

6. The Birthday Problem

Introduction

The Sampling Model

As in the [basic sampling model](#), suppose that we select n numbers at random, *with* replacement, from the population $D = \{1, 2, \dots, m\}$. Thus, our outcome vector is

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

where $X_i \in D$ is the i^{th} number chosen. Recall that our basic modeling assumption is that \mathbf{X} is uniformly distributed on the sample space

$$S = D^n = \{1, 2, \dots, m\}^n$$

In this section, we are interested in the number of population values missing from the sample, and the number of (distinct) population values in the sample. The computation of probabilities related to these random variables are generally referred to as **birthday problems**. Often, we will interpret the sampling experiment as a distribution of n balls into m cells; X_i is the cell number of ball i . In this interpretation, our interest is in the number of empty cells and the number of occupied cells.

Multinomial Distribution

For $i \in D$, let $Y_{n,i}$ denote the number of times that i occurs in the sample:

$$Y_{n,i} = \#\{j \in \{1, 2, \dots, n\} : X_j = i\}$$

1. Show that $\mathbf{Y}_n = (Y_{n,1}, Y_{n,2}, \dots, Y_{n,m})$ has the [multinomial distribution](#) with parameters n and

$$\left(\frac{1}{m}, \frac{1}{m}, \dots, \frac{1}{m}\right):$$

$$\mathbb{P}(Y_{n,1} = k_1, Y_{n,2} = k_2, \dots, Y_{n,m} = k_m) = \binom{n}{k_1, k_2, \dots, k_m} \frac{1}{m^n} \quad \text{for } (k_1, k_2, \dots, k_m) \in \mathbb{N}^m \text{ with } \sum_{i=1}^m k_i = n$$

Random Variables

We will now define the main random variables of interest: the number of population values missing in the sample:

$$U_{m,n} = \#\{j \in \{1, 2, \dots, m\} : Y_{n,j} = 0\}$$

and the number of (distinct) population values that occur in the sample:

$$V_{m,n} = \#\left(\{j \in \{1, 2, \dots, m\} : Y_{n,j} > 0\}\right)$$

Clearly we must have $U_{m,n} + V_{m,n} = m$ so once we have the probability distribution and moments of one variable, we can easily find them for the other variable. However, we will first solve the simplest version of the birthday problem.

The Simple Birthday Problem

The event that there is at least one duplication in the sample can be written as

$$B_{m,n} = \{V_{m,n} < n\} = \{U_{m,n} > m - n\}$$

The (simple) **birthday problem** is to compute the probability of this event. For example, suppose that we choose n people at random and note their birthdays. If we ignore leap years and assume that birthdays are uniformly distributed throughout the year, then our sampling model applies with $m = 365$. In this setting, the birthday problem is to compute the probability that at least two people have the same birthday (this special case is the origin of the name).

The solution of the birthday problem is an easy exercise in combinatorial probability.

2. Use the [multiplication rule of combinatorics](#) to show that

$$\mathbb{P}(B_{m,n}) = \begin{cases} 1 - \frac{m^{(n)}}{m^n}, & n \leq m \\ 1, & n > m \end{cases}$$

Hint: The complementary event $S \setminus B_{m,n}$ occurs if and only if the outcome vector X forms a permutation of size n from $D = \{1, 2, \dots, m\}$.

The fact that the probability is 1 for $n > m$ is sometimes referred to as the **pigeonhole principle**: if more than m pigeons are placed into m holes then at least one hole has 2 or more pigeons.

A Recurrence Relation

3. Let $p_{m,n}$ denote the probability of the complementary event, $S \setminus B_{m,n}$, that the sample variables are distinct. Prove the following recursion relation in two ways: first starting with the result in [Exercise 2](#), and then by using a conditional probability argument.

a. $p_{m,1} = 1$

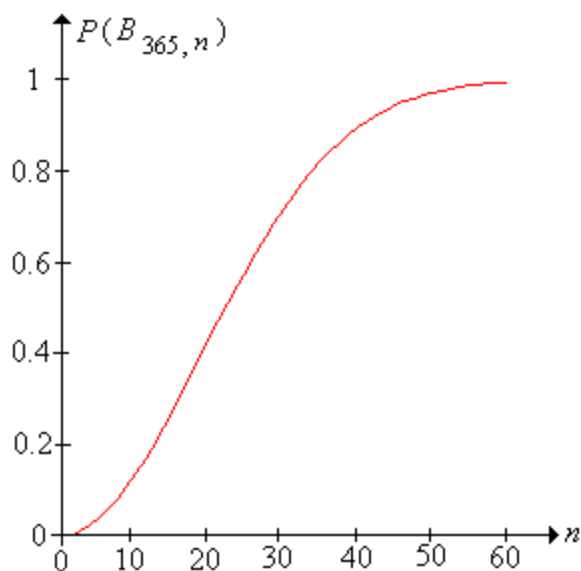
b. $p_{m,n+1} = \frac{m-n}{m} p_{m,n}$

Examples

4. Let $m = 365$ (the standard birthday problem). Verify the following birthday probabilities:

- $\mathbb{P}(B_{365,10}) = 0.117$
- $\mathbb{P}(B_{365,20}) = 0.411$
- $\mathbb{P}(B_{365,30}) = 0.706$
- $\mathbb{P}(B_{365,40}) = 0.891$
- $\mathbb{P}(B_{365,50}) = 0.970$
- $\mathbb{P}(B_{365,60}) = 0.994$

5. Graph the values in the previous exercise as a function of n . When smoothed (for the sake of appearance), your curve should look like the graph below.



