Applications of Stochastics — Simulation projects

Gábor Pete
http://www.math.bme.hu/~gabor

November 22, 2020

Most problems are best for pairs of students, but some are also good for triples (e.g., No 11 probably requires more work than average), or solos (e.g., No 13 and 14 are quite easy).

1. **Aldous' theorem.** Show (by simulations) that the vector of the two largest cluster sizes $C_1(n), C_2(n)$ in the critical Erdős-Rényi graph $G(n, 1/n)$, scaled by $n^{2/3}$, converges in distribution to the vector of the two longest excursions of a Brownian motion with parabolic drift, $B_t - t^2/2$, away from its running minimum (see PGG Theorem 12.23). Do not get frightened by Brownian motion: $(B_t)_{t\geq 0}$ should be just simulated as the limit of $X_{nt}/\sqrt{n}$ as $n \to \infty$, where X_i is simple symmetric random walk on $\mathbb{Z}$.

2. **Persistence of disconnectedness.** Recall that there is a sharp phase transition at $p = p_n = \frac{\ln n}{n}$ for the connectedness of the Erdős-Rényi graph $G(n, p)$.
   - **(a)** Estimate the probability of connectedness at p_n via simulations.
   - **(b)** What is the probability of being disconnected at $p_n(t) := p_n + \frac{t}{n}$, in the $n \to \infty$ limit? For instance, how does it behave as $t \to \infty$? Note that you can get an explicit mathematical guess by looking at the expected number of isolated vertices, which is $\sim e^{-t}$, for large n.
   - **(c)** Now, starting from a configuration at p_n, consider the dynamics where, at each step, a uniform random edge of K_n is chosen and resampled: independently of whether it was present or not, let it be present with probability p_n. Fixing a large $t > 0$, what is the probability that this dynamic random graph is disconnected all along the first $tn/2$ steps? Note (via a math argument) that this probability is at least as large as the previous off-critical probability, but the question is if it is much larger; say, only subexponentially small in t. (I do not know the answer.)

3. **Noise sensitivity in the Minimal Spanning Tree.** Assign to each edge e of the complete graph K_n an independent $U_e \sim \mathsf{Unif}[0, 1]$ label, and let MST be the spanning tree T that minimizes the total weight $\sum_{e\in T} U_e$. Let us denote this minimal total weight by $\mathsf{W}(\{U_e\})$, and the diameter of MST (in terms of the graph metric, not in terms of the labels) by $\mathsf{diam}(\{U_e\})$. Recall that MST can be effectively sampled by Kruskal's or similar greedy algorithms.
   - **(a)** Plot the distributions of $\mathsf{W}(\{U_e\})$ and $\mathsf{diam}(\{U_e\})$ for several values of n. How do the means and standard deviations scale with n?
   - **(b)** Now introduce a small noise to the labels: fix a small $\epsilon > 0$, and for each $e \in E(K_n)$, let $\tilde{U}_e$ be equal to U_e with probability $1 - \epsilon$, and an independent $\mathsf{Unif}[0, 1]$ variable with probability ϵ. How do the correlations $\mathrm{Corr}\big(\mathsf{W}(\{U_e\}), \mathsf{W}(\{\tilde{U}_e\})\big)$ and $\mathrm{Corr}\big(\mathsf{diam}(\{U_e\}), \mathsf{diam}(\{\tilde{U}_e\})\big)$ behave for fixed $\epsilon > 0$ as $n \to \infty$? I expect that the first remains close to 1, while the second goes to 0; i.e., the macroscopic geometry of the tree is noise sensitive, but the total weight is not.

4. **Virus infection on an expander graph.** Take a uniformly random 4-regular graph on n vertices (there are algorithms and packages doing that). The infection starts spreading from a single vertex (the 0th generation), in the following way: with probability $1 - p$ it infects a uniformly random neighbour,

while with probability p it infects two independently chosen uniform random neighbours (possibly the same one). Each of the newly infected neighbours (the 1st generation) infects 1 or 2 of its own neighbours, independently, with the same distribution as before (and possibly choosing already infected neighbours). Then this 2nd generation gives rise to the 3rd generation, and so on. (One may call this a branching random walk.) Stop the process when the set of vertices that have ever got infected reaches $\lfloor un \rfloor$, for some fixed $u \in (0,1)$; maybe just take $u = 1/2$. (Perform the infections from generation i to $i+1$ sequentially, so that when we reach $\lfloor un \rfloor$, we can stop the infection in the middle of the generation.)

