1 Introduction: within and beyond the normal approximation

1a Mathematical prelude

Tossing n fair coins we get a random variable S_n distributed binomially,

$$\mathbb{P}\big(S_n = k\big) = 2^{-n}\binom{n}{k} = \frac{n!}{2^n k!(n-k)!} \quad \text{for } k = 0, 1, \dots, n\,;$$

it is asymptotically normal: for all $x \in \mathbb{R}$,

$$\mathbb{P}\Big(\frac{2S_n - n}{\sqrt{n}} \ge x\Big) \to \underbrace{\int_x^\infty \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-u^2/2}\,\mathrm{d}u}_{1-\Phi(x)} \quad \text{as } n \to \infty\,;$$

and on the other hand,

$$1 - \Phi(x) = \frac{1}{x} \cdot \underbrace{\frac{1}{\sqrt{2\pi}} \mathrm{e}^{-x^2/2}}_{\varphi(x)=\Phi'(x)} \cdot \big(1 + o(1)\big) \quad \text{as } n \to \infty\,.$$

Does it mean that $\mathbb{P}\big(\frac{2S_n-n}{\sqrt{n}} \ge x\big) \approx \frac{1}{x}\varphi(x)$ for large n and x? Yes and no. "Yes" if "$\approx$" means a small absolute error; but this is trivial: both sides are ≈ 0. "No" if it means a small relative error; indeed, for $x > \sqrt{n}$ the binomial probability is 0, while its normal approximation is not. Well, what happens for $x = \sqrt{n}$? The binomial probability is 2^{-n}, while its normal approximation is roughly $\mathrm{e}^{-n/2}$; quite bad: $\frac{1}{2} \ne \frac{1}{\sqrt{\mathrm{e}}}$.

The Stirling formula

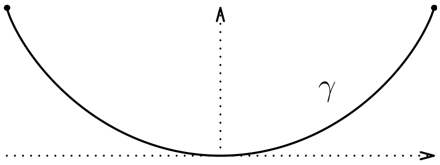
$$n! = n^n \mathrm{e}^{-n} \sqrt{2\pi n} \bigg(1 + \mathcal{O}\Big(\frac{1}{n}\Big)\bigg)$$

leads to

$$\mathbb{P}\big(S_n = k\big) = \frac{2}{\sqrt{n}} \frac{1}{\sqrt{1-a^2}} \frac{1}{2\pi} \mathrm{e}^{-n\gamma(a)} \bigg(1 + \mathcal{O}\Big(\frac{1}{n(1-|a|)}\Big)\bigg) \quad \text{where } a = \frac{2k-n}{n}$$

for all n and $k = 0, 1, \dots, n$; here

$$\gamma(a) = \frac{1}{2}(1+a)\ln(1+a) + \frac{1}{2}(1-a)\ln(1-a) \quad \text{for } a \in (-1,1)\,, \tag{1a1}$$
$$\gamma(-1) = \gamma(+1) = \ln 2\,.$$



Taking into account that

$$\frac{\mathbb{P}\big(S_n = k+1\big)}{\mathbb{P}\big(S_n = k\big)} = \frac{n-k}{k+1} \le \frac{1-a}{1+a}\,, \quad \frac{\mathbb{P}\big(S_n = k+2\big)}{\mathbb{P}\big(S_n = k\big)} \le \Big(\frac{1-a}{1+a}\Big)^2, \dots$$

we get for $k > n/2$

$$1 \le \frac{\mathbb{P}\big(S_n \ge k\big)}{\mathbb{P}\big(S_n = k\big)} \le \frac{1+a}{2a}\,,$$

thus, $\mathrm{e}^{-n\gamma(a)}$ as an approximation to $\mathbb{P}\big(\frac{2S_n-n}{n} \ge a\big)$ is rather crude, but still much better than the normal approximation in the limit $n \to \infty$, $a = \text{const} > 0$. Some numeric data for $n = 50$:

k	26	30	40	44	46	48
P	0.444	0.101	$1.19 \cdot 10^{-5}$	$1.62 \cdot 10^{-8}$	$2.23 \cdot 10^{-10}$	$1.13 \cdot 10^{-12}$
MD$/P$	0.9998	1.0022	1.72	5.2	15	87
LD$/P$	2.2	3.6	5.5	5.1	4.5	3.5

Here $P = \mathbb{P}\big(S_n \ge k\big)$, $\mathrm{MD} = 1 - \Phi\big(\frac{2k-1-n}{\sqrt{n}}\big)$, $\mathrm{LD} = \exp\big(-n\gamma(\frac{2k-n}{n})\big)$.

