CHAPTER 7

THE CENTRAL GAUSSIAN, OR NORMAL, DISTRIBUTION

"My own impression ⋯ is that the mathematical results have outrun their interpretation and that some simple explanation of the force and meaning of the celebrated integral ⋯ will one day be found ⋯ which will at once render useless all the works hitherto written." - - - Augustus de Morgan (1838)

Here, de Morgan was expressing his bewilderment at the "curiously ubiquitous" success of methods of inference based on the gaussian, or normal, "error law" (sampling distribution), even in cases where the law is not at all plausible as a statement of the actual frequencies of the errors. But the explanation was not forthcoming as quickly as he expected.

In the middle 1950's the writer heard an after–dinner speech by Professor Willy Feller, in which he roundly denounced the practice of using gaussian *probability* distributions for errors, on the grounds that the *frequency* distributions of real errors are almost never gaussian. Yet in spite of Feller's disapproval, we continued to use them, and their ubiquitous success in parameter estimation continued. So 145 years after de Morgan's remark the situation was still unchanged, and the same surprise was expressed by George Barnard (1983): "*Why have we for so long managed with normality assumptions?*"

Today we believe that we can, at last, explain (1) the inevitably ubiquitous use, and (2) the ubiquitous success, of the gaussian error law. Once seen, the explanation is indeed trivially obvious; yet to the best of our knowledge it is not recognized in any of the previous literature of the field, because of the universal tendency to think of probability distributions in terms of frequencies. We cannot understand what is happening until we learn to to think of probability distributions in terms of their demonstrable *information content* instead of their imagined (and as we shall see, irrelevant) frequency connections.

A simple explanation of these properties – stripped of past irrelevancies – has been achieved only very recently, and this development changed our plans for the present work. We decided that it is so important that it should be inserted at this somewhat early point in the narrative, even though we must then appeal to some results that are established only later. In the present Chapter, then, we survey the historical basis of gaussian distributions and get a quick preliminary understanding of their functional role in inference. This understanding will then guide us directly – without the usual false starts and blind alleys – to the computational procedures which yield the great majority of the useful applications of probability theory.

The Gravitating Phenomenon

We have noted an interesting phenomenon several times in previous Chapters; in probability theory there seems to be a central, universal distribution

$$\varphi(x) \equiv \frac{1}{\sqrt{2\pi}} \exp(-x^2/2) \tag{7–1}$$

toward which all others gravitate under a very wide variety of different operations – and which, once attained, remains stable under an even wider variety of operations. The famous Central Limit Theorem, derived below, concerns one special case of this. In Chapter 4, we noted that a binomial or beta sampling distribution goes asymptotically into a gaussian when the number of trials becomes large. In Chapter 6 we noted a virtually universal property, that posterior distributions for parameters go into gaussians when the number of data values increases.

In physics these gravitating and stability properties have made this distribution the universal basis of kinetic theory and statistical mechanics; in biology, it is the natural tool for discussing population dynamics in ecology and evolution. We cannot doubt that it will become equally fundamental in economics, where it already enjoys ubiquitous use, but somewhat apologetically, as if there were some doubt about its justification. We hope to assist this development by showing that its range of validity for such applications is far wider than usually supposed.

This distribution is called the Gaussian, or Normal distribution for historical reasons discussed in our closing Comments. Both names are inappropriate and misleading today; all the correct connotations would be conveyed if we called it, simply, the *Central distribution* of probability theory.† We consider first three derivations of it that were important historically and conceptually, because they made us aware of three important properties of the gaussian distribution.

The Herschel–Maxwell Derivation

One of the most interesting derivations, from the standpoint of economy of assumptions, was given by the astronomer John Herschel (1850). He considered the two–dimensional probability distribution for errors in measuring the position of a star. Let x be the error in the longitudinal (east–west) direction and y the error in the declination (north–south) direction, and ask for the joint probability distribution $\rho(x,y)$. Herschel made two postulates (P1, P2) that seemed required intuitively by conditions of geometrical homogeneity:

(P1): Knowledge of x tells us nothing about y. That is, probabilities of errors in orthogonal directions should be independent; so the undetermined distribution should have the functional form

$$\rho(x,y)\,dx\,dy = f(x)\,dx \cdot f(y)\,dy\,. \tag{7–2}$$

We can write this equally well in polar coordinates r, θ defined by $x = r\cos\theta$, $y = r\sin\theta$:

$$\rho(x,y)\,dx\,dy = g(r,\theta)\,r\,dr\,d\theta\,. \tag{7–3}$$

(P2): This probability should be independent of the angle: $g(r,\theta) = g(r)$. Then (7–2), (7–3) yield the functional equation

$$f(x)\,f(y) = g(\sqrt{x^2+y^2}) \tag{7–4}$$

Now setting $y = 0$, this reduces to $g(x) = f(x)\,f(0)$, so (7–4) becomes the functional equation

$$\log\frac{f(x)}{f(0)} + \log\frac{f(y)}{f(0)} = \log\frac{f(\sqrt{x^2+y^2})}{f(0)} \tag{7–5}$$

But the general solution of this is obvious; a function of x plus a function of y is a function only of x^2+y^2. The only possibility is that $\log[f(x)/f(0)] = ax^2$. We have a normalizable probability only if a is negative, and then normalization determines $f(0)$; so the general solution can only have the form

$$f(x) = \sqrt{\frac{\alpha}{\pi}}\,e^{-\alpha x^2}\,, \qquad \alpha > 0 \tag{7–6}$$

with one undetermined parameter. The only two–dimensional probability density satisfying Herschel's invariance conditions is a circular symmetric gaussian:

---

† However, it is general usage outside probability theory to denote any function of the general form $\exp(-ax^2)$ as a *gaussian function*, and we shall follow this.

$$\rho(x, y) = \frac{\alpha}{\pi} \exp[-\alpha(x^2 + y^2)] \,. \tag{7–7}$$

Ten years later, James Clerk Maxwell (1860) gave a three–dimensional version of this same argument to find the probability distribution $\rho(v_x,\, v_y,\, v_z) \propto \exp[-\alpha(v_x^2 + v_y^2 + v_z^2)]$ for velocities of molecules in a gas, which has become well known to physicists as the 'Maxwellian velocity distribution law' fundamental to kinetic theory and statistical mechanics.

The Herschel–Maxwell argument is particularly beautiful because two qualitative conditions, incompatible in general, become compatible for just one quantitative distribution, which they therefore uniquely determine. Einstein (1905) used the same kind of argument to deduce the Lorentz transformation law from his two qualitative postulates of relativity theory.‡

The Herschel–Maxwell derivation is economical also in that it does not actually make any use of probability theory; only geometrical invariance properties which could be applied equally well in other contexts. Gaussian functions are unique objects in their own right, for purely mathematical reasons. But now we give a famous derivation that makes explicit use of probabilistic intuition.

The Gauss Derivation

We estimate a location parameter θ from $(n+1)$ observations $(x_0 \cdots x_n)$ by maximum likelihood. If the sampling distribution factors: $p(x_0 \cdots x_n|\theta) = f(x_0|\theta) \cdots f(x_n|\theta)$, the likelihood equation is

$$\sum_{i=0}^{n} \frac{\partial}{\partial\theta} \log f(x_i|\theta) = 0 \tag{7–8}$$

or, writing

$$\log f(x|\theta) = g(\theta - x) = g(u) \tag{7–9}$$

the maximum likelihood estimate $\hat\theta$ will satisfy

$$\sum_i g'(\hat\theta - x_i) = 0 \,. \tag{7–10}$$

Now intuition may suggest to us that the estimate ought to be also the arithmetic mean of the observations:

$$\hat\theta = \overline{x} = \frac{1}{n+1} \sum_{i=0}^{n} x_i \,, \tag{7–11}$$

but (7–10) and (7–11) are in general incompatible [(7–11) is not a root of (7–10)]. Nevertheless, consider a possible sample, in which only one observation x_0 is nonzero: if in (7–11) we put

$$x_0 = (n+1)u \,, \qquad x_1 = x_2 = \cdots = x_n = 0 \,, \qquad (-\infty < u < \infty) \,, \tag{7–12}$$

then $\hat\theta = u$, $\hat\theta - x_0 = -nu$, whereupon (7–10) becomes $g'(-nu) + n\, g'(u) = 0$, $n = 1, 2, 3, \cdots$. The case $n = 1$ tells us that $g'(u)$ must be an antisymmetric function: $g'(-u) = -g'(u)$, so this reduces to

‡ These are: (1) The laws of physics take the same form for all moving observers; and (2) The velocity of light has the same constant numerical value for all such observers. These are also contradictory in general, but become compatible for one particular quantitative law of transformation of space and time to a moving coordinate system.

$$g'(nu) = n\,g'(u)\,, \qquad (-\infty < u < \infty), \quad n = 1, 2, 3, \cdots . \tag{7–13}$$

Evidently, the only possibility is a linear function:

$$g'(u) = au\,, \qquad\qquad g(u) = \frac{1}{2}\,au^2 + b\,. \tag{7–14}$$

Converting back by (7–9), a normalizable distribution again requires that a be negative, and normalization then determines the constant b. The sampling distribution must have the form

$$f(x|\theta) = \sqrt{\frac{\alpha}{2\pi}}\; e^{-\frac{1}{2}\alpha(x-\theta)^2}\,, \qquad (0 < \alpha < \infty) \tag{7–15}$$

Since (7–15) was derived assuming the special sample (7–12), we have shown thus far only that (7–15) is a necessary condition for the equality of maximum likelihood estimate and sample mean. Conversely, if (7–15) is satisfied, then the likelihood equation (7–8) always has the unique solution (7–11); and so (7–15) is the necessary and sufficient condition for this agreement. The only freedom is the unspecified scale parameter α.

Historical Importance of Gauss' Result

This derivation was given by Gauss (1809), as little more than a passing remark in a work concerned with astronomy. It might have gone unnoticed but for the fact that Laplace saw its merit and the following year published a large work calling attention to it and demonstrating the many useful properties of (7–15) as a sampling distribution. Ever since, it has been called the 'gaussian distribution'.

Why was the Gauss derivation so sensational in effect? Because it put an end to a long – and, it seems to us today, scandalous – psychological hangup suffered by some of the greatest mathematicians of the time. The distribution (7–15) had been found in a more or less accidental way already by de Moivre (1733), who did not appreciate its significance and made no use of it. Throughout the 18'th Century it would have been of great value to astronomers faced constantly with the problem of making the best estimates from discrepant observations; yet the greatest minds failed to see it. Worse, even the qualitative fact underlying data analysis – cancellation of errors by averaging of data – was not perceived by so great a mathematician as Leonhard Euler.

Euler (1749) trying to resolve the 'Great Inequality of Jupiter and Saturn' found himself with what was at the time a monstrous problem (described briefly in our closing Comments). To determine how the longitudes of Jupiter and Saturn had varied over long times he had 75 observations over a 164 year period (1582–1745), and eight orbital parameters to estimate from them.