The question is how uniformly spread locally the infection is. That is, take a large but fixed r (at least 3, but at most half the diameter of the entire graph), and for each vertex $i = 1, \dots, n$, record the ratio u_i of the vertices in the ball $B_r(i)$ that are infected. Plot the histogram of these u_i's. The average is of course u (check this as a sign that your code is correct), but does the histogram get concentrated around u as $n \to \infty$, or not at all? I do not know for sure, but I expect that there is a phase transition near $p = p_c = 2/\sqrt{3} - 1 \approx 0.155$: for $p < p_c$, the histogram is concentrated, while for $p > p_c$ it is not.

5. **The shape of infection in the plane.** Let μ be some probability distribution supported on $[1, 10]$, continuous or discrete, whatever. To each edge e of the square lattice $\mathbb{Z}^2$, assign an independent random length t_e with distribution μ. This may be considered as the time it takes from one endpoint of the edge, when it gets infected, to transfer the infection to the other endpoint. Thinking of these weights as lengths, take the radius R ball around the origin in this metric space. With the time interpretation, this is the set of vertices infected by time R. There is a general theorem stating that this ball, as $R \to \infty$, has a limiting shape. Play with the possible distributions μ and make pictures, to see what limit shapes you can get. How close can you get to a Euclidean disk? Besides just eyeballing the simulation results, one possible quantification of this is to let x_R be the Euclidean distance of the infection boundary on the x axis, z_R be the Euclidean distance on a diagonal, and to look at the ratio x_R/z_R.

6. **A random walk in Manhattan.** In Manhattan, all streets are one-way. So, for each infinite line of $\mathbb{Z}^2$, flip a fair coin, orienting it one way or the other. Given this random environment, let X_n be the random walk that, at each corner, chooses one of the two possibilities (continuing straight or turning, respecting the one-way direction) with probability $1/2$. Is this walk recurrent? How far is typically X_n from the origin, for large n? Draw pictures of the trajectory. [Barnabás and Olivér]

7. **Random walk in a changing random environment.** Consider critical dynamical bond percolation on the $n \times n$ discrete torus $(\mathbb{Z}/n\mathbb{Z})^2$: at the beginning, each edge is open or closed, with probability $1/2$ each, independently, then at each time step, one edge is chosen at random and its status is flipped. Now consider a particle that starts from the origin and performs a random walk in this changing maze with "infinite speed": that is, it is always uniformly distributed in its current cluster. That is, let $X_0 = (0,0)$, and let C_0 be the cluster of X_0. Then let X_1 be a uniform random vertex in C_0. Then flip the status of a random edge. The new cluster of X_1 will be C_1. Then let X_2 be a uniform random vertex in C_1. Then flip the status of a random edge. The new cluster of X_2 will be C_2, and so on.
How many steps are needed for the particle X_t to be approximately uniformly distributed on the torus?

8. **PageRank and degrees for Barabási-Albert.** Take the BA graph with m outgoing edges per vertex, at a large time n, first with $m = 1$ (so that we get a tree), then with $m = 4$. Look at six vectors:
   - the times of arrivals $(1, 2, \dots, n)$;
   - the PageRank scores $(R_1, \dots, R_n)$, indexed by the arrival times;
   - the first eigenvector centrality $(v_1, \dots, v_n)$;
   - the indegrees $(id_1, \dots, id_n)$ — the outdegrees are constant m, so not interesting;
   - the average indegree $(ad_1, \dots, ad_n)$ in a large oriented r-neighbourhood of each vertex, which is the set of vertices that can be reached by r steps respecting the orientation of the edges;

- the average indegree $(bd_1, \ldots, bd_n)$ in a large unoriented r-neighbourhood of each vertex, which is the set of vertices that can be reached by r steps using edges in any orientation.