We see that the normal approximation is better for $k \le 40$ ("moderate deviations") and worse for $k \ge 46$ ("large deviations").[1]

Much better approximations are available, the so-called strong moderate deviations and strong large deviations:[2]

$$\mathrm{sMD} = \mathrm{MD} \cdot \exp\Big(\frac{x^2}{2} - n\gamma\Big(\frac{x}{\sqrt{n}}\Big)\Big) \quad \text{where } x = \frac{2k-1-n}{\sqrt{n}}\,;$$

[1] There is no "official" definition of "moderate". In the context of sums of i.i.d. random variables (with exponential moments) "moderate" means $a = o(1)$. For more general thoughts, see: Inglot et al. 1992, Ann. Prob. **20**:2, 987–1003.

[2] The word "strong" is overloaded. Here I use it, following Chaganty and Sethuraman 1993, Ann. Prob. **21**:3, 1671–1690. But sometimes it means that convergence of distributions results from convergence of random variables (as in: Inglot et al. 1992).

$$\mathrm{sLD} = \mathrm{LD} \cdot \frac{1}{\sqrt{2\pi n}} \frac{1}{a} \sqrt{\frac{1+a}{1-a}} \quad \text{where } a = \frac{2k-n}{n} \,.$$

Now, for $n = 50$ again,

k	26	30	40	44	46	48
P	0.444	0.101	$1.19 \cdot 10^{-5}$	$1.62 \cdot 10^{-8}$	$2.23 \cdot 10^{-10}$	$1.13 \cdot 10^{-12}$
sMD/P	0.9998	0.9978	0.9942	0.9916	0.9887	0.9815
sLD/P	3.2	1.25	1.029	1.022	1.025	1.044

and for $n = 1000$,

k	501	511	561	600	800	950
P	0.487	0.253	$6.39 \cdot 10^{-5}$	$1.36 \cdot 10^{-10}$	$8.23 \cdot 10^{-86}$	$9.32 \cdot 10^{-217}$
sMD/P	0.999 998	0.999 945	0.999 795	0.999 765	0.999 660	0.999 049
sLD/P	12.9	1.82	1.053	1.019	1.0015	1.0018

A wonder: sMD looks better than sLD in all cases.[1]

Also a wonder: we can compute easily such probability as $9.32 \cdot 10^{-217}$.

However, what is it really good for? Does it matter that $\mathbb{P}\big(S_{1000} \ge 950\big)$ is $9.32 \cdot 10^{-217}$ rather than $9.35 \cdot 10^{-217}$? Moreover, does it matter that it is not 0? Tossing 1000 fair coins we may be pretty sure that "heads" will not appear 950 times. Not even once in any feasible number of trials.

Is it reasonable to say that, for all practical purposes,

(a) $9.32 \cdot 10^{-217} \approx 10^{-217}$?

(b) $9.32 \cdot 10^{-217} \approx 0$?

My answers: (b) sometimes it is, but not always; (a) I am not sure; maybe, always.

The reason is related to statistical physics.

1b Physical prelude

To understand why rare events are important at all, one only has to think of a lottery to be convinced that rare events (such as hitting the jackpot) can have an enormous impact.

Amir Dembo and Ofer Zeitouni[2]

The numbers that arise in statistical mechanics can defeat your calculator. A googol is 10^{100} *(one with a hundred zeros after it). A googolplex is* 10^{googol}.

James P. Sethna[3]

[1] Really, I do not know, why. A feature of the (symmetric) binomial case? Or a manifestation of a more general phenomenon?

[2] See page 1 in the book "Large deviations techniques and applications", Jones and Bartlett Publ., 1993.

[3] See page 54 in the book "Statistical mechanics: entropy, order parameters, and complexity", Oxford, 2006.

Small probabilities, such as 10^{-6}, are important for lotteries, reliability etc., which cannot be said about much smaller probabilities, such as $10^{-1\,000\,000\,000\,000\,000\,000\,000}$. However, these monsters do appear in statistical physics (as e^{-cn} where the number of particles like $n = 10^{20}$ is quite usual).

