CSE 6311

Notes by chandrashekar vijayraenu for the class on April 23rd 2009

Occupancy problem

Bins and Balls Throw n balls into n bins at random.

1. Pr[Bin 1 is empty] = $\left(1-\frac{1}{n}\right)^n \sim \frac{1}{e}$
2. Pr[Bin 1 has k balls] =$\binom{n}{k}\left(\frac{1}{n}\right)^k\left(1-\frac{1}{n}\right)^{n-k} \le \frac{1}{e*k!}$

Sterling's Approximations

$$(\frac{n}{k})^k \le \binom{n}{k} \le (\frac{ne}{k})^k$$

Thus, letting $A_{i,k}$ be the event that bin i contains at least k balls, we have

$$\mathbf{Pr}(A_{i,k}) \;=\; \sum_{i=k}^{n}\binom{n}{i}^i\left(\frac{i}{n}\right)^i\left(1-\frac{i}{n}\right)^{n-k}$$

Thus, by the union bound,

$$\mathbf{Pr}(\text{any bin contains more than } k \text{ balls}) \;\le\; \sum_{i=1}^{n}\mathbf{Pr}(A_{i,k})$$

In order to approximate this, we need to derive a simple upper bound for Pr($A_{i,k}$). We'll make use of the following elementary inequality, for any i <= n:

$$\left(\frac{n}{i}\right)^i \le \binom{n}{i} \le \left(\frac{ne}{i}\right)^i$$

Using this we can easily derive the bound

$$\begin{aligned}\mathbf{Pr}(A_{i,k}) &\le \sum_{i=k}^{n}\left(\frac{ne}{i}\right)^i\left(\frac{1}{n}\right)^i\\ &= \left(\frac{e}{i}\right)^k\left(1+\frac{e}{k}+\left(\frac{e}{k}\right)^2+\cdots\right)\\ &= \left(\frac{e}{k}\right)^k\frac{1}{1-e/k}\end{aligned}$$

Now comes the tedious part. Let k = [(3 log n)/ log log n]. Then

$$\begin{aligned}
\mathbf{Pr}(A_{i,k}) &\leq \left(\frac{e}{k}\right)^k \frac{1}{1-e/k} \\
&\leq 2\left(\frac{e}{3\log n/\log\log n}\right)^k \\
&\leq 2\left(e^{1-\log 3-\log\log n+\log\log\log n}\right)^k \\
&\leq 2\left(e^{-\log\log n+\log\log\log n}\right)^k \\
&\leq 2\left(e^{-3\log n+3\frac{\log\log\log n}{\log\log n}\log n}\right) \\
&\leq 2\left(e^{-2\log n}\right) \\
&= \frac{2}{n^2}
\end{aligned}$$

for n sufficiently large that (log log log n)/ log log n < 1/3.
It follows that

$$\begin{aligned}
\mathbf{Pr}(\text{no bin contains more than } \lceil (3\log n)/\log\log n \rceil \text{ balls}) &= 1-\sum_{i=1}^{n}\mathbf{Pr}(A_{i,k}) \\
&\geq 1-\frac{2}{n}
\end{aligned}$$

Stable marriage problem

Given n men and n women, where each person has ranked all members of the opposite sex with a unique number between 1 and n in order of preference, marry the men and women off such that there are no two people of opposite sex who would both rather have each other than their current partners. If there are no such people, all the marriages are "stable".

Coupon collector problem

In probability theory, the coupon collector's problem describes the "collect all coupons and win" contests. It asks the following question: Suppose that there n coupons, from which coupons are being collected with replacement. What is the probability that more than t sample trials are needed to collect all n coupons? The mathematical analysis of the problem reveals that the expected number of trials needed grows asO(nlog(n)). For example, when n = 50 it takes about 225 samples to collect all 50 coupons.

