SUGGESTED PROJECT TOPICS

DATA 602, Winter 2025 (A. R. de Leon)

Project 1: The Distribution of Birthdays and the Birthday Paradox

The ***birthday paradox*** (see, e.g., Szekely, 1986) is about random matches of birthdays among N people and the surprise many express upon learning that it is more likely that there are at least 2 people sharing the same birthday in a small party of $N = 23$ guests than for all the guests to have different birthdays. It is an example widely used to illustrate how our intuition can fail us when assessing the likelihood of seemingly unlikely events. Indeed, when available empirical data are compared with the relevant probabilities, even when the calculations were done based on mathematically convenient assumptions, many so-called coincidences have been shown to occur no more or less likely than expected.

In this project, you will carry out some empirical work to validate the assumption of ***uniform birthday distribution*** on which typical discussions of the birthday paradox rely. Specifically, you will need to carry out the following tasks:

1. You are hosting a party to which you invited N guests. Assume that your guests were all born in a non-leap year, so that the 365 possible birthdays (i.e., day and month) can be numbered as follows:

   January 1 is 1, January 2 is $2, \cdots,$ December 30 is 364, and December 31 is 365.

   Viewed as a ***random experiment***, its ***sample space*** $\mathcal{S}$ is represented as follows:

   $$\mathcal{S} \;=\; \Big\{\boldsymbol{b} : \boldsymbol{b} = (b_1, \cdots, b_N), \text{ such that } b_i = 1, \cdots, 365, \text{ for } i = 1, \cdots, N\Big\},$$

   where $\boldsymbol{b} = (b_1, \cdots, b_N)$ is an (***ordered***) ***n-tuple***, with b_i representing the birthday of guest $i = 1, \cdots, N$, so that there are a total of $365 \cdots 365 = 365^N$ outcomes in $\mathcal{S}$ (i.e., 365^N different ways the N guests can have their birthdays). For a small party of $N = 4$ guests, say, the outcome $(b_1, b_2, b_3, b_4) = (10, 360, 35, 300)$ indicates that Guest 1, Guest 2, Guest 3, and Guest 4 have their respective birthdays on January 10, December 26, February 4, and October 27.

   Next, assume that your guests' birthdays share a common ***discrete uniform distribution*** over the set $\mathcal{X} = \{1, 2, \cdots, 365\}$: The birthday of any guest is ***equally likely*** to fall in any one of the 365 days of his/her birthyear, whence we get $P\big(\text{outcome } \boldsymbol{b}\big) = 1/365^N$, $\forall \boldsymbol{b} \in \mathcal{S}$. Verify the following:

   $$\begin{aligned} P\begin{pmatrix}\text{at least 2 of your } N \text{ guests} \\ \text{share the same birthday}\end{pmatrix} &= P\big(\text{all } N \text{ guests have distinct birthdays}\big), \\ &= 1 - \frac{N!\binom{365}{N}}{365^N}. \end{aligned}$$

Evaluate the above probability of shared birthdays using the functions **factorial()** and **choose()**, for $N = 20(1)29, 30(5)100$, and present the results in a table and/or plot.

2. Write a short **R script** that implements the following Monte Carlo approximation of the probability of shared birthdays in (1). Use the function **sample()** to simulate the outcome $\boldsymbol{b}$ representing the birthdays (as integers in $\mathcal{X}$) of the N guests, for $N = 20(1)29, 30(5)100$. Given outcome $\boldsymbol{b}$, obtain the gap $G(b_i, b_{i'})$ as the number of days separating birthdays b_i and $b_{i'}$, for any pair of birthdays b_i and $b_{i'}$ in $\boldsymbol{b}$, keeping in mind that years are cyclical, so that December 31, the last day of the year, and January 1, the first day of the year, are considered consecutive. For the outcome $\boldsymbol{b} = (10, 360, 35, 300)$ in (1), for example, we have

$$G(b_i, b_{i'}) \quad = \quad \begin{cases} 15 & \text{, for } i = 1, i' = 2, \\ 25 & \text{, for } i = 1, i' = 3, \\ 75 & \text{, for } i = 1, i' = 4, \\ 40 & \text{, for } i = 2, i' = 3, \\ 60 & \text{, for } i = 2, i' = 4, \\ 100 & \text{, for } i = 3, i' = 4, \end{cases}$$

and the smallest gap is $G_{min}(\boldsymbol{b}) = min\big\{G(b_i, b_{'i}) \,:\, b_i, b_{i'} \in \boldsymbol{b}\big\} = G(b_1, b_2) = 15$. Clearly, the event that no 2 guests share a birthday corresponds to $G_{min}(\boldsymbol{b}) = 0$.

