The Normal Distribution

1. From Binomial to Normal

S Ramasubramanian

The normal (or Gaussian) distribution is derived as a convenient approximation for binomial probabilities for large n. An optimal way of choosing sample size in an opinion poll is indicated using the normal distribution.

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.

Introduction

In this article, the ubiquitous normal distribution is introduced as a convenient approximation for computing binomial probabilities for large values of n. Stirling's formula and DeMoivre-Laplace theorem establish that it is an appropriate approximation; these are proved in all their gory mathematical details later in this article. The pain of having to go through the details with 'paper and pen' the only way (known to the author) of learning a mathematical topic may be compensated by a sense of participation in the discovery! The reader may find the section on application to opinion poll amusing.

Need for Approximation

If a fair coin is tossed n times, then it is an easy exercise to check that

Probability of getting exactly k heads

$$= \binom{n}{k} \left(\frac{1}{2}\right)^k \left(\frac{1}{2}\right)^{n-k}, k = 0, 1, \quad , n , \qquad (1)$$

where $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ is the number of ways of choosing k objects from n objects.

Even fairly sophisticated calculators cannot handle products of more than 70 consecutive integers.

More generally let S_n denote the number of successes in n trials of a success-failure experiment; each trial can result only in success or failure and each trial is assumed to be independent of other trials. Suppose the probability of success in any individual trial is p where $0 < p < 1$. Again it is quite simple to show that

$$\text{Prob}(S_n \leq k) = \sum_{j=0}^{k} \binom{n}{j} p^j (1-p)^{n-j}, k = 0, 1, 2, \quad , n . \quad (2)$$

This is the well known *binomial distribution* with parameters n and p; (why is it called the binomial distribution?)

Computing (1) or (2) would involve calculating factorials or equivalently products of strings of consecutive integers; as anyone who has played around with a calculator will be aware, even fairly sophisticated calculators cannot handle products of more than 70 consecutive integers. It, therefore, becomes important to have a good approximation for (2) which can be easily computed even for large values of n and k.

Standardisation

The average number of successes in the above set up in np; that is, expectation of the random variable S_n is np; this is denoted $E(S_n) = np$. Thus, as $n \to \infty$ the average number of successes also goes to ∞. Since np can be easily computed, it is reasonable to ask how S_n fluctuates about its expectation for large n; that is, np can serve as a centering parameter for S_n.

However $S_n - np$ is not good enough. Consider Prob.$(S_n = k)$ as a function of k; this is the probability function (or discrete density function) of S_n. Similarly one can have the probability function of the 'centred' random variable $S_n - np$. Observe that the graph of the probability function of $S_n - np$ is merely that of S_n shifted by a distance of np. Mere shifting is unlikely to bring out any asymptotic regularity.

Thus, besides the centering parameter, one also needs a proper scaling factor. That is, one seeks a function g of n such that the probability law of $(S_n - np)/g(n)$ as $n \to \infty$,

will hopefully exhibit a pattern!

From the above it is clear that the constant function $g(n) \equiv c$ is not suitable. What about another possible 'natural' choice viz. $g(n) = n$? We will show that this also does not serve the purpose.

Define the random variable $X_i, i = 1, 2,$ by

$$X_i = \begin{cases} 1 & \text{if the } i\text{-th trial results in success} \\ 0 & \text{if the } i\text{-th trial results in failure.} \end{cases} \quad (3)$$

Clearly $P(X_i = 1) = p$, $P(X_i = 0) = 1 - p$, $i = 1, 2,$ and the $X_i's$ are independent random variables; each X_i is called a *Bernoulli* random variable with parameter p. Observe that $S_n = X_1 + X_2 + \cdots + X_n$; that is, S_n can be expressed as a sum of n independent random variables each having a Bernoulli distribution with parameter p. (There seems to be some method in our madness!)

It is easy to check that $E(X_i) = p$ for all i. So by the law of large numbers (see Karandikar's article in Suggested Reading), for any $\epsilon > 0$,

$$P\left(\left|\frac{S_n - np}{n}\right| > \epsilon\right) = P\left(\left|\frac{X_1 + \quad + X_n}{n} - p\right| > \epsilon\right) \to 0 \text{ as } n \to \infty. \quad (4)$$

Thus the function $g(n) = n$ grows much too fast compared to $S_n - np$ to be of much use.

