculated by use of the formula (2.1), which was already derived by Laplace in the 1770s. The exact value of this probability is (for $n = 10$, $v = 0.1$) equal to 0.41096, whereas the approximation by the normal distribution gives the value 0.4161. Similar calculations for other v show that, already for $n = 10$, the difference between the exact and the approximate value is less than 1/100. If $n > 10$, for a comparison with the case of 10 observations we have to use values of v which decrease in the ratio $\sqrt{10/n}$. For $n = 20$, $v = 0.1 \cdot \sqrt{0.5}$, and $r = 0.9$ the minimum sum $C_1 + C_2 + C_3$ is roughly 0.16 ($\Theta \approx 13.4$); for $n = 100$, $v = 0.1 \cdot \sqrt{0.1}$, and $r = 0.9$ the minimum sum is still about 6/100 ($\Theta \approx 33.2$). A critical numerical discussion of this kind was certainly within Cauchy's reach, and his above-quoted reference to the "true significance" of his "formulae" might point in this direction.

Thus, within the framework of observational practice, by applying Cauchy's bounds one was able to confirm Laplace's point of view only if a really large number of observations appeared. Certainly, a great many observations were occasionally available in the context of astronomical problems. Bessel [1818, 18–21], in his comparison of the frequency distributions of the residuals of direct observations on the one hand, and normal distributions on the other, had used two series with 300 observations each, and one with 470.³⁹ Alexis Bouvard had considered approximately 130 equations of condition for Jupiter and another 130 for Saturn in his determination of the orbit elements of these planets. This work was described by Laplace [1812/20/86, 516] as an "immense travail."⁴⁰ In most cases, however, the number of observations was far below 100. Gauss [1811], for example, determined his "improvements" of elliptical elements of Pallas from only 11 equations of condition.

³⁸ In our special case the "fonction auxiliaire" is $\varphi(z) = \frac{\sin z}{z}$. For $z^2 \geq 10$ the estimate $\frac{1}{1+0.9z^2} \geq \frac{1}{z^2}$, and thus, $[\varphi(z)]^2 \leq \frac{1}{1+0.9z^2}$ is valid.

³⁹ See [Stigler 1986, 204; Hald 1998, 361–363].

⁴⁰ For a summary of Bouvard's work see [Bouvard 1821].

There exists a brief report [Cauchy 1853g'] in the *Comptes rendus* referring to Cauchy's remarks on Bienaymé's defense [1853e] (see Sect. 2.5.3) of Laplace's approach to least squares. Concerning Laplace's analytical methods, we read:

The analysis by which he [Laplace] has established the properties of the method of least squares uses series expansions whose convergence is not proven. M. Cauchy has replaced this analysis by exact and rigorous formulae.

Thus, we can see that Cauchy clearly stressed his “new” analytical rigor as an exceptional merit as opposed to Laplace's style of reasoning. But, from the practical point of view of error theory, he neither succeeded in improving Laplace's analysis by establishing sufficiently close bounds for the error of approximation, nor did he succeed in giving convincing counterexamples concerning the method of least squares. We should not forget that Cauchy's main interest was originally to give an effective procedure for astronomical calculations (see Sect. 2.5.1). Thus, his turn toward an “abstract” point of view which scarcely considered questions like general applicability or computational simplicity was—in a certain sense—contrary to his original aims. From a purely mathematical point of view, however, Cauchy's contribution even enforced the Laplacian preference to least squares in the case of bounded errors. Naturally, Cauchy could not exclude the possibility of bounds more appropriate than his own (which in fact can be derived by modern methods). Bienaymé, however, whose analytical abilities were likewise at a respectable level, was unable to give an exact mathematical argument in favor of Laplace's position. On the contrary, by showing that (in modern terminology) the estimator obtained from *any* system of multipliers (if these are of an order of magnitude inversely proportional to the number of observations) converges stochastically to the true value, he showed at the same time, that the method of least squares could be, presupposing a “very large” number of observations, only slightly superior (according to Laplace's criterion) to other methods. Thus, the end of the scientific controversy was not so much determined by Cauchy's hypothetic fear of coming too close to Bienaymé's position, but rather by a situation in which neither of the two scientists was able to make further substantial contributions. In this sense, the dispute ended in a tie.

