Occupancy Problems

Occupancy problems deal with pairings of objects. The basic occupancy problem is about placing m balls into n bins. This seemingly ordinary problem has a vast number of applications.

Let X_i be the random variable which counts the number of balls in bin i (so X_i is not an indicator random variable). Clearly

$$\sum_{i=1}^{n} X_i \quad = \quad m$$

so $\mathrm{E}[\sum X_i] = n$ and by linearity of expectation, $\mathrm{E}[X_i] = m/n$. If $m = n$ we expect to see one ball in each bin, but how many bins actually have a ball in them? How many have more than one? These questions are more interesting and harder to answer.

First of all, X_i has the Binomial distribution. To see this, let X_{ij} be the indicator random variable for ball j going into bin i, so that $X_i = \sum X_{ij}$ and

$$X_{ij} = \begin{cases} 1 & \text{if ball } j \text{ goes into bin } i \\ 0 & \text{otherwise} \end{cases}$$

Then each X_{ij} represents a Bernoulli trial with probability $p = 1/n$, which is the probability of ball j going into bin i. Since X_i is a sum of Bernoulli trials, it has the binomial distribution. Specifically, it has a distribution of the form:

$$\Pr[X_i = k] \;=\; \binom{m}{k} p^k (1-p)^{m-k} \;=\; \binom{m}{k} \left(\frac{1}{n}\right)^k \left(1 - \frac{1}{n}\right)^{m-k} \tag{1}$$

Approximation for large m, n

If m and n are both large compared to k, the distribution of balls in bins is well-approximated by the Poisson distribution. To see this, start with the binomial distribution form:

$$\Pr[X_i = k] \;=\; \binom{m}{k} p^k (1-p)^{m-k} \;=\; \binom{m}{k} \left(\frac{1}{n}\right)^k \left(1 - \frac{1}{n}\right)^{m-k}$$

We can approximate the final expression above assuming $m, n >> k$ as:

$$\Pr[X_i = k] \approx \frac{m^k}{k!} \left(\frac{1}{n}\right)^k \left(\left(1 - \frac{1}{n}\right)^n\right)^{m/n} \approx \frac{1}{k!} \left(\frac{m}{n}\right)^k e^{-m/n}$$

Which we recognize as the Poisson distribution $\lambda^k e^{-\lambda}/k!$ with $\lambda = m/n$.

In the next lecture and later in the course, we will want to know probabilities over a range of values, e.g. the probability that a running time or subarray size or other random object exceeds a certain value. Many distributions are too complicated to compute such probabilities directly, so we will use a variety of approximation techniques. Lets start this by analyzing the probability that the number of balls in the i^{th} bin exceeds a certain value. Before we proceed, it will help to introduce a very useful approximation for factorials:

Stirling's Formula

The following formula gives a very good approximation for factorials. Notice that it is expressed as an equality with a big-O bound embedded in it. You should treat the $O(1/k^2)$ term as a proxy for a function $f(k)$ which is bounded by a constant times $1/k^2$ for almost all k, which is the usual definition of a big-O bound.

$$k! \quad = \quad \sqrt{2\pi k}\left(\frac{k}{e}\right)^k\left(1+\frac{1}{12k}+O\left(\frac{1}{k^2}\right)\right)$$

Stirling's approximation implies a simpler inequality on $k!$ which is:

$$k! \geq \left(\frac{k}{e}\right)^k$$

and we can substitute this into the formula for $\binom{m}{k}$ to give:

$$\binom{m}{k} \leq \frac{m^k}{k!} \leq \left(\frac{me}{k}\right)^k$$

We can now simplify the binomial probability formula (1) to get a bound on $\Pr[X_i = k]$:

$$\Pr[X_i = k] \leq \left(\frac{me}{k}\right)^k\left(\frac{1}{n}\right)^k\left(1-\frac{1}{n}\right)^{m-k} \leq \left(\frac{me}{kn}\right)^k$$

This gives an upper bound on the probability that bin i contains exactly k balls, and it is simple enough that we can use it in a sum for a range of k values. So we sum over k:

$$\Pr[X_i \geq k] \;\leq\; \sum_{j=k}^{m}\left(\frac{me}{jn}\right)^j \;\leq \sum_{j=k}^{\infty}\left(\frac{me}{kn}\right)^j$$

which is an infinite geometric series. Substituting for the sum of that series gives:

$$\Pr[X_i \geq k] \;\leq\; \left(\frac{me}{kn}\right)^k\left(\frac{1}{1-me/kn}\right) = \left(\frac{\lambda e}{k}\right)^k\left(\frac{1}{1-\lambda e/k}\right) \tag{2}$$

Warning The substitution we just made is only valid if the geometric series converges, and that assumes that $kn > me$, In fact it isnt very accurate until the ratio is somewhat larger than one. If this isnt the case, then k must be "small". For small values of k, we can compute $\Pr[X_i \geq k]$ exactly as $1 - \Pr[X_i < k]$, since $\Pr[X_i < k]$ would be a sum of a small number of terms.

