A Brief History of Probability Theory

XIAN Sicheng, Qiuzhen College

Contents

1 Classical Probability

People ask questions about probability every day, like "What are the odds of raining tomorrow?" or "What's the chance of surviving from a car accident?" Nowadays, probability theory is one of the major research areas of mathematics, but it's quite surprising that probability was not regarded as a branch of mathematics or seriously studied until quite late, when Fermat and Pascal studied the well-known problem proposed by De Méré in 17th century.

A century before that, Girolamo Cardano had written the book *Liber de ludo aleae* (the book on games of chance), which is in fact essentially a guidebook for gamblers, in which he studied the probability problems emerging from gambling. This book is seen as the earliest work related to probability(chance) theory.

It is widely believed that classical probability begins with the Pascal-Fermat correspondence. However both mathematicians didn't publish their results; the first book on probability theory attributes to Huygens. Later in 18th century Bernoulli and De-Moivre established respectively the first law of large numbers and the first central limit theorem. In 19th century Laplace summarized these works in his book *Théorie Analytique de Probabilités*, which completes the foundations for classical probability.

In this chapter, we will review the history of classical probability theory, introducing the earliest versions of law of large numbers and central limit theorem, which will be discussed more detailedly in Chapter 2 and Chapter 3.

1.1 Cardano and gambling

Cardano(1501-1576), also known as Jerome Cardan, is an Italian mathematician, famous for his work *Ars magna* (The great art), in which he gave the solutions of cubic and quartic equations.

Cardano's father, Fazio, was a lawyer but also an expert in math. Fazio had taught Cardan mathematics, which inspired him of his academic career. Soon after Fazio's death, Cardan squandered the small legacy from his father, and turned to gambling. Card games, dices and chess were the methods he used to make a living. His understanding

made him win more than he lost.

Figure 1: Cardano

In the book *Liber de Ludo Aleae*, Cardano carefully studied the dice games. For example, he stated that the "most important principle" of all in gambling is simply equal conditions, requiring the die to be even, which is. He observed that in the case of throwing 2 dice, the point 10 can be obtained by (5,5) and (4,6), but the latter one can occur in two ways, so that the whole number of ways obtaining 10 will be $\frac{1}{12}$ out of circuit. In this way, he calculated the chances for each number of points for throwing 3 dice as well as 2 dice.

This book is the first study of things such as dice rolling, based on the premise that there are fundamental scientific principles governing the likelihood of achieving the elusive 'double six', outside of mere luck or chance. He also came of the idea of law of large numbers - he wrote -

A repeated succession, for example, in 3600 casts, the equality is 1/2 of that number, namely, 1800 casts; in such a number of casts the desired result may or may not happen. Accordingly, this knowledge is based on conjecture which yields only an approximation, yet it happens in the case of many circuits that the matter falls out very close to conjecture.

He realized that there's something certain behind those completely

random games of chance - the likelihood of winning or losing.

Unfortunately, the manuscripts wasn't discovered until nearly a century after his death. His work had no direct influence on the development of probability theory since by the time *Liber de Ludo Aleae* was first published in 1660s, the theory of probability had reached the stage beyond Cardano's work. At that time, the basic foundations of the probability theory was already established by Pascal and Fermat.

1.2 Pascal-Fermat correspondence

Chevalier de Méré, a 17th century gambler, a philosopher of noble society, made an acquaintance of mathematician Blaise Pascal in 1650s, during a trip to Poitou, France. During the 1650s, de Méré asked several questions to Pascal, including the "dice problem" and the "division problem". The dice problem asks how many times one must throw before expecting a double six, while the other "division problem", is stated as follows.

> *Two gamblers agree to toss a fair coin until one of them wins 6 rounds. When the first player wins 5 rounds and the second player wins 2 rounds, the game comes to an abrupt halt and cannot be finished. How shall one divide the prize money equitably then?*

This problem has many versions, and it can date back to the book *Summa de arithmetica, geometrica, propotioni et proportionalita* by Italian mathematician Luca Pacioli. In the 200 years after it was proposed, many mathematicians tried to give their solutions, including Pacioli himself and Cardano, but all their answers were wrong. The problem of division remained unsolved until Pascal set out to solve it.

Blaise Pascal(1623-1662), French mathematician who contributed to many areas of mathematics. He was the third child of Étienne Pascal, a lawyer and amateur mathematician. Étienne had unorthodox educational views and decided to teach Blaise himself. He didn't want Blaise to study mathematics too early and he removed all the math texts from their house. Blaise however, raised by his curiosity, started to work on geometry himself. When his father found out, he relented and allowed Blaise a copy of Euclid. Early in the age of 16, he discovered and proved the well-known Pascal's hexagon theorem.

Figure 2: Pascal

In his twentieth Blaise Pascal began a series of experiments on atmospheric pressure. In 1648 he discovered that the atmospheric pressure decreases with height and predicted the existence of vacuum. In 1653 he worked on *Treatise on the Equilibrium of Liquids* and explained the Pascal's law of pressure.

Pascal also contributed to the study of binomial coefficients. The work on Pascal's triangle also helped him solve the problem asked by De Méré.

When the two players needed a and b more rounds to win respectively, we imagine they play $(a+b-1)$ more rounds, so that exactly one of them wins, even if the result may have been determined in less than $(a+b-1)$ rounds. Let $n = a+b-1$, then in total of 2^n cases, the first player wins

$$\binom{0}{n} + \binom{1}{n} + \cdot + \binom{b-1}{n}$$

cases, and the second players wins the rest

$$\binom{b}{n} + \binom{b+1}{n} + \cdot + \binom{n}{n}$$

cases. And this is the correct solution to the 200-year-lasted question. However, Pascal himself wasn't quite sure about his own answer, since mathematicians have been giving different (wrong) answers for centuries, so he contacted another mathematician, Fermat, in the year of 1654.

Figure 3: Fermat

Pierre de Fermat(1601-1665), a French lawyer, government official and (amateur) mathematician, most remembered for his work in number theory, especially the Fermat's Last Theorem, and the quote *"I have discovered a truly remarkable proof which this margin is too small to contain."*

In July, 1954, Pascal and Fermat began writing letters to each other. Pascal was cheered by finding Fermat's results were in complete agreement with his: "I see that the truth is the same in Toulouse and Paris." The Pascal-Fermat correspondence remarks the beginning of the theory of probability theory, although in their correspondence there are no rigorous definitions of probability, or even the word "probability" itself, let alone rigorous theorems and proofs. However, mathematicians knew very well in their minds what "probability" should be - they regard it as the number of equi-possible events - and achieved many important results, without axiomatic systems been built (see Chapter 5), similar to the theory of calculus and analysis in 17th and 18th century.

1.3 Huygens and expectation

Figure 4: Huygens

Christian Huygens(1629-1695), a Dutch mathematician, noticed the deep ideas behind the simple "games of chance".

I would like to believe that if someone studies these things a little more closely, then he will almost certainly come to the conclusion that it is not just a game which has been treated here, but that the principles and the foundations are laid of a very nice and very deep speculation.

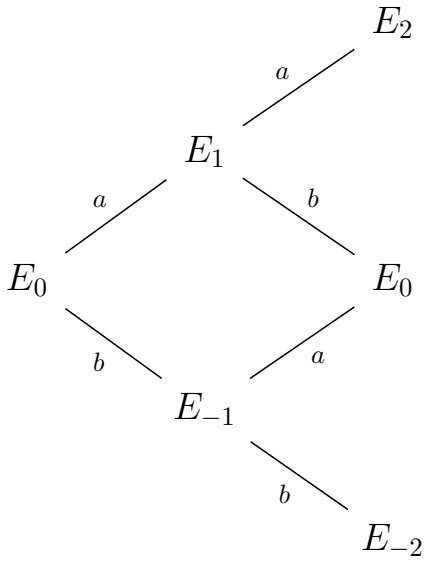
Although Huygens is famous for his contributions on physics and astronomy (He invented the pendulum clock, established his wave theory for light, improved telescopes and observed the rings of Saturn.), he also contributed to mathematics, including the probability theory. On a visit to Paris in 1655 he became acquainted with the division problem and later studied on it. In 1657 he published his treatise *On Reckoning at Games of Chance*, where he first proposed the concept of **expectation**. (When mathematicians haven't defined "probability" yet.)

For example, he stated that "The total expectation for a number with p cases turning out to be a and q cases of b shall be $\frac{ap+bq}{p+q}$"

In the book Huygens proposed the problem known as the "gambler's ruin" (Originally mentioned in Pascal-Fermat correspondence), generalized description follows:

Let two men play a game, where each player starts with n points. In each round, the chance for each to score a point and simultaneously subtract a point from his opponent is $a : b$. Then what's the chance of victory (when one's opponent reaches 0 points.) for each player?

Here is a solution by Huygens himself:



Let the expectation for player A be E_t when net transfer of points to A is t. ($t < 0$ implies net points gained by A is negative.) As the diagram shows above, the expectations must satisfy

$$E_0 = \frac{a^2 E_2 + 2abE_0 + b^2 E_{-2}}{(a+b)^2}$$

which is equivalent to

$$E_0 = \frac{a^2 E_2 + b^2 E_{-2}}{a^2 + b^2}$$

Generally for different indices, we have

$$E_k = \frac{a^2 E_{k+2} + b^2 E_{k-2}}{a^2 + b^2}, \quad E_k = \frac{aE_{k+1} + bE_{k-1}}{a+b}$$

and the victory condition requires

$$E_m = 1, E_{-m} = 1, \forall m \geqslant n$$

Through some computation, one may solve that

$$E_k = \frac{a^r E_{k+r} + b^r E_{k-r}}{a^r + b^r}$$

Which implies the chance ratio for victory should be $a^n : b^n$.

This result implies the following fact that a gambler is doomed to lose all his money if he insists on playing a game with negative expectation, hence the name.

1.4 Jakob Bernoulli's contributions

Huygens' was the only available text in probability for the next fifty years, (and served as standard textbook) until it was incorporated as Part I of Jakob Bernoulli's *Ars Conjectandi*(The Art of Conjecturing).

Jakob Bernoulli(1655-1705) was a Swiss mathematician, member of the distinguished Bernoulli family which produced 8 mathematicians who made significant contributions. Here we omit the introduction of the Bernoulli family but instead focus on Jakob's contributions only, especially on probability theory.

