More Occupancy Problems

Assume that m balls are placed randomly is n boxes.

- How many boxes are empty?
- How many boxes have at least k balls?
- What is the maximum number of balls in any box?
- ...

The Birthday Paradox

Having thirty people in the room, is it more likely or not that some two people in the room share the same birthday?

Assume birthdays uniformly distributed in $[1, \ldots, 365]$.

Count the configurations where no two people a birthday. The probability that all birthdays are distinct is:

$$\frac{\binom{365}{30} 30!}{365^{30}} \approx 0.29. \tag{1}$$

We can also calculate this probability by considering one person at a time:

$$\left(1 - \frac{1}{365}\right) \cdot \left(1 - \frac{2}{365}\right) \cdot \left(1 - \frac{3}{365}\right) \ldots \cdot \left(1 - \frac{29}{365}\right)$$

More generally, if there are m people and n possible birthdays, the probability that all m have different birthdays is

$$\left(1-\frac{1}{n}\right)\cdot\left(1-\frac{2}{n}\right)\cdot\left(1-\frac{3}{n}\right)\cdots\cdot\left(1-\frac{m-1}{n}\right)$$

$$=\prod_{j=1}^{m-1}\left(1-\frac{j}{n}\right).$$

Using $1-\frac{k}{n}\approx \mathrm{e}^{-k/n}$ when $k << n$,

$$\begin{aligned}\prod_{j=1}^{m-1}\left(1-\frac{j}{n}\right) &\approx \prod_{j=1}^{m-1}\mathrm{e}^{-j/n}\\ &= \mathrm{e}^{-\sum_{j=1}^{m-1} j/n}\\ &= \mathrm{e}^{-m(m-1)/2n}\\ &\approx \mathrm{e}^{-m^2/2n}.\end{aligned}$$

We place m balls randomly into n bins, how many bins remain empty?
The probability that a given bin is missed by all m balls is

$$\left(1-\frac{1}{n}\right)^m \approx \mathrm{e}^{-m/n}$$

Let $X_j = 1$ if the j-th bin is empty else $X_j = 0$.

$E[X_j] = (1-\frac{1}{n})^m$.

$$E[X] = E\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} E[X_i] = n\left(1-\frac{1}{n}\right)^m \approx n\mathrm{e}^{-m/n}.$$

$$\Pr(X=0) = ?$$

How many bins have r balls?
The probability that a given bin has r balls is

$$p_r = \binom{m}{r} \left(\frac{1}{n}\right)^r \left(1 - \frac{1}{n}\right)^{m-r}$$
$$= \frac{1}{r!} \frac{m(m-1)\cdots(m-r+1)}{n^r} \left(1 - \frac{1}{n}\right)^{m-r}.$$

For $m, n >> r$,

$$p_r \approx \frac{\mathrm{e}^{-m/n}(m/n)^r}{r!},$$

The Poisson distribution

Definition

A discrete Poisson random variable X with parameter μ is given by the following probability distribution on $j = 0, 1, 2, \ldots$.

$$\Pr(X = j) = \frac{e^{-\mu}\mu^j}{j!}.$$

$$\begin{aligned}
\sum_{j=0}^{\infty} \Pr(X = j) &= \sum_{j=0}^{\infty} \frac{\mathrm{e}^{-\mu}\mu^j}{j!} \\
&= \mathrm{e}^{-\mu} \sum_{j=0}^{\infty} \frac{\mu^j}{j!} \\
&= 1,
\end{aligned}$$

Expectation

$$
\begin{aligned}
\mathbf{E}[X] &= \sum_{j=1}^{\infty} j \Pr(X = j) \\
&= \sum_{j=1}^{\infty} j \frac{\mathrm{e}^{-\mu} \mu^j}{j!} \\
&= \mu \sum_{j=1}^{\infty} \frac{e^{-\mu} \mu^{j-1}}{(j-1)!} \\
&= \mu \sum_{j=0}^{\infty} \frac{e^{-\mu} \mu^j}{j!} \\
&= \mu.
\end{aligned}
$$

Lemma

The sum of a finite number of independent Poisson random variables is a Poisson random variable.

Proof.