Number of empty bins

Next define

$$Z_i = \begin{cases} 1 \text{ if bin } i \text{ contains } \textit{zero} \text{ balls} \\ 0 \text{ otherwise} \end{cases}$$

Now as we just saw (Poisson approximation),

$$\Pr[Z_i = 1] = \Pr[X_i = 0] \;=\; \left(1 - \frac{1}{n}\right)^m \approx e^{-m/n} = e^{-\lambda}$$

with $\lambda = m/n$ and because Z_i is an indicator r.v., we have that

$$\mathrm{E}[Z_i] = \Pr[Z_i = 1] \approx e^{-\lambda}$$

so we can appeal to linearity of expected value to argue that $Z = \sum Z_i$, the total number of empty bins, satisfies:

$$\mathrm{E}[Z] \;=\; \sum_{i=1}^{n} \mathrm{E}[Z_i] \;\approx\; \sum_{i=1}^{n} e^{-\lambda} = ne^{-\lambda}$$

So for instance, if $m = n$, we have $\lambda = 1$ and about $1/e$ or 37% of the bins are empty on average. With twice as many balls as bins $m = 2n$, the fraction of empty bins is e^{-2} or about 13.5%. While the number of empty bins decreases exponentially with λ, quite a few empty bins remain unless m is significantly larger than n. This is the "coupon collectors problem" that we will study next week.

The Birthday Paradox

The birthday paradox comes from the observation that birthday collisions are likely to happen with relatively few people.

The birthday paradox can be viewed as an occupancy problem. Each of the m people is a ball, assigned independently at random to one of $n = 365$ bins, which are the days of the year. We want to determine the probability of some bin containing two or more balls, and find the range of m for which this probability is high.

Its easier to compute the probability that *no* bin contains two or more balls. We assume that balls are placed one at a time into bins. Let E_i be the event that ball number i goes into an empty bin. Then the probability that no bin contains two balls is equal to the probability that every ball goes into an empty bin, or

$$\Pr[E_1 \wedge \cdots \wedge E_m]$$

We have to proceed carefully because the E_i are not independent. Furthermore, its difficult to compute $\Pr[E_i]$ directly without knowing how many empty bins remain when we place ball i. But what we can compute easily is $\Pr[E_k | E_1 \wedge E_2 \wedge \cdots E_{k-1}]$, the probability that ball k goes into an

empty bin *given* that the earlier balls went into empty bins. And the probability we are looking for has a simple expression in terms of those conditional probabilities:

$$\Pr[E_1 \wedge \cdots \wedge E_m] = \Pr[E_1]\ \Pr[E_2|E_1]\ \Pr[E_3|E_1 \wedge E_2] \cdots \Pr[E_m|E_1 \wedge \cdots \wedge E_{m-1}]$$

You can check this yourself using Bayes rule $\Pr[A|B] = \Pr[A \wedge B]/\Pr[B]$. The product "telescopes" with all but the last numerator cancelling.

Now when the k^{th} ball is placed, if we assume earlier balls went into empty bins there are exactly $n - k + 1$ empty bins left. So the probability that this last ball goes into an empty bin is

$$\Pr[E_k|E_1 \wedge E_2 \wedge \cdots E_{k-1}] = (n - k + 1)/n = 1 - (k - 1)/n$$

So the probability of all m balls going into empty bins is

$$\Pr[E_1 \wedge \cdots \wedge E_m] = 1 \cdot \left(1 - \frac{1}{n}\right)\left(1 - \frac{2}{n}\right) \cdots \left(1 - \frac{m-1}{n}\right) = \prod_{i=1}^{m-1}\left(1 - \frac{i}{n}\right)$$

This is a tricky product to compute. One way to deal with a difficult product is to turn it into a sum. We can do that by introducing the exponential function. In this case, we use the inequality $1 - x \leq e^{-x}$ (check it yourself using calculus). Then

$$\Pr[E_1 \wedge \cdots \wedge E_m] \leq \prod_{i=1}^{m-1} \exp\left(-\frac{i}{n}\right) \ = \ \exp\left(-\frac{1}{n}\sum_{i=1}^{m-1} i\right) \ = \ \exp(-m(m-1)/2n)$$

Now if this probability is small ($<< 1$), then we have a high probability that some bin contains two or more balls. So we want $m(m-1)/2n > 1$ or in other words

$$m \geq \sqrt{2n}$$

So as long as the number of people is greater than the square root of twice the number of days, there is a good chance of a birthday collision. e.g. if the number of people is 40, the probability of a collision is at least $1 - \exp(-2.13) > 0.88$, nearly 90 percent.