Figure 5: Jacob Bernoulli

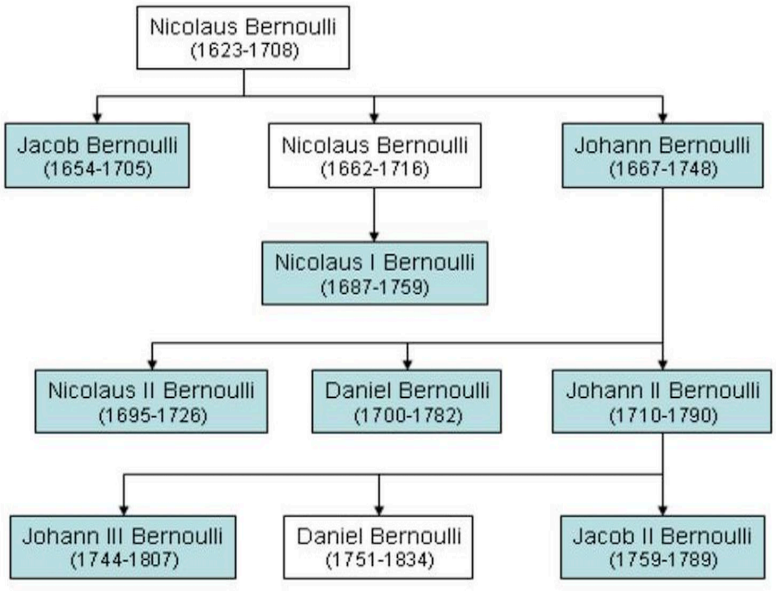


Figure 6: The Bernoullis

Ars Conjectandi consists of 4 parts. The first part is basically a reprint of Huygens' *On Reckoning at Games of Chance* with a bit modification. The second and third parts focus on the theory of permutations and combinations as well as their applications, in which he developed a lot tools and formulas about combinatorial numbers including what now we call "Bernoulli numbers", which are the coefficients B_n satisfying

$$\sum_{i=1}^{n} i^k = \frac{1}{n+1} \sum_{i=0}^{n} B_i \binom{n+1}{i} n^{n+1-i}$$

which now are usually defined as coefficients in the expansion

$$\frac{x}{\mathrm{e}^x - 1} = \sum_{n=0}^{\infty} B_n \frac{x^n}{n!}$$

In the last part of the book, Bernoulli applied the content in the previous parts to the field of probability, stated and proved the well-known **Bernoulli's law of large numbers**. He introduced a different way of viewing probability. Instead of regarding probability as

an objective concept, he views it as subjective lack of information - through a more precise knowledge of parameters of throwing a dice, it would be possible to determine the outcome in advance; different people may describe a same event with different chance depending on their knowledge of information.

He distinguished two different ways of determining probability, exactly or approximately. The former one requires pre-assumptions of equipossibility for elementary events, like drawing either one ball from a bag containing n balls. However there are events that cannot be reduced to elementary and equi-possible ones, which is the second case. Bernoulli stated that we can estimate its probability by experiments - the unknown possibility can be approximated by the occurrence in a series of repeated and independent trials.

This statement is guaranteed by what he called *theorema aureum*("golden theorem") and named as "Bernoulli's law of large numbers" later by Poisson. More precisely, denote h_n as the frequency of an event with (unknown) probability p in repeating n independent trials. Then for arbitrary positive real number ε,δ>0, for sufficient big n, we have

$$\Pr\{|h_n - p| < \varepsilon\} > 1 - \delta.$$

Detailed discussion and rigorous proof are left to next chapter.

Ars Conjectandi wasn't published until 1713 by his nephew Nikolaus Bernoulli. Again, it appeared too late to offer something brand new, and Nikolaus offered a modified proof for the main theorem, the law of large numbers. Nevertheless, it stimulated de Moivre to find his central limit theorem.

1.5 De Moivre and Gaussian distribution

Abraham de Moivre(1667-1754) was a French mathematician, often remembered for his formula for

$$(\cos x + i \sin x)(\cos y + i \sin y) = \cos(x + y) + i \sin(x + y)$$

He pioneered the development of probability theory. He first published *The Doctrine of Chances* in 1718. The 1756 edition of the book contained one of the most significant work by De Moivre, which is the

the approximation of normal distribution by the binomial distribution. In fact, this is a special case of the general **central limit theorem**, which will be carefully studied in Chapter 3. He first published this result in 1733 aiming to improve Bernoulli's law of large numbers. This motivated the study of normal distribution and the central limit theorem. De Moivre also seems to have recognized the parameter now called standard derivation (which is an important parameter for normal distribution).

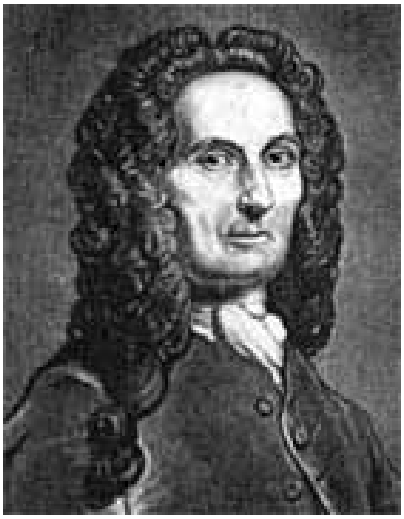


Figure 7: Abraham de Moivre

Moreover, in 1733 he used the Stirling's formula to derive the approximation of normal distribution by binomial.

$$n! \sim \frac{n^n}{e^n}\sqrt{2\pi n} \tag{1}$$

It is first discovered by de Moivre with 2π in the formula being an unknown constant. Later in 1738 de Moivre wrote in his book that his friend Stirling determined the constant 2π in the formula.

The normal distribution is also called as Gaussian distribution in honer of the celebrated German mathematician **Johann Carl Friedrich Gauss**(1777-1855)

Gauss independentlt discovered this distribution in the study of random errors in measurements. Suppose a measurement contains sufficiently many steps that leads to random errors. Then the total

error can be estimated by a random variable with Gaussian distribution - which is a corollary of central limit theorem.

It is also believed that Gauss came up with the law of *least squares* when he studied errors. Nowadays this method is widely used in statistics. With the law of least squares, he argued that the most likely value of a series of measurements is taken to be the arithmetic mean.

Figure 8: Carl Friedrich Gauss

In mathematical language, a normal distribution is a continuous distribution with probability density function

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

whose image looks like a bell curve

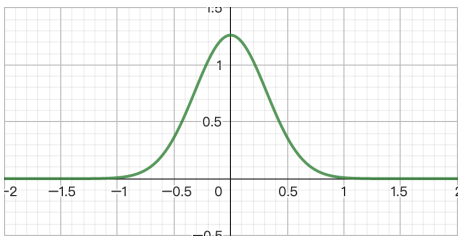


Figure 9: normal distribution

However, so far our theory on probability is not compatible with continuous stuff, because probability with infinitely many events is unacceptable. ($\frac{\infty}{\infty}$isn't well defined.) We'll leave the details in Chapter 3.

1.6 Laplace's contributions

Laplace's work *Théorie Analytique de Probabilités* "summarized the results of the classical probability theory and gave a decisive thrust to its further development", said Rényi. It established the foundation of probability theory in 19th century.

Figure 10: Pierre-Simon Laplace

Pierre-Simon Laplace(1749-1827) was a French mathematician and astronomer. He made significant contributions to analysis, especially in applying analytical tools (calculus for example) to physical problems. In his masterpiece *Méchanique Céleste* he proved the stability of planetary orbits by setting differential equations and solving them to describe the resulting motions. In *Méchanique Céleste*, the equation which was named after him

$$\nabla^2 f = 0 \tag{3}$$

appears and now we define the Laplacian operater $\Delta = \nabla \cdot \nabla$. (Here $\nabla\cdot$ is the divergence operator and ∇ is the gradient operator.) Such results stimulates the study of harmonic functions and potential theory. Laplace equation is also used in heat conduction.

Laplace applied analytical tools in theory of probability as well.

It's an interesting fact that the modern version of what we now call "Bayes's Theorem" is actually presented by Laplace as well. In 1774 he proved that

$$\Pr(A_i|B) = \frac{\Pr(B|A_i)\Pr(A_i)}{\sum_j \Pr(B|A_j)\Pr(A_j)} \tag{4}$$

where A_i's is a complete partition of all possible events. (Or in modern language, $\bigcup_i A_i = \Omega$ the full sample space, and $A_i \cap A_j = \emptyset$.) Here $\Pr(A|B)$ represents the probability of A in the condition that event B happens. Often this theorem is also stated as (a simplified but equivalent version)

$$\Pr(A|B) = \frac{\Pr(B|A)\Pr(A)}{\Pr(B)} \tag{5}$$

Figure 11: Bayes

English statistician **Thomas Bayes**(1702-1761) first discovered a special case of this theorem, hence the name. The problem with

which Bayes concerned himself was the following. Given the number of times in which an unknown event has happened and failed: required the chance that the probability of its happening in a single trial lies somewhere between any two degrees of probability. The solution is given by (expressed in modern notation)

$$\begin{aligned}&\Pr\left(p_1 < p < p_2 | \text{ Event A has happened } m \text{ times out of } m+n \text{ trials}\right)\\&= \int_{p_1}^{p_2} \binom{m+n}{m} x^m (1-x)^n \mathrm{d}x \Big/ \int_0^1 \binom{m+n}{m} x^m (1-x)^n \mathrm{d}x\end{aligned} \tag{6}$$

Here p is the "prior" probability of event A. (However in fact he supposed implicitly that the probability is "uniform")

Bayes gives a new idea of viewing probability (what he himself calls "chance"). He defines it to be the "ratio between the expectation of something uncertain and the value of the thing expected if it happens", which is particularly useful when it comes to conditional probability. Therefore, Bayesian probability (and related concepts including prior/postprior probability) theory is widely used in statistics inference.

However, Laplace chose a different approach to probability.

In 1812, Laplace published *Théorie Analytique de Probabilités*, summarizing his results on probability theory throughout his entire life. Early in 1780s he had been working on analysis (such as generating functions and asymptomatic approximation of definite integrals, which we will see later) that was close to probability. These contents are all included in the first edition of Laplace's *Théorie Analytique*. He begins with a (relatively) acceptable definition of probability. He writes,

…probabilité est relative à la subdivision de tous les cas en d'autres également possibles. Soient N la somme de tous les cas ainsi subdivisés, et n la somme de ces cas qui sont favorables; on aura (Probability is relative to the subdivision of all the cases into others that are equally possible. Let N be the sum of all the cases thus subdivided, and n the sum of the cases that are wanted; we will have)

$$p = \frac{n}{N}$$

In *Théorie Analytique* he discussed some mathematical tools that

he used to solve probability problems, including generating functions and asymptotic approximations (which is still taught in university today).