2.5.7 Conclusion: Steps Toward Modern Probability

Laplace's version of the CLT served mainly as a tool of “good sense” and therefore its importance was primarily determined by a field beyond mathematics. Around the mid-19th century, due to the contributions of Dirichlet and Cauchy, the CLT became part of mathematics in the narrow sense. In Dirichlet's work, it served as an illustration of special analytical techniques, whereas Cauchy used it for his approach to an error theory which was mainly determined by purely mathematical goals. In adjusting Laplacian approximation techniques to an analytical style different from algebraic analysis, they contributed to the development of new standards within analysis. Poisson, with his contributions to the CLT, however

still according to the principles of classical probability, considerably influenced Dirichlet's and Cauchy's work through his innovative analytical techniques and through his discussion of the validity of normal approximations.

On the one hand, in Dirichlet's and Cauchy's contributions the CLT obtained a substantial intramathematical role. In Cauchy's work, it was connected with a rather abstract and therefore almost "modern" perspective of error theory. On the other hand, it had not yet reached an entirely independent status within mathematics. In particular, general statements independent of the original context of applications were still lacking. Full autonomy, according to Mehrtens [1990] an essential characteristic of the modernization of mathematics, was not reached for the CLT until Lyapunov published his epochal work on the "Theorem of Laplace" in 1900/1901.

Appendix: Original Text of Dirichlet's Proof of the Central Limit Theorem According to Lecture Notes from 1846

The following text is a transcription⁴¹ of pages 25 to 31 of the lecture notes [Dirichlet 1846] (for closer bibliographic details see References and [Fischer 1994]). To the greatest extent possible, the original wording is reproduced to the letter, and the original punctuation is kept as well. As a rule, "mistakes" are therefore not due to misprints. The original page numbers are also referred to.

Seite 25

..., Es möge sich bei einer bestimmten Gattung von Beob. das Fehlergesetz von Beob. zur Beob. beliebig ändern, dabei aber doch, was ja immer erreicht werden kann, indem man nur das größte als Norm nimmt, sämtliche Fehlergesetze $f_1(x_1), f_2(x_2), f_3(x_3) \dots f_n(x_n)$ zw. festen Grenzen $\pm a$ enthalten sein, man soll bestimmen wie groß die Wahrscheinlichk. ist, daß wenn man die Fehler der einzelnen Beobachtungen $x_1, x_2, x_3 \dots$ mit den respectiven Constanten $\alpha_1, \alpha_2, \alpha_3 \dots$ multipliziert, die Productsumme zw. gegebenen Gränzen g und h liege; daß man also habe:"

$$g < \alpha_1 x_1 + \alpha_2 x_2 + \dots + \alpha_n x_n < h$$

Zur Lösung dieser Aufgabe bemerken wir, daß in Folge des Vorhergegangenen die Probabilitäten, daß der erste Fehler zwischen

Seite 26

den Grenzen x_1 u. $x_1 + \partial x_1$, der zweite zw. den Grenzen x_2 u. $x_2 + \partial x_2$, der n^{te} zw. den Grenzen x_n u. $x_n + \partial x_n$ liege, ausgedrückt sind durch

$$f_1(x_1)\partial x_1, f_2(x_2)\partial x_2 \dots f_n(x_n)\partial x_n$$

für die Größe der Probabilität, daß diese Fehler zw. den Grenzen g und h enthalten sind, hat man die Ausdrücke:

$$\int_g^h f_1(x_1)\partial x_1, \int_g^h f_2(x_2)\partial x_2 \dots \int_g^h f_n(x_n)\partial x_n$$

und die zusammengesetzte Wahrscheinlichk., daß diese Grenzen bei allen gleichzeitig Statt finden, ist gleich dem Vielfachen Integrale:

$$\iiint \dots \iiint \int_g^h f_1(x_1) f_2(x_2) \dots f_n(x_n) \partial x_1 \partial x_2 \dots \partial x_n$$

Zur Discussion dieses Integrals, wollen wir fürs erste den Anfangsp. von dem aus man die Grenzen g und h zehlt in den P. $\frac{g+h}{2}$ verlegen. Es wird dann wenn man

⁴¹ Courtesy of Institut für Geschichte der Naturwissenschaften, Universität München, Professor M. Folkerts.