A physical question

A system of n so-called spin-1/2 particles is described by the configuration space $\{-1,1\}^n$. Each configuration $(x_1,\dots,x_n) \in \{-1,1\}^n$ has its energy[1]

$$H_n(x_1,\dots,x_n) = nf\Big(\frac{x_1+\dots+x_n}{n}\Big)\,,$$

where $f : [-1,1] \to \mathbb{R}$ is a given smooth function (not depending on n). If the system is in thermal equilibrium with a heat bath at temperature T, then each configuration $(x_1,\dots,x_n)$ appears with the probability

$$\mathrm{const}_n \cdot \exp\bigg(-\frac{1}{k_{\mathrm{B}}T} H_n(x_1,\dots,x_n) \bigg)\,,$$

where k_{B} $(= 1.38 \cdot 10^{-23} \mathrm{J}/K)$ is the so-called Boltzmann constant. For large n, up to small fluctuations, the energy per particle $f(\frac{x_1+\dots+x_n}{n})$ is a function of the temperature. Find this function.

A solution

The distribution P_β of the number k of spins $+1$, corresponding to the so-called inverse temperature $\beta = \frac{1}{k_{\mathrm{B}}T}$, is

$$P_\beta(k) = P_0(k) \cdot \mathrm{const}_{\beta,n} \cdot \exp\bigg(-\beta n f\Big(\frac{2k-n}{n}\Big) \bigg)\,,$$

where P_0 is the "fair coin" binomial distribution (treated in Sect. 1a). We note that

$$P_0(k) = \exp\bigg(-n\Big(\gamma\Big(\frac{2k-n}{n}\Big) + o(1)\Big) \bigg)$$

[1] Assuming that the spins interact only with the same magnetic field $g((x_1+\dots+x_n)/n)$ that depends on the mean field $(x_1+\dots+x_n)/n$ via a function g describing (generally, nonlinear) magnetic properties of the environment. Thus, $f(s) = sg(s)$. See also Sect. 9 in: R.S. Ellis, "The theory of large deviations and applications to statistical mechanics", 2006, http://www.math.umass.edu/∼rsellis/pdf-files/Dresden-lectures.pdf; and Sect. 7.3.2 in: D. Yoshioka, "Statistical physics", Springer, 2007.

where $o(1)$ (as $n \to \infty$) is uniform over all k such that $|\frac{2k-n}{n}|$ is bounded away from 1.[1] Thus,

$$P_\beta(k) = \mathrm{const}_{\beta,n} \cdot \exp\biggl(-n\biggl(\gamma\Bigl(\frac{2k-n}{n}\Bigr) + \beta f\Bigl(\frac{2k-n}{n}\Bigr) + o(1) \biggr)\biggr) .$$

Assuming that the function $\gamma + \beta f$ has a single minimum $a_\beta \in (-1, 1)$ we see that P_β concentrates (for large n) on k such that $\frac{2k-n}{n} \approx a_\beta$. The energy per particle is therefore $f(a_\beta) + o(1)$.

Consider, for example, the simple case $f(a) = a$ (an external magnetic field only). We have $(\gamma + \beta f)'(a_\beta) = 0$, that is, $\gamma'(a_\beta) = -\beta$; generally $\gamma'(a) = \frac{1}{2} \ln \frac{1+a}{1-a}$; thus, $\frac{1+a_\beta}{1-a_\beta} = \mathrm{e}^{-2\beta}$;

$$a_\beta = -\frac{\mathrm{e}^\beta - \mathrm{e}^{-\beta}}{\mathrm{e}^\beta + \mathrm{e}^{-\beta}} = -\tanh\beta .$$

We see that $a_\beta \to -1$ as $\beta \to \infty$, and no wonder; at low temperature the energy is roughly minimal.

Note that P_β is concentrated on a set (of k) of very small probability P_0; indeed, exponentially small (in n). Taking into account that $n = 10^{20}$ is usual, we observe the probability about $\exp(-10^{20})$. Surely a number that can defeat a calculator!

Why does such a tiny probability matter? Because of the interplay between different probability measures related via exponentially small or large numbers. This is why we cannot replace a small probability with zero. On the other hand, a rough approximation, of the form $\exp\bigl(-n\bigl(\gamma(\dots) + o(1)\bigr)\bigr)$, is all we need. It means that, for instance, 10^{-217} is a reasonably good approximation for $9.32 \cdot 10^{-217}$, since their *logarithms* are *relatively* close. Likewise, $\exp(-10^{20})$ is a good approximation for $\exp(10^{15}) \exp(-10^{20})$ in this framework.

What about the distinction between $9.32 \cdot 10^{-217}$ and $9.35 \cdot 10^{-217}$? Can it matter in another framework? In principle, why not; but I did not face such situations.

[1] Still, a neighborhood of ±1 does not harm; for now I do not explain, why.