Today, a desk–top microcomputer could solve this problem by an algorithm to be given in Chapter 19, and print out the best estimates of the eight parameters and their accuracies, in about one minute [the main computational job is the inversion of an (8×8) matrix]. Euler failed to solve it, but not because of the magnitude of this computation; he failed even to comprehend the principle needed to solve it. Instead of seeing that by combining many observations their errors tend to cancel, he thought that this would only 'multiply the errors' and make things worse. In other words, Euler concentrated his attention entirely on the worst possible thing that could happen, as if it were certain to happen – which makes him perhaps the first really devout believer in Murphy's Law.†

Yet practical people, with experience in actual data taking, had long perceived that this worst possible thing does *not* happen. On the contrary, averaging our observations has the great

† "If anything *can* go wrong, it *will* go wrong."

advantage that the errors tend to cancel each other.[‡] Hipparchus, in the second Century B. C., estimated the precession of the equinoxes by averaging measurements on several stars. In the late sixteenth Century, taking the average of several observations was the routine procedure of Tycho Brahe. Long before it had any formal theoretical justification from mathematicians, intuition had told observational astronomers that this averaging of data was the right thing to do.

But some thirty years after Euler's effort another competent mathematician, Daniel Bernoulli (1777), still could not comprehend the procedure. Bernoulli supposes an archer is shooting at a vertical line drawn on a target, and asks how many shots land in various vertical bands on either side of it:

> "Now is it not self–evident that the hits must be assumed to be thicker and more numerous on any given band the nearer this is to the mark? If all the places on the vertical plane, whatever their distance from the mark, were equally liable to be hit, the most skillful shot would have no advantage over a blind man. That, however, is the tacit assertion of those who use the common rule [the arithmetic mean] in estimating the value of various discrepant observations, when they treat them all indiscriminately. In this way, therefore, the degree of probability of any given deviation could be determined to some extent *a posteriori*, since there is no doubt that, for a large number of shots, the probability is proportional to the number of shots which hit a band situated at a given distance from the mark."

We see that Daniel Bernoulli (1777), like his uncle James Bernoulli (1713), saw clearly the distinction between probability and frequency. In this respect his understanding exceeded that of John Venn 100 years later. Yet he fails completely to understand the basis for taking the arithmetic mean of the observations as an estimate of the true 'mark'. He takes it for granted (although a short calculation, which he was easily capable of doing, would have taught him otherwise) that if the observations are given equal weight in calculating the average, then one must be assigning equal probability to all errors, however great. Presumably, many others made intuitive guesses like this, unchecked by calculation, making this part of the folklore of the time. Then one can appreciate how astonishing it was when Gauss, 32 years later, proved that the condition

$$(\textit{maximum likelihood estimate}) \;=\; (\textit{arithmetic mean})$$

uniquely determines the gaussian error law, not the uniform one.

In the meantime, Laplace (1783) had investigated this law as a limiting form of the binomial distribution, derived its main properties, and suggested that it was so important that it ought to be tabulated; yet lacking the above property demonstrated by Gauss, he still failed to see that it was the natural error law (the Herschel derivation was still 77 years in the future). Laplace persisted in trying to use the form $f(x) \propto \exp(-a|x|)$ which caused no end of analytical difficulties. But he did understand the qualitative principle that combination of observations improves the accuracy of estimates, and this was enough to enable him to solve, in 1787, the problem of Jupiter and Saturn, on which the greatest minds had been struggling since before he was born.

Twenty–two years later, when Laplace saw the Gauss derivation, he understood it all in a flash – doubtless mentally kicked himself for not seeing it before – and hastened (Laplace, 1810, 1812) to give the Central Limit Theorem and the full solution to the general problem of reduction of observations, which is still how we analyze it today. Not until the time of Einstein did such a simple mathematical argument again have such a great effect on scientific practice.

The Landon Derivation

A derivation of the gaussian distribution that gives us a very lively picture of the process by which a gaussian frequency distribution is built up in Nature was given in 1941 by Vernon D. Landon, an electrical engineer studying properties of noise in communication circuits. We give a generalization of his argument, in our current terminology and notation.

[‡] If positive and negative errors are equally likely, then the probability that ten errors all have the same sign is $(0.5)^9 \simeq 0.002$.

The argument was suggested by the empirical observation that the variability of the electrical noise voltage $v(t)$ observed in a circuit at time t seems always to have the same general properties even though it occurs at many different levels (say, mean square values) corresponding to different temperatures, amplifications, impedance levels, and even different kinds of sources – natural, astrophysical, or man–made by many different devices such as vacuum tubes, neon signs, capacitors, resistors made of many different materials, *etc.* Previously, engineers had tried to characterize the noise generated by different sources in terms of some "statistic" such as the ratio of peak to RMS (Root Mean Square) value, which it was thought might identify its origin. Landon recognized that these attempts had failed, and that the samples of electrical noise produced by widely different sources " $\cdots$ *cannot be distinguished one from the other by any known test.*" ‡

Landon reasoned that if this distribution of noise voltage is so universal, then it must be better determined theoretically than empirically. To account for this universality but for magnitude, he visualized not a single probability distribution for the voltage at any given time, but a hierarchy of distributions $p(v|\sigma)$ characterized by a single scale parameter σ^2, which we shall take to be the expected square of the noise voltage. The stability seems to imply that if the noise level σ^2 is increased by adding a small increment of voltage, the probability distribution still has the same functional form, but only moved up the hierarchy to the new value of σ. He discovered that for only one functional form of $p(v|\sigma)$ will this be true.

Suppose the noise voltage v is assigned the probability distribution $p(v|\sigma)$. Then it is incremented by a small extra contribution ϵ, becoming $v' = v + \epsilon$ where ϵ is small compared to σ, and has a probability distribution $q(\epsilon)d\epsilon$, independent of $p(v|\sigma)$. Given a specific ϵ, the probability for the new noise voltage to have the value v' would be just the previous probability that v should have the value $(v - \epsilon)$. Thus by the product and sum rules of probability theory, the new probability distribution is the convolution

$$f(v') = \int p(v' - \epsilon|\sigma)\, q(\epsilon)\, d\epsilon\,. \tag{7–16}$$

Expanding this in powers of the small quantity ϵ and dropping the prime, we have

$$f(v) = p(v|\sigma) - \frac{\partial p(v|\sigma)}{\partial v} \int \epsilon\, q(\epsilon)\, d\epsilon + \frac{1}{2}\frac{\partial^2 p(v|\sigma)}{\partial v^2} \int \epsilon^2\, q(\epsilon)\, d\epsilon + \cdots \tag{7–17}$$

or, now writing for brevity $p \equiv p(v|\sigma)$,

$$f(v) = p - \langle\epsilon\rangle \frac{\partial p}{\partial v} + \frac{1}{2}\langle\epsilon^2\rangle \frac{\partial^2 p}{\partial v^2} + \cdots \tag{7–18}$$

This shows the general form of the expansion; but now we assume that the increment is as likely to be positive as negative; $\langle\epsilon\rangle = 0$.⋆ At the same time, the expectation of v^2 is increased to $\sigma^2 + \langle\epsilon^2\rangle$, so Landon's invariance property requires that $f(v)$ should be equal also to

‡ This universal, stable type of noise was called "grass" because that is what it looks like on an oscilloscope. To the ear, it sounds like a smooth hissing without any discernible pitch; today this is familiar to everyone because it is what we hear when a television receiver is tuned to an unused channel. Then the automatic gain control turns the gain up to the maximum, and both the hissing sound and the flickering 'snow' on the screen are the greatly amplified noise generated by random thermal motion of electrons in the antenna according to the Nyquist law noted below.

⋆ If the small increments all had a systematic component in the same direction, one would build up a large "D. C." noise voltage, which is manifestly not the present situation. But the resulting solution might have other applications; see Exercise 7.1.

$$f(v) = p + \langle \epsilon^2 \rangle \frac{\partial p}{\partial \sigma^2} \,. \tag{7–19}$$

Comparing (7–18) and (7–19), we have the condition for this invariance:

$$\frac{\partial p}{\partial \sigma^2} = \frac{1}{2} \frac{\partial^2 p}{\partial v^2} \,. \tag{7–20}$$

But this is a well–known differential equation (the "diffusion equation"), whose solution with the obvious initial condition $p(v|\sigma = 0) = \delta(v)$ is

$$p(v|\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{v^2}{2\sigma^2}\right] , \tag{7–21}$$

the standard Gaussian distribution. By minor changes in the wording, the above mathematical argument can be interpreted either as *calculating* a probability distribution, or as *estimating* a frequency distribution; in 1941 nobody except Harold Jeffreys and John Maynard Keynes took note of such distinctions. As we shall see, this is, in spirit, an incremental version of the Central Limit Theorem; instead of adding up all the small contributions at once, it takes them into account one at a time, requiring that at each step the new probability distribution has the same functional form (to second order in ϵ).

This is just the process by which noise is produced in Nature – by addition of many small increments, one at a time (for example, collisions of individual electrons with atoms, each collision radiating another tiny pulse of electromagnetic waves, whose sum is the observed noise). Once a gaussian form is attained, it is preserved; this process can be stopped at any point and the resulting final distribution still has the Gaussian form. What is at first surprising is that this stable form is independent of the probability distributions $q(\epsilon)$ of the small increments; that is why the noise from different sources could not be distinguished by any test known in 1941.†

Today we can go further and recognize that the reason for this independence was that only the second moment $\langle \epsilon^2 \rangle$ of the increments mattered for the updated point distribution (that is, the probability distribution for the voltage *at a given time* that we were seeking). Even the magnitude of the second moment did not matter for the functional form; it determined only how far up the σ^2–heirarchy we moved. But if we ask a more detailed question, involving time–dependent correlation functions, then noise samples from different sources are no longer indistinguishable. The second order correlations of the form $\langle \epsilon(t)\epsilon(t') \rangle$ are related to the power spectrum of the noise through the Wiener–Khinchin theorem, which was just in the process of being discovered in 1941; they give information about the duration in time of the small increments. But if we go to fourth order correlations $\langle \epsilon(t_1)\epsilon(t_2)\epsilon(t_3)\epsilon(t_4) \rangle$ we obtain still more detailed information, different for different sources even though they all have the same Gaussian point distribution and the same power spectrum.‡

---

† Landon's original derivation concerned only a special case of this, in which $q(\epsilon) = [\pi\sqrt{a^2 - \epsilon^2}]^{-1}$, $|\epsilon| < a$ corresponding to an added sinusoid of amplitude a and unknown phase. But the important thing was his *idea* of the derivation, which anyone can generalize once it is grasped. In essence he had discovered independently, in the expansion (7–18), what is now called the Fokker–Planck equation of statistical mechanics, a powerful method which we shall use later to show how a nonequilibrium probability distribution relaxes into an equilibrium one. It is now known to have a deep meaning, in terms of continually remaximized entropy.

‡ Recognition of this invalidates many naïve arguments by physicists who try to prove that "Maxwell Demons" are impossible by assuming that thermal radiation has a universal character, making it impossible to distinguish the source of the radiation. But only the second order correlations are universal; a demon

Exercise 7.1. The above derivation established the result to order $\langle \epsilon^2 \rangle$. Now suppose that we add n such small increments, bringing the variance up to $\sigma^2 + n\langle \epsilon^2 \rangle$. Show that in the limit $n \to \infty$, $\langle \epsilon^2 \rangle \to 0$, $n\langle \epsilon^2 \rangle \to$ const., the gaussian distribution (7–21) becomes exact (the higher terms in the expansion (7–17) become vanishingly small compared to the terms in $\langle \epsilon^2 \rangle$).