Simulate $\boldsymbol{b}$ repeatedly $R = 10000$ times, and obtain $G_{min}(\boldsymbol{b}_r)$, $\forall r$, where $\boldsymbol{b}_r$ is the simulated value of $\boldsymbol{b}$ in repeat $r = 1, \cdots, R$. Then we get

$$\frac{\#\big\{\boldsymbol{b}_r\colon\ G_{min}(\boldsymbol{b}_r) = 0\big\}}{R} \quad \underset{R \to +\infty}{\longrightarrow} \quad 1 - P\left(\begin{array}{c}\text{at least 2 of your } N \text{ guests} \\ \text{share the same birthday}\end{array}\right).$$

Present the approximate probabilities of shared birthdays for $N = 20(1)29, 30(5)100$ along with the exact probabilities in (1) in the best way that you think facilitates comparison of the exact and Monte Carlo approximated probabilities.

3. People who share birthdays are called ***birthmates***. Define the RV Y_i as follows:

$$Y_i \quad = \quad \begin{cases} 1 & \text{, if Guest } i \text{ has at least 1 birthmate,} \\ 0 & \text{, otherwise,} \end{cases}$$

for $i = 1, \cdots, N$, so that RV $Y = \sum_{i=1}^{N} Y_i$ is the ***number of guests who have birthmates*** in a party with N guests. Evaluate $E(Y)$ first analytically, and then numerically, using the same Monte Carlo simulation in (2), for $N = 20(1)29, 30(5)100$. What is the minimum number N of guests at which we expect at least 2 guests to have birthmates? Is this surprising?

4. You will next check the validity of the assumption of a ***uniform birthday distribution***. For this, you will need real data on daily numbers of (live) births in a town, city, province/state, or country, where such data are available. The City of Calgary does not keep records of daily and monthly births in the city. The Province of Alberta only has data on numbers of births by month and municipality, which are available through the province's ***Open Government Program***. Statistics Canada also provides data on the

number of births in Canada by month and by mother's province of residence. Malaysia is the only country, as far as I am aware, that makes data on daily births from 1920 until 2022 available online through its ***Open Data Portal***. I suggest you use data from Malaysia by downloading the data file from https://data.gov.my/data-catalogue/births.

Summarize the data over the M years in terms of the following statistics:

$$\begin{matrix}\text{Average \# of births on}\\ \text{day } j \text{ of month } k\end{matrix} = \sum_{m=1}^{M} \begin{pmatrix}\text{\# of births on day } j\\ \text{of month } k \text{ in year } m\end{pmatrix} \times \frac{\begin{pmatrix}\text{total \# of births on day } j\\ \text{of month } k \text{ over } M \text{ years}\end{pmatrix}}{\begin{pmatrix}\text{total \# of births in}\\ \text{month } k \text{ over } M \text{ years}\end{pmatrix}},$$

$$\begin{matrix}\text{Average daily birth}\\ \text{frequency in month } k\end{matrix} = \sum_{m=1}^{M} \frac{\begin{pmatrix}\text{\# of births on day } j\\ \text{of month } k \text{ in year } m\end{pmatrix}}{\text{\# of days in month } k} \times \frac{\begin{pmatrix}\text{total \# of births in}\\ \text{month } k \text{ over } M \text{ years}\end{pmatrix}}{\begin{pmatrix}\text{total \# of births}\\ \text{over } M \text{ years}\end{pmatrix}},$$

for $j = 1$ (Monday), $\cdots, 7$ (Sunday), and for $k = 1$ (January), $\cdots, 12$ (December), where $m \in \{1, \cdots, M\}$ indicates the years for which data on daily live births were obtained. Present the summaries graphically and on tables. Comment on the validity of the assumption of a uniform birthday distribution for the daily births data.

Observe that the annual daily births distribution is a ***multinomial distribution*** with the 365 days of the year as ***multinomial categories***, with corresponding probabilities $p_1, \cdots, p_{365}$, where $p_j = P\big(\text{a birth occurs on day } j\big)$, for $j = 1, \cdots, 365$. To determine if the ***observed daily births distribution*** is ***statistically significantly different*** from the ***assumed discrete uniform daily births distribution***, with $p_j = 1/365$, $\forall j$, carry out a ***goodness-of-fit test*** of the latter to see how well it fits the former. This can be done using the R function **chisq.test()**, either on data for each of the M or on data aggregated over the M years. For example, you have **chisq.test(x = x.day, p = rep(1/length(x.day), length(x.day)))**, where **x.day** is the 365×1 vector containing the daily numbers of births, in a given year, or averaged over the M years.

