Revista Brasileira de Ensino de Física, vol. 44, e20220087 (2022)
www.scielo.br/rbef
DOI: https://doi.org/10.1590/1806-9126-RBEF-2022-0087

Articles

Structure of the theories of probability

Mário J. de Oliveira*[1]

[1]Universidade de São Paulo, Instituto de Física, 05508-090, São Paulo, SP, Brasil.

Received on March 14, 2022. Revised on March 27, 2022. Accepted on April 06, 2022.

We examine the concept of probability from its emergence within the realm of the games of chance and the development of the theory of probability until the appearance of the treatise of Kolmogorov on this subject. The discipline related to that theory is framed as the science of aleatory events. Probability is understood as a primitive concept represented by a measure assigned to the space of events and obeying the fundamental postulates of the theory. The measurement of probability is the ratio of the number of the observed favorable outcomes and the total number of observed outcomes when these numbers are larger enough.
Keywords: Theory of probability, chance in dice games, Bernoulli trial, central limit theorem, Markov chain.

1. Introduction

The theory of probability as well as geometry are sometimes considered pure mathematical theories. However, geometry can be understood as the science related to the distance, shape, and size of figures in real space. This understanding is not recent and is at the very beginning of the science of geometry. A similar appreciation can be made of the theory of probability which can be framed as the *science of real aleatory events*. The connection between the concept of probability and aleatory events is also at the genesis of the science of probability which, in retrospect, is to be found in the realm of the games of chance.

The recognition of the theory of probability as the science of aleatory events is accomplished by the understanding that the probability of an event is interpreted as the *frequency* of the occurrence of that event. Notice that frequency is neither the probability nor the definition of probability. Given the possible outcomes of an aleatory experiment, the probability of an event is determined by *assigning* a number between zero and unity to each one of the mutually exclusive elementary events. The possible outcomes of an aleatory event constitute the space of events or the sample space. In the case of the throwing of a die, it consists of the set of numbers from one to six.

A relevant aspect of probability that we wish to emphasize here is that it should be understood as a *primitive concept* of the theory. A primitive concept such as time, space and mass cannot be defined in terms of other more fundamental concepts [1]. Nevertheless, they can and are apprehended by our minds, and are consolidated once we know how they are measured. Time is measured by a clock, distance by a ruler, and weight by a balance. The probability of an event is measured by the frequency of its occurrence in a real experiment. Notice that these forms of measurement are not the definitions of the concepts.

The concept of probability grew up within the realm of the games of chance and the earliest probability calculation were based on a general rule of equiprobability of the elementary outcomes in the throwing of dice. Since we do not expect to find data of observed frequencies associated to these calculations, a question then arises as to how the probabilistic calculations were compared with real frequencies. This question is answered by bearing in mind that estimates of the real frequencies were empirically known to the experienced dice players.

The probability related to frequency is not the only concept of probability that has been conceived. There are other concepts with the name of probability. Carnap claimed that there are two fundamentally different kinds of probability which he calls logical probability and statistical probability [2]. The first kind of probability is understood as a degree of confirmation, which he also termed inductive probability because it is related to what he called inductive inference.

Hacking pointed out [3] the existence of two kinds of probability, which he called the duality aspect of probability. One type of probability he called epistemic probability, understood as the "degrees of belief in propositions quite devoid of statistical background" [3]. The duality was also pointed out by Poisson and by Cournot who used the French words *chance* for the aleatory concept and *probabilité* for the epistemic concept [3]. Both Poisson and Cournot remarked that the former is an objective concept whereas the latter depends on our knowledge about the event and is thus subjective [4, 5].

The term probability and its cognates in other languages did not always have the current scientific meaning, which emerged in the second half of the seventeenth century [3]. The earlier meaning of the term probability

* Correspondence email address: oliveira@if.usp.br

was the state of being approved or worthy of approval [3]. The French word *probabilité* appeared in the modern sense in the last pages of a book on logic, known as the Port Royal Logic, published in 1662 [6]. Although the term probability in the modern sense appeared around the second half of the seventeenth century, the concept of probability related to frequency appeared before this date, as we have seen above, and was called by other names such as chance.

In the following we analyze the development of the concept of probability and of the theory of probability [7–12] starting from its emergence within the realm of the games of chance, including Cardano and Galileo. It is usual to proclaim that the theory of probability started with the exchange of letters between Pascal and Fermat in 1654, an attribution that was made by Laplace. In his treatise on the theory of probability of 1812 [13], he stated that the birth of this science is due to these two Frenchmen, and that it was subsequently extended by Huygens and developed by Jacob Bernoulli, Montmort and Moivre. However, the concept of probability is much older. Laplace himself conceded that before Pascal and Fermat, people were determining for quite a long time the ratio between favorable and unfavorable cases in games of chance [14, 15].

In addition to the works of the authors mentioned by Laplace, we analyze his treatise on the theory of probability, and the works of Daniel Bernoulli, Lagrange, Poisson, Bertrand, Borel, Markov, ending with the treatise of Kolmogorov on the foundations of the probability theory, published in 1933. Our focus will rest on the exposition of the theoretical aspects which in many cases are identified as the methods of probability calculation.

2. Chance in Dice Games

Un coup de dés n'abolira
jamais le hasard
Toute pensée émet un coup de dés
Mallarmé, 1897 [16]

2.1. *De Vetula*

De Vetula [17] is the name of a poem written in Latin in France in the mid-thirteenth century containing probability calculations [19]. It was ascribed to a medieval author who wrote the poem in the form of an autobiography of the Roman poet Ovid [19]. The oldest printed versions are from the last decades of the fifteenth centuries [19]. In the first of three parts of the book the author describes the calculation of chances in the throw of three dice. According to Bellhouse [19], the poem was well known and some readers may well have understood clearly the probability calculations contained in the poem, and that an elementary probability calculus may have been established in Europe from the second half of the thirteenth century.

The translation of the first half of the passage of the poem dealing with the throw of dice is as follows [19]: *Perhaps, however, you will say that certain numbers are better /Than others which players use, for the reason that, /Since a die has six sides and six single numbers, /On three dice there are eighteen, /Of which only three can be on top of the dice. /These vary in different ways and from them, /Sixteen compound numbers are produced. They are not, however, /Of equal value, since the larger and the smaller of them /Come rarely and the middle ones frequently, /And the rest, the closer they are to the middle ones, /The better they are and more frequently they come.*

In these lines the author explains that in a throw of three dice, there are 16 possible outcomes, which are the numbers from 3 to 18, as shown in the last column of the upper table of Figure 1. Each outcome is the sum of the pips on the top of the dice. The outcomes are not of equal value, that is, they do not have the same frequency. The largest and the smallest are less frequent and the middle ones are more frequent and are thus considered better numbers.

The translation of the second half is as follows [19]: *These, when they occur, have only one configuration of pips on the dice, /Those have six, and the remaining ones have configurations midway between the two, /Such that there are two larger numbers and just as many smaller ones, /And these have one configuration. The two which follow, /The one larger, the other smaller, have two configurations of pips on the dice apiece. /Again, after them they have three apiece, then four apiece. /And five apiece, as they follow them in succession approaching /The four middle numbers which have six configurations of pips on the dice apiece.*

In these lines the author explains that each outcome corresponds to one or more configurations, as can be seen in the upper table of Figure 1. The smallest and largest outcomes correspond to a small number of configurations. The middle ones corresponds to a larger number of configurations. For instance, the outcome 17 corresponds to just one configuration which is (665). The outcome 15 correspond to three configurations which are (663), (654), and (555). The total number of configurations is 56, which are the possible configurations of the pips on the top of the three dice.

The author then explains that each configuration can come in various ways of falling, except those configurations in which the numbers are equal such as (333) in which case there is just one way of falling. A configuration where two numbers are equal and one is different correspond to three ways of falling. For instance the configuration (332) corresponds to the ways (3,3,2), (3,2,3), and (2,3,3). A configuration where the three numbers are different, the number of ways of falling is six. For instance, the configuration (532) corresponds to the ways (5,3,2), (5,2,3), (2,5,3), (2,3,5), (3,2,5), and (3,5,2).

 DOI: https://doi.org/10.1590/1806-9126-RBEF-2022-0087

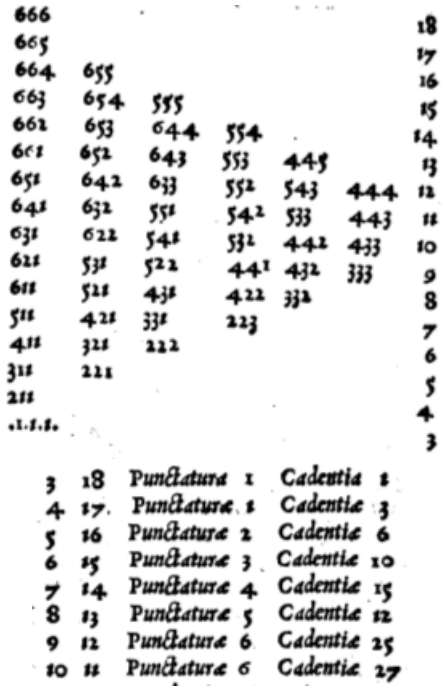


Figure 1: Two tables from a printed version of 1534 of the poem *De Vetula* [17], written around 1250. The upper table shows the possible configurations of three dice and the possible outcome in the last column. The bottom table shows the possible outcomes in the first and second column, the number of configurations (*punctaturae*) and the number of ways of falling (*cadentiae*).

Using the three rules above and taking into account the configurations shown in the upper table of Figure 1, we may determine the ways of falling for each outcome of the last column, and they are shown in the bottom table of Figure 1. For instance the outcome 10 corresponds to 27 ways of falling. The total number of ways of falling is 216. The chance a certain number is thus related to ways of is falling which is the number of its permutations.

A question that should be raised here is whether the ways of falling (*cadentiae*) are not just combinatorial calculations without relation to actual throw of dice. This does not seem to be the case because the words rarely and frequently, which are understood as related to actual throw of dice, were explicitly mentioned in the poem.

2.2. Commentary to the Divine Comedy

In a commentary on the Divine Comedy of Dante called *L'Ottimo Commento* one finds an indication of the calculation of chances in a throw of three dice [20, 21]. The preface of a printed version of the commentary, published in 1827–1829 [20], says that the author of the commentary was a contemporary of Dante and that the antiquity of the commentary is attested by the mentioning of some events that occurred in 1333.

The passage related to the game of dice corresponds to the first lines of the sixth canto of the Purgatorio, which reads [22]:

> *Quando si parte il gioco de la zara,*
> *colui che perde si riman dolente,*
> *repetendo le volte, e tristo impara.*

When the game of zara starts, the one who loses is sore, repeating the throws, and in sadness learns. The game of zara [22, 23] was usually played with three dice. Each player throws the dice and at the same time tries to guess the number of the sum of the pips on the top of the dice by declaring the number aloud. If the guess is correct, the player gets a number of coins equal to the number called. Otherwise, the player pays a number of coins equal to the number that came out.

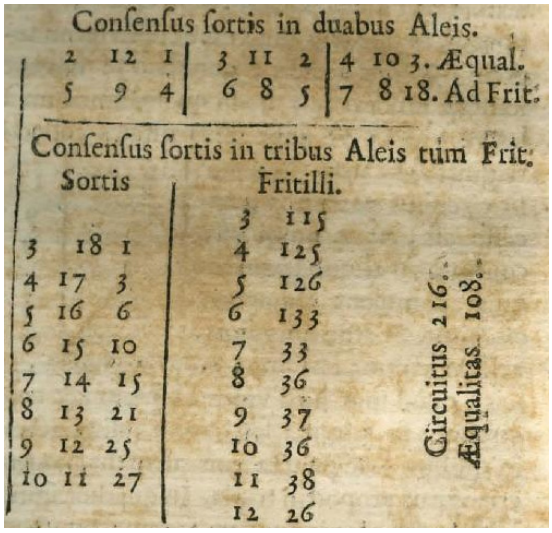
The author of the commentary states that the number three comes in one way. The number four comes in a way such that one die has the number two and the other two dices have aces. Thus we cannot expect too much from these numbers. They are called unlucky and are not considered in the game. The same can be said about the numbers eighteen and seventeen which are equally unlucky. The number between them can come in multiple ways and the best ones are those that come in more that one way. These comments indicate that the more frequent numbers are those that can be obtained in more ways. However, no quantitative calculations are given, and it is unclear whether the ways are related to partitions or permutations.

2.3. Cardano

Cardano (1501–1576) wrote a book about the games of chance which was found among his manuscripts and published posthumously in 1663 within his collected works [24–26]. The book began to be written when he was twenty-five years of age and was rewritten when he was sixty-four [24]. The book presents a calculation of the chances in the throw of dice which we describe here.

In the throw of two dice there are six results such that the numbers on the top faces are equal, and 15 in which the results are distinct. Doubling this last results we find 30 which added to six make up 36 cases (permutations), which Cardano calls the cycle. In the upper table of Figure 2, it is shown the number of cases corresponding to a certain sum of points of the top faces of two dice. For 2 and 12 points there is just one case; for 3 and 11, two cases; for 4 and 10, three cases, for 5 and 9, four cases, for 6 and 8, five cases; and for 7, six cases. For instance, for 4 points, the cases are (1,3), (2,2), and (3,1).

In the throw of three dice, there are 6 results in which the number are all equal. There are 30 results in which two number are equal and one different, and they occur in three ways, which make up 90 cases. The number of results with three different numbers is 20, each one occurring in six ways, which makes 120 cases. The total number of cases (permutations) is 216 which is the cycle.


Conſenſus ſortis in duabus Aleis.

2	12	1	3	11	2	4	10	3. Æqual.
5	9	4	6	8	5	7	8	18. Ad Frit.

Conſenſus ſortis in tribus Aleis tum Frit.

Sortis			Fritilli.		
			3	115	
3	18	1	4	125	
4	17	3	5	126	
5	16	6	6	133	Circuitus 216.
6	15	10	7	33	Æqualitas 108.
7	14	15	8	36	
8	13	21	9	37	
9	12	25	10	36	
10	11	27	11	38	
			12	26	

Figure 2: A table from the book of Cardano on games of chance published in 1663 [26]. The upper part shows the number of cases in the throw of two dice. There is a misprint: at the right of the number 7 should be the number 6 and not 8. The lower part shows the number of cases in the throw of three dice.

In the lower table of Figure 2, it is shown the number of cases corresponding to a certain sum of points of the top faces of three dice. For 3 and 18 points there just one case; for 4 and 17, three cases; for 5 and 16, six cases, for 6 and 15, ten cases, for 7 and 14, fifteen cases; for 8 and 13, twenty one cases; for 9 and 12, twenty five cases; and for 10 and 11, twenty seven cases. For instance, for 5 points, the cases are (1,1,3), (1,3,1), (3,1,1), (1,2,2), (2,1,2), and (2,2,1).

Cardano then proceeds to determine the number of cases for some combinations of points. The number of cases containing at least one ace is 11 out of 36. The number of cases containing one of the points 1 and 2 is 20, that of the points 1 to 3 is 27, that of 1 to 4 is 32, that of 1 to 5 is 35, and that of 1 to 6 is 36. Thus, if someone wants a one, or a two or a three, the number of favorable cases is 27 out of 36.

Cardano repeats this reasoning for the case of three dice finding respectively the numbers 91, 152, 189, 208, 215, and 216 cases out of 216. Thus for getting at least one ace in the throw of three dice the chance is 91/216. Cardano then argues that if one wishes to get at least one ace in two successive throws of three dice the chance is the square of this number. In three successive throws, it is the cube of this number. It becomes clear that Cardano is using a power formula for the chances in successive throws.

It is worth mentioning that Cardano established a general rule that guided his calculations. This rules concerns the possible cases of what he called the cycle, which are 6 for one die, 36 for two and 216 for three dice. According to this rule, the chances are determined in proportion of the favorable to the unfavorable cases in the cycle. For instance in the calculation of the chance of getting at least one ace in a throw of three dice there are 91 favorable cases out of 216. They are as follows. One case with all faces equal to 1; 15 cases with exactly two aces, such as (1,3,1); and 75 cases, such as (1,3,5). In other terms, Cardano is assuming as a principle that the cases of a cycle have equal chances.