The most significant result is that he gave a proof of an improved version of De Moivre's central limit theorem, now known as De Moivre-Laplace theorem. It states that the total successes of repeating trials (with same probability p of success) can be approximated by normal distribution. Written in precise and modern language, denote the sum S_n of n Bernoulli random variables (with same parameter p probability of success), then random variable $\frac{S_n - np}{\sqrt{np(1-p)}}$ converges to the standard normal distribution $e^{-\frac{x^2}{2}}$ in probability. Or equivalently,

$$\lim_{n\to\infty} \Pr\left(a < \frac{S_n - np}{\sqrt{np(1-p)}} < b\right) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-x^2/2} \mathrm{d}x \tag{7}$$

This is considered as the most important result in classical probability theory. We will come back to this result in Chapter 3.

Théorie Analytique also covers applied probability theory in physical, economic and sociological problems. For example, Laplace use probability tools to study measure errors, estimate population size, life expectancy, duration of marriages, insurance and more. Probability theory no longer focuses on problems within gambling and games, but also solves real-life problems. This is actually the characteristic of 18th century mathematics - applying analytical results to different fields, concerning physics problems, but making little of rigorous proofs and mathematical foundations.

2 Law of Large Numbers

Early in 16th century Cardano noticed that as one repeat throwing dice over and over again, the proportion of showing up certain points(or other patterns) tends to become stable. It is so natural for people to believe that as the number of trials increases, the sample average tends to converge to theoretical mean, or the mathematical expectation. And this is exactly the story that the **law of large numbers** (**LLN** for short) is telling.

In mathematical language, we say a sequence of random variables $\{X_n\}_{n=1}^{\infty}$ follows the law of large numbers if there exists a sequence $\{a_n\}_{n=1}^{\infty}$ such that

$$\lim_{n\to\infty} \Pr\left\{\left|\frac{1}{n}\sum_{i=1}^{n} X_i - a_n\right| < \varepsilon\right\} = 1 \tag{8}$$

Such law of large numbers is also called the "weak law of large numbers", in contrast with the "strong law of large numbers", in which the condition is strengthened as

$$\Pr\left\{\lim_{n\to\infty}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = a\right\} = 1, \text{ for some fixed } a. \tag{9}$$

As we mentioned in Chapter 1, the first law of large numbers in history was Bernoulli's. At that time, terms like "random variable" and rigorous foundations for probability theory weren't established yet. In convenience, we may use modern terminologies to prove the following theorems; however the key ideas in the proofs are essentially the same, regardless of which language we use. One may check the supplementary materials in the last section (2.5) of this chapter about the inequalities and lemmas we might use without proof.

2.1 Bernoulli's law of large numbers

Let's restate the Bernoulli's law of large numbers in terms of random variables. We say a random variable X follows **Bernoulli distribution** if it has probability p to attain outcome $X = 1$ (success)

and probability $1-p$ to attain $X=0$ (failure). Then Bernoulli's LLN can be stated as:

Independent random variables $\{X_n\}_{n=1}^{\infty}$ *following indentical Bernoulli distribution follows the law of large numbers.*[1]

In this case, $\frac{1}{n}\sum_{i=1}^{n} X_i$ represents the frequency of success trials out of the first n trials.

We may just consider the case when $p \in \mathbb{Q}$, since irrational number can always be approximated by rational numbers, so that

$$\lim_{n\to\infty} \Pr\left\{\left|\frac{1}{n}\sum_{i=1}^{n} X_i - a\right| < \frac{\varepsilon}{2}\right\} = 1$$

for some rational a **s.t.** $|a-p| < \frac{1}{2}$ implies

$$\lim_{n\to\infty} \Pr\left\{\left|\frac{1}{n}\sum_{i=1}^{n} X_i - p\right| < \varepsilon\right\} = 1$$

Here's the original proof by Jakob Bernoulli for p rational. Pick integers t, r, s, **s.t.**

$$t > \frac{1}{\varepsilon}, \quad r = tp, \quad s = t - tp.$$

Consider binomial

$$1 = \left(\frac{r+s}{t}\right)^{nt} = \sum_{i=-nr}^{ns} T_i, \quad T_i = \binom{nt}{nr+i} r^{nr+i} s^{ns-i} t^{-nt}.$$

T_i is the number of possibilities for $nr+i$ successful trials out of nt. So

$$\Pr\left\{\left|\frac{1}{n}\sum_{i=1}^{n} X_i - p\right| \leqslant \frac{1}{t}\right\} = \sum_{i=-n}^{n} T_i$$

Now we need to estimate $\sum_{i=-n}^{n} T_i$. One may verify following facts easily from basic properties of combinatorial numbers.

[1]Sometimes the condition "independent identically distributed" is abbreviated as "**i.i.d.**"

$$T_{-nr} < \cdot < T_{-1} < T_0 > T_1 > \cdot > T_{ns} \tag{10}$$

$$\frac{T_0}{T_1} < \frac{T_1}{T_2} < \cdots < \frac{T_{ns-1}}{T_{ns}}; \quad \frac{T_0}{T_{-1}} < \frac{T_{-1}}{T_{-2}} < \cdots < \frac{T_{-nr+1}}{T_{-nr}} \tag{11}$$

$$\frac{T_0}{T_n} < \frac{T_i}{T_{i+n}}; \quad \frac{T_0}{T_{-n}} < \frac{T_{-i}}{T_{-i-n}}, \text{ for any } i > 0 \tag{12}$$

Then

$$\frac{\sum_{i=1}^{n} T_i}{\sum_{i=n+1}^{ns} T_i} \geqslant \frac{\sum_{i=1}^{n} T_i}{(s-1)\sum_{i=n+1}^{2n} T_i} > \frac{1}{s-1}\frac{T_0}{T_n} \tag{13}$$

where

$$\begin{aligned} \frac{T_0}{T_n} &= \frac{(nr+1)(nr+2)\cdots(nr+n)}{(ns-n)(ns-n+1)\cdots(ns-1)}\frac{s^n}{r^n} \\ &= \prod_{t=1}^{n} \frac{nrs+ts}{nrs-(n+1-t)r} \\ &> \prod_{t=1}^{n}\left(1+\frac{(n+1-t)r+ts}{nrs}\right) > \left(1+\frac{1}{\max\{r,s\}}\right)^n \end{aligned} \tag{14}$$

Similarly,

$$\frac{\sum_{i=-n}^{-1} T_i}{\sum_{i=-nr}^{-n-1} T_i} > \frac{1}{r-1}\frac{T_0}{T_{-n}} > \frac{1}{r-1}\left(1+\frac{1}{\max\{r,s\}}\right)^n \tag{15}$$

So when $n \to \infty$,

$$\frac{\sum_{i=1}^{n} T_i}{\sum_{i=n+1}^{ns} T_i}, \quad \frac{\sum_{i=-n}^{-1} T_i}{\sum_{i=-nr}^{-n-1} T_i} \to \infty$$

Based on this fact, one can deduce

$$\Pr\left\{\left|\frac{1}{n}\sum_{i=1}^{n} X_i - p\right| \leqslant \frac{1}{t}\right\} = \sum_{i=-n}^{n} T_i = \frac{\sum\limits_{i=-n}^{n} T_i}{\sum\limits_{i=-nr}^{ns} T_i} \to 1 \tag{16}$$

as $n \to \infty$, which finishes the proof of Bernoulli's law of large numbers.

2.2 Poisson's law of large numbers

Poisson is the one who first gave the name "law of large numbers". Poisson's LLN is a generalization of Bernoulli's. It doesn't require the variables to be of identical distribution:

Let X_n be a sequence of independent random variables following Bernoulli distribution of probability p_n, then $\{X_n\}_{n=1}^{\infty}$ follows the law of large numbers:

$$\lim_{n\to\infty} \Pr\left\{\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{n}\sum_{i=1}^{n} p_i\right| < \varepsilon\right\} = 1 \tag{17}$$

We are not going to give the proof of Poisson's law of large numbers here, since itself can be seen as a special case of Chebyshev's result. Nevertheless, let's take a look at its history.

Figure 12: Siméon Denis Poisson

Siméon Denis Poisson(1781-1840) is a French mathematician. He wasn't born in a noble family, but bourgeoisie instead. He was eight years old when the French revolution began in July, 1789. His father wanted Poisson to become a surgeon, but Poisson later found out he had no interest in medical profession. In 1796, when France

had become a republic, he enrolled on the École Centrale and soon discovered his talent in academic research. He passed the entrance examinations for the École Polytechnique and began to study mathematics under Laplace and Lagrange.

Poisson's achievements covered almost all of the physics and mathematical analysis of the period. For instance, his works on differential equations of mechanics became the basis for researches of Hamilton 30 years later.

His major work in probability is the book *Recherches sur la probabilité*, which is largely a treatise in tradition of and in sequel of Laplace's *Théorie Analytique*. He proposed the term "Law of Large Numbers"(*Loi des grand nombres* in French) for the first time, as well as his version of Law, which is a generalization of Bernoulli's, as we've mentioned.

As for the Poisson distribution, it may be found in his treatise in 1837. It was named after him although De Moivre had already found the distribution before him. Given λ the expected number of success, the number of Bernoulli trials N approaches $+\infty$ while probability of success in each trial $p = \frac{\lambda}{N}$. Then the distribution of total success approaches

$$\begin{aligned} P_\lambda(n) &= \lim_{N\to\infty} \frac{N!}{n!(N-n)!} p^n (1-p)^{N-n} \\ &= \lim_{N\to\infty} \frac{N(N-1)\cdots(N-n+1)}{N^n} \frac{\lambda^n}{n!} \left(1-\frac{\lambda}{N}\right)^{N-n} \\ &= 1 \cdot \frac{\lambda^n}{n!} \cdot \mathrm{e}^{-\lambda} \\ &= \frac{\lambda^n \mathrm{e}^{-\lambda}}{n!} \end{aligned} \tag{18}$$

which is the Poisson distribution. As expected, $\sum_{n=0}^{\infty} P_\lambda(n) = 1$ and expectation $\sum_{n=0}^{\infty} P_\lambda(n) \cdot n = \lambda$. The distribution is used for calculating the possibilities for an event with average rate of value, or describing events with low probability but large quantity. Thus someone also call it "the law of small numbers."

