12.6: 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 $D = \{1, 2, \ldots, m\}$. Thus, our outcome vector is $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ where X_i is the ith number chosen. Recall that our basic modeling assumption is that $\boldsymbol{X}$ is uniformly distributed on the sample space $S = D^n = \{1, 2, \ldots, m\}^n$

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.

For $i \in D$, let Y_i denote the number of times that i occurs in the sample:

$$Y_i = \#\{j \in \{1, 2, \ldots, n\} : X_j = i\} = \sum_{j=1}^{n} \mathbf{1}(X_j = i) \tag{12.6.1}$$

$\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_m)$ has the multinomial distribution with parameters n and $(1/m, 1/m, \ldots, 1/m)$

$$\mathbb{P}(Y_1 = y_1, Y_2 = y_2, \ldots, Y_m = y_m) = \binom{n}{y_1, y_2, \ldots, y_n} \frac{1}{m^n}, \quad (y_1, y_2, \ldots, y_m) \in \mathbb{N}^n \text{ with } \sum_{i=1}^{m} y_i = n \tag{12.6.2}$$

Proof

This follows immediately from the definition of the multinomial distribution, since $(X_1, X_2, \ldots, X_n)$ is an independent sequence, and X_i is uniformly distributed on $\{1, 2, \ldots, m\}$for each i.

We will now define the main random variables of interest.

The number of population values missing in the sample is

$$U = \#\{j \in \{1, 2, \ldots, m\} : Y_j = 0\} = \sum_{j=1}^{m} \mathbf{1}(Y_j = 0) \tag{12.6.3}$$

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

$$V = \#\{j \in \{1, 2, \ldots, m\} : Y_j > 0\} = \sum_{j=1}^{m} \mathbf{1}(Y_j > 0) \tag{12.6.4}$$

Also, U takes values in $\{\max\{m - n, 0\}, \ldots, m - 1\}$ and V takes values in $\{1, 2, \ldots, \min\{m, n\}\}$

Clearly we must have $U + V = m$ 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 when a sample of size n is chosen from a population of size m is

$$B_{m,n} = \{V < n\} = \{U > m - n\} \tag{12.6.5}$$

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.

The probability of the birthday event is

$$\mathbb{P}(B_{m,n}) = 1 - \frac{m^{(n)}}{m^n}, \quad n \leq m \tag{12.6.6}$$

and $P(B_{m,n}) = 1$ for $n > m$

Proof

The complementary event B^c occurs if and only if the outcome vector $\boldsymbol{X}$ forms a permutation of size n from $\{1, 2, \ldots, m\}$. The number of permutations is $m^{(n)}$ and of course the number of samples is m^n.

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. The following result gives a recurrence relation for the probability of distinct sample values and thus gives another way to compute the birthday probability.

Let $p_{m,n}$ denote the probability of the complementary birthday event B^c, that the sample variables are distinct, with population size m and sample size n. Then $p_{m,n}$ satisfies the following recursion relation and initial condition:

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

Examples

Let $m = 365$ (the standard birthday problem).

1. $\mathbb{P}(B_{365,10}) = 0.117$
2. $\mathbb{P}(B_{365,20}) = 0.411$
3. $\mathbb{P}(B_{365,30}) = 0.706$
4. $\mathbb{P}(B_{365,40}) = 0.891$
5. $\mathbb{P}(B_{365,50}) = 0.970$
6. $\mathbb{P}(B_{365,60}) = 0.994$

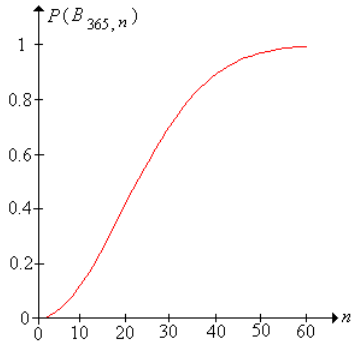

Figure 12.6.1: $P(B_{365,n})$ as a function of n, smoothed for the sake of appearance

In the birthday experiment, set $n = 365$ and select the indicator variable I. For $n \in \{10, 20, 30, 40, 50, 60\}$ run the experiment 1000 times each and compare the relative frequencies with the true probabilities.

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! With just 23 people, the birthday event is about $\frac{1}{2}$; specifically $\mathbb{P}(B_{365,23}) = 0.507$. Mathematically, the rapid increase in the birthday probability, as n increases, is due to the fact that m^n grows much faster than $m^{(n)}$.

