CHAPTER 7

# NUMBERS
## at work and play

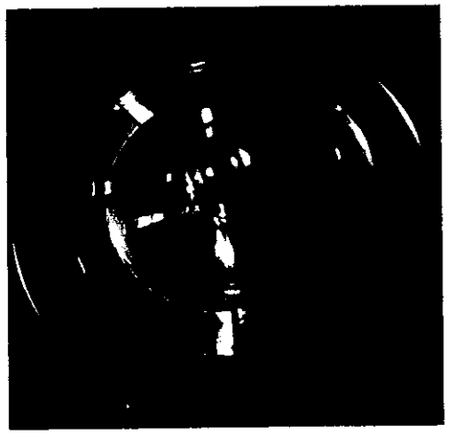
**Calculus and analysis are a long way from the everyday encounters with numbers that many of us have. Even though most of the science we come into contact with, most of the products we use and much of the world around us depends on activities in higher mathematics, our everyday encounters are more likely to lie with statistics and probability. In finance, gambling, games, the economy and many other spheres, numbers as predictors and risk assessors help us to make decisions – whether about buying a lottery ticket, taking out life insurance or flying on a plane.**

*Numbers, and the possibilities they offer, are with us all the time.*

## Cheer up, it may never happen

Humankind has played games of chance for millennia. This is playing with numbers; the fall of the dice or roll of the roulette wheel are effectively random, and winning at these games demands either large slices of luck or great proficiency in calculating probabilities and risks.

Very simple probabilities are easy to see – if we toss a coin, there is a 1 in 2 chance that it will land heads and the same chance that it will land tails. If we toss a coin a large number of times, we will probably get about as many heads as tails. This was first noted by the Swiss mathematician Jakob Bernoulli in a treatise published posthumously in 1713. He did acknowledge that the result is so patently obvious that even a very stupid person would notice it, but he is still given credit for it as he spent 20 years developing a rigorous demonstration of why it is true. He called it his Golden Theorem, but it is generally known now as the Law of Large Numbers. Casinos depend on it; although an individual gambler may have a run of good luck, over time a casino can expect to keep 5.3 per cent of all the money bet on a roulette wheel.



*Although it is possible to 'beat the bank' over a short period of time, the casino is fairly certain to win in the long run.*

Between the obvious probabilities and the Law of Large Numbers, problems of probability become more complex. What are the chances of getting tails exactly five times in a row? If we throw three dice, what is the chance of getting three sixes?

We need to do a little work with probability to be able to calculate these; the chance of getting tails five times in a row is 1 in $2^5$ = 1 in 32; the chance of throwing three sixes is 1 in $6^3$ = 1 in 216.

For most of the many thousands of years that people have been playing games of chance, they had no way of working out the probabilities of different outcomes beyond the few that are very obvious or for which it is easy to enumerate the possibilities.

### DICE AND CHAOS

Although the fall of dice or spin of a roulette wheel are effectively random, they are actually determined events. The starting position and all prevailing conditions, including the direction and force of the throw, the surface of the table and the exact features of the dice will determine the outcome. However, there are too many conditions, and their measurement is too difficult, for the outcome to be modelled or calculated.



### A GAME OF CHANCE

Probability – the chance or likelihood of an event happening – entered mathematics in the 17th century and it was in the context of a game of chance. Although Gerolamo Cardano had written on games of chance in the 1520s (see page 132–3), his work was not published until 1633 so he lost out to Fermat and Pascal. In a series of letters, the pair discussed a problem proposed by a gambler, the Chevalier de Méré:

*Two players are playing a game of pure chance on which each has bet 32 coins. The first to win three times in a row claims the pot. However, their game is interrupted after only three games. Player A has won twice and player B has won once. How can they divide the pot fairly?*

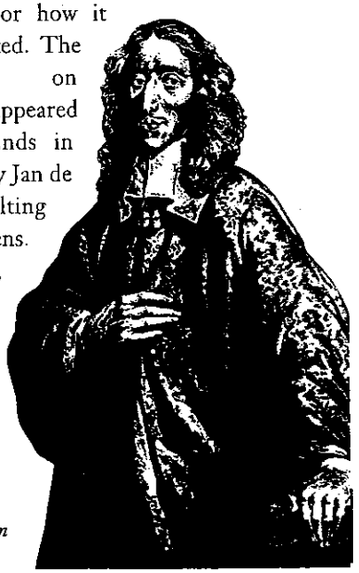
The two mathematicians both came up with a 3:1 distribution in favour of player A, though they arrived at the solution by different methods.

Fermat gave his answer in terms of probabilities. Two more games is the most that would be needed to decide the match, and there are four possible outcomes AA, AB, BA, BB. Only the last would make B the overall winner, so he has a one in four chance and should receive a quarter of the winnings. Pascal proposed a solution based on expectation. Assuming B wins the next round, each player would have an equal claim to 32 coins. Player A should receive 32 coins anyway as he definitely has two wins. The chance of B winning this next game is 50 per cent so he should have half of the remaining 32 coins. Player A also has a 50 per cent chance of winning and should have the last 16 coins. Again, player A receives 48 and player B receives 16 coins. Pascal's strategy was the one which won approval among mathematicians dealing with chance events.