Exercise 7.2. Repeat the above derivation without assuming that $\langle \epsilon \rangle = 0$ in (7–18). The resulting differential equation is a Fokker–Planck equation. Show that there is now a superimposed steady drift, the solutions having the form $\exp[-(v - a\sigma^2)^2/2\sigma^2]$. Suggest a possible useful application of this result. *Hint*: σ^2 and v may be given other interpretations, such as time and distance.

Why the Ubiquitous Use of Gaussian Distributions?

We started this Chapter by noting the surprise of de Morgan and Barnard at the great and ubiquitous success that is achieved in inference – particularly, in parameter estimation – through the use of gaussian sampling distributions, and the reluctance of Feller to believe that such success was possible. It is surprising that to understand this mystery requires almost no mathematics – only a conceptual re–orientation toward the idea of probability theory as logic.

Let us think in terms of the *information* that is conveyed by our equations. Whether or not the long–run frequency distribution of errors is in fact gaussian is almost never known empirically; what the scientist knows about them (from past experience or from theory) is almost always simply their general magnitude. For example, today most accurate experiments in physics take data electronically, and a physicist usually knows the mean–square error of those measurements because it is related to the noise energy and temperature by the well–known Nyquist thermal fluctuation law.* But he seldom knows any other property of the noise. If one assigns the first two moments of a noise probability distribution to agree with such information, but has no further information and therefore imposes no further constraints, then a gaussian distribution fit to those moments will, according to the Principle of Maximum Entropy as discussed in Chapter 11, represent most honestly his state of knowledge about the noise.

Put differently, we shall find in Chapter 11 that when we use a gaussian sampling distribution for the noise, we are in effect telling the robot: "The only thing I know about the noise is its first two moments; so please take that into account in assigning your probability distribution, but be careful *not* to assume anything else about the noise." We shall see presently how well the robot, applying Bayes' theorem, obeys this instruction.†

This does not mean that the frequency distribution of the noise is to be ignored if it happens to be known. Probability theory as logic does not conflict with conventional probability theory if

---

who perceives fourth order correlations in thermal radiation is far from blind about the details of his surroundings. Indeed, the famous Hanbury Brown–Twiss interferometer (1956), invokes just such a fourth–order demon, in space instead of time and observing $\langle \epsilon^2(x_1)\epsilon^2(x_2)\rangle$ to measure the angular diameters of stars. Conventional arguments against Maxwell demons are logically flawed and prove nothing.

* A circuit element of resistance $R(\omega)$ ohms at angular frequency ω develops across its terminals in a small frequency band $\Delta\omega = 2\pi\Delta f$ a fluctuating mean–square open–circuit voltage $V^2 = 4kTR\Delta f$, where f is the frequency in Hz (cycles per second), $k \equiv 1.38 \times 10^{-23}$ joules/degree is Boltzmann's constant, and T is the Kelvin temperature. Thus it can deliver to another circuit element the maximum noise power $P = V^2/4R = kT\Delta f$. At room temperature, $T = 300$K, this is about 4×10^{-15} watts per megahertz bandwidth. Any signal of lower intensity than this will be lost in the thermal noise and cannot be recovered, ordinarily, by any amount of amplification. But prior information about the kind of signal to be expected will still enable a Bayesian computer program to extract far weaker signals than this, as the work of Bretthorst (1988) demonstrates. We study this in Chapter 23.

† If we have further pieces of information about the noise, the robot can take these into account also by assigning generalized gaussian – that is, general maximum entropy – distributions.

we actually have the information that conventional theory presupposes; but it continues to operate when we do not have such information (as is the case in the vast majority of real problems), using whatever information we do have.

Exercise 7.3. Suppose that the long–run frequency distribution of the noise is known to be the gaussian (7–21) and we have no other information about the noise. Show, by reasoning like that leading to (4–48) and using Laplace's Rule of Succession (6–70), that in the limit of a very large amount of frequency data, our *probability* distribution for the noise becomes numerically equal to the observed frequency distribution.

In other fields such as analysis of economic data, knowledge of the noise may be more crude, consisting of an approximate general magnitude of the noise and nothing else. But a gaussian distribution fit to that magnitude is still a good approximation to one's state of knowledge. If even that knowledge is lacking, a gaussian with σ an undetermined nuisance parameter to be estimated from the data is an appropriate and useful starting point. Indeed, as Bretthorst (1988) demonstrates, this is often the safest procedure even in a physics experiment, because the noise may not be the theoretically well understood Nyquist noise (a defective apparatus may generate noise far above the Nyquist value).

We shall find, in the Central Limit Theorem, still another strong justification for using gaussian error distributions. But if the gaussian law is nearly always a good representation of our state of knowledge about the errors *in our specific data set*, it follows that inferences made from it are nearly always the best ones that could have been made from the information that we actually have. Those inferences could be improved upon only by one who had additional information about the actual errors in his specific data set, *beyond what is known from the data.*

For this reason, whether our inferences are successful or not, unless such extra information is at hand there is no justification for adopting a different error law; and indeed, no principle to tell us which different one to adopt. This explains the ubiquitous use. Since the time of Gauss and Laplace, the great majority of all inference procedures with continuous probability distributions have been conducted – necessarily and properly – with gaussian sampling distributions. Those who disapproved of this, whatever the grounds for their objection, have been unable to offer any alternative that was not subject to a worse objection; so already in the time of de Morgan, some 25 years after the work of Laplace, use of the gaussian rule had become ubiquitous by default, and this continues today. Recognition of this considerably simplifies our expositions of Bayesian inference.

It is now clear that the most ubiquitous reason for using the gaussian sampling distribution is not that the error frequencies are known or believed to be gaussian; but rather because those frequencies are *unknown*. One sees what a totally different outlook this is than that of Feller and Barnard; "normality" was not an *assumption* of physical fact at all. It was a *valid description* of our state of information. In most cases, had we done anything different we would be making an unjustified, gratuitous assumption (violating one of our Chapter 1 desiderata of rationality). But this still does not explain why the procedure is so successful; so we need to explain much more.

Why the Ubiquitous Success?

By 'ubiquitous success' we mean that for nearly two Centuries, the gaussian sampling distribution has continued to be much easier to use and to yield better results (more accurate parameter estimates) than any alternative sampling distribution that anyone has been able to suggest. To explain this requires that analysis that de Morgan predicted would one day be found.

As a start, note that we are going to use some function of the data as our estimate; then whether our present inference – here and now – is or is not successful, depends entirely on what that function is, and on the actual errors that are present in the one specific data set that we are analyzing. Therefore to explain its success requires that we examine that specific data set. The frequency distribution of errors in other data sets that we might have got but did not – and which we are therefore not analyzing – is irrelevant to the question we now seek to answer, unless (a) it is actually known, not merely imagined; and (b) it tells us something about the errors in our specific data set that we would not know otherwise. Those who have not perceived this persist in trying to relate the gaussian distribution to assumed frequencies in imaginary data sets that do not exist; thus they continue to look in the wrong place, and so are unable to find any explanation of the success.

We are trying to estimate a location parameter μ, and our data D consist of n observations: $D = \{y_1 \cdots y_n\}$. But they have errors that vary in a way that is uncontrolled by the experimenter and unpredictable from his state of knowledge.‡ In the following we denote the unknown true value by μ_0, and use μ as a general running variable. Then

$$y_i = \mu_0 + e_i\,, \qquad (1 \le i \le n) \tag{7–22}$$

where e_i is the actual error in the i'th measurement. Now if we assign an independent gaussian sampling distribution:

$$p(D|\mu, \sigma, I) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left[-\frac{\sum(y_i-\mu)^2}{2\sigma^2}\right] \tag{7–23}$$

we have

$$\sum_{i=1}^{n} (y_i - \mu)^2 = n[(\mu - \overline{y})^2 + s^2] \tag{7–24}$$

where

$$\begin{aligned} \overline{y} &\equiv \frac{1}{n}\sum y_i = \mu_0 + \overline{e} \\ s^2 &\equiv \overline{y^2} - (\overline{y})^2 = \overline{e^2} - (\overline{e})^2 \end{aligned} \tag{7–25}$$

The consequence of assigning the gaussian error distribution is that *only the first two moments* of the data are going to be used for inferences about μ (and about σ, if it is unknown). They are called *sufficient statistics*, and we study the general phenomenon of sufficiency in Chapter 8. From (7–25) it follows that only the first two moments of the noise values $\{e_1 \cdots e_n\}$:

$$\overline{e} = \frac{1}{n}\sum_i e_i\,, \qquad \overline{e^2} = \frac{1}{n}\sum_i e_i^2 \tag{7–26}$$

can matterfor the error in our estimate. We have, in a sense, the simplest possible connection between the errors in our data and the error in our estimate, allowing the maximum possible opportunity for that error cancellation to take place.

> *When we assign an independent gaussian sampling distribution to additive noise, what we achieve is that all other aspects of the noise beyond $\overline{e}$ and $\overline{e^2}$ are made irrelevant to our inferences; they can contribute nothing to the numerical value or the accuracy of our estimates.*

Feller, thinking exclusively in terms of sampling distributions for estimators, thought that unless our sampling distribution correctly represented the actual frequencies of errors, our estimates would be in some way unsatisfactory. Now there is an important and closely related truth here: *The*

‡ This does not mean that they are 'not determined by anything' as is so often implied by those suffering from the Mind Projection Fallacy; it means only that they are not determined by any circumstances that the experimenter is observing. Whether controlling factors could or could not be observed in principle is irrelevant to the present problem, which is to reason as best we can *in the state of knowledge that we have specified.*

actual variability of the estimate in the long run over all possible data sets, is indeed determined by the actual long run frequency distribution of the errors. But does it follow that our assigned sampling distribution must be equal to that frequency distribution in order to get satisfactory estimates? Might not the estimates of μ be made still better in the long run (*i.e.*, more closely concentrated about the true value μ_0) by a different choice of sampling probability distribution? There is a big gap in the logic here, even from a pure sampling–theory standpoint.

But more fundamental than the logical error is the conceptual disorientation; the scenario envisaged by Feller is not the real problem facing a scientist. His job is not to fantasize about an imaginary 'long run' which will never be realized; but to estimate the parameters in the one real case before him, from the one real data set that he actually has.*

In the absence of cogent prior information about μ (that is, if the prior density $p(\mu|I)$ is substantially constant over the region of high likelihood), then Bayes' theorem will lead us to make the posterior (mean) $\pm$ (standard deviation) estimate†

$$(\mu)_{est} = \left\{ \begin{array}{ll} \overline{y} \pm \sigma/\sqrt{n}\,, & \sigma \text{ known} \\ \overline{y} \pm s/\sqrt{n-3}\,, & \sigma \text{ unknown} \end{array} \right\} \tag{7–27}$$

Thus the actual error we shall make in the estimate is the average of the individual errors in our specific data set:‡

$$\Delta = \overline{y} - \mu_0 = \overline{e} \tag{7–28}$$

Note that $\overline{e}$ is not an average over any probability distribution; it is the average of the *actual* errors; and this result holds however the actual errors e_i are distributed. For example, whether a histogram of the e_i closely resembles the assigned gaussian (7–23) or whether all of the error happens to be in e_1 does not matter in the least; (7–28) remains correct.