2.4. Galileo

In a text about the throw of dice written by Galileo (1564–1642) but which he left unpublished, he discusses the reason why some numbers are more frequent than others in a throw of three dice. The text appeared in his collected works published in 1718 [27, 28]. Galileo gives the example that the number 9 has the same number of partition as the number 10 but the dice-players consider the 10 more advantageous than 9. The partitions of 9 are six in number: (126), (135), (144), (225), (234), and (333). The partitions of 10 are also six: (136), (145), (226), (235), (244), and (334). If we consider that the partitions have equal chances, than the frequency of 9 and 10 would be the same. But Galileo argues in a different direction.

The results that are to be considered with equal chances are not the partitions but the permutation one can make with each one of the three dice. For the number 9 they are 25 in number: (1,2,6), (1,6,2), (3,3,3) and so on. That is, for each partition we have to consider the possible permutations. For the number 10 they are 27 in number, which is larger than 25.

We see that the reasoning of Galileo is the same as that of Cardano. The results that are to be consider with equal chance are the permutations and not the partitions. It is interesting to note that Leibniz gave examples of the throw of dice that leads one to think that he considered that the partitions have equal chances and not the permutations [29]. He said that in the throw of two dice, the number 12 is as feasible as the number 11, and that the number 7 is three times more feasible than those numbers [29].

It is worth mentioning that Galileo reported that dice players claim that in a game of three dice, the number 10 is more advantageous than the number 9. As the difference between the probabilities corresponding to these outcomes is about one percent, we see that the dice players mentioned by Galileo estimated fairly precisely the frequencies of the outcomes.

3. Pascal and Fermat

The solution of a problem related to the games of chance was the subject of a series of letters exchanged between Pascal and Fermat in 1654 [30–32]. The problem, known as the problem of points (*problème des partis*), concerned the division of a stake between two players of a game of chance when the game was interrupted before its close. The game has several rounds and the players have equal

 DOI: https://doi.org/10.1590/1806-9126-RBEF-2022-0087

chance to win each round. The winner of the game is the player who wins a certain predetermined number of rounds.

The division of a stake is not properly a probabilistic problem but becomes one when if it is assumed that the division should be proportional to the chances each player has of winning the game if it proceeded. This was the point of view taken by Pascal and Fermat. The problem is thus reduced to finding in how many ways each player could win the game if the game had not been interrupted.

The solution given by Fermat is as follows. Suppose that one player need two rounds to win the game and the other needs three rounds. In four rounds one or the other player will win the game. Thus we have to consider all the 16 possible results of the game in four rounds, which are represented by

aaaa	*aaab*	*aaba*	*aabb*
abaa	*abab*	*abba*	*abbb*
baaa	*baab*	*baba*	*babb*
bbaa	*bbab*	*bbba*	*bbbb*

where the letters from left to right indicate who is the winner in the successive rounds. The letter a and b indicate that the winner of a round is the first or the second player, respectively. We see that the number of possible results with at least two letters a is 11 and with at least three letters b is 5. As the first player needs two rounds, he will win the game in 11 and the second in 5 out of the 16 possibilities. Therefore the stake is divided in the proportion of 11 to 5.

Pascal then applies the Fermat method to the case of three players. The first player needs one round to win, and the second and the third need two rounds each. In this case the game ends in three rounds and there are 27 possible cases which are

aaa	*aab*	*aac*	*aba*	*abb*	*abc*	*aca*	*acb*	*acc*
baa	*bab*	*bac*	*bba*	*bbb*	*bbc*	*bca*	*bcb*	*bcc*
caa	*cab*	*cac*	*cba*	*cbb*	*cbc*	*cca*	*ccb*	*ccc*

where the letter a, b and c indicate that the winner of a round is the first, the second, or the third player, respectively. As the first player needs just one round he wins in 17 cases, whereas the other two players win in 5 cases each, as they need two rounds. It should be remarked that in the case *abb*, it is the first player who wins the game. The second player wins the last two rounds but the first player had already won the game as he needed just one round. Notice that according to the convention used here, the first letter to the left represent the first round.

In the general case of two players, let us suppose that the first player needs n rounds to win the game and the second player needs m rounds, where m is larger than n. After $\ell = n + m - 1$ rounds, one of the players will certainly win the game. We then consider the sequences of the letters a and b, which represent the results of the rounds of the game. Let us consider the sequences in

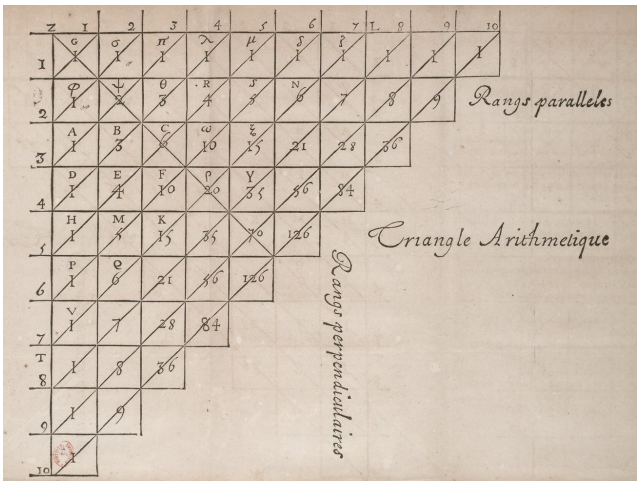


Figure 3: The arithmetic triangle from a figure of the treatise of Pascal on this subject [33]. The number in a square is the sum of the number inside the upper and left squares in accordance with the property (3).

which the letter a appear k times, which represents the number of results in which the first player wins k rounds and consequently the second player wins $\ell - k$ rounds. The first player wins the game if $k \geq m$ because in this case the second player wins at most $n-1$ rounds, which is insufficient to win the game. Therefore, if we denote by P_k^ℓ the number of of sequences in which the letter a appear k times, then the second player wins the game in a number of times equal to

$$B = \sum_{k \geq m} P_k^\ell, \tag{1}$$

and the first player wins the game in a number of times equal to

$$A = \sum_{k < m} P_k^\ell. \tag{2}$$

The stake is divided in proportion A and B for the first and second players, respectively. In the above example, $n = 2$, and $m = 3$ and $\ell = 4$, and $P_0^4 = P_4^4 = 1$, $P_1^4 = P_3^4 = 4$, $P_2^4 = 6$, from which follows $B = 4 + 1 = 5$ and $A = 1 + 4 + 6 = 11$.

We should remark that each one of the possible cases, which is represented by a sequence of letters, is assumed to have equal chances. This assumption, which is to be understood as a principle of the theory, is not explicit mentioned by either Pascal or Fermat but is implicit in their reasonings.

Pascal showed that the numbers P_k^ℓ are the numbers appearing in the arithmetic triangle shown in Figure 3. The demonstration appeared in a treatise on the arithmetic triangle [33] published posthumously in 1665 but written probably by the end of 1654 [32]. The numbers shown in this triangle are defined by

$$P_k^{\ell+1} = P_k^\ell + P_{k-1}^\ell, \tag{3}$$

and $P_0^\ell = 1$ and $P_\ell^\ell = 1$. That is, all the numbers in the square cells of the first row and first column are set equal to one. The others are obtained by the rule (3) which states that the number in a certain cell is the sum of the numbers in the upper and left cells.

From the rule (3), Pascal demonstrated several results for P_k^ℓ. One of them, which was stated as a problem, is to determine a formula for P_k^ℓ. As an example, he showed that P_4^6 equals $6.5.4.3/4.3.2.1 = 15$. A generalization of this result is

$$P_k^\ell = \frac{\ell(\ell-1)\cdots(k+1)}{k(k-1)\cdots 2\cdot 1}. \tag{4}$$

Pascal was not the originator of the arithmetic triangle but he developed much of its property and applied to the problem of the points as we have seen above. Before Pascal, the triangle appeared in many works [34], as in the second volume of a mathematical textbook by Hérigone [35] published in 1634 and cited by Pascal.

We remark that at the time that Pascal and Fermat solved the problem of points, the problem was not new. It had been considered by Pacioli, Tartaglia, Cardano, Peverone, and Forestani [8, 11, 36]. However, they failed to understand the problem as a probabilistic problem and treated the problem as an exercise in proportion or in geometric progression.

The ideas of Pascal concerning probability can be found in the book known as Port Royal Logic, [6], written by two of his friends. In this book we find the two aspect aspect of probabilities, related to the degree of belief and to frequency [3]. In fact the authors of the book use the second aspect to justify and explain the first aspect. According to the authors "to judge what one should do to obtain a good or to avoid an evil, one must not only consider the good and the evil in themselves, but also the probability that it will or not happen". To explain the meaning of probability, they immediately give the following example. In a certain game of chance, each one of ten people bets a crown and only one of them wins the total and the others lose. Individually, the game seems to be advantageous as one bets one and can win nine. However, each player has nine degrees of probability of losing a crown and just one of winning nine crowns.

4. Huygens

The findings of Pascal and Fermat become widely known through a small treatise of Huygens on the calculations in games of chance called *Ratiociniis in Ludo Aleae*, published in 1657 [37, 38]. The Dutch version was published three years later in 1660 [39]. The treatise won recognition and became the standard text in probability for the next five decades [11]. The treatise is divided in fourteen propositions where Huygens treats the problem of the division of a stake with two and three players, and problems related to the throw of dice.

The first and the second propositions are particular cases of the third proposition which reads [11]: *If the number of chances of getting a is p, and the number of chances of getting b is q, assuming always that any chance occurs equally easy, then this is worth* $(ap + bq)/(p + q)$ *to me*.

It is clear that $(ap+bq)/(p+q)$ is the definition of what we call expectation or average. However, for Huygens this is not just a mathematical definition. It seems that he is giving a real connotation to the expression in the following sense. In the problems of division of a stake, the division is made in accordance with the chances that each player has to win the game if it proceeded. The chance of a player in turn is understood as proportional to the number of games won by this player if the game would be repeated several times. Since in each game there is a certain stake, this player would get the stake a number of time equal to the number of games won. The arithmetic mean of the stakes can be understood as a real estimate of the mathematical expectation defined by Huygens. However, the arithmetic meaning has no meaning in the sense that if the game was actually carried out, the player would not get the arithmetic mean. The player would get either the whole stake or nothing. However, if the game does not proceed, the average can be used as a criterion to divide the stake.

In propositions from four to seven, he treats the problem of the division of a stake involving two players. He considers the following particular cases. The first player needs 1 round and the second player needs 2 to win the game; then 1 and 3; 1 and 4; 2 and 3; and 2 and 4. The method is similar to that of Pascal and Fermat. If the stake is denoted by a, then the shares found by Huygens for the first and the second players are $3a/4$ and $a/4$, respectively; $7a/8$ and $a/8$; $15a/16$ and $a/16$; $11a/16$ and $5a/16$; $13a/16$ and $3a/16$. In propositions eight ant nine, Huygens treats the case of three players. He consider several particular cases, some of which are shown in the table of Figure 4. Again the method of solution is similar to that of Pascal and Fermat.

The propositions from ten to fourteen deal with the chances in the throw of dice. For one die, there are 6 throws, each of which have equal chance. For two dice, there are $36 = 6 \times 6$ throws with equal chance each. For three dice, there are $216 = 36 \times 6$ throws, and so on for other number of dice. Let us call these throws, the elementary throws. To determine the chance that in certain throw the sum of the top pips of the dice is a certain given number, Huygens determines in how many ways this number can be produced. That is, he determined the number of elementary throws which yields the given number.

Huygens determined the expectation of the occurrence of a certain event at least once in a certain number of successive throw of dice. From his numerical calculation it follows that this number is

$$1 - p^n, \tag{5}$$

 DOI: https://doi.org/10.1590/1806-9126-RBEF-2022-0087

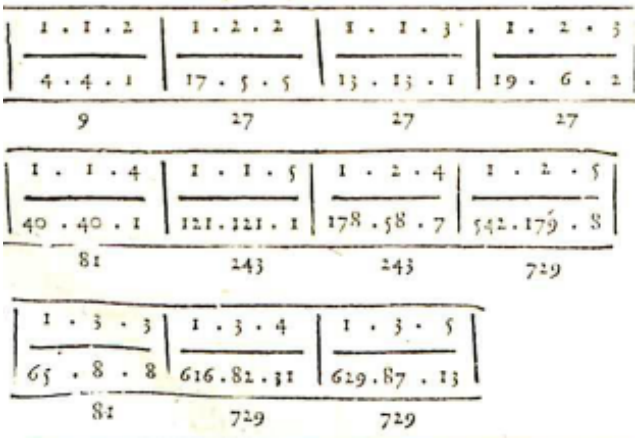


Figure 4: Table from the Huygens treatise [37] showing the results for the problem of the division of a stake between three players. The upper part of each rectangle shows the numbers of rounds needed for winning the game. The lower part shows the respective share corresponding to each player.

where p is the probability of not occurring the event in one throw, and n is the number of successive throws. For instance, the expectation of getting the number 6 at least once in 3 successive throws is $1-(5/6)^3 = 91/216$. The expectation of getting the number 12 at least once in the 2 successive throws of two dice is $1-(35/36)^2 = 71/1296$.

In the last proposition, number fourteen, two individuals A and B play a game in several rounds. They play the rounds alternatively starting with the player A. In the rounds played by A, he has a chance p of winning the round, and in the rounds played by B, he has a chance q of winning the round. The winner of the game is the player who win a round before his opponent. The problem is to find the chances of each player of winning the game. The solution given by Huygens is as follows. Let x be the chance of B winning the game so that the chance of A winning the game is $1-x$.

When it is the turn of A to play the round, x will express the chance of B. When it is the turn of B, the chance of B will be different, a value we denote by y. If A is about to play the round then the chance of B to win the game will be ry since $1-p=r$ is the chance that A does not win the round. Therefore

$$x = ry. \tag{6}$$

Now suppose that B is about to play. Then the chance of B to win the game will be $q+sx$, where $s=(1-q)$, and

$$y = q + sx. \tag{7}$$

Solving these two equations, we find

$$x = \frac{rq}{1-rs}. \tag{8}$$

The example given by Huygens corresponds to the case where the player A needs a six and B needs a seven in the throw of two dice. This gives $p=5/36$ and $q=6/36$ from which follows $x=31/61$.

At the end of the treatise Huygens proposed five problems related to games chance. The fifth problem is a gambler's ruin problem and was stated as follows. Two players A and B each have 12 tokens and play with three dice. If 11 is thrown A gives a token to B, and if 14 is thrown B gives a token to A. The games continues until one of the players is in possession of all tokens. This problem was originally proposed by Pascal and Fermat. Huygens became aware of the problem by a letter he received from Carcavy in 1656. He immediately sent a letter to Carcavy with the solution. Latter in 1676, he elaborate a more satisfactory solution of the problem [39].

The Huygens treatise continued to be the best account on the subject until the beginning of the eighteen century when there appeared the treatises of Montmort in 1708, Jacob Bernoulli in 1713, and De Moivre in 1718 [7, 8].

5. Jacob Bernoulli

Jacob Bernoulli was one of the earliest supporter of the differential calculus in the form introduced by Leibniz, giving numerous contribution to this field, and he was also one of the proponents of the calculus of variations. His most original contribution was in the field of probability. His treatise on this field called *Ars Conjectandi* [40, 41] was published posthumously in 1713, but he had been writing the book from 1685 until his death in 1705 [3]. In this treatise we find his fundamental limit theorem which was later named by Poisson the law of large numbers [4].

