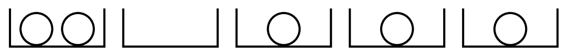
Balls and Bins Intro

Note 12
Note 13

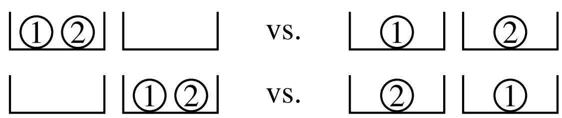
Recall the *balls and bins* (aka *stars and bars*) technique we used to count the number of ways we can sample with replacement, where order does not matter. Consider a process where we throw balls into bins, where each ball is independent and equally likely to land in any bin.



$\star\star \mid \mid \star \mid \star \mid \star\star$

Unfortunately, outcomes have different probabilities. For instance, with 2 balls and 2 bins, it's less likely for both balls to be in bin 1 than for one ball to be in each bin. We don't have a uniform probability space.

To work with a uniform probability space, we will *label* the balls. Order does matter now: throwing the first ball into bin 1 and the second ball into bin 2 (top right) is a different outcome from throwing the first ball into bin 2 and the second ball into bin 1 (bottom right).



Solution: You may ask, "why does labeling the balls give us a uniform probability space?"

For throwing m labeled balls into n bins, the probability of any outcome is

$$\underbrace{\frac{1}{n}}_{\text{ball 1 lands in its bin}} \times \underbrace{\frac{1}{n}}_{\text{ball 2 lands in its bin}} \times \cdots \times \underbrace{\frac{1}{n}}_{\text{ball } m \text{ lands in its bin}} = \frac{1}{n^m}$$

where the $\frac{1}{n}$ comes from the n bins being equally likely to be landed in, and it's the same for all balls since the balls are independent. Since there are n^m possible outcomes (each of the m balls has n bins it can land on), each outcome is equally likely—a uniform probability space. For instance, all four outcomes with labeled balls above have probability $\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$.

Whereas if the balls were not labeled, then the probabilities of each outcome are different. For example, the probability of both balls landing in bin 1 is $\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$, while the probability of one ball in each bin is $2 \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{2}$ (multiply by 2 since there are two "labeled ball" ways to get the same "unlabeled" ball outcome). Stars and bars would give you $\binom{2+2-1}{2} = 3$ outcomes, but the probabilities of these unlabeled ball outcomes are *not* $\frac{1}{3}$.

1 Balls and Bins

Note 14

Suppose you throw b balls into n labeled bins one at a time, where $b \geq 1$ and $n \geq 2$.

(a) What is the probability that the first bin is empty?

(b) What is the probability that the first k bins are empty?

(c) If k bins were randomly selected, what is the probability that they are all empty?

(d) Let A be the event that at least k bins are empty. Let m be the number of subsets of k bins out of the total n bins. If we assume A_i is the event that the ith subset of k bins is empty. Then we can write A as the union of A_i's:

$$A = \bigcup_{i=1}^{m} A_i.$$

Compute m in terms of n and k, and use the union bound to give an upper bound on $\mathbb{P}[A]$.

(e) What is the probability that the second bin is empty given that the first one is empty?

(f) Are the events that "the first bin is empty" and "the first two bins are empty" independent?

(g) Are the events that "the first bin is empty" and "the second bin is empty" independent?

Solution: Since the balls are thrown one at a time, there is an ordering, and so we are sampling with replacement where order matters rather than where it doesn't (which would correspond to each configuration in the stars and bars setup being equally likely).

(a) Note that this is a uniform sample space, with outcomes representing all possible ways to throw each ball individually into the bins. Here, $|\Omega| = n^b$, as each of the b balls has n possible bins to fall into, and out of these possibilities, $(n-1)^b$ of them leave the first bin empty—each ball would then have $n-1$ possible bins to fall into. This gives us an overall probability $\left(\frac{n-1}{n}\right)^b$ that the first bin is empty.

Equivalently, we can note that each throw is independent of all of the other throws. Since the probability that ball i does not land in the first bin is $\frac{n-1}{n}$, the probability that all of the balls do not land in the first bin is $\left(\frac{n-1}{n}\right)^b$.

(b) Similar to the previous part, we have the same uniform sample space of size n^b. Now, there are a total of $(n-k)^b$ possible ways to throw the balls into bins such that the first k bins are empty—each ball has $n-k$ possible bins to fall into.

Alternatively, we can similarly make use of independence. Since the probability that ball i does not land in the first k bins is $\frac{n-k}{n}$, the probability that all of the balls do not land in the first k bins is $\left(\frac{n-k}{n}\right)^b$.