  By a large r-neighbourhood, I mean an r that is fixed (does not depend on n) and is at most half of the diameter of the entire graph. Ideally, I would take $r = 10$, then take n so large that the diameter is at least 20, but this would probably mean an impossibly large n.

How do the correlation coefficients between these vectors behave, as $n \to \infty$? (The correlation between vectors $(v_1, \ldots, v_n)$ and $(w_1, \ldots, w_n)$ is understood as the correlation between the random variables v_U and w_U, when $U \sim \mathsf{Unif}(\{1, \ldots, n\})$. These correlations between our random vectors are now random themselves, but quite concentrated, my guess is. You can anyway take the expected correlation.) Does the arrival time of the highest ranked vertex (in each of the five centrality measures, R, v, id, ad, bd) go to infinity with the growth of the graph?

9. **Random genetic drift drives a population towards genetic uniformity.** Consider the Wright-Fisher model, as follows. A certain gene can have two alleles, A and B. At the beginning, the two alleles are represented equally in the gene pool given by N diploid individuals: there are altogether N copies of A and N copies of B. In the next generation, we again have N individuals, with each of their altogether $2N$ genes drawn independently at random from all the genes in the old generation. And so on, repeated forever.

   (a) How many generations does it typically take to eradicate one of the alleles from the gene pool?

   (b) Now assume that, in each generation, each individual may go dormant, independently with probability λ/N, some $\lambda \in (0, \infty)$ fixed, and stays dormant for an independent time ξ with distribution $\mathbf{P}[\,\xi \geq t\,] = t^{-\beta}$, $t = 1, 2, 3, \ldots$, some $\beta > 0$. When D individuals are dormant, then the reproduction is like before, just with the $N - D$ non-dormant individuals participating. When an individual wakes up, it will take part in the reproduction, and thus may re-introduce a seemingly extinct allele. For what values of λ and β is the time scale to get complete uniformity significantly larger than before?

10. **Positive overshoots with negative drift.** Consider a random walk $S_n = X_1 + \cdots + X_n$ on $\mathbb{R}$, with iid increments satisfying $\mathbf{E}X_i < 0$, but $\mathbf{P}[\,X_i > 0\,] > 0$, moreover, with $\mathbf{E}(X_i^+)^2 = \infty$, where $x^+ := \max\{x, 0\}$. (In particular, the size-biased version of X_i^+ exists, but has infinite expectation.) Let $T := \inf\{n > 0 : S_n > 0\}$, where the infimum is defined to be infinite if the set is empty.

    (a) Does it seem to be always true that $\mathbf{E}[\,S_T \mid T < \infty\,] < \infty$?

    (b) Does it seem to be always true that $\mathbf{E}[\,S_T \mid T < \infty\,] = \infty$?

11. **A liquid crystal (the math behind LCD screens).** In $\mathbb{R}^2/(n\mathbb{Z})^2$, the 2-dimensional continuum torus of side length n, let $X_1, X_2, \ldots$ be iid uniform random points. From each X_i iteratively, draw a unit vector at a uniform random angle, unless it intersects some previously drawn vector. Do this until we have n^2 vectors drawn.

    (a) How many tries are needed typically?

    (b) In a typical subsquare of side-length m, there are of order m^2 vectors. One can say that they are pointing roughly in the same direction (there is long range order in this subsquare) if their vector sum has length of order m^2. What is the largest $m = m(n)$ for which most subsquares have long range order?

    (c) Make pictures.

12. **Gaussian copula.** Consider the following data from the last 100 days for the prices of a pair of stocks:

    {197.353,196.091}, {199.994,198.65}, {199.072,199.348}, {200.708,201.}, {200.913,200.886}, {198.658,200.963}, {197.991,198.945}, {196.623,195.647}, {174.145,173.292}, {195.316,197.539}, {198.094,197.832}, {199.989,199.081}, {195.803,194.361}, {198.876,199.206}, {200.673,198.304}, {199.18,199.571}, {199.31,199.408}, {198.183,198.941}, {195.385,194.721}, {194.352,193.637}, {200.305,200.425}, {200.364,198.983}, {193.307,192.28}, {199.938,199.881},