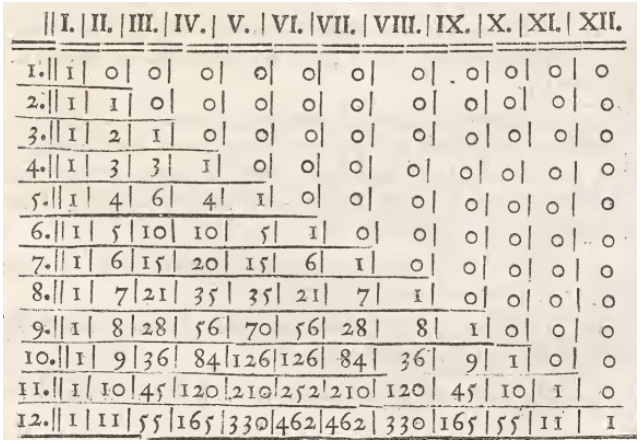
The treatise is divided in four parts. The first part contained the reprint of the Huygens treatise on the same subject along with extensive commentaries, other proofs of the main propositions and solutions of problems proposed by Huygens. The demonstration of the proposition fourteen of the Huygens treatise is as follows. The expectations that the game finish at rounds 1, 2, 3, and so on, is respectively p, rq, rsp, r^2sq, r^2s^2p, r^3s^2q, r^3s^3p, and so on, where p and q have the meaning given above and $r=1-p$ and $s=1-q$. The player A wins the game when the game finishes at the odd rounds and the player B, in the even games. Thus the expectation of the players A and B are, respectively, the given by the sums

$$p + rsp + r^2s^2p + r^3s^3p + \ldots, \tag{9}$$

$$rq + r^2sq + r^2s^2q + \ldots, \tag{10}$$

which are infinite geometric series with ratio rs. The sums of the series are $p/(1-rs)$ and $rq/(1-rs)$. The former result is the expectation of player A and the later result is the expectation of player B, which coincides with (8).

The second part contains the theory of permutations and combinations. Bernoulli starts with the



	I.	II.	III.	IV.	V.	VI.	VII.	VIII.	IX.	X.	XI.	XII.
1.	1	0	0	0	0	0	0	0	0	0	0	0
2.	1	1	0	0	0	0	0	0	0	0	0	0
3.	1	2	1	0	0	0	0	0	0	0	0	0
4.	1	3	3	1	0	0	0	0	0	0	0	0
5.	1	4	6	4	1	0	0	0	0	0	0	0
6.	1	5	10	10	5	1	0	0	0	0	0	0
7.	1	6	15	20	15	6	1	0	0	0	0	0
8.	1	7	21	35	35	21	7	1	0	0	0	0
9.	1	8	28	56	70	56	28	8	1	0	0	0
10.	1	9	36	84	126	126	84	36	9	1	0	0
11.	1	10	45	120	210	252	210	120	45	10	1	0
12.	1	11	55	165	330	462	462	330	165	55	11	1

Figure 5: Table of combinations or figurative numbers from the Bernoullli treatise [40].

permutations of a sequence of letters. For three letter a, b, and c, there are six permutations: abc, acb, bac, bca, cab, and cba. For four letters the number of permutations is 24. Generally, for n letters this number is $1{\cdot}2{\cdot}3{\cdot}4\cdots n$. If a certain a certain letter appears twice then the number of permutation is $1\cdot2\cdot3\cdot4\cdots n/2$. If in addition another letter appears three times then, $1\cdot2\cdot3\cdot4\cdots n/2\cdot6$.

The numbers in the table of combinations of Figure 5, or figurative numbers, are constructed as follows. We consider a certain number of rows and place in the first row the letter a. Then we write the letter b and adhere to this letter the previous letter a forming the second row: b, ba. Next, we write the letter c and adhere it to the terms of the two previous rows forming the third row: c, ca, cb, cba. The fourth row is obtained by writing the letter d and adhering it to the terms of *all* previous rows: d, da, db, dba dc, dca, dcb, $dcba$. This procedure is then repeated as many time as we wish. The figurative numbers in the table is obtained by counting for each row, the number of terms with one letter, two letters, three letters and so one. For the fourth row, one has 1 term with one letter, 3 terms with two letters, 3 terms with three letters, and just one term with four letters. In the Table 5, the number of letters in each term is denoted by a Roman numeral.

Next, Bernoulli demonstrates several properties of the figurative numbers. One of them is the formula for these number, given by

$$\frac{n(n-1)\cdots(n-c-1)}{1\cdot2\cdot3\cdot c}, \tag{11}$$

where n denotes the row and c the column. These numbers are understood as the combinations of n objects taken c at at time.

In the third part, consisting of 24 problems, Bernoulli deals with the games of chances, which illustrate the theory of combinations described in the second part. Problem five is the third problem of the Huygens treatise, which was solved in the first part but now Bernoulli solves it by using the theory of combinations. The problem is to pick up 4 cards from a deck of 40 cards such that there is one card of each suit. The number of choosing 4 objects out of 40 is $(40{\cdot}39{\cdot}38{\cdot}37)/(1{\cdot}2{\cdot}3{\cdot}4) = 91390$. Of these cases there are 10000 that meets the condition of the problem. Thus there are 10000 favorable cases and 81390 unfavorable cases.

Problems from 11 to 15 deal with throwing of dice. The problems from 16 to 24 concern known games: *cinq et neuf*, a kind of a roulette, *trijaques capriludium* or *Bockspiel*, *bassette*, and *blind Würffel*.

5.1. Limit theorem

The fourth part is the most important part of the book. In the last chapter of this last part, Bernoulli demonstrates his fundamental limit theorem. But before we analyze this theorem, it is useful to describe the following example given by Bernoulli, which is closely related to the theorem.

Suppose that a urn has 500 tokens, some of them are white and some are black. To estimate the number of tokens of each color, one takes out one token at a time, putting back the token before another one is taken out. This procedure is carried out as many times one wishes. The ratio between the numbers of times a white and a black token are chosen is equal to the ratio between the white and black token inside the urn. It clear that both ratios will not be exactly equal, but we expect that they become equal as the number of observation increases.

Next Bernoullli treats the problem that is called Bernoulli trial. Let us suppose that a certain outcome of an experiment occurs with probability p and fails to happen with the complementary probability $q = 1 - p$. If this experiment is repeated m times, we expect that the ratio k/m between the number of the favorable cases k and the total number m of cases is close to p. The Bernoulli fundamental theorem states that this is indeed the case and that the ratio k/m approaches the probability p as the number m increases indefinitely.

The demonstration given by Bernoulli is as follows. Let us write $p = r/t$ and $q = s/t$ where r and s are integers and $t = r + s$, and consider the binomial expansion

$$(r+s)^m = \sum_k P_k s^k r^{m-k}, \tag{12}$$

where m is equal to t times an integer n, that is, $m = nt$. The summation extends from zero to m, and

$$P_k = \frac{m(m-1)\cdots(m-k+1)}{1\cdot2\cdot3\cdots k}, \tag{13}$$

with $P_0 = 1$. If we compare the terms in the summation we see that they increase as k increases until k reaches

 DOI: https://doi.org/10.1590/1806-9126-RBEF-2022-0087

the value ns. From this value of k, they decrease as one increases k. Thus the largest term in the summation is the term corresponding to $k = ns$, namely the term

$$M = P_{ns} s^{ns} r^{nr}. \tag{14}$$

Bernoulli then consider the terms L and Λ that are a distance n from the left and to the right of the largest, respectively, which are given by

$$L = P_{ns-n} s^{ns-n} r^{nr+n}, \tag{15}$$

$$\Lambda = P_{ns+n} s^{ns+n} r^{nr-n}. \tag{16}$$

For sufficient large n the ratio M/L and M/Λ becomes

$$\frac{M}{L} = \left(\frac{rs+s}{rs-r}\right)^n, \tag{17}$$

$$\frac{M}{\Lambda} = \left(\frac{rs+r}{rs-s}\right)^n. \tag{18}$$

Therefore these ratios approaches infinity when n goes to infinity. Based on these results, Bernoulli argues that the sum A of the terms between L and Λ, including these two terms, divided by the remaining terms also approaches infinity.

If we denote by B the sum of all the terms, that is, $B = (r+s)^m$, the remaining terms equals $B - A$. Thus we may say that $A/(B-A)$ approaches infinity.

Each term appearing in the binomial expansion divided by the sum $B = (r+s)^m$ is the probability of the occurrence of k favorable cases in m cases. Let us consider the sum of the probabilities such that k is between $ns-n$ and $ns+n$, which equals A/B. But from the result shown above, the ratio $A/(B-A)$ approaches infinity which means that A/B approaches the unity. Therefore, the ratio k/m approaches p as the number of cases m increases indefinitely.

The fundamental theorem is to be understood as a probability calculation similar to those related to the throw of dice. The distinction here is that the number of dice is very large and for this reason there are some events whose chance of occurrence approaches the certainty. But the relevance of the theorem is that it gives an experimental method to find the probability or better to measure the probability. Let us suppose one wants to determine experimentally the probability p of the ace in a certain die. We throw the die a certain number m of times and count how many times the ace appeared. If we denote by k this number, then the ratio k/m will be an estimate of the probability p. Of course there is an experimental error, which decreases as one increases m.

5.2. Concept of probability

In the first four chapters of the fourth part of the treatise, Bernoulli discusses the concept of probability and proposes its use in practical problems such as the applications to civil, moral and economic problems.

Bernoulli starts by establishing the meaning of certainty and probability, necessity and contingency. He states that probability is degree of certainty, and differs from the latter as a part differs from the whole. We may represent certainty by the unity and probability by a fraction, which argues for the future existence of some outcome, and the complementary fraction against it. Something is necessary if it cannot not exist. A thing that may not exist is contingent. For Bernoulli, the things that are certain are said to be understood, but those that are not certain, we have to conjecture, which is the measure of its probability. Thus the art of conjecture, or stochastics, is the art of measuring probability.

It is interesting to note that Bernoulli considered the contingency as subjective. As an example he says that the eclipses are necessary, but before the principles of astronomy were known, they were contingent phenomena. On the other hand, the fall of dice or the future weather, which are contingent, could be considered necessary if we know all data which determine the subsequent effects.

As probability is related to conjecture, which is subjective, the probability for Bernoulli is subjective and depend on our knowledge. In fact, this idea of probability is reasonable if one wants, as Bernoulli did, to apply the theory to events that depend on opinions as those encountered in civil and moral contexts.

6. De Moivre

Jacob Bernoulli was not the only one that published solutions of the problems posed by Huygens. Among those who published solutions for some of these problems we find Montmort and de Moivre. Montmort wrote an essay where he analyzed the games of chance, which was published in 1708 [42]. A second enlarged edition appeared in 1713 [43]. The work contains the theory of combinations, and calculations of chance in card games and games of dice, in addition to the solution of various problems proposed by Huygens.

In 1711, De Moivre published a paper on probability [44] which was expanded and became a book called *Doctrine of Chances*, published in 1718 [45]. A second edition appeared in 1738 [46] and a third edition was published in 1756 [47]. The work of De Moivre was influenced by both Jacob Bernoulli and Montmort and he generalized and extended much of their work [8].

In the introduction of the book, De Moivre gives the meaning of the probability of an event as the number of chances that the event may happen compared to the number of all chances by which it may or not may happen. The basic rule of the theory concerns independent events. Let p be the number of chances of the occurrence of a certain event and q the number

of chances that it does not occur. Analogously, let r and s the numbers of chances of the occurrence and not occurrence of another event which is independent of the previous one. The product of $p+q$ by $r+s$ which is $pr + ps + qr + qs$ contains the chances of all four combinations of the two events. If two players A and B play a game such that A wins if the two independent events happen than the probability that A is the winner is $pr/(pr + ps + qr + qs)$.

This rule results in the method which is based on the expansion of the power of a binomial which in many cases solves the problem more easily then the method of combinations, says De Moivre. Suppose that the chance of a certain event to happen is a and to fail is b. If this is repeated n times then $(a+b)^n$ will given the chances of all possible cases. The chance that the event never occurs is the last term of the expansion, b^n, and the chance that the event happens in at least once is $(a+b)^n - b^n$. The chance that it fail to happen at most once equals the sum of the last two terms, $nab^{n-1} - b^n$, and the chance that it happens at least twice is $(a+b)^n - nab^{n-1} - b^n$. In all cases the corresponding probabilities are obtained by dividing these results by $(a+b)^n$.

Let us consider one of the problems treated by De Moivre, called duration of play. This problem is a generalization of the fifth problem proposed by Huygens in his treatise and is understood as a gambler's ruin problem. It was considered by Jacob Bernoulli in his treatise of 1713, and by Montmort in the first and in second edition of his book, where he stated that Nicolas Bernoulli sent to him a letter in 1711 with the solution of the problem. It was considered by De Moivre in his paper of 1711 and in the three editions of his book.

De Moivre formulates the problem as follows. Two players A and B have n tokens each and they play a game in several rounds. When a player wins a round he gets one token from the loser. The games ends when one of the player gets all the tokens of the other. The chance of A and B winning a round is a and b, respectively, which means that the probabilities are $p = a/(a+b)$ and $q = b/(a+b)$, respectively. The problem is to find the probability that the games ends in a certain number of rounds $n + d$.

Let us consider the case $n = 2$ and $d = 1$. There are 8 possible outcomes of the game in three rounds, which are represented by

$$\begin{array}{cccc} AAA & AAB & ABA & ABB \\ BAA & BAB & BBA & BBB \end{array}$$

where the letters from left to right indicate who is the winner in the successive rounds. The probabilities of each one of these outcomes are contained in the expansion

$$(p+q)^3 = p^3 + 3p^2q + 3pq^2 + q^3. \tag{19}$$

The player A wins in AAA, AAB, the player A wins in BBA, BBB, and neither player wins in ABA, ABB, BAA, BAB. The probabilities of these three events are, respectively $p^3 + p^2q$, $pq^2 + q^3$, and $2pq^2 + 2p^2q$. This last result can be obtained by calculating $(a+b)^2$, rejecting the extremes a^2 and b^2, and dividing the result by $(a+b)^2$.

Let us consider the case $n = 2$ and $d = 2$. Now there 16 possible outcomes which are represented by

$$\begin{array}{cccc} AAAA & AAAB & AABA & AABB \\ ABAA & ABAB & ABBA & ABBB \\ BAAA & BAAB & BABA & BABB \\ BBAA & BBAB & BBBA & BBBB \end{array}$$

The probabilities of these outcomes are contained in the expression

$$(p+q)^4 = p^4 + 4p^3q + 6p^2q^2 + 3pq^3 + q^4. \tag{20}$$

There is no winner in cases $ABAB$, $ABBA$, $BAAB$, $BABA$, $BBAA$, which corresponds to the probability $4p^2q^2$. This result can be obtained by using the rules. Raise $(a+b)$ to the second power and reject the extremes a^2 and b^2. Multiply the result by $(a+b)^2$, which gives $2a^3b + 4a^2b^2 + 2ab^3$, and the extremes of this last expression. Divide the result by $(a+b)^4$ to obtain $4a^2b^2/(a+b)^4$.

The rules used above are particular case of the general rule put forward by De Moivre, which is stated as follows [7]. Let consider first the case where d is even, that is, $d = 2\ell$. Calculate $(p+q)^n$ an reject the two terms p^n and q^n. Multiply the result by $(p+q)^2$ and reject the extreme terms. Repeat the last procedure $\ell - 1$ times. Finally divide the result by $(a+b)^{n+2\ell}$. If d is odd, we first determine ℓ by $d = 2\ell + 1$ and use the same rule, which amounts to say that the desired probability corresponding to $d = 2\ell + 1$ is the same as that corresponding to $d = 2\ell$.

6.1. Approximation to the binomial expansion

In the second and third editions of his Doctrine of Chances, De Moivre added a section where he presented an approximation for the terms of a binomial expansion which leaded him to an expression which we recognize as the normal curve. The added section was essentially a translation from Latin of a note he had written earlier, which had appeared as a printed pamphlet in 1733 [48–50]. In this note, De Moivre considers the middle term of the expansion of $(1+1)^n$ which is $n!/[(n/2)!]^2$. He states that this term divided by 2^n is

$$\frac{2}{B\sqrt{n}}, \tag{21}$$

where

$$\ln B = 1 - \frac{1}{12} + \frac{1}{360} - \frac{1}{1260} + \frac{1}{1680} \ldots \tag{22}$$

To reach this expression, he used an asymptotic expansion for the factorial that appeared as a supplement of a book he published in 1730 [7]. In the

 DOI: https://doi.org/10.1590/1806-9126-RBEF-2022-0087

supplement, De Moivre arrives at a result for $\ln(m-1)!$, which we write in the form

$$\left(m-\frac{1}{2}\right)\ln m - m + a + b, \tag{23}$$

where

$$a = \frac{1}{12m} - \frac{1}{360m^3} + \frac{1}{1260m^5} - \frac{1}{1680m^7} + \ldots, \tag{24}$$

$$b = 1 - \frac{1}{12} + \frac{1}{360} - \frac{1}{1260} + \frac{1}{1680}\ldots. \tag{25}$$

For large values of m, we may neglect a but not b. De Moivre obtained a numerical value for b, but he became aware of a similar asymptotic expansion by Stirling, published in 1730 [51]. The Stirling series for $\ln(z-1/2)$ is

$$z\ln z - z + \ln\sqrt{2\pi} - \frac{1}{2\cdot 12z} + \frac{7}{8\cdot 360z^3} - \frac{31}{32\cdot 1260z^5} + \ldots. \tag{26}$$

Comparing the two expressions for the logarithm of the factorials, De Moivre concludes that $b = \ln\sqrt{2\pi}$ and that $B = \sqrt{2\pi}$. The expression (21) becomes

$$\frac{2}{\sqrt{2\pi n}}. \tag{27}$$

Next, De Moivre determines the term distant from the middle term by an interval ℓ, which is

$$\frac{n!}{(n/2+\ell)!(n/2-\ell)!}. \tag{28}$$

Using the above approximation for the factorial he arrives at the following expression

$$e^{-2\ell^2/n}, \tag{29}$$

for the ratio between the above term and the middle term. Although De Moivre understood this result as an expression of a geometrical curve, one cannot draw the conclusion that he understood it as a probability density function [10].

