The Normal Distribution

2. Some Roles of Normality

S Ramasubramanian

The computational aid of part 1 of this series turns out to be a basic feature of nature, thanks to the central limit theorem. The role of normal distribution in statistics, velocity distribution of an ideal gas, and the phenomenon of Brownian motion is briefly illustrated.

S Ramasubramanian is with the Bangalore Centre of the Indian Statistical Institute. He received his Ph.D. from the Indian Statistical Institute in 1982. His research interests centre around diffusion processes.

The previous article of this series was:
1. From binomial to normal, June 1997.

Introduction

To compensate for the hard work done in part I of this series, we basically pontificate in this article. Mathematical details are side-stepped and we indulge in a lot of 'hand-waving', especially in the section on Brownian motion.

It is pointed out that the DeMoivre-Laplace theorem is just a special case of a far reaching general result known as 'the central limit theorem' in probability theory. Consequently our convenient computational aid turns out to be actually a basic principle of nature. To illustrate this point we cite three instances where the normal distribution plays a major role. First, we indicate how the central limit theorem provides a mathematical framework for the bulk of Statistics.

Next, we show that the normal distribution in a sense governs the velocity distribution of an ideal gas, and the phenomenon of Brownian motion; in both these cases one starts out with reasonable qualitative assumptions which are compatible with experience. (In our presentation the constants involved are either suppressed or standardised - something a physicist will justifiably disapprove of. However, as the present author is incapable of understanding physics, he may be excused!)

Central Limit Theorem

The discussion in part 1 of this article raises the question: Is it possible to have reasonable approximations for other random phenomena so that computations may become simple? To indicate an answer we rephrase the De Moivre-Laplace theorem as follows.

Theorem : Let $X_1, X_2, \cdots$ be independent random variables each having a Bernoulli distribution with expectation p and variance $p(1-p)$. Let $S_n = X_1 + X_2 + \cdots + X_n$ and $Z_n = \frac{S_n - E(S_n)}{\sqrt{Var(S_n)}}, n = 1, 2, \cdots$. Then for any $a < b$, $\lim_{n\to\infty}$ Prob $(a < Z_n \leq b) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx$.

It is a deep result in probability theory that the above conclusion holds not just for Bernoulli distribution but for *any* distribution.

Central limit theorem: Let $X_1, X_2 \cdots$, be independent random variables having the same distribution, with common finite expectation μ and common finite nonzero variance σ^2. Set $S_n = X_1 + X_2 + \cdots + X_n$ and $Z_n = \frac{S_n - E(S_n)}{\sqrt{Var(S_n)}}, n = 1, 2, \cdots$. Then for any $a < b$,

$$\lim_{n\to\infty} \text{Prob } (a < Z_n \leq b) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx \,. \qquad (1)$$

Remark : The hypotheses of the above result can be weakened considerably. In fact it is enough to assume that X_1, X_2, ... are independent with finite expectations and finite nonzero variances satisfying certain mild technical conditions; even with such an assumption the conclusion (1) of the above theorem holds; (see Feller or Parthasarathy in Suggested Reading). (In the sequel when we talk about CLT we mean the above theorem along with the generalization alluded to in this remark.)

Before making a few comments concerning CLT, let us point out that the right side of (1) gives a probability distribution,

viz. one can have a random variable Z such that

$$\text{Prob } (Z \text{ takes value in} (a,b]) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx\,. \qquad (2)$$

Such a probability distribution is called the *standard normal distribution* or the *standard Gaussian distribution* (in honour of Gauss who encountered it in 'the theory of errors'). The integrand in (2) is called the standard normal density. It can be shown that for such a Z,

$$\begin{aligned} E(Z) &= \int_{-\infty}^{\infty} x \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx = 0 \\ Var(Z) &= E(Z^2) - (E(Z))^2 \\ &= \int_{-\infty}^{\infty} x^2 \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx = 1\,. \qquad (3) \end{aligned}$$

