Threshold functions and the birthday paradox

DAVID BEVAN

In memory of
Norman Evans (1927–2020), who taught me A level maths,
and Peter Neumann (1940–2020), one of my undergraduate tutors.

Our goal is to illustrate the idea of a threshold function in the context of the birthday paradox. We do this by exploring the asymptotics of binomial coefficients.

To begin, we consider the limit definition of the exponential function. It is well known that $\lim_{n\to\infty}\left(1+\frac{x}{n}\right)^n = e^x$ for any real constant x. But what happens when x grows with n? For what functions $x = x(n)$ is it the case that $\left(1+\frac{x}{n}\right)^n$ behaves asymptotically like e^x? More generally, given a function x, what adjustment factor $A_n(x)$ is needed so that $\left(1+\frac{x}{n}\right)^n \sim A_n(x)e^x$, where we write $f(n) \sim g(n)$ to denote that $\lim_{n\to\infty} f(n)/g(n) = 1$?

For fixed x and n, by taking logs and using the Taylor expansion for $\ln(1+z)$, we have

$$\begin{aligned}\left(1+\frac{x}{n}\right)^n &= \exp\left[n\ln\left(1+\frac{x}{n}\right)\right]\\ &= \exp\left[n\left(\frac{x}{n}-\frac{x^2}{2n^2}+\frac{x^3}{3n^3}-\dots\right)\right], \qquad \text{if } 0 \leqslant x < n,\\ &= \exp\left(-\frac{x^2}{2n}+\frac{x^3}{3n^2}-\dots\right)e^x\\ &= e^{S_n(x)}e^x,\end{aligned}$$

where $S_n(x) = \sum_{r=1}^{\infty}(-1)^r t_{n,r}(x)$ and $t_{n,r}(x) = \frac{x^{r+1}}{(r+1)n^r}$.

Case 1. Suppose first that $x \ll \sqrt{n}$, where $f(n) \ll g(n)$ denotes that $\lim_{n\to\infty} f(n)/g(n) = 0$ (that is, "f grows more slowly than g"). Then,

$$\sum_{r=1}^{\infty} t_{n,r}(x) \;<\; \sum_{r=1}^{\infty}\frac{x^{r+1}}{n^r} \;=\; \frac{x^2}{n-x},$$

which tends to zero as n increases since $x \ll \sqrt{n}$. Thus, $\lim_{n\to\infty} S_n(x) = 0$ and $\left(1+\frac{x}{n}\right)^n \sim e^x$. So, if x grows more slowly than $\sqrt{n}$ there is no need for an adjustment factor. For example, $\left(1+\frac{c\ln n}{n}\right)^n \sim e^{c\ln n} = n^c$, for any constant $c > 0$.

Case 2. Now suppose that $x \sim c\sqrt{n}$ for some positive constant c. In this case, we have $t_{n,1}(x) = x^2/2n \sim c^2/2$, but

$$\sum_{r=2}^{\infty} t_{n,r}(x) \;<\; \sum_{r=2}^{\infty}\frac{x^{r+1}}{n^r} \;=\; \frac{x^3}{n(n-x)},$$

which tends to zero since $x \ll n^{2/3}$. So $\lim_{n\to\infty} S_n(x) = -c^2/2$, yielding an adjustment factor of $e^{-c^2/2}$:

$$\left(1+\frac{x}{n}\right)^n \;=\; \left(1+\frac{c}{\sqrt{n}}\right)^n \;\sim\; e^{-c^2/2}e^{c\sqrt{n}}.$$

Case 3. If $\sqrt{n} \ll x \ll n^{2/3}$, then the analysis is as in Case 2, yielding an adjustment factor of $A_n(x) = e^{-x^2/2n}$. For example, if $x = \sqrt{n \ln n}$, then

$$\left(1+\frac{x}{n}\right)^n \;=\; \left(1+\sqrt{\frac{\ln n}{n}}\right)^n \;\sim\; e^{-\ln\sqrt{n}}e^x \;=\; e^x/\sqrt{n}.$$