De Moivre added the general result valid for the case of the expansion of $(a+b)^n$. In this case the maximum term divided by $(a+b)^n$ is

$$\frac{a+b}{\sqrt{2\pi abn}}, \tag{30}$$

and the term distant from from the maximum by an interval ℓ, divided by the maximum term, is

$$e^{-(a+b)^2\ell^2/2abn}. \tag{31}$$

They reduce to the results above when $a = b = 1$.

7. Daniel Bernoulli

The contribution of Daniel Bernoulli to the theory of probability is contained in several papers where he dealt with games of chance, astronomy, theory of errors, and the urn models [52]. One of the problems he treated was the so called Petersburg problem. He also was interested in the economic and demographic statistics such as those related to the smallpox and inoculation, duration of marriages and relative frequency of male and female births.

His investigations of the urn models be identified in retrospect as a Markov process. However, his equations involved only the evolution of the averages on the number of ball in each urn and not the probability itself. The evolution of the probability was later considered by Laplace. It is worth mentioning in addition that a similar urn model was used by Ehrenfest in 1907 to illustrate the approach to equilibrium of a thermodynamic system [53].

The urn model was introduced and discussed by Bernoulli in paper published in 1770 [52]. There are several urns each one with the same number n of balls of different colors. An operation which Bernoulli calls a permutation, in fact a cyclic permutation, is carried as follows. A ball is taken out at random from each one of the urns. After that, the ball taken from the first urn is placed on the second urn, that taken from the second is place on the third, and so on until the ball taken from the last urn is placed on the first urn. This operation is repeated several times. The number of balls of each color is n and the number of different colors is equal to the number of urns. Initially the balls in each urn have the same color, the first urn with all balls of the white color. One wishes to determine the average number of white balls in each urn after a certain number of operations.

Bernoulli first solves the case of two urns by determining the average number x of white balls in the first urn. He solves the problem for a small number of operations and then generalizes the result for r operations, which is

$$x = \frac{n}{2}(1+m^r), \tag{32}$$

where $m = 1-2/n$. The average number y of white balls in the second urn is

$$y = \frac{n}{2}(1-m^r) \tag{33}$$

because $x + y = n$.

For the case of three urns, the balls are of three colors one of them being white. Bernoulli determines the numbers x, y, and z of white balls in each urn by writing the solution for small number of operations. He writes down explicit solutions for x, y, and z up to $r = 9$. He then perceives that the expressions of x, y, and z contains terms that are similar to those of the expansion of $(a+b)^r$, where $a = 1-1/n$ and $b = 1/n$. In modern

notation the expansion is

$$(a+b)^r = \sum_{j=0}^{r} \frac{r!}{(r-j)!j!} a^{r-j} b^j. \tag{34}$$

The solution for x contains the terms of this expansion corresponding to j equal to a multiple of 3, that is,

$$x = n \sum_{j} \frac{r!}{(r-j)!j!} a^{r-j} b^j, \tag{35}$$

where the summation is over j equal to $0, 3, 6, \ldots$. The same expressions is valid for y and z except that the summation is over $j = 1, 4, 7, \ldots$ and $j = 2, 5, 8, \ldots$, respectively.

For the case of two urns the solution for x is also given by (35) except that the summation is over j even, that is, $j = 0, 2, 4, \ldots$ The solution for y, the expression is the same, except that the sum is over $j = 1, 3, 5, \ldots$ To see that this general solution agrees with x and y given by (32) and (33), we use the results $a-b = m$ and $a+b = 1$ to write (32) and (33) in the form

$$x = n\frac{1}{2}[(a+b)^r + (a-b)^r], \tag{36}$$

$$y = n\frac{1}{2}[(a+b)^r - (a-b)^r]. \tag{37}$$

It is clear that these expressions agree with the general solution.

After that, Bernoulli used a continuous approach, which is appropriate for large values of n. For the case of two urns, he writes the variation in x as

$$dx = -\frac{x}{n}dr + \frac{n-x}{n}dr, \tag{38}$$

where the first term corresponds to the removing of a white ball whereas the second corresponds to the introducing of a white ball. The solution is obtained by writing this equation in the form

$$\frac{dx}{2x-n} = -\frac{dr}{n}, \tag{39}$$

from which follows the solution

$$x = \frac{n}{2}(1 + e^{-2r/n}). \tag{40}$$

The continuous solution for the case of three urns, the equations for the variation in the average numbers of white balls in the first and second urns are

$$dx = -\frac{x}{n}dr + \frac{z}{n}dr, \tag{41}$$

$$dy = -\frac{y}{n}dr + \frac{x}{n}dr. \tag{42}$$

The equation for z is not necessary because $z = n - x - y$. Bernoulli solves this set of differential equations and finds

$$x = \frac{n}{3} + \frac{2n}{3}\cos\frac{r\sqrt{3}}{2n}e^{-3r/2n}, \tag{43}$$

$$y = \frac{n}{3} + \frac{n}{\sqrt{3}}\sin\frac{r\sqrt{3}}{2n}e^{-3r/2n} - \frac{n}{3}\cos\frac{r\sqrt{3}}{2n}e^{-3r/2n}. \tag{44}$$

After an infinite number r of operations, the average number of white balls in each urn reaches the asymptotic value $n/3$.

It is worth mentioning that Bernoulli makes an analogy of the urn model with the mixing of two fluids. If the number of balls is large enough the urns may be understood as vessels containing distinct fluids. The transfer of a ball from one urn to the other corresponds to placing a communication between one vessel to another.

8. Lagrange

Lagrange is most remembered for his fundamental treatise on analytical mechanics and for his works on mathematical analysis. He also published some papers on the theory of probability. We examine here the one related to the theory of errors and the multinomial distribution published in 1776 [54]. The paper is written in the form of ten problems.

Let us consider an experiment with three possible outcomes occurring with chances a, b, and c. The experiment is repeated n times and one wishes to determine the probability that each one of the three outcomes occur with a certain number of times which we denote by i, j, and k, with $i+j+k = n$. Next we consider the expansion

$$(a+b+c)^n = \sum_{ijk} \frac{n!}{i!j!k!} a^i b^j c^k, \tag{45}$$

where the summation over i, j, and k is carried out with the restriction $i+j+k = n$. The desired probability is identified with each one of the terms in the expansion, divided by $(a+b+c)^n$. We are using the modern notation $n!$ for the factorial and a compact form in terms of the capital sigma letter, which were not used by Lagrange.

Lagrange treats the case where the outcomes are the errors of an observation. The errors are zero with chance a and $+1$ with chance b and -1 with equal chance $c = b$. After n repetition, the total error will be $m = j-k$. The problem is to determine the probability of the occurrence of the error m after n repetitions. To solve this problem Lagrange starts by writing

$$(a + bx + bx^{-1})^n = \sum_{ijk} \frac{n!}{i!j!k!} a^i b^{j+k} x^m, \tag{46}$$

where $m = j-k$. The desired probability is related to the coefficient of x^m which is given by

$$A = \sum_{ijk} \frac{n!}{i!j!k!} a^i b^{j+k}, \tag{47}$$

 DOI: https://doi.org/10.1590/1806-9126-RBEF-2022-0087

where the sum is restricted to $i+j+k=n$ and $j-k=m$, and can be written in the form

$$A=\sum_j \frac{n!}{(n+m-2j)!j!(j-m)!}a^{n+m-2j}b^{2j-m}, \quad (48)$$

where the summation in j is restricted to $j \geq m$ and $j \leq (n+m)/2$. Or

$$A=\sum_\ell \frac{n!}{(n-m-2\ell)!(\ell+m)!\ell!}a^{n-m-2\ell}b^{2\ell+m}, \quad (49)$$

where the summation in ℓ is restricted to $\ell \geq 0$ and $\ell \leq (n-m)/2$. When $m=0$, which corresponds to no errors,

$$A=\sum_j \frac{n!}{(n-2j)!j!j!}a^{n-2j}b^{2j}, \quad (50)$$

where the summation in j is restricted to $j \geq 0$ and $j \leq n/2$. In a explicit form written by Lagrange

$$A=a^n+n(n-1)a^{n-2}b^2+\frac{n(n-1)(n-2)(n-3)}{2\cdot 2}a^{n-4}b^4 + \frac{n(n-1)(n-2)(n-3)(n-4(n-5)}{2\cdot 3\cdot 2\cdot 3}a^{n-6}b^6+\ldots. \quad (51)$$

Lagrange generalizes the problem as follows. The errors are $p,q,r,s,\ldots$, and the chances are $a,b,c,d,\ldots$, respectively. On this case we have to consider the expansion of

$$(ax^p+bx^q+cx^r+dx^s+\cdots)^n, \quad (52)$$

in powers of x. The probability of the sum of the errors m is the coefficient of x^m in the expansion, divided by $(a+b+c+d+\ldots)^n$. The seventh problem that Lagrange proposes to solve corresponds to the case where the errors are $-\alpha,\ldots,-2,-1,0,1,2,\ldots,\beta$ and the probabilities are all equal. This problem was the same as that proposed by de Moivre and continued by Simpson [7].

In the eighth problem, the errors are $-\alpha,\ldots,-2,-1,0,1,2,\ldots,\alpha$, which occurs with chances $1,2,\ldots,\alpha,\alpha+1,\alpha,\ldots,2,1$. In this case we have to consider the expansion of the n power of

$$x^{-\alpha}+2x^{-\alpha+1}+\ldots+(\alpha+1)+\alpha x+\ldots+2x^{\alpha-1}+x^\alpha, \quad (53)$$

in powers of x, and it is required to find the coefficient of x^m. This expression can be simplified and written in the form

$$\frac{x^{-\alpha}(x^{\alpha+1}-1)^2}{(1-x)^2}. \quad (54)$$

Therefore, the desired coefficient is obtained from the expansion of $(x^{\alpha+1}-1)/(1-x)$ raised to the power $2n$.

In the final part, Lagrange treats the case where the errors are continuous in a certain interval. Denoting by ℓ the errors, then the summation above becomes the integral

$$\int_{-\alpha}^{\alpha} p(\ell)x^\ell d\ell, \quad (55)$$

and we seek for the z power of x in the integral above raised to the n power. Lagrange show that the n power of this integral can be written as

$$\int_{-n\alpha}^{n\alpha} fx^z dz, \quad (56)$$

where f is a function of z and is the desired coefficient of x^z.

9. Laplace

Laplace is well known for his superlative contributions to physical sciences which are exposed in his treatise on celestial mechanics. Laplace also gave a relevant contribution to probability which was presented in his treatise on analytical theory of probability published in 1812 [13]. A second edition was published in 1814 and a third edition in 1820. Laplace also wrote a philosophical essay on probability, published in 1814 [14, 15], where he presented his concept of probability and causal determinism. The essay became an introduction to the second and third editions of the treatise.

The treatise on probability is divided in two books and is largely based on his papers on the subject which he published for about fifteen years starting from 1774. In the first book, Laplace is concerned with the development of some mathematical techniques that he used in book two. In the first part of the first book, he develops the method of generating function. If f_n a function of n, then the generating function of f_n is defined by

$$u=\sum_{n=0}^{\infty} f_n t^n. \quad (57)$$

A similar definition is made for the case of two variables.

In the second part of the first book he develops approximative methods for the integration of functions containing factors raised to higher powers. An example is the integral

$$\int y dx, \quad (58)$$

where $y=\phi u^n$ with ϕ and u some functions of x and n a certain power. It is assumed that y has a maximum value in the interval of integration. Assuming that $y=Ye^{-t^2}$ where t is a function of x and Y is the maximum value of y, then the integral becomes

$$Y\int e^{-t^2}\frac{dx}{dt}dt. \quad (59)$$

We remark that throughout the treatise, Laplace uses the letter c and not e as we are doing. Assuming that

$$x = B_0 + B_1 t + B_2 t^2 + B_3 t^3 + B_4 t^4 + B_5 t^5 + \ldots, \quad (60)$$

then the integral becomes

$$Y \int e^{-t^2} (B_1 + 2B_2 t + 3B_3 t^2 + 4B_4 t^3 + 5B_5 t^4 + \ldots) dt. \quad (61)$$

Supposing that this integral extends from minus infinity to infinity, the terms corresponding to odd power of t vanishes and the integration of the other terms gives

$$Y\sqrt{\pi}\left(B_1 + \frac{3}{2}B_3 + \frac{5\cdot 3}{2^2}B_5 t^4 + \ldots\right). \quad (62)$$

Laplace shows the following result

$$\int e^{-t^2} dt = \frac{\sqrt{\pi}}{2}, \quad (63)$$

where the integral extends from zero to infinity. The demonstration of this result starts by considering the double integral

$$\int\int e^{-s(1+x^2)} ds dx = \int \frac{dx}{1+x^2} = \frac{\pi}{2}, \quad (64)$$

where both variable take values from zero to infinity, and we have integrated first in s and then in x. Now one defines $t = \sqrt{s}x$ and change variable from x to t. The double integral becomes

$$\int\int e^{-s-t^2} \frac{ds}{\sqrt{s}} dt = \int e^{-s} \frac{ds}{\sqrt{s}} \int e^{-t^2} dt. \quad (65)$$

The first integral on the right-hand side is equal to two times the second integral. Therefore, the double integral, which is equal to $\pi/2$, is equal to two times the integral that we wish to determine, from which follows the equality (63).

Next, Laplace shows the following result

$$\int \cos rx e^{-a^2x^2} dx = \frac{\sqrt{\pi}}{2a} e^{-r^2/4a^2}, \quad (66)$$

where the variable x extends from zero to infinity. The integral is equal to

$$\frac{1}{2}\int e^{-a^2x^2+irx} dx + \frac{1}{2}\int e^{-ax^2-irx} dx. \quad (67)$$

We are using the modern notation i where Laplace uses $\sqrt{-1}$. Making the change of variables $t = ax - ir/2a$ in the first and $t = ax - ir/2a$ in the second, the integral becomes

$$\frac{1}{a} e^{-r^2/4a^2} \int e^{-t^2} dt = \frac{\sqrt{\pi}}{2a} e^{-r^2/4a^2}, \quad (68)$$

where we have used the result (63).

It should be noted that the last integral involves values imaginaries of the variable t. To circumvent this inconvenience, Laplace uses a distinct reasoning to justify the correctness of the result (66). Denoting the integral in (66) by y and deriving it with respect to r, we find

$$\frac{dy}{dr} = -\frac{ry}{2a^2}, \quad (69)$$

where we have performed in integration by parts. The integration of this equation gives

$$y = Be^{-r^2/4a^2}. \quad (70)$$

The constant of integration is found by considering that when $r = 0$, the integral in (66) equals $\sqrt{\pi}/2a$ by the use of (63).