2.3 Chebyshev's law of large numbers

Figure 13: Pafnuty Chebyshev

Pafnuty Lvovich Chebyshev(1821-1894) was a Russian mathematician, largely remembered for his contributions in number theory and probability theory.

One major contribution of him is the *Chebyshev theorem* in number theory. Proven in 1852, which states that

$$\forall n \geqslant 3, \text{ there is a prime } p, n < p < 2n \tag{19}$$

Chebyshev entered Moscow university in 1837, and he was greatly influenced by professor Nikolai Brashman whose interests were wide ranging, including the calculus of probability.

In 1845 Poisson's works came to Chebyshev's attention. He published his treatise *Demonstration Elémentaire d'une Proposition Générale de la Théorie des Probabilités* in 1846, studying Poisson's results on law of large numbers. In the treatise, he "demonstrates rigorously this proposition by some totally elementary considerations." In 1866 he proved the well known Chebyshev inequality, which can be used to give an even simpler proof for **Chebyshev's law of large numbers**,

to which Poisson's result is just a special case:

$$\Pr\left[\left|\frac{1}{n}\sum_{i=1}^{n}(X_n - E(X_n))\right| < \varepsilon\right] \geqslant 1 - \frac{c}{\varepsilon^2 n}, \tag{20}$$

requiring variance $\mathrm{Var}(X_n) \leqslant c$.

The Chebyshev inequality in probability theory (in modern interpretation) states that

$$\Pr\left[|X - \mu| \geqslant k\right] \leqslant \frac{\sigma^2}{k^2}. \tag{21}$$

where $\mu = E(X)$ is the expectation of the random variable X, meanwhile $\sigma = \sqrt{\mathrm{Var}(X)}$ is the standard deviation.

The inequality was first formulated (without proof) by Chebyshev's friend and colleague Bienaymé in 1853, and later given a proof by Chebyshev.(So sometimes this inequality is also referred to as Bienaymé-Chebyshev inequality.)

In Chebyshev's treatise in 1866 he writes (simplified and translated),

If one designates by $a, b, c, \ldots$ the mathematical expectation of the quantity $x, y, z, \ldots$ and $a_1, b_1, c_1, \ldots$ the mathematical expectation of their squares $x^2, y^2, z^2, \ldots$, then the probability that the sum $x + y + z + \ldots$ is contained between

$$a + b + c + \cdots + \alpha\sqrt{a_1 - a^2 + b_1 - b^2 + c_1 - c^2 + \ldots}$$

$$a + b + c + \cdots - \alpha\sqrt{a_1 - a^2 + b_1 - b^2 + c_1 - c^2 + \ldots}$$

which is essentially the same theorem as the modern interpreted one. And we can also observe that mathematicians have already noticed the additivity of variance (for independent variables). See Section 2.5 for the proof of Chebyshev inequality. Now let's see the proof of Chebyshev's LLN using Chebyshev inequality. Let's restate Chebyshev's result more detailedly once again.

Given independent random variables $\{X_n\}_{n=1}^{\infty}$ *(not essentially identically distributed) whose variances are uniformly bounded by a constant* $c > 0$*, i.e.* $\mathrm{Var}(X_i) \leqslant c$ *for all* X_i*'s, then*

$$\Pr\left[\left|\frac{1}{n}\sum_{i=1}^{n}(X_i - E(X_i))\right| < \varepsilon\right] \geqslant 1 - \frac{c}{\varepsilon^2 n}, \tag{22}$$

and by taking limit, one obtains standard equation for law of large numbers:

$$\lim_{n\to\infty} \Pr\left[\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{n}\sum_{i=1}^{n} E(X_i)\right| < \varepsilon\right] = 1 \tag{23}$$

Essentially it suffices to prove 22. Apply Chebyshev inequality directly, one obtains

$$\begin{aligned}
&\Pr\left[\left|\frac{1}{n}\sum_{i=1}^{n}(X_n - E(X_n))\right| < \varepsilon\right] \\
=&\Pr\left[\left|\sum_{i=1}^{n} X_i - E\left(\sum_{i=1}^{n} X_i\right)\right| < n\varepsilon\right] \\
\geqslant& 1 - \frac{\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right)}{(n\varepsilon)^2} \text{ (apply Chebyshev inequality)} \\
=&1 - \frac{\sum_{i=1}^{n}\mathrm{Var}(X_i)}{(n\varepsilon)^2} \text{ (by independence)} \\
\geqslant& 1 - \frac{nc}{(n\varepsilon)^2} = 1 - \frac{c}{\varepsilon^2 n}.
\end{aligned} \tag{24}$$

And we are done with the proof. This method is also known as the method of "moments", which is used by later mathematicians in many other problems.

In the proof (24) of Chebyshev's law of large numbers, we used the independence condition for random variables X_i so that $\mathrm{Var}(\sum X_i)$ can be decomposed into $\sum \mathrm{Var}(X_i)$, which can be controlled. Our goal is to control $\mathrm{Var}(\sum X_i)$, so that as long as we have

$$\lim_{n\to\infty} \frac{\mathrm{Var}(\sum_{i=1}^{n} X_i)}{n^2} = 0 \tag{25}$$

then $\{X_n\}$ follows LLN. This result is known as **Markov's law of large numbers**:

Given random variables $\{X_n\}_{n=1}^{\infty}$ (not essentially identically distributed or independent) satisfying condition (25) then

$$\lim_{n\to\infty} \Pr\left[\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{n}\sum_{i=1}^{n} E(X_i)\right| < \varepsilon\right] = 1.$$

2.4 From weak LLN to strong LLN

Khinchin's LLN is the most general law of large numbers for independent identically distributed (i.i.d.) random variables, stated as following:

Given i.i.d. random variables $\{X_n\}_{n=1}^{\infty}$ with expectation satisfying expectation $\mu = \mathrm{E}(X) < +\infty$ then $\{X_n\}_{n=1}^{\infty}$ follows the law of large numbers.

It's notable that Khichin's law of large numbers doesn't require X to have a finite variance. In other words, $\mathrm{Var}(X) = +\infty$ is acceptable (which is not acceptable in Chebyshev's law of large numbers.)

It is a very late LLN, compared to other versions, not proven until 20th century. However this result can also be further strengthened to *strong law of large numbers*, known as **Kolmogorov's Strong Law of Large Numbers**:

$\{X_n\}$ is a sequence of i.i.d. random variables with expectation $\mathrm{E}(X_i) = \mu < \infty$, then

$$\Pr\left[\lim_{n\to\infty} X_n = \mu\right] = 1$$

However this expression of the limit of random variables is not rigor without axiomatic probability theory (introduced in Chapter 5)

This result can also be further strengthened for X_n aren't necessarily identically distributed. The condition (sometimes called the Kolomogorv criterion)

$$\sum \frac{\sigma_k^2}{k^2} < +\infty$$

is sufficient.

The result above is proven by William Filler in 1968.

2.5 Supplementary material: Terminology and Definition

This section serves as the appendix of mathematical tools for Chapter 2. Most of terms are defined in modern language.

Classical probability theory only handles finite sample spaces. Nevertheless, similar concepts and terms can be applied to modern axiomatic ones (as long as we take integrals instead of sums). So we'll focus on finite ones in this section.

A (discrete) **random variable** X is a function assigning each real number to a non-negative number (the probability of obtaining that number) such that $\sum_{x\in\mathbb{R}} \Pr[X = x] = 1$. It's easy to see that at only countably many points X has non-zero probability.

The **expectation** of (discrete) random variable X is defined to be

$$\mathrm{E}(X) = \sum_{x\in\mathbb{R}} x \cdot \Pr[X = x]$$

which is an intuitive definition. However this (infinite) sum may not exist. So we usually only consider those random variable X with $\sum_{x\in\mathbb{R}} |x| \cdot \Pr[X = x] < +\infty$. Similar assumptions apply to variance as well.

The **variance** of X is defined to be

$$\mathrm{Var}(X) = \mathrm{E}\Big((X - \mathrm{E}(X))^2\Big) = \mathrm{E}(X^2) - (\mathrm{E}(X))^2$$

The **standard deviation** (usually denoted as σ) of X is defined to be

$$\sigma = \mathrm{SD}(X) = \sqrt{\mathrm{Var}(X)}$$

When it comes to two (discrete) random variables, for example X, Y, (they may have some relations), can be described as a function $\mathbb{R}^2 \to \mathbb{R}_{\geqslant 0}$

In this sense, we can define the sum of two random variables

$$\Pr[X + Y = x] = \sum_{a+b=x} \Pr[X = a, Y = x - a]$$

Similarly we can define products and so on.

It is easy to check that expectation is linear, i.e. $\mathrm{E}(X+Y) = \mathrm{E}(X)+\mathrm{E}(Y)$.

We say two random variables are **independent** if

$$\Pr[X=x, Y=y] = \Pr[X=x]\cdot\Pr[Y=y], \text{ for all } x, y.$$

An intuitive understanding of "independence" is that the value of one variable doesn't affect the other. Similarly we can define the independence for more than 2 variables.

If some two random variables satisfy $\mathrm{E}(X)\mathrm{E}(Y) = \mathrm{E}(XY)$, then we say they are **uncorrelated**.

The **covariance** of two variables is defined as

$$\mathrm{Cov}(X,Y) = \mathrm{E}\Big((X-\mathrm{E}(X)(Y-\mathrm{E}(Y)))\Big) = \mathrm{E}(XY)-\mathrm{E}(X)\mathrm{E}(Y)$$

So $\mathrm{Cov}(X,Y)=0$ is equivalent to X, Y are uncorrelated.

Notice that, independence implies uncorrelation:

$$\begin{aligned}
\mathrm{E}(XY) &= \sum_{t\in\mathbb{R}} t\cdot\sum_{ab=t}\Pr[X=a,Y=b] \\
&= \sum_{a\in\mathbb{R},b\in\mathbb{R}} ab\cdot\Pr[X=a,Y=b] \\
&= \Big(\sum_{a\in\mathbb{R}} a\cdot\Pr[X=a]\Big)\Big(\sum_{b\in\mathbb{R}} b\cdot\Pr[Y=b]\Big) = \mathrm{E}(X)\mathrm{E}(Y)
\end{aligned}$$

We are also interested in the variance of the sum of random variables:

$$\begin{aligned}
\mathrm{Var}\Big(X+Y\Big) &= \mathrm{E}\Big((X+Y)^2\Big) - \Big(\mathrm{E}(X+Y)\Big)^2 \\
&= \mathrm{E}(X^2)+2\mathrm{E}(XY)+\mathrm{E}(Y^2) - \Big(\mathrm{E}(X)+\mathrm{E}(Y)\Big)^2 \\
&= \mathrm{E}(X^2)-\mathrm{E}(X)^2+\mathrm{E}(Y^2)-\mathrm{E}(Y)^2+2\mathrm{E}(XY)-2\mathrm{E}(X)\mathrm{E}(Y) \\
&= \mathrm{Var}(X)+\mathrm{Var}(Y)+2\mathrm{Cov}(X,Y)
\end{aligned}$$

Therefore, when two variables are uncorrelated (or in particular, independent), the variance of the sum equals to the sum of variance. This result also holds for more than two independent variables, and is used in the proof (24).