Let T be the time to collect all n coupons, and let ti be the time to collect the i-th coupon after i − 1 coupons have been collected. Think of Tand ti as random variables. Observe that the probability of collecting a new coupon given i − 1 coupons is pi = (n − i + 1)/n.
Therefore, ti hasgeometric distribution with expectation 1/pi. By the linearity of expectations we have:

$$\begin{aligned}\mathrm{E}(T) &= \mathrm{E}(t_1) + \mathrm{E}(t_2) + \cdots + \mathrm{E}(t_n) = \frac{1}{p_1} + \frac{1}{p_2} + \cdots + \frac{1}{p_n} \\ &= \frac{n}{n} + \frac{n}{n-1} + \cdots + \frac{n}{1} = n \cdot \left(\frac{1}{1} + \frac{1}{2} + \cdots + \frac{1}{n}\right) \;=\; n \cdot H_n.\end{aligned}$$

Here Hn is the harmonic number. Using the asymptotics of the harmonic numbers, we obtain:

$$\mathrm{E}(T) = n \cdot H_n = n \ln n + \gamma n + \frac{1}{2} + o(1), \quad \text{as } n \to \infty,$$

where $\gamma \approx 0.5772156649$ is the Euler–Mascheroni constant.

Now one can use the Markov inequality to bound the desired probability:

$$\mathrm{P}(T \geq c\, n H_n) \leq \frac{1}{c}.$$

Calculating the variance

Using the independence of random variables ti, we obtain:

$$\begin{aligned}\mathrm{Var}(T) &= \mathrm{Var}(t_1) + \mathrm{Var}(t_2) + \cdots + \mathrm{Var}(t_n) \\ &= \frac{1-p_1}{p_1^2} + \frac{1-p_2}{p_2^2} + \cdots + \frac{1-p_n}{p_n^2} \\ &\leq \frac{n^2}{n^2} + \frac{n^2}{(n-1)^2} + \cdots + \frac{n^2}{1^2} \\ &\leq n^2 \cdot \left(\frac{1}{1^2} + \frac{1}{2^2} + \cdots\right) = \frac{\pi^2}{6} n^2 \leq 2n^2,\end{aligned}$$

where the last equality is the value of the Riemann zeta function known as the Basel problem.

Now one can use the Chebyshev inequality to bound the desired probability:

$$\mathrm{P}\left(|T - nH_n| \geq c\,n\right) \leq \frac{2}{c^2}.$$

Tail estimates

A different upper bound can be derived from the following observation. Let Z_i^r denote the event that the ith coupon was not picked in the first r trials. Then

$$P\left[Z_i^r\right] = \left(1 - \frac{1}{n}\right)^r \leq e^{-r/n}$$

Thus, for r = βnlogn, we have $P\left[Z_i^r\right] \leq e^{(-\beta n \log n)/n} = n^{-\beta}$.

$$P\left[T > \beta n \log n\right] \le P\left[\bigcup_i Z_i^{\beta n \log n}\right] \le n \cdot P[Z_1] \le n^{-\beta+1}$$

The coupon collector's problem can be solved in several different ways. The generating function approach is a combinatorial technique that allows to obtain precise results.

We introduce the probability generating function (PGF) G(z) where [zq]G(z) is the probability that we take q steps to collect the n coupons i.e.T = q and the expectation is given by

$$\mathrm{E}(T) = \left.\frac{\mathrm{d}}{\mathrm{d}z}G(z)\right|_{z=1}.$$

We can calculate G(z) explicitly. We have

$$G(z) = \frac{n}{n}z\frac{1}{1-\frac{1}{n}z}\frac{n-1}{n}z\frac{1}{1-\frac{2}{n}z}\frac{n-2}{n}z\frac{1}{1-\frac{3}{n}z}\frac{n-3}{n}z\cdots\frac{1}{1-\frac{n-1}{n}z}\frac{n-(n-1)}{n}z$$

To see what this means, note that

$$\frac{1}{1-pz} = 1 + pz + p^2z^2 + p^3z^3 + \cdots,$$

so that this is the PGF of an event that has probability p occurring zero or more times, with the exponent of z counting the number of times. We split the sequence of coupons into segments. A new segment begins every time a new coupon is retrieved for the first time. The PGF is the product of the PGFs of the individual segments. Applying this to G(z), we see that it represents the following sequence of events:

- retrieve the first coupon (no restrictions at this time)
- retrieve the first coupon some number of times
- retrieve the second coupon (probability (n − 1) / n))
- retrieve a mix of the first and second coupons some number of times
- retrieve the third coupon (probability (n − 2) / n)
- retrieve a mix of the first, second, and third coupons some number of times
- retrieve the fourth coupon (probability (n − 3) / n)
- . . .
- retrieve the last coupon (probability (n − (n − 1)) / n).
- We simplify G(z) before we compute the expectation, getting

$$G(z) = z^n \frac{n-1}{n-z} \frac{n-2}{n-2z} \frac{n-3}{n-3z} \cdots \frac{n-(n-1)}{n-(n-1)z}.$$

Now we use the fact that

$$\frac{\mathrm{d}}{\mathrm{d}z} \frac{n-k}{n-kz} = \frac{k(n-k)}{(n-kz)^2}$$

to obtain the derivative of G(z)

$$\frac{\mathrm{d}}{\mathrm{d}z} G(z) = G(z) \left(\frac{n}{z} + \frac{1}{n-z} + \frac{2}{n-2z} + \frac{3}{n-3z} \cdots + \frac{n-1}{n-(n-1)z} \right)$$

.

This yields

$$\mathrm{E}(T) = \left. \frac{\mathrm{d}}{\mathrm{d}z} G(z) \right|_{z=1} = G(1) \left(n + \frac{1}{n-1} + \frac{2}{n-2} + \frac{3}{n-3} \cdots + \frac{n-1}{n-(n-1)} \right)$$

or

$$\mathrm{E}(T) = n + \sum_{k=1}^{n-1} \frac{k}{n-k}.$$

Finally, some simplification:

$$\sum_{k=1}^{n-1} \frac{k}{n-k} = \sum_{k=1}^{n-1} \left(\frac{k}{n-k} - \frac{n}{n-k} \right) + nH_{n-1} = nH_{n-1} - (n-1)$$

so that

$$\mathrm{E}(T) = n + nH_{n-1} - (n-1) = nH_{n-1} + 1 = nH_n, \quad \text{QED}.$$

The PGF G(z) makes it possible to obtain an exact value for the variance. Start with

$$\mathrm{Var}(T) = \mathrm{E}(T(T-1)) + \mathrm{E}(T) - \mathrm{E}(T)^2,$$

which consists entirely of factorial moments that can be calculated from the PGF. We already have the value of $\mathrm{E}(T)$. For the remainder, use

$$\mathrm{E}(T(T-1)) = \left. \left(\frac{\mathrm{d}}{\mathrm{d}z} \right)^2 G(z) \right|_{z=1}.$$

The derivative is

$$G(z)\left(\frac{n}{z}+\frac{1}{n-z}+\frac{2}{n-2z}+\frac{3}{n-3z}\cdots+\frac{n-1}{n-(n-1)z}\right)^2$$

$$+\,G(z)\left(-\frac{n}{z^2}+\frac{1^2}{(n-z)^2}+\frac{2^2}{(n-2z)^2}+\frac{3^2}{(n-3z)^2}\cdots+\frac{(n-1)^2}{(n-(n-1)z)^2}\right)$$

Evaluation at z = 1 yields

$$n^2H_n^2-n+\sum_{k=1}^{n-1}\frac{k^2}{(n-k)^2}=n^2H_n^2-n+\sum_{k=1}^{n-1}\frac{(n-k)^2}{k^2}$$

$$=n^2H_n^2-n+n^2H_{n-1}^{(2)}-2nH_{n-1}+(n-1).$$

The conclusion is that

$$\begin{aligned}\mathrm{Var}(T) &= \ n^2H_n^2-1+n^2H_{n-1}^{(2)}-2nH_{n-1}+nH_{n-1}+1-n^2H_n^2\\ &= \ n^2H_{n-1}^{(2)}-nH_{n-1}<\frac{\pi^2}{6}n^2, \quad \text{QED.}\end{aligned}$$