In the second book, Laplace considers several problems in probability. We start with the problems of the division of a stake known as the problem of points, treated in the chapter 2. Two players A and B need to win n and m rounds, respectively, to win the game. Their probabilities of winning a round is p and q, respectively, with $p + q = 1$. Laplace solves the problem by means of the generating function. Let us denote by P_{nm} the probability that player A will reach first the number of points n necessary for winning the game. In the next round, if he wins the round the probability becomes $P_{n-1,m}$, and if he loses, it becomes $P_{n,m-1}$. Since his chance of winning the next round is p and of losing is q, then

$$P_{nm} = pP_{n-1,m} + qP_{n,m-1}, \quad (71)$$

where $n, m = 0, 1, 2, \ldots$. The following conditions should be taken into account: $P_{n0} = 0$ because when $m = 0$, the player B wins the game and A loses, and $P_{0m} = 1$ because in this case A wins the game. The generating function by

$$g = \sum_{nm} P_{nm} x^n y^m. \quad (72)$$

Multiplying (71) by $x^n y^m$ and summing in n and m we find an equation such that the left hand side equals g and the right hand side contain two summations. In one of them we change the summation variable from n to $n + 1$ and in the other we change from m to $m + 1$. By this procedure, the right hand side becomes equal to $f + pxg + qyg$ from which we get

$$g = \frac{f}{1 - px - qy}, \quad (73)$$

where f is a function of y only. To determine f, we observe that when $x = 0$ the generating function becomes $y/(1 - y)$ because $P_{0m} = 1$. Therefore,

$$f = \frac{y(1 - qy)}{1 - y}, \quad (74)$$

 DOI: https://doi.org/10.1590/1806-9126-RBEF-2022-0087

and

$$g = \frac{y(1-qy)}{(1-y)(1-px-qy)}. \tag{75}$$

Expanding g in power of x,

$$g = \frac{y}{1-y}\sum_n \frac{p^n x^n}{(1-qy)^n}. \tag{76}$$

Expanding again in powers of y, we find the coefficient of $x^m y^m$ to be

$$P_{nm} = \sum_{j=0}^{m-1} \frac{n(n+1)\cdots(n+j-1)}{1\cdot 2\cdot 3\cdots j}. \tag{77}$$

In chapter 3 of the second book Laplace considers the problem related to the Bernoulli fundamental theorem. In a certain trial, an event occurs with a certain probability p and fails with probability $q = 1-p$. The probability that the events occur j times in n trials is

$$P_{jk} = \frac{n!}{j!k!}p^j q^k, \tag{78}$$

where $k = n - j$. The problem is to calculate the probability that the number of events occurring in n trials lie within a certain interval.

Laplace performed the calculation by considering that the number of trial is large enough as to use the formula

$$n! = n^{n+1/2}e^{-n}\sqrt{2\pi}\left(1 + \frac{1}{12n} + \ldots\right). \tag{79}$$

Using this approximation one finds

$$P_{jk} = \frac{1}{\sqrt{2\pi}}\frac{n^{n+1/2}p^j q^k}{j^{j+1/2}k^{k+1/2}}. \tag{80}$$

The largest value of the terms P_{jk}, is the one corresponding to j equal to the integer as nearly as possible equal to pn. Defining m as the deviation of j from this number, and expanding P_{jk} around $m = 0$ one finds

$$P_{jk} = \frac{e^{-m^2/2pqn}}{\sqrt{2\pi pqn}}. \tag{81}$$

Now we determine the probability that j is within the interval between $pn - \ell$ and $pn + \ell$, or the m is in the interval between $-\ell$ and ℓ. Thus we have to sum the expression from $m = -\ell$ until $m = \ell$. Laplace replaces the sum by an integral plus a correction term

$$\frac{2}{\sqrt{\pi}}\int_0^\tau e^{-t^2}dt + \frac{e^{-\tau^2}}{\sqrt{2\pi pqn}}, \tag{82}$$

where $\tau = \ell/\sqrt{2pqn}$. Thus the above expression gives the probability that the fraction of successful events in n trials is between

$$p - \tau\sqrt{2pq/n} \qquad \text{and} \qquad p + \tau\sqrt{2pq/n}. \tag{83}$$

Taking into account that the expression (82) approaches the unity even for moderate values of τ, we may conclude that the deviation of the fraction of successful events from p is of the order $1/\sqrt{n}$.

In the same chapter 3 of the second book, Laplace treats the following problems of urns. Let us consider two urns A and B each one with n balls of two colors, white and black, with the same number of balls of each color. A ball is drawn from each urn and placed in the other urn. The operation is repeated a certain number of times and one asks for the probability that the urn A has ℓ white balls after r number of operations.

Let $P_{\ell r}$ be this probability. Laplace argues that this probability fulfills the equation

$$P_{\ell,r+1} = a_{\ell+1}^2 P_{\ell+1,r} + 2a_\ell b_\ell P_{\ell r} + b_{\ell-1}^2 P_{\ell-1,r}, \tag{84}$$

where $a_\ell = \ell/n$, and $b_\ell = 1 - \ell/n$. This equation gives the probability at any value of r if it is known at $r = 0$. Laplace transforms this equation into a partial differential equation by considering that n is large. To this end, he defines the variables x and t by $\ell = (n+x\sqrt{n})/2$ and $r = nt$. Performing the expansion of in $1/n$ and neglecting terms of order $1/n^2$, the equation for $U(x,t) = P_{\ell r}$ becomes

$$\left(\frac{dU}{dt}\right) = 2U + 2x\left(\frac{dU}{dx}\right) + \left(\frac{d^2U}{dx^2}\right). \tag{85}$$

To solve this equation, Laplace uses the transformation

$$U = \int e^{-xz}\phi dz, \tag{86}$$

which he had introduced in 1773, known today as the Laplace transformation. After deriving an equation for ϕ and solving it, he finds the following expression for U,

$$U = \frac{2}{\sqrt{n\pi(1+\beta)}}e^{-x^2/(1+\beta)}, \tag{87}$$

where $\beta = \beta_0 e^{-4t}$. That this is ideed a solution can be verified by a direct substitution in the differential equation for U. When t becomes infinity, U approaches the value

$$U = \frac{2}{\sqrt{n\pi}}e^{-x^2}. \tag{88}$$

Chapter 4 of the second book is the most important in the Laplace work [7]. It deals with the probability of errors in the mean of the results obtaining from a great number of observations. He shows that the distribution of the mean is normal with a standard deviation proportional to the inverse of the square root of the number of ... n. He also deals with the method of least squares.

Let us suppose that the possible errors of an observation are $a_1, a_2, \ldots$ occurring with probabilities p_i. One

wishes to determine the probability distribution of the sum errors in n observations. The desired probability is obtained by determining the n-th power of

$$X = p_1 e^{ika_1} + p_2 e^{ika_2} + \ldots, \tag{89}$$

and finding the coefficient of $e^{ik\ell}$ of X^n, where ℓ denotes the sum of the errors in n trails.

Laplace considers the errors to be $-s, -s+1, \ldots, s-1, s$ occurring with the same probability equal to $1/b$ where $b = 2s+1$. We have to consider

$$X = \frac{1}{b}(e^{-iks} + \ldots + e^{ik(s-1)} + e^{iks}), \tag{90}$$

which is equal to

$$X = \frac{\sin kb/2}{b \sin k/2}. \tag{91}$$

To determine the coefficient of $e^{ik\ell}$ in this expression Laplace multiply X by $e^{-ik\ell}/\pi$ and integrate in k from zero to π. By this procedure all terms will vanish except the one we want. Denoting by P_ℓ this coefficient

$$P_\ell = \frac{1}{\pi}\int e^{-ik\ell} X^n dk = \frac{1}{\pi}\int \cos k\ell X^n dk. \tag{92}$$

Since X has a maximum at $k = 0$, then X^n is very sharped at $k = 0$ which means that the integral comes from its values around $k = 0$. The expansion of X around $k = 0$ gives

$$X = e^{-k^2 s(s+1)/6}, \tag{93}$$

and the integral becomes

$$P_\ell = \frac{1}{\pi}\int \cos k\ell \, e^{-nk^2 s(s+1)/6} dk. \tag{94}$$

Using the result (66) one reaches the result

$$P_\ell = \frac{\sqrt{3} e^{-3\ell^2/2ns(s+1)}}{\sqrt{2\pi ns(s+1)}}. \tag{95}$$

The distribution of errors (95) became known as normal distribution and sometimes as Gaussian distribution since Gauss also considered it. Pearson proposed to call it Laplace-Gaussian curve to avoid the word normal, which would imply that the other curves are abnormal [55].

The method used by Laplace to reach the distribution (95) is explained in moder terminology as follows. First one construct the characteristic function related to one trial, which is the Fourier transform of the probability distribution, given by (90). The distribution of the sum of errors is the probability whose characteristic function is X^n.

In the philosophical essay on probability, Laplace stated that *the theory of chance consists in reducing all the events of the same kind to a certain number of cases equally possible*, and *in determining the number of cases favorable to the event whose probability is sought. The ratio of this number to that of all cases possible is the measure of probability*. The first part corresponds to the construction of a sample space whose elementary events have equal probability. The second part gives the rule to determine the probability of an event by considering how many elementary events compose that event. The word measure here is not used in the sense of physical measurement but simply as numerical value.

10. Poisson

Poisson made contributions to several areas of the physical sciences and also to the theory of probability. His major work on this field was a book published in 1837 [4]. A large part of the book is a treatise on probability with emphasis on events with a large number of trials. The other part deals with the application of the theory to the judgment in criminal and civil matters.

The definitions and rules of the theory are stated by Poisson as follows. (1) Probability of an event is the motif that we have to believe that it will take place or takes place. (2) The measure of the probability of an event is the ratio of the number of favorable cases and the total number of cases. If an urn has four white balls and six black balls and the other has ten white balls and fifteen black balls, then the probability of white balls is the same for both urns and equal 2/5. (3) The sum of the probability p of an event E and the probability q of the contrary event F is equal to the unity, $p + q = 1$. (4) The certitude of an event is represented by a probability equal to the unity. (5) If the probabilities of two independent events are p and p', then the probability of their concurrence or of an event composed by the two events is equal to the product pp'. (7) The probability that an event E occurs m times in a row is p^m. (8) The probability that an event E occurs at least once in m trials is $1-(1-p)^n$. (9) If two events E and E' are not independent, that is, the arrival of an event influences the other, then the probability of the compound events E and E' is equal to the product pp' where p' is the probability that if E has arrived than E' will come next. (10) If an event takes place in several distinct and independent ways the total probability is the sum of the probabilities of the ways of occurrence.

If E and F are contrary events, occurring with probabilities p and q, with $p + q = 1$, then the probability that the occurrence of m events E and n events F in any order is

$$\frac{\mu!}{m!n!} p^m q^n, \tag{96}$$

where $\mu = m + n$ and we are using the moder notation for the factorial. They correspond to the terms of the development of $(p + q)^\mu$ in powers of p and q. The generalization for three independent events E_1, E_2, and

 DOI: https://doi.org/10.1590/1806-9126-RBEF-2022-0087

E_3 occurring with probabilities p_1, p_2, and p_3

$$\frac{\mu}{n_1!n_2!n_3}p_1^{n_1}p_2^{n_2}p_3^{n_3}, \tag{97}$$

which is the probability of the occurrence of n_1, n_2, and n_3 times the events E_1, E_2, and E_3, respectively, independent of the order. A similar expression is valid for more than three independent events, and we are using the modern notation for the factorial.

If in an event E one gets a quantity g, in an event E' a quantity g' and so on, the mathematical expectation is

$$gp + g'p' + g''p'' + \ldots, \tag{98}$$

where p, p', and son on are the probabilities of the events.

To find the asymptotic expansion of the binomial distribution

$$U = \frac{\mu!}{m!n!}p^m q^n. \tag{99}$$

Poisson uses the expansion

$$n! = n^n e^{-n}\sqrt{2\pi n}\left(1 + \frac{1}{12n} + \frac{1}{288n^2} + \ldots\right). \tag{100}$$

Using this expansion, one finds the following approximation for U

$$U = \left(\frac{p\mu}{m}\right)^m \left(\frac{q\mu}{n}\right)^n \sqrt{\frac{\mu}{2\pi mn}}, \tag{101}$$

valid for large values of n, m, and μ. Poisson remarks that the maximum value of the probability U occurs when $p = m/\mu$ and $q = n/\mu$.

Defining g such that

$$m = p\mu + g\sqrt{pq\mu}, \qquad n = q\mu - g\sqrt{pq\mu}, \tag{102}$$

we reach the following expression for U

$$U = \frac{e^{-g^2/2}}{\sqrt{2\pi pq\mu}}. \tag{103}$$

In the derivation of of the asymptotic expansion for the binomial distribution, it was assumed that p and q are not small quantities. Poisson then consider the case were one of them, say q, is very small so that $\omega = q\mu$ cannot be considered a large quantity. Accordingly, we consider n finite and μ large so that n/μ is a small fraction. Thus, writing

$$\frac{\mu!}{(\mu - n)!} = \mu(\mu - 1)\cdots(\mu - n + 1), \tag{104}$$

we see that it can be approximated by μ^n and the binomial distribution (99) becomes

$$U = \frac{\omega^n}{n!}e^{-\omega}, \tag{105}$$

where, within the same approximation, we have replace $p^m = (1 - q)^m$ by $e^{-\omega}$. This is known as the Poisson distribution. In fact, Poisson derives the distribution in its cumulative form

$$P = (1 + \omega + \frac{\omega^2}{2!} + \frac{\omega^3}{3!} + \cdots + \frac{\omega^n}{n!})\, e^{-\omega}. \tag{106}$$

Let us consider a series of trials each of which an event succeeds with probability p and fails with probability q and let

$$X = pe^{ix} + qe^{-ix}. \tag{107}$$

Poisson writes $\sqrt{-1}$ for the imaginary unity. In a certain number μ of trials, the probability U of the occurrence m successful events and n falling events is the coefficient of $e^{ix(m-n)}$ in the expansion of X^μ. It is obtained by

$$U = \frac{1}{2\pi}\int_{-\pi}^{\pi} X^\mu e^{-ix(m-n)}dx. \tag{108}$$

Expressing X in the form $X = Ye^{iy}$, where

$$Y^2 = 1 - 4pq\sin^2 x, \tag{109}$$

and $\tan y = (p - q)\tan x$, then

$$U = \frac{1}{2\pi}\int_{-\pi}^{\pi} Y^\mu \cos[\mu y - x(m - n)]dx. \tag{110}$$

When μ is large, Y is very small except for small values of x. Thus we may use the approximation $\ln Y = -2pqx^2$ and the integral above becomes

$$U = \frac{1}{2\pi}\int e^{-2pq\mu x^2}\cos gx dx, \tag{111}$$

where we have also approximated y in the argument of the cosine by $y = (p-q)x$ and we used the abbreviation $g = \mu(p-q) - (m-n)$. The integration is extended from minus infinity to plus infinity because the integrand is negligible outside the interval where x is small. Carrying out the integral we find

$$U = \frac{e^{-g^2/2(4pq\mu)}}{\sqrt{2\pi 4pq\mu}}. \tag{112}$$

Defining $k^2 = 2pq$

$$U = \frac{e^{-g^2/(4k^2\mu)}}{\sqrt{\pi 4k^2\mu}}. \tag{113}$$

11. Curve of Errors

The method of least squares is a well known and widely used technique in statistical analysis. It consists in minimizing the square of the difference between the observed value and the value provided by a model. The method was invented by Legendre and presented in 1805 [56] as an appendix to his book concerning the

determination of the orbits of comets. Legendre writes the errors as

$$E = a + bx + cy + \ldots, \tag{114}$$

$$E' = a' + b'x + c'y + \ldots, \tag{115}$$

where $x, y, \ldots$ are unknown quantities and their coefficients in all equations are known. To find the unknown quantities, Legendre proposes to minimize the square of the errors

$$E^2 + E'^2 + \ldots, \tag{116}$$

by varying the unknown quantities. This is obtained by setting to zero the derivative of this expression with respect to the unknown quantities, leading to a set of linear equations to be solved.

A similar approach was used by Gauss to find the approximate orbit of a planet from observed data. His approach appeared in his book on the motion of bodies around the sun, published in 1809 [57, 58]. The distinction feature of the Gauss approach was an explicit use of a probabilistic reasoning [10]. He regarded the errors $\Delta, \Delta', \ldots$ as distributed according to the same distribution $\varphi(\Delta), \varphi(\Delta'), \ldots$ and sought to maximize the product

$$\Omega = \varphi(\Delta)\varphi(\Delta')\ldots. \tag{117}$$

Gauss then argued that distribution should be

$$\frac{h}{\sqrt{\pi}} e^{-h^2\Delta^2}, \tag{118}$$

which as we have seen above had been considered by Laplace. The derivation of the error distribution (118) was also provided by Adrain in 1808 using a reasoning similar to that of Gauss [11].