Consider two independent Poisson random variables X and Y with means μ_1 and μ_2. Now

$$\begin{aligned}
\Pr(X+Y=j) &= \sum_{k=0}^{j} \Pr((X=k)\cap(Y=j-k)) \\
&= \sum_{k=0}^{j} \frac{\mathrm{e}^{-\mu_1}\mu_1^k}{k!}\frac{\mathrm{e}^{-\mu_2}\mu_2^{(j-k)}}{(j-k)!} \\
&= \frac{\mathrm{e}^{-(\mu_1+\mu_2)}}{j!}\sum_{k=0}^{j}\frac{j!}{k!(j-k)!}\mu_1^k\mu_2^{(j-k)} \\
&= \frac{\mathrm{e}^{-(\mu_1+\mu_2)}}{j!}\sum_{k=0}^{j}\binom{j}{k}\mu_1^k\mu_2^{(j-k)} \\
&= \frac{\mathrm{e}^{-(\mu_1+\mu_2)}(\mu_1+\mu_2)^j}{j!}.
\end{aligned}$$

□

Another Proof

$$\begin{aligned} E[\mathrm{e}^{tX}] &= \sum_{k=0}^{\infty} \frac{\mathrm{e}^{-\mu}\mu^k}{k!}\mathrm{e}^{tk} \\ &= \mathrm{e}^{\mu(\mathrm{e}^t-1)} \sum_{k=0}^{\infty} \frac{\mathrm{e}^{-\mu\mathrm{e}^t}\left(\mu\mathrm{e}^t\right)^k}{k!} = \mathrm{e}^{\mu(\mathrm{e}^t-1)}. \end{aligned}$$

$$\begin{aligned} M_{X+Y}(t) &= M_X(t) \cdot M_Y(t) = \mathrm{e}^{\mu_1(\mathrm{e}^t-1)} \cdot \mathrm{e}^{\mu_2(\mathrm{e}^t-1)} \\ &= \mathrm{e}^{(\mu_1+\mu_2)(\mathrm{e}^t-1)}. \end{aligned}$$

which is the moment generating function of a Poisson distribution with expectation $\mu_1 + \mu_2$.

Chernoff bound

Theorem

Let X be a Poisson random variable with parameter μ.

1. *If $x > \mu$, then $\Pr(X \geq x) \leq \frac{e^{-\mu}(e\mu)^x}{x^x}$;*
2. *If $x < \mu$, then $\Pr(X \leq x) \leq \frac{e^{-\mu}(e\mu)^x}{x^x}$;*

Proof.

For any $t > 0$ and $x > \mu$,

$$\Pr(X \geq x) = \Pr(e^{tX} \geq e^{tx}) \leq \frac{E[e^{tX}]}{e^{tx}}.$$

Hence

$$\Pr(X \geq x) \leq e^{\mu(e^t - 1) - xt}.$$

Choosing $t = \ln(x/\mu) > 0$ gives

$$\begin{aligned} \Pr(X \geq x) &\leq e^{x - \mu - x\ln(x/\mu)} \\ &= \frac{e^{-\mu}(e\mu)^x}{x^x}. \end{aligned}$$

□

Proof.

For any $t < 0$ and $x < \mu$,

$$\Pr(X \leq x) = \Pr(e^{tX} \geq e^{tx}) \leq \frac{E[e^{tX}]}{e^{tx}}.$$

Hence

$$\Pr(X \leq x) \leq e^{\mu(e^t - 1) - xt}.$$

Choosing $t = \ln(x/\mu) < 0$, gives

$$\begin{aligned} \Pr(X \leq x) &\leq e^{x - \mu - x\ln(x/\mu)} \\ &= \frac{e^{-\mu}(e\mu)^x}{x^x}. \end{aligned}$$

Maximum per bin

Lemma

When n balls are thrown independently and uniformly at random into n bins, the probability that the maximum load is more than $3 \ln n / \ln \ln n$ is at most $1/n$ for n sufficiently large.

The probability that bin 1 receives at least M balls is at most

$$\binom{n}{M} \left(\frac{1}{n}\right)^M.$$

$$\binom{n}{M} \left(\frac{1}{n}\right)^M \leq \frac{1}{M!} \leq \left(\frac{\mathrm{e}}{M}\right)^M.$$

We use:

$$\frac{k^k}{k!} < \sum_{i=0}^{\infty} \frac{k^i}{i!} = \mathrm{e}^k,$$

which gives:

$$k! > \left(\frac{k}{\mathrm{e}}\right)^k.$$

The probability that any bin receives at least $M \geq 3 \ln n / \ln \ln n$ balls is bounded above by

$$\begin{aligned} n\left(\frac{\mathrm{e}}{M}\right)^{M} &\leq n\left(\frac{\mathrm{e} \ln \ln n}{3 \ln n}\right)^{3 \ln n / \ln \ln n} \\ &\leq n\left(\frac{\ln \ln n}{\ln n}\right)^{3 \ln n / \ln \ln n} \\ &= \mathrm{e}^{\ln n}\left(\mathrm{e}^{\ln \ln \ln n - \ln \ln n}\right)^{\frac{3 \ln n}{\ln \ln n}} \\ &= \mathrm{e}^{-2 \ln n + \frac{3(\ln n)(\ln \ln \ln n)}{\ln \ln n}} \\ &\leq \frac{1}{n} \end{aligned}$$