{196.373,200.394}, {198.139,198.212}, {198.429,200.204}, {195.85,195.527}, {199.789,197.688}, {142.878,144.063}, {197.9,199.182}, {199.062,198.951}, {199.45,198.405}, {199.155,199.998}, {200.273,199.752}, {195.985,196.035}, {194.796,195.318}, {146.416,146.557}, {201.217,198.965}, {181.586,178.65}, {197.829,198.288}, {199.705,199.521}, {196.436,198.504}, {198.789,197.327}, {199.322,199.112}, {197.326,196.97}, {196.636,198.376}, {198.896,200.127}, {196.368,196.261}, {199.445,199.997}, {196.488,197.711}, {201.327,200.46}, {199.445,198.858}, {202.185,198.577}, {198.497,199.534}, {187.733,187.746}, {202.017,199.699}, {197.905,196.714}, {200.163,200.675}, {199.892,200.168}, {189.9,189.625}, {199.831,197.95}, {199.754,199.582}, {197.078,198.139}, {194.82,195.171}, {190.081,192.552}, {201.011,199.33}, {199.266,200.368}, {198.476,199.991}, {198.325,199.554}, {201.485,200.171}, {200.068,199.977}, {191.163,190.332}, {198.721,197.111}, {199.126,199.662}, {200.361,200.111}, {200.368,200.463}, {185.678,183.188}, {198.889,196.268}, {196.492,197.666}, {198.766,198.719}, {199.475,196.836}, {199.234,198.996}, {194.382,195.764}, {199.488,200.936}, {199.055,198.705}, {193.661,194.99}, {200.075,199.714}, {200.656,199.61}, {197.777,198.142}, {197.921,198.226}, {196.327,195.933}, {182.735,183.658}, {199.297,198.142}, {199.786,198.945}, {198.017,199.}, {94.6431,92.8557}, {197.519,196.789}, {199.518,200.268}, {198.256,198.966}

(I admit that this is actually iid data from a certain bivariate distribution, so we see things that would not happen for real stock prices. In real life, a day like $\{94.6431, 92.8557\}$ could happen without any warning signs beforehand, but would not be followed by completely normal days.)

(a) Calculate the sample mean vector μ and covariance matrix Σ for this data.

(b) Assuming that the distribution is bivariate normal, with the parameters (μ, Σ) just obtained, make a random sample how the next 100 days may look like.

(c) Estimate the marginal distributions of the data, then using the Gaussian copula with parameters (μ, Σ), make a random sample for the next 100 days.

(d) Vice versa, calculate the sample copula of the data, then assuming that the marginals are normal, with marginal parameters obtained above, make a random sample for the next 100 days.

(e) Now use the marginals and the copula obtained from the data, and make a random sample for the next 100 days.

(f) Plot all the data, in five separate two-dimensional pictures: (1) the original; (2) bivariate normal; (3) estimated marginals, Gaussian copula; (4) estimated copula, Gaussian marginals; (5) estimated copula, estimated marginals. How similar are these to each other?

(g) We go bankrupt in the future if both prices go below 0. For which model does this seem to be the most likely? (Of course, since the minimum number in the entire data is 142.269, it's not really possible to estimate this probability. I'm just asking for simple-minded intuition, which is what many traders would also rely on.)

13. In a **k-step Markov chain** $X_0, X_1, X_2, \ldots$, by definition, k is the smallest value such that, for every $n \geq k$, the distribution of X_n depends only on the previous k steps:

$$X_n \,|\, X_0, X_1, \ldots, X_{n-1} \stackrel{d}{=} X_n \,|\, X_{n-k}, \ldots, X_{n-1}.$$

The following sequence is the first 2000 steps of a k-step Markov chain. Make a guess what k is and what the transition probabilities are.