The **distribution function** F_X of a random variable X (not necessarily discrete) is defined as

$$F_X(a) = \Pr[X \leqslant a] \tag{26}$$

Sometimes people encounter so called "probability density function" for "continuous" random variables. By this people mean an integrable function f on $\mathbb{R}$, satisfying

(1) non-negative, i.e. $f \geqslant 0$ everywhere on $\mathbb{R}$ and

(2) $\int_{\mathbb{R}} f = 1$.

(3) The probability for variable X to belong to some interval (a, b) is given by $\Pr[a < X < b] = \int_a^b f$

In this case, the integral of the density function $F_X = \int f$ is exactly the distribution function.

In this sense, we can also define expectation $\mathrm{E}(X) = \int_{\mathbb{R}} x \cdot f(x)$, and also the sum, product of variables and so on, satisfying all the properties just as the discrete variables.

The general normal distribution $X = \mathcal{N}(\mu, \sigma)$ with density function

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}} \tag{27}$$

is an example. We can calculate its expectation and variance

$$\mathrm{E}(X) = \int_{\mathbb{R}} x \cdot f(x) = \mu + \int_{\mathbb{R}} \frac{x-\mu}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}} = \mu.$$

$$\begin{aligned}
\mathrm{Var}(X) &= \int_{\mathbb{R}} (x-\mu)^2 \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}} \\
&= \int_{\mathbb{R}} \frac{\sigma^2}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}} - \int_{\mathbb{R}} ((x-\mu)^2 - \sigma^2) \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}} \\
&= \sigma^2 - \left((x-\mu) \frac{\sigma^2}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}} \right) \Bigg|_{-\infty}^{+\infty} = \sigma^2
\end{aligned}$$

As you can see, the μ and σ in the expression $\mathcal{N}(\mu, \sigma)$ respectively represent exactly the expectation and the standard deviation.

Now we give some common inequalities in probability theory.
Markov inequality is stated as:

X is a non-negative random variable, given $a > 0$ then

$$\Pr[X \geqslant a] \leqslant \frac{\mathrm{E}(X)}{a} \tag{28}$$

The proof is quite easy:

$$\begin{aligned}\mathrm{E}(X) &= \sum_{x\in\mathbb{R}} \Pr[X = x] \cdot x = \sum_{x\geqslant 0} \Pr[X = x] \cdot x \\ &\geqslant \sum_{x\geqslant a} \Pr[X = x] \cdot x \geqslant \sum_{x\geqslant a} \Pr[X = x] \cdot a \\ &= a \cdot \Pr[X \geqslant a]\end{aligned}$$

Now comes the **Chebyshev inequality** (21), which is equivalent to

$$\Pr[|X - \mathrm{E}(X)| \geqslant k] \leqslant \frac{\mathrm{Var}(X)}{k^2}. \tag{29}$$

Apply Markov inequality (28) to random variable $(X - \mathrm{E}(X))^2$

$$\Pr\Big[\Big(X - \mathrm{E}(X)\Big)^2 \geqslant k^2\Big] \leqslant \frac{\mathrm{E}\Big((X - \mathrm{E}(X))^2\Big)}{k^2} = \frac{\mathrm{Var(X)}}{k^2}$$

finishes the proof.

Both LLN and CLT study the "convergence" of random variables. So it is important to introduce different modes of convergence in probability theory.

We say a sequence of random variables $\{X_n\}_{n=1}^{\infty}$ converges to X **in probability** if

$$\forall \varepsilon > 0, \lim_{n\to\infty} \Pr\Big[|X_n - X| > \varepsilon\Big] = 0$$

This is the mode of convergence in weak law of large numbers.

We say a sequence of random variables $\{X_n\}_{n=1}^{\infty}$ converges to X **almost surely** if

$$\Pr\Big[\lim_{n\to\infty} X_n = X\Big] = 1$$

or equivalently,

$$\lim_{n\to\infty} \Pr\left[\bigcap_{m=n}^{\infty} |X_m - X| < \varepsilon\right] = 1$$

which is the mode of convergence for strong law of large numbers. However this is not a well-defined concept unless we introduce modern language for probability theory. (What is the convergence of random variables?)

We say a sequence of random variables $\{X_n\}_{n=1}^{\infty}$ converges to X **in distribution** if

$$\forall a \in \mathbb{R}, \lim_{n\to\infty} \Pr[X_n \leqslant a] = \Pr[X \leqslant a]$$

This is a very weak version of convergence, weaker than convergence in probability. However we'll see the mode in the following introduction to central limit theorem.

There are also many other types of convergence for random variables; for we're introducing the history rather than probability theory itself, there's no need to list them all.

3 Central Limit Theorem

Central limit theorem (CLT for short) is considered one of the most important theorems in probability theory. It indicates that the distribution of normalized sample mean converges to standard normal distribution, even if the original variable itself is not normally distributed. This is also why normal distribution called "normal". It appears everywhere as long as you take average of a large number of samples.

The law of large numbers already tells us the sample mean approaches the expectation in probability. But the central limit theorem moreover tells us the distribution of sample average near the expectation. Similar to LLN, it is a collective term for many different versions.

Given independent random variables $\{X_n\}_{n=1}^{\infty}$ then the sum $S_n = \sum_{i=1}^{n} X_n$ has expectation

$$\mathrm{E}(S_n) = \sum_{i=1}^{n} \mathrm{E}(X_n)$$

and variance

$$\mathrm{Var}(S_n) = \sum_{i=1}^{n} \mathrm{Var}(X_n)$$

The CLT states that, if X_n satisfies some good properties, then S_n will approach normal distribution. In precise, the normalized variable

$$\frac{S_n - \mathrm{E}(S_n)}{\sqrt{\mathrm{Var}(S_n)}} \to \mathcal{N}(0,1)$$

approaches the standard normal distribution with expectation $\mu = 0$ and variance $\sigma^2 = 1$. ($\mathcal{N}(0,1)$ stands for normal distribution (27))

Now let's take a look at the different versions of central limit theorem in history.

3.1 De Moivre-Laplace Theorem

The earliest CLT is given by De-Moivre when he study the results of flipping fair coins. The result can be completely described by

binomial distribution:

$$\Pr[X = k] = \frac{1}{2^n}\binom{n}{k}$$

Then he estimated using stirling's formula that this distribution can be approximated by a normal curve.

Laplace improved his result for general "unfair coins" in his *Théorie Analytique* in 1812, although he didn't list it out as a theorem. We have mentioned this result (7) in Chapter 1. In this case the binomial distribution has parameter $p, q = 1 - p$:

$$\Pr[X = k] = \binom{n}{k} p^k q^{n-k}$$

That is to say, independent identically distributed Bernoulli variables $\{X_n\}$ follows central limit theorem. It's the first time people became aware of the generality and importance of normal distribution.

Now this result is known as the De Moivre-Laplace theorem:

Given independent Bernoulli random variables $\{X_n\}_{n=1}^{\infty}$ *with probability* p *to obtain* $X = 1$ *and* $1 - p$ *to obtain* $X = 0$*, denote* $S_n = \sum_{i=1}^{n} X_i$ *(which follows binomial distribution), then*

$$\frac{S_n - np}{\sqrt{np(1-p)}} \to \mathcal{N}(0, 1)$$

In the proof, Laplace introduced the method of Fourier transformations into probability theory. What we now call the "characteristic function" of a random variable X is Fourier transformation

$$\varphi_X(t) = \mathrm{E}(e^{itX}) = \int_{-\infty}^{\infty} e^{itX} f(x)\mathrm{d}x = \int_{-\infty}^{\infty} e^{itX}\mathrm{d}F_X(x) \tag{30}$$

where $f(x)$, $F_X(x)$ are density function and distribution function. (see (26))

Although the theorem itself can be proved using elementary approaches without Fourier transformation and characteristic function, this method is useful in proving the latter, more complex and general versions of central limit theorem and other theorems in probability theory.

Now let's sketch a direct proof of de Moivre-Laplace theorem. By using Stirling's formula

$$\begin{aligned}
\binom{n}{k}p^k q^{n-k} &= \frac{n!}{k!(n-k)!}p^k q^{n-k} \\
&\simeq \frac{n^n e^{-n}\sqrt{2\pi n}}{k^k e^{-k}\sqrt{2\pi k}(n-k)^{n-k}e^{-(n-k)}\sqrt{2\pi(n-k)}}p^k q^{n-k} \\
&= \sqrt{\frac{n}{2\pi k\,(n-k)}}\frac{n^n}{k^k\,(n-k)^{n-k}}p^k q^{n-k} \\
&= \sqrt{\frac{n}{2\pi k\,(n-k)}}\left(\frac{np}{k}\right)^k\left(\frac{nq}{n-k}\right)^{n-k} \\
&\simeq \frac{1}{\sqrt{2\pi npq}}\left(\frac{np}{k}\right)^k\left(\frac{nq}{n-k}\right)^{n-k}
\end{aligned}$$

Finally we use the Taylor series of $\ln(1+x)$ at $x=0$. We substitute $x=\frac{k-np}{\sqrt{npq}}$.

$$\begin{aligned}
&\ln\left(\binom{n}{k}p^k q^{n-k}\right) \\
\simeq &-\ln(\sqrt{2\pi npq}) + \left(-k\ln\left(\frac{k}{np}\right) + (k-n)\ln\left(\frac{n-k}{nq}\right)\right) \\
= &-\ln(\sqrt{2\pi npq}) - k\left(\frac{np+x\sqrt{npq}}{np}\right) + (k-n)\ln\left(\frac{n-np-x\sqrt{npq}}{nq}\right) \\
= &-\ln(\sqrt{2\pi npq}) - k\ln\left(1+x\sqrt{\frac{q}{np}}\right) + (k-n)\ln\left(1-x\sqrt{\frac{p}{nq}}\right) \\
\simeq &-\ln(\sqrt{2\pi npq}) - k\left(x\sqrt{\frac{q}{np}} - \frac{x^2 q}{2np}\right) + (k-n)\left(-x\sqrt{\frac{p}{nq}} - \frac{x^2 p}{2nq}\right) \\
= &-\ln(\sqrt{2\pi npq}) - (np+x\sqrt{npq})\left(x\sqrt{\frac{q}{np}} - \frac{x^2 q}{2np}\right) \\
&-(nq-x\sqrt{npq})\left(-x\sqrt{\frac{p}{nq}} - \frac{x^2 p}{2nq}\right) \\
= &-\ln(\sqrt{2\pi npq}) - \frac{1}{2}x^2
\end{aligned}$$

Which finishes the proof.