Case 4. Suppose $x \ll n^{3/4}$. Now, $\lim_{n\to\infty} \sum_{r=3}^{\infty} t_{n,r}(x) = 0$, but we need to include $t_{n,2}(x)$ in the adjustment factor, giving $A_n(x) = e^{-x^2/2n+x^3/3n^2}$. For example, if $x = n^{5/7}$, we have

$$\left(1+n^{-2/7}\right)^n \;\sim\; \exp\left(n^{5/7} - \tfrac{1}{2}n^{3/7} + \tfrac{1}{3}n^{1/7}\right).$$

General case. In general, each function of the form $n^{1-1/p}$ acts as a threshold beyond which the adjustment factor needs an additional term. Specifically, if x grows as fast as $n^{(p-1)/p}$ but slower than $n^{p/(p+1)}$, then $p-1$ terms are required:

$$\text{If } \; x \ll n^{p/(p+1)} \;\text{ then }\; A_n(x) \;=\; \exp\left(\sum_{r=1}^{p-1}(-1)^r\frac{x^{r+1}}{(r+1)n^r}\right).$$

Now let's apply this to the asymptotics of the binomial coefficient $\binom{n}{k}$ when k grows with n. We exploit Stirling's approximation for the factorial, $n! \sim \sqrt{2\pi n}\, n^n e^{-n}$.

If $k \ll n$, then

$$\begin{aligned}
\binom{n}{k} \;&\sim\; \frac{1}{k!}\sqrt{\frac{n}{n-k}}\,\frac{n^n}{(n-k)^{n-k}}\,\frac{e^{-n}}{e^{-(n-k)}} \\
&\sim\; \frac{n^k}{k!}\,\frac{n^{n-k}}{(n-k)^{n-k}}e^{-k}, \qquad\qquad \text{since } \lim_{n\to\infty} n/(n-k) = 1, \\
&=\; \frac{n^k}{k!}\left(1+\frac{k}{n-k}\right)^{n-k}e^{-k} \\
&\sim\; A_{n-k}(k)\frac{n^k}{k!}.
\end{aligned}$$

Let $B_n(k) = A_{n-k}(k)$ denote this binomial adjustment factor. To establish an explicit expression for $B_n(k)$, note first that

$$t_{n-k,r}(k) \;=\; \left(1-\frac{k}{n}\right)^{-r} t_{n,r}(k) \;=\; \sum_{j=0}^{\infty}\binom{r-1+j}{r-1}\frac{k^{r+1+j}}{(r+1)n^{r+j}}.$$

Given that $S_{n-k}(k) = \sum_{r=1}^{\infty} (-1)^r t_{n-k,r}(k)$, the coefficient of k^{t+1}/n^t in the expansion for $S_{n-k}(k)$ is given by

$$\sum_{s=1}^{t} \frac{(-1)^s}{s+1} \binom{t-1}{s-1} \;=\; -\frac{1}{t(t+1)}.$$

This identity is equivalent to $\sum_{s=1}^{t} (-1)^s s \binom{t+1}{s+1} = -1$, which can be established by differentiating the binomial expansion of $(1+x)^{t+1}$ and setting $x = -1$.

Thus, the binomial adjustment factor behaves as follows:

$$\text{If } \; k \ll n^{p/(p+1)} \; \text{ then } \; B_n(k) \;=\; \exp\left(-\sum_{t=1}^{p-1} \frac{k^{t+1}}{t(t+1)n^t} \right).$$

So if $k \ll \sqrt{n}$, then $\binom{n}{k} \sim n^k/k!$. And as another illustration,

$$\binom{m^2}{m} \;\sim\; e^{-1/2} \, \frac{m^{2m}}{m!} \;\sim\; \frac{e^m m^m}{\sqrt{2\pi e m}}.$$

We've now laid sufficient groundwork to turn to our application. Suppose we have n boxes and m balls, and we throw the balls into the boxes at random. What is the probability $p(n,m)$ that two balls end up in the same box? When $n = 365$, this is the *birthday paradox* in another guise: If $m \geqslant 23$, then $p(n,m) > 0.5$, so in a random group of 23 or more people, it is more likely than not that some pair of them will have the same birthday.

