Chapter 5

Normal approximation to the Binomial

5.1 History

In 1733, Abraham de Moivre presented an approximation to the Binomial distribution. He later (de Moivre, 1756, page 242) appended the derivation of his approximation to the solution of a problem asking for the calculation of an expected value for a particular game. He posed the rhetorical question of how we might show that experimental proportions should be close to their expected values:

> *From this it follows, that if after taking a great number of Experiments, it should be perceived that the happenings and failings have been nearly in a certain proportion, such as of 2 to 1, it may safely be concluded that the Probabilities of happening or failing at any one time assigned will be very near in that proportion, and that the greater the number of Experiments has been, so much nearer the Truth will the conjectures be that are derived from them.*
>
> *But suppose it should be said, that notwithstanding the reasonableness of building Conjectures upon Observations, still considering the great Power of Chance, Events might at long run fall out in a different proportion from the real Bent which they have to happen one way or the other; and that supposing for Instance that an Event might as easily happen as not happen, whether after three thousand Experiments it may not be possible it should have happened two thousand times and failed a thousand; and that therefore the Odds against so great a variation from Equality should be assigned, whereby the Mind would be the better disposed in the Conclusions derived from the Experiments.*

> *In answer to this, I'll take the liberty to say, that this is the hardest Problem that can be proposed on the Subject of Chance, for which reason I have reserved it for the last, but I hope to be forgiven if my Solution is not fitted to the capacity of all Readers; however I shall derive from it some Conclusions that may be of use to every body: in order thereto, I shall here translate a Paper of mine which was printed November 12, 1733, and communicated to some Friends, but never yet made public, reserving to myself the right of enlarging my own Thoughts, as occasion shall require.*

De Moivre then stated and proved what is now known as the normal approximation to the Binomial distribution. The approximation itself has subsequently been generalized to give normal approximations for many other distributions. Nevertheless, de Moivre's elegant method of proof is still worth understanding. This Chapter will explain de Moivre's approximation, using modern notation.

A Method of approximating the Sum of the Terms of the Binomial $\overline{a+b}\backslash^n$ ***expanded into a Series, from whence are deduced some practical Rules to estimate the Degree of Assent which is to be given to Experiments.***

> *Altho' the Solution of problems of Chance often requires that several Terms of the Binomial* $\overline{a+b}\backslash^n$ *be added together, nevertheless in very high Powers the thing appears so laborious, and of so great difficulty, that few people have undertaken that Task; for besides James and Nicolas Bernouilli, two great Mathematicians, I know of no body that has attempted it; in which, tho' they have shown very great skill, and have the praise that is due to their Industry, yet some things were further required; for what they have done is not so much an Approximation as the determining very wide limits, within which they demonstrated that the Sum of the Terms was contained. Now the method ...*

5.2 Pictures of the binomial

Suppose X_n has a $\text{Bin}(n, p)$ distribution. That is,

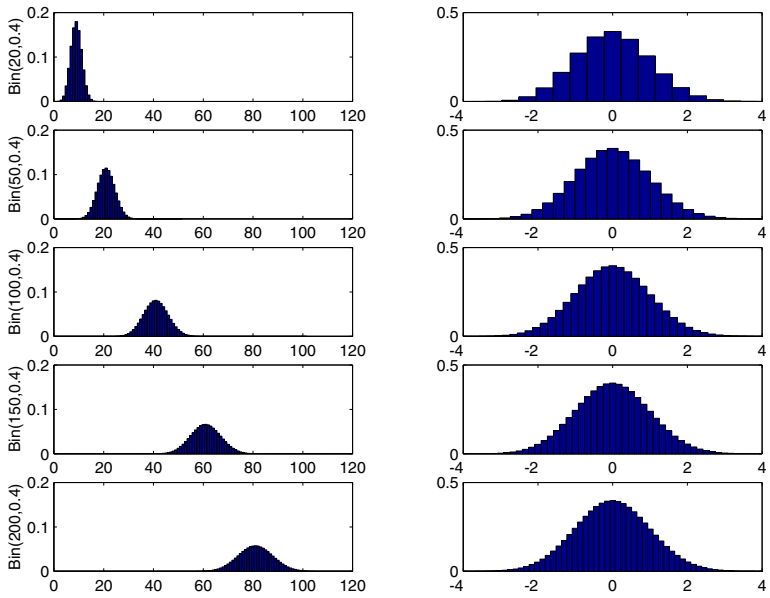
$$b_n(k) := \mathbb{P}\{X_n = k\} = \binom{n}{k} p^k q^{n-k} \qquad \text{for } k = 0, 1, \ldots, n, \text{ where } q = 1 - p,$$