$g = -\lambda$ setzt offenbar $h = +\lambda$, und unsere Ungleichung geht über in

$$-\lambda < \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_n x_n < \lambda$$

$$\text{oder} \quad -1 < \frac{1}{\lambda}(\alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_n x_n) < +1$$

Wir wenden nun das bekannte Verfahren eines Multipliers an. Man hat nemlich

$$\int_0^\infty \frac{\sin \varphi}{\varphi} \partial \varphi = \frac{\pi}{2}$$

oder wenn man $l\varphi$ st φ schreibt:

$$\frac{2}{\pi} \int_0^\infty \frac{\sin l\varphi}{\varphi} \partial \varphi = \pm 1$$

je nachdem l eine positive oder negative Constante vorstellt. Mittelst dieses Integrals kann man nun leicht sich den gewünschten Multiplier verschaffen. Es ist nemlich:

$$\frac{2}{\pi} \int_0^\infty \frac{\sin \varphi}{\varphi} \cos k\varphi \partial \varphi = \frac{2}{\pi} \left\{ \frac{1}{2} \int_0^\infty \frac{\sin(1+k)\varphi}{\varphi} \partial \varphi + \frac{1}{2} \int_0^\infty \frac{\sin(1-k)\varphi}{\varphi} \partial \varphi \right\}$$

woraus man mit Hilfe des vorhergehenden Integrales erhält:

Seite 27

$$\frac{2}{\pi} \int_0^\infty \frac{\sin \varphi}{\varphi} \cos k\varphi \partial \varphi = \begin{cases} 0 & \text{für } k > 1 \text{ absolut genommen} \\ 1 & \text{„ } -1 < k < 1. \end{cases}$$

In Folge unserer Ungleichheitsbedingung $-1 < \frac{1}{\lambda}(\alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_n x_n) < 1$ kann man mit diesem Integral das zu untersuchende durch Multiplication verbinden, wodurch man erhält:

$$\begin{aligned} \frac{2}{\pi} \iint \cdots \iiint \int_{-a}^a \int_0^\infty f_1(x_1) f_2(x_2) \cdots f_n(x_n) \frac{\sin \varphi}{\varphi} \times \\ \times \cos(\alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_n x_n) \frac{\varphi}{\lambda} \partial \varphi \partial x_1 \partial x_2 \partial x_3 \cdots \end{aligned}$$

Nun ist bekanntlich: $\sqrt{-1} \int_{-\mu}^\mu \sin \frac{\alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_n x_n}{\lambda} \varphi \partial \varphi = 0$ und hieraus und dem obigen Ausdrucke wird durch Addition:

$$\begin{aligned} \frac{2}{\pi} \iint \cdots \iiint \int_{-a}^a \int_0^\infty f_1(x_1) f_2(x_2) \cdots f_n(x_n) \frac{\sin \varphi}{\varphi} \times \\ \times e^{(\alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_n x_n) \frac{\varphi}{\lambda} \sqrt{-1}} \partial x_1 \partial x_2 \partial x_3 \cdots \partial \varphi \end{aligned}$$

Setzt man $\lambda\varphi' = \varphi$, und läßt man nach geschehener Subst die Accente wieder weg, so erhält man:

$$\frac{2}{\pi} \iint \cdots \iint \int_{-a}^a \int_0^\infty f_1(x_1) f_2(x_2) \cdots f_n(x_n) \frac{\sin \lambda \varphi}{\varphi} \times \\ \times e^{(\alpha_1 x_1 + \alpha_1 x_1 \cdots \alpha_n x_n) \varphi \sqrt{-1}} \partial x_1 \partial x_2 \cdots \partial x_n \partial \varphi$$

was man offenbar auch in folgender Form schreiben kann:

$$\frac{2}{\pi} \int_0^\infty \frac{\sin \lambda \varphi}{\varphi} \left\{ \int_{-a}^a f(x) e^{\alpha_1 x_1 \varphi \sqrt{-1}} \partial x \right\} \left(\int_{-a}^a f_2(x_2) e^{\alpha_2 x_2 \varphi \sqrt{-1}} \partial x_2 \right) \cdots \\ \cdots \left(\int_{-a}^a f_n(x_n) e^{\alpha_n x_n \varphi \sqrt{-1}} \partial x_n \right) \Big\} \partial \varphi \quad (1)$$