### ALL'S FAIR...

Although games of chance continued to interest mathematicians, another impetus was the legal idea of a fair contract. In a fair contract, the parties have equal expectations. This was an important concept because fair expectations were at the heart of the justification for money-lending. Christian doctrine bans usury – profiting from lending money. To get around the difficulty, lenders were considered to be investors who put in money at their own risk and could fairly expect to share in the profits.

Until the 17th century, the rates for loans and annuities were fixed with no regard for any mathematical concept of risk or how it might be calculated. The first treatise on calculating risk appeared in the Netherlands in 1671, produced by Jan de Wit after consulting Christiaan Huygens. At the time, annuities were sold by the state to raise money, often to finance wars. The return



*Jan de Wit realized that risk should govern rates of return.*

was always a seventh of the value of the annuity, paid each year until the holder's death. The age or health of the holder was not taken into account. Clearly, without any assessment of how long the state may have to pay the annuity holder, this could be expensive. Even though de Wit could see the flaws in the system, there were at the time no data on mortality at different ages, so little could be done to improve the system – and little was done. It was not until 1762 that an insurance company in London, Equitable, began to price its policies on the basis of calculated risk, or probability.

## GOD EXISTS – PROBABLY

Probability did not become an exact mathematical concept until the 18th century, and was still generally considered an indistinct idea based on common sense into the 19th century. The French mathematician Pierre-Simon de Laplace (see page 174) referred to probability as 'good sense reduced to calculation'.

Interestingly, a link between chance and religion became a central interest of natural theology in the 18th century. John Arbuthnot (1667–1735) produced evidence that God definitely exists from a study of christening statistics in London between 1629 and 1710. He showed that there were slightly more boys born than girls – 14 boys christened for every 13 girls – yet by the age of marriage the balance of the sexes was equal. If we assume that the chance of a child being born a boy is 0.5, the chance of more boys than girls being born every year for 82 years is $0.5^{82}$. The same pattern of more male births is found throughout the world. Arbuthnot took this as incontrovertible evidence of Divine Providence at work, setting up society with the perfect balance. (It doesn't seem to have occurred to him that Divine Providence could equally well have killed fewer boys on the path to adulthood, thus avoiding the suffering of bereaved parents at the same time as achieving the required balance.) The argument was generally adopted and refined. However, Nicolas Bernoulli, the more rational Swiss mathematician, suggested that perhaps the probability of a male birth was not 0.5 at all but 0.5169, which would produce exactly the required result with no need for divine intervention.

### PASCAL'S WAGER

In 1657-58 Blaise Pascal wrote a philosophical essay in which he described the 'wager' a sceptic should make. The penalty for not believing in God (the Christian God, for Pascal) could be eternal damnation; however, the cost of believing in God if He turns out not to exist is slight. At most, the person who chooses to believe may relinquish a few fleeting pleasures and spend a few fruitless hours in church. Although the sceptic may feel that the chance of God's existing is very small, the cost of losing the wager is so high and the price of belief so comparatively low, that it is a better bet to believe than not believe.

## MAKING DECISIONS

As with Pascal's wager, many decisions that may be influenced by a knowledge of probability are also affected by a more subjective perception of desirable outcomes and the concept known as 'marginal utility'. Imagine a national lottery, in which tickets cost one ducat (a coin in use in much of Europe in the 18th century) and the prize is a million ducats. For a poor man, a ducat is very valuable, and the payout immensely so. For a rich man, a ducat is of little consequence, though the payout is still valuable to him. The rich man can better afford to bet a ducat than the poor man, but as he has less need of the prize he might not bother. Although the probability of winning is equal for both, the decision about whether to buy a ticket is very different for each.

In the 1750s and 1760s, inoculation against smallpox was a topical subject of debate. The inoculation used live smallpox virus and in a small number cases produced smallpox (Jenner's vaccine produced from cow pox was a later and safer introduction). Smallpox was very common, often deadly and, even when not fatal, frequently led to lifelong damage such as blindness or brain damage. Someone who did not have the vaccine stood a high chance of contracting smallpox at some time in the future, and a 1 in 7 chance of dying from it. Someone who chose to have the vaccine stood a small chance (not measured) of dying immediately of smallpox brought on by the inoculation, but otherwise virtually no chance of dying of smallpox in the future. The purely mathematical calculation, carried out by Daniel Bernoulli, suggested that there was only one sensible choice – inoculation. But the French mathematician Jean Le Rond d'Alembert, among others, argued that many people may prefer the better chance of surviving the next week or two to the assurance of safety in the future. (Today, plenty of people prefer the immediate advantage of long-haul flights to the long-term benefit of still having a planet to live on.)

## INDEPENDENCE

People are not only affected by marginal utility and the preference d'Alembert noted for short-term benefit. They may also be swayed by superstition that has no grounds in statistical probability at all.