0, 1, 1, 0, 1, 1, 1, 1, 0, 1, 1, 0, 1, 1, 0, 0, 1, 0, 0, 0, 1, 0, 0, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 1, 0, 0, 1, 1, 0, 1, 0, 0,
0, 1, 1, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0,
1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 1, 1, 0, 0, 1, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 0, 0,
0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 1, 1, 0, 0, 0, 1, 1, 0, 0, 1, 1, 1, 0, 0, 1, 0, 1, 0, 0, 0,
1, 1, 1, 1, 0, 0, 1, 1, 0, 1, 1, 1, 1, 1, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 1, 0, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 1, 0, 1, 0, 0,
0, 0, 0, 1, 0, 0, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 1, 0, 0, 0, 0, 1, 1,
1, 1, 1, 0, 0, 1, 0, 0, 0, 1, 0, 0, 1, 1, 1, 1, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 1, 1, 0, 0, 1, 0, 0, 1, 1, 0, 0, 0, 0, 0, 1, 1,
1, 1, 0, 1, 0, 0, 1, 0, 0, 0, 0, 1, 1, 1, 1, 1, 0, 1, 0, 0, 0, 1, 0, 0, 1, 1, 1, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,

1, 1, 0, 1, 0, 1, 1, 1, 1, 0, 0, 1, 0, 0, 0, 1, 1, 1, 1, 1, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 1, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0,
1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 0, 1, 0, 1, 0, 0, 1, 1, 0, 1, 0, 0, 1, 1, 1, 0, 0, 1, 1, 1, 1, 1, 1, 0, 0, 1, 0, 0, 0, 1, 0,
0, 0, 1, 1, 1, 1, 0, 1, 0, 0, 0, 1, 0, 1, 1, 1, 0, 0, 0, 0, 0, 0, 1, 0, 1, 1, 0, 0, 1, 1, 1, 0, 0, 0, 0, 1, 0, 0, 1, 1, 0, 0,
0, 0, 1, 1, 1, 1, 0, 0, 1, 1, 1, 1, 0, 1, 0, 0, 0, 0, 1, 1, 0, 0, 0, 1, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 1, 1, 1,
1, 1, 1, 0, 0, 1, 1, 1, 1, 0, 0, 0, 1, 0, 1, 1, 1, 1, 0, 0, 1, 0, 0, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 1, 0, 1, 1, 1, 0, 0, 1, 1, 1, 0, 0, 1, 1, 1, 1, 1, 1, 1, 0, 0, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 1, 0, 0, 0, 1, 0, 0, 0,
1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 1, 0, 0, 1, 1, 1, 1, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 1, 0, 1,
1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 1, 1, 0, 0, 1, 1, 1, 1, 1, 1, 0, 1, 0, 1, 0, 0, 1, 1, 1, 0, 1, 0, 0, 1, 0, 1,
1, 1, 1, 1, 1, 0, 0, 0, 1, 1, 1, 1, 0, 0, 1, 1, 1, 1, 0, 0, 0, 1, 1, 0, 1, 1, 0, 0, 1, 0, 0, 1, 0, 0, 1, 1, 1, 0, 0, 0, 1, 1,
1, 1, 1, 0, 1, 0, 0, 0, 1, 0, 0, 1, 1, 1, 1, 1, 0, 1, 1, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 1, 0, 1, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0,
1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 0, 0, 1, 1, 0, 0, 1, 0, 0, 1, 1, 1, 1, 0, 0, 1, 0, 0, 0,
1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 1,
0, 1, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 1, 1, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 1, 1, 0, 0, 0, 1, 0, 0, 1, 0, 0, 1, 1, 0, 1, 0, 1,
1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 1, 0, 0, 0, 1, 0, 1,
0, 1, 1, 1, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 0, 1, 0, 0, 1, 0, 0, 1, 0, 0, 1, 1, 0, 0, 1, 1, 0, 1,
1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 1, 0, 0, 1, 1, 1, 0, 0, 1, 0, 0, 1, 0, 0, 1, 1, 1, 0, 1, 0, 0, 1, 0, 1, 0, 0, 1, 0, 0, 0, 1, 1,
1, 1, 0, 0, 0, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 1, 1, 1, 0, 0, 1, 0, 1, 0, 1, 0, 0, 1, 0, 1, 1, 1, 1, 1, 0, 1, 1,
0, 1, 0, 0, 0, 0, 1, 1, 1, 1, 0, 0, 1, 1, 0, 0, 0, 1, 0, 0, 1, 0, 0, 1, 1, 1, 0, 0, 1, 0, 1, 0, 0, 1, 0, 0, 1, 0, 0, 1, 1, 1,
1, 1, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 0, 0, 0, 1, 1, 0, 0, 1, 1, 1, 0, 0, 0, 0, 1, 1, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 1, 1, 0, 0,
0, 1, 0, 1, 1, 1, 1, 1, 0, 0, 1, 1, 1, 1, 1, 1, 1, 0, 0, 1, 0, 0, 0, 0, 1, 1, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1,
1, 0, 1, 1, 0, 0, 0, 0, 0, 1, 1, 1, 0, 1, 0, 0, 0, 0, 0, 1, 0, 1, 1, 1, 1, 0, 0, 1, 1, 1, 1, 1, 0, 1, 1, 1, 0, 0, 1, 0, 0, 1,
1, 1, 1, 1, 1, 1, 1, 0, 1, 0, 0, 1, 0, 0, 0, 1, 1, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 1, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 1,
0, 0, 0, 1, 1, 0, 0, 1, 1, 0, 1, 1, 1, 1, 1, 0, 0, 1, 0, 0, 1, 1, 1, 1, 1, 0, 0, 1, 0, 0, 1, 0, 1, 0, 1, 0, 0, 1, 0, 1, 0, 0,
0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0,
1, 0, 0, 1, 1, 0, 0, 1, 0, 0, 0, 1, 1, 1, 1, 0, 1, 0, 0, 0, 1, 0, 0, 1, 0, 1, 0, 1, 1, 1, 1, 0, 1, 0, 0, 0, 0, 1, 1, 1, 1, 1,
0, 0, 1, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 1, 1, 1,
0, 0, 0, 0, 0, 1, 1, 0, 0, 1, 1, 1, 1, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 1, 0, 0, 1, 0, 0, 0, 1, 1,
1, 0, 0, 0, 0, 0, 1, 1, 0, 0, 1, 1, 1, 1, 1, 0, 1, 1, 1, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 1, 1, 1, 1, 0, 0, 1, 0, 0,
0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 1, 1, 0, 1, 0, 0, 1, 1, 1, 1, 0, 0, 1, 1, 0, 0, 1, 0, 1, 1, 1, 1, 1, 1, 1,
1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 1, 1, 0, 0, 0, 0, 1