Recall that we can think of X_n as a sum of independent random variables $Y_1 + \cdots + Y_n$ with $\mathbb{P}\{Y_i = 1\} = p$ and $\mathbb{P}\{Y_i = 0\} = q$. From this representation it follows that

$$\mathbb{E}X_n = \sum\nolimits_i \mathbb{E}Y_i = n\mathbb{E}Y_1 = np$$
$$\operatorname{var}(X_n) = \sum\nolimits_i \operatorname{var}(Y_i) = n \times \operatorname{var}(Y_1) = npq$$

Recall also that Tchebychev's inequality suggests the distribution should be clustered around np, with a spread determined by the standard deviation, $\sigma_n := \sqrt{npq}$.

What does the Binomial distribution look like? The plots in the next display, for the $\text{Bin}(n, 0.4)$ distribution with $n = 20, 50, 100, 150, 200$, are typical. Each plot on the left shows bars of height $b_n(k)$ and width 1, centered at k. The maxima occur near $n \times 0.4$ for each plot. As n increases, the spread also increases, reflecting the increase in the standard deviations $\sigma_n = \sqrt{npq}$ for $p = 0.4$. Each of the shaded regions on the left has area to one because $\sum_{k=0}^{n} b_n(k) = 1$ for each n.



The plots on the right show represent the distributions of the standardized random variables $Z_n = (X_n - np)/\sigma_n$. The location and scaling effects of the increasing expected values and standard deviations (with $p = 0.4$ and various n) are now removed. Each plot is shifted to bring the location of the maximum close to 0 and the horizontal scale is multiplied by a factor $1/\sigma_n$.

A bar of height $\sigma_n \times b_n(k)$ with width $1/\sigma_n$ is now centered at $(k-np)/\sigma_n$. The plots all have similar shapes. Each shaded region still has area 1.

5.3 De Moivre's argument

Notice how the standardized plots in the last picture settle down to a symmetric 'bell-shaped' curve. You can understand this effect by looking at the ratio of successive terms:

$$\begin{aligned} b_n(k)/b_n(k-1) & \\ &= \left(\frac{n!}{k!(n-k)!}p^k q^{n-k}\right) / \left(\frac{n!}{(k-1)!(n-k+1)!}p^{k-1}q^{n-k+1}\right) \\ &= (n-k+1)p/(kq) \qquad \text{for } k = 1, 2, \ldots, n. \end{aligned}$$

As a consequence, $b_n(k) \geq b_n(k-1)$ if and only if $(n-k+1)p \geq kq$, that is, iff $(n+1)p \geq k$. For fixed n, the probability $b_n(k)$ achieves its largest value at $k_{\max} = \lfloor (n+1)p \rfloor \approx np$. The probabilities $b_n(k)$ increase with k for $k \leq k_{\max}$ then decrease for $k > k_{\max}$. That explains why each plot on the left has a peak near np.

