Quantiles in the coupon collector problem

The classical "coupon collector" problem can be rephrased as follows. We repeatedly roll a balanced n-sided die, and we let X be the number of rolls needed to see each face at least once. Describe the distribution of the random variable X.

If we just want the expected value of X, the standard trick is to write $X = X_1 + \cdots + X_n$ where X_i is the number of "extra" rolls needed to see the ith "new" face after the $(i-1)$th "new" face has been seen. Then $X_1, X_2, \ldots, X_n$ are independent geometric variables with parameters $\frac{n}{n}, \frac{n-1}{n}, \ldots, \frac{1}{n}$, so we have $\mathbf{E}(X) = n(1+\frac{1}{2}+\cdots+\frac{1}{n-1}+\frac{1}{n}) \approx n \log n$. (Here and throughout, "log" means natural log.)

Suppose we want an expression for

$P = \mathbf{P}(X > m) = \mathbf{P}(\text{at least one face has not been seen in the first } m \text{ rolls}).$

For $i = 1, \ldots, n$, let A_i be the event that face i has not been seen in the first m rolls. Using inclusion-exclusion, we have

$$\begin{aligned} P = \mathbf{P}(A_1 \cup \cdots \cup A_n) &= \mathbf{P}(A_1) + \cdots + \mathbf{P}(A_n) \\ &\quad - \Big(\mathbf{P}(A_1 \cap A_2) + \cdots + \mathbf{P}(A_{n-1} \cap A_n)\Big) \\ &\quad + \cdots \\ &\quad + (-1)^{n-1}\Big(\mathbf{P}(A_1 \cap \cdots \cap A_n)\Big) \\ &= n\Big(1 - \frac{1}{n}\Big)^m \\ &\quad - \binom{n}{2}\Big(1 - \frac{2}{n}\Big)^m \\ &\quad + \cdots \\ &\quad + (-1)^{n-1}\binom{n}{n}\Big(1 - \frac{n}{n}\Big)^m \\ &= \sum_{k=1}^{n} (-1)^{k-1}\binom{n}{k}\Big(1 - \frac{k}{n}\Big)^m. \end{aligned}$$

Now let $m = n\log n + cn$ where c is a constant. In 1961, Erdős and Rényi [1] observed that since

$$\left(1 - \frac{k}{n}\right)^{n(\log n + c)} \approx (e^{-k})^{\log n + c} = \frac{1}{n^k}(e^{-c})^k$$

and since $\binom{n}{k} \approx \frac{n^k}{k!}$, we can say

$$P \approx \sum_{k=1}^{n}(-1)^{k-1}\,\frac{n^k}{k!}\,\frac{1}{n^k}(e^{-c})^k = \sum_{k=1}^{n}(-1)^{k-1}\frac{(e^{-c})^k}{k!} \approx 1 - e^{-e^{-c}}.$$

However, when making this rigorous, the details are nontrivial. We follow the presentation in Section 3.6.3 of Motwani and Raghavan [2].

Lemma. If $0 < k \le k^2 < n$, then

$$e^{-k}\left(1 - \frac{k^2}{n}\right) < \left(1 - \frac{k}{n}\right)^n < e^{-k}.$$

Proof. We start with the series expression

$$\log(1-x) = -x - \frac{x^2}{2} - \frac{x^3}{3} - \frac{x^4}{4} - \cdots$$

which is valid for $-1 < x < 1$. This implies

$$\log\left(1 - \frac{k}{n}\right) = -\frac{k}{n} - \frac{k^2}{2n^2} - \frac{k^3}{3n^3} - \frac{k^4}{4n^4} - \cdots$$

$$\implies n\log\left(1 - \frac{k}{n}\right) = -k - \frac{k^2}{2n} - \frac{k^3}{3n^2} - \frac{k^4}{4n^3} - \cdots$$

so certainly $n\log\left(1 - \frac{k}{n}\right) < -k$, which implies $\left(1 - \frac{k}{n}\right)^n < e^{-k}$. Also,

$$n\log\left(1 - \frac{k}{n}\right) > -k - \frac{k^2}{n} - \frac{k^4}{2n^2} - \frac{k^6}{3n^3} - \cdots$$

$$\implies n\log\left(1 - \frac{k}{n}\right) > -k + \log\left(1 - \frac{k^2}{n}\right)$$

which implies $e^{-k}\left(1 - \frac{k^2}{n}\right) < \left(1 - \frac{k}{n}\right)^n$, completing the proof of the lemma.

