12.7: The Coupon Collector Problem

Basic Theory

Definitions

In this section, our random experiment is to sample repeatedly, *with* replacement, from the population $D = \{1, 2, \ldots, m\}$. This generates a sequence of independent random variables $\boldsymbol{X} = (X_1, X_2, \ldots)$, each uniformly distributed on D

We will often interpret the sampling in terms of a *coupon collector*: each time the collector buys a certain product (bubble gum or Cracker Jack, for example) she receives a coupon (a baseball card or a toy, for example) which is equally likely to be any one of m types. Thus, in this setting, $X_i \in D$ is the coupon type received on the ith purchase.

Let V_n denote the number of distinct values in the first n selections, for $n \in \mathbb{N}_+$. This is the random variable studied in the last section on the Birthday Problem. Our interest is in this section is the sample size needed to get a specified number of distinct sample values

For $k \in \{1, 2, \ldots, m\}$, let

$$W_k = \min\{n \in \mathbb{N}_+ : V_n = k\} \tag{12.7.1}$$

the sample size needed to get k distinct sample values.

In terms of the coupon collector, this random variable gives the number of products required to get k distinct coupon types. Note that the set of possible values of W_k is $\{k, k+1, \ldots\}$. We will be particularly interested in W_m, the sample size needed to get the entire population. In terms of the coupon collector, this is the number of products required to get the entire set of coupons.

In the coupon collector experiment, run the experiment in single-step mode a few times for selected values of the parameters.

The Probability Density Function

Now let's find the distribution of W_k. The results of the previous section will be very helpful

For $k \in \{1, 2, \ldots, m\}$, the probability density function of W_k is given by

$$\mathbb{P}(W_k = n) = \binom{m-1}{k-1} \sum_{j=0}^{k-1} (-1)^j \binom{k-1}{j} \left(\frac{k-j-1}{m}\right)^{n-1}, \quad n \in \{k, k+1, \ldots\} \tag{12.7.2}$$

Proof

Note first that $W_k = n$ if and only if $V_{n-1} = k-1$ and $V_n = k$. Hence

$$\mathbb{P}(W_k = n) = \mathbb{P}(V_{n-1} = k-1)\mathbb{P}(V_n = k \mid V_{n-1} = k-1) = \frac{m-k+1}{m}\mathbb{P}(V_{n-1} = k-1) \tag{12.7.3}$$

Using the PDF of V_{n-1} from the previous section gives the result.

In the coupon collector experiment, vary the parameters and note the shape of and position of the probability density function. For selected values of the parameters, run the experiment 1000 times and compare the relative frequency function to the probability density function.

An alternate approach to the probability density function of W_k is via a recursion formula.

For fixed m, let g_k denote the probability density function of W_k. Then

1. $g_k(n+1) = \frac{k-1}{m} g_k(n) + \frac{m-k+1}{m} g_{k-1}(n)$
2. $g_1(1) = 1$

Decomposition as a Sum

We will now show that W_k can be decomposed as a sum of k independent, geometrically distributed random variables. This will provide some additional insight into the nature of the distribution and will make the computation of the mean and variance easy.

For $i \in \{1, 2, \ldots, m\}$, let Z_i denote the number of additional samples needed to go from $i-1$ distinct values to i distinct values. Then $\boldsymbol{Z} = (Z_1, Z_2, \ldots, Z_m)$ is a sequence of independent random variables, and Z_i has the geometric distribution on $\mathbb{N}_+$ with parameter $p_i = \frac{m-i+1}{m}$. Moreover,

$$W_k = \sum_{i=1}^{k} Z_i, \quad k \in \{1, 2, \ldots, m\} \tag{12.7.4}$$

This result shows clearly that each time a new coupon is obtained, it becomes harder to get the next new coupon.

In the coupon collector experiment, run the experiment in single-step mode a few times for selected values of the parameters. In particular, try this with m large and k near m.

Moments

The decomposition as a sum of independent variables provides an easy way to compute the mean and other moments of W_k.

The mean and variance of the sample size needed to get k distinct values are

1. $\mathbb{E}(W_k) = \sum_{i=1}^{k} \frac{m}{m-i+1}$
2. $\var(W_k) = \sum_{i=1}^{k} \frac{(i-1)m}{(m-i+1)^2}$

Proof

These results follow from the decomposition of W_k as a sum of independent variables and standard results for the geometric distribution, since $\mathbb{E}(W_k) = \sum_{i=1}^{k} \mathbb{E}(Z_i)$ and $\var(W_k) = \sum_{i=1}^{k} \var(Z_i)$.

In the coupon collector experiment, vary the parameters and note the shape and location of the mean $\pm$ standard deviation bar. For selected values of the parameters, run the experiment 1000 times and compare the sample mean and standard deviation to the distribution mean and standard deviation.

The probability generating function of W_k is given by

$$\mathbb{E}\left(t^{W_k}\right) = \prod_{i=1}^{k} \frac{m-i+1}{m-(i-1)t}, \quad |t| < \frac{m}{k-1} \tag{12.7.5}$$

Proof

This follows from the decomposition of W_k as a sum of independent variables and standard results for the geometric distribution on $\mathbb{N}_+$, since $\mathbb{E}\left(t^{W_k}\right) = \prod_{i=1}^{k} \mathbb{E}\left(t^{Z_i}\right)$.

Examples and Applications

Suppose that people are sampled at random until 40 distinct birthdays are obtained. Find each of the following:

1. The probability density function of the sample size.
2. The mean of the sample size.
3. The variance of the sample size.
4. The probability generating function of the sample size.

Answer

Let W denote the sample size.

1. $\mathbb{P}(W = n) = \binom{364}{n} \sum_{j=0}^{30} (-1)^j \binom{39}{j} \left(\frac{39-j}{365}\right)^{n-1}$ for $n \in \{40, 41, \ldots\}$

2. $\mathbb{E}(W) = 42.3049$
3. $\text{var}(W) = 2.4878$
4. $\mathbb{E}\left(t^W\right) = \prod_{i=1}^{40} \frac{366-i}{365-(i-1)t}$ for $|t| < \frac{365}{39}$

Suppose that a standard, fair die is thrown until all 6 scores have occurred. Find each of the following:

1. The probability density function of the number of throws.
2. The mean of the number of throws.
3. The variance of the number of throws.
4. The probability that at least 10 throws are required.

Answer

Let W denote the number of throws.

1. $\mathbb{P}(W = n) = \sum_{j=0}^{5} (-1)^j \binom{5}{j} \left(\frac{5-j}{6}\right)^{n-1}$ for $n \in \{6, 7, \ldots\}$
2. $\mathbb{E}(W) = 14.7$
3. $\text{var}(W) = 38.99$
4. $\mathbb{P}(W \geq 10) = \frac{1051}{1296} \approx 0.81096$

A box of a certain brand of cereal comes with a special toy. There are 10 different toys in all. A collector buys boxes of cereal until she has all 10 toys. Find each of the following:

1. The probability density function of the number boxes purchased.
2. The mean of the number of boxes purchased.
3. The variance of the number of boxes purchased.
4. The probability that no more than 15 boxes were purchased.

Answer

Let W denote the number of boxes purchased.

1. $\mathbb{P}(W = n) = \sum_{j=0}^{9} (-1)^j \binom{9}{j} \left(\frac{9-j}{10}\right)^{n-1}$, for $n \in \{10, 11, \ldots\}$
2. $\mathbb{E}(W) = 29.2897$
3. $\text{var}(W) = 125.6871$
4. $\mathbb{P}(W \leq 15) = 0.04595$