(These can be proved using gamma integrals); so any random variable Z whose probability law is given by (2) is said to be normally distributed with mean 0 and variance 1. This suggests how one may define normal distribution with mean μ and variance σ^2; indeed this can be done by putting

$$f(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\{-\frac{1}{2\sigma^2}(x-\mu)^2\} \qquad (4)$$

for $-\infty < x < \infty$, and the corresponding distribution is denoted as the $N(\mu,\sigma^2)$ distribution. (Justify the notation! For any nonnegative function g such that $\int_{-\infty}^{\infty} g(x)\,dx = 1$, one can define a probability law corresponding to that. How?)

Unlike the De Moivre-Laplace theorem, in the CLT we do not specify the distribution of X_i. In fact the assumption on the distribution of X_i is basically qualitative and very general; however, the conclusion is quantitative! Thus the CLT implies the following interpretation.

Interpretation : Suppose the randomness in a phenomenon is due to the *cumulative* effects of randomness (or error in measurement) in a *large number* of parameters; further suppose that these parameters are taken to be *independent* of

Suppose the randomness in a phenomenon is due to the cumulative effects of randomness in a large number of parameters; suppose that these are taken to be independent of each other. In such a case the random phenomenon follows a normal (or Gaussian) distribution.

each other. In such a case the random phenomenon follows a normal (or Gaussian) distribution.

So, what began essentially as a computational aid has turned out to be a basic principle of nature, because as a well known text book in physics says "... the generality of the result accounts for the fact that so many phenomena in nature obey approximately a Gaussian distribution"; (see Reif in Suggested Reading).

The proof of CLT is quite involved. One method uses the Fourier transforms; an interested reader may refer to Feller or Parthasarathy (Suggested Reading).

To appreciate the significance of the normal distribution we present three situations where it plays a major role. But before that, let us mention that vector valued random variables can also be defined. For example, we say that (X, Y, Z) is a random vector with probability density function g if

$$\text{Prob } ((X,Y,Z) \text{ takes value in } A \times B \times C) = \int_{A \times B \times C} g(x,y,z)dxdydz \ , \qquad (5)$$

where the integral is a 3-dimensional integral. Note that X, Y, Z are one dimensional random variables; also X, Y, Z are said to be *independent* if

$$g(x,y,z) = g_1(x)g_2(y)g_3(z) \text{ for all } x, y, z \qquad (6)$$

where g_1, g_2, g_3 respectively are the probability density functions of X, Y and Z.

Statistics

An important objective of the discipline of Statistics is to be able to draw reasonable conclusions from *incomplete* information, like predicting the outcome of an election from an opinion poll, forecasting the future demand/prices based

on the past data, and so on. As statisticians are not blessed with clairvoyant powers (and as they are likely to be held accountable for their predictions), it is desirable to be able to ascertain the margin of error/uncertainty in their forecasts! In principle this is a slightly tricky situation for the following reason.

An important objective of the discipline of Statistics is to draw reasonable conclusions from incomplete information.

When one draws a conclusion about the entire population based on a sample, or predicts the future based on the past, etc., one resorts to *inductive logic*. Bulk of science is basically governed by inductive logic. If several repetitions of an experiment lead to the same observable pattern/products/conclusion, one *infers a cause-effect* relationship between the constituents. However, rules of *deductive logic* would not permit that; even if 10 million repetitions of an experiment obey the same pattern, there is no guarantee that the next trial would follow suit. As opposed to Science, Mathematics is firmly anchored in deductive logic. Now, if we want to indicate the degree of uncertainty in the predictions, such quantitative measures can be defined in a meaningful way only if there is a mathematical framework.