Imagine flipping a coin ten times; the probability of getting heads each time is 1 in $2^{10}$. Suppose the first time it is heads. Now the probability of all ten flips being heads is 1 in $2^9$. If the first nine come up heads, the probability of ten heads, by the last time, is 1 in 2. Now suppose you want to fly on a plane. You know that the chances of dying in a plane crash are, say, 1 in a million on any particular flight (this is not the real probability). You have already made 1,000 flights

*Human beings seem happy to gamble with their long-term future if it means gains in the short term.*

### HERD IMMUNITY

Some diseases have been completely or nearly eradicated by national inoculation programmes. An example is measles, once endemic in the western world but now rare in countries with inoculation programmes. However, worries about the safety of the vaccine in the 1990s led to a reduced take-up of childhood vaccination in the UK and measles began to take hold again. While the vast majority of a population has immunity, a few unprotected individuals benefit from the 'herd immunity' as the disease can't get a foothold amongst the inoculated population. However, as the number of unprotected individuals rises, the presence of the disease increases to the point where it can spread amongst the uninoculated population.

The dilemma facing parents who were unconvinced about the safety of the vaccine mirrored that of the people making a choice about the early smallpox vaccine. For society as a whole, there was a moral dimension – was it right that a few individuals should avoid the (possible) risk posed by the vaccine and depend on benefiting from the herd immunity acquired at the cost of everyone else taking that risk? For mathematicians and medics, there was a different question: what proportion of the population could remain unvaccinated before their safety was compromised?

safely. Your chances of dying this time are still one in a million – the previous flights do not affect this one. In this case the events are independent; even if you had made 999,999 flights safely – or ten million – the chances of dying in the next flight would still be only 1 in a million. But it doesn't feel like that to many people. The perception is often that if we have been 'lucky' up to now, our luck is due to run out. It can work the other way, too. People may pick the same lottery number each week because they believe their number 'must come up sooner or later'. Few people pick numbers 1, 2, 3, 4, 5 and 6 because they believe (irrationally) that this combination is less likely to be drawn than any other. This tendency is not so far removed from the Ancients who believed the number 3 had special properties, or who wore a magic square for protection.

### INTERDEPENDENCE

When choosing whether to board a plane, people are dealing with random events – they have no control over whether the plane will crash. A situation that is harder for



*John von Neumann was a member of the Institute for Advanced Study at Princeton, a group of academics affectionately known as the 'demi-gods'.*

mathematicians to model is that in which one person's actions are dependent on or linked with those of another person (such as the decision about whether to vaccinate a child). This is addressed by game theory, developed in the 1940s by the Hungarian-American mathematician John von Neumann and the German-American Oskar Morgenstern.

Despite its name, game theory is concerned with the serious pursuits of economics rather than the frivolity of games. Morgenstern and von Neumann saw that the mathematical models developed for systems in physics and other areas of science were poor tools for working with economics and other studies that involve human behaviour because they were based on the actions of disinterested parties. When people make choices, they try to maximize the benefit for themselves. They may also try to minimize the detriment to others – or they may pay no attention to the impact on others, or even act to spite them.

Game theory tries to take account of the motives and insights of people acting in the situation that is modelled, as well as many other relevant aspects. For example, players – which may be individuals, groups, nations or corporations, for example – may be in direct competition or may cooperate to a greater or lesser degree. They may be competing for a finite resource or infinite resources. They may be in full possession of all relevant information, including the actions of other players, or have only partial access to information. There are different game theory models to cover these and other possibilities. Game theory often produces a matrix of outcomes which can then be analyzed.

### BACKWARD REASONING

Proofs such as that of Arbuthnot that God exists work backwards from effects to causes – there are equal numbers of marriageable men and women, therefore God exists. Jakob Bernoulli demonstrated that, if the probability of an event is not known, it can be inferred from looking at the results of experiment or observation as long as the observer has sufficient knowledge and experience. He gave as an example the fact that if a coin is tossed enough times, the ratio of heads to tails approaches ever more closely the ideal 1:1. A formal demonstration of probability in this way was made independently by Thomas Bayes and by Laplace and is now known as Bayes' theorem. Laplace famously used it to argue

the probability of the sun rising tomorrow, given our knowledge that it has risen every day for the last 6,000 years (which in 1744 was considered to be the age of the Earth).

Laplace and his contemporaries tried to put probability at the heart of the moral sciences, though their attempt was somewhat dubious. Enlightenment philosophers and reformers were concerned with the value of the judgements made by electorates and juries – would they reach the right decision or elect the best candidate? They addressed this as a problem in probability. Assuming that each juror acted independently (French juries did not deliberate) and had a greater than 0.5 chance of reaching the right verdict, they worked out the optimum size of jury and the majority needed to reach a safe conviction. The practice of deciding jury size and majority using probability continued until the 1830s. By then the system was coming into disrepute and a pupil of Laplace, Siméon-Denis Poisson, used new statistics to produce a better model.

Before probability could be used effectively in any area, though, reliable information was necessary. Statistics and probability go hand in hand.

### PIERRE-SIMON, MARQUIS DE LAPLACE (1749–1827)

The French scientist and mathematician Pierre-Simon de Laplace was most famous for his work on astronomy and his application of probability to scientific problems. He was the son of a peasant farmer, who revealed mathematical ability while at a military academy in Beaumont. In 1766 he went for one year to the University of Caen, but left for Paris, where Jean d'Alembert helped him to secure a professorship at the École Militaire. He taught there until 1776.

