

# Odds of a Given Playlist Occurring at Random

MusicStrands, Inc.

March 27, 2006

## 1 General Problem Statement

Given a population of  $L$  songs from which  $K$  playstreams of length  $M$  are selected, what are the odds that a given playlist  $\mathcal{P}$  of  $N$  songs occurs in a playstream? We assume that all songs are equally popular since popularity effects significantly complicate the analyses we present here. We also assume that playlist "occurs" in a playstream of length  $M \geq N$  songs if the playstream includes the  $N$  songs distributed in the  $M$  songs with any pattern and any order.

## 2 Scenario #1

In the first scenario we assume that no song is repeated in a playstream. Under the assumptions, we can reduce the first step, determining the probability of the playlist of  $N$  songs occurring in a single playstream of length  $M$  problem to an *urn* problem. In this case we assume we have an urn that contains  $L - N$  balls with the number "0" on them, and  $N$  balls with the numbers "1", ..., "N" on them. If we do  $M$  draws without replacement, and on each draw note the number on the ball, what is the probability we have noted the numbers "1", ..., "N" in the  $M$  draws?

The probability of drawing a particular number of each type of balls is given by the *multivariate hypergeometric distribution*. We can simplify our problem in this case, however, because we want to know the probability we draw all  $N$  of the balls with the numbers "1", ..., "N". We can assume that instead we have  $N$  balls numbered "1". The probability of drawing  $n$  balls numbered "1" in  $M$  draws without replacement is given by the value of the *hypergeometric distribution* for the random variable  $\mathbf{N}$

$$\Pr\{\mathbf{N} = n; L, M, N\} = \frac{\binom{N}{n} \binom{L - N}{M - n}}{\binom{L}{M}}$$

In this case,  $\mathbf{N} = N$ , so we have

$$\Pr\{\mathbf{N} = N; L, M, N\} = \frac{\binom{N}{N} \binom{L-N}{M-N}}{\binom{L}{M}} = \frac{\binom{L-N}{M-N}}{\binom{L}{M}}$$

In addition, if  $M = N$ , since by definition

$$\binom{L-N}{0} = 1$$

we have

$$\Pr\{\mathbf{N} = N; L, N, N\} = \binom{L}{N}^{-1}$$

We then determine that the probability we do not see the playlist in  $K$  playstreams is

$$\Pr\{\mathcal{P} \text{ does not occur in } K \text{ playstreams}\} = 1 - \Pr\{\mathbf{N} = N; L, M, N\}^K$$

and the probability we do see the playlist is

$$\Pr\{\mathcal{P} \text{ does occur in } K \text{ playstreams}\} = 1 - [1 - \Pr\{\mathbf{N} = N; L, M, N\}^K]$$

The odds of seeing the playlist then are:

$$\frac{\Pr\{\mathcal{P} \text{ does occur in } K \text{ playstreams}\}}{\Pr\{\mathcal{P} \text{ does not occur in } K \text{ playstreams}\}} = \frac{1 - [1 - \Pr\{\mathbf{N} = N; L, M, N\}^K]}{1 - \Pr\{\mathbf{N} = N; L, M, N\}^K}$$

Evaluating this exactly, given  $L = 1,000,000$ ,  $K = 1$ ,  $N=10$ , and say on average  $M = 6000$ , is difficult because the probability of seeing the playlist is so small. We can approximate the value in a two step process. First, for large  $L$  such as this, the hypergeometric distribution  $\Pr\{\mathbf{N} = n; L, M, N\}$  can be approximated by the binomial distribution

$$\Pr\{\mathbf{N} = n; M, p = N/L\} = \binom{M}{n} p^n (1-p)^{M-n}$$

If  $M * p > 10$  and  $M * (1-p) > 10$ , which unfortunately is not the case here, we could approximate this binomial distribution by the normal distribution  $\mathbb{G}(n; Mp, Mp(1-p))$ . We could reformulate the problem such that we want to find a last one occurrence of the playlist in a single playstream of  $K * M$  songs, but our simple urn model wouldn't apply.