What we intend to explore here is the manner in which $p(n,m)$ increases from near zero to near one when n is large. It turns out that we've done most of the necessary work already. There are a total of n^m ways of placing the m balls in the n boxes, and $\binom{n}{m} m!$ ways of doing so without there being two balls in the same box. So, if we let $q(n,m) = 1 - p(n,m)$ be the probability that each of the m balls is in a distinct box, then

$$q(n,m) \;=\; \binom{n}{m} \frac{m!}{n^m} \;\sim\; B_n(m).$$

Now, we have established that

- if $m \ll \sqrt{n}$, then $B_n(m) = 1$,
- if $m \sim c\sqrt{n}$, then $B_n(m) = e^{-c^2/2}$, and
- if $m \gg \sqrt{n}$, then $B_n(m) \leqslant e^{-m^2/2n}$, which tends to zero as n increases.

Hence,

$$p(n,m) \;\sim\; \begin{cases} 0 & \text{if } m \ll \sqrt{n}, \\ 1 - e^{-c^2/2} & \text{if } m \sim c\sqrt{n}, \\ 1 & \text{if } m \gg \sqrt{n}. \end{cases}$$

So we observe that the property of there being a box containing more than one ball becomes likely rather abruptly when m is of the order of $\sqrt{n}$. If m grows more slowly than $\sqrt{n}$ then the property holds asymptotically almost never, whereas if m grows faster than $\sqrt{n}$ then it holds asymptotically almost surely. We say that $\sqrt{n}$ is a *threshold function* for this property, where $m^* = m^*(n)$ is a threshold function for a property that holds with probability $p(n, m)$ if the following condition is satisfied:

$$\lim_{n\to\infty} p(n,m) \;=\; \begin{cases} 0 & \text{if } m \ll m^*, \\ 1 & \text{if } m \gg m^*. \end{cases}$$

This is a common phenomenon. There is a very general theory that guarantees the existence of threshold functions (and thus of the associated abrupt changes in the probability that a property holds) for many properties in a variety of contexts. In (models of) physical, chemical and biological systems, threshold functions correspond to the occurrence of *phase transitions*, such as those between ice and liquid water, and between liquid water and steam.

In our balls-in-boxes model, another important property having a threshold function is that of there being a ball in every box. This is the *coupon collector problem* in another form: If there are n distinct coupons, how many do you need to collect before you have a complete set? The threshold for this property (which requires rather more sophisticated techniques to establish) occurs at $n \ln n$. Moreover, it is a *sharp threshold* in the sense that, for any $\varepsilon > 0$, the property holds asymptotically almost never if $m \leqslant (1-\varepsilon) n \ln n$ and asymptotically almost surely if $m \geqslant (1+\varepsilon) n \ln n$, thus satisfying a stronger condition than that required for a threshold in general.

We conclude with some suggested further reading. For a presentation of threshold functions in the context of random graphs (the arena in which they have been studied the most), the introductory textbook [1] is recommended. Another area in which threshold phenomena are observed is in shuffling a pack of cards. For example, if a pack of n cards is riffle-shuffled significantly more than $s_n = \frac{3}{2} \log_2 n$ times then the probability of the cards being in any particular order is close to $1/n!$, whereas shuffling significantly fewer than s_n times results in a distribution that is far from uniform. Formal definitions and an overview of the topic can be found in [2]. More generally, the short monograph [3] provides a relatively gentle introduction to the kind of asymptotic analysis undertaken above.

References

[1] Alan Frieze and Michał Karoński. *Introduction to Random Graphs*. Cambridge University Press, 2015.

[2] Persi Diaconis. Mathematical developments from the analysis of riffle shuffling. In Alexander Ivanov, Martin Liebeck, and Jan Saxl, *Groups, Combinatorics and Geometry*. World Scientific Publishing, 2003.

[3] Joel Spencer and Laura Florescu. *Asymptopia*. American Mathematical Society, 2014.

DAVID BEVAN
Department of Mathematics and Statistics,
University of Strathclyde,
26 Richmond Street,
Glasgow G1 1XH
e-mail: *david.bevan@strath.ac.uk*