In a publication of 1823, [59–63] Gauss improved the method of least squares and provided a new demonstration of the distribution of errors (118). The paper starts by presenting some preliminary concepts concerning continuous distributions including that of the probability density function. Let $\varphi(x)dx$ denote the probability that an error made on certain observation lies between x and $x + dx$. The integral

$$\int_a^b \varphi(x)dx, \tag{119}$$

represents the probability that the error lies between a and b. The value of the integral taken from minus infinity to plus infinity equals the unity.

The mean value of x is given by the integral

$$k = \int x\varphi(x)dx. \tag{120}$$

It vanishes when the negative and positive errors are equally likely because in this case $\varphi(-x) = \varphi(x)$. When k is positive, there is some cause of error that produce positive errors with greater likelihood then negative errors. The mean value of x squared is

$$m^2 = \int x^2\varphi(x)dx, \tag{121}$$

where m is called the mean error. This quantity is more appropriate to quantify the uncertainty of the observations.

As an example, let $\varphi(x) = 1/2a$, a constant between $-a$ and a, and zero otherwise. Then $m = a/\sqrt{3}$, and $\mu = \lambda/\sqrt{3}$. As a second example, $\varphi(x) = (a - x)/a^2$ for x between zero and a, and $\varphi(x) = (a + x)/a^2$ for x between $-a$ and zero. In this case $m = a/\sqrt{6}$ and $\mu = \lambda\sqrt{2/3} - \lambda^2/6$. If

$$\varphi(x) = \frac{e^{-x^2/h^2}}{h\sqrt{\pi}}, \tag{122}$$

then $m = h/\sqrt{2}$. Defining

$$\Theta(z) = \frac{2}{\sqrt{\pi}} \int_0^z e^{-x^2} dx, \tag{123}$$

then $\mu = \Theta(\lambda/\sqrt{2})$.

Given a function U of the quantities $V_1, V_2, \ldots$, we wish to find the mean error M in the estimative of U when instead of the true values of these quantities we use the observed values having mean errors $m_1, m_2, \ldots$ Let us denote by $e_1, e_2, \ldots$ the errors in the observed values of $V_1, V_2, \ldots$. The resulting error is represented by

$$E = \sum_i \lambda_i e_i, \tag{124}$$

where $\lambda_i = dU/dV_i$, from which follows that the mean value of E is zero. Now, M^2 is the mean of

$$\sum_{ij} \lambda_i\lambda_j e_i e_j, \tag{125}$$

so that M^2 is the mean value of

$$\sum \lambda_i^2 e_i^2, \tag{126}$$

because the mean of the cross terms vanish.

To reach the distribution (118), Gauss assumed that the best estimator of a certain quantity V is obtained by maximization of the probability

$$\varphi(\Delta)\varphi(\Delta')\varphi(\Delta'')\ldots, \tag{127}$$

where $\Delta = M - V, \Delta' = M' - V, \Delta'' = M'' - V, \ldots$, and $M, M', M'', \ldots$ are the observed values of V. The maximization of the above expression gives

$$\frac{\varphi'(\Delta)}{\varphi(\Delta)} + \frac{\varphi'(\Delta')}{\varphi(\Delta')} + \frac{\varphi'(\Delta'')}{\varphi(\Delta'')} + \ldots = 0. \tag{128}$$

 DOI: https://doi.org/10.1590/1806-9126-RBEF-2022-0087

Comparing this expression with the arithmetic mean of the errors,

$$\Delta + \Delta' + \Delta'' + \ldots = 0. \tag{129}$$

Gauss concludes that φ'/φ should be proportional to Δ from which follows, after integration, that φ is proportional to $e^{-h^2\Delta^2}$, leading to the expression (118).

In 1850, Herschel sketched a derivation of (118) by assuming that the error distribution should be a function of the sum of the squares of the errors. This derivation appeared in an account of the Quetelet book on the application of the theory of probability to moral and political sciences that he wrote for the Edinburgh Review [64]. Herschel argues by considering the free fall of a ball from a certain height from a point situated vertically above a mark on the ground. The deviation from the mark is the error and the probability of the error should be a function of the sum of squares of the deviations determined by a rectangular frame of references.

Ellis provided in 1850 [65] a mathematical demonstration based on the Herschel reasoning as follows. Denoting the error by r, then using a rectangular coordinate we may write $r^2 = x^2 + y^2$. Thus the error function is expressed by $f(r^2) = f(x^2 + y^2)$. Assuming that the errors are independent of the direction, the function f has to satisfy the equation

$$f(x^2)f(y^2) = f(0)f(x^2 + y^2), \tag{130}$$

the solution of which is $f(x^2)$ proportional to e^{mx^2}, where m is a constant, from which follows

$$f = \frac{h}{\sqrt{\pi}} e^{-h^2x^2}. \tag{131}$$

In 1860 [66], Maxwell introduced the distribution of the velocity of the molecules of a gas. In his demonstration, he used a reasoning similar to that given above. He assumes that the rectangular components x, y, and z of the velocity of a particle are statistically independent so that probability density function related to x, y, and z is $f(x)f(y)f(z)$. As the direction of the coordinates are arbitrary the distribution must depends on the distance from the origin, that is,

$$f(x)f(y)f(z) = \phi(x^2 + y^2 + z^2). \tag{132}$$

The solution of this equation is $f(x)$ proportional to e^{Ax^2}. Writing $A = -1/\alpha^2$ and after normalization, Maxwell reaches the result

$$f(x) = \frac{1}{\alpha\sqrt{\pi}} e^{-x^2/\alpha^2}. \tag{133}$$

From this expression, Maxwell derives the probability density function related to the velocity v, which is the square root of $x^2 + y^2 + z^2$,

$$\frac{4v^2}{\alpha^3\sqrt{\pi}} e^{-v^2/\alpha^2}. \tag{134}$$

According to Hald [11], "There is nothing new in Maxwell's argument" but "it is the application to a new field that is revolutionary".

The work that exerted a great influence on Maxwell on the development of the velocity distribution was the reveiw of Quetelet book by Herschel that we mentioned above [67]. The account of Quetelet book by Herschel provided the analogy that Maxwell needed [68]. "What had occurred to no-one before Maxwell was that statistical laws could also apply to *physical processes*" [68].

12. Central Limit Theorem

Laplace showed that the probability distribution related to a certain number n of independent trials approaches a normal distribution when n is large. The demonstration was presented in his treatise in probability, as shown above, and before that in a publication of 1810 [69]. It was improved by Poisson in 1824 and 1829 and appeared in his treatise of probability of 1837, as shown above. After that, it was considered by Bessel in 1838, by Ellis in 1844, and by Cauchy in 1853 [11]. The Cauchy demonstration is based on the use of the characteristic function as did Laplace and Poisson.

In the period that include the last decades of the nineteenth century and the first decade of the twentieth century the problem was considered by Chebychev and by Markov by the method of moments and by Lyapunov by means of characteristic function. After this period, the theorem was discussed by von Mises, Pólya, Lindeberg, Lévy, Cramér, Kolmogorov, Khinchin and Feller [11]. Pólya called the theorem the central limit theorem because it is of central importance.

Let us consider a random variable ξ with a probability distribution $\rho(\xi)$. The characteristic function is defined by the integral

$$f(k) = \int e^{ik\xi}\rho(\xi)d\xi, \tag{135}$$

or by a summation if the random variable is discrete.

The moments μ_ℓ and the cumulants κ_ℓ of the distribution, if they exist, can be obtained from the expansion

$$f(k) = \sum_\ell \frac{\mu_\ell}{\ell!}(ik)^\ell, \tag{136}$$

$$\ln f(k) = \sum_\ell \frac{\kappa_\ell}{\ell!}(ik)^\ell. \tag{137}$$

Let us denote by $\xi_1, \xi_2, \ldots, \xi_n$ independent random variables with the same probability distribution $\rho(\xi)$. The distribution is symmetric which means that the odd moments and the odd cumulants vanish. We wish to determine the probability distribution of the sum of the random variables

$$x = \xi_1 + \xi_2 + \ldots + \xi_n. \tag{138}$$

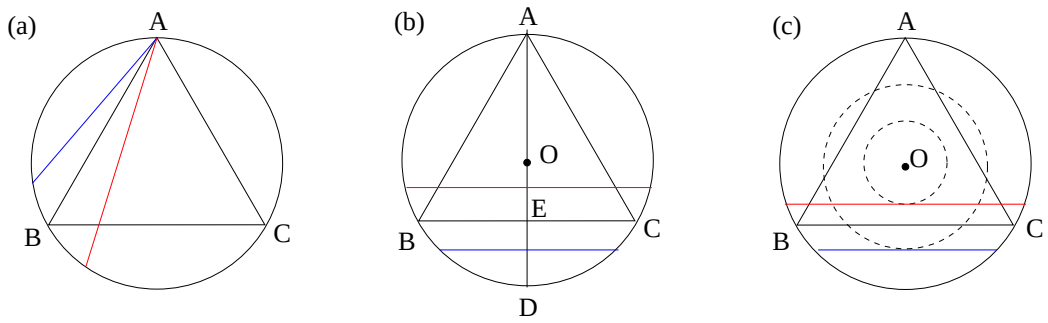


Figure 6: Three ways of randomly drawing a chord on a circle. (a) Two points on the circumference are chosen at random. (b) One radius is chose and random and one point of the radius is randomly chosen to place the middle point of the chord. (c) A point inside the circle is chosen at random to be the middle point of the chord.

Since the random variables are independent, the characteristic functions is

$$F(k) = [f(k)]^n, \tag{139}$$

and the desired distribution of the sum (138) is

$$\frac{1}{2\pi}\int F(k)e^{-ixk}dk. \tag{140}$$

From (139), we see that the cumulants of this distribution are n times the cumulants κ_ℓ.

Let us consider now the random variable $z = x/\sqrt{n}$. The probability of this distribution is

$$P(z) = \frac{\sqrt{n}}{2\pi}\int F(k)e^{-izk\sqrt{n}}dk. \tag{141}$$

Changing the variable of integration $k = w\sqrt{n}$,

$$P(z) = \frac{1}{2\pi}\int \Phi(w)e^{-izw}dw, \tag{142}$$

where $\Phi(w) = F(w/\sqrt{n})$ is the characteristic function related to the distribution $P(z)$.

Let us denote by σ_ℓ the cumulants associated to $\Phi(w)$. Taking into account that $\Phi(w) = F(w/\sqrt{n})$ and that the cumulants associated to $F(k)$ are $n\kappa_\ell$, we conclude that $\sigma_\ell = (n\kappa_\ell)/n^{\ell/2}$. Therefore, $\sigma_2 = \kappa_2$, $\sigma_4 = \kappa_4/n$, and so on, and in the limit where n increases without bounds, all cumulants vanish except σ_2. The characteristic function becomes

$$\Phi(w) = e^{-\sigma^2 w^2/2}, \tag{143}$$

where we wrote $\sigma_2 = \sigma^2$. Performing the integration in (142) we find

$$P(z) = \frac{e^{-z^2/2\sigma^2}}{2\pi\sigma^2}, \tag{144}$$

which is a normal distribution and an expression of the central limit theorem.

13. Bertrand

Bertrand wrote a book on the calculus of probability that was published in 1889 [70]. In this book he presented a paradox in probability that became connected to his name. It is presented as follows. Let us consider a circle and a let us draw a chord at random. The problem is to find the probability that the chord is longer than the side of an equilateral inscribed in the circle. Bertrand stated that there are three ways of reasoning that lead to three different answers which are 1/3, 1/2, and 1/4.

(a) In the first reasoning, one choses at random two points on the circumference of the circle and draw the chord. One of them is A and the other will be in any point of the circumference, and let us draw the equilateral triangle with a vertex at the point A, as shown in Figure 6a. From the figure, we see that the second point of a chord that are longer than the side of the triangle lies in the arc BC. Thus, the desired probability equals the ratio of length of the arc BC and the length of the circumference, which is 1/3.

(b) In the second reasoning, one choses a radius OD, and then one point at the radius chosen at random. Through this point one draws a chord perpendicular to the radius. The chords that are longer than the side of the triangles lie between O and E, as shown in Figure 6b. Those that are shorter lie between E and D. As OE is equal to ED, then the desired probability is the ratio betwee OE and OD, which is 1/2.

(c) In the third reasoning, one choses at random a point within the circle and draw a chord such that the chosen point is the midpoint of the chord. Let us draw a dashed circumference at the chosen point as shown in the Figure 6c. After that, one draws the equilateral triangle with one side parallel to the chord. The chords that are longer than the side of the equilateral triangle correspond to dashed circumferences with radius smaller than the radius OA. In this case the probability is related to the area of the dashed circle and thus the desired probability is 1/4.

The three reasonings correspond two three distinct ways of assigning the probability distribution.

 DOI: https://doi.org/10.1590/1806-9126-RBEF-2022-0087

To understand this, we determine for each one of the three cases the probability $\rho(\theta)d\theta$ that the angle subtended by the chord is between θ and $\theta + d\theta$. The probability p that the chord is longer that the inscribed equilateral triangle is equal to the probability that θ is between 120 and 180 degrees, that is,

$$p = \int_{2\pi/3}^{\pi} \rho(\theta)d\theta. \tag{145}$$

In the first case, ρ is constant and given by

$$\rho = \frac{1}{\pi}, \tag{146}$$

and we find $p = 1/3$. In the second case, the probability that the middle point of the chord is located in any point of a radius is constant. Denoting by x the distance of this middle point to the center of the circle, then the relation of x to the angle θ is given by $x = \cos\theta/2$, where we are considering a circle of unit radius. Taking into account that the probability density of x is constant then $\rho d\theta = dx$ which leads us to the result

$$\rho = \frac{1}{2}\sin\frac{\theta}{2}, \tag{147}$$

from which follow that $p = 1/2$.

In the third case, the middle point of the chord can be in any point of the circle. Denoting the rectangular coordinates of the middle point by x and y then the probability of finding the middle point finding the middle point inside the area $dxdy$ is $dxdy/\pi$ as the area of the circle of unit radius is π. From this result, it follows that the probability that the middle point is a distance between r and $r+dr$ form the center of the circle is $2rdr$. Taking into account that θ is related to r by $r = \cos\theta/2$, we find

$$\rho = \frac{1}{2}\sin\theta, \tag{148}$$

from which we get $p = 1/4$.

These examples shows that the probability is to be assigned according to the model that we wish to construct which will describe a certain aleatory real phenomenon.

14. Borel

In his book on the theory of probability published in 1909 [71], Borel gives a definition of probability as follows. Probability is the ratio between the number of favorable cases and the possible cases, regarding all cases as equally probable. This definition contains an apparent vicious circle, says Borel, as probability is being defined by the use of the terms equally probable, which encloses the very concept that is being defined. In fact, there is no vicious circle because these terms are not being used in the same cognitive level of the term being defined. In fact, they are being used in the vulgar meaning whereas the probability is being used in its mathematical sense. To break the vicious circle, we see that Borel is appealing to reasonings that are outside the scope of the theory. However, we may remedy this by merely assuming that equally probable is a primitive concept of the theory.

The postulates regarding probability are stated by Borel as follows. (1) The probability of an event that can occur in several different mutually exclusive ways is the sum of the probabilities corresponding to these various ways. (2) The probability of successive events is the product of the probabilities of these events, on the assumption that the preceding ones have occurred. In the case of simultaneous events, the probability is also the product if the events are independent of each other.

Let us consider a series of n trials where in each trial a favorable alternative occurs with probability p and the contrary alternative with probability $q = 1 - p$. The probability of occurrence of k favorable alternatives in n trials is

$$P = \frac{n!}{k!(n-k)!}p^k q^{n-k}. \tag{149}$$

Writing

$$k = np + t\sqrt{n}, \tag{150}$$

and using the Stirling formula in the approximation

$$n! = n^n e^{-n}\sqrt{2\pi n}, \tag{151}$$

one finds

$$P = \frac{1}{2\pi npq}e^{-t^2/2pq}. \tag{152}$$

From this expression, Borel derives the Jacob Bernoulli theorem, or law of large numbers, in the following form. Given a number ε as small as desired, the probability that the difference between the observed ratio $k/(n-k)$ and the theoretical ratio p/q is greater in absolute value than ε, tends towards zero when the number n of trials increases indefinitely.