Example: Number of Empty Bins

Theorem

Assume that m balls are placed randomly in n boxes. Assume that $m, n \to \infty$, such that the quantity $\lambda = ne^{-m/n}$ is bounded. Let $p_r(n, m)$ be the probability that exactly r boxes are empty. For each fixed r,

$$\lim_{n.m \to \infty} p_r(n, m) = P(r, \lambda) = e^{-\lambda} \frac{\lambda^r}{r!}.$$

Poisson Distribution:

X has a Poisson distribution with parameter λ ($P(\lambda)$):

$$Pr(X = i) \; = \; e^{-\lambda} \frac{\lambda^i}{i!}.$$

Inclusion-Exclusion

Let $E_1, ..., E_n$ be arbitrary events:

$$\begin{aligned} Pr(\cup_{i=1}^{n} E_i) &= \sum_{i=1}^{n} Pr(E_i) - \sum_{i<j} Pr(E_i \cap E_j) \\ &+ \sum_{i<j<k} Pr(E_i \cap E_j \cap E_k) \\ &- \cdots + (-1)^{\ell+1} \sum_{i_1<i_2<\cdots<i_\ell} Pr(\cap_{r=1}^{\ell} E_r) + \cdots \end{aligned}$$

Proof of the Theorem

We first compute $P_0(m, n)$.
Let E_1 be the event "box i is empty".

$$1 - P_0(m.n) \;=\; Pr(\cup_{i=1}^{n} E_i)$$

$$Pr(E_i) = (1 - \frac{1}{n})^m$$

$$Pr(\cap_{i=1}^{k} E_i) = (1 - \frac{k}{n})^m$$

$$\sum_{i_1 < i_2 < \cdots < i_k} Pr(\cap_{j=1}^{k} E_{i_j}) = \binom{n}{k}(1 - \frac{k}{n})^m$$

Lemma

For any fixed $k > 0$,

$$\lim \binom{n}{k}(1-\frac{k}{n})^m = \frac{\lambda^k}{k!}$$

Proof.

$$\frac{n^k}{k!}(1-\frac{k}{n})^k \leq \binom{n}{k} \leq \frac{n^k}{k!}$$

$$e^t(1-\frac{t^2}{n}) \leq (1+\frac{t}{n})^n \leq e^t$$

$$P_0(m,n) \;=\; 1-\lambda+\frac{\lambda^2}{2}-\frac{\lambda^3}{3!}+.....$$

$$P_0(m,n) \;=\; \sum_{i=0}^{n}(-1)^i\frac{\lambda^i}{i!}$$

$$P_0(m,n) \rightarrow e^{-\lambda}$$

$$P_r(m,n) = \binom{n}{r}(1-\frac{r}{n})^m P_0(m,n-r)$$

$$P_r(m,n) \rightarrow \frac{\lambda^r}{r!}e^{-\lambda}$$

Limit of Binomial Distribution

Theorem

Let X_n be a binomial random variable with parameters n and p, where p is a function of n and $\lim_{n\to\infty} np = \lambda$ is a constant independent of n. Then for any fixed k,

$$\lim_{n\to\infty} \Pr(X_n = k) = \frac{\mathrm{e}^{-\lambda}\lambda^k}{k!}.$$

Random Graphs

Many important computation problems are defined on graphs.

Many of these problems are NP-Complete but are solved 'efficiently" in practice.

A probabilistic model of graphs for probabilistic analysis of graph algorithms.

Random Graph Process

Consider the following stochastic process:

- Start with n vertices, no edges.
- In each step add one edge between a randomly chosen pair of vertices.

If we stop the process after N steps (i.e. after adding N random edges):

1. Is the graph connected?
2. Does the graph have a large connected component?
3. Are there isolated vertices?

A graph property is **monotone** if for any two graphs $G = (V, E)$ and $G' = (V, E')$, such that $E \subseteq E'$, if G has the property also G' has that property.

Monotone properties:

1. No isolated vertices;
2. Connectivity;
3. Perfect Matching;
4. Hamiltonian Path;
5. ....

The $G_{n,N}$ model:

- The set of all graphs on n vertices with exactly N edges.
- All graphs in this set have equal probability.
- There are $T = \binom{\binom{n}{2}}{N}$ graphs on n vertices with exactly N edges.
- The probabilistic space $G_{n,N}$ has T simple events, each with probability $\frac{1}{T}$.