Now for the shape. At least for $k = k_{\max} + i$ near $k_{\max}$ we get a good approximation for the logarithm of the ratio of successive terms using the Taylor approximation: $\log(1+x) \approx x$ for x near 0. Indeed,

$$\begin{aligned} b(k_{\max}+i)/b(k_{\max}+i-1) &= \frac{(n-k_{\max}-i+1)p}{(k_{\max}+i)q} \\ &\approx \frac{(nq-i)p}{(np+i)q} \\ &= \frac{1-i/(nq)}{1+i/(np)} \qquad \text{after dividing through by } npq. \end{aligned}$$

The logarithm of the last ratio equals

$$\log\left(1-\frac{i}{nq}\right) - \log\left(1+\frac{i}{np}\right) \approx -\frac{i}{nq} - \frac{i}{np} = -\frac{i}{npq}.$$

By taking a product of successive ratios we get the ratio of the individual Binomial probabilities to their largest term. On a log scale the calculation

is even simpler. For example, if $m \geq 1$ and $k_{\max} + m \leq n$,

$$\begin{aligned}
&\log \frac{b(k_{\max}+m)}{b(k_{\max})} \\
&\quad = \log\left(\frac{b(k_{\max}+1)}{b(k_{\max})} \times \frac{b(k_{\max}+2)}{b(k_{\max}+1)} \times \cdots \times \frac{b(k_{\max}+m)}{b(k_{\max}+m-1)}\right) \\
&\quad = \log \frac{b(k_{\max}+1)}{b(k_{\max})} + \log \frac{b(k_{\max}+2)}{b(k_{\max}+1)} + \cdots + \log \frac{b(k_{\max}+m)}{b(k_{\max}+m-1)} \\
&\quad \approx \frac{-1-2-\cdots-m}{npq} \\
&\quad \approx -\tfrac{1}{2}\frac{m^2}{npq}.
\end{aligned}$$

The last line used the fact that $1+2+3+\cdots+m = \frac{1}{2}m(m+1) \approx \frac{1}{2}m^2$.

In summary,

$$\mathbb{P}\{X = k_{\max}+m\} \approx b(k_{\max}) \exp\left(-\frac{m^2}{2npq}\right) \qquad \text{for } m \text{ not too large.}$$

An analogous approximation holds for $0 \leq k_{\max}+m \leq k_{\max}$.

5.4 The largest binomial probability

Using the fact that the probabilities sum to 1, for $p = 1/2$ de Moivre was able to show that the $b(k_{\max})$ should decrease like $2/(B\sqrt{n})$, for a constant B that he was initially only able to express as an infinite sum. Referring to his calculation of the ratio of the maximum term in the expansion of $(1+1)^n$ to the sum, 2^n, he wrote (de Moivre, 1756, page 244)

> *When I first began that inquiry, I contented myself to determine at large the Value of B, which was done by the addition of some Terms of the above-written Series; but as I perceived that it converged but slowly, and seeing at the same time that what I had done answered my purpose tolerably well, I desisted from proceeding further till my worthy and learned Friend Mr. James Stirling, who had applied himself after me to that inquiry, found that the Quantity B did denote the Square-root of the Circumference of a Circle whose Radius is Unity, so that if that Circumference be called c, the Ratio of the middle Term to the Sum of all the Terms will be expressed by* $2\sqrt{n}c$.

In modern notation, the vital fact discovered by the learned Mr. James Stirling asserts that

$$n! \approx \sqrt{2\pi}\, n^{n+1/2} e^{-n} \qquad \text{for } n = 1, 2, \ldots$$

in the sense that the ratio of both sides tends to 1 (very rapidly) as n goes to infinity. See Feller (1968, pp52-53) for an elegant, modern derivation of the Stirling formula.

