

Homework 1: Birthday Problem

Stats285

Fall 2017, Stanford

Due date: Oct 19, 2017

* You may use the `farmshare` compute clusters to run your experiments if you don't have access to any other cluster on campus.

* All issues, and questions regarding ClusterJob must be posted on <https://groups.google.com/forum/#!forum/clusterjob>

Problems

This homework is intended to familiarize you with the ClusterJob environment. We consider a classic probability problem, *the Birthday Problem*, and solve it by conducting computational experiments.

The birthday problem, which was proposed by Richard von Mises in 1939 determines *the chance of finding, in a group of n randomly chosen people, two people with the same birthday* In this assignment, the quantities we would like to compute at each particular n are the expected number of matching pairs and the empirical probability of finding at least a match. More precisely, let x_i denote the number of matches in i -th trial ($1 \leq i \leq N$), then the empirical mean and probability are given by,

$$\bar{x} = \frac{\sum_{i=1}^N x_i}{N}, \quad \text{Prob}(\text{match}) = \frac{\sum_{i=1}^N \mathbf{1}(x_i > 0)}{N}$$

where $\mathbf{1}(\cdot)$ is the indicator function for finding at least a match. Note that each trial involves sampling n people (birthdays) with replacement from a total of D days in a year (here we assume $D = 365$ ignoring the leap years).

The classic answer to the birthday problem assumes that all 365 days of a year are equally probable. However, in reality there are seasonal trends and weekend/weekday/holiday effects for birth rates. Figure 1 shows the empirical probability of birth across the days of the year. Evidently, there is less birth chance on holidays such as the new year, Christmas and Independence day, and higher chance in September. With this document, you are given the probability (weight) vector in a CSV file: `empirical_birth_pdf.csv`. This data is generated by a query to [Google's BigQuery WebUI](#) Google has the public birth data from 1/1/1969 to 12/31/1988 at `publicdata:samples.natality`. This data is provided by [USA Centers for Disease Control and Prevention](#)

We are interested to see whether considering the seasonal and weekend/weekday effect would change the classical solution. Therefore, You will compute and compare the quantities of interest for two sampling scenarios:

- The probabilities of birth for all 365 days are equal (**uniform**)
- The probabilities of birth for 365 day are given by empirical data (**empirical**)

Please note that Persi Diaconis and Susan Holmes (Stanford Statistics Professors) studied such problems analytically in their 2002 paper available at <http://statweb.stanford.edu/~susan/papers/feller.ps>. Interested students may read this paper to compare their conclusions with those of this paper.

1 Install ClusterJob

Visit <http://clusterjob.org> and claim your CJID and CJKey. Then, follow the instruction at <http://clusterjob.org/documentation/> to install CJ on your machines. To test that your installation is successful:

1. Issue `cj who` command. You should get your CJID and AgentID back.
2. run `simpleExample.m` in `clusterjob/example/MATLAB` by
`$ cj run simpleExample.m corn -m 'test'`

2 Implement the birthday problem

Write a MATLAB code that implements the birthday problem for $n = 23$. In particular, for each random instance of n people, your code should output a comma-delimited text (CSV) file, say `results.csv`, that contains at least the following 7 quantities:

1. Your SUID (e.g., 'monajemi')
2. Sampling type ('empirical' or 'uniform')
3. Number of people, n
4. Number of assumed days in a year, D
5. Total number of possible pairs
6. Total number matched pairs
7. Time elapsed

In MATLAB, you can implement this as:

```
fprintf(fid, '%s,%s,%i,%i,%i,%i,%f\n', ...
        SUID,samplingType,n, D, total_pairs, matched_pairs, tElapsed);
```