3.2 Classical Central Limit Theorem

It is a very natural guess that the Bernoulli variable in de Moivre-Laplace Theorem can be replaced with arbitrary random variable with finite expectation and variance. Mathematicians have been concerning this question after Laplace. Poisson, Chebyshev, Markov all provided their versions of CLT and their proofs. Now we know the answer is *YES*. And people often use (Classical) central limit theorem to refer to this result, i in contrast with the other theorems (like Lindeberg's) that we will cover later.

The CLT is stated as following:

Random variables $\{X_n\}_{n=1}^{\infty}$ *i.i.d., with* $\mathrm{E}(X_n) = \mu < +\infty, \mathrm{Var}(X_n) = \sigma^2 < +\infty$. *Define* $Y_n^* = \dfrac{\sum_{i=1}^{n} X_n - n\mu}{\sqrt{n}\sigma}$. Then

$$\lim_{n\to\infty} \Pr[Y_n^* < a] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a} e^{-\frac{t^2}{2}} \mathrm{d}t \tag{31}$$

This theorem is also called *Lindeberg-Lévy CLT*, named after Finnish mathematician **Jarl Waldemar Lindeberg**(1876-1932) and French mathematician **Paul Pierre Lévy**(1886-1971).

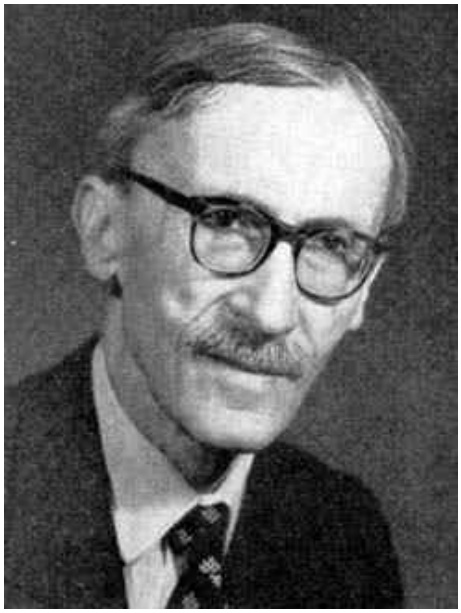


Figure 14: Paul Lévy

The proof of theorem requires the method of characteristic function (30)

We may use the fact that the convergence of characteristic functions imply the convergence of probability in distribution. By computing $\varphi_{Y_n^*}(t)$ we obtain

$$\begin{aligned}\varphi_{Y_n^*}(t) &= \mathrm{E}(e^{itY_n^*}) = \mathrm{E}\Big(\exp\Big(it\sum_{k=1}^{n}\frac{X_i-\mu}{\sqrt{n}\sigma}\Big)\Big)\\ &= \prod_{k=1}^{n}\mathrm{E}\Big(\exp\Big\{it\frac{X_i-\mu}{\sqrt{n}\sigma}\Big\}\Big) = \Big(\varphi_{X_k-\mu}\Big(\frac{t}{\sqrt{n}\sigma}\Big)\Big)^n\\ &= \Big(1+\frac{1}{2}\varphi''_{X_k-\mu}(0)\Big(\frac{t}{\sqrt{n}\sigma}\Big)^2+o(\frac{1}{n})\Big)^n\\ &= \Big(1-\frac{1}{2}\mathrm{Var}(X_k)\frac{t^2}{n\sigma^2}+o(\frac{1}{n})\Big)^n \to e^{-\frac{1}{2}t^2}\end{aligned}$$

And this finishes the proof.

The CLT above can actually be generalized to cases when variables are not identically distributed. It's in fact a special case of *Lindeberg-Feller Theorem.*

3.3 Lindeberg's CLT and More

Lindeberg's theorem focus on *triangular arrays* of random variables, which is of the form

$$\begin{array}{llll} X_{11} & & & \\ X_{21} & X_{22} & & \\ X_{31} & X_{32} & X_{33} & \\ \dots & & & \end{array}$$

Where the random variables in each row are independent, with zero expectation and finite variance. What we are interested in is the sum of each row, $S_n = \sum_{i=1}^n X_{ni}$. Obviously $\mathrm{E}(Z_n) = 0$ and $s_n^2 = \mathrm{Var}(S_n) = \sum_{i=1}^n \mathrm{Var}(X_{ni}) = \sum_{i=1}^n \sigma_{ni}^2$

The *Lindeberg condition* on a triangular array of random variables $\{X_{mn}\}$ is the following:

$$\frac{1}{s_n^2}\sum_{i=1}^{n}\mathrm{E}\Big(X_{ni}^2\cdot 1_{|X_{ni}|\geqslant\varepsilon s_n}\Big) = \frac{1}{s_n^2}\sum_{i=1}^{n}\int_{|X_{ni}|\geqslant\varepsilon s_n} X_{ni}^2 \to 0 \qquad (32)$$

Lindeberg-Feller's CLT states that if $\{X_{mn}\}$ satisfies the Lindeberg condition, then $S_n^* = \dfrac{S_n}{s_n}$ converges to standard normal distribution $\mathcal{N}(0,1)$ in distribution. The result is proven by William Feller, using characteristic functions. We may omit the proof.

It is notable that the Lindeberg condition is somewhat an equivalent condition for central limit theorem. That is to say, Lindeberg's CLT is the strongest version. Feller proved that for triangular array $\{X_{ni}\}$ satisfying $S_n^* \to \mathcal{N}(0,1)$ and $\max\limits_i \sigma_{ni}^2/s_n^2 \to 0$, then the Lindeberg condition must hold.

Lyapunov's CLT is a corollary of Lindeberg's CLT. Given $\delta > 0$, the Lyapunov's condition for a triangular array of random varibale $\{X_{mn}\}$ is

$$\lim_{n\to\infty} \frac{1}{s_n^{2+\delta}} \sum_{i=1}^{n} \mathrm{E}\Big(|X_{ni}|^{2+\delta}\Big) = \lim_{n\to\infty} \frac{1}{s_n^{2+\delta}} \sum_{i=1}^{n} \int_{|X_{ni}|\geqslant \varepsilon s_n} X_{ni}^2 = 0$$

If a triangular array $\{X_{mn}\}$ satisfies Lyapunov's condition, then $S_n^* \to \mathcal{N}(0,1)$ in distribution.

However, all the discussions above restricts to 1-dimensional variables, i.e. random variables on $\mathbb{R}$. Central limit theorem applies to higher dimensional random variables as well. $X = (X_1, X_2, \ldots, X_k)$ is a k-dimensional variable. Similarly we can define characteristic function

$$\varphi(t) = \mathrm{E}(e^{it\cdot x}) = \mathrm{E}\Big(\exp\Big(i\sum_{j=1}^{k} t_j X_j\Big)\Big)$$

where $t = (t_1, \ldots, t_k)$

A random varibale X follows *multivariable normal distribution* if there is a vector μ and a positive definite matrix Σ, s.t. $\varphi(t) = \exp(i\mu^T t - \frac{1}{2}t^T\Sigma t)$. The standard multivariable normal distribution $\mathcal{N}(0, I)$ takes $\Sigma = I$ the unit matrix, so that the probability distribution is

$$f(x) = \frac{1}{(2\pi)^{\frac{k}{2}}} e^{-\frac{||x||^2}{2}}$$

In this sense, the classical central limit theorem can be generalized to multivarible versions. For i.i.d. random variables $\{X_n =$

$(X_{n1}, \ldots, X_{nk})\}_{n=1}^{\infty}$ with expectancy μ and finite second momentum (analog to variance). Let $S_n = X_1 + \cdots + X_n$, then we have:

$$\frac{S_n = n\mu}{\sqrt{n}} \to \mathcal{N}(0, \Sigma)$$

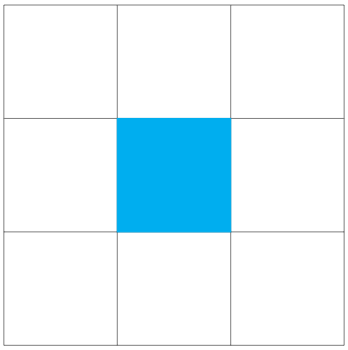
in probability, for some matrix Σ.

There are also many versions of generalizations of CLT in different senses. For example in cases when variables are dependent. We won't list each of them out.

4 Geometric Probability

Laplace pioneered in the fields of analytic probability theory. He influenced the main direction of probability theory research in later ages. The methods of many mathematicians like Poisson and Chebyshev are basically within the framework constructed by him.

However, not all probability problems can be well-solved in classical theory of probability. A good example is the geometric probability, for the foundation of his theory is limited to discrete cases and is unable to handle infinite cases. It's a different model of probability (from Laplace's). People have noticed it very early, but no rigor theory has been established.



If one one throws dart *randomly* at a target with a 3×3 grid (as shown above), what would be the probability that the dart hits the middle grid which is colored cyan? It's very natural to believe that as long as the probability distribution is "uniform", the chance that the dart hits the middle grid would be $\frac{1}{9}$, the area ratio.

As you can see, the essence of geometric probability is "area", in other words, integral, or measure. This can also explain why the modern axiomatic probability theory (in Chapter 5) is based on measure theory,

4.1 Buffon's needle problem

French scientist **Georges Louis Leclerc Comte de Buffon**(1707-1788) is one of the earliest ones who study geometric probabilities (although he himself might not think it's anything pioneering). He

was born in a wealthy family. Besides mathematician, he was also a botanist and naturalist. He aimed to publish 50 volumes of works on nature history (*Histoire naturelle, générale et particulière*), but only finished 36 by the time of his death.

His major contribution to probability is an experiment that he proposed in 1733, now named after him, called the Buffon's needle experiment.