(c) In part (b), it wasn't important which k bins we picked, so we can use the same reasoning to obtain $\left(\frac{n-k}{n}\right)^b$.

(d) We use the union bound. Then

$$\mathbb{P}[A] = \mathbb{P}\left[\bigcup_{i=1}^{m} A_i\right] \le \sum_{i=1}^{m} \mathbb{P}[A_i].$$

We know the probability of the first k bins being empty from part (b), and this is true for any set of k bins from part (c), so

$$\mathbb{P}[A_i] = \left(\frac{n-k}{n}\right)^b.$$

Then,

$$\mathbb{P}[A] \le m \cdot \left(\frac{n-k}{n}\right)^b = \binom{n}{k}\left(\frac{n-k}{n}\right)^b.$$

(e) Using Bayes' Rule:

$$\begin{aligned}\mathbb{P}[\text{2nd bin empty} \mid \text{1st bin empty}] &= \frac{\mathbb{P}[\text{2nd bin empty} \cap \text{1st bin empty}]}{\mathbb{P}[\text{1st bin empty}]} \\ &= \frac{(n-2)^b/n^b}{(n-1)^b/n^b} \\ &= \left(\frac{n-2}{n-1}\right)^b\end{aligned}$$

Alternate solution: We know bin 1 is empty, so each ball that we throw can land in one of the remaining $n-1$ bins. We want the probability that bin 2 is empty, which means that each ball cannot land in bin 2 either, leaving $n-2$ bins. Thus for each ball, the probability that bin 2 is empty given that bin 1 is empty is $\frac{n-2}{n-1}$. For b total balls, this probability is $\left(\frac{n-2}{n-1}\right)^b$.

(f) They are dependent. Knowing the latter means the former happens with probability 1.

(g) In part (e) we calculated the probability that the second bin is empty given that the first bin is empty: $\left(\frac{n-2}{n-1}\right)^b$. The probability that the second bin is empty (without any prior information) is $\left(\frac{n-1}{n}\right)^b$. Since these probabilities are not equal, the events are dependent.

2 Monopoly Card Collector

Note 15

It's that time of the year again—Safeway is offering its Monopoly Card promotion. Each time you visit Safeway, you are given one of n different Monopoly Cards with equal probability. You need to collect them all to redeem the grand prize.

(a) Suppose you have k unique Monopoly Cards already, where $k \le n$. What is the probability that you'll get a new unique Monopoly Card on your next visit to Safeway?

Now, suppose that you have visited Safeway r times already.

(b) Let X_i be the event that you are missing Monopoly Card i, where $i \in \{1, \dots, n\}$. Find $\mathbb{P}[X_i]$.

(c) Let T be the event that you are missing any Monopoly Card. Express T in terms of the events X_i. *Hint*: Events are subsets of the probability space Ω, so you should be thinking of set operations.

(d) Show that $ne^{-r/n}$ is an upper bound for $\mathbb{P}[T]$. That is, show that $\mathbb{P}[T] \le ne^{-r/n}$. (One inequality you may find helpful is $1 - x \le e^{-x}$. Is there another inequality that would be helpful?)

(e) Find the minimum number of Safeway visits r such that part (d) guarantees that $\mathbb{P}[T]$ is at most $\frac{1}{n^2}$?

Solution:

(a) On your next visit to Safeway, there are n possible cards you can get. Since k of them are the same as the cards you already have, there are $n-k$ cards that would be new for your collection.

Thus, the probability of getting a new unique card on your next visit is $\frac{n-k}{n} = 1 - \frac{k}{n}$.

(b) For you to be missing card i after all r visits, you need to not get it on visit 1, visit 2, ..., **and** visit r. The probability that you do get card i in any one visit is $\frac{1}{n}$, so the probability that you do not get card i in one visit is $1 - \frac{1}{n}$.

Since the probabilities are the same for each visit, we can use independence to find $\mathbb{P}[X_i]$:

$$\begin{aligned}
\mathbb{P}[X_i] &= \mathbb{P}[\text{miss card } i \text{ after } r \text{ visits}] \\
&= \mathbb{P}[\bigcap_{j=1}^{r} (\text{miss card } i \text{ on visit } j)] \\
&= \prod_{j=1}^{r} \mathbb{P}[\text{miss card } i \text{ on visit } j] && \text{(Independence)} \\
&= \prod_{j=1}^{r} (1 - \frac{1}{n}) \\
&= (1 - \frac{1}{n})^r
\end{aligned}$$