In the second part of the book, Borel considers the case of continuous probability, or geometric probability. If we consider a straight line segment A, the probability that a point is found in the segment R contained in A is equal to the ratio of the length of R and the lenght of A. This definition is extended to the case of a plane. The probability of finding a point inside the surface region R enclosed in a surface region A is the ratio between the area of R and the area of A. In three dimensions, the probability is the ratio between the volume of A and the volume of R.

Borel treats several problems on the geometric probability. Let A a given point on the surface or a sphere and B another point chosen at random. The problem is to determine the probability that the length of the arc AB is smaller than α. It is understood that the arc AB

refers to the small arc of the great circle passing through A and B. The point B must be within the surface of a spherical cap of area equal to $2\pi R^2(1-\cos\alpha)$. The desired probability is the ratio between the area of the cap and the area $4\pi R^2$ of the spherical surface, that is, $(1-\cos\alpha)/2 = \sin^2\alpha/2$.

The needle problem is as follows. Suppose that in a sheet of papers we draw several parallel lines separated by the same distance. A needle is dropped on the sheet and we ask for the probability that the needle will cross one of the lines. This is a well known problem and was given a correct answer by Buffon, says Borel. We denote by $2a$ be the distance between the lines and by 2ℓ the length of the needle. We suppose that $a < \ell$ so that the needle will cross at most one line.

Let M be the middle point of the needle. The probability that the distance of this point to one of the line is between x and $x+dx$ is equal to dx/a. Let us denote by θ the angle that the needle makes with the perpendicular to the lines. The probability that the angle is between θ and $\theta + d\theta$ is equal to $2d\theta/\pi$. The needle will cross the line if $x \leq \ell\cos\theta$. The probability that this happens is

$$\int_0^{\pi/2}\int_0^{\ell\cos\theta}\frac{2dxd\theta}{a\pi} = \frac{2\ell}{a\pi}\int_0^{\pi/2}\cos\theta d\theta = \frac{2\ell}{a\pi}. \quad (153)$$

Following Poincaré [72], Borel states that it is possible to introduce an arbitrary function to represent the probability. In this case one denotes by $\varphi(x)dx$ the probability that the a point in the straight line is found between x and $x + dx$. The only condition imposed on this function is that it is positive. But there is a second condition represented by the integral

$$\int_{-\infty}^{\infty}\varphi(x)dx = 1. \quad (154)$$

The probability that the point lies between a and b is

$$\int_a^b \varphi(x)dx. \quad (155)$$

The expressions above are generalized to two or more dimensions. In two dimensions and the probability that a point is found in a surface S is

$$\int\int_S \varphi(x,y)dxdy, \quad (156)$$

where the integral of $\varphi(x,y)$ over the whole plane equals one. By changing variable from x and y to $\alpha = f(x,y)$ and $\beta = g(x,y)$ this probability can be expressed by the integral

$$\int\int_\Sigma \Phi(\alpha,\beta)d\alpha d\beta. \quad (157)$$

Borel notes that it is always possible to find a transformation such that $\Phi(\alpha,\beta) = 1$ in which case the probability is proportional to the area of the surface Σ. Therefore, given an arbitrary probability function there is a convenient transformation of variables that leads to a constant probability and thus justifying the apparent arbitrariness of the Laplace equally probable principle.

15. Markov

Markov [73, 74] contributed to several branches of mathematics particularly probability theory. In 1900 there appeared the first edition of his book on calculus of probability which went through three more editions [75, 76]. Markov was a political activist and was involved in many political and social issues. In 1913, when officials celebrated the three-hundred years of the House of Romanov, he organized a celebration of the two-hundred years of the Jacob Bernoulli law of large numbers.

Markov created a new field of research which was later called Markov chains. The urn problems studied by Daniel Bernoulli and by Laplace can be identified in retrospect as Markov chains. However, these earlier studies were not his motivation to the study of Markov chains. The main motivation of Markov was the extension of the central limit theorem to the case of dependent random variables [74]. His results showed that the independence of random variables is not a necessary condition for the validity of this fundamental theorem.

The concept of chains appeared in a paper of 1906 where Markov extended the law of large numbers to random variables depending on each other [77]. A second paper on the same subject followed in 1907 [78, 79], and in a third paper, published in 1908, he extended the limit theorem to the sum of dependent random variables[80]. This paper appeared as an appendix of the German translation of 1912 of his book on probability theory [76]. An English translation appeared in 1971 [81]. In the following we examine this third paper.

Let us consider a sequence of random variables $x_0, x_1, x_2, x_3 \ldots$, each one taking the values of a set of discrete values. The probability that x_{n+1} takes the value i when x_n takes the values j is denoted by p_{ij} and it is assumed that none of them equals the unity and

$$\sum_j p_{ij} = 1. \quad (158)$$

Denoting by P_i^n the probability that the variable x_n takes the value i then

$$P_j^{n+1} = \sum_i P_i^n p_{ij}, \quad (159)$$

and we may determine these probability from the initial values.

We also define the sum of the variables

$$s_n = x_1 + x_2 + \ldots + x_n. \quad (160)$$

 DOI: https://doi.org/10.1590/1806-9126-RBEF-2022-0087

Denoting by $Q^n_{s,i}$ the probabilities that the variable x_n takes the value i and that s_n takes the value m, Markov establishes the following equation

$$Q^{n+1}_{m,j} = \sum_i Q^n_{m-j,i} p_{ij}. \tag{161}$$

To solve this equation, we define the generating function

$$\phi^n_i = \sum_m Q^n_{m,i} t^m. \tag{162}$$

From equation (161), the following equation can be derived

$$\phi^{n+1}_j = \sum_i \phi^n_i p_{ij} t^j. \tag{163}$$

Once ϕ^n_i is found, we determine the function

$$\Phi_n = \sum_i \phi^n_i. \tag{164}$$

From this quantity we may find the probability R^n_m of s_n being equal to m, which is the coefficient of t^m in the expansion of Φ^n in powers of t.

The solution of equation (163) tell us that ϕ^n_i is a linear combination of the n-th power of the eigenvalues λ_k of the matrix A with elements $p_{ij}t^j$. The same can be said of the function Φ_n, that is,

$$\Phi_n = \sum_k a_k \lambda^n_k, \tag{165}$$

where a_k as well as λ_k depend on t. Defining

$$\Psi = \sum_n \Phi^n z^n, \tag{166}$$

we find

$$\Psi = \sum_{nk} a_k (\lambda_k z)^n. \tag{167}$$

Carrying out the summation in n,

$$\Psi(t,z) = \sum_k \frac{a_k}{1-\lambda_k z}, \tag{168}$$

and R^n_m is the coefficient of $t^m z^n$ in the expansion of Ψ.

Markov writes Ψ in the following form. Let $\delta_{ij} - p_{ij}t^j z$ be the elements of the matrix B and F its determinant. We see that the eigenvalues of B are $1-\lambda_k z$ so that

$$F(t,z) = \prod_k (1-\lambda_k z), \tag{169}$$

and we may write

$$\Psi(t,z) = \frac{f(t,z)}{F(t,z)}, \tag{170}$$

Table 1: Frequency of vowels and consonants in the *Eugene Onegin* according to Markov [82, 83]. The first table shows the frequencies of vowel (A) and the frequency of consonant (B). The second and third tables shows the frequencies of a vowel (A) or a consonant (B) following the group of letter shown in the firs row. We are indicating by A any vowel and by B any consonant.

0.432	A
0.568	B

A	B	↴
0.128	0.663	A
0.872	0.337	B

AA	AB	BA	BB	↴
0.104	0.546	0.131	0.868	A
0.896	0.454	0.869	0.132	B

where

$$f = \sum_k a_k \prod_{k'(\neq k)} (1-\lambda_{k'} z), \tag{171}$$

and it is clear that f is a polynomial in z, the largest power of z being the number of possible values taken by the random variable x_i, minus one.

When $t = 1$, the equation (163) becomes equal to (159) and we may identify ϕ^n_i with P^n_i. Since the sum of P^n_i in i equals the unit so does the sum of ϕ^n_i which is Φ^n. Therefore, Φ^n equals the unity when $t = 1$ in which case we find

$$\Psi(1,z) = \frac{1}{1-z}. \tag{172}$$

The generating function gives the probability R^n_m that s_n equals m. From this quantity one determines the moments M_ℓ of $y = (m-an)/\sqrt{n}$. Markov, shows that the

$$\lim_{n\to\infty} M_\ell = C^{1/2} \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} t^\ell e^{-t^2} dt, \tag{173}$$

where C is a constant. Therefore the probability that

$$t_1\sqrt{C} < y < t_2\sqrt{C} \tag{174}$$

is given by

$$\frac{1}{\sqrt{\pi}} \int_{t_1}^{t_2} t^\ell e^{-t^2} dt, \tag{175}$$

which proves the central limit theorem.

In 1913, Markov published a paper where he showed an interesting application of his theory to the analysis of a literary text [82, 83]. Markov analyzed the sequence of 20,000 letters of the Pushkin's poem *Eugeny Onegin* and determined the frequency of vowels shown in Table 1. One of these quantities is simply the frequency of vowels in the text. Another quantity is the frequency of vowels that are preceded by a group of letters. In one case the preceding group is composed of just one letter. In the second case the preceding group is composed by two

letters. The former case is understood as a Markov chain of range one whereas the second is of range two.

The Markov process is usually employed to describe a random process that occurs in time, called stochastic process or stochastic dynamics [84]. In this sense the Markov equation (159) is understood as the evolution of probability in discrete time and discrete space of states, and the coefficients p_{ij} are the transition probabilities from state i to state j. It is possible to consider the time as a continuous variable in which case the equation is called a master equation [84]. It is also possible to consider an equation for a continuous space of states as is the case of the Fokker-Planck equation and the Kolmogorov equation [84]. All these equations, discrete or continuous, are derived from the theory advanced by Markov [84].

As we have said above, the problems of urn studied by Daniel Bernoulli and Laplace can be understand in retrospect as a Markov chain. In fact, the equation (84) that was proposed by Laplace is an example of a Markov equation, and the equation (85) also proposed by Laplace is a Fokker-Planck equation. We also add that the fundamental equation introduced by Boltzmann within the kinetic theory can be understood in retrospect, although in an approximate form, as a Markov equation in continuous time and continuous space [85].

16. Kolmogorov

In 1933, Kolmogorov published his book on the foundation of the theory of probability, which is considered as the foundation of modern probability theory [86]. The book appeared in German and was translated into Russian in 1936 and into English in 1950 [87]. The theory of probability presented in the book has the systematic and analytic structure in which the theorems are derived from the fundamental postulates or axioms. Kolmogorov states that the theory of probability should be treated as other theories such as geometry or algebra.

In some sense, the systematic and analytic structure of the theory of probability is found in the theories that we have examined above, although the fundamental propositions are not explicit given in these theories. But the main distinguishing feature of the Kolmogorov theory is a clear identification of probability as a *measure* (not to be confused with physical measurement) defined on the space of events or the sample space. Examples of physical quantities that are understood as measures defined on the real space is length, area, volume, and mass. They are all non negative real quantities that increases monotonically in the following sense. If A and B are subsets of the measurable space such that A is a subset B then the measure of A is smaller or equal to the measure of B.

The Kolmogorov theory is based on the understanding that the events that we wish to describe make up a space of events, or the sample space. This space of events is identified as a set in the mathematical sense of the set theory whose elements are the *elementary events*. This set is called U and any event is a subset of U. Kolmogorov gives the correspondence of propositions involving the events and those involving sets, as shown in Table 2.

Table 2: Correspondence between sets and events. We are using the following notation: $\cap$ and $\cup$ for the union and intersection of sets instead of $+$ and juxtaposition used by Kolmogorov. The empty set is denoted by V, and the complementary set of A by $\bar{A}$.

Events	Sets
collection of all elementary events	U
an elementary event	$e, \quad e \in U$
an event	$A, \quad A \subset U$
simultaneous occurrence of events	$A \cap B = X$
impossible event	V
incompatible events	$A \cap B = V$
occurrence of at least one event	$A \cup B = Y$
non-occurrence of event A	$\bar{A}$
event A follows inevitably from B	$B \subset A$

If two events A and B are incompatible it means these subsets do not intersect, or that that their intersection $A \cap B$ equals the empty set. If C is an event defined as the simultaneous occurrence of A and B, then the subset C is the intersection of the subsets A and B, that is, $C = A \cap B$. If C is the defined as the occurrence of at least one of the two events A and B, then the subset C is the union of A and B, that is, by $C = A \cup B$. The event corresponding to the non-occurrence of an event A is the complementary subset $\bar{A}$.

The postulates are as follows. (1) To each subset A of U one assigns a non-negative real number $P(A)$, the probability of A. (2) The probability of U is equal to one, $P(U) = 1$. (3) If A and B have no element in common, that is, if $A \cap B = V$ then $P(A \cup B) = P(A) + P(B)$. Some elementary consequences follows immediately. From $U = U \cup V$, it follows that $P(V) = 0$, which means that the probability of an impossible event is zero. Since $U = A + \bar{A}$, then $P(\bar{A}) = 1 - P(A)$. If B is included in A, then $P(B) \leq P(A)$.

An event which consists of the occurrence of an event A but not the occurrence of an event B corresponds to the set C whose elements belongs to set A but not to set B. These last elements compose the intersection $A \cap B$. Since C and $A \cap B$ are mutually exclusive sets and $A = C + A \cap B$ then $P(A) = P(C) + P(A \cap B)$. But C and B are mutually exclusive sets and they compose the union of the sets A and B. Therefore $P(A \cup B) = P(C) + P(B)$, and

$$P(A \cup B) = P(A) + P(B) - P(A \cap B). \tag{176}$$

These axioms are not sufficient to specify completely the actual values of the probability of each event. The establishment of these values will depend on the model one construct to describe a specific aleatory phenomena, which is the assignment of probability. For instance,

in the case of a dice, the sample space U is the set $\{1, 2, 3, 4, 5, 6\}$. But the postulates say nothing about the probability that we may assign to each element.

The conditional probability is defined by

$$P_A(B) = \frac{P(A \cap B)}{P(A)}, \tag{177}$$

provided $P(A) > 0$, from which follows $P(A \cap B) = P(A)P_A(B)$. In an analogous manner $P_B(A) = P(A \cap B)/P(B)$ from which follows the Bayes theorem

$$P_B(A) = \frac{P(A)P_A(B)}{P(B)}. \tag{178}$$

The concept of independence is a central concept of the theory of probability. Independent events are defined as events such that $P(A \cap B) = P(A)P(B)$. In this case the probability of the occurrence of either events is the sum of the probabilities $P(A \cup B) = P(A) + P(B)$.

The concept of random variables is a relevant concept in the theory of probability as it allows an easier analytical approach to the probabilistic problem. The basic idea is to associate a real number to the elementary events of the set U of all events. A random variable ξ is defined as a function of the elementary events e of U, $\xi = f(e)$. It maps the set U into the set of the real numbers.

Let A_x be the subset of U such that $\xi = f(e) < x$ for all elements e belonging to A_x. Then the probability of A_x, which we denote by $F(x)$ is the probability that $\xi < x$, and is called the distribution function of the random variable ξ. The probability that $x_1 \leq \xi < x_2$ is $F(x_2) - F(x_1)$ and since this must be a non-negative quantity it follows that $F(x)$ is nondecreasing function of x. If the random variables takes values in the interval between a and b, then $F(x)$ equals zero and one, when x approaches a and b, respectively. For continuous random variables we define the probability density function $\rho(x)$ by $\rho = dF/dx$ so that

$$F(x) = \int_a^x \rho(z)dz. \tag{179}$$

The concepts introduced above can be generalized to the case where one associates a random vector to an elementary event.

Several books were written in accordance with the theory of probability developed by Kolmogorov. We mention the book of Gnedenko which was published in Russian in 1950 [88]. A German translation was published in 1957 [89] and an English translation in 1962 [90]. The book went through several editions in several languages. A general treatise on the theory of probability was published by Feller in 1950 [91]. It contains a clear exposition of the basic theory and a number of applications. It was followed by a second volume published in 1966 [92].