And the CLT provides a way out in many many situations. If the randomness/ uncertainty in the situation under consideration can be attributed to the cumulative effect of a large number of uncertain factors, then CLT says that a Gaussian model is justified. For example, uncertainty in price levels of foodgrains next year can be taken to be the cumulative effect of uncertainties in climatic conditions, political situation, prices of auxiliary commodities like fertilisers, etc. over the year; each of these factors in turn depends on a large number of other factors; (see P K Das in Suggested Reading concerning Gaussian assumption in weather prediction). In such a case the underlying phenomenon can be assumed to obey a Gaussian law with unknown mean μ and unknown variance σ^2; various procedures for estimating these parameters or for testing hypotheses concerning these parameters (along with the degree of uncertainty in the procedures) can be outlined using mathematical/deductive principles. A large chunk of the discipline of Statistics is based on the normality assump-

Bulk of science is basically governed by inductive logic.

tion; no wonder the bell shaped curve of the normal density adorns the cover of many books on statistics.

A word of caution is in order. There are situations when normality assumption would be untenable, as for example, if the hypotheses of CLT are far from reasonable to assume. Even in the context of De Moivre-Laplace theorem, if p is very close to 0 or 1, the so called 'Poisson approximation' is more suitable.

Nevertheless it will not be an exaggeration to say that CLT has provided a 'raison d'être' for Statistics!

Velocity of an Ideal Gas

Let $V = (V_1, V_2, V_3)$ denote the velocity of a particle of an *ideal gas*. (An ideal gas basically means a collection of a large number of particles in motion whose interaction is so weak that it may be disregarded; such an idealisation is meaningful when one considers 'dilute gas media', especially at low pressures and/or high temperatures, when the good old Boyle's law holds!) As there are a large number of particles, V may be considered a 3-dimensional random variable; for the sake of mathematical simplicity let us assume that V has a probability density function $f(v_1, v_2, v_3)$, with $(v_1, v_2, v_3) \in \mathbb{R}^3$.

We make the following qualitative assumptions:

- The three components of V i.e. V_1, V_2, V_3 are independent (one dimensional) random variables; let g_i denote the probability density function of $V_i, i = 1, 2, 3$.
- Velocity depends only on kinetic energy; that is, f is a function only of $v_1^2 + v_2^2 + v_3^2$.

Note that the above assumptions mean that

$$g_1(v_1)g_2(v_2)g_3(v_3) = f(v_1, v_2, v_3) = h(v_1^2 + v_2^2 + v_3^2) \qquad (7)$$

for some function h. From (7), by freezing some variables, it is easy to see that $g_2(v) = \left(\frac{g_2(v_0)}{g_1(v_0)}\right) g_1(v)$ for all $-\infty < v < \infty$, where v_0 is a fixed point such that $g_1(v_0) \neq 0$. As $\int_{-\infty}^{\infty} g_1(v)dv = \int_{-\infty}^{\infty} g_2(v)dv = 1$ (why?) we now get $g_1 \equiv g_2$. Similarly $g_1 \equiv g_3 = g$ (say). So (7) implies that

$$\log g(v_1) + \log g(v_2) + \log g(v_3) = \log h(v_1^2 + v_2^2 + v_3^2).$$

Differentiating the above with respect to v_1, v_2, v_3 separately we get

$$\frac{g'(v_1)}{v_1 g(v_1)} = \frac{g'(v_2)}{v_2 g(v_2)} = \frac{g'(v_3)}{v_3 g(v_3)} = \frac{2h'(r)}{h(r)} \qquad (8)$$

for any $(v_1, v_2, v_3) \neq 0$ (r denoting $v_1^2 + v_2^2 + v_3^2$). This is possible only if $[h'(r)/h(r)]$ is a constant. (Why?) Thus we get the differential equation

$$\frac{g'(v)}{g(v)} = kv, \qquad -\infty < v < \infty; \qquad (9)$$

(luckily one doesn't need to know anything about differential equations to solve (9)). Observe that (9) is just $\frac{d}{dv} \log g(v) = kv$ and hence $\log g(v) = kv^2 + c$; therefore $g(v) = Ce^{kv^2}$ for some constants C and k. As g should be nonnegative and integrable we should have $C > 0$ and $k < 0$. Writing $k = -\frac{1}{2\sigma^2}$ for some $\sigma > 0$, as $\int g(v)dv = 1$ we get