With these numbers we find

$$\begin{aligned}\Pr\{\mathbf{N} = 10; M = 6000, p = 10/1000000\} &\approx \binom{6000}{10} (10^{-5})^{10} (1 - 10^{-5})^{5990} \\ &= \binom{6000}{10} (9.9999 \cdot 10^{-6040/5990})^{5990}\end{aligned}$$

We still can't compute this directly with reasonable accuracy, so instead we first compute this by first computing

$$\begin{aligned}\ln \Pr\{\mathbf{N} = 10; M = 6000, p = 10/1000000\} \\ &= \ln \Pr\{\mathbf{N} = 10; M = 6000, p = 10/1000000\} \\ &= \ln \binom{6000}{10} - 50 \ln 10 + 5990 \ln 9.9999 - 5990 \ln 10\end{aligned}$$

We can then this value with the aid of Stirling's approximation  $\ln n! \approx n \ln n - n$  to

$$\begin{aligned}\ln \Pr\{\mathbf{N} = 10; M = 6000, p = 10/1000000\} \\ &= [(6000 \ln 6000 - 6000) - (5990 \ln 5990 - 5990) - (10 \ln 10 - 10)] \\ &\quad - 6040 \ln 10 + 5990 \ln 9.9999 \\ &= 6000 \ln 6000 - 5990 \ln 5990 - 6050 \ln 10 + 5990 \ln 9.9999 \\ &\approx 52197.088489 - 52100.101680 - 13930.639813 + 13792.424807 \\ &= -41.228197\end{aligned}$$

so that  $\Pr\{\mathbf{N} = 10; M = 6000, p = 10/1000000\} \approx 1.244003 \cdot 10^{-18}$ . The odds then of seeing the playlist in single day's playstream ( $K = 1$ ) are:

$$\begin{aligned}\frac{\Pr\{\mathcal{P} \text{ does occur in } K \text{ playstreams}\}}{\Pr\{\mathcal{P} \text{ does not occur in } K \text{ playstreams}\}} &= \frac{1 - (1 - 1.244003 \cdot 10^{-18})^1}{(1 - 1.244003 \cdot 10^{-18})^1} \\ &\approx \frac{1}{8.038564 \cdot 10^{17}}\end{aligned}$$

If we repeat the process and ask the same question for a year of playstreams ( $K = 365$ ), we get

$$\frac{\Pr\{\mathcal{P} \text{ does occur in } K \text{ playstreams}\}}{\Pr\{\mathcal{P} \text{ does not occur in } K \text{ playstreams}\}} = \frac{1 - (1 - 1.244003 \cdot 10^{-18})^{365}}{(1 - 1.244003 \cdot 10^{-18})^{365}}$$

Using the approximation  $\ln(1 + x) \approx x$  for small  $x$ , we have

$$\begin{aligned}\ln(1 - 1.244003 \cdot 10^{-18})^{365} &= 365 \ln(1 - 1.244003 \cdot 10^{-18}) \\ &\approx 365 \cdot -1.244003 \cdot 10^{-18} \\ &= -4.540611 \cdot 10^{-16}\end{aligned}$$

so that

$$\begin{aligned} \frac{\Pr\{\mathcal{P} \text{ does occur in } K \text{ playstreams}\}}{\Pr\{\mathcal{P} \text{ does not occur in } K \text{ playstreams}\}} &\approx \frac{4.540611 \cdot 10^{-16}}{(1 - 4.540611 \cdot 10^{-16})} \\ &\approx \frac{1}{2.202347 \cdot 10^{15}} \end{aligned}$$

In rough terms then, the odds of the playlist occurring in a playstream in one day are 1 to  $8.038564 \cdot 10^{17}$  *against*. The odds of the playstream in one year are roughly 1 to  $2.202347 \cdot 10^{15}$  *against*. Highly unlikely by any estimate.

### 3 Scenario #2

In this second scenario, we assume songs can occur more than once in a playstream. We use the original formulation of the urn problem in which we assume the urn contains  $L - N$  balls with the number “0” on them, and  $N$  balls with the numbers “1”, ..., “ $N$ ” on them. If we do  $M$  draws, and on each draw note the number on the ball and then replace it in the urn, what is the probability we have noted the numbers “1”, ..., “ $N$ ” in the  $M$  draws?