Wir müssen uns nun fürs erste mit der Discussion des Integrales

$$\int_{-a}^a f(x) e^{\alpha x \varphi \sqrt{-1}} \partial x = \int_{-a}^a f(x) \cos(\alpha x \varphi) \partial x + \sqrt{-1} \int_{-a}^a f(x) \sin(\alpha x \varphi) \partial x \\ = \int_{-a}^a f(x) \cos(\alpha x \varphi) \partial x = \int_{-a}^a f(x) \cos \beta x \partial x \quad \alpha \varphi = \beta$$

beschäftigen. Dasselbe erreicht für $\beta = 0$ sein Max, wo es dann, da $f(x)$ als Ausdruck einer Wahrscheinlichk., nie negativ werden kann, aus lauter positiven Elementen besteht. Die Reihe

$$\int_{-a}^a f(x) \cos \beta x \partial x = \int_{-a}^a f(x) \partial x - \frac{\alpha^2 \varphi^2}{2} \int_{-a}^a x^2 f(x) \partial x + \cdots$$

Seite 28

in der α eine gegebene endliche Constante bezeichnet, convergirt für sehr kleine Werthe von φ , so schnell, daß fast der ganze Werth des Integrales in den beiden ersten Gliedern der Reihe enthalten ist, wodurch bewirkt wird, daß der ganze Werth unseres Integralproductes sich im Anfange concentrirt. Setzt man zur Abkürzung

$$\frac{1}{2} \int_{-a}^a x_v^2 f(x_v) \partial x_v = k_v$$

wo k_v eine Constante bezeichnet, die sich nur für die betreffenden Beobachtungen, von deren Fehlergesetz es abhängt, verändert, so erhält man wegen der Relation

$$\int_{-a}^a f(x) \partial x = 1$$

für einen Factor obigen Doppelintegrales den Ausdruck:

$$1 - k_v \alpha_v^2 \varphi^2 + \dots$$

Nimmt man den Neper'schen Logarithmus, so bekommt man, wenn man dieselben in Reihen auflöst:

$$\log \text{nat}(1 - k_v \alpha_v^2 \varphi^2 + \dots) = -k_v \alpha_v^2 \varphi^2 + \dots$$

wo man für v die Zahlen $1, 2, 3, \dots, n$ zu setzen hat, und die so erhaltenen Ausdrücke sodann alle zu addiren. Geht man dann von den Logarithmen wieder zu den Zahlen über, so erhält man einen Ausdruck für das oben behandelte Produkt aus Integralfactoren. Man hat aber hiebei auch dafür zu sorgen, daß die weggelassenen Glieder der Reihe absolut klein seien, nicht bloß klein im Verhältniße zum ersten Gliede. Denn in einer Exponentialgröße $e^{\alpha+\beta} = e^\alpha e^\beta$, wie sie hier auftritt, darf man offenbar nur dann den zweiten Theil β des Exponenten vernachlässigen, wenn es eine absolut verschwindende Größe ist.