Laplace applied Newton's theory of gravitation to the movement of the planets. He perfected the contemporary model of the solar system and demonstrated that apparent changes are not cumulative, but occur and correct themselves in predictable cycles. (Isaac Newton had suggested that divine intervention was sometimes needed to put the solar system right!) Laplace was the first to suggest that the solar system was formed by the cooling of a vast cloud of gases.

His explanation of planetary motions made him a celebrity. Laplace was president of the Board of Longitude, helped to organize the development and introduction of the metric system, and for six weeks was minister of the interior under Napoleon.



## Samples and statistics

Without information on which to base decisions, it is possible to calculate only the most basic probabilities. Astonishingly, it was not until the late 17th century that people began to recognize the true value of collecting numeric information about populations and economies. Suddenly, statistics were everywhere and computing with them gave new insights into how societies might work. For the first time, the guesswork was taken out of planning and the burgeoning science of statistical analysis had material to work with and aims to work towards.

### PEOPLE COUNTING

Collecting information about the number of people living in an area by taking a census has been practised intermittently for thousands of years. The Babylonians, Ancient Chinese, Egyptians, Greeks and Romans all held population censuses. In Christian tradition, the parents of Jesus travelled to Bethlehem immediately before His birth because the five-yearly census required everyone in the Roman Empire to return to their place of birth to be counted.

The very basic information collected in these early censuses allowed rulers to work out how much money could be collected in taxes, how many people could be recruited for an army or building project and how much food could be produced or would be needed. In Egypt, it was also used to redistribute land after the annual flooding of the Nile. But no additional analysis of population data was carried out and only the most basic details were collected. Often, the census data were not reliable. If people expected to be taxed on the basis of how many lived in a house, a few might be missed out, for example.

In 1066, after the conquest of Britain by Norman invaders, William the Conqueror held a thorough audit of his new lands. This included a census and a listing of every item of property in the land. It was written up in the *Domesday Book* – a massive undertaking for the 11th century and one which still provides valuable statistics for historians. Thereafter, there was no enthusiasm for regular census-taking. Although bishops in many parts of Europe were supposed to keep count of the families in their dioceses, there was little information about population levels. Some people even believed that taking a census was sacrilegious, citing a story from the Bible in which King David attempted a census which was interrupted by a terrible plague and never completed.

The first regular census in modern times was carried out in Quebec, Canada in 1666. In Europe, Iceland was the first in 1703, followed by Sweden in 1749. The US held its first ten-yearly census in 1790 and the UK in 1801; the US had just under 4 million inhabitants and the UK 10 million (previous estimates had put the UK population at between 8 and 11 million).

### THE RISE OF STATISTICS

In 1662, the English statistician John Graunt published a set of statistics drawn from mortality records in London, and in the 1680s the political economist William Petty published a series of essays on 'political arithmetic' which provided statistical records with calculations – some

### THE CENSUS AND COMPUTERS

The demands of census-taking were a considerable spur to the development of technological aids to calculating. The first machine for working with census data was used in 1870. Census data were transcribed on to a rolling paper tape displayed through a small window. In 1884 Herman Hollerith (1860–1929) acquired the first patent for storing data on punched cards and organized the health records for Baltimore, Maryland, New York City and New Jersey, which won him the contract to tabulate the 1890 census. The huge success of this census opened other markets to Hollerith and his machines were used in Europe and Russia. He incorporated his Tabulating Machine Company in 1896, which later became IBM.

*Hollerith produced a mechanical tabulator based on the idea that all personal data could be coded numerically.*

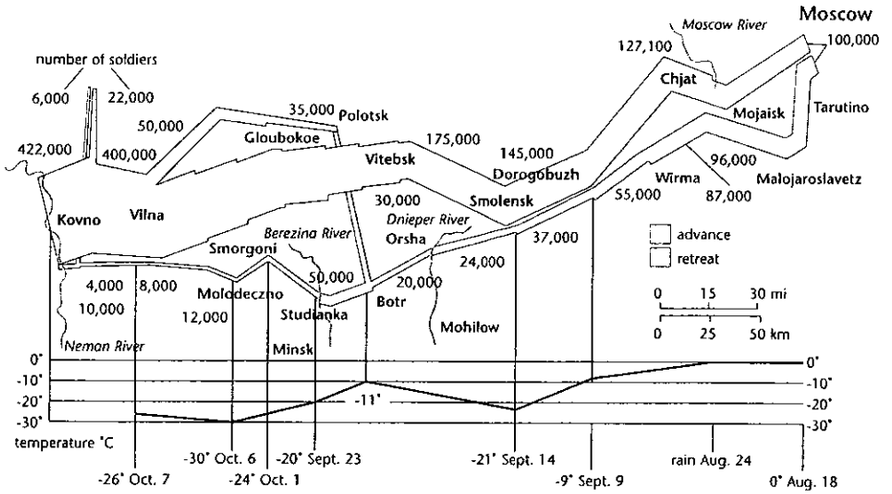
quite bizarre, such as the monetary value of all people in Ireland. On the whole, governments encouraged or financed statistical surveys and guarded the results jealously, using them to increase the power of the state. They were still inextricably tied up with superstition and followed very unscientific methods. One of the most famous 'political arithmeticians' was the Prussian Johann Sissmilch, who published three volumes over more than twenty years, ending in 1765, proving again the existence of God revealed in the harmony of social statistics.