Figure 15: Buffon

Buffon's needle problem asks for the probability for arbitrarily thrown needles to intersect the (equi-distantly distributed) parallel lines on a floor.

We may assume the length of the needle is l and the distance between parallel lines is d, as shown in figure 16.

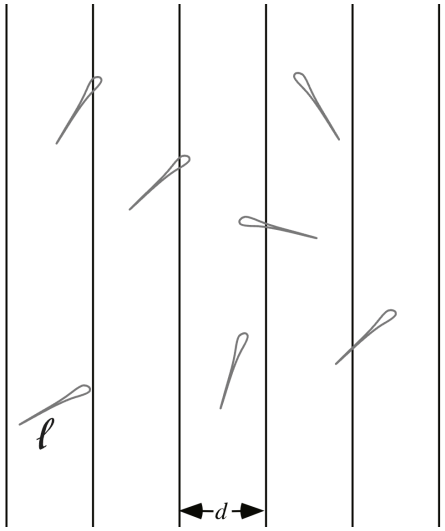


Figure 16: Buffon's Needle Problem

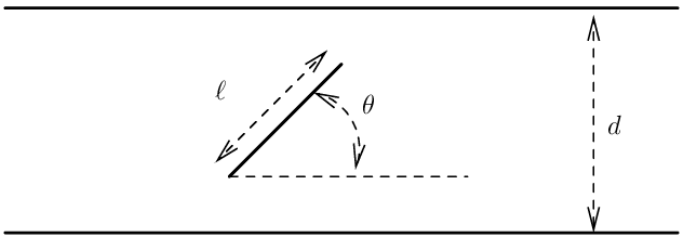
In the special case when $l = \frac{d}{2}$, the probability

$$p = \frac{1}{\pi}. \tag{33}$$

Buffon himself did an experiment with 2212 throws and 702 of them intersects, with the inverse of probability $2212/702 \simeq 3.142$.

Buffon provided an explanation in his work in 1777.

In convenience we assume $l \leqslant d$ so that every needle doesn't intersect more than 1 lines (which is a slightly more complicated case).



For the needles are thrown randomly, the tilted angle θ can be seen as a (continuous) random variable that is equally distributed in $[0, 2\pi)$. Then the projection of the needle is of length $l|\sin\theta|$. For the position of the needle is random, the probability for the needle (in this angle) to intersect the parallel lines should be $\frac{l|\sin\theta|}{d}$, so that the net

probability should be

$$\int_0^{2\pi} \frac{l|\sin\theta|}{d}\frac{\mathrm{d}\theta}{2\pi} = \frac{2l}{\pi d}$$

which explains the phenomenon 33. As Bernoulli's law of large numbers have shown, as the number of independent trails increase, the frequency converges (in probability) to theoretical value.

This experiment provided mathematicians a new way to calculate π or other probability: to generate a large number of random data to simulate the probability a random process. This method of approximating probability is now called **Monte Carlo method**, and is widely used in statistics and computer simulations.

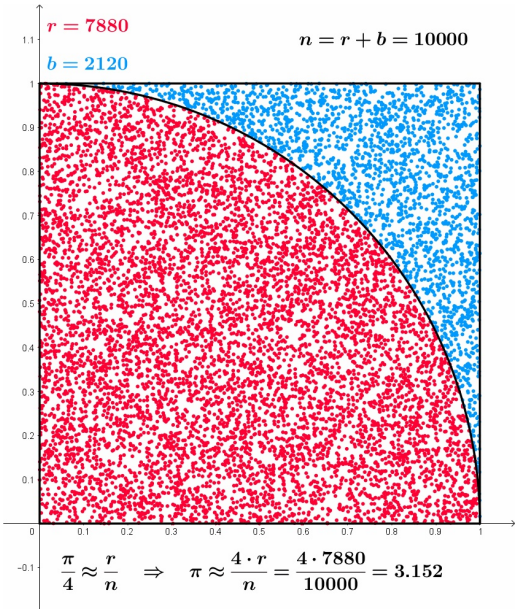


Figure 17: Monte Carlo Method, calculating π

4.2 Bertrand's paradox

Although such explanation of (geometric) probability is intuitive, sometimes it leads to paradoxes. In 1889, French mathematician **Joseph Louis François Bertrand**(1822-1900) introduced one of the paradoxes.

Figure 18: Joseph Bertrand

Joseph Bertrand was the son of physician Alexandre Jacques François Bertrand and the brother of archaeologist Alexandre Bertrand.

Joseph translated Gauss's work on theory of errors and method of least squares into French. He also worked on number theory: In 1845 he conjectured that there is at least one prime number between n and $2n-2$ for every $n > 3$, which was proven by Chebyshev (19).

Bertrand's problem asks the probability of a randomly chosen chord of a circle to be longer than $\sqrt{3}$ times of the circle's radii (in other words, longer than the side of an inscribed equilateral triangle.)

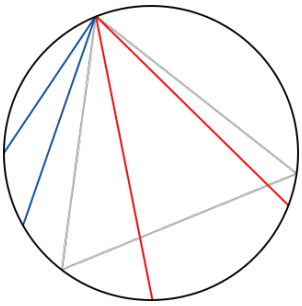


Figure 19: method 1

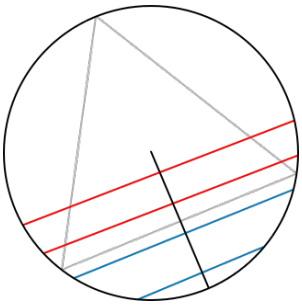


Figure 20: method 2

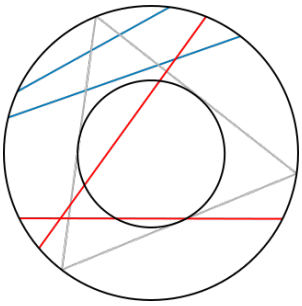


Figure 21: method 3

Figure 22: Different methods of calculating probability (from wikipedia)

Bertrand provided three approaches to this problem.

The first one starts with choosing a random endpoint, and then the other. Observe that the chord is longer than $\sqrt{3}$ times radii if and only if the other endpoint lies on the arc between the endpoints of the triangle side opposite to the first point. The probability is proportional to arc length, hence the probability that a random chord is longer than $\sqrt{3}r$ is $\frac{1}{3}$

The second approach is to choose a random radius and a random point a the radius, then construct the chord through this point and perpendicular to the radius. Observe that the chord is longer than the side of triangle if and only if the chosen point is nearer to the center than the point where the side of triangle intersects the radius. Therefore, the probability should be $\frac{1}{2}$.

The last method begins with choosing a random point anywhere within the circle and the construct a chord with the chosen point being its midpoint. In this case, the constructed chord is longer than $\sqrt{3}r$ only if the chosen midpoint falls inside a concentric circle with radius $\frac{1}{2}r$, and with area being $\frac{1}{4}$ of the larger one, so that the probability is $\frac{1}{4}$

We obtain different answers with different approaches, so what's wrong?

All the calculations above are correct. The problem is our definition of "random". When there are (uncountably) infinitely many cases, we can't assign each specific case to a non-zero probability (in other words, the probability for every single event is zero). Essentially, we are doing integrals, (although we often do this implicitly), and we need a integral measure. We actually used different measures in the three different approaches, which leads to the different answers. None of them is wrong, however, they are answers of different questions. Such paradox is solved by probability based on measure theory instead of classical probability theory.

5 Modern Probability Theory

Starting in late 19th century, mathematicians began to pay attention to the rigor in analysis. The axiomatization for set theory and foundations of mathematics, as well as the need of rigorous theory in statistics and other sciences stimulated mathematicians to find rigorous axiomatic foundations for probability theory. In 1900, Hilbert proposed his famous 23 problems, in which the sixth problem asks for the axiomatization in probability theory and mechanics. The greatest contribution to this problem is made by Russian mathematician Kolmogorov in 1930s. Modern probability theory is built on real analysis and measure theory.

In modern language, a **probability space** $(\Omega, \mathcal{A}, \mathbf{P})$ consists of a *sample space* Ω, an *event space* $\mathcal{A}$, and a probability function (measure) $\mathbf{P}$ on $\mathcal{A}$, satisfying certain conditions (or axioms).

In this sense, both Laplace's classical probability theory and different types of geometric probability can seen as explicit models of axiomatic ones. In other words, they just realize the axioms in different ways.

After the proposal of axiomatic theory, probability theory has developed fast; stochastic process, Markov process, martingale theory and other fields of probability theory has emerged and grown rapidly till today. Probability theory has been applied to different sciences including statistics, physics, biology and more, and it is still under development.

5.1 Measure theory

We'll briefly review the development of measure from 19th century to 20th century. The need of a more general integral theory stimulated the development of measure theory. The first definition of the measure of an arbitrary set was given by **Cantor** (in 1883) and **Stolz** (1884), and supplemented by **Peano** (in 1887) and **Jordan** (in 1892).

In 1898, **Emile Borel** formulated postulates which became the rules for defining measures (of sets) in his book *Leçons sur la Théorie des Fonctions*, which are as follows:

1. A measure is nonnegative

2. The measure of the disjoint union of finite (or countable) number of sets equals the sum of their measures

3. The measure of the difference of two sets equals to the difference of measures.

4. The measure of a set consisting of one point is zero

All the properties above are very natural for what we would expect the "measure" (length, in particular) to have. However, Borel's definition of length is a descriptive one, which leads to the question of existence of such object. Such question is solved by **Lebesgue**.

In the Lebesgue's thesis published in 1902, he proposed the modern description of "Lebesgue measure".

Firstly, we define the measure of interval $[a, b]$ (or (a, b)) is $(b-a)$, then the measure $m(E)$ of open sets in $\mathbb{R}$ can be defined.

Given an arbitrary set $E \subset \mathbb{R}$, we may define the *outer measure* of E:

$$m_e(E) = \inf_{\text{open } G \supset E} m(G)$$

Similarly outer measure of E:

$$m_i(E) = \sup_{\text{open } G \subset E} m(G)$$

A set $E \subset \mathbb{R}$ is said to be measurable if $m_e(E) = m_i(E)$, and we call this value to be the *Lebesgue measure* of E.

In general, a measurable space $(X, \mathcal{A})$ consists of a set X and $\mathcal{A} \subset 2^X$ satisfying:

1. $\mathcal{A}$ is closed under complement (of X), and $X \in \mathcal{A}$

2. $\mathcal{A}$ is closed under countable union and countable intersection.

Such $\mathcal{A}$ is also called a *σ-algebra on X*.