Wenn man nun unsere n Gleichungen

$$\log \text{nat}(1 - k_v \alpha_v^2 \varphi^2 + \dots) = -k_v \alpha_v^2 \varphi^2 + l \varphi^4 + \dots$$

Seite 29

summirt, so wird die Summe der Glieder der vierten Ordnung immer $\leq n l \varphi^4$ sein, wenn man nemlich mit l den größten vorhandenen Entwicklungscoeff bezeichnet, und also eine gewisse Constante sein wird in Beziehung auf φ , und n als der Index der einzelnen Beobachtungen als eine immer wachsende Größe gedacht werden muß. Nach dem oben Gesagten muß nun unser Bestreben immer dahin gehen, daß die Summe der höheren Entwicklungscoeff immer eine absolut kleine Zahl bleibt, und dieß bewerkstelligen wir dadurch daß wir in dem willkürlich eingeführten Integrale mit der Variablen φ diese letztere bloß soweit wachsen lassen, daß das Produkt $n \delta^2$, wo δ einen Zustand von φ bezeichnet, immer wie groß auch n werden möge, sich der Grenze Null nähert. Dann reduzirt sich in vorstehender Formel das Product der Co... auf die Summe ihrer ersten Entwicklungsglieder, und man erhält statt unserer Formel (1) sogleich:

$$\frac{2}{\pi} \int_0^\delta e^{-\varphi^2 \sum_{v=1}^{v=n} k_v \alpha_v^2} \frac{\sin \lambda \varphi}{\varphi} \partial \varphi + \frac{2}{\pi} \int_\delta^\infty e^{-\varphi^2 \sum_{v=1}^{v=n} k_v \alpha_v^2} \frac{\sin \lambda \varphi}{\varphi} \partial \varphi = p + q$$

wo also δ durch die Bedingung bestimmt wird, daß $\delta \sqrt[4]{n}$ für ein zunehmendes n immer kleiner und kleiner wird. Beschäftigen wir uns zuerst mit dem ersten Integrale

$$p = \frac{2}{\pi} \int_0^\delta e^{-\varphi^2 \sum k_v \alpha_v^2} \frac{\sin \lambda \varphi}{\varphi} \partial \varphi$$

so wir für $\varphi = \frac{\psi}{\sqrt{n}}$ sogleich:

$$p = \frac{2}{\pi} \int_0^{\delta \sqrt{n}} e^{-\psi^2 \frac{\sum k_v \alpha_v^2}{n}} \frac{\sin \frac{\lambda}{\sqrt{n}} \psi}{\psi} \partial \psi$$

Nach unserer Bedingung soll nun $\delta \sqrt[4]{n}$ beständig abnehmen, womit aber durchaus nicht gesagt ist, daß dies auch mit $\delta \sqrt{n}$ der Fall sein müsse, welches im Gegenteil sogar immer größer werden kann, wie dieß z.B. für die Annahme $\delta = \frac{1}{\sqrt[3]{n}}$ statt findet. Bestimmen wir nun

Seite 30

das δ so, daß für ein zunehmendes n das Product $\delta \sqrt[4]{n}$ immer abnimmt, $\delta \sqrt{n}$ aber wächst, so erhalten wir für eine sehr große Anzahl von Beobachtungen offenbar:

$$p = \frac{2}{\pi} \int_0^\infty e^{-\psi^2 \frac{\sum k_v \alpha_v^2}{n}} \frac{\sin \lambda \psi}{\psi} \partial \psi$$

wobei noch zu bemerken ist, daß man hier, wie auch geschehen ist, auch das λ mit wachsendem n zunehmen muß, indem offenbar die Wahrscheinlichkeit, daß bei unendlich vielen Beobachtungen der Fehler zw. gegebenen festen Grenzen liege, Null ist. Wir setzen deshalb $\lambda \sqrt{n}$ statt λ . Nun ist aber zur Vereinfachung dieses Resultates bekanntlich:

$$\int_0^\infty e^{-c^2 \varphi^2} \cos \lambda \varphi \partial \varphi = \frac{\sqrt{\pi}}{2c} e^{-\frac{\lambda^2}{4c^2}}$$

$$\int_0^\lambda \partial \lambda \int_0^\infty e^{-c^2 \varphi^2} \cos \lambda \varphi \partial \varphi = \int_0^\infty e^{-c^2 \varphi^2} \frac{\sin \lambda \varphi}{\varphi} \partial \varphi = \frac{\sqrt{\pi}}{2c} \int_0^\lambda e^{-\frac{s^2}{4c^2}} \partial s$$