$$g(v) = \frac{1}{\sqrt{2\pi}\sigma} \exp\{-\frac{v^2}{2\sigma^2}\}, \quad -\infty < v < \infty$$

and hence

$$f(v_1, v_2, v_3) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^3 \exp\{-\frac{1}{2\sigma^2}(v_1^2 + v_2^2 + v_3^2)\}. \quad (10)$$

Thus each V_i has $N(0, \sigma^2)$ distribution and $V = (V_1, V_2, V_3)$ has a 3-dimensional normal distribution. (In physics literature the 3-dimensional distribution given by (10) is also called the *Maxwell distribution*, in honour of James Maxwell who first derived the velocity distribution. In the case of an

ideal monatomic gas, the parameter σ can be expressed in terms of absolute temperature, mass of a molecule of gas, Avogadro's number, etc; see A M Vasilyev in Suggested Reading).

Regarding the assumptions, the second hypothesis holds if the system is invariant under rotations. It seems that the first assumption is not tenable at velocities close to that of light.

Mathematically speaking, the above discussion implies: if V_1, V_2, V_3 are independent random variables such that their joint distribution is invariant under rotations, then each V_i has $N(0, \sigma^2)$ distribution. (The Gaussian distribution has many such interesting characterizations).

Brownian Motion

The last example highlighting the importance of Gaussian distribution concerns the phenomenon of Brownian motion. This phenomenon (as most readers would be aware) is a ceaseless chaotic movement encountered in colloidal solutions; (such a movement has been observed whenever microscopic particles are suspended in liquids/gases). For example, a pollen grain suspended in water executes such a motion and this can be observed through a microscope; (this was observed first by Robert Brown, a botanist, and hence the name). For a lucid, non-mathematical account of the role and importance of Brownian motion in physics, see Albert Einstein and Leopold Infeld in Suggested Reading.

A pollen grain is 'too big' compared to the water molecule, but not big enough to be sedimented at the bottom of the container; on the other hand the water molecules are too numerous. Since all the particles are in motion, the pollen grain is kicked around by the bombardment of water molecules resulting in Brownian motion. We now present a heuristic account of Brownian motion suppressing a lot of mathematical details. We assume that no external force is acting on the

system, the medium is homogeneous and that the motion of the Brownian particle (that is, pollen grain) is purely due to the fluctuations caused by the bombarding water molecules. Also for the sake of simplicity we consider only one coordinate of the Brownian particle, that is, one-dimensional Brownian motion. Further there are no boundaries and the particle can move anywhere on the real axis.

Let $X(t)$ denote the position of the Brownian particle at time t. As the motion is haphazard and unrelenting, the following assumptions can be taken to be reasonable and based on experimental observations:

- For each $t > 0$, $X(t)$ is a random variable; let $X(0) \equiv x$ be the initial position of the Brownian particle.
- $X(t)$ is continuous in t.
- For $0 \leq s \leq t$, the displacement $X(t) - X(s)$ during the time interval (s, t) is independent of the 'history' upto time s.

The first and the third assumptions imply that even after observing the Brownian particle for some time one cannot predict its future course; at best one can hope to make some probabilistic statements. The third hypothesis means that for $0 < t_1 < t_2 < \cdots < t_k$, the random variables $X(t_k) - X(t_{k-1}), X(t_{k-1}) - X(t_{k-2}), \cdots, X(t_1) - X(0)$ are independent. The second assumption, though very crucial, is just a physical reality concerning continuity of motion.