1. What is the probability that a graph in $G_{n,N}$ has isolated vertices?
2. What is the probability that a graph in $G_{n,N}$ is connected?
3. What is the probability that a graph in $G_{n,N}$ has a Hamiltonian cycle?
4. How fast can an algorithm find a Hamiltonian cycle in $G \in G_{n,N}$?

Isolated Vertices

Theorem

Let $N = \frac{1}{2}(n \log n + cn)$*, the probability that* $G \in G_{n,N}$ *has isolated vertices is* $1 - e^{-e^{-c}}$*.*

Proof.

Coupon collector.

□

The $G_{n,p}$ model:

- The set of all graphs on n vertices.
- The probability of a graph with N edges is $p^N(1-p)^{\binom{n}{2}-N}$
- For $N = \binom{n}{2}p$
  - $G_{n,N}$ and $G_{n,p}$ have similar monotone properties.

Algorithm for Finding a Hamiltonian Path

A **simple** path is a path with no loops, i.e. a vertex is visited no more than once.
A **Hamiltonian Path** is a simple path that visits every vertex of the graph.
A **Hamiltonian Cycle** is a cycle that visits every vertex in the graph exactly once.
Given a graph G, deciding if G has a Hamiltonian path/cycle is NP-Complete.

Rotation

Let G be an undirected graph. Assume that

$$P = v_1, v_2, ..., v_k$$

is a simple path in G and (v_k, v_i) is an edge of G then

$$P' = v_1, ..., v_i, v_k, v_{k-1}, ..., v_{i+2}, v_{i+1}$$

is a simple path is G.

Algorithm

Assume that each vertex has its list of adjacent edges, in a random order.

1. Choose an arbitrary vertex x_0 to start the path. $HEAD = x_0$.
2. Repeat until all vertices are connected
   1. Let $(HEAD, u)$ be the first edge in $HEAD$'s list.
   2. Remove $(HEAD, u)$ from $HEAD$'s and u's lists.
   3. If u not in the path $HEAD := u$, else use the edge to "rotate" the path.

Modified Algorithms

Consider a "less efficient" algorithm that for each vertex u keeps two lists:

1. $new\_edges(u)$ - adjacent edges that were not used yet;
2. $old\_edges(u)$ - edges that were already used.

When u is at the head of the path we choose

- A random element in $old\_edges(u)$ with probability $\frac{|old\_edges(u)|}{n}$.
- With probability $\frac{1}{n}$ the tail of the path becomes the head of the path.
- Else, (with probability $1 - \frac{1}{n} - \frac{|old\_edges(u)|}{n}$), the head of $new\_edges(u)$ list (and move it to $old\_edges(u)$).

Modified Hamiltonian Cycle Algorithm:

1. Start with a random vertex as the head of the path.
2. Repeat until the rotation edge closes a Hamiltonian cycle or the unused-edges list of the head of the path is empty:
   1. Let the current path be $P = v_1, v_2, \ldots, v_k$, with v_k being the head.
   2. Execute i, ii or iii below with probabilities $\frac{1}{n}$, $\frac{|\text{used-edges}(v_k)|}{n}$, and $1 - \frac{1}{n} - \frac{|\text{used-edges}(v_k)|}{n}$, respectively:
      1. Reverse the path, and make v_1 the head.
      2. Choose uniformly at random an edge from used-edges(v_k); if the edge is (v_k, v_i), rotate the current path with (v_k, v_i) and set v_{i+1} to be the head. (If the edge is (v_k, v_{k-1}), then no change is made.)
      3. Select the first edge from unused-edges(v_k), call it (v_k, u). If $u \neq v_i$ for $1 \leq i \leq k$, add $u = v_{k+1}$ to the end of the path and make it the head. Otherwise, if $u = v_i$, rotate the current path with (v_k, v_i), and set v_{i+1} to be the head. (This step closes the Hamiltonian path if $k = n$ and the chosen edge is (v_n, v_1).)
   3. Update the used-edges and unused-edges lists appropriately.
3. Return a Hamiltonian cycle if one was found or failure if no cycle was found.

Lemma

The probability that a given vertex becomes HEAD at a given iteration of the modified algorithm is $\frac{1}{n}$.

Proof.

Clear for the tail of the paths and for neighbors of the current head in old edges.
The probability of using the *new_edge*() list is

$$1 - \frac{1}{n} - \frac{|old\_edges(u)|}{n}$$

and that edge is connected to a vertex chosen uniformly at random from a set of

$$n - 1 - |old\_edge(u)|$$

vertices. □

Theorem

Suppose the input to the modified Hamiltonian cycle algorithm initially has unused edge-lists where each edge (v, u) with $u \neq v$ is placed on v's list independently with probability $q \geq \frac{20 \ln n}{n}$. Then the algorithm successfully finds a Hamiltonian cycle in $O(n \ln n)$ iterations of the repeat loop (step 2) with probability $1 - O(n^{-1})$.