6. In the **birthday experiment**, set $m = 365$, and select the indicator variable I . For $n \in \{10, 20, 30, 40, 50, 60\}$ run the experiment 1000 times each and compute the relative frequency of the event that the sample contains a duplication. Compare the relative frequencies with the probabilities computed in the previous exercise.

In spite of its easy solution, the birthday problem is famous because, numerically, the probabilities can be a bit surprising. Note that with a just 60 people, the event is almost certain! Mathematically, the rapid increase in the birthday probability, as n increases, is due to the fact that m^n grows much faster than $m^{(n)}$.

7. Four fair, standard dice are rolled. Find the probability that the scores are distinct.



8. In the **birthday experiment**, set $m = 6$ and select the indicator variable I . Vary n with the scrollbar and note graphically how the probabilities change. Now with $n = 4$, run the experiment 1000 times with an update frequency of 10. Note the apparent convergence of the relative frequency of the event to the

corresponding probability.

9. Five persons are chosen at random.

- Find the probability that at least 2 have the same birth *month*.
- Criticize the sampling model in this setting



10. In the **birthday experiment**, set $m = 12$ and select the indicator variable I . Vary n with the scrollbar and note graphically how the probabilities change. Now with $n = 5$, run the experiment 1000 times with an update frequency of 10. Note the apparent convergence of the relative frequency of the event to the corresponding probability.

11. A fast-food restaurant gives away one of 10 different toys with the purchase of a kid's meal. A family with 5 children buys 5 kid's meals. Find the probability that the 5 toys are different.



12. In the **birthday experiment**, set $m = 10$ and select the indicator variable I . Vary n with the scrollbar and note graphically how the probabilities change. Now with $n = 5$, run the experiment 1000 times with an update frequency of 10. Note the apparent convergence of the relative frequency of the event to the corresponding probability.

13. Let $m = 52$. Find the smallest value of n such that the probability of a duplication is at least $\frac{1}{2}$.



The General Birthday Problem

We now return to the more general problem of finding the distribution of the number of distinct sample values and the distribution of the number of excluded sample values.

The Probability Density Function

For $j \in D$, consider the event that j does not occur in the sample: $A_{n,j} = \{Y_{n,j} = 0\}$. Now let $K \subseteq D$ with $\#(K) = k$. Using the **multiplication rule of combinatorics**, it is easy to count the number of samples that do not contain any elements of K :

14. Show that

$$\#(\bigcap_{j \in K} A_{n,j}) = (m - k)^n$$

Now the **inclusion-exclusion rule of combinatorics** can be used to count the number samples that are missing at least one population value:

15. Show that

$$\#(\bigcup_{j=1}^m A_{n,j}) = \sum_{k=1}^m (-1)^{k-1} \binom{m}{k} (m - k)^n$$

Once we have this, it's simple to count the number samples that contain all population values:

16. Show that

$$\#(\bigcap_{j=1}^m A_{n,j}^c) = \sum_{k=0}^m (-1)^k \binom{m}{k} (m-k)^n$$

Now we can use a two-step procedure to generate all samples that exclude exactly j population values:

1. First, choose the j values that are to be excluded.
2. Then select a sample of size n from the remaining population values so that none are excluded.

Thus, we can use the multiplication principle of combinatorics to count the number of samples that exclude j population values.

17. Show that

$$\#(\{U_{m,n} = j\}) = \binom{n}{j} \sum_{k=0}^{m-j} (-1)^k \binom{m-j}{k} (m-j-k)^n$$

Finally, since the probability distribution of X on the sample space S is uniform, we can find the probability density function of the number of excluded values:

18. Show that

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

We can now easily find the probability density function of the number distinct values in the sample:

19. Show that

$$\mathbb{P}(V_{m,n} = j) = \binom{n}{j} \sum_{k=0}^j (-1)^k \binom{j}{k} \left(\frac{j-k}{m}\right)^n, \quad j \in \{1, 2, \dots, \min\{m, n\}\}$$

20. In the **birthday experiment**, select the number of distinct sample values. Vary the parameters and note the shape and location of the probability density function. For selected values of the parameters, run the simulation 1000 times updating every 10 runs and note the apparent convergence of the relative frequency function to the probability density function.

A Recurrence Relation

The distribution of the number of excluded values can also be obtained by a recursion argument.

21. Let $p_{m,n}(j) = \mathbb{P}(U_{m,n} = j)$ for $j \in \{\max\{m-n, 0\}, \dots, m-1\}$. Use probability arguments to show that

- a. $p_{m,1}(m-1) = 1$

$$b. p_{m,n+1}(j) = \frac{m-j}{m} p_{m,n}(j) + \frac{j+1}{m} p_{m,n}(j+1)$$

Moments

Now we will find the [means](#) and [variances](#). The number of excluded values and the number of distinct values are counting variables and hence can be written as sums of indicator variables. As we have seen in many other models, this representation is frequently the best for computing moments.