(c) For you to be missing any Monopoly Card, you can be missing card 1, card 2, ..., **or** card n. These are precisely the events X_i from part (b), so T is the union of these events.

$$T = \bigcup_{i=1}^{n} X_i$$

(d) We can also use the union bound, since we found in part (c) that T is a union of events X_i.

$$\begin{aligned}
\mathbb{P}[T] &= \mathbb{P}[\bigcup_{i=1}^{n} X_i] && \text{(part (c))} \\
&\leq \sum_{i=1}^{n} \mathbb{P}[X_i] && \text{(Union Bound)} \\
&= \sum_{i=1}^{n} (1 - \frac{1}{n})^r && \text{(part (b))} \\
&= n \cdot (1 - \frac{1}{n})^r \\
&\leq n \cdot (e^{-1/n})^r && \text{(Inequality)} \\
&= n \cdot e^{-r/n}
\end{aligned}$$

where the inequality still holds when raising to the power of r, as the base $1 - \frac{1}{n}$ is non-negative.

(e) We want to find the minimum r such that $\mathbb{P}[T] \leq \frac{1}{n^2}$. Using part (d) and solving for r:

$$\begin{aligned}
\mathbb{P}[T] \leq n \cdot e^{-r/n} &\leq \frac{1}{n^2} \\
e^{-r/n} &\leq \frac{1}{n^3} \\
-\frac{r}{n} &\leq \ln \frac{1}{n^3} = -3 \ln n \\
r &\geq 3n \ln n
\end{aligned}$$

and we will round $3n \ln n$ up to the nearest integer, since the number of visits must be an integer.

3 Planetary Party

Note 15

For this question, you may use a calculator.

(a) Suppose we are at party on a planet where every year is 2849 days. If 30 people attend this party, what is the exact probability that two people will share the same birthday? You may leave your answer as an unevaluated expression.

(b) From lecture, we know that given n bins and m balls, $\mathbb{P}[\text{no collision}] \approx \exp(-m^2/(2n))$. Using this, give an approximation for the probability in part (a).

(c) Using the approximation in part (b), what is the approximate minimum number of people that need to attend this party to ensure that the probability that any two people share a birthday is at least 0.5?

(d) Now suppose that 70 people attend this party. What the is probability that none of these 70 individuals have the same birthday? You can use the approximation you used in the previous parts.

Solution:

(a) Let's compute the probability that no two partygoers have the same birthday. We know the second person at the party cannot share a birthday with the first person, the third person at the party cannot share a birthday with the first two, etc. Thus

$$\mathbb{P}\left[\text{no collision}\right] = \left(1 - \frac{1}{2849}\right)\left(1 - \frac{2}{2849}\right)\left(1 - \frac{3}{2849}\right)\cdots\left(1 - \frac{29}{2849}\right)$$

Thus $\mathbb{P}[\text{collision}] = 1 - \mathbb{P}[\text{no collision}] = 1 - \left(1 - \frac{1}{2849}\right)\left(1 - \frac{2}{2849}\right)\left(1 - \frac{3}{2849}\right)\cdots\left(1 - \frac{29}{2849}\right)$.

(b) From lecture, we know that given n bins and m balls, $\mathbb{P}\left[\text{no collision}\right] \approx \exp(-m^2/(2n))$. Therefore in this case, if we want to find the probability of collision, we must find $1 - \mathbb{P}\left[\text{no collision}\right]$.

$$\mathbb{P}\left[\text{no collision}\right] = \exp\left(-\frac{30^2}{2 \cdot 2849}\right) = 0.854$$

This means that there is a 0.146 chance that two people share the same birthday in the group of 30.

(c) Rephrasing the question in terms of balls and bins, we want to find the minimum number of balls (m) such that there is at least 0.5 probability of collision when we have $n = 2849$ bins, which is the same as **at most** 0.5 probability of **no** collisions.

$$\begin{aligned}
\mathbb{P}[\text{no collisions}] \approx \exp\left(\frac{-m^2}{2n}\right) &\leq 0.5 \\
\implies \frac{-m^2}{2n} &\leq \ln 0.5 \\
\implies m &\geq \sqrt{(-2\ln 0.5)n} \\
&= 62.845
\end{aligned}$$

Since m must be an integer which is at least 62.485, we need at least $\boxed{63}$ people at the party.

(d) Once again we need to find $\mathbb{P}\left[\text{no collisions}\right]$ given that $m = 70$.

$$\mathbb{P}\left[\text{no collision}\right] = \exp\left(-\frac{70^2}{2 \cdot 2849}\right) = 0.423$$

There is about a 42% chance that 70 people don't share the same birthday.