The Near–Irrelevance of Sampling Distributions

Another way of looking at this is helpful. As we have seen before, in a repetitive situation the probability of any event is usually the same as its expected frequency (using, of course, the same basic probability distribution for both). Then given a sampling distribution $f(x|\theta)$, it tells us that $\int_R f(x|\theta)\,dx$ is the expected frequency of the event $x \in R$, *before the data are known.*

But if, as always supposed in elementary parameter estimation, the parameters are held fixed throughout the taking of a data set, then the variability of the data *is also, necessarily*, the variability of the actual errors in that data set. If the noise is additive, as is assumed by the model

---

* Curiously, in that same after–dinner speech Feller also railed against those who fail to distinguish between the long run and the individual case, yet it appears to us that it was Feller who failed to make that distinction properly. He would judge the merit of an individual case inference by its imagined long run properties. But it is not only possible, but common as soon as we depart from gaussian sampling distributions, that an estimator which is proved to be as good as can be obtained, as judged by its long run success over all data sets, may nevertheless be very poor for our particular data set and should not be used for it. This introduces us to the phenomenon of ancillarity, pointed out by R. A. Fisher in the 1930's and discussed in Chapter 8. It is now known that Bayes' theorem automatically detects this situation and does the right thing here, choosing for each data set the optimal estimator *for that data set.*

† The detailed derivation of this result is given in our discussion of Estimation with a Gaussian Distribution in Chapter 20; however, we might regard it as an easy exercise for the reader.

‡ Of course, probability theory tells us that this is the best estimate we can make if, as supposed, the only information we have about μ comes from this one data set. If we have other information (previous data sets, other prior information) we should take it into account; but then we are considering a different problem.

(7–22), the exact distribution of the errors is known from the data to within a uniform translation: $e_i - e_j = x_i - x_j$. Then what use remains for the sampling distribution, which yields only the prior expectations of the error frequencies? Whatever error distribution we might have expected before seeing the data, is rendered irrelevant by the information in the data! What remains significant is the likelihood function – how the probability of the data varies with the parameters θ.

Although all these results are mathematically trivial, we stress their nontrivial consequences by repeating them in different words. A gaussian distribution has a far deeper connection with the arithmetic mean than that shown by Gauss: If we assign the independent gaussian error distribution, then the error in our estimate is always the arithmetic mean of the true errors in our data set; and whether the frequency distribution of those errors is or is not gaussian is totally irrelevant. Any error vector $\{e_1 \cdots e_n\}$ with the same first moment $\overline{e}$ will lead us to the same estimate of μ; and any error vector with the same first two moments will lead us to the same estimates of both μ and σ and the same accuracy claims, *whatever the frequency distributions of the individual errors*. This is the first part of the answer to de Morgan, Feller, and Barnard.

This makes it clear that what matters to us functionally – that is, what determines the actual error of our estimate – is not whether the gaussian error law correctly describes the frequency distribution of the errors; but rather whether that error law correctly describes our *prior information* about the errors. If it does, then the above calculations are the best we can do; and there is nothing more to be said.

The only case where we should – or indeed, could – do anything different is when we have additional prior information about the errors beyond their first two moments. For example, if we know that they are simple digitizing errors with digitizing interval δ, then we know that there is a rigid upper bound to the magnitude of any error: $|e_i| \le \delta/2$. Then if $\delta < \sigma$, use of the appropriate truncated sampling distribution instead of the gaussian (7–23) will almost surely lead to more accurate estimates of μ. This kind of prior information can be very helpful (although it complicates the analytical solution, this is no deterrent to a computer), and we consider a problem of this type below, under 'Accuracy of Computation'.

But, closer to the present issue, in what sense and under what conditions does the gaussian error law 'correctly describe' our information about the errors?

The Remarkable Efficiency of Information Transfer

Again, we anticipate a few results from later Chapters in order to get a quick, preliminary view of what is happening, which will improve our judgment in setting up real problems. The gaussian distribution

$$p(e|\alpha,\beta) = \frac{1}{\sqrt{2\pi\beta^2}} \exp\left[-\frac{(e-\alpha)^2}{2\beta^2}\right] \tag{7–29}$$

has maximum entropy $H = -\int p(e) \log p(e)\, de$ subject to the constraints of fixed expectations

$$\langle e \rangle = \alpha, \qquad \langle e^2 \rangle = \alpha^2 + \beta^2 \tag{7–30}$$

in which the brackets $\langle\ \rangle$ now denote averages over the probability distribution (7–29). So a state of prior information which leads us to prescribe the expected first and second moments of the noise–+and nothing else–+uniquely determines the gaussian distribution. Then it is eminently satisfactory that this leads to sufficient statistics that are just the first and second moments of the actual errors. When we assign error probabilities by the principle of maximum entropy, *the only properties of the errors that are used in our Bayesian inference are the properties about which we had some prior information.* This is a very important second part of that answer.

In this example we have stumbled for the first time onto a fundamental feature of probability theory as logic: if we assign probabilities to represent our information, then circumstances about which we have no information, are not used in our subsequent inferences. But it is not only true of this example; we shall find in Chapter 11 that it is a general theorem that the class of sampling distributions which have sufficient statistics is precisely the class generated by the maximum entropy principle; and that the resulting sufficient statistics are precisely the constraints which determined that maximum entropy distribution.

Put differently, our rules for extended logic automatically use all the information that we have, and avoid assuming information that we do not have. Indeed, our Chapter 1 desiderata require this. In spite of its extremely simple formal structure in the product and sum rules, probability theory as logic has a remarkable sophistication in applications. It perceives instantly what generations of statisticians and probabilists failed to see; for a probability calculation to have a useful and reliable function in the real world, it is by no means required that the probabilities have any relation to frequencies.*

But once this is pointed out, it seems obvious that something about which we have no information *cannot* be of any use to us in inference. *Ad hoc* rules for inference which fail to recognize this and try to introduce such quantities as error frequencies into the calculation, even when we have no information about them, are claiming, in effect, to get something for nothing (in fact, they are injecting arbitrary – and therefore almost certainly false – information). As we shall see in Chapter 17, such devices yield wrong and/or misleading conclusions if applied outside the small class of problems for which they were invented.

On the other hand, probability theory as logic is always safe and conservative, in the following sense; it always yields the conclusions that are justified by the information *which was put into it*, because it always spreads the probability out over the full range of conditions allowed by that information. The robot can return vague or incomplete answers if we give it vague or incomplete information; but then *it warns us of that fact by returning posterior distributions so wide that they still include the true value*. It cannot actually mislead us – in the sense of assigning a high probability to a false conclusion – unless we have given it false information.

For example, if we assign a sampling distribution which supposes the errors to be far smaller than the actual errors, then we have put false information into the problem, and the consequence will be, not necessarily bad estimates of parameters, but false claims about the accuracy of those estimates and – often more serious – the theory can hallucinate, artifacts of the noise being misinterpreted as real effects. As de Morgan (1872, p. 113) put it, this is the error of "attributing to the motion of the moon in her orbit all the tremors which she gets from a shaky telescope".

Conversely, if we use a sampling distribution which supposes the errors to be much larger than the actual errors, the result is not necessarily bad estimates, but overly conservative accuracy claims for them and – often more serious – blunt perception, failing to recognize effects that are real, by dismissing them as part of the noise. This would be the opposite error of attributing to a shaky telescope the real and highly important deviation of the moon from her expected orbit. If we use a sampling distribution that reflects the true average errors and the true mean square errors, we have the maximum protection against both of these extremes of misperception, steering the safest possible middle course between them.

Nuisance Parameters as Safety Devices

As an example of this principle, if we do not have actual knowledge about the magnitude σ of our errors, then it could be dangerous folly to assume some arbitrary value; the wisest and safest procedure is to adopt a model which honestly acknowledges our ignorance by allowing for various possible values of σ; we should assign a prior $p(\sigma|I)$ which indicates the range of values that σ

* This is not to say that probabilities are *forbidden* to have any relation to frequencies; the point is rather that whether they do or do not depends on the problem. We shall see, in the work of Galton below, an example where a clear frequency connection is present, and analysis of the general conditions for this will appear in Chapter 9.

might reasonably have. Then in the Bayesian analysis we shall find first the joint posterior *pdf* for both parameters:

$$p(\mu, \sigma|D, I) = p(\mu, \sigma|I) \frac{p(D|\mu, \sigma, I)}{p(D|I)} . \tag{7–31}$$

But now notice how the product rule rearranges this:

$$p(\mu, \sigma|D, I) = p(\sigma|I)\, p(\mu|\sigma, I) \frac{p(D|\sigma, I)\, p(\mu|\sigma, D, I)}{p(D|I)\, p(\mu|\sigma, I)} = p(\mu|\sigma, D, I)\, p(\sigma|D, I) \tag{7–32}$$

So, if we now integrate out σ as a nuisance parameter, we obtain the marginal posterior *pdf* for μ alone in the form:

$$p(\mu|D, I) = \int p(\mu|\sigma, D, I)\, p(\sigma|D, I)\, d\sigma\,, \tag{7–33}$$

a weighted average of the *pdf*'s $p(\mu|\sigma, D, I)$ for all possible values of σ, weighted according to the marginal posterior *pdf* $p(\sigma|D, I)$ for σ, which represents everything we know about σ.

Thus when we integrate out a nuisance parameter, we are not throwing away any information relevant to the parameters we keep; on the contrary, probability theory automatically estimates the nuisance parameter for us from all the available evidence, and takes that information fully into account in the marginal posterior *pdf* for the interesting parameters (but it does this in such a slick, efficient way that one may not realize that this is happening, and think that he is losing something). In the limit where the data are able to determine the true value $\sigma = \sigma_0$ very accurately, $p(\sigma|D, I) \to \delta(\sigma - \sigma_0)$ and $p(\mu|D, I) \to p(\mu|\sigma_0, D, I)$; the theory yields, as it should, the same conclusions that we would have if the true value were known from the start.

This is just one example showing that, whatever question we ask, probability theory as logic automatically takes into account all the possibilities *allowed by our model*. Then, of course, the onus is on us to choose a model wisely so that the robot is given the freedom to estimate for itself, from the prior information and the evidence in the data, any parameter that we do not know. If we fail to recognize the existence of a relevant but uninteresting parameter and so leave it out of the model (thus in effect assuming it to be zero), then the robot is crippled and cannot return the optimal inferences to us. The marginalization paradox, discussed in Chapter 15, and the data pooling paradox of Chapter 8, exhibit some of the things that can happen if we leave a relevant parameter out of the model; the robot's conclusions are still the best ones that could have been made *from the information we gave it*, but they are not the ones that simple common sense would make.

In practice, we find that recognition of a relevant, but unknown and uninteresting parameter by including it in the model and then integrating it out again as a nuisance parameter, can greatly improve our ability to extract the information we want from our data – often by orders of magnitude. By this means we are forewarning the robot about a possible disturbing complication, putting it on the lookout for it; and the rules of probability theory then lead the robot to make the optimal allowance for it. The recent success of Bayesian Spectrum Analysis (Bretthorst, 1988) is based on this recognition, as we shall see in detail when we take up spectrum/shape analysis.

This point is extremely important in some current problems of estimating environmental hazards or the safety of new drugs or food additives, where inattention to all of the relevant prior information that scientists have about the model and the possible errors, can cause the danger to be grossly overestimated or underestimated.