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

Answer

$\frac{5}{18}$

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 and compare the relative frequency of the event to the corresponding probability.

Five persons are chosen at random.

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

Answer

1. $\frac{89}{144}$
2. The number of days in a month varies, so the assumption that a person's birth month is uniformly distributed over the 12 months not quite accurate.

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 and compare the relative frequency of the event to the corresponding probability.

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.

Answer

$\frac{189}{625}$

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 and comparethe relative frequency of the event to the corresponding probability.

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

Answer

$n = 9$

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

The number of samples with exactly j values excluded is

$$\#\{U = j\} = \binom{m}{j} \sum_{k=0}^{m-j} (-1)^k \binom{m-j}{k} (m-j-k)^n, \quad j \in \{\max\{m-n, 0\}, \ldots, m-1\} \tag{12.6.7}$$

Proof

For $i \in D$, consider the event that i does not occur in the sample: $A_i = \{Y_i = 0\}$. 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:

$$\#\left(\bigcap_{i \in K} A_i\right) = (m-k)^n \tag{12.6.8}$$

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

$$\#\left(\bigcup_{i=1}^m A_i\right) = \sum_{k=1}^m (-1)^{k-1} \binom{m}{k} (m-k)^n \tag{12.6.9}$$

Once we have this, we can use DeMorgan's law to count the number samples that contain all population values:

$$\#\left(\bigcap_{i=1}^m A_i^c\right) = \sum_{k=0}^m (-1)^k \binom{m}{k} (m-k)^n \tag{12.6.10}$$

Now we can use a two-step procedure to generate all samples that exclude exactly j population values: First, choose the j values that are to be excluded. The number of ways to perform this step is $\binom{m}{j}$. Next select a sample of size n from the remaining population values so that none are excluded. The number of ways to perform this step is the result in the last displayed equation, but with $m - j$ replacing m. The multiplication principle of combinatorics gives the result.

The distributions of the number of excluded values and the number of distinct values are now easy.

The probability density function of U is given by

$$\mathbb{P}(U = j) = \binom{m}{j} \sum_{k=0}^{m-j} (-1)^k \binom{m-j}{k} \left(1 - \frac{j+k}{m}\right)^n, \quad j \in \{\max\{m-n, 0\}, \ldots, m-1\} \tag{12.6.11}$$

Proof

Since the samples are uniformly distributed, $\mathbb{P}(U = j) = \#\{U = j\} / m^n$ and so the result follows from the previous exercise.

The probability density function of the number of distinct values V is given by

$$\mathbb{P}(V = j) = \binom{m}{j} \sum_{k=0}^{j} (-1)^k \binom{j}{k} \left(\frac{j-k}{m}\right)^n, \quad j \in \{1, 2, \ldots, \min\{m, n\}\} \tag{12.6.12}$$

Proof

This follows from the previous theorem since $\mathbb{P}(V = j) = \mathbb{P}(U = m - j)$.

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 and compare the relative frequency function to the probability density function.

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

Let $f_{m,n}$ denote the probability density function of the number of excluded values U, when the population size is m and the sample size is n. Then

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

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.

For $j \in \{0, 1, \ldots, m\}$, let $I_j = \mathbf{1}(Y_j = 0)$, the indicator variable of the event that j is not in the sample. Note that the number of population values missing in the sample can be written as the sum of the indicator variables:

$$U = \sum_{j=1}^m I_j \tag{12.6.13}$$

For distinct i, $j \in \{1, 2, \ldots, m\}$,

1. $E(I_j) = \left(1 - \frac{1}{m}\right)^n$
2. $\text{var}(I_j) = \left(1 - \frac{1}{m}\right)^n - \left(1 - \frac{1}{m}\right)^{2n}$
3. $\text{cov}(I_i, I_j) = \left(1 - \frac{2}{m}\right)^n - \left(1 - \frac{1}{m}\right)^{2n}$

Proof

Since each population value is equally likely to be chosen, $\mathbb{P}(I_j = 1) = (1 - 1/m)^n$. Thus, parts (a) and (b) follow from standard results for the mean and variance of an indicator variable. Next, $I_i I_j$ is the indicator variable of the event that i and j are both excluded, so $\mathbb{P}(I_i I_j = 1) = (1 - 2/m)^n$. Part (c) then follows from the standard formula for covariance.

The expected number of excluded values and the expected number of distinct values are

1. $\mathbb{E}(U) = m\left(1 - \frac{1}{m}\right)^n$
2. $\mathbb{E}(V) = m\left[1 - \left(1 - \frac{1}{m}\right)^n\right]$

Proof

Part (a) follows from the previous exericse and the representation $U = \sum_{j=1}^n I_j$. Part (b) follows from part (a) since $U + V = m$.

