Birthdays Project: Some general comments on counting the possibilities for birthdays

John McCuan

October 19, 2023

Abstract

This is a project based on Problem 3 from Orlof and Booth's

`18.05 Problem Set 1, Spring 2014`.

I would like to both understand various aspects of the problem and also all elements in the code of the `18.05 function` colMatches. This code is contained in a file `colMatches.r` which is easily downloaded from the internet, and since I am an expert on neither the statistics package R nor computer programming in general[1] the analysis of `colMatches.r` may be a little tedious and involve several attempts/iterations with details that may be extraneous to many of my readers.

I had intended to include this discussion in my course notes, but for the moment I think it may be better to produce a standalone document. This is primarily due to the cumbersome discussion of the r coding and the anticipated extensive use of color text in reference to that discussion.

[1]It would probably be a more poetic description (and more accurate) to say I know nothing about R and nothing about computer programming. Neither description is entirely accurate.

Say we have a collection of k people with birthdays modeled by the natural numbers in

$$S = \{1, 2, 3, \ldots, n\}.$$

For example if the birthdays January 1, January 2, January 3, ..., December 31 in a standard year are taken as possibilities, then we may take $S = \{1, 2, 3, \ldots, 365\}$. This will be the default example for the discussion below though most if not all formulas and simulations can be given for general n.

The basic question with which we wish to start is of the following form:

> *What is the probability that among k people, at least m of them have the same birthday?*

A first observation is that the problem only really makes sense for

$$k \geq m \geq 2.$$

Alternatively, if $m = 1$, then the probability becomes 1 or 100%, and if $k < m$, then the probability becomes 0 because it is impossible for k people to share m birthdays in this case. A second observation is that if k is too large, then the probability again becomes 1 or 100%. This cutoff is at $(m-1)n$, so the interesting cases correspond to the conditions

$$2 \leq m \leq k \leq (m-1)n. \tag{1}$$

The significance of these inequalities may be somewhat difficult to visualize and/or appreciate especially when the value of n takes a large value like 365. Figure 1 and Figure 2 below represent an effort to illustrate the basic bounds of (1). In an effort to see the upper limits of interest for large values of n, we may introduce a logarithmic scale on the horizontal axis. The region of interest is then bounded above by $\ln[(m-1)n]$. Specifically, if we introduce the paramter $w = \ln k$, then interesting values of k correspond to $\ln m \leq w \leq \ln[(m-1)n]$ as illustrated for $n = 365$ in Figure 2.

1 Modeling

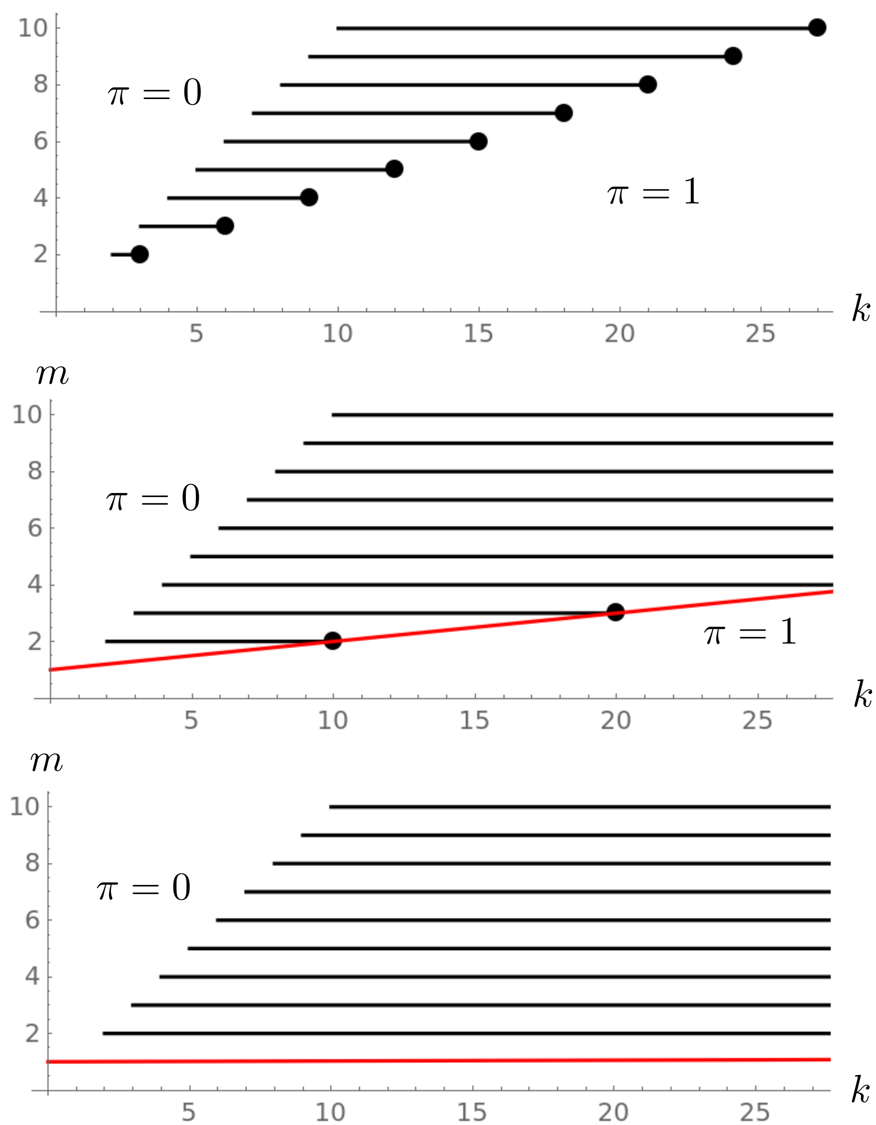

Figure 1: The values of k and m of interest for $n = 3$ (top), $n = 10$ (middle), and $n = 365$ (bottom). For each fixed m (desired number of birthday matches) the values of interest start at $m = k$ and terminate along a line of slope $1/n$ through $(k, m) = (0, 1)$. When $n = 365$ this line appears horizontal in the scale of the plot.

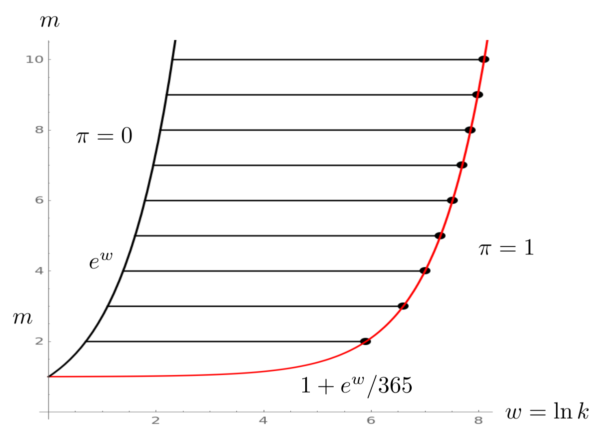

Figure 2: The values of $w = \ln k$ and m of interest for $n = 365$.