More General Properties

Although the Gauss derivation was of the greatest historical importance, it does not satisfy us today because it depends on intuition; *why* must the 'best' estimate of a location parameter be the sample mean? Why must it be even a linear function of the observations? Evidently, in view of the Gauss derivation, if the sampling distribution is not gaussian, the best estimate of the location parameter will *not* be the sample mean. It could have a wide variety of other functional forms; then under what circumstances is Laplace's prescription the one to use?

We have just seen the cogent pragmatic advantages of using a gaussian sampling distribution. Today, anticipating a little from later Chapters, we would say that its unique theoretical position derives not from the Gauss argument, but rather from four mathematical stability properties which have fundamentally nothing to do with probability theory or inference; and a fifth which has everything to do with them, but was not discovered until the mid–twentieth Century:

(A) Any smooth function with a single rounded maximum, if raised to higher and higher powers, goes into a gaussian function. We saw this in Chapter 6.
(B) The product of two gaussian functions is another gaussian function
(C) The convolution of two gaussian functions is another gaussian function.
(D) The fourier transform of a gaussian function is another gaussian function.
(E) A gaussian probability distribution has higher entropy than any other with the same variance; therefore any operation on a probability distribution which discards information but conserves variance, leads us inexorably closer to a gaussian. The Central Limit Theorem, derived below, is the best known example of this, in which the operation being performed is convolution.

Properties (A) and (E) explain why a gaussian form is approached more and more closely by various operations; properties (B) – (D) explain why that form, once attained, is preserved.

Convolution of Gaussians

The convolution property (C) is shown as follows. Expanding now the notation of (7–1)‡

$$\varphi(x-\mu|\sigma) \equiv \frac{1}{\sigma}\,\varphi\left(\frac{x-\mu}{\sigma}\right) = \sqrt{\frac{1}{2\pi\sigma^2}}\,\exp\left[-\,\frac{(x-\mu)^2}{2\sigma^2}\right] = \sqrt{\frac{w}{2\pi}}\,\exp\left[-\frac{w}{2}\,(x-\mu)^2\right] \qquad (7\text{–}34)$$

in which we introduce the "weight" $w \equiv 1/\sigma^2$ for convenience, the product of two such functions is

$$\varphi(x-\mu_1|\sigma_1)\,\varphi(y-x-\mu_2|\sigma_2) = \frac{1}{2\pi\,\sigma_1\,\sigma_2}\,\exp\left\{-\frac{1}{2}\left[\left(\frac{x-\mu_1}{\sigma_1}\right)^2 + \left(\frac{y-x-\mu_2}{\sigma_2}\right)^2\right]\right\} \qquad (7\text{–}35)$$

but we bring out the dependence on x by rearranging the quadratic form:

$$\left(\frac{x-\mu_1}{\sigma_1}\right)^2 + \left(\frac{y-x-\mu_2}{\sigma_2}\right)^2 = (w_1+w_2)\,(x-\hat{x})^2 + \frac{w_1\,w_2}{w_1+w_2}\,(y-\mu_1-\mu_2)^2 \qquad (7\text{–}36)$$

where $\hat{x} \equiv (w_1\mu_1 + w_2 y - w_2\mu_2)/(w_1+w_2)$. The product is still a gaussian with respect to x; so on integrating out x we have the convolution law:

‡ This notation is not quite inconsistent, since $\varphi(\;\;)$ and $\varphi(\;\;|\;\;)$ are different functional symbols.

$$\int_{-\infty}^{\infty} \varphi(x-\mu_1|\sigma_1)\,\varphi(y-x-\mu_2|\sigma_2)\,dx = \varphi(y-\mu|\sigma) \tag{7–37}$$

where $\mu \equiv \mu_1 + \mu_2$, $\sigma^2 \equiv \sigma_1^2 + \sigma_2^2$. Two gaussians convolve to make another gaussian, the means μ and variances σ^2 being additive.

The question whether nongaussian distributions also have parameters additive under convolution leads us to the notion of *cumulants* discussed in Appendix C. Now we note some important applications in biology, which require only the above convolution law.

Galton's Discovery

Although from our present standpoint (7–37) is only a straightforward integration formula, which we may write for present purposes in the form

$$\int_{-\infty}^{\infty} \varphi(x|\sigma_1)\,\varphi(y-ax|\sigma_2)\,dx = \varphi(y|\sigma) \tag{7–38}$$

where we have made the scale changes $x \to ax$, $\sigma_1 \to a\sigma_1$, and so now

$$\sigma = \sqrt{a^2\sigma_1^2 + \sigma_2^2} \tag{7–39}$$

it became in the hands of Francis Galton (1886) a major revelation about the mechanism of biological variation and stability.* We use the conventional language of that time, which did not distinguish between the notions of probability and frequency, using the words interchangeably. But this is not a serious matter because his data were, in fact, frequencies and as we shall see in Chapter 9, strict application of probability theory as logic would then lead to *probability* distributions that are substantially equal to the *frequency* distributions. Consider, for example, the frequency distribution of heights h of adult males in the population of England. He found that this could be represented fairly well by a gaussian

$$\varphi(h-\mu|\sigma)dh = \varphi\left(\frac{h-\mu}{\sigma}\right)\frac{dh}{\sigma} \tag{7–40}$$

with $\mu = 68.1$ inches, $\sigma = 2.6$ inches. Then he investigated whether children of tall parents tend to be tall, *etc.* To keep the number of variables equal to two in spite of the fact that each person has two parents, he determined that the average height of men was about 1.08 times that of women, and defined a person's "midparent" as an imaginary being of height

$$h_{mid} \equiv \frac{1}{2}\,(h_{father} + 1.08\,h_{mother}) \tag{7–41}$$

He collected data on 928 adults born of 205 midparents and found as expected that children of tall parents do indeed tend to be tall, etc. but that children of tall parents still show a spread in heights, although less than the spread ($\pm\sigma$) of the entire population.

But if the children of each selected group of parents still spread in height, why does the spread in height of the entire population not increase continually from one generation to the next? Because of the phenomenon of "reversion"; the children of tall parents tend to be taller than the average person, but less tall than their parents. Likewise children of short parents are generally shorter than the average person, but taller than their parents. If the population as a whole is to be stable,

* A photograph of Galton, with more details of his work and a short biographical sketch, may be found in Stigler (1986). His autobiography (Galton, 1908) has additional details.

this 'systematic' tendency to revert back to the mean of the entire population must exactly balance the 'random' tendency to spreading. Behind the smooth facade of a constant overall distribution of heights, an intricate little time–dependent game of selection, drift, and spreading is taking place constantly.

In fact, Galton (with some help from mathematicians) could predict the necessary rate of reversion theoretically, and verify it from his data. If $(x = h - \mu)$ is the deviation from the mean height of the midparents, let the population as a whole have a height distribution $\varphi(x/\sigma_1)$, while the sub–population of midparents of height $(x+\mu)$ tend to produce children of height $(y+\mu)$ with a frequency distribution $\varphi[(y-ax)/\sigma_2]$. Then the height distribution of the next generation will be given by (7–38). If the population as a whole is to be stable, it is necessary that $\sigma = \sigma_1$, or the reversion rate must be

$$a = \pm\sqrt{1 - \frac{\sigma_2^2}{\sigma_1^2}}\,. \tag{7–42}$$

which shows that a need not be positive; if tall people tended to have unusually short children, this would bring about an alternation from one generation to the next, but there would still be equilibrium for the population as a whole.

We see that equilibrium is not possible if $|a| > 1$; the population would explode. Although (7–38) is true for all a, equilibrium would then require $\sigma_2^2 < 0$. The boundary of stability is reached at $\sigma_2 = 0$, $|a| = 1$; then each sub–population breeds true, and whatever initial distribution of heights happened to exist, would be maintained thereafter. An economist might call the condition $a = 1$ a "unit root" situation; there is no reversion and no spreading.†

Of course, this analysis is in several obvious respects an oversimplified model of what happens in actual human societies. But that involves only touching up of details; Galton's analysis was, historically, of the greatest importance in giving us a general understanding of the kind of processes at work. For this, its freedom from nonessential details was a major merit.

Exercise 7.1. Galton's device of the midparent was only to reduce the computational burden, which would otherwise have been prohibitive in the 1880's. But today computing power is so plentiful and cheap that one can easily analyze the real three–variable problem, in which the heights of father, mother, and child are all taken into account. Reformulate Galton's problem to take advantage of this; what hypotheses about spreading and reversion might be considered and tested today? As a class project, one might collect new data (perhaps on faster–breeding creatures) and write the computer program to analyze them and estimate the new spreading and reversion coefficients.

Twenty years later this same phenomenon of selection, drift, and spreading underlying equilibrium was perceived independently by Einstein (1905) in physics. The steady thermal Boltzmann distribution for molecules at temperature T to have energy E is $\exp(-E/kT)$. Being exponential in energies $E = u + (mv^2/2)$ where $u(x)$ is potential energy, this is gaussian in particle velocities v. This generates a time–dependent drift in position; a particle which is at position x at time $t = 0$ has at time t the conditional probability to be at y of

$$p(y|x,t) \propto \exp\left(-\frac{(y-x)^2}{4Dt}\right)$$

from random drift alone, but this is countered by a steady drift effect of external forces $F = -\nabla u$, corresponding to Galton's reversion rate.

† It is a currently popular theory among some economists that many economic processes, such as the stock market, are very close to the unit root behavior, so that the effects of momentary external perturbations like wars and droughts tend to persist instead of being corrected.

Although the details are quite different, Galton's Eq. (7–42) is the logical equivalent of Einstein's relation $D = \lambda kT$ connecting diffusion coefficient D representing random spreading of particles, with the temperature T and the mobility λ (velocity per unit force) representing the systematic reversion rate counteracting the diffusion. Both express the condition for equilibrium as a balance between a random "spreading" tendency, and a systematic counter–drift that holds it in check.

Population Dynamics and Darwinian Evolution

Galton's type of analysis can explain much more than biological equilibrium. Suppose the reversion rate does not satisfy (7–42). Then the height distribution in the population will not be static, but will change slowly. Or, if short people tend to have fewer children than do tall people, then the average height of the population will drift slowly upward. Do we have here the mechanism for Darwinian evolution? The question could hardly go unasked, since Francis Galton was a cousin of Charles Darwin.

A new feature of probability theory has appeared here, that is not evident in the works of Laplace and Gauss. Being astronomers, their interests were in learning facts of astronomy and telescopes were only a tool toward that end. The vagaries of telescopes themselves were for them only 'errors of observation' whose effects were to be eliminated as much as possible; and so the sampling distribution was called by them an 'error law'.

But a telescope maker might see it differently. For him, the errors it produces are the objects of interest to study, and a star is only a convenient fixed object on which to focus his instrument for the purpose of determining those errors. Thus a given data set might serve two entirely different purposes; one man's 'noise' is another man's 'signal'.

But then in any science the 'noise' might prove to be not merely something to get rid of, but the essential phenomenon of interest. It seems curious (at least, to a physicist) that this was first seen clearly not in physics, but in biology. In the late Nineteenth Century many biologists saw it as the major task confronting them to verify Darwin's theory of evolution by exhibiting the detailed mechanism by which evolution takes place. For this purpose the journal *Biometrika* was founded by Karl Pearson and Walter Frank Raphael Weldon, in 1901. It started (Volume 1, page 1) with an editorial setting forth the journal's program, in which Weldon wrote:

> "The starting point of Darwin's theory of evolution is precisely the existence of those differences between individual members of a race or species which morphologists for the most part rightly neglect. The first condition necessary, in order that a process of Natural Selection may begin among a race, or species, is the existence of differences among its members; and the first step in an enquiry into the possible effect of a selective process upon any character of a race must be an estimate of the frequency with which individuals, exhibiting any degree of abnormality with respect to that character, occur."