One approach is based on viewing the set of  $M$  draws as independent events with  $N + 1$  outcomes  $i = 0, 1, \dots, N$ , such that the probability of the individuals outcomes are

$$p_0 = \frac{L - N}{L} \text{ (balls with "0")} \quad p_1 = \dots = p_N = \frac{1}{L} \text{ (balls with "1", \dots, "N")}$$

If we let  $r_0, r_1, \dots, r_N$  denote the number of outcomes of each type, then the probability of a specific set of outcomes is

$$\Pr\{r_0, r_1, \dots, r_N\} = \frac{M!}{r_0! r_1! \dots r_N!} p_0^{r_0} p_1^{r_1} \dots p_N^{r_N} \quad \sum_{i=0}^N p_i = 1, \quad \sum_{i=0}^N r_i = M$$

The probability of seeing the  $N$  songs on the playlist in the playstream is

$$\Pr\{\mathcal{P} \text{ occurs in playstream}\} = \sum_{r_0=0}^{M-N} \sum_{\substack{r_1 + \dots + r_N \\ = M - r_0 \\ r_i \geq 1}} \frac{M!}{r_0! r_1! \dots r_N!} p_0^{r_0} p_1^{r_1} \dots p_N^{r_N}$$

Direct evaluation of this equation for the probability that  $\mathcal{P}$  occurs in the playstream clearly is difficult. However, we can re-frame this scenario as in the first scenario by taking advantage of the fact the  $p_i$  are the same for  $i = 1, \dots, N$ . As in the first case we can consider first the probability of drawing  $r_0$  balls with the number “0” on them from the urn in  $M$  draws. This is described by the binomial distribution

$$\Pr\{\text{drawing } r_0 \text{ balls with "0"}\} = \binom{L}{r_0} p_0^{r_0} (1 - p_0)^{M - r_0}$$

If we have drawn  $M - r_0$  balls with a number other than “0” on them, we can ignore the numbers on those balls and instead determine how many ways we can randomly distribute those balls between  $N$  urns numbered “1”, “2”,  $\dots$ , “ $N$ ” so that there are at least  $\ell$  balls in each urn. This number is

$$C_{M-r_0, N, \ell} = \binom{M - r_0 - \ell N}{N}$$

Since the  $p_i$  are equal for  $i = 1, \dots, N$ , each distribution of the balls amongst the  $N$  urns is equally likely. The conditional probability of a distribution with at least 1 ball in each urn then is

$$\Pr\{\text{distribution with at least } \ell \text{ balls in each urn}\} = \frac{\binom{M - r_0 - \ell N}{N}}{\binom{M - r_0}{N}}$$

We conclude that

$$\begin{aligned} \Pr\{\mathcal{P} \text{ occurs in playstream}\} &= \sum_{r_0=0}^{M-N} \frac{\binom{L}{r_0} \binom{M - r_0 - N}{N}}{\binom{M - r_0}{N}} p_0^{r_0} (1 - p_0)^{M-r_0} \\ &= \sum_{r_0=0}^{M-N} \frac{(M - r_0 - N)!^2 L!}{(M - r_0)! (M - r_0 - 2N)! r_0! (L - r_0)!} p_0^{r_0} (1 - p_0)^{M-r_0} \end{aligned}$$

We can approximate individual terms in the summation by taking the log and using Stirling’s approximation as before (we have to be careful about the case  $r_0 = M - N$  since  $0! = 1$  and Stirling’s approximation is undefined):

$$\begin{aligned} \ln \frac{\binom{L}{r_0} \binom{M - r_0 - N}{N}}{\binom{M - r_0}{N}} p_0^{r_0} (1 - p_0)^{M-r_0} &\approx [2(M - r_0 - N) \ln(M - r_0 - N) - 2(M - r_0 - N) + L \ln L - L] \\ &\quad - (M - r_0) \ln(M - r_0) + (M - r_0) \\ &\quad - (M - r_0 - 2N) \ln(M - r_0 - 2N) + (M - r_0 - 2N) \\ &\quad - r_0 \ln r_0 + r_0 - (L - r_0) \ln(L - r_0) + (L - r_0) \\ &\quad + r_0 \ln p_0 + (M - r_0) \ln(1 - p_0) \\ &\approx 2(M - r_0 - N) \ln(M - r_0 - N) + L \ln L \\ &\quad - (M - r_0) \ln(M - r_0) - (M - r_0 - 2N) \ln(M - r_0 - 2N) \\ &\quad + r_0 \ln r_0 + (L - r_0) \ln(L - r_0) - 2L + r_0 \ln p_0 + (M - r_0) \ln(1 - p_0) \end{aligned}$$

The calculator work and the final odds computation in a manner analogous to the previous case are left as an exercise for the reader.