The variance of the number of exluded values and the variance of the number of distinct values are

$$\text{var}(U) = \text{var}(V) = m(m-1)\left(1 - \frac{2}{m}\right)^n + m\left(1 - \frac{1}{m}\right)^n - m^2\left(1 - \frac{1}{m}\right)^{2n} \tag{12.6.14}$$

Proof

Recall that $\text{var}(U) = \sum_{i=1}^m \sum_{j=1}^m \text{cov}(I_i, I_j)$. Using the results above on the covariance of the indicator variables and simplifying gives the variance of U. Also, $\text{var}(V) = \text{var}(U)$ since $U + V = m$.

In the birthday experiment, select the number of distinct sample values. Vary the parameters and note the size and location of the mean $\pm$ standard-deviation bar. For selected values of the parameters, run the simulation 1000 times and compare the sample mean and variance to the distribution mean and variance.

Examples and Applications

Suppose that 30 persons are chosen at random. Find each of the following:

1. The probability density function of the number of distinct birthdays.
2. The mean of the number of distinct birthdays.
3. The variance of the number of distinct birthdays.
4. The probability that there are at least 28 different birthdays represented.

Answer

1. $\mathbb{P}(V = j) = \binom{30}{j} \sum_{k=0}^j (-1)^k \left(\frac{j-k}{365}\right)^{30}, \quad j \in \{1, 2, \ldots, 30\}$
2. $\mathbb{E}(V) = 28.8381$
3. $\text{var}(V) = 1.0458$
4. $\mathbb{P}(V \geq 28) = 0.89767$

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.

Suppose that 10 fair dice are rolled. Find each of the following:

1. The probability density function of the number of distinct scores.
2. The mean of the number of distinct scores.
3. The variance of the number of distinct scores.

4. The probability that there will 4 or fewer distinct scores.

Answer

1. $\mathbb{P}(V = j) = \binom{10}{j} \sum_{k=0}^{j} (-1)^k \binom{j}{k} \left(\frac{j-k}{6}\right)^{10}, \quad j \in \{1, 2, \ldots, 6\}$
2. $\mathbb{E}(V) = 5.0310$
3. $\var(V) = 0.5503$
4. $\mathbb{P}(V \le 4) = 0.22182$

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

A fast food restaurant gives away one of 10 different toys with the purchase of each kid's meal. A family buys 15 kid's meals. Find each of the following:

1. The probability density function of the number of toys that are missing.
2. The mean of the number of toys that are missing.
3. The variance of the number of toys that are missing.
4. The probability that at least 3 toys are missing.

Answwer

1. $\mathbb{P}(U = j) = \binom{15}{j} \sum_{k=0}^{10-j} (-1)^k \binom{10-j}{k} \left(1 - \frac{j+k}{10}\right)^{15}, \quad j \in \{0, 1, \ldots, 9\}$
2. $\mathbb{E}(U) = 2.0589$
3. $\var(U) = 0.9864$
4. $\mathbb{P}(U \ge 3) = 0.3174$

In the birthday experiment, set $m = 10$ and $n = 15$. Run the experiment 1000 times and compute the relative frequency of the event in part (d).

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. Find each of the following:

1. The probability density function of the number of distinct answers.
2. The probability that the students get away with their deception.
3. The mean of the number of distinct answers.
4. The standard deviation of the number of distinct answers.

Answer

1.

j	1	2	3
$\mathbb{P}(V = j)$	$\frac{1}{16}$	$\frac{9}{16}$	$\frac{6}{16}$

2. $\mathbb{P}(V = 1) = \frac{1}{16}$
3. $\mathbb{E}(V) = \frac{37}{16}$
4. $\sd(V) = \sqrt{\frac{87}{256}} \approx 0.58296$

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. Find each of the following:

1. The probability density function of the number of ducks that are killed.
2. The mean of the number of ducks that are killed.
3. The standard deviation of the number of ducks that are killed.

Answer

1.

j	1	2	3	4	5
$\mathbb{P}(V=j)$	$\frac{1}{10\,000}$	$\frac{27}{2000}$	$\frac{9}{50}$	$\frac{63}{125}$	$\frac{189}{625}$

2. $\mathbb{E}(V)=\frac{40\,951}{10\,000}=4.0951$
3. $\mathrm{sd}(V)=0.72768$