Other statistics were collected by scientists, professionals of different types and humanitarians. Indeed, there was a growing enthusiasm for statistics, which became something of a mania during the early 19th century. Suddenly, everything was studied, counted, audited – the weather, agriculture, population movements, the tides, the land, the Earth's magnetism... The European countries that had empires surveyed their new acquisitions and took censuses in their colonies. As Americans moved westward, claiming more land, they charted it and logged its resources.

### SOCIETY IS TO BLAME

The Belgian mathematician Adolphe Quetelet (1796–1874) was a champion of statistics as the basis of the social study which he termed 'social physics'. He examined data of all kinds, using the techniques common in some scientific disciplines of amassing a vast collection of data and looking for emergent patterns. To his surprise, he found them everywhere, not just in the areas where Divine Providence






might be expected to operate. In particular, he was impressed to find that crime figures followed a predictable pattern. He conjectured that they are a product of society rather than individuals and that, while an individual criminal may be able to resist the urge to commit a crime, the overall pattern of crime rates is altered little by individual actions. He felt that the proper study was of crime rates rather than criminals and that the proper remedy to crime lay in social action, including education and an improved judicial system. Careful use of statistics to examine the effects of changes and suggest directions for future change would, he felt sure, produce the desired results.

Quetelet's thesis prompted some debate on the apparent conflict between statistics and the doctrine of free will – if crime rates can be determined by statistical methods and are unchanging over time, how much freedom do individuals really have over their actions?

*One of the most accomplished graphical representations of statistics ever made is Charles Minard's graph of Napoleon's disastrous campaign in Russia in 1812. It shows mortality on the way to and from Moscow and correlated with temperature. The width of the green and orange lines represents the size of the army, showing how it dwindles. Only 4 per cent returned from the campaign.*

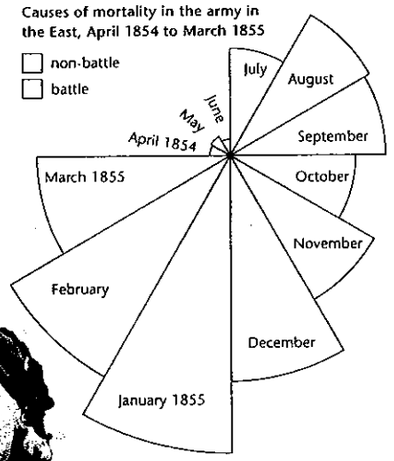
### STATISTICS MEET SCIENCE

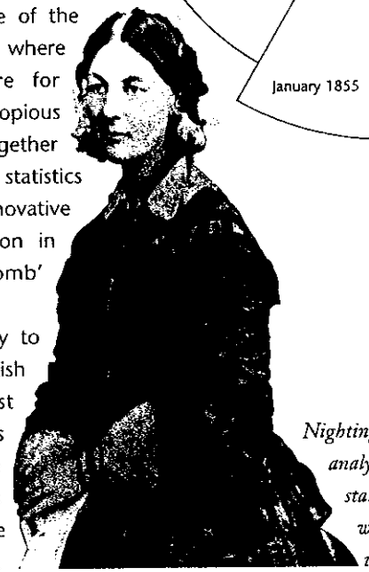
Perhaps surprisingly, it was not until the middle of the 19th century that statistics began to be applied to science with the same enthusiasm and rigour as they had been applied to social science. In the 1870s the Scottish physicist James Clerk Maxwell often explained his theory of gases with reference to social statistics. From the very large number of random movements of molecules he derived thermodynamic laws – order from chaos. He argued that, just as statistics relating to crime or suicide can yield consistent results from the unordered acts of individuals, so predictable outcomes

### FLORENCE NIGHTINGALE (1820–1910)

Florence Nightingale enjoyed a privileged childhood in England, where her father taught her languages, philosophy, history and mathematics. She claimed to have had a message from God telling her she had a vocation and later wanted to train as a nurse. Her family resisted and she became instead an expert on public health. She did later train as a nurse and, during the Crimean War (1854–56), was put in charge of the hospital at Scutari, in Turkey, where she revolutionized healthcare for wounded soldiers. She kept copious notes and after the war put together an extensive report from the statistics she had gathered. She used innovative ways of presenting information in graphs, such as the 'coxcomb' graph (above right).

Nightingale worked tirelessly to improve conditions in the British army. She founded the first training school for nurses anywhere in the world, the Nightingale School for Nurses in London, and established the professional footing of nursing.



*Nightingale was a pioneer in the analysis and presentation of statistics. 'Coxcomb' graphs were designed to be understood by everybody.*

on a large scale could be extracted from acts that are unpredictable on the small scale. But before statistics could be applied, it had to develop as a mathematical discipline. Mathematical methods specifically applying to statistics began to emerge from the end of the 18th century and proliferated rapidly.