Let $I_{n,j} = \mathbf{1}(A_{n,j})$. Thus, $I_{n,j} = 1$ if $A_{n,j}$ occurs, which means that j is not in the sample, and $I_{n,j} = 0$ otherwise. Note that the number of population values missing in the sample can be written as the sum of the indicator variables:

$$U_{m,n} = \sum_{j=1}^m I_{n,j}$$

22. Show that

$$a. \mathbb{E}(I_{n,j}) = \left(1 - \frac{1}{m}\right)^n \text{ for } j \in \{1, 2, \dots, m\}$$

$$b. \mathbb{E}(I_{n,i} I_{n,j}) = \left(1 - \frac{2}{m}\right)^n \text{ for } (i, j) \in \{1, 2, \dots, m\}^2 \text{ with } i \neq j$$

23. Use the results of [Exercise 22](#) to show that

$$a. \mathbb{E}(U_{m,n}) = m \left(1 - \frac{1}{m}\right)^n$$

$$b. \mathbb{E}(V_{m,n}) = m \left(1 - \left(1 - \frac{1}{m}\right)^n\right)$$

24. Use the result of [Exercise 22](#) to show that

$$a. \text{var}(I_{n,j}) = \left(1 - \frac{1}{m}\right)^n - \left(1 - \frac{1}{m}\right)^{2n} \text{ for } j \in \{1, 2, \dots, m\}$$

$$b. \text{cov}(I_{n,i}, I_{n,j}) = \left(1 - \frac{2}{m}\right)^n - \left(1 - \frac{1}{m}\right)^{2n} \text{ for } (i, j) \in \{1, 2, \dots, m\}^2 \text{ with } i \neq j$$

25. Use the results of the [Exercise 24](#), and basic properties of variance to show that

$$\text{var}(U_{m,n}) = \text{var}(V_{m,n}) = m(m-1) \left(1 - \frac{2}{m}\right)^n + m \left(1 - \frac{1}{m}\right)^n - m^2 \left(1 - \frac{1}{m}\right)^{2n}$$

26. In the [birthday experiment](#), select the number of distinct sample values. Vary the parameters and note the size and location of the mean/standard-deviation bar. For selected values of the parameters, run the simulation 1000 times updating every 10 runs and note the apparent convergence of the sample mean and variance to the distribution mean and variance.

Examples and Applications

27. Suppose that 30 persons are chosen at random.

- Find the probability density function of the number of distinct birthdays.
- Find the mean of the number of distinct birthdays.
- Find the variance of the number of distinct birthdays.
- Find the probability that there are at least 28 different birthdays represented.



28. In the **birthday experiment**, set $m = 365$. and $n = 30$, run the experiment 1000 times with an update frequency of 10 and compute the relative frequency of the event in part (d) of the last exercise.

29. Suppose that 10 fair dice are rolled.

- Find the probability density function of the number of distinct scores.
- Find the mean of the number of distinct scores.
- Find the variance of the number of distinct scores.
- Find the probability that there will 4 or fewer distinct scores.



30. In the **birthday experiment**, set $m = 6$ and $n = 10$, run the experiment 1000 times with an update frequency of 10 and compute the relative frequency of the event in part (d) of the last exercise.

31. A fast food restaurant gives away one of 10 different types of toy with the purchase of each kid's meal. A family buys 15 kid's meals.

- Find the probability density function of the number of toy types that are missing.
- Find the mean of the number of toy types that are missing.
- Find the variance of the number of toy types that are missing.
- Find the probability that at least 3 toy types are missing.



32. In the **birthday experiment**, set $m = 10$ and $n = 15$, run the experiment 1000 times with an update frequency of 10 and compute the relative frequency of the event in part (d).

33. **The lying students problem**. Suppose that 3 students, who ride together, miss a mathematics exam. They decide to lie to the instructor by saying that the car had a flat tire. The instructor separates the students and asks each of them which tire was flat. The students, who did not anticipate this, select their answers independently and at random.

- Give the probability density function of the number of distinct answers.
- In particular, give the probability that the students get away with their deception.
- Give mean of the number of distinct answers.
- Give the standard deviation of the number of distinct answers.



34. **The duck hunter problem.** Suppose that there are 5 duck hunters, each a perfect shot. A flock of 10 ducks fly over, and each hunter selects one duck at random and shoots.
- Give the probability density function of the number of ducks that are killed.
 - Give mean of the number of ducks that are killed.
 - Give the standard deviation of the number of ducks that are killed..



[Virtual Laboratories](#) > [12. Finite Sampling Models](#) > [1](#) [2](#) [3](#) [4](#) [5](#) **[6](#)** [7](#) [8](#) [9](#)

[Contents](#) | [Applets](#) | [Data Sets](#) | [Biographies](#) | [External Resources](#) | [Keywords](#) | [Feedback](#) | ©