$$\text{oder} \quad \int_0^\infty e^{-c^2 \varphi^2} \frac{\sin \lambda \varphi}{\varphi} \partial \varphi = \sqrt{\pi} \int_0^{\frac{\lambda}{2c}} e^{-s^2} \partial s$$

und hiemit wird:

$$p = \frac{2}{\sqrt{\pi}} \int_0^{\frac{\lambda}{2\sqrt{\frac{\sum k_v \alpha_v^2}{n}}}} e^{-s^2} \partial s.$$

als Ausdruck der Wahrscheinlichk. daß die GröÙe $\sum_{v=1}^{v=n} \alpha_v x_v$ zw. den Gränzen $\pm \lambda \sqrt{n}$ enthalten sei, oder daß sei

$$-\lambda \sqrt{n} < \sum_{v=1}^{v=n} \alpha_v x_v < \lambda \sqrt{n}$$

wenn n immer größer wird. Es ist hiedurch allein schon diese Wahrscheinlichkeit ausgedrückt, weil das zweite Integral

$$q = \int_\delta^\infty e^{-\varphi^2 \sum k_v \alpha_v^2} \frac{\sin \lambda \varphi}{\varphi} \partial \varphi$$

sich unaufhörlich der Null nähert. Diese Behauptung erweist man folgender maßen. Wir haben bei der Form (1) bemerkt, daß $2 \int_0^a f(x) \cos(\alpha \varphi x) \partial x$ für $\alpha x = c$ ein absolutes Maximum und $= 1$ sei

Seite 31

und niemals mehr unter den später eintreten Maximis ein diesen an Größe gleich kommendes sich befinde. Man kann sich nun das Intervall so klein denken, daß die Function innerhalb deßselben nicht allein beständig abnimmt, sondern auch noch größer bleibt, als sie später irgendwo noch werden kann, *a fortiori* also das Product aller dieser analog gebildeten Functionen, und man kann diesen Zustand der Ungleichheit durch Verkleinerung des Intervalles von 0 bis δ soweit treiben, als verlangt wird, woraus unsere Behauptung folgt. Strenger läßt sie sich aber auf folgende Weise rechtfertigen. Man zerlege das Integral \int_{δ}^{∞} in die Summe zweier andren $\int_{\delta}^{\Delta} + \int_{\Delta}^{\infty}$, so wird immer im ersteren die Function am Anfange größer sein, als irgendwo später. Es wird also das Integral kleiner sein, als die Differenz der Grenzen $\Delta - \delta$, also um so mehr kleiner als der Anfangswerth der Function, und es nähert sich deswegen, wenn man das Δ einer positiven Potenz von n proportional nimmt, der Werth des Productes der Grenze Null. Was nun noch das zweite Integral betrifft, so hat man durch partielle Integration:

$$\int \cos(\alpha\varphi x) f(x) dx = \frac{\sin(\alpha\varphi x)}{\alpha\varphi} f(x) - \int \frac{\sin(\alpha\varphi x)}{\alpha\varphi} f'(x) dx$$

$$\int_{-a}^a \cos(\alpha\varphi x) f(x) dx = \frac{2 \sin(\alpha\varphi a)}{\alpha\varphi} f(a) - \int_{-a}^a \frac{\sin(\alpha\varphi x)}{\alpha\varphi} f'(x) dx$$

wo wir also jetzt auch noch annehmen müssen, daß $f(x)$ innerhalb der Integralgränzen endlich bleibt, welche Annahme durch die Natur der Fehlercurve vollkommen gerechtfertigt ist. Die Zähler beider Ausdrücke schwanken immer zw. gewissen Grenzen hin und her, und der Werth des Integrales wird daher kleiner als $\frac{c}{\varphi}$, und folglich der Werth des Productes n solcher Integrale $< \frac{c}{\varphi^n}$ welcher sich also mit wachsendem φ der Null nähert.

<http://www.springer.com/978-0-387-87856-0>

A History of the Central Limit Theorem
From Classical to Modern Probability Theory

Fischer, H.

2011, XVI, 402 p., Hardcover

ISBN: 978-0-387-87856-0