Table 3: The author together with the abbreviated name of the main work where it is found, the year of its public presentation or publication, and references. Roman numerals indicate the century when the work appeared. Numbers between square brackets indicate the approximate date when the work was written.

Author	Work	Year	Ref.
	De Vetula	XIII	[17]
	commentary on Divine Comedy	XIV	[20]
Cardano	Book on Games of Chance	[1550]	[26]
Galileo	Findings on the game of dice	[1620]	[27]
Pascal and Fermat	correspondence	1654	[30]
Huygens	Calculation in Games of Chance	1657	[37]
J. Bernoulli	Art of Conjecturing	1713	[40]
Montmort	Essay on the Games of Chance	1708	[42]
De Moivre	Doctrine of Chances	1718	[45]
D. Bernoulli	conjectural problem	1770	[52]
Lagrange	calculation of probability	1776	[54]
Laplace	Theory of Probability	1812	[13]
Poisson	Research on Probability	1837	[4]
Bertrand	Calculus of Probability	1889	[70]
Borel	Theory of Probability	1909	[71]
Markov	dependent variables	1906	[77]
Kolmogorov	Theory of Probability	1933	[86]

17. Conclusion

In Table 3 we show the works that we have analyzed here concerning the development of the concept of probability and the theory of probability. The development of the theory of probability shows that it can be understood as the science of aleatory events. The quantity called probability obeys certain fundamental postulates, such as those advanced by Kolmogorov, which are stated in terms of the concept of space of events or sample space. It is usual to say that these postulates define the concept of probability, but we deem it more convenient to assume it as primitive concept such as time, space, or mass. A second essential point of the science of probability is that probability is measured by interpreting it as the ratio of the favorable observable outcomes and the total observed outcomes, or in short the frequency of favorable observed outcomes.

The fundamental postulates alone are not sufficient to determine the probability related to a specific aleatory phenomena. We have to assign a probability to each one of the elementary events of the sample space and this will depend on the model we construct. Thus, to describe a set of aleatory event by a probabilistic theory, we have to set up a sample space and assign a probability distribution to this sample. The failure to clearly carry out these procedures may result in paradoxes such as the Bertrand paradox explained above.

From the development of the concept of probability described above, we may say that the calculation of probability was usually carried out by setting up a sample space such that the elementary events of this space have equal probability. This equiprobability rule

was not always explicitly stated but eventually it was transformed into the basis of the Laplace theory of probability. According to this principle the probability is the ratio between the *possible* favorable outcomes and the total outcomes. This ratio should not be confused with the frequency of *observed* favorable outcomes in certain number of observed trials.

There are some reasonings used in the calculation of probabilities that are considered errors or fallacies [93]. An example is the Leibniz reasoning in the calculation of probability in a throw of two dice, that we have mentioned above. He argues that the number 12 has the same probability as the number 11. One interpretation of this result is to say that the equiprobable sample space used by Leibniz consists of partitions and not permutations. The numbers 12 and 11 have each one just one partition which are (66) and (56), respectively. Although, there is one permutation for the number 12, there are two permutations for the number 11 which are $(5,6)$ and $(6,5)$. The problem here is that the appropriate model for the throw of two dice is not the partition model used by Leibniz but the permutation model. But this can only be verified by an experiment and cannot be decided by theoretical reasoning.

The theories of probability that we have examined above were and are applied to the analysis of statistical data coming from various areas of research and to the analysis of statistical errors in experimental measurements. But we wish to remark that the theory of probability is an essential part of probabilistic physical theories, better known as statistical physical theories. That was the case of the kinetic theory developed by Clausius, Maxwell and Boltzmann and of the statistical mechanics developed by Gibbs. If we interpret the Markov chains as evolution of probability in time, then the theory of Markov can be understood as an essential part of the stochastic theory of Brownian motion and more generically of stochastic dynamics. More recently is was essential part of the stochastic mechanics, also known as stochastic thermodynamics.

Acknowledgement

I wish to acknowledge Pedro Tomé for drawing my attention to the Mallarmé poem.

References

[1] M.J. de Oliveira, "The structure of the scientic theories", Rev. Bras. Ens. Fis. **43**, e20200506 (2021).

[2] R. Carnap, *An Introduction to the Philosophy of Science* (Basic Books, New York, 1966).

[3] I. Hacking, *The Emergence of Probability* (Cambridge University Press, Cambridge, 1975).

[4] S.D. Poisson, *Recherches sur la Probablité de Jugements em Matière Criminelle et en Matière Civile* (Bachelier, Paris, 1837).

[5] A.A. Cournot, *Exposition de la Théorie de Chances et des Probabilités* (Hachette, Paris, 1843).

[6] A. Arnauld and P. Nicole, *La Logique ou l'Art de Penser* (Savreux, Paris, 1662).

[7] I. Todhunter, *A History of the Mathematical Theory of Probability* (Macmillan, Cambridge, 1865).

[8] F.N. David *Games, Gods and Gambling* (Hafner, New York, 1962).

[9] K. Pearson, *The History of Statistics in the 17th and 18th Centuries* (Griffin, London, 1978).

[10] S.M. Stigler, *The History of Statistics* (Belknap Press, Cambridge, 1986).

[11] A. Hald, *A History of Probability and Statistics and Their Applications before 1750* (Wiley, New York, 1990).

[12] A. Hald, *A History of Mathematical Statistics from 1750 to 1930* (Wiley, New York, 1998).

[13] P.S. Laplace, *Théorie Analytique des Probabilités* (Courcier, Paris, 1812).

[14] P.S. Laplace, *Essai Philosophique sur les Probabilités* (Courcier, Paris, 1814).

[15] P.S. Laplace, *A Philosophical Essay on Probabilities* (Wiley, New York, 1902).

[16] S. Mallarmé, "Un coup de dés jamais n'abolira le hasard", Cosmopolis **5**, 417 (1897).

[17] P. Ovidii Nasonis, *De Vetula* (1534).

[18] M.G. Kendal, "The beginning of a probability calcullus", Biometrika **43**, 1 (1956).

[19] D.R. Bellhouse, "*De Vetula*: a medieval manuscript containing probability calculations", International Statistical Review **68**, 123 (2000).

[20] A. Torri, *L'Ottimo Commento della Divina Commedia* (Niccolò Capurro, Pisa, 1827), 3 volumi.

[21] G. Libri, *Histoire des Sciences Mathématiques en Italie depuis la Renaissance de Lettres jusqu'à la fin du Disseptieme Siècle* (Renouard, Paris, 1838).

[22] D. Aliguieri, *Divina Commedia* (Le Monnier, Firenze, 2009).

[23] "Zara" in *Enciclopedia Dantesca* (Treccani, Rome, 1970), available in: www.treccani.it/enciclopedia/elenco-opere/Enciclopedia_Dantesca

[24] O. Ore, *Cardano, The Gambling Scholar* (Princeton University Press, Princeton, 1952).

[25] G. Cardano, *The Book on Games of Chance* (Holt, Rinehart and Winston, New York, 1961).

[26] H. Cardani, *Opera Omnia* (Lugduni, 1663), p. 262.

[27] G. Galilei, in: *Opere di Galileo Galilei* (Tartini e Franchi, Firenze, 1718), v. 3, p. 119.

[28] G. Galilei, in: *Le Opere di Galileo Galilei* (Barbèra, Firenze, 1898), v. 8, p. 591.

[29] G.W. Leibniz, *Opera Omnia* (Tournes, Genevae, 1768), v. 6, p. 217.

[30] P. Tannery and C. Henry, *Oeuvres de Fermat* (Gauthier-Villars, Paris, 1894), v. 8, p. 288.

[31] D.E. Smith, *A Source Book in Mathematics* (McGraw-Hill, New York, 1929).

[32] D.J. Struik, *A Source Book in Mathematics 1200-1800* (Princeton University Press, Princeton, 1986).

[33] B. Pascal, *Traité du Triangle Arithmetic* (Desprez, Paris, 1665).

[34] A.W.F. Edwards, *Pascal's Arithmetical Triangle* (Johns Hopkins University Press, Baltimore, 1987).

[35] P. Hérigone, *Cursus Mathematicus* (Paris, 1634).

[36] E. Coumet, "Le probleme des partis avant Pascal", Arch. Intern. d'Histoire des Sciences **18**, 245 (1965).

[37] F. Schooten, *Exercitationum Mathematicarum* (Elsevirii, Lugduni Batauorum, 1657).

[38] C. Huygens, *The Value of all Chances in Games of Fortune* (Keimer, London, 1714).

[39] C. Huygens, *Oeuvres Complètes de Christiaan Huygens* (Martinus Nijhoff, La Haye, 1920), tome quatorzième.

[40] J. Bernoulli, *Ars Conjectandi* (Thurnisii fratres, Basileae, 1713).

[41] J. Bernoullli, *The Art of Conjecturing* (John Hopkins University Press, Baltimore, 2006).

[42] P.R. Montmort, *Essay d'Analyze sur les Jeux de Hazards* (Jacque Quillau, Paris, 1708).

[43] P.R. Montmort, *Essay d'Analyze sur les Jeux de Hazards* (Jacque Quillau, Paris, 1713), seconde édition.

[44] A. Moivre, "De mensura sortis", Philosophical Transactions **27**, 213 (1711).

[45] A. Moivre, *The Doctrine of Chances* (Pearson, London, 1718).

[46] A. Moivre, *The Doctrine of Chances* (Woodfall, London, 1738), 2nd. ed.

[47] A. Moivre, *The Doctrine of Chances* (Millar, London, 1756), 3rd. ed.

[48] K. Pearson, "Historical note on the origin of the normal curve of errors", Biometrika **16**, 402 (1924).

[49] R.C. Archibald, "A rare pamphlet of Moivre and some of his discoveries", Isis **8**, 671 (1926).

[50] R.H. Daw and E.S. Pearson, "Abraham De Moivre's 1933 derivation of the normal curve: a bibliographic note", Biometrika **59**, 677 (1972).

[51] J. Stirling, *Methodus Differentialis, sive Tractatus de Summatione et Interpolatione Serierum Infinitarum* (Bowyer, Londini, 1730).

[52] D. Bernoulli, "Disquisitiones analyticae de novo problemate conjecturali", Novi Commentarii Academiae Scientarum Imperialis Petropolitanae **14**, 3 (1770).

[53] P. Ehrenfest, "Über zwei bekannte Einwände gegen das Boltzmannsche H-Theorem", Physikalische Zeitschrift **8**, 311 (1907).

[54] J.L. Lagrange, "Memoire sur l'utilite de la methode de prendre le milieu entre les resultats de plusiers observations", Miscellanea Taurinensia **5**, 167 (1776).

[55] K. Pearson, "Notes on the history of correlation", Biometrika **13**, 25 (1920).

[56] A.M. Legendre, *Nouvelles Mèthodes pour la Détermination des Orbites des Comètes* (Firmin Didot, Paris, 1805).

[57] C.F. Gauss, *Theoria Motus Corporum Coelestium in Sectionibus Conicis Solem Ambientium* (Perthes et Besser, Hamburg, 1809).

[58] C.F. Gauss, *Theory of the Motion of Heavenly Bodies Moving about the Sun in Conic Sections* (Little, Brown and Company, Boston, 1857).

[59] C.F. Gauss, "Theoria combinationis observationum erroribus minimis abnoxiae: Pars prior", Commentatines Societatis Regiae Scientarum Gottingensis Recentiores **5**, 33 (1823).

[60] C.F. Gauss, "Theoria combinationis observationum erroribus minimis abnoxiae: Pars posterior", Commentatines Societatis Regiae Scientarum Gottingensis Recentiores **5**, 63 (1823).

[61] C.F. Gauss, "Supplementum theoriae combinationis observationum erroribus minimis abnoxiae: Pars posterior", Commentatines Societatis Regiae Scientarum Gottingensis Recentiores **6**, 57 (1829).

[62] C.F. Gauss, *Méthode des Moindres Carrés* (Mallet-Bachelier, Paris, 1855).

[63] C.F. Gauss, *Theory of Combination of Observations Least Subject to Errors. Part One, Part Two, Supplement* (SIAM, Philadelphia, 1995).

[64] J.F.W. Herschel, "Quetelet on probabilities", Edinburgh Review **92**, 1 (1850).

[65] R.L. Ellis, "Remarks on an alleged proof of the method of least squares", Philosophical Magazine **37**, 321 (1850).

[66] J.C. Maxwell, "Illustrations of the dynamic theory of gases", Philosophical Magazine **19**, 19; **20**, 21 (1860).

[67] E. Garber, "Aspects of the introduction of probability into physics", Centaurus **17**, 11 (1973).

[68] B. Mahon, *The Man who Changed Everything* (Wiley, New York, 2004).

[69] P.S. Laplace, Mémoires de la Classe des Sciences Mathématiques et Physiques de l'Institut de France, année 1809 **353** (1810).

[70] J. Bertrand, *Calcul des Probabilités* (Gauthier-Villars, Paris, 1889).

[71] E. Borel, *Éléments de la Théorie des Probabilités* (Hermann, Paris, 1909).

[72] H. Poincaré, *Calcul de Probabilités* (George Carré, Paris, 1896).

[73] C.C. Gillispie, *Dictionary of Scientific Biography* (Scribner, New York, 1974), v. 9, p. 124.

[74] G. P. Basharin, A.N. Langville and V.A. Naumov, "The life and work of A. A. Markov", Linear Algebra and its Applications **386**, 3 (2004).

[75] A.A. Markov, *Ischislenie Veroyatnostej* (1900).

[76] A.A. Markov, *Wahrscheinlichkeits-Rechnung* (Teubner, Leipzig, 1912).

[77] A.A. Markov, "Rasprostranenie zakona bol'shih chisel na velichiny zavisyaschie drug ot druga", Izvestiya Fiziko-matematicheskogo obschestva pri Kazanskom universitete **15**, 135 (1906).

[78] A.A. Markov, "Issledovanie zamechatel'nogo sluchaya zavisimyh ispytanij", Izvestiya Akademii Nauk, **1** 61 (1907).

[79] A.A. Markov, "Recherches sur un cas remarquable d'epreuves dependantes", Acta Mathematica **33**, 87 (1910).

[80] A.A. Markov, "Rasprostranenie predel'nyh teorem ischisleniya veroyatnostej na summu velichin svyazannyh v cep", Zapiski Akademii Nauk po Fiziko-matematicheskomu otdeleniyu **22** (1908).

[81] R.A. Howard, *Dynamic Probabilistic Systems* (Wiley, New York, 1971).

[82] A.A. Markov, "Primer statisticheskogo issledovaniya nad tekstom 'Evgeniya Onegina', illyustriruyuschij svyaz' ispytanij v cep' ", Izvestiya Akademii Nauk **7**, 153 (1913).

[83] A.A. Markov, "An example of statistical investigation of the text Eugene Onegin concerning the connection of samples in chains", Science in Context **19**, 591 (2006).

[84] T. Tomé and M.J. de Oliveira, *Stochastic Dynamics and Irreversibility* (Springer, Cham, 2015).

[85] M.J. de Oliveira, "Boltzmann stochastic thermodynamics", Phys. Rev. E **99**, 052138 (2019).

[86] A. Kolmogoroff, *Grundbegriffe der Wahrscheinlichkeitsrechnung* (Springer, Berlin, 1933).

[87] A.N. Kolmogorov, *Foundations of the Theory of Probability* (Chelsea, New York, 1950).

[88] B.V. Gnedenko, *Kurs Teorii Veroyatnostey*, (Tekhniko-Teoreticheskoi Literatury, Moskva, 1950).

[89] B.W. Gnedenko, *Lehrbuch der Wahrscheinlichkeitsrechnung* (Akademie, Berlin, 1957).

[90] B.V. Gnedenko, *The Theory of Probability* (Chelsea, New York, 1962).

[91] W. Feller, *An Introduction to Probability Theory and Its Applications* (Wiley, New York, 1950), v. 1.

[92] W. Feller, *An Introduction to Probability Theory and Its Applications* (Wiley, New York, 1950), v. 2.

[93] P. Gorroochurn, "Errors of probability in historical context", The American Statistician **65**, 246 (2011).