The Coupon Problem

The Problem: There are n different types of coupons and the collector wishes to collect all n coupons. At each trial a coupon is chosen at random. Each coupon is equally likely and the choices are independent. The question is what is the *waiting time* to collect all n coupons?

The Solution: Let T_n denote the random variable defined to be the number of trials required to collect all n coupons. Our first task will be to compute $E(T_n)$. Let $C_1, C_2, \ldots, C_{T_n}$ denote the sequence of trials where $C_i \in \{1, 2, \ldots, n\}$. (The coupons are labelled 1 through n.) The process stops when we have all n different coupons. Define C_i to be a *success* if the type C_i was not drawn in any of the first $i-1$ selections. Clearly, C_1 and C_{T_n} are always successes.

We divide the time interval $[0, T_n]$ into n subintervals of time. Subinterval i is, by definition, the time following the i^{th} success and ends with the trial on which we obtain the $(i+1)^{th}$ success. Let X_i denote the number of trials in the i^{th} subinterval. Thus $X_0 = 1$. Then

$$T_n = \sum_{i=0}^{n-1} X_i \, .$$

The X_i are independent random variables with geometric distribution. The probability of success during the epoch i is

$$p_i = \frac{n-i}{n}, \;\; i = 0, 1, \ldots, n-1.$$

We know that if X is a random variable with geometric distribution then

$$E(X) = \frac{1}{p}$$

where p is the probability of success. Thus

$$\begin{aligned} E(T_n) &= \sum_{i=0}^{n-1} E(X_i) \\ &= \sum_{i=0}^{n-1} \frac{n}{n-i} \\ &= n \sum_{i=1}^{n} \frac{1}{i} \\ &= nH_n. \end{aligned}$$

Here H_n is the harmonic series

$$H_n := \sum_{i=1}^{n} \frac{1}{i}\,.$$

In calculus it is proved that for large n

$$H_n = \log n + \gamma + \frac{1}{2n} + \mathrm{O}(\frac{1}{n^2})$$

where $\gamma = 0.57721\ldots$ is Euler's constant. Thus for $n \to \infty$

$$E(T_n) = n \log n + \gamma n + \frac{1}{2} + \mathrm{O}(\frac{1}{n})\,.$$

For example, if $n = 100$ an exact evaluation gives $100\,H_{100} = 518.738$ the asymptotic expansion gives 518.739.

We can also calculate the variance of T_n. We first recall that if X is a random variable with geometric distribution that

$$\mathrm{Var}(X) = \frac{q}{p^2}\,.$$

This is easy to prove once one establishes

$$\sum_{n=1}^{\infty} n^2 q^{n-1} = \frac{1+q}{(1-q)^3}\,.$$

Thus

$$\mathrm{Var}(X_i) = \frac{q_i}{p_i^2} = (\frac{i}{n})\,\frac{n^2}{(n-i)^2} = \frac{ni}{(n-i)^2}\,.$$

Since the X_i are independent,

$$\begin{aligned}
\mathrm{Var}(T_n) &= \sum_{i=0}^{n-1} \mathrm{Var}(X_i) \\
&= \sum_{i=0}^{n-1} \frac{ni}{(n-i)^2} \\
&= \sum_{i=1}^{n} \frac{n(n-i)}{i^2} \\
&= n^2 \sum_{i=1}^{n} \frac{1}{i^2} - nH_n.
\end{aligned}$$

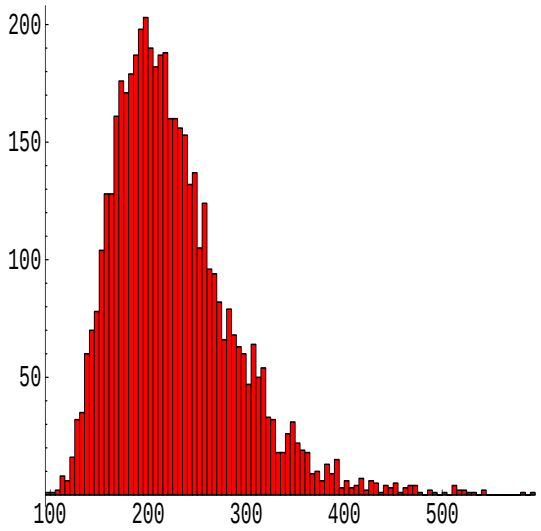

Figure 1: Histogram of waiting times for $n = 50$ and 5000 simulations.

Since

$$\sum_{i=1}^{\infty} \frac{1}{i^2} = \frac{\pi^2}{6}\,,$$

we have

$$\mathrm{Var}(T_n) = \frac{\pi^2}{6}\, n^2 - n \log n + \mathrm{O}(n),$$

and hence

$$\sigma_{T_n} \sim \frac{\pi}{\sqrt{6}}\, n\,.$$

This is a large variance so we can expect to see considerable fluctuations from the mean $E(T_n)$. We turn to simulations to see what is happening. For $n = 50$ (number of coupons) a simulation was run 5000 times. The maximum waiting time observed was 592 and the minimum waiting time observed was 99. The average of the 5000 waiting times was 225.161. This should be compared with the theoretical result $E(T_{50}) = 50\, H_{50} = 224.96$. The standard deviation is 61.95. In Figure 1 a histogram of the 5000 waiting times is shown.

A second set of simulations were performed this time for $n = 100$ (number of coupons) and $10,000$ simulations. Here the maximum waiting time observed

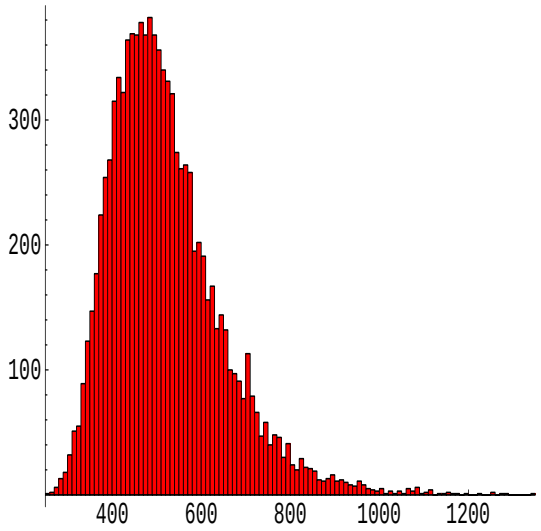

Figure 2: Histogram of waiting times for $n = 100$ and $10,000$ simulations.

was 1346, the minimum waiting time was 250 with an average waiting time of 519.73. This last number compares well with the theoretical result $100H_{100} = 518.74$. The standard deviation was 125.8. In Figure 2 a histogram of the $10,000$ observed waiting times is shown. The distribution in both cases is asymmetric thus ruling out a gaussian distribution.

Here is the Mathematica program that produced the simulation of one waiting time.

```
(* Coupon Problem -- simulates the time required to collect n coupons  *)


waitingTime[noCoupons_]:=Module[{couponList={},i,remainingCoupons,
coupon,j,time=0},
For[i=1,i<=noCoupons,i++,
couponList=Append[couponList,i]];
remainingCoupons=couponList;
While[remainingCoupons!={},
coupon=Random[Integer,{1,noCoupons}];time++;
If[MemberQ[remainingCoupons,coupon]==True,
j=Position[remainingCoupons,coupon];
remainingCoupons=Delete[remainingCoupons,j]]
];time];


mean[n_]:=N[n*Sum[1/i,{i,1,n}]];

stdDev[n_]:=N[n*Sum[(n-i)/i^2,{i,1,n}]];
```