Appendix A

The Birthday Problem

The setting is that we have q balls. View them as numbered, $1, \ldots, q$. We also have N bins, where $N \geq q$. We throw the balls at random into the bins, one by one, beginning with ball 1. At random means that each ball is equally likely to land in any of the N bins, and the probabilities for all the balls are independent. A collision is said to occur if some bin ends up containing at least two balls. We are interested in $C(N, q)$, the probability of a collision.

The birthday paradox is the case where $N = 365$. We are asking what is the chance that, in a group of q people, there are two people with the same birthday, assuming birthdays are randomly and independently distributed over the days of the year. It turns out that when q hits $\sqrt{365}$ the chance of a birthday collision is already quite high, around 1/2.

This fact can seem surprising when first heard. The reason it is true is that the collision probability $C(N, q)$ grows roughly proportional to q^2/N. This is the fact to remember. The following gives a more exact rendering, providing both upper and lower bounds on this probability.

Theorem A.1 [Birthday bound] Let $C(N, q)$ denote the probability of at least one collision when we throw $q \geq 1$ balls at random into $N \geq q$ buckets. Then

$$C(N, q) \quad \leq \quad \frac{q(q-1)}{2N}$$

and

$$C(N, q) \quad \geq \quad 1 - e^{-q(q-1)/2N} .$$

Also if $1 \leq q \leq \sqrt{2N}$ then

$$C(N, q) \quad \geq \quad 0.3 \cdot \frac{q(q-1)}{N} . \blacksquare$$

In the proof we will find the following inequalities useful to make estimates.

Proposition A.2 The inequality

$$\left(1 - \frac{1}{e}\right) \cdot x \;\leq\; 1 - e^{-x} \;\leq\; x \;.$$

is true for any real number x with $0 \leq x \leq 1$. ▮

Proof of Theorem A.1: Let C_i be the event that the i-th ball collides with one of the previous ones. Then $\Pr[C_i]$ is at most $(i-1)/N$, since when the i-th ball is thrown in, there are at most $i-1$ different occupied slots and the i-th ball is equally likely to land in any of them. Now

$$\begin{aligned} C(N,q) &= \Pr[C_1 \vee C_2 \vee \cdots \vee C_q] \\ &\leq \Pr[C_1] + \Pr[C_2] + \cdots + \Pr[C_q] \\ &\leq \frac{0}{N} + \frac{1}{N} + \cdots + \frac{q-1}{N} \\ &= \frac{q(q-1)}{2N} \;. \end{aligned}$$

This proves the upper bound. For the lower bound we let D_i be the event that there is no collision after having thrown in the i-th ball. If there is no collision after throwing in i balls then they must all be occupying different slots, so the probability of no collision upon throwing in the $(i+1)$-st ball is exactly $(N-i)/N$. That is,

$$\Pr[D_{i+1} \mid D_i] \;=\; \frac{N-i}{N} \;=\; 1 - \frac{i}{N} \;.$$

Also note $\Pr[D_1] = 1$. The probability of no collision at the end of the game can now be computed via

$$\begin{aligned} 1 - C(N,q) &= \Pr[D_q] \\ &= \Pr[D_q \mid D_{q-1}] \cdot \Pr[D_{q-1}] \\ &\vdots \quad \vdots \\ &= \prod_{i=1}^{q-1} \Pr[D_{i+1} \mid D_i] \\ &= \prod_{i=1}^{q-1} \left(1 - \frac{i}{N}\right) \;. \end{aligned}$$

Note that $i/N \leq 1$. So we can use the inequality $1 - x \leq e^{-x}$ for each term of the above expression. This means the above is not more than

$$\prod_{i=1}^{q-1} e^{-i/N} \;=\; e^{-1/N - 2/N - \cdots - (q-1)/N} \;=\; e^{-q(q-1)/2N} \;.$$

Putting all this together we get

$$C(N,q) \ \geq \ 1 - e^{-q(q-1)/2N} \ ,$$

which is the second inequality in Proposition A.1. To get the last one, we need to make some more estimates. We know $q(q-1)/2N \leq 1$ because $q \leq \sqrt{2N}$, so we can use the inequality $1 - e^{-x} \geq (1 - e^{-1})x$ to get

$$C(N,q) \ \geq \ \left(1 - \frac{1}{e}\right) \cdot \frac{q(q-1)}{2N} \ .$$

A computation of the constant here completes the proof. ▮