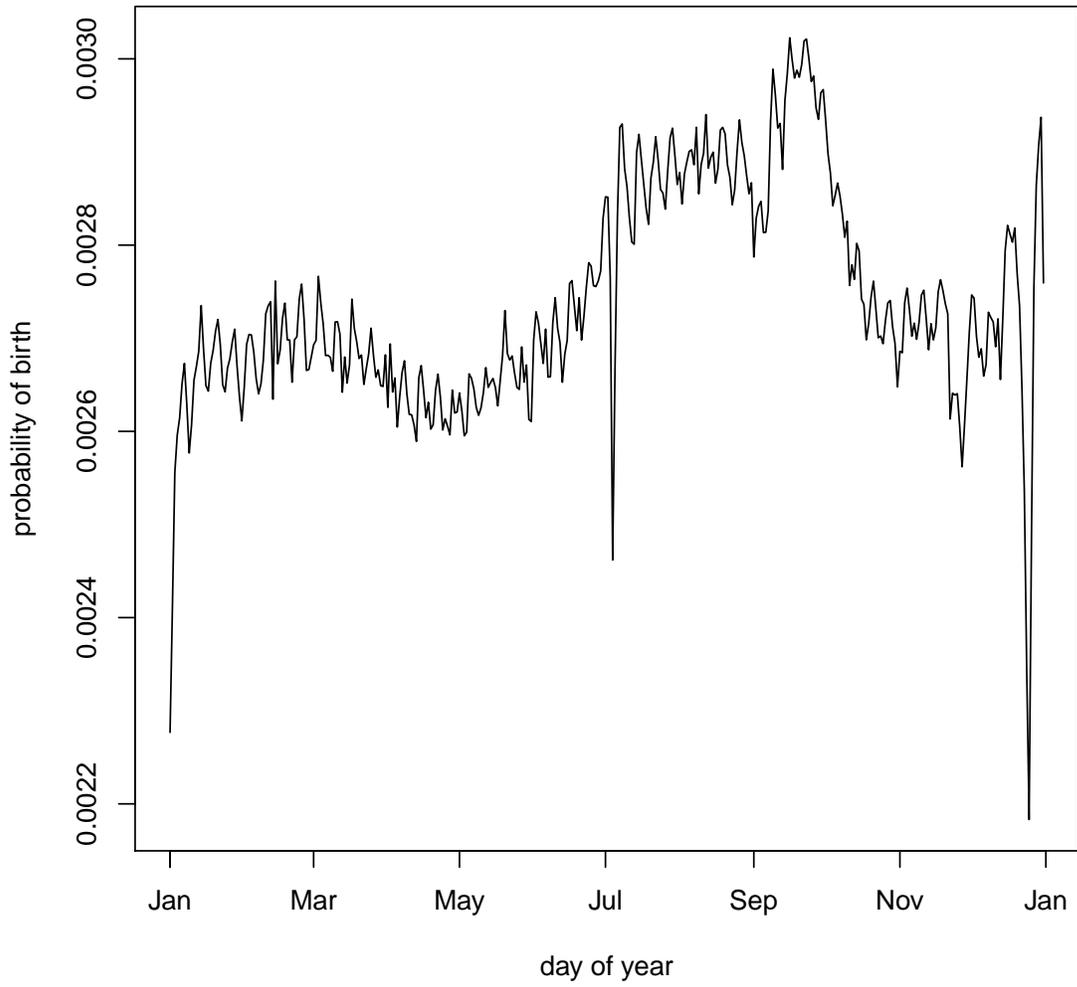


Figure 1: Empirical birth probability in US based on data from 1969-1988

You may use the following code as a starter ([download code here](#)):

```
% The birthday problem
% Author: YOUR NAME
% Date: DATE
clear all
clc

SUID          = 'YOUR SUID';
n             = 23;
D            = 365;
nMonte       = 10000;
samplingType  = 'empirical';    %'uniform'

file = 'results.csv';

for i = 1:nMonte
    tic;
    %-----
    % Draw a realization of n people.
    %-----
    if (strcmp(samplingType,'uniform'))
        birthdays_sample =
    else if(strcmp(samplingType,'empirical'))
        birthdays_sample =
    else
        error('sampling type "%s" not recognized',samplingType)
    end

    %-----
    % compute the number of matches
    %-----
    matched_pairs =

    tElapsed = toc;

    fid = fopen(file,'at');
        fprintf(fid, '%s,%s,%i,%i,%i,%i,%f\n', ...
            SUID,samplingType,n, D, total_pairs, matched_pairs, tElapsed);
    fclose(fid);
end
```

3 Serial Run

- Run your code on your local machine for $n = 150$ and $nMonte = 100,000$, `samplingType='uniform'`.

- Run the same experiment as above on a cluster using `cj run` command.
- Bring the results back to your local machine using `cj get` command.
- Report the total time it took to compute the results on your local machine and on the cluster. Do you see a difference?

4 Parallel Run

- Run your code for a list of n values taken from the list `nlist = 2:365`, a list of sampling types `sampling_list={'uniform','empirical'}` and `nMonte=10,000` on a cluster using ClusterJob's `parrun` command. If you are using `farmshare`, run 364 independent jobs, as the maximum number of jobs per user in the queue is 480. If you have account on `sherlock`, you may run 728 independent jobs up to a maximum of 3000 jobs. How would you change your main script to make it compatible with ClusterJob?
- Harvest your results using `reduce` command.
- Bring the results back to your local machine using `get` command.

Congratulations! You have just executed **7.28 million** individual instances of the birthday problem, and so your `results.csv` must have this many rows. You can check the number of rows by running `wc -l results.csv` in your terminal.

5 Analyze

Using the output of your experiments, `results.csv`, answer the following questions. Use your language of choice.

1. Plot the total compute time versus the number of people n for the two sampling types in two separate plots. Comment on the difference.
2. Approximately, how much time would you need to run these 7.28 million instances serially? How long did it take to run these experiment in parallel? Give an estimated speed up factor.
3. Plot histogram of matched pairs for $n = 23,200$ and the two prior probabilities you assumed for birthday problem. Report and comment on your four plots. Do you see major differences? What type of known probability distributions do you think would fit your data?
4. Plot the following quantities on one plot:
 - the empirical probability of finding a match as a function of the number of people n for uniform sampling.
 - the empirical probability of finding a match as a function of the number of people n for empirical sampling.

- probabilities when sampling is uniform based on analytically available formulas.

At what n do these probabilities exceeds 0.5 (give three numbers for the three plots above). Report the analytical formula you are using. Does the theoretical answer match your computational finding? Comment on the difference between sampling from the uniform distribution versus the empirical distribution.

5. Plot the following quantities on one plot:

- the empirically derived expected number of matches as a function of the number of people n for uniform sampling.
- the empirically derived expected number of matches as a function of the number of people n for empirical sampling.
- expected number of matches when sampling is uniform based on analytically available formulas.

Report the analytical formula you are using. At what n the expected number of matched pairs exceeds 1.0 (give three numbers for the three plots above). Does theoretical answer match your computational finding? Comment on the difference between sampling from the uniform distribution versus the empirical distribution.

6 Report

- Submit your solution and all reproducible codes to Canvas. Please put all your files in a directory and then use
`tar -czvf <your-suid>-stats285-hw1.tar.gz /path/to/results/directory`
 to compress and archive your results. **Upload only this one archive file.**