5. Is there a weekly cyclical component to the numbers of daily births, similar to what McFarlane et al. (2019) described as "a regular weekly cycle with the numbers of births each day increasing from Mondays to Fridays, with lower numbers of births on Saturdays and the lowest numbers of births on Sundays"?

   It is possible that by "averaging" the observed proportion of daily births on a given day (in a given year) with those of the 6 days following it, the assumption of discrete uniform distribution becomes more tenable, albeit for aggregate weekly births,, since the data are "smoothed" (Nunnikhoven, 1992). Verify that this is indeed the case by plotting the "smoothed" weekly birth frequencies in a given year, or aggregated over M years.

   As in (4), the annual weekly births distribution is a ***multinomial distribution*** with the 52 weeks of the year as ***multinomial categories***, with corresponding probabilities

$P_1, \cdots, P_{52}$ given by

$$P_\ell \;=\; P\big(\text{a birth occurs on week } \ell\big) \;=\; \begin{cases} \sum\limits_{j=7\ell-6}^{7\ell} p_j = \frac{7}{365} & \text{, for } \ell = 1, \cdots, 51, \\ \sum\limits_{j=358}^{365} p_j = \frac{8}{365} & \text{, for } \ell = 52. \end{cases}$$

To determine if the ***observed weekly births distribution*** is ***statistically significantly different*** from the ***assumed discrete "uniform" annual weekly births distribution***^{a}, with $P_\ell = 7/365$, for $\ell = 1, \cdots, 51$, and $P_{52} = 8/365$, carry out a ***goodness-of-fit test*** of the latter to see how well it fits the former. This can be done using the R function **chisq.test(x = x.wk, p = c(rep(7/365, times = 51), 8/365)**, on data **x.wk**, the 52×1 vector containing the weekly numbers of births in each of the M years or aggregated over the M years.

6. As the final task of this project, you will calculate the exact probability of at least 1 shared birthday for a non-uniform distribution of birthdays. For a non-uniform distribution of birthdays, we have $p_j \neq 1/365$, $\exists j$ (i.e., not all the p_j's are equal to 1/365). Consider the extreme case, where $p_j \neq 1/365$, $\forall j$ (i.e., all 365 probabilities are distinct). In a party of N guests, it follows that

$$P\big(\text{all } N \text{ guests have distinct birthdays}\big) \;=\; N! \sum_{\boldsymbol{i} \in \mathcal{A}} \left(\prod_{j=i_1}^{i_N} p_j \right),$$

where $\boldsymbol{i} = (i_1, \cdots, i_N)$ is an ***ordered subset*** of $\mathcal{X} = \big\{1, 2, \cdots, 365\big\}$, and

$$\mathcal{A} \;=\; \Big\{ \boldsymbol{i} : i_1 < \cdots < i_N, \text{ such that } \forall i_h \in \mathcal{X} \Big\},$$

so that there are $\binom{365}{N}$ elements in $\mathcal{A}$, or equivalently, there are $\binom{365}{N}$ terms in the above sum. It is then easy to see that the above probability simplifies to the probability given in (1), if $p_j = 1/365$, $\forall j$ (i.e., all probabilities are the same):

$$N! \sum_{\boldsymbol{i} \in \mathcal{A}} \left(\prod_{j=i_1}^{i_N} p_j \right) \Bigg|_{p_j = 1/365, \forall j} \;=\; N! \sum_{\boldsymbol{i} \in \mathcal{A}} \frac{1}{365^N} \;=\; \frac{N! \binom{365}{N}}{365^N}.$$

If the probabilities are all distinct, calculation of the above sum becomes computationally challenging due to $\binom{365}{N}$, the number of terms to evaluate, being too large, even for moderate N. For example, if $N = 25$, $\binom{365}{25} > 10^{36}$!