By Stirling's formula, for $k = k_{\max} \approx np$,

$$\begin{aligned} b_n(k) &= \frac{n!}{k!(n-k)!} p^k q^{n-k} \\ &\approx \frac{1}{\sqrt{2\pi}} \frac{n^{n+1/2}}{(np)^{np+1/2}(nq)^{nq+1/2}} p^{np} q^{nq} \\ &= \frac{1}{\sqrt{2\pi npq}}. \end{aligned}$$

De Moivre's approximation becomes

$$\mathbb{P}\{X_n = k_{\max} + m\} \approx \frac{1}{\sqrt{2\pi npq}} \exp\left(-\frac{m^2}{2npq}\right),$$

or, substituting np for $k_{\max}$ and writing k for $k_{\max} + m$,

$$\mathbb{P}\{X_n = k\} \approx \frac{1}{\sqrt{2\pi npq}} \exp\left(-\frac{(k-np)^2}{2npq}\right) = \frac{1}{\sigma_n \sqrt{2\pi}} \exp\left(-\frac{(k-np)^2}{2\sigma_n^2}\right).$$

That is, $\mathbb{P}\{X_n = k\}$ is approximately equal to the area under the smooth curve

$$f(x) = \frac{1}{\sigma_n \sqrt{2\pi}} \exp\left(-\frac{(x-np)^2}{2\sigma_n^2}\right),$$

for the interval $k - 1/2 \le x \le k + 1/2$. (The length of the interval is 1, so it does not appear in the previous display.)

Similarly, for each pair of integers with $0 \le a < b \le n$,

$$\mathbb{P}\{a \le X_n \le b\} = \sum_{k=a}^{b} b_n(k) \approx \sum_{k=a}^{b} \int_{k-1/2}^{k+1/2} f(x)\, dx = \int_{a-1/2}^{b+1/2} f(x)\, dx.$$

A change of variables, $y = (x - np)/\sigma_n$, simplifies the last integral to

$$\frac{1}{\sqrt{2\pi}} \int_{\alpha}^{\beta} e^{-y^2/2} dy \qquad \text{where } \alpha = \frac{a - np - 1/2}{\sigma_n} \text{ and } \beta = \frac{b - np + 1/2}{\sigma_n}.$$

Remark. It usually makes little difference to the approximation if we omit the $\pm 1/2$ terms from the definitions of α and β.

5.5 Normal approximations

How does one actually perform a normal approximation? Back in the olden days, I would have interpolated from a table of values for the function

$$\Phi(x) := \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-y^2/2} dy,$$

which was found in most statistics texts. For example, if X has a Bin$(100, 1/2)$ distribution,

$$\begin{aligned} \mathbb{P}\{45 \le X \le 55\} &\approx \Phi\left(\frac{55.5 - 50}{5}\right) - \Phi\left(\frac{44.5 - 50}{5}\right) \\ &\approx 0.8643 - 0.1356 = 0.7287 \end{aligned}$$

These days, I would just calculate in R:

```
> pnorm(55.5, mean = 50, sd = 5) - pnorm(44.5, mean = 50, sd = 5)
[1] 0.7286679
```

or use another very accurate, built-in approximation:

```
> pbinom(55,size = 100, prob = 0.5) - pbinom(44,size = 100, prob = 0.5)
[1] 0.728747
```

5.6 Continuous distributions

At this point, the integral in the definition of $\Phi(x)$ is merely a reflection of the Calculus trick of approximating a sum by an integral. Probabilists have taken a leap into abstraction by regarding Φ, or its derivative $\phi(y) := \exp(-y^2/2)/\sqrt{2\pi}$, as a way to define a probability distribution

<5.1> **Definition.** *A random variable Y is said to have a* ***continuous distribution*** *(on $\mathbb{R}$) with* ***density function*** *$f(\cdot)$ if*

$$\mathbb{P}\{a \le Y \le b\} = \int_a^b f(y)\, dy \qquad \text{for all intervals } [a, b] \subseteq \mathbb{R}.$$

Equivalently, for each subset A of the real line,

$$\mathbb{P}\{Y \in A\} = \int_A f(y)\, dy = \int_{-\infty}^{\infty} \mathbb{I}\{y \in A\} f(y)\, dy$$

Notice that f should be a nonnegative function, for otherwise it might get awkward when calculating $\mathbb{P}\{Y \in A\}$ for the set $A = \{y \in \mathbb{R} : f(y) < 0\}$:

$$0 \le \mathbb{P}\{Y \in A\} = \int_A f(y)\,dy \le 0.$$

Remark. By putting A equal to $\mathbb{R}$ we get

$$1 = \mathbb{P}\{-\infty < Y < +\infty\} = \int_{-\infty}^{\infty} f(y)\,dy$$

That is, the integral of a density function over the whole real line equals one.