He had here reached a very important level of understanding. Morphologists, thinking rather like astronomers, considered individual variations as only 'noise' whose effects must be eliminated by averaging, in order to get at the significant 'real' properties of the species as a whole. Weldon, learning well from the example of Galton, saw it in just the opposite light; those individual variations *are the engine that drives the process of evolutionary change*, which will be reflected eventually in changes in the morphologist's averages. Indeed, without individual variations, the mechanism of Natural Selection has nothing to operate on. So to demonstrate the mechanism of evolution at its source, and not merely the final result, it is the distribution of individual variations that must be studied.

Of course, at that time they had no conception of the physical mechanism of mutations induced by radioactivity (much less by genetic recombination or errors in DNA replication), and they expected that evolution would be found to take place via slow, nearly continuous changes.†

† The necessity for evolution to be particulate (by discrete steps) was perceived later by several people, including R. A. Fisher (1930b). Evolutionary theory taking this into account is often called *Neo–Darwinism*.

Nevertheless, the program of studying the individual variations would be the correct one to find the fundamental mechanism of evolution, whatever form it took. The scenario is somewhat like the following:

Evolution of Humming–Birds and Flowers. Consider a population of humming–birds in which the "noise" consists of a distribution of different length beaks. The survival of birds is largely a matter of finding enough food; then a bird that finds itself with the mutation of an unusually long beak will be able to extract nectar from deeper flowers. If such flowers are available it will be able to nourish itself and its babies better than others because it has a food supply not available to other birds; so the long–beak mutation will survive and become a greater portion of the bird population, in more or less the way Darwin imagined.

But this influence works in two directions; a bird is inadvertently fertilizing flowers by carrying a few grains of pollen from one to the next. A flower that happens to have the mutation of being unusually deep will find itself sought out preferentially by long–beaked birds because they need not compete with other birds for it. Therefore its pollen will be carried systematically to other flowers of the same species where it is effective, instead of being wasted on the wrong species. As the number of long–beaked birds increases, deep flowers thus have an increasing survival advantage, ensuring that their mutation is present in an increasing proportion of the flower population; and this in turn gives a still greater advantage to long–beaked birds; and so on.

Over millions of years, this back–and–forth reinforcement of mutations goes through hundreds of cycles, resulting eventually in a symbiosis so specialized – a particular species of bird and a particular species of flower that seem designed specifically for each other – that it appears to be a miraculous proof of a guiding purpose in Nature, to one who does not think as deeply as did Darwin and Galton.‡ Yet short–beaked birds do not die out, because birds patronizing deep flowers leave the shallow flowers for them; the process tends to an equilibrium distribution of populations of short and long beaked birds, coupled to distributions of shallow and deep flowers.

Discussion. As noted, the role of "noise" as the mechanism driving a slow change in a system was perceived independently by Einstein (of course, he knew about Darwin's theory, but we think it highly unlikely that he would have known about the work of Galton or Weldon in Switzerland in 1905). Random thermal fluctuations caused by motion of individual atoms are not merely 'noise' to be averaged out in our predictions of mass behavior; they are *the engine that drives irreversible processes in physics*, and eventually brings about thermal equilibrium.

The visions of Weldon and Einstein represented such a major advance in thinking that today, some 100 years later, many have not yet comprehended them or appreciated their significance in either biology or physics. We still have biologists⋆ who try to account for evolution by a quite unnecessary appeal to the second law of thermodynamics; and physicists who try to account for the second law by appealing to quite unnecessary modifications in the equations of motion. The operative mechanism of evolution is surely Darwin's original principle of natural selection and any effects of the second law would only hinder it.† Natural Selection is a process entirely different

‡ The unquestioned belief in such a purpose pervades even producers of biological research products who might be expected to know better. In 1993 there appeared in biological trade journals a full–page ad with a large color photograph of a feeding hummingbird and the text: "**Specific purpose.** The sharply curved bill of the white–tipped sickle–billed hummingbird is specifically adapted to probe the delicate tubular flowers of heliconia plants for the nectar on which the creature survives. High–performance [*Brand Name*] Taq DNA polymerase is designed and tested with an equally specific goal in mind $\cdots$."

⋆ For example, see Weber, *et al*, (1988). Here the trouble is that it goes in the wrong direction; if the second law were the driving principle, evolution would proceed inexorably back to the primordial soup, which has a much higher entropy than would any collection of living creatures that might be made from the same atoms.

† This is not to say that natural selection is the *only* process at work; random drift is still an operative cause of evolution with or without subsequent selection. For an extensive discussion of the evidence and later

from the second law of thermodynamics. The purposeful intervention of man can suspend or reverse natural selection – as we observe in wars, medical practice, and dog breeding – but it can hardly affect the second law.

The remarkable – almost exact – analogy between the processes that bring about equilibrium in physics and in biology surely has other important implications, particularly for theories of equilibrium in economics, not yet exploited. It seems likely, for example, that the "turbulence" of individual variations in economic behavior is the engine that drives macroeconomic change toward the equilibrium envisaged by Adam Smith. The existence of this turbulence was recognized by John Maynard Keynes (1936), who called it 'animal spirits' which cause people to behave erratically; but he did not see in this the actual cause that moves the economy toward equilibrium.

In the next level of understanding we see that Adam Smith's equilibrium is never actually attained in the real world because of what a physicist would call 'external perturbations', or an economist 'exogenous variables' which vary on the same time scale. That is, wars, droughts, taxes, tariffs, bank reserve requirements, discount rates and other disturbances come and go on about the same time scale as would the approach to equilibrium in a perfectly 'calm' society. As we see it, this is the basic reason why economic data are very difficult to interpret; even if relevant and believable data were easy to gather, the rules of the game and the conditions of play are changing constantly.

Resolution of Distributions into Gaussians

The tendency of probability distributions to gravitate into the gaussian form suggests that we might view the appearance of a gaussian, or 'normal' frequency distribution as loose evidence (but far from proof) that some kind of equilibrium has been reached. This view is also consistent with (but by no means required by) the results of Galton and Einstein. In the first attempts to apply probability theory in the biological and social sciences (for example, Quetelet, 1835, 1869) serious errors were made through supposing firstly that the appearance of a normal distribution in data indicates that one is sampling from a homogeneous population, and secondly that any departure from normality indicates an inhomogeneity in need of explanation; by resolving it into gaussians Quetelet thought that one would be discovering the different sub–species, or varieties, that were present in the population. If this were true reliably, we would indeed have a powerful tool for research in many different fields. But later study showed that the situation is not that simple.

We have just seen how one aspect of it was corrected finally by Galton (1886), in showing that a normal frequency distribution by no means proves homogeneity; from (7–38) a gaussian of width σ can arise inhomogeneously – and in many different ways – from the overlapping of narrower gaussian distributions of various widths σ_1, σ_2. But those sub–populations are in general merely mathematical artifacts like the sine waves in a Fourier transform; they have no individual significance for the phenomenon unless one can show that a particular set of sub–populations has a real existence and plays a real part in the mechanism underlying stability and change. Galton was able to show this from his data by measuring those widths.

The second assumption, that non–normal distributions can be resolved into gaussian subdistributions, turns out to be not actually wrong (except in a nit–picking mathematical sense); but without extra prior information it is ambiguous in what it tells us about the phenomenon.

We have here an interesting problem, with many useful applications: is a nongaussian distribution explainable as a mixture of gaussian ones? Put mathematically, if an observed data histogram is well described by a distribution $g(y)$, can we find a mixing function $f(x) \geq 0$ such that $g(y)$ is seen as a mixture of gaussians:

---

research efforts by many experts, see the massive three–volume work *Evolution After Darwin* (Tax, 1960) produced to mark the Centenary of the publication of Darwin's *Origin of Species*, or the more informal work of Dawkins (1987).

$$\int \varphi(y - x|\sigma)\, f(x)\, dx = g(y)\,, \qquad -\infty \le y \le \infty \tag{7–43}$$

Neither Quetelet nor Galton was able to solve this problem, and today we understand why. Mathematically, does this integral equation have solutions, or unique solutions? It appears from (7–41) that we cannot expect unique solutions in general, for in the case of gaussian $g(y)$, many different mixtures [many different choices of a, σ_1, σ_2] will all lead to the same $g(y)$. But perhaps if we specify the width σ of the gaussian kernel in (7–43) there is a unique solution for $f(x)$.

Solution of such integral equations is rather subtle mathematically. We give two arguments; the first depends on the properties of hermite polynomials and yields a class of exact solutions, the second appeals to fourier transforms and yields an understanding of the more general situation.

Hermite Polynomial Solutions. The rescaled hermite polynomials $R_n(x)$ may be defined by the displacement of a gaussian distribution $\varphi(x)$, which gives the generating function:

$$\frac{\varphi(x-a)}{\varphi(x)} = e^{xa - a^2/2} = \sum_{n=0}^{\infty} R_n(x)\, \frac{a^n}{n!} \tag{7–44}$$

or, solving for R_n, we have the Rodriguez form

$$R_n(x) = \frac{d^n}{da^n}\left[e^{xa-a^2/2}\right]_{a=0} = (-1)^n\, e^{x^2/2}\, \frac{d^n}{dx^n}\, e^{-x^2/2}\,. \tag{7–45}$$

The first few of these polynomials are: $R_0 = 1$, $R_1 = x$, $R_2 = x^2 - 1$, $R_3 = x^3 - 3x$, $R_4 = x^4 - 6x^2 + 3$. The conventional hermite polynomials $H_n(x)$ differ only in scaling: $H_n(x) = 2^{n/2}\, R_n(x\sqrt{2})$.

Multiplying (7–44) by $\varphi(x)\exp(xb - b^2/2)$ and integrating out x, we have the orthogonality relation

$$\int_{-\infty}^{\infty} R_m(x)\, R_n(x)\, \varphi(x)\, dx = n!\, \delta_{mn} \tag{7–46}$$

and in consequence these polynomials have the remarkable property that convolution with a gaussian function reduces simply to

$$\int_{-\infty}^{\infty} \varphi(y-x)\, R_n(x)\, dx = y^n\,. \tag{7–47}$$

Therefore if $g(y)$ is represented by a power series

$$g(y) = \sum_n a_n\, y^n \tag{7–48}$$

we have immediately a formal solution of (7–43):

$$f(x) = \sum_n a_n \sigma^n\, R_n\left(\frac{x}{\sigma}\right)\,. \tag{7–49}$$

Since the coefficient of x^n in $R_n(x)$ is unity, the expansions (7–48) and (7–49) converge equally well. So if $g(y)$ is any polynomial or entire function (*i.e.* one representable by a power series (7–48) with infinite radius of convergence), the integral equation has the unique solution (7–49).