*'[Statisticians] have already overrun every branch of science with a rapidity of conquest rivalled only by Attila, Mohammed and the Colorado beetle.'*

Maurice Kendall, 1942

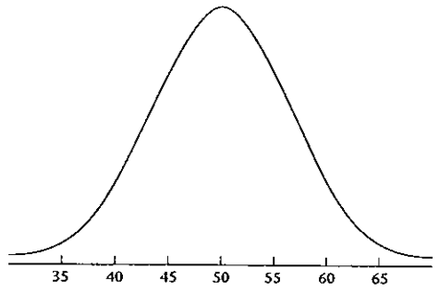


## Statistical mathematics

*Abraham de Moivre was a pioneer in analytic geometry and the theory of probability, being first to notice the normal distribution curve.*

### WHAT'S NORMAL?

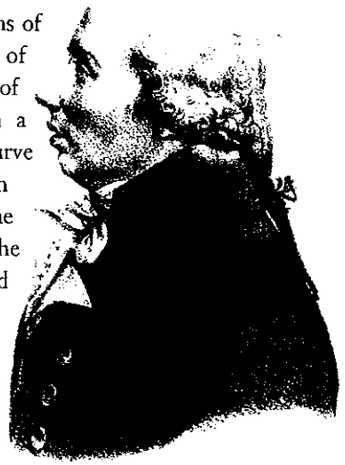
Abraham de Moivre (1667–1754) was the first person to notice the characteristic bell curve of the normal distribution (see below). The curve plots the frequency or probability of values against the values themselves. The most frequently occurring results are at the top, representing the mean value; the results that deviate most from this norm and occur least frequently are on the lower arms of the curve. The slope of the curve is determined by the degree of variation within the sample. Approximately 68 per cent of the values in the normal distribution are said to fall within one standard deviation of the norm.




The normal distribution curve and the concept of standard deviation from the norm were widely used to assess statistics in many different fields. Laplace used the model, too, in his probability studies, particularly in applying probability to very large numbers of events. Quetelet argued that virtually all human traits conformed to the normal distribution curve, from physical attributes such as height to characteristics of psychological profiling such as propensity to get married or commit suicide.

### WORKING WITH ERROR

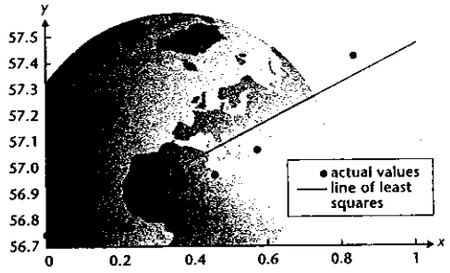
The early 19th century saw a rapid rise in mathematical methods involving statistics. Work on measuring the Earth's longitudinal circumference in order to determine the length of a metre (to be 1/40,000,000 of the circumference) needed statistical methods to deal with errors and inconsistencies in geodetic measurements. In 1805, the French mathematician Adrien-Marie Legendre (1752–1833) proposed a technique which has come to be known as the 'least squares' method. He took values that minimized the sums of the squares of deviations in a set of observations from a point, line or curve drawn through them. Gauss became interested in the method and showed in 1809 that it



*Adrien-Marie Legendre has a crater on the Moon named after him.*

### METHOD OF LEAST SQUARES

The method of least squares calculates the best line through a set of points by working out the smallest possible sum of the squares of deviations from the line of all the points. Squares are used to remove the difficulty of dealing with both positive and negative deviations, since when squared they will both give a positive result.

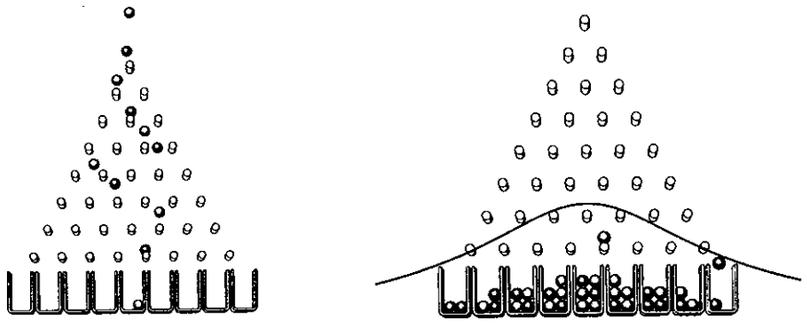



gave the best possible estimate if we assume that the errors in measurement follow the normal distribution. The method of least squares was applied to statistics in all fields and became the principal tool of statisticians in the 19th century. It was often used to estimate whole populations from a study of a small sample.

## PERFECTING HUMANITY

Francis Galton, a cousin of Charles Darwin, took an interest in the variation highlighted by normal distribution and standard deviations. He used a model, known as the Galton board, to show how a normal distribution is achieved (see below). A set of pegs is arranged in a triangle above a row of cups. Ball bearings dropped at the top of the triangle bounce down through the pegs to fall into a cup. A few fall into outlying cups but most fall into the cups in the middle of the board, forming a normal distribution curve.

Galton applied statistical ideas to heredity to show how variation tends to be bred out, and generations of an organism tend to revert to similar levels of variance.



