

7. The Coupon Collector Problem

Basic Theory

Random Variables

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

$$\mathbf{X} = (X_1, X_2, \dots)$$

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 i^{th} purchase.

Let $V_{m,n}$ 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 in this section is the sample size needed to get k distinct sample values, for $k \in \{1, 2, \dots, m\}$. Thus, let

$$W_{m,k} = \min \{n \in \mathbb{N}_+ : V_{m,n} = k\}$$

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 $W_{m,k}$ is $\{k, k+1, \dots\}$. We will be particularly interested in $W_{m,m}$, 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.

1. 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 $W_{m,k}$. The results of the previous section will be very helpful

2. Argue that $W_{m,k} = n$ if and only if $V_{m,n-1} = k-1$ and $V_{m,n} = k$.

3. Use Exercise 2 and a conditional probability argument to show that

$$\mathbb{P}(W_{m,k} = n) = \frac{m-k+1}{m} \mathbb{P}(V_{m,n-1} = k-1)$$

4. Use the result of the last exercise and the distribution of $V_{m,n-1}$ from the last section to show that

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

5. 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 with an update frequency of 10 and note the apparent convergence of the relative frequency function to the probability density function.

Recursion Formula

An alternate approach to the distribution of the sample size needed to get k distinct values is via a recursion formula.

6. Let $p_{m,k}(n) = \mathbb{P}(W_{m,k} = n)$ Use a conditional probability argument to show that

$$p_{m,k}(n+1) = \frac{k-1}{m} p_{m,k}(n) + \frac{m-k+1}{m} p_{m,k-1}(n)$$

Decomposition as a Sum

We will now show that $W_{m,k}$ 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 $i \in \{1, 2, \dots, m\}$, let $Z_{m,i}$ denote the number of additional samples needed to go from $i-1$ distinct values to i distinct values.

7. Argue that

- $(Z_{m,1}, Z_{m,2}, \dots, Z_{m,m})$ is a sequence of independent random variables.
- $Z_{m,i}$ has the geometric distribution on \mathbb{N}_+ with parameter $p_{m,i} = \frac{m-i+1}{m}$
- $W_{m,k} = \sum_{i=1}^k Z_{m,i}$

Exercise 7 shows clearly that each time a new coupon is obtained, it becomes harder to get the next new coupon.

8. 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

9. Use the results of **Exercise 7** to show that

- a. $\mathbb{E}(W_{m,k}) = \sum_{i=1}^k \frac{m}{m-i+1}$
- b. $\text{var}(W_{m,k}) = \sum_{i=1}^k \frac{(i-1)m}{(m-i+1)^2}$

10. In the **coupon collector experiment**, vary the parameters and note the shape and location of the mean/standard deviation bar. For selected values of the parameters, run the experiment 1000 times with an update frequency of 10 and note the apparent convergence of the sample mean and standard deviation to the distribution mean and standard deviation.

11. Use the Result of [Exercise 7](#) to show that the probability generating function of $W_{m,k}$ is

$$\mathbb{E}(t^{W_{m,k}}) = \prod_{i=1}^k \frac{m - (i-1)}{m - (i-1)t}, \quad |t| < \frac{m}{k-1}$$

Examples and Applications

12. Suppose that people are sampled at random until 40 distinct birthdays are obtained.

- Find the probability density function of the sample size.
- Find the mean of the sample size.
- Find the variance of the sample size.
- Find the probability generating function of the sample size.



13. Suppose that a standard, fair die is thrown until all 6 scores have occurred.

- Find the probability density function of the number of throws.
- Find the mean of the number of throws.
- Find the variance of the number of throws.
- Find the probability that at least 10 throws are required.



14. 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 the probability density function of the number boxes purchased.
- Find the mean of the number of boxes purchased.
- Find the variance of the number of boxes purchased.
- Find the probability that no more than 15 boxes were purchased.