Note that we did not assume that the input random graph has a Hamiltonian cycle.

$\mathcal{E}_1$: The algorithm run $3n \ln n$ iterations with no unused-edges list becoming empty, but failed to construct a Hamiltonian path.
$\mathcal{E}_2$: At least one unused-edges list became empty during the first $3n \ln n$ iterations of the loop.
We first bound $\Pr(\mathcal{E}_1)$.
The probability that any vetex was not chosen in $2n \ln n$ iterations is

$$n\left(1-\frac{1}{n}\right)^{2n\ln n} \leq \mathrm{e}^{-\ln n} = \frac{1}{n}.$$

The probability that the path does not become a cycle within the next $n \ln n$ iterations is

$$\left(1-\frac{1}{n}\right)^{n\ln n} \leq \mathrm{e}^{-\ln n} = \frac{1}{n}.$$

$$\Pr(\mathcal{E}_1) \leq \frac{2}{n}.$$

$\Pr(\mathcal{E}_2) =$ the probability that an unused-edges list is empty in the first $3n \ln n$ iterations.

$\mathcal{E}_{2a}$: At least $9 \ln n$ edges were removed from the unused-edges list of at least one vertex in the first $3n \ln n$ iterations of the loop.

$\mathcal{E}_{2b}$: At least one vertex has fewer than $10 \ln n$ edges.

$$\Pr(\mathcal{E}_2) \leq \Pr(\mathcal{E}_{2a}) + \Pr(\mathcal{E}_{2b}).$$

We bound $\Pr(\mathcal{E}_{2a})$.

Let X_j^i be a Bernoulli random variable that is 1 if the i-th vertex is adjacent to the edge used in the j-th iteration of the loop and 0 otherwise.

$X^i = \sum_{j=1}^{3n \ln n} X_j^i$.

$\mathbf{E}[X_j^i] = \frac{1}{n}$ and $\mathbf{E}[X^i] \leq 3 \ln n$.

$$\Pr(X^i \geq 9 \ln n) \leq \left(\frac{\mathrm{e}^2}{27}\right)^{3 \ln n} \leq \frac{1}{n^2}.$$

$\Pr(\mathcal{E}_{2a}) \leq 1/n$.

$\mathcal{E}_{2b}$: At least one vertex has $10 \ln n$ or fewer edges initially in its unused-edges list.
$Y^i =$ number of edges initially in vertex i unused-edges list.
$\mathbf{E}[Y^i] = (n-1)q \geq 20(n-1)\ln n/n \geq 19 \ln n$ for sufficiently large n.

$$\Pr(Y^i \leq 10 \ln n) \leq \mathrm{e}^{-19 \ln n (9/19)^2/2} < \frac{1}{n^2}$$

$$\Pr(\mathcal{E}_{2b}) < \frac{1}{n},$$

$$\Pr(\mathcal{E}_2) \leq \frac{1}{n} + \frac{1}{n} = \frac{2}{n}.$$

$$\Pr(\mathcal{E}_1) + \Pr(\mathcal{E}_2) \leq \frac{2}{n} + \frac{2}{n} = \frac{4}{n}.$$

Corollary

By putting edges on the unused-edges lists appropriately, the algorithm finds a Hamiltonian cycle on a graph chosen randomly from $G_{n,p}$ with probability $1 - O(1/n)$ whenever $p \geq 40 \ln n/n$.

Let $q \in [0, 1]$ be such that $p = 2q - q^2$.

For any edge (u, v) do exactly one of the following:

1. With probability $q(1 - q)/(2q - q^2)$, place the edge on u's unused edge-list, but not v's;
2. With probability $q(1 - q)/(2q - q^2)$, place the edge on v's unused edge-list, but not u's;
3. With probability $q^2/(2q - q^2)$, the edge is placed on both unused-edge lists.

For edge (x, y), the probability that it is initially placed in the unused-edge list for x is

$$p\left(\frac{q(1-q)}{2q-q2}+\frac{q2}{2q-q2}\right)=q;$$

The probability that it is placed in both x's and y's lists is:

$$\frac{pq^2}{2q-q^2}=q^2,$$

so events are independent.
We need $q \geq 20 \ln n/n$.
When $p \geq 40 \ln n/n$, we have $q \geq p/2 \geq 20 \ln n/n$.