Therefore we want g so that $\left|\frac{S_n - np}{g(n)}\right|$ neither collapses to a constant nor does it explode to ∞. If $E\left[\left(\frac{S_n - np}{g(n)}\right)^2\right]$ stays bounded and bounded away from 0, (that is, if there exist constants c_1, c_2 such that for all $n, 0 < c_1 \leq E\left[\left(\frac{S_n - np}{g(n)}\right)^2\right] \leq c_2 < \infty$), then one can expect $\frac{S_n - np}{g(n)}$ not to collapse or explode; (can you think of a heuristic explanation for this?). This suggests the choice $g(n) = \sqrt{np(1-p)}$ = standard deviation of S_n.

Put $Z_n = \frac{S_n - np}{\sqrt{np(1-p)}}$; clearly $E(Z_n) = 0$ and $\mathrm{Var}(Z_n) = 1$; Z_n may be called *standardised version* of S_n. We want to study the distribution of Z_n, as $n \to \infty$.

It is instructive to keep in mind that our choice of $g(n)$ above has had the benefit of hindsight, viz. that it is going to work! Needless to say, quite a bit of 'trial and error' must have gone in to the search for appropriate scale.

Stirling's Formula

We will now derive an estimate for $n!$ which will be used later to prove the required approximation theorem for the binomial probabilities. The approach taken here is as in Feller's book (see Suggested Reading). For this we estimate $\log(n!) = \log 1 + \log 2 + \cdots + \log n$.

As log is a monotone function note that for $k \geq 1$

$$\int_{k-1}^{k} \log x \, dx < \log k < \int_{k}^{k+1} \log x \, dx$$

Summing the above over $1 \leq k \leq n$ we get

$$\int_{0}^{n} \log x \, dx < \log(n!) < \int_{1}^{n+1} \log x \, dx$$

Since $\frac{d}{dx}(x \log x) = 1 + \log x$, we now have

$$(n \log n) - n < \log(n!) < ((n+1)\log(n+1)) - n \,. \qquad (5)$$

The two sided inequality (5) suggests comparing $\log(n!)$ with the arithmetic mean of the extreme members of (5) or with some quantity 'close' to the arithmetic mean.

Note that

$$\frac{(n \log n) + (n+1)\log(n+1)}{2} - (n + \frac{1}{2}) \log n \longrightarrow \frac{1}{2}.$$

So we consider comparing $\log(n!)$ with $(n + \frac{1}{2}) \log(n) - n$; (why is it a reasonable thing to do?).

Put

$$d_n = \log(n!) - (n + \frac{1}{2}) \log(n) + n \,. \qquad (6)$$

Using the expansion $\log\left(\frac{1+t}{1-t}\right) = 2\sum_{j=0}^{\infty} \frac{1}{(2j+1)} t^{(2j+1)}$, for $|t| < 1$, we now get

$$\begin{aligned} d_n - d_{n+1} &= (n+\frac{1}{2})\log\left(\frac{n+1}{n}\right) - 1 \\ &= (n+\frac{1}{2})\log\left(\frac{1+\frac{1}{2n+1}}{1-\frac{1}{2n+1}}\right) - 1 \\ &= \frac{1}{3(2n+1)^2} + \frac{1}{5(2n+1)^4} + \cdots \\ &< \frac{1}{3}\left[\frac{1}{(2n+1)^2} + \frac{1}{(2n+1)^4} + \quad \cdot\right] \\ &= \frac{1}{3}\frac{1}{((2n+1)^2-1)} = \frac{1}{12n(n+1)} \\ &= \frac{1}{12n} - \frac{1}{12(n+1)}\,. \end{aligned} \tag{7}$$

By the first three equalities in (7) we get $d_n > d_{n+1}$, whereas from the inequality of the extreme terms of (7) we get $d_n - \frac{1}{12n} < d_{n+1} - \frac{1}{12(n+1)}$. These imply that the sequence $\{d_n\}$ converges to a finite limit. (Why?)