A **measure** on a measurable space $(X, \mathcal{A})$ is a function $\mu : \mathcal{A} \to \mathbb{R} \cup \{+\infty\}$ satisfying the following conditions (which are similar to previous (1)-(4)):

- $\mu(E) \geqslant 0, \forall E \in \mathcal{A}$
- $\mu(\varnothing) = 0$

- For any countable collection $\{E_n\}_{n=1}^{\infty}$ of disjoint sets,
  $\mu\Big(\bigcup_{n=1}^{\infty} E_n\Big) = \sum_{n=1}^{\infty} \mu(E_n)$

A map f from one measurable space $(X, \mathcal{A})$ to another $(X', \mathcal{A}')$ is said to be **measurable** if $f^{-1}(E) \in \mathcal{A}$ for all $E \in \mathcal{A}'$. (Which is somewhat similar to the definition of continuity.)

With this theory, we now can do integration on not only standard $\mathbb{R}^n$ but also general measure spaces.

Mathematicians have been trying to re-establish probability theory based on set and measure theory since 1920s. In 1925, French mathematician **Paul Lévy**(1886-1971) noticed that probability is essentially a measure (of possible events), and he introduced measure to probability in his book *Cacul des probabilités*, which is several years before Kolmogorov's work.

At that time there was no mathematical theory of probability - only a collection of small computational problems. Now it is a fully-fledged branch of mathematics using techniques from all branches of modern analysis and making its own contribution of ideas, problems, results and useful machinery to be applied elsewhere. If there is one person who has influenced the establishment and growth of probability theory more than any other, that person must be Paul Lévy.

5.2 Kolmogolov's axiomatic probability theory

Soviet mathematician **Andrey Nikolaevich Kolmogorov**(1903-1987) is considered the most influential mathematician in history of modern probability theory. He also contributed to other fields of mathematics like topology.

Figure 23: Andrey Kolmogorov

Kolmogorov was born in Tambov in 1903. His mother died in childbirth at Kolmogorov's birth, and his father was an agriculturist who took part in Russian Revolution and died in 1919. Kolmogorov was brought up by his aunt Vera Yakovlena.

Kolmogorov entered Moscow university in 1920. At this stage, he studied multiple subjects including not only mathematics but also history. Once his teacher said in regarding his scientific thesis on Russian history that *You have supplied one proof of your thesis, and in the mathematics that you study this would perhaps suffice, but we historians prefer to have at least ten proofs.* That is probably partly the reason why he decided to commit to mathematics.

In 1925, Kolmogorov published his first paper on probability, which contains his "three series theorem" on which he worked together with Khinchin:

Given random variables $\{X_n\}_{n=1}^{\infty}$, *denote*

$$Y_n = \begin{cases} X_n, & |X_n| \leqslant A \\ 0, & else \end{cases}$$

If we have

- $\sum_{n=1}^{\infty} \Pr[X_n > A] < +\infty$
- $\sum_{n=1}^{\infty} \mathrm{Var}[Y_n] < +\infty$
- $\sum_{n=1}^{\infty} \mathrm{E}[Y_n] < +\infty$
- X_n converges to X in distribution.

Then we will have X_n converges to X almost surely (see 2.5).

In 1929 Kolmogorov finished his doctorate and was appointed a professor at Moscow university in 1931. Two years later in 1933, he published his monograph on probability theory *Grundbegriffe der Wahrscheinlichkeitsrechnung* (Basic concepts of probability theory), which remarks the establishment of a rigorous probability theory based on fundamental axioms just as how Euclid treated geometry 2000 years ago.

Given any set Ω (also called the *sample space*) and a σ-algebra $\mathcal{A}$ (the *event space*) on it (whose elements are called *measurable sets*), a probability on measure space $(\Omega, \mathcal{A})$ is essentially a measure P with $P(\Omega) = 1$. The three objects, set Ω, σ-algebra $\mathcal{A}$ and measure P consist what is called the **probability space** $(\Omega, \mathcal{A}, P)$.

The axioms of probability can be summarized as the following:

- $P(E) \in \mathbb{R}, P(E) \geqslant 0, \forall E \in F$
- $P(\Omega) = 1$
- $P\Big(\bigcup\limits_{k=1}^{\infty} E_k\Big) = \sum\limits_{k=1}^{\infty} P(E_k)$

Based on these facts, we can infer several properties of probability:

- $A \subset B$ then $P(A) \leqslant P(B)$
- $0 \leqslant P(E) \leqslant 1$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

In axiomatic theory, a **random variable** X is a measurable map:

$$X : (\Omega, \mathcal{A}) \to \mathbb{R}$$

Here $\mathbb{R}$ is equipped with standard (Lebesgue) measure. The expectation of random variable can be defined as

$$\mathrm{E}(X) = \int X \mathrm{d}P$$

which is the integral with respect to the measure P. In the easiest case when X is a random variable on $(\mathbb{R}, \mathcal{A}, \mu)$ with standard Lebesgue measure, then the expectation is naturally defined as $\mathrm{E}(X) = \int_{\mathbb{R}} X$.

One may see that not every subset of the sample space Ω is measurable and thus has a probability. Also, it is possible that a non-empty event's probability non-zero. (A singleton's Lebesgue measure is 0, for example.)

Kolmogorov's axiomatic probability is a great contribution to Hilbert's sixth problem:

To treat in the same manner, by means of axioms, those physical sciences in which already today mathematics plays an important part; in the first rank are the theory of probabilities and mechanics.

5.3 Contents of modern probability theory

Kolmogorov's work is a symbolic turning point in theory of probability, establishing the fundamental theory, and can be considered as the **roots** of modern probability theory. The major content of probability theory in 20th century consists of multiple methods and tools of stochastic processes, including stochastic integral, markov process, martingale theory and so on. They are the **branches** of probability theory. The specific models, like random graphs, random matrices and percolation theory are **leafs**. The leafs grow, blossom and yield fruit, which then grows into new trees.

We'll briefly introduce some branches and leafs of modern probability theory.

The theory of stochastic processes studies random variables indexed by time.

Markov process is named after Russian mathematician **Andrei Andreyevich Markov**(1856-1922). A simple model of Markov process, the Markov chain, is proposed by him in 1907. The basic property

of these processes is that given present state, the future is independent of the past.

Figure 24: Andrey Markov

A sequence of random variables X_n is called a (discrete) Markov chain if

$$\Pr\Big[X_{n+1}\Big|X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n\Big] = \Pr\Big[X_{n+1}\Big|X_n = x_n\Big]$$

A continuous-time chain $\{X_t\}_{t\geqslant 0}$ is Markov if for all i, j, t and $h \to 0$,

$$\Pr\Big[X_{t+h} = j\Big|X_t = i\Big] = \delta_{ij} + q_{ij}h + o(h)$$

for some q_{ij} and δ_{ij} the Kronecher delta function. The constant q_{ij} measures the speed of transition from i to j happens.

Markov chains can be applied in a variety of different subjects such as statistics, physics and chemistry. It often appears in thermodynamics, statistical mechanics and statistical inference and so on.

Martingale is also an important concept in modern probability theory. It describes a stochastic process in which the expectation of the next value is equal to the present value. A (discrete-time) martingale is a sequence of random variables $\{X_n\}$ with

- $|\mathrm{E}(X_n)| < +\infty, \forall n$

- $\mathrm{E}(X_{n+1}|X_1, X_2, \ldots, X_n) = \mathrm{E}(X_n), \forall n$

This concept appears very early in history, and is generalized from concepts in gambling. In a fair coin game or fair random walk, X_{n+1} has probability $\frac{1}{2}$ to take $X_n + 1$ and has equal probability to take $X_n - 1$.

The concept was introduced to probability theory by Lévy in 1934, although he didn't name it. The term "martingale" is named by Jean Ville in 1939. Systematic works and more developments are done by **Joseph L. Doob**(1910-2004). Now it is widely used in economics and finance, and can be applied to problems like random walks.

Stochastic calculus is a branch of probability theory that studies the integral and differential equations of stochastic processes, for example, the Brownian motion.

A Brownian motion is a continous-time stochastic process $\{W_t\}_{t\geqslant 0}$ that is a Markov process and a martingale at the same time. Moreover it satisfies

- The sample path W_t is almost surely continous.
- The increment $W_t - W_s \sim \mathcal{N}(0, \sigma^2(t-s))$, where standard deviation σ describes how "fast" the motion is.

The theory of stochastic calculus is mainly established by Japanese mathmatician **Ito Kiyoshi**(1915-2008). In 1942 Ito published his paper *On stochastic processes* establishing the foundations of modern theory of stochastic process. He also pionneered in constructing connections between stochastic processes and differential geometry.

Nowadays, specific probability models like random graphs and interdisciplines with other fields like stochastic geometry became more important topics. Mathematicians like **Harry Kesten**(1931-2019) and **Oded Schramm**(1961-2018) made great contributions in this time. They solved problems in random walks, percolations and related topics.

References

[1] Hardy Grant, Israel Kleiner. Turning Points in the History of Mathematics. Birkhäuser New York, NY. Springer. (2015)

[2] D Kendall, G K Batchelor, N H Bingham, W K Hayman, J M E Hyland, G G Lorentz, H K Moffatt, W Parry, A A Razborov, C A Robinson and P Whittle, Andrei Nikolaevich Kolmogorov (1903-1987), Bull. London Math. Soc. 22 (1) (1990), 31-100.

[3] Eddie Shoesmith. Huygens' Solution to the Gambler's Ruin Problem. Historia Mathematica 13 (1986), 157-164.

[4] Jakob Bernoulli (1713). Wahrscheinlichkeitsrechnung (Ars Conjectandi). Ostwald's Klassiker Der Exakten Wisserschaften Nr.108.

[5] M. Le Marquis de Laplace; Courcier, Ve. Théorie Analytique des Probabilités, troisième Édition. (1820)

[6] Pafnuti Chebyshev; Liouville. Des Valeurs Moyennes. Journal de mathématiques pures et appliqués. 2nd Series. XII. (1867),pp. 177-184.

[7] Eugene Seneta. A Tricentenary history of the Law of Large Numbers. arXiv:1309.6488v1. DOI: 10.3150/12-BEJSP12

[8] Hans Fischer. A History of the Central Limit Theorem, From Classical to Modern Probability Theory. Springer New York, NY. (2010)

[9] I. Grattan-Guinness (Editor), Ivo Schneider et al. Landmark Writings in Western Mathematics 1640-1940. Elsevier Science (2005). https://doi.org/10.1016.B978-0-444-50871-3.X5080-3