We can see the solution (7–49) a little more explicitly if we invoke the expansion of R_n, deducible from (7–44) by expanding $\exp(xa - a^2/2)$ in a power series in x:

$$R_n\left(\frac{x}{\sigma}\right) = \sum_{m=0}^{M}(-1)^m \frac{n!}{2^m m!\,(n-2m)!}\left(\frac{x}{\sigma}\right)^{n-2m} \tag{7–50}$$

where $M = (n-1)/2$ if n is odd, $M = n/2$ if n is even. Then noting that

$$\frac{n!}{(n-2m)!}\left(\frac{x}{\sigma}\right)^{n-2m} = \sigma^{2m-n}\frac{d^{2m}}{dx^{2m}}x^n \tag{7–51}$$

we have the formal expansion

$$\begin{aligned} f(x) &= \sum_{m=0}^{\infty}\frac{(-1)^m\sigma^{2m}}{2^m m!}\frac{d^{2m}}{dx^{2m}}g(x) \\ &= g(x) - \frac{\sigma^2}{2}\frac{d^2g(x)}{dx^2} + \frac{\sigma^4}{8}\frac{d^4g(x)}{dx^4} - \cdots . \end{aligned} \tag{7–52}$$

An analytic function is differentiable any number of times, so if $g(x)$ is an entire function this will converge to the unique solution. If $g(x)$ is a very smooth function, it converges very rapidly, so the first two or three terms of (7–52) are already a good approximation to the solution. This gives us some insight into the workings of the integral equation; as $\sigma \to 0$, the solution (7–52) relaxes into $f(x) \to g(x)$, as it should. The first two terms of (7–52) are what would be called, in image reconstruction, 'edge detection'; for small σ the solution goes into this. The larger σ, the more the higher order derivatives matter; that is, the more fine details of the structure of $g(y)$ contribute to the solution. Intuitively, the broader the gaussian kernel, the more difficult it is to represent fine structure of $g(y)$ in terms of that kernel.

Evidently, we could continue this line of thought with much more analytical work, and it might seem that the problem is all but solved; but now the subtlety starts. Formal solutions like (7–49) and (7–52), although correct mathematically, ignore some facts of the real world; is $f(x)$ non–negative when $g(y)$ is? Is the solution stable, a small change in $g(y)$ inducing only a small change in $f(x)$? What if $g(x)$ is not an entire function but is piecewise continuous; for example, rectangular?

Fourier Transform Relations. For some insight into these questions, let us look at the integral equation from the fourier transform viewpoint. Taking the transform of (7–43) according to

$$F(k) \equiv \int_{-\infty}^{\infty} f(x)\,e^{ikx}\,dx \tag{7–53}$$

(7–43) reduces to

$$\exp\left(-\frac{k^2\sigma^2}{2}\right)F(k) = G(k) \tag{7–54}$$

which illustrates that the fourier transform of a gaussian function is another gaussian function, and shows us at once the difficulty of finding more general solutions than (7–49). If $g(y)$ is piecewise continuous, then as $k \to \infty$, from the Riemann–Lebesgue lemma $G(k)$ will fall off only as $1/k$. Then $F(k)$ must blow up violently, like $\exp(+k^2\sigma^2/2)/k$, and one shudders to think what the function $f(x)$ must look like (infinitely violent oscillations of infinitely high frequency?). If $g(y)$ is continuous, but has discontinuous first derivatives, like a triangular distribution, then $G(k)$ falls off as k^{-2}, and we are in a situation about as bad. Evidently, if $g(y)$ has a discontinuity in any derivative, there is no solution $f(x)$ that would be acceptable in the physical problem. This is evident also from (7–52).

In order that we can interpret $g(y)$ as a mixture of possible gaussians, $f(x)$ must be non–negative. But we must allow the possibility that the $f(x)$ sought is a sum of delta-functions; indeed, to resolve $g(y)$ into a discrete mixture of gaussians $g(y) = \sum a_j\, \varphi(x - x_j)$ was the real goal of Quetelet and Galton. If this could be achieved uniquely, their interpretation might be valid. Then $F(k)$ does not fall off at all as $k \to \pm\infty$, so $G(k)$ must fall off as $\exp(-\,k^2\sigma^2/2)$. In short, in order to be resolvable into gaussians of width σ with positive mixture function $f(x)$, the function $g(y)$ must itself be at least as smooth as a gaussian of width σ. This is a formal difficulty.

There is a more serious practical difficulty. If $g(y)$ is a function determined only empirically, we do not have it in the form of an analytic function; we have only a finite number of approximate values g_i at discrete points y_i. We can find many analytic functions which appear to be good approximations to the empirical one. But because of the instability evident in (7–52) and (7–54) they will lead to greatly different final results $f(x)$. Without a stability property and a criterion for choosing that smooth function, we really have no definite solution in the sense of inversion of an integral equation.*

In other words, finding the appropriate mixture $f(x)$ to account for an empirically determined distribution $g(y)$ is not a conventional mathematical problem of inversion; *it is itself a problem of inference, requiring the apparatus of probability theory.* In this way, a problem in probability theory can generate a hierarchy of sub–problems, each involving probability theory again but on a different level.

There is Hope After All. But following up this idea, the original goal of Quetelet has now been very nearly realized by analysis of the integral equation as a problem of Bayesian inference instead of mathematical inversion; and useful examples of analysis of real data by this have now been found. Sivia and Carlile (1992) report the successful resolution of noisy data into as many as nine different gaussian components, representing molecular excitation lines, by a Bayesian computer program.† We consider the theory of this in connection with the general spectrum/shape problem in Chapter 23.

It is hardly surprising that Quetelet and Galton could not solve this problem in the last Century; but it is very surprising that today many scientists, engineers, and mathematicians still fail to see the distinction between inversion and inference, and struggle with problems like this that have no deductive solutions, only inferential ones. The problem is, however, very common in current applications; it is known as a "generalized inverse" problem, and today we can give unique and useful inferential solutions to many such problems.

This suggests another interesting mathematical problem; for a given entire function $g(y)$, over what range of σ is the solution (7–49) nonnegative? There are some evident clues; when $\sigma \to 0$, we have $\varphi(x - y|\sigma) \to \delta(x - y)$ and so as noted above, $f(x) \to g(x)$, so for σ sufficiently small $f(x)$ will be nonnegative if $g(y)$ is. But when $\sigma \to \infty$ the gaussians in (7–43) become very broad and

---

* For other discussions of the problem, see Andrews & Mallows (1974); Titterington, *et al*, (1985).

† We noted already in Chapter 1 that most of the computer programs used in this field are only intuitive *ad hoc* devices that make no use of the principles of probability theory; therefore in general they are usable in some restricted domain, but they fail to extract all the relevant information from the data and are subject to both the errors of hallucination and blunt perception. One commercial program for resolution into gaussians or other functions simply reverts to empirical curve–fitting. It is advertised [*Scientific Computing*, July 1993, p. 15] with a provocative message. It depicts two scientists with the same data curve showing two peaks; by hand drawing one could resolve it very crudely into two gaussians. The ad proclaims: "Dr. Smith found two peaks . . . Using [our program] Dr. Jones found *three* peaks . . .. **Guess who got the grant?** We are encouraged to believe that we can extract money from the Government by first allowing the software company to extract \$500 from us for this program, whose output would indeed be tolerable for noiseless data. But it would surely degenerate quickly into dangerous, unstable nonsense as the noise level increases. A Bayesian inference program will continue to return the best resolution possible from the data and the model, without instability, whatever the noise level (Chapter 23).

smooth; and so if $f(x)$ is nonnegative, the integral in (7–43) must be at least as broad. Thus when $g(y)$ has detailed structure on a scale smaller than σ, there can be no solution with nonnegative $f(x)$; and it is not obvious whether there can be any solution at all.

Exercise 7.2. From the above arguments we conjecture that there will be some upper bound σ_{max} such that the solution $f(x)$ is nonnegative when and only when $0 \le \sigma < \sigma_{max}$. It will be some functional $\sigma_{max}[g(y)]$ of $g(y)$. Prove or disprove this conjecture; if it is true, give a verbal argument by which we could have seen this without calculation; if it is false, give a specific counter–example showing why. It appears that (7–52) might be useful in this endeavor.

But this suggests that the original goal of Quetelet and Galton was ambiguous; any sufficiently smooth nongaussian distribution may be generated by many different superpositions of different gaussians of different widths. Therefore a given set of sub–populations, even if found mathematically, would have little biological significance unless there were additional prior information pointing to gaussians of that particular width σ as having a "real" existence and playing some active role in the phenomena. Of course, this *caveat* applies equally to the aforementioned Bayesian solution; but Sivia and Carlile did have that prior information.

The above applications required only the single convolution formula (7–37). Now we turn to an application which uses repeated convolutions.

The Central Limit Theorem

The reader who has not yet studied Appendix C on convolutions and cumulants, should do so now. If the functions $f_i(x)$ to which we apply that theory are probability distributions, then they are necessarily nonnegative and normalized: $f_i(x) \ge 0\,, \int f_i(x)dx = 1$. Then the zero'th moments are all $Z_i = 1$, and the fourier transforms

$$F_i(\alpha) \equiv \int_{-\infty}^{\infty} f_i(x)\, e^{i\alpha x}\, dx \tag{7–55}$$

are absolutely convergent for real α. Note that all this remains true if the f_i are discontinuous, or contain delta–functions; therefore the following derivation will apply equally well to the continuous or discrete case or any mixture of them.‡

******************** MUCH MORE TO COME HERE! *********

Accuracy of Computations

As a useful application of the Central Limit Theorem, consider a computer programmer deciding on the accuracy to be used in a program. This is always a matter of compromise between misleading, inaccurate results on the one hand, and wasting computation facilities with more accuracy than needed on the other.

Of course, it is better to err on the side of a little more accuracy than really needed. Nevertheless, it is foolish (and very common) to tie up a large facility with a huge computation to double precision (16 decimal places) or even higher, when the user has no use for anything like that accuracy in the final result. The computation might have been done in less time but with the same result on a desktop microcomputer, had it been programmed for an accuracy that is reasonable for the problem.

Programmers can speed up and simplify their creations by heeding what the CLT tells us. In probability theory we seldom have any serious need for more than three–figure accuracy in our

‡ At this point, the reader who has been taught to disbelieve in delta–functions must unlearn that by reading Appendix B. In fact, without the free use of delta–functions and other generalised functions, real applications of fourier analysis are in an almost helpless, crippled condition, as noted in Appendix F.

final results, so we shall be on the safe side if we strive to get four figure accuracy reliably in our computations.

As a simple example, suppose we are computing the sum

$$S \equiv \sum_{n=1}^{N} a_n \tag{7–56}$$

of N terms a_n, each one positive and of order unity. To get a given accuracy in the sum, what accuracy do we need in the individual terms?

Our computation program or lookup table necessarily gives each a_n digitized to some smallest increment ϵ, so this will be actually the true value plus some error e_n. If we have a_n to six decimal digits, then $\epsilon = 10^{-6}$; if we have it to sixteen binary digits, then $\epsilon = 2^{-16} = 1/65536$. The error in any one entry is in the range $(-\epsilon/2 < e_n \le \epsilon/2)$, and in adding N such terms the maximum possible error is $N\epsilon/2$. Then it might be thought that the programmer should ensure that this is acceptably small.

But if N is large, this maximum error is enormously unlikely; this is just the point that Euler failed to see. The individual errors are almost certain to be positive and negative about equally often, giving a high degree of mutual cancellation, so that the net error should tend to grow only as $\sqrt{N}$.