Put $\log C_0 = \lim_n d_n$. Then by (6) we have $\frac{n!}{C_0 e^{-n} n^{n+\frac{1}{2}}} \longrightarrow 1$ as $n \to \infty$. Thus we have proved that there is a constant $C_0 > 0$ such that

$$\lim_{n\to\infty} \frac{n!}{C_0 e^{-n} n^{n+\frac{1}{2}}} = 1\,. \tag{8}$$

It will turn out later, (as a consequence of De Moivre-Laplace limit theorem), that $C_0 = \sqrt{2\pi}$; (8) is known as *Stirling's formula.*

De Moivre-Laplace Limit Theorem

Let a, b be arbitrary real numbers such that $a < b$; these are fixed for the following discussion. We want to study $\lim_{n\to\infty} P(a < Z_n \leq b)$.

Note that

$$P(a < Z_n \leq b) = \sum_{r \in A_n} \frac{n!}{r!(n-r)!} p^r q^{n-r} \tag{9}$$

where $q = 1 - p$, and

$$A_n = \{r : r \text{ integer}, \ 0 \leq r \leq n$$

$$\text{and } np + a\sqrt{npq} < r \leq np + b\sqrt{npq}.$$

Let $m_n = \min\{r : r \in A_n\}, \ell_n = \min\{n - r : r \in A_n\}$. Observe that, as a and b are fixed, we have

$$\lim_{n\to\infty} m_n = \infty, \quad \lim_{n\to\infty} \ell_n = \infty. \tag{10}$$

Let $\epsilon > 0$. By Stirling's formula (8) and (10) there is a positive integer $K_1(\epsilon)$ such that for all $n \geq K_1(\epsilon)$,

$$\left| \left\{ \frac{n!}{C_0 e^{-n} n^{n+\frac{1}{2}}} \frac{C_0 e^{-r} r^{r+\frac{1}{2}}}{r!} \frac{C_0 e^{-(n-r)}(n-r)^{n-r+\frac{1}{2}}}{(n-r)!} \right\} - 1 \right| < \epsilon$$

for all $r \in A_n$. Consequently for all $n \geq K_1(\epsilon)$ and for all $r \in A_n$,

$$\left| \left[\left(\frac{n!}{r!(n-r)!} p^r q^{n-r} \right) \Big/ \right. \right.$$

$$\left. \left. \left\{ \left(\frac{n}{C_0^2 r(n-r)} \right)^{\frac{1}{2}} \left(\frac{np}{r} \right)^r \left(\frac{nq}{n-r} \right)^{n-r} \right\} \right] - 1 \right| < \epsilon. \tag{11}$$

For $n = 1, 2, \quad$ and $r \in A_n$, write $\xi_{r,n} = \frac{r-np}{\sqrt{npq}}$; note that $\xi_{r,n} \in (a, b]$. Writing $\xi = \xi_{r,n}$ for typographical reasons it is easily seen that

$$\frac{n}{C_0^2 r(n-r)} = \frac{n}{C_0^2 (np + \xi\sqrt{npq})(nq - \xi\sqrt{npq})} \tag{12}$$

$$\left(\frac{r}{np} \right)^{-r} = \left(1 + \xi \sqrt{\frac{q}{np}} \right)^{-(np + \xi\sqrt{npq})} \tag{13}$$

$$\left(\frac{n-r}{np}\right)^{-(n-r)} = \left(1-\xi\sqrt{\frac{p}{nq}}\right)^{-(nq-\xi\sqrt{npq})} \qquad (14)$$

As $\xi = \xi_{r,n} \in (a,b]$ remains bounded, we see that

$$\lim_{n\to\infty} \frac{n}{C_0^2(np+\xi\sqrt{npq})(nq-\xi\sqrt{npq})}.C_0^2 npq = 1 \qquad (15)$$

Therefore by (11) (15), there exists a positive integer $K_2(\epsilon)$ such that for all $n \geq K_2(\epsilon)$ and all $r \in A_n$.

$$\left|\left[\left(\frac{n!}{r!(n-r)!}p^r q^{n-r}\right) / \left(\frac{1}{C_0\sqrt{npq}} f_n(\xi_{r,n})\right)\right] - 1\right| < \epsilon\,, \quad (16)$$

where

$$\begin{aligned} -\log f_n(\xi_{r,n}) &= (np+\xi\sqrt{npq})\log\left(1+\xi\sqrt{\frac{q}{np}}\right) \qquad (17)\\ &\quad +(nq-\xi\sqrt{npq})\log\left(1-\xi\sqrt{\frac{p}{nq}}\right) \end{aligned}$$