*Ball bearings dropped on to the Galton board at the top are deflected into the cups at the bottom. The distribution of ball bearings in cups demonstrates a normal distribution curve.*



So although the children of exceptional parents may be exceptional themselves, at least in some ways, on the whole, they tend to regress towards the general population as a whole. Galton took his research in an alarming direction, becoming the founder of the eugenics movement which aimed to guide human evolution towards perfection. He wanted to breed in 'good genes' in the way that breeders select the best genes in farm animals and crops.

Although originally he was interested primarily in genetics and heredity, Galton recognized the application of his statistical methods to other areas and stressed the adaptability of the tools he developed.

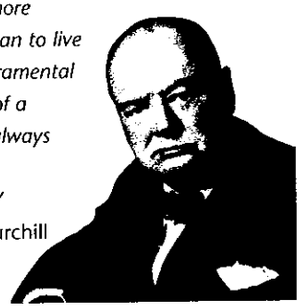
## COURTING RANDOMNESS

Developments in statistics aimed to enable information from a small sample of data to be extrapolated or applied to a larger population. By deciding the rate of crime or marriage or an inherited disease in a sample of the human population, for example, researchers hoped to reach conclusions about the rate in the whole population. The results of any statistical survey depend, of course, on the quality of the sample measured. The head of the Norwegian Central Bureau of Statistics, A. N. Kiaer, aimed to draw samples that covered the full range of representative variables in the population, such as old and young, rich and poor. The English statistician Arthur Bowley was one of the first to try to introduce randomness into sampling. The Polish statistician Jerzy Neyman brought these two concerns together in 1934, trying to ensure that a sample included representatives of major variables but that the individuals included should be chosen at random. The first triumph of this technique of stratified sampling came in 1936 when George Gallup's poll predicted the re-election of Franklin D. Roosevelt in the US, while a larger, unstratified sample, confidently (and wrongly) predicted the opposite result. Gallup drew on a sample of only 3,000 voters, while Literary Digest, the opposing pollsters, polled 10 million. Roosevelt won with the largest landslide in American history. A large sample is no guarantee of a representative sample or an accurate result.

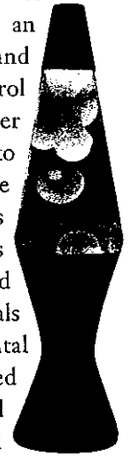
Experimental design went hand in hand with the development of statistical tools. The use of a control group to compare with

*The landslide re-election of Franklin D. Roosevelt in 1936 came as no surprise to Gallup who had used stratified sampling to predict such a result.*

*'Nothing is more dangerous than to live in the temperamental atmosphere of a Gallup Poll, always taking one's temperature.'*
Winston Churchill



the experimental group, and the random allocation of individuals to the control or experimental group, emerged as standard procedure during the early years of the 20th century. In particular, the British geneticist and statistician Sir Ronald Aylmer Fisher (1890–1962) reshaped experiment design in many fields, including psychology, medicine and ecology in the years after the Second World War. He began his research in genetics, where he used statistical analysis to reconcile inconsistencies in Darwin's theory of evolution that had been thrown up by the experimental work on inheritance of the Austrian botanist Gregor Mendel. He developed the method – which now seems ridiculously obvious – of varying only one condition in an experiment at a time and comparing results with a control group. Although earlier experimenters had done this to some degree, it was felt to be immoral where human subjects were concerned, and so rigorous use of control groups and random allocation of individuals to the control or experimental group had not been practised previously. Fisher also advocated repeating experiments and



*The lavarand random number generator developed by Bob Mende in 1996 produced random numbers using a computer program seeded with digital photographs of the patterns produced by lava lamps.*

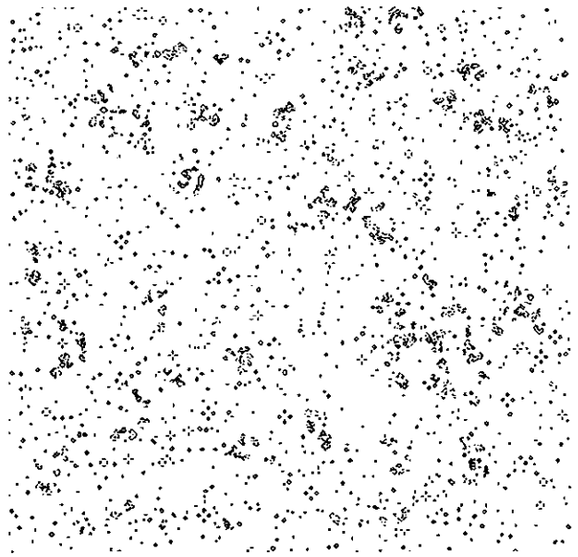
### THE DIFFICULTY OF BEING RANDOM

It is not only in sampling that randomness is actively sought. In games of chance for high stakes, there is an imperative to make sure that events that are supposed to be random are in fact just that. Cryptography also demands randomly generated numbers. This is much harder than it at first appears. As chaos theory demonstrates, many events that look random are actually not, but are governed by complex laws and a large number of variables.

