

The iPod Shuffle
Stat 341 - Fall 2008

As we saw with the birthday problem, probabilities of events can be non-intuitive. Events that seem to have small probabilities actually have larger probabilities than expected. This was definitely true for the following problem. In January 2005, Steven Levy, the Technologist for Newsweek magazine, wrote an article about the random shuffle feature of the iPod. The link to this article is available on the course webpage. His experiences with the shuffle feature led him to question its randomness.

In this article, he describes using iTunes to randomly select songs to fill the smaller iPod Shuffles. At the time, the iPod Shuffle would hold about 120 songs. When he used iTunes to select songs for the Shuffle, he noticed that several of the songs selected were from the same album. “The first few times . . . , I found some disturbing clusters in the songs chosen. More than once the “random” playlist included three tracks from the same album! Since there are more than 3000 tunes in my library, this seemed to defy the odds.”

There are two ways to look at this statement. The first way is to choose an album and look at the 120 songs selected to fill the Shuffle to see how many songs are selected from **this particular album** each time. Assume the album contains 12 songs and the library of songs has a total of 3,000 songs. We can simulate the probability that this particular album will have 3 or more songs selected for the Shuffle in much the same way as we simulated the probability of obtaining heads when flipping a coin. The variable **album** will indicate whether or not the songs in the library are from the particular album of interest.

```
album<- c(rep(1,12), rep(0, 2988))
```

Randomly selecting the 120 songs for the Shuffle involves sampling from the variable album 120 times. Since songs are not repeated in the songs selected for the Shuffle, the sampling is done without replacement. We can then use the sum command to calculate how many songs from our particular album have been selected for the Shuffle.

```
shuffle<- sample(album, 120, replace = F)  
sum(shuffle)
```

In my shuffle, only one song was selected from this particular album. We can, of course, repeat this sampling to obtain an empirical probability of obtaining at least three songs from this particular album. Here is the R code.

```
numsongs<- rep(0, 10000)  
  
for (i in 1:10000){  
  shuffle<- sample(album, 120, replace = F)  
  numsongs[i]<- sum(shuffle)  
}
```

Below is a table of observed values for the number of songs selected from the particular album of interest. As you can see, selection of three or more songs occurs only 109 times ($102 + 7 + 1$) out of 10,000 trials for an empirical probability of 0.0109. Focusing on just one particular album makes the probability small.

0	1	2	3	4	5
6157	3034	699	102	7	1

However, this is not the event described in Levy's article. (I contacted him by email just to make sure). The event he describes is the event where there are three or more songs from *any* album in the 120 songs selected for the Shuffle. This is the same as grouping his library of 3000 songs by album and looking at the maximum number of songs from any one album in the 120 songs selected for the Shuffle. We can simulate this situation in a similar way to the simulation of the birthday problem. We will assume to start that the library of 3000 songs contains 250 albums, with each album having 12 songs. Which album each of the 3000 songs belongs to is indicated by number in the variable `albums` below.

```
albums<- rep(1:250, 12)
```

Just as before, randomly selecting the 120 songs for the Shuffle is like sampling from the variable `albums` without replacement. Here is an example.

```
shuffle<- sample(albums, 120, replace = F)
122 240 90 66 18 212 97 158 69 98 233 134 52 70 83 128 76 77
227 225 163 173 239 221 205 136 95 239 242 198 144 210 29 8 230 153
67 92 230 247 16 186 163 208 188 117 89 245 98 28 92 164 10 237
68 211 143 113 186 103 176 199 40 143 133 72 233 129 97 156 22 82
12 29 162 183 59 90 3 31 124 9 90 194 7 10 130 71 127 148
192 227 107 130 5 90 38 52 216 156 9 204 138 224 71 55 127 96
236 143 162 233 66 196 145 86 101 218 200 178
```

Scanning through the list of albums belonging to the 120 songs selected for repeated albums is difficult. However, similar to the birthday problem, we can use histograms to help with this task. There are 250 different albums, so we will have breaks in the histogram centered on these 250 values.

```
albumbreaks<- c(1:251) - 0.5
hist(shuffle, breaks = albumbreaks)$counts
```

In my example, one album has four songs selected, two albums have three songs selected, several albums have two songs selected, and many albums have one song selected. If you repeat this simulation, you will get different albums selected, and different numbers of songs selected from the different albums. However, what you will notice is that the maximum number of songs selected from *any* album is usually at least three.

What is the probability of this event? We can obtain an empirical probability in R similar to our other examples. Here is the code.

```

maxnumsongs<- rep(0, 10000)

for (i in 1:10000){
  shuffle<- sample(albums, 120, replace = F)
  albumcounts<- hist(shuffle, breaks = albumbreaks, plot = F)$counts
  maxnumsongs[i]<- max(albumcounts)
}

```

The variable **maxnumsongs** contains the maximum number of songs from **any** one album selected for the Shuffle. As you can see from the table, in our 10,000 trials, there were always repeated albums in the 120 songs. While it is possible for each of the 120 songs to come from a different album, this did not happen in the 10,000 trials. The event of interest, three or more songs from **any** album occurred in 9,439 out of the 10,000 trials ($7252 + 2040 + 138 + 8 + 1$), for an empirical probability of 0.9439.

2	3	4	5	6	7
561	7252	2040	138	8	1

In his article, Levy stated this event seemed to defy the odds. In fact, the probability of this event is very high, 0.944, give or take. In performing this simulation, we made one assumption about the library of songs available to select from for the Shuffle. Many people do not have an equal number of songs per album in their libraries of songs. Full albums can range anywhere from 8 to as many as 20 songs. The advent of iTunes has made full albums much rarer, making libraries with just a few songs from a particular album much more common. The assumption of a simple library versus a more realistic one, like the one Levy had, doesn't make the event less likely; it makes it more likely to occur. Even though we don't have the details of Levy's library at the time, the event of seeing three or more songs from an album when selecting songs for the Shuffle is virtually certain to occur. So much for defying the odds!