I prefer to think of densities as being defined on the whole real line, with values outside the range of the random variable being handled by setting the density function equal to zero in appropriate places. If a range of integration is not indicated explicitly, it can then always be understood as $-\infty$ to ∞, with the zero density killing off unwanted contributions.

Distributions defined by densities have both similarities to and differences from the sort of distributions I have been considering up to this point in Stat 241/541. All the distributions before now were ***discrete***. They were described by a (countable) discrete set of possible values $\{x_i : i = 1, 2, \dots\}$ that could be taken by a random variable X and the probabilities with which X took those values:

$$\mathbb{P}\{X = x_i\} = p_i \qquad \text{for } i = 1, 2, \dots.$$

For any subset A of the real line

$$\mathbb{P}\{X \in A\} = \sum\nolimits_i \mathbb{I}_{\{x_i \in A\}} \mathbb{P}\{X = x_i\} = \sum\nolimits_i \mathbb{I}_{\{x_i \in A\}} p_i$$

Expectations, variances, and things like $\mathbb{E}g(X)$ for various functions g, could all be calculated by conditioning on the possible values for X.

For a random variable X with a continuous distribution defined by a density f, we have

$$\mathbb{P}\{X = x\} = \int_x^x f(y)\,dy = 0$$

for every $x \in \mathbb{R}$. We cannot hope to calculate a probability by adding up (an uncountable set of) zeros. Instead, as you will see in Chapter 7, we must pass to a limit and replace sums by integrals when a random variable X has a continuous distribution.

5.7 Appendix: The mysterious $\sqrt{2\pi}$

The $\sqrt{2\pi}$ appeared in de Moivre's approximation by way of Stirling's formula. It is slightly mysterious why it appears in that formula. The reason for both appearances is the fact that the constant

$$C := \int_{-\infty}^{\infty} \exp(-x^2/2)\, dx$$

is exactly equal to $\sqrt{2\pi}$, as I now explain.

Equivalently, the constant $C^2 = \iint \exp(-(x^2+y^2)/2)\, dx\, dy$ equal to 2π. (Here, and subsequently, the double integral runs over the whole plane.) We can evaluate this double integral by using a small Calculus trick.

Using the fact that

$$\int_0^{\infty} \mathbb{I}\{r \le z\} e^{-z}\, dz = e^{-r} \qquad \text{for } r > 0,$$

we may rewrite C^2 as a triple integral: replace r by $(x^2 + y^2)/2$, then substitute into the double integral to get

$$\begin{aligned} C^2 &= \iint \left(\int_0^{\infty} \mathbb{I}\{x^2+y^2 \le 2z\} e^{-z}\, dz \right) dx\, dy \\ &= \int_0^{\infty} \left(\iint \mathbb{I}\{x^2+y^2 \le 2z\}\, dx\, dy \right) e^{-z}\, dz. \end{aligned}$$

With the change in the order of integration, the double integral is now calculating the area of a circle centered at the origin and with radius $\sqrt{2z}$. The triple integral reduces to

$$\int_0^{\infty} \pi \left(\sqrt{2z}\right)^2 e^{-z}\, dz = \int_0^{\infty} \pi 2 z e^{-z}\, dz = 2\pi.$$

That is, $C = \sqrt{2\pi}$.

References

de Moivre, A. (1756). *The Doctrine of Chances: or, A Method of Calculating the Probabilities of Events in Play* (Third ed.). New York: Chelsea. Third edition (fuller, clearer, and more correct than the former), reprinted in 1967. First edition 1718.

Feller, W. (1968). *An Introduction to Probability Theory and Its Applications* (third ed.), Volume 1. New York: Wiley.