The systems used to pick numbers for large-scale gambling ventures such as national lotteries are very carefully designed and engineered to remove, as far as possible, all bias in the selection methods. It is very difficult to produce computer algorithms for picking random numbers, so most lotteries use mechanical methods instead. (These also have the advantage of looking more spectacular than computers.) Computers that can generate genuinely random numbers do so using a physical source such as atmospheric noise (e.g., www.random.org).



looking at the variation in results to determine the margin of error. The most influential statistician of the 20th century, Fisher summed up his findings in the highly influential text *Statistical Methods and Scientific Inference* (1956). One of his most important developments was in the analysis of variance (called ANOVA) which looks at the points in a sample which vary from the norm. It is used to assess whether or not results are statistically significant – that is, whether they are likely to reflect a real trend, change or cause, or whether they could have come about by chance.



*British mathematician John Conway's 'Game of Life' quickly gained a cult following by simulating life, death and change in a 'society of living organisms'.*

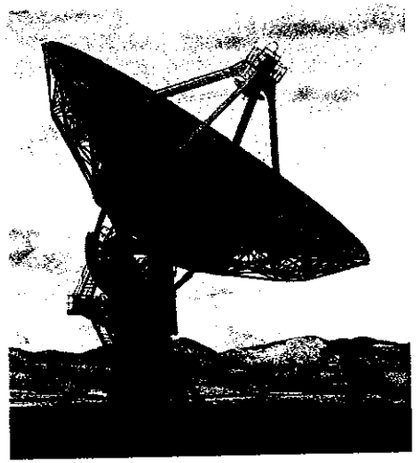
## COMPUTERIZATION

The burden of calculating with very large sets of data has been made easier by the widespread use of computers. While earlier statisticians had the laborious task of carrying out calculations for each data point by hand, their modern counterparts can feed all their data directly into a computer and leave it to apply the necessary statistical tools and provide the analysis and graphs. Often, the data are even collected by computers directly from sensors. We can now handle immense data sets, so large that they could not have been handled in a whole lifetime without computers. It means that statistical analysis can be applied in all areas of life, determining patterns and projecting outcomes in areas as diverse as the effect of early education on crime rates, the likely spread of epidemic disease and the effects of global warming.

A famous illustration of the importance of initial conditions is John Conway's 'Game of Life' (1970). This is a cellular automaton – a computer simulation of an evolving population or universe in which an initial organism or automaton makes copies of itself which succeed or fail according to various conditions (such as overcrowding, lack of resources, etc.). Conway created it in response to a problem presented by John von Neumann in the 1940s relating to constructing a machine that could make copies of itself. The 'Game of Life' is not a game in the usual sense of the word, in that there are no active players. After the

### SETI@HOME

The SETI project – Search for Extra-Terrestrial Intelligence – collects radio data from space on a continual basis, and is starting to look also for pulses of laser light. Its stated aim is 'to explore, understand and explain the origin, nature and prevalence of life in the universe'. SETI's task is to examine the constantly growing data set for patterns that might indicate a deliberate radio transmission. To do this, it asks volunteers around the world to install a screensaver which imports chunks of data from SETI over the Internet and processes them on the computer while it is not being used. In this way, SETI makes use of millions of hours of free computer time on personal computers around the world. Each PC reports its results back to SETI and any possible patterns are flagged for further investigation. An unimaginably large task in statistical analysis is being carried out at very little cost and much more quickly than it could be managed using dedicated computers.



*Looking for signs of life: radio antennae which form part of the Very Large Array astronomical observatory in New Mexico, USA.*

#### The SETI equation

The Drake equation (1961) is suggested as a way of calculating the likely number of planets that have intelligent life in the Milky Way:

$N = R^* \times fp \times ne \times fl \times fi \times fc \times L$

where

N = the number of civilizations in the Milky Way whose electromagnetic emissions are detectable.

R* = the rate of formation of stars suitable for the development of intelligent life.

fp = the fraction of those stars with planetary systems.

ne = the number of planets in each solar system with an environment suitable for life.

fl = the fraction of suitable planets on which life actually appears.

fi = the fraction of life-bearing planets on which intelligent life emerges.

fc = the fraction of civilizations that develop a technology that releases detectable signs of their existence into space.

L = the length of time these civilizations release detectable signals.



*'Nothing in the universe is unique and alone, and therefore in other regions there must be other Earths inhabited by different tribes of men and breeds of beasts.'* Lucretius, 50BC

instigator sets up initial conditions, the game runs, producing generations that flourish or perish according to the consequences of the starting conditions. The original game used populations of coloured squares in a grid, but it spawned a whole industry of computer simulation games, some of them immensely complicated, that produce populations of creatures or other entities. The interest in cellular automata that grew out of Conway's game has found applications in many fields, including research into human, animal and viral populations, growth of crystals, economic problems and many other areas in which complex patterns develop organically.

### MOVING ON

Much of the work on statistics in the last hundred years or so has led to analysis of groups or sets of data in quite complex ways. The behaviour of sets – whether of numbers or anything else – is the subject of set theory, first developed in the second half of the 19th century. The appearance of set theory has been one of the most important developments in the history of mathematics.