Once again as $\xi_{r,n}$ remains bounded, using the expansion of $\log(1+t)$ for $|t|<1$ we get for large n

$$-\log f_n(\xi_{r,n}) = \frac{1}{2}\xi_{r,n}^2 + \delta_n\,, \qquad (18)$$

where $|\delta_n| \leq K_3\frac{1}{\sqrt{n}}$ for some constant K_3 independent of n. Thus from (16) (18) we see that for $\epsilon > 0$ there is $K_4(\epsilon)$ such that for all $n \geq K_4(\epsilon)$ and all $r \in A_n$

$$\left|\left[\left(\frac{n!}{r!(n-r)!}p^r q^{n-r}\right)/\left(\frac{1}{C_0\sqrt{npq}}e^{-\frac{1}{2}\xi_{r,n}^2}\right)\right] - 1\right| < \epsilon\,. \qquad (19)$$

Now from (9) it is seen that

$$P(a < Z_n \leq b) = \sum_{r\in A_n} \frac{1}{C_0}\frac{1}{\sqrt{npq}}e^{-\frac{1}{2}\xi_{r,n}^2}$$

$$+\sum_{r\in A_n}\frac{1}{C_0}\frac{1}{\sqrt{npq}}e^{-\frac{1}{2}\xi_{r,n}^2}\left[\frac{b(n,r,p)}{\frac{1}{C_0}\frac{1}{\sqrt{npq}}e^{-\frac{1}{2}\xi_{r,n}^2}} - 1\right] \qquad (20)$$

where $b(n,r,p) = \frac{n!}{r!(n-r)!}p^r q^{n-r}$ Observe that $\{\xi_{r,n} : r \in A_n\}$ forms a partition of $(a,b]$ into subintervals of length

$\leq \frac{1}{\sqrt{npq}}$; (all except the two end subintervals have length $\frac{1}{\sqrt{npq}}$). Hence by the definition of Riemann integral

$$\lim_{n\to\infty} \sum_{r\in A_n} \frac{1}{\sqrt{npq}} \left(\frac{1}{C_0} e^{-\frac{1}{2}\xi_{r,n}^2}\right) = \int_a^b \frac{1}{C_0} e^{-\frac{1}{2}x^2} dx \qquad (21)$$

Since the integral on the r.h.s. of (21) is finite, by (19) we get that for all $n \geq K_4(\epsilon)$,

$$| \text{ second expression on the r.h.s. of (20) } | \leq \frac{(b-a)}{C_0}\epsilon . \qquad (22)$$

By (20)–(22) we now obtain

$$\lim_{n\to\infty} P(a < Z_n \leq b) = \int_a^b \frac{1}{C_0} e^{-\frac{1}{2}x^2} dx \qquad (23)$$

We still have to find the mysterious constant C_0. For this letting $a \to -\infty$, $b \to \infty$ we see that l.h.s. of (23) is unity. Hence

$$C_0 = \int_{-\infty}^{\infty} e^{-\frac{1}{2}x^2} dx \qquad (24)$$

Note that we can write

$$C_0^2 = \left(\int_{-\infty}^{\infty} e^{-\frac{1}{2}x^2} dx\right)\left(\int_{-\infty}^{\infty} e^{-\frac{1}{2}y^2} dy\right)$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-\frac{1}{2}(x^2+y^2)} dx\, dy .$$

To evaluate the double integral, using polar coordinates viz. $x = r\cos\theta, y = r\sin\theta$ (and not forgetting the Jacobian of the transformation) we get

$$C_0^2 = \int_0^{\infty}\int_0^{2\pi} re^{-\frac{1}{2}r^2} d\theta dr = 2\pi$$

Thus we have now completely proved the following important result.

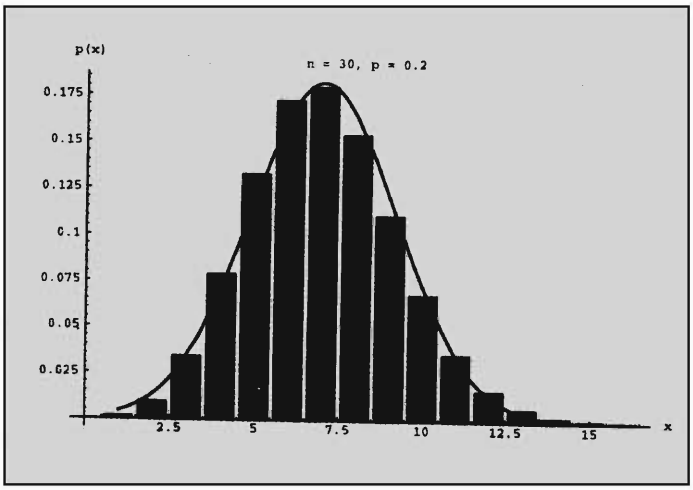

Figure 1 Normal approximation to the binomial.