We thus need to reduce the number of distinct probabilities from 365 to some smaller number. To do this, we can assume that certain probabilities are equal. For example, it might be the case that $p_1 = \cdots = p_{31} = p(1) > 0$, $p_{32} = \cdots = p_{59} = p(2) > 0, \cdots$, and $p_{335} = \cdots = p_{365} = p(12) > 0$ (i.e., D_k days of month k have the same probability of a birthday falling on one of them, $k = 1, \cdots, 12$). It is also possible that certain months, say, the summer months of June, July, August, and September, and the winter months of December, January, and February, are such that $p(6) = p(7) = p(8) = p(9)$ and $p(12) = p(1) = p(2)$. If this is the case, then there are only the following 7, instead of 365, distinct probabilities: $\pi_1 = p(1) = p(2) = p(12)$, $\pi_2 = p(3), \pi_3 = p(4), \pi_4 = p(5)$, $\pi_5 = p(6) = p(7) = p(8) = p(9)$, $\pi_6 = p(10)$, and $\pi_7 = p(11)$. That is,

$$\begin{aligned}
\pi_1 &= P\begin{pmatrix}\text{birthday falls on 1 of the } \delta_1 = D_1 + D_2 + D_{12} = 90 \\ \text{days between December 1 and February 28}\end{pmatrix},\\
\pi_2 &= P\Big(\text{birthday falls on 1 of the } \delta_2 = D_3 = 31 \text{ days in March}\Big),\\
\pi_3 &= P\Big(\text{birthday falls on 1 of the } \delta_3 = D_4 = 30 \text{ days in April}\Big),\\
\pi_4 &= P\Big(\text{birthday falls on 1 of the } \delta_4 = D_5 = 31 \text{ days in May}\Big),\\
\pi_5 &= P\begin{pmatrix}\text{birthday falls on 1 of the } \delta_5 = D_6 + D_7 + D_8 + D_9 = 122 \\ \text{days between June 1 and September 30}\end{pmatrix},\\
\pi_6 &= P\Big(\text{birthday falls on 1 of the } \delta_6 = D_{10} = 31 \text{ days in October}\Big),\\
\pi_7 &= P\Big(\text{birthday falls on 1 of the } \delta_7 = D_{11} = 30 \text{ days in November}\Big),
\end{aligned}$$

where $\sum_{s=1}^{7} \delta_s = 365$ days. The probability of the N guests all having distinct birthdays is then given by

$$P\Big(\text{all } N \text{ guests have distinct birthdays}\Big) = N! \sum_{\boldsymbol{n} \in \mathcal{B}} \left(\prod_{s=1}^{7} \binom{\delta_s}{n_s} \pi_s^{n_s} \right),$$

where $\boldsymbol{n} = (n_1, \cdots, n_7)$, such that $n_s \geqslant 0$ are integers, and

$$\mathcal{B} = \left\{ \boldsymbol{n} : \sum_{s=1}^{7} n_s = N \right\},$$

so that there are $\binom{N+6}{N}$ elements in $\mathcal{B}$, or equivalently, there are $\binom{N+6}{N}$ terms in the above sum. If $N = 5$, say, then $\binom{N+6}{N} = \binom{11}{5} = 462 \ll 52,521,291,823 = \binom{365}{5}$. The calculations involved in evaluating the required probabilities are now more manageable and could more easily be implemented numerically, say, in R.

Using the ***(average) daily birth frequencies*** for month $k = 1, \cdots, 12$ you calculated from your data in (4), define π_1 as the average of the daily birth frequencies for December, January, and February; π_2 as the daily birth frequency for March; π_3 as that for April; π_4 as that for May; π_5 as the average of those for June, July, August, and September; π_6 as that for October; and π_7 as that for November.

To make the calculations manageable, evaluate the above probability that all your N guests have distinct birthdays, for $N = 5, 6, 7$. Note that to evaluate the sum over $\mathcal{B}$ (i.e., for all N-combinations or subsets of size N from $N + 6$), you need to be able to list all such subsets. For $N = 5$, there are $\binom{N+7-1}{N} = \binom{11}{5} = 462$ such subsets; for $N = 6$, there are $\binom{N+7-1}{N} = \binom{12}{6} = 924$ such subsets; and for $N = 7$, there are $\binom{N+7-1}{N} = \binom{13}{7} = 1,716$

such subsets. Fortunately, function **combn()** in package **combinat** enumerates all such subsets. You will need to calculate as well the same probability for the uniform birthday distribution, like you did in (1), but for $N = 5, 6, 7$, this time. Compare the 2 sets of probabilities graphically. Does the assumption of a uniform birthday distribution make a big difference in the value of the probability?

[a] Note that I put "uniform" in quotes since the weekly probabilities are only roughly, not exactly, all equal to 7/365, due to 365 being ***relatively prime*** to 7.