14. **Benford's law.** The first digits in many real-life numerical datasets follow Benford's distribution: the probability that the first digit is k is about $\log_{10}(k+1) - \log_{10} k$, for $k = 1, \ldots, 9$. During the week following Biden's victory in the US elections, many right-wing news outlets (including in Hungary) picked up the news from "some mathematicians" that the dataset consisting of the number of people voting for Biden in each district in certain key states does not follow Benford's law, while the numbers for Trump do, which is a proof of election fraud: Biden's number must have been artificially altered. The following exercise explains how much unfounded this is. Part (a) describes models where we can expect to see Benford's law; real-life examples include prices of various goods, sizes of cities, distances to stars, surface areas of lakes. The most important requirement is that the database should contain numbers of several

orders of magnitude. Part (b) explains that the election numbers are not like that.

(a) Let $X_1, \ldots, X_{10}$ be independent random variables, not necessarily identically distributed, but each having a distribution that is supported in the interval $[1, 9]$, continuous or discrete, whatever, but should be reasonably random — say, with a standard deviation at least 2. Let $Y \in \{1, \ldots, 9\}$ be the first digit of the product $X_1 \cdots X_{10}$. Make 1000 independent samples, and plot the empirical distribution of Y. Play with the 10 distributions for $X_1, \ldots, X_{10}$, and see how often the histogram for Y is similar to Benford's law.

(b) Now take 1000 integers, $N_1, \ldots, N_{1000}$ each between 200 and 1000, in any way you want. Let B_i, for $i = 1, \ldots, 1000$ be independent random variables with distribution $\mathsf{Binom}(N_i, 2/3)$, and let $T_i := N_i - B_i$. These variables model the number of votes for Biden and Trump, respectively, in the ith electoral district with N_i voters, in a Biden-leaning county of Pennsylvania (exactly as in the news). Plot the empirical distribution of the first digit of the $B_i's$, and the empirical distribution of the first digit of the T_i's. Can you make these two plots resemble Benford's law, by choosing the N_i's? Typically, which of the two plots is more similar to Benford's law?