Theorem : For any $a < b$,

$$\lim_{n\to\infty} \text{Prob } (a < Z_n \leq b) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx. \qquad (25)$$

The proof given above is basically adapted from K R Parthasarathy's book (see Suggested Reading). The above approximation can be proved to be quite good even for $n = 25$ when p is not too close to 0 or 1. Though there is no closed form expression for the r.h.s. of (25), it can be calculated to any desired accuracy using quadrature formulae like Simpson's rule/trapezoidal rule with a calculator. These are also extensively tabulated, given in the last few pages of practically every book on statistics. In *Figure 1* we have given the probability histograms of the Binomial distribution with parameters $n = 30$ and $p = 0.2$ and the approximating normal distribution.

How small can n be so that the estimated value differs from the true value by more than 0.05 only in less than 2% of the time this procedure is adopted?

Application to Opinion Poll

Suppose we want to estimate the unknown proportion p of an electorate voting for a particular party, say A. Common sense suggests that we take a 'representative sample' from the population and ascertain their preference; and the proportion in the sample preferring A can be taken to be an

Suggested Reading

- K R Parthasarathy. *Introduction to Probability and Measure.* Mac Millan (India). New Delhi, 1977.
- B Gnedenko. *The Theory of Probability.* Mir Publishers. Moscow, 1978.
- P Billingsley. *Probability and Measure.* John Wiley, 1984.
- W Feller. *An Introduction to Probability Theory and its Applications.* Vol. 1. Wiley-Eastern, 1993.
- R L Karandikar. On Randomness and Probability. *Resonance.* Vol. 1. No. 2, pp 55-68, 1996.
- Mohan Delampady and V R Padmawar. Sampling, Probability Models and Statistical Reasoning. *Resonance.* Vol. 1. No. 5, pp 49-58, 1996.

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

estimate for p. (Indeed, this procedure can be given a justification using the law of large numbers. This can also be shown to be the so called 'maximum likelihood estimator'; see Delampady & Padmawar in Suggested Reading). Now, De Moivre-Laplace Theorem enables us to do even better, viz. to choose the sample size n in an optimal manner.

The probability of a person chosen at random supporting A is the unknown parameter p. We assume that a person is chosen at random, her/his preference noted and she/he returned to the population. This procedure is repeated n times, and it is called *sampling with replacement.* Let S_n denote the number of persons preferring A among the sampled ones. A moment's reflection tells us that S_n has a binomial distribution with parameters n and p. And our 'common sense estimate' for p is $\frac{1}{n}S_n$. How small can the sample size n be so that we are still assured that

$$\text{Prob}\left(\left|\frac{S_n}{n} - p\right| > 0.05\right) \leq 0.02\,? \tag{26}$$

This can be interpreted as : How small can n be so that the estimated value differs from the true value by more than 0.05 only in less than 2% of the time this procedure is adopted? Such questions are very relevant because taking a representative sample can be tedious and expensive.

By De Moivre-Laplace limit theorem

$$\begin{aligned}\text{Prob}\left(\left|\frac{S_n}{n} - p\right| \leq 0.05\right) &= \text{Prob}\left(\frac{-0.05\sqrt{n}}{\sqrt{pq}} \leq Z_n \leq \frac{0.05\sqrt{n}}{\sqrt{pq}}\right) \\ &\cong \int_{-a}^{a} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx \end{aligned} \tag{27}$$

where $a = \frac{0.05\sqrt{n}}{\sqrt{pq}}$ and $\cong$ denotes approximate equality.

From tables of normal distribution, it can be found that the r.h.s. of (27) $\cong 0.98$ if $a = 2.33$. Since $p(1-p) \leq \frac{1}{4}$ for all $0 \leq p \leq 1$, it follows that (26) will be assured if $n \geq 543$. In other words a sample of size 543 would suffice for our purpose.