Independent random variables and building probability models

Just a recap, there are 3 steps in building a probability model of a real-world problem:

1. Identify the basic random variables.
2. Define the sample space $\Omega$.
3. Assign probabilities to the outcomes in the sample space.

In earlier examples, it was quite easy to assign probabilities to the outcomes because in many cases each outcome has the same probability, so the probability of each outcome is simply $1/|\Omega|$. But things can get more complicated when the probabilities of the outcomes are not the same. For example, in the case of $n$ coin flips, there are $2^n$ outcomes (all possible $n$-tuples of H’s and T’s), and assigning probabilities to each one of the outcome will take a long time when $n$ is large. However, in many models, it is reasonable to assume that many of the random variables are independent.

1. **Coin Tosses**: In this example, let $X_i$ be the result of the $i$th flip. It is reasonable to assume that the $X_i$’s are mutually independent. Also, it is reasonable to assume that the bias of each flip of getting a Heads is the same, say $p$. So the probability of getting an outcome say HHHTTH can be computed by just multiplying the probabilities:

$$P(HHHTTH) = P(X_1 = H)P(X_2 = H)P(X_3 = H)P(X_4 = T)P(X_5 = T)P(X_6 = H)$$
$$= p \times p \times p \times (1 - p) \times (1 - p) \times p$$
$$= p^4(1 - p)^2.$$  

Note that the number of parameters needed to specify the probabilities in this model has been drastically reduced from $2^n - 1$ to 1 by assuming independence of the flips and that the bias of each flip is the same (because we are flipping the same coin over and over again).

2. **Balls in Bins**: Let $X_i$ be the index of the bin in which the $i$th ball ends up. Then it is reasonable to assume that the $X_i$’s are mutually independent random variables. We note that we can also consider the random variables $Y_j$’s, the number of balls in bin $j$, as basic random variables to model our problem, but these random variables are no longer independent and the model becomes much more complicated.

3. **Monty Hall**: Here we want 3 random variables $X$, the location of the car, $Y$, the contestant’s choice, and $Z$ Carol’s choice. It is reasonable to model $X$ and $Y$ as independent. But $Z$ is not independent of $X$ and $Y$ because Carol’s choice depends on where the car is and the contestant’s choice. So this is an example of a probability model where some of the random variables are independent but others are dependent on them, a pretty common situation.
**Example: The coupon collection problem**

Suppose that when we buy a box of cereal, as a marketing ploy, the cereal manufacturer has included a random baseball card. Suppose that there are \( n \) baseball players who appear on some card, and each cereal box contains a card chosen uniformly and independently at random from these \( n \) possibilities. Assume that Babe Ruth is one of the \( n \) baseball players, and I am a huge fan of him: I really want his baseball card. How many boxes of cereal will I have to buy, to have at least a 90% chance of obtaining a Babe Ruth card? This problem can be analyzed as follows. Suppose we buy \( m \) boxes of cereal. Let \( E \) be the event that I do receive a Babe Ruth card in any of the \( m \) boxes. The card in each of the \( m \) boxes is independent, and for each box the chances that I don’t receive Babe Ruth’s card from that box is \( 1 - \frac{1}{n} \). We first find \( m \) such that \( P(E) \geq 0.9 \). \( E^c \) is the event that none of the \( m \) boxes contain BR. Let \( F_i \) be the event that the \( i \)th box does not have Babe Ruth. We can write \( E^c = F_1 \cap F_2 \cdots F_m \).

Now, we have that
\[
P(E^c) = P(F_1 \cap F_2 \cdots \cap F_m) = P(F_1) \cdots P(F_m) = P(F_1)^m
\]
where \( P(F_i) = 1 - \frac{1}{n} \). Therefore, \( P(E) = 1 - (1 - \frac{1}{n})^m \).

We would like to get \( P(E) \geq 0.9 \), so we get
\[
1 - (1 - \frac{1}{n})^m = 0.9 \quad \implies (1 - \frac{1}{n})^m = 0.1 \quad \implies m = \frac{\log(0.1)}{\log(1 - \frac{1}{n})}
\]
hence \( m \approx n \log(10) \). Note that \( m \) depends linearly on \( n \): if there are twice as many cards, you need to buy twice as many boxes in order to guarantee with 90% chance you will get Babe Ruth.

Next suppose that what I really want is a complete collection: I want at least one of each of the \( n \) cards. Let \( A \) be the event that we get the whole collection after \( m \) boxes. Then, \( A^c \) is the event that we are missing at least one of the cards after \( m \) boxes. This event can be broken down into simpler events. How is \( A^c \) related to \( E \)?

Let's give a number to each player and suppose Babe Ruth is player number 1 (who else?). Let \( E_1^c \) be the event that BR is not among the cards in the \( m \) boxes. Similarly, we let \( E_i \) be the event that player \( i \) is not among the cards in the \( m \) boxes and \( E_i^c \) is the event that player \( i \) is not among the cards in the \( m \) boxes. \( A^c \) can be expressed in terms of \( E_i \) as
\[
A^c = E_1^c \cup \cdots E_n^c
\]
Now, by symmetry,
\[
P(E_i^c) = (1 - \frac{1}{n})^m, \quad i \in 1, \ldots, m.
\]
We are interested in finding an \( m \) that is large enough such that \( P(A^c) \leq 0.1 \). Therefore, we only need to bound \( P(A^c) \) from above.

Using the union bound (i.e. \( P(A \cup B) \leq P(A) + P(B) \) for any events \( A \) and \( B \)), we get that
\[
P(A^c) \leq P(E_1^c) + \cdots + P(E_n^c) = n(1 - \frac{1}{n})^m.
\]
Therefore it is sufficient to solve for a value of \( m \) such that \( n(1 - \frac{1}{n})^m = 0.1 \), which gives \( m \approx n \log(10n) \).

In the case of collecting a specific card (say Babe Ruth), the number of boxes one needs to collect grows linearly with the number of cards. In the case of collecting a whole collection, however, the number of boxes

---

1This section is adapted from Berkeley EECS 70 notes.
needed grows *super-linearly* with the number of cards. This means that if there are twice as many cards to collect, you need to buy twice the number of boxes to collect them. This is because it becomes more and more difficult to get a new card once we have already collected several cards.

Perhaps the most interesting aspect of the coupon collector’s problem above is that it illustrates the use of the union bound to upper-bound the probability that something bad happens. In this case, the bad event is that we fail to obtain a complete collection of baseball cards, but in general, this methodology is a powerful way to prove that some bad event is not too likely to happen.