Now let ε be any positive real number. There exists a positive integer T such that $\sum_{k=1}^{t}(-1)^{k-1}\frac{(e^{-c})^k}{k!}$ is within $\frac{\varepsilon}{2}$ of $1-e^{-e^{-c}}$ for all $t > T$.

Let $2r-1$ and $2r$ be greater than T. By the Bonferroni inequalities,

$$A := \sum_{k=1}^{2r}(-1)^{k-1}\binom{n}{k}\left(1-\frac{k}{n}\right)^m$$

is an underestimate for P, and

$$B := \sum_{k=1}^{2r-1}(-1)^{k-1}\binom{n}{k}\left(1-\frac{k}{n}\right)^m$$

is an overestimate for P.

Now suppose $n > 4r^2$. By the lemma, for each $k = 1, \ldots, 2r$, we have

$$e^{-k}\left(1-\frac{k^2}{n}\right) < \left(1-\frac{k}{n}\right)^n < e^{-k}$$

$$\implies (e^{-k})^{\log n + c}\cdot\left(1-\frac{k^2}{n}\right)^{\log n + c} < \left(1-\frac{k}{n}\right)^m < (e^{-k})^{\log n + c}$$

$$\implies \frac{1}{n^k}(e^{-c})^k\left(1-\frac{k^2}{n}\right)^{\log n + c} < \left(1-\frac{k}{n}\right)^m < \frac{1}{n^k}(e^{-c})^k$$

$$\implies \binom{n}{k}\frac{1}{n^k}(e^{-c})^k\left(1-\frac{k^2}{n}\right)^{\log n + c} < \binom{n}{k}\left(1-\frac{k}{n}\right)^m < \binom{n}{k}\frac{1}{n^k}(e^{-c})^k.$$

It is straightforward to show $\lim_{n\to\infty}\binom{n}{k}\frac{1}{n^k} = \frac{1}{k!}$ and $\lim_{n\to\infty}\left(1-\frac{k^2}{n}\right)^{\log n} = 1$. It then follows that we have $\lim_{n\to\infty}\binom{n}{k}\left(1-\frac{k}{n}\right)^m = \frac{(e^{-c})^k}{k!}$ for each k.

We now choose n large enough that $\binom{n}{k}\left(1-\frac{k}{n}\right)^m$ is within $\frac{\varepsilon}{4r}$ of $\frac{(e^{-c})^k}{k!}$ for each $k = 1, \ldots, 2r$.

It then follows that A is within $\frac{\varepsilon}{2}$ of $\sum_{k=1}^{2r}(-1)^{k-1}\frac{(e^{-c})^k}{k!}$, and that B is within $\frac{\varepsilon}{2}$ of $\sum_{k=1}^{2r-1}(-1)^{k-1}\frac{(e^{-c})^k}{k!}$.

But then both A and B are within ε of $1-e^{-e^{-c}}$. This completes the rigorous proof that

$$\lim_{n\to\infty} \mathbf{P}(X > n\log n + cn) = 1 - e^{-e^{-c}}.$$

References

[1] P. Erdős and A. Rényi, *On a classical problem of probability theory*, Magyar Tud. Akad. Mat. Kutató Int. Közl. **6** (1961), 215–220.

[2] R. Motwani and P. Raghavan, *Randomized Algorithms*, Cambridge University Press, 1995.