Now let $t > 0$ be fixed. For any positive integer n, as $X(0) = x$, note that

$$X(t) - x = X(t) - X(\frac{n-1}{n}t) + X(\frac{n-1}{n}t) - X(\frac{n-2}{n}t)$$

$$+ \cdots + X(\frac{1}{n}t) - X(0).$$

By the third assumption $X(t)$ is thus a sum of n independent random variables. Since this happens for any n, we have

$$X(t) - x = \lim_{n\to\infty} \sum_{k=1}^{n} X(\frac{k}{n}t) - X(\frac{k-1}{n}t); \qquad (11)$$

that is, $X(t)$ is a limit of sums of independent random variables. This rings a bell! However one has to be careful. Note that right side of (11) is *not* expressed as a limit of partial sums of a sequence of independent random variables; that is, summands occurring in (11) are different for different values of n. Fortunately, because of continuity in the time variable, it can be shown that the conclusion of CLT is still valid in this case; (intuitively, continuity keeps things under control and the behaviour of (11) is like the limit of partial sums of a sequence of independent random variables.) This is a far reaching generalization of CLT (see L Breiman in Suggested Reading). Thus $X(t)$ can be taken to have a normal distribution for each $t > 0$.

(Without the continuity assumption, the conclusion is false, as anyone who is familiar with Poisson distribution and Poisson processes can easily see.)

Having come this far, it is a bit tempting to go a little further. Note that a normal distribution is completely determined once the mean and variance are known. As the medium is homogeneous and there is no preferred direction (because of the absence of external forces), the particle is as likely to move to the left as to the right at any instant. As expectation denotes the 'mean' position and $X(0) \equiv x$, we can take $\mathrm{E}(X(t)) = x$ for all t. Next, the particle is more likely to wander farther from its mean position as time goes by. Since variance gives a measure of dispersion, it follows that $\mathrm{Var}(X(t))$ should be an increasing function of t. The simplest increasing function one can think of is $\mathrm{Var}(X(t)) = t$; this just means that we are choosing a convenient scale. So we may take that $X(t)$ has $N(x,t)$ distribution whose probability density function is

$$p(t,x,z) = \frac{1}{\sqrt{2\pi t}} \exp\left(-\frac{(z-x)^2}{2t}\right) \qquad (12)$$

for $t > 0, x, z \in \mathbb{R}$. Note that $\int_A p(t,x,z)dz$ denotes the probability that the Brownian particle is in A at time t if it has started from x at time 0. It is easily checked that p satisfies

$$\frac{\partial p}{\partial t}(t,x,z) = \frac{1}{2}\frac{\partial^2 p}{\partial z^2}(t,x,z) \qquad (13)$$

which is the well known *heat equation (or diffusion equation)*, one of the trinity in the theory of 'basic partial differential equations'. In fact p is called the *fundamental solution* of the heat equation as any solution of that equation can be got from p. This shows a connection between probability theory and the theory of partial differential equations; and what we have indicated is just the proverbial tip of the iceberg.

Suggested Reading

- L Breiman. *Probability*. Addison-Wesley, 1968.
- K R Parthasarathy. *Introduction to Probability and Measure*. MacMillan (India). New Delhi, 1977.
- B Gnedenko. *The Theory of Probability*. Mir Publishers. Moscow, 1978.
- A M Vasilyev. *An Introduction to Statistical Physics*. Mir Publishers, 1983.
- P Billingsley. *Probability and Measure*. John Wiley, 1984.
- W Feller. *Introduction to Probability Theory and its Applications*. Vol. 2. Wiley-Eastern, 1984.
- F Reif. *Fundamentals of Statistical and Thermal Physics*. McGraw Hill International Student Edition. p. 39, 1985.
- Albert Einstein and Leopold Infeld. *The Evolution of Physics*. Universal Book Stall. New Delhi. pp 58-62, 1995.
- P K Das. The Earth's Changing Climate. *Resonance*. Vol. 1. No. 3. pp 54-65, 1996.

Address for Correspondence
S Ramasubramanian
Statistics & Mathematics Unit
Indian Statistical Institute
RVCE Post
Bangalore 560 059, India.