The CLT tells us what is essentially a simple combinatorial fact, that out of all conceivable error vectors $\{e_1, \ldots, e_N\}$ that could be generated, the overwhelming majority have about the same degree of cancellation, that indicated by the $\sqrt{N}$ rule. If we consider each individual error equally likely to be anywhere in $(-\epsilon/2, \epsilon/2)$, this corresponds to a rectangular probability distribution on that interval, leading to an expected square error per datum of

$$\frac{1}{\epsilon}\int_{-\epsilon/2}^{\epsilon/2} x^2\, dx = \frac{\epsilon^2}{12} \tag{7–57}$$

Then by the CLT the probability distribution for the sum S will be gaussian with a variance $N\epsilon^2/12$, while S is approximately N. So if N is large so that the CLT is accurate, the probability that the magnitude of the net error will exceed $\epsilon\sqrt{N}$, which is $\sqrt{12} = 3.46$ standard deviations, is about

$$2[1 - \Phi(3.46)] \simeq 0.0006 \tag{7–58}$$

where $\Phi(x)$ is the cumulative normal distribution. One will almost never observe an error that great. Since $\Phi(2.58) = 0.995$, there is about a 1% chance that the net error magnitude will exceed 2.58 standard deviations, or $0.74\epsilon\sqrt{N}$.

Therefore if we strive, not for certainty, but for greater than 99% probability, that our sum S is correct to four figures, this indicates the value of ϵ that can be tolerated in our algorithm or lookup table. We require $\epsilon\sqrt{N} \le 10^{-4}N$, or

$$\epsilon \le 10^{-4}\sqrt{N} \tag{7–59}$$

The perhaps surprising result is that if we are adding $N = 100$ roughly equal terms, to achieve a virtual certainty of four–figure accuracy in the sum we require only three–figure accuracy in the individual terms. Thus we can get by with a considerably shorter computation for the individual terms, or a smaller lookup table, than might be supposed.

This simple calculation can be greatly generalized, as indicated by Exercise 7.4. But we should note an important proviso to be investigated in Exercise 7.5; this holds only when the individual errors e_n are logically independent. Given ϵ in advance, if knowing e_1 then tells us anything about any other e_n, then the errors are correlated in our probability assignment, the CLT no longer applies, and a different analysis is required. In practice this is almost never a serious limitation because the individual a_n are determined by some continuously variable algorithm and differ among themselves by amounts large compared to ϵ, making it impossible to determine any e_n given any other e_m.

Exercise 7.4. Suppose that we are to evaluate a fourier series $S(\theta) = \sum a_n \sin n\theta$. Now the individual terms vary in magnitude and are themselves both positive and negative. In order to achieve four–figure accuracy in $S(\theta)$ with high probability, what accuracy do we now require in the individual values of a_n and $\sin n\theta$?

Exercise 7.5. Show that if there is a positive correlation in the e_n, then the error in the sum may be much greater than indicated by the CLT. Try to make a more sophisticated probability analysis taking correlations into account, which would be helpful to a computer programmer striving for the greatest efficiency for a given accuracy.

COMMENTS

Terminology Again. As we are obliged to point out so often, this field seems to be cursed more than any other with bad and misleading terminology which seems impossible to eradicate. The Electrical Engineers have solved this problem very effectively; every few years an Official Committee issues a Revised Standard Terminology, which is then enforced by editors of their journals (witness the meek acceptance of the change from 'megacycles' to 'megahertz' which was accomplished almost overnight a few years ago).

In probability theory there is no Central Authority with the power to bring about dozens of needed reforms, and it would be self–defeating for any one author to try to do this by himself; he would only turn away readers. But we can offer tentative suggestions in the hope that others may see merit in them.

The literature gives conflicting evidence about the origin of the term "Normal distribution". Karl Pearson (1920) claimed to have introduced it "many years ago", in order to avoid an old dispute over priority between Gauss and Legendre; but he gave no reference. Hilary Seal (1967) attributes it instead to Galton, but again fails to give a reference. However, the term had long been associated with the general topic; given a linear model $y = X\beta + e$, Gauss (1823) called the system of equations $X'X\hat{\beta} = X'y$, which give the least squares parameter estimates $\hat{\beta}$, the "normal equations" and the ellipsoid of constant probability density was called the "normal surface." It appears that somehow the name got transferred from the equations to the sampling distribution that leads to those equations.

Presumably, Gauss meant "normal" in its mathematical sense of "perpendicular" expressing the geometric meaning of those equations. The minimum distance from a point (the estimate) to a plane (the constraint) is the length of the perpendicular. But, as Pearson himself observes, the term "normal distribution" is a bad one because the common colloquial meaning of 'normal' is *standard* or *sane*, implying a value judgment. This leads many to think – consciously or subconsciously – that all other distributions are in some way abnormal.

Actually, it is quite the other way; it is the so–called 'normal' distribution that is abnormal in the sense that it has many unique properties not possessed by any other. Almost all of our experience in inference has been with this abnormal distribution, and much of the folklore that we must counter here was acquired as a result. For decades, workers in statistical inference have been misled, by that abnormal experience, into thinking that methods such as confidence intervals, that happen to work satisfactorily with this distribution, should work as well with others.

The alternative name "gaussian distribution" is equally bad for a different reason, although there is no mystery about its origin. Stigler (1980) sees it as a general law of eponymy that *no discovery is named for its original discoverer*. Our terminology is in excellent compliance with this law, since the distribution itself had been found by de Moivre (1733) before Laplace was born; and the fundamental nature of this distribution and its main properties were noted by Laplace (1783) when Gauss was six years old. But as we noted, the distribution became popularized by the work of Gauss (1809) who gave a derivation of it that was simpler than previous ones and seemed very compelling intuitively at the time. This is the derivation that we gave above, Eq. (7–15), and which resulted in his name becoming attached to it.

The term "Central distribution" would avoid both of these objections while conveying a correct impression; it is the final 'stable' or 'equilibrium' distribution toward which all others gravitate under a wide variety of operations (large number limit, convolution, stochastic transformation, etc.) and which, once attained, is maintained through an even greater variety of transformations, some of which are still unknown to statisticians because they have not yet come up in their problems.

Fopr example, Boltzmann showed that collisions in a gas tend to bring about a "Maxwellian" or Gaussian, distribution of velocities. Then Kennard (1936, Chap. 3) showed that this distribution, once attained, is maintained automatically, without any help from collisions, as the molecules move about, constantly changing their velocities, in any conservative force field. Thus this distribution has stability properties considerably beyond anything yet utilized by statisticians, or yet demonstrated in the present work.

While venturing to use the term 'Central Distribution' in a cautious, tentative way, we continue to use also the bad but traditional terms, preferring "gaussian" for two reasons. Ancient questions of priority are no longer of interest; far more important today, "gaussian" does not imply any value judgment. Use of emotionally loaded terms appears to us a major cause of the confusion in this field, leading workers to adhere to nice–sounding principles like 'unbiased' or 'admissible' or 'uniformly most powerful', in spite of the nonsensical results they can give in practice. But also, we are writing for an audience that includes both statisticians and scientists. Everybody understands what "Gaussian distribution" means; but only statisticians are familiar with the term "Normal distribution".

The fundamental Boltzmann distribution of statistical mechanics, exponential in energies, is of course Gaussian or Maxwellian in particle velocities. The general central tendency of probability distributions toward this final form is now seen as a consequence of their maximum entropy properties (Chapter 11). If a probability distribution is subjected to some transformation that discards information but leaves certain quantities invariant, then under very general conditions, if the transformation is repeated, the distribution tends to the one with maximum entropy subject to the constraints of those invariant quantities.

This brings us to the term "Central Limit Theorem" which we have derived as a special case of the phenomenon just noted – the behavior of probability distributions under repeated convolutions. This name was introduced by George Pólya (1920), with the intention that the adjective 'central' was to modify the noun 'theorem'; *i.e.* it is the limit theorem which is *central to probability theory.* Almost universally, students today think that 'central' modifies 'limit', so that it is instead a theorem about a '*central limit*', whatever that means.†

But in view of the equilibrium phenomenon, it appears that Pólya's choice of words was after all fortunate in a way that he did not foresee. Our suggested terminology takes advantage of this; looked at in this way, the terms 'Central distribution' and 'Central limit theorem' both convey the right connotations to one hearing them for the first time. One can read 'Central limit' as meaning a limit toward a central distribution, and will be invoking just the right intuitive picture.

The Great Inequality of Jupiter and Saturn

An outstanding problem for 18'th Century science was noted by Edmund Halley in 1676. Observation showed that the mean motion of Jupiter (30.35 degrees per year) was slowly accelerating, that

† The confusion does not occur in the original German, where Pólya's words were: *Über den zentralen Grenzwertsatz der Wahrscheinlichkeitsrechnung*, an interesting example where the German habit of inventing compound words removes an ambiguity in the literal English rendering.

of Saturn (12.22 deg/yr) decelerating. But this was not just a curiosity for astronomers; it meant that Jupiter was drifting closer to the sun, Saturn farther away. If this trend were to continue indefinitely, then eventually Jupiter would fall into the sun, carrying with it the earth and all the other inner planets. This seemed to prophesy the end of the world – and in a manner strikingly like the prophesies of the Bible.

Understandably, this situation was of more than ordinary interest, and to more people than astronomers. Its resolution called forth some of the greatest mathematical efforts of 18'th Century savants, either to confirm the coming end; or preferably to show how the Newtonian laws would eventually put a stop to the drift of Jupiter and save us.

Euler, Lagrange, and Lambert made heroic attacks on the problem without solving it. We noted above how Euler was stopped by a mass of overdetermined equations; 75 simultaneous but inconsistent equations for 8 unknown parameters. If the equations were all consistent, he could choose any eight of them and solve (this would still involve inversion of an 8×8 matrix), and the result would be the same whatever eight he chose. But the observations all had unknown errors of measurement, and so there were

$$\binom{75}{8} \simeq 1.69 \times 10^{10}$$

possible choices. There are over sixteen billion different sets of estimates for the parameters, with apparently nothing to choose between them.* At this point, Euler managed to extract reasonably good estimates of two of the unknowns (already an advance over previous knowledge); and simply gave up on the others. For this work (Euler, 1749) he won the French Academy of Sciences prize.

The problem was finally solved in 1787 by one who was born that same year. Laplace (1749–1827) "saved the world" by using probability theory to estimate the parameters accurately enough to show that the drift of Jupiter was not secular after all; the observations at hand had covered only a fraction of a cycle of an oscillation with a period of about 880 years. This is caused by an "accidental" near resonance in their orbital periods:

$$2 \times (\text{period of Saturn}) \simeq 5 \times (\text{period of Jupiter})$$

Indeed, from the above mean motion data we have

$$2 \times \frac{360}{12.22} = 58.92\, yr\,, \qquad 5 \times \frac{360}{30.35} = 59.32\, yr\,.$$

In the time of Halley their difference was only about 0.66 percent and decreasing.

So long before it becomes a danger to us, Jupiter indeed reverses its drift and returns to its old orbit. Presumably, Jupiter and Saturn have repeated this seesaw game several million times since the solar system was formed. The first half–cycle of this oscillation to be observed by man will be completed in about the year 2012.

* Our algorithm for this in Chapter 19 [Eq's (19–24), (19–37)] actually calculates a weighted average over all these billions of estimates; but in a manner so efficient that one is unaware of this.