arXiv:2302.09643v1 [math.HO] 19 Feb 2023

参议院中的三重生日匹配：谎言、该死的谎言，以及 ChatGPT *

Rick Durrett，杜克大学 †

2023年2月21日

摘要

我们的问题是："至少有三名参议院成员在同一天生日的概率是多少？"疫情之前，我曾在全国各地大学举办的几场面向大众的数学讲座中提出过这个问题。受 ChatGPT 在回答这一问题时糟糕表现的启发，我最近重新研究了这个问题，并得到了一个更令人满意的答案；这在很大程度上要归功于我从 Wolfram MathWorld 的一个页面中学到的内容，而这个页面是我通过 Google 搜索找到的。

1 引言

我先从　　我在2018年4月于圣母大学举办的一场人人数学 演讲中给出的答案

- 某一天出生的参议员人数服从二项分布(100,1/365)，因此近似服从均值为 100/365 的泊松分布。
- 某一天恰有三位参议员出生的概率

$$e^{-100/365}(100/365)^3/3! = 0.002606$$

365 天内三人同日生日的预期次数为 $= 0.9512$。如果我们假设事件 $B_j^3 =$ 恰有三位参议员在第 j 天出生相互独立，那么这一次数将服从均值为 0.9512 的泊松分布，且至少有一天出现这种情况的概率为 $1 -$exp$(-0.9512) = 0.6137$

- 某一天恰有两位参议员出生的概率

$$e^{-100/365}(100/365)^2/2! = 0.028536$$

在 365 天中的预期数量为 $= 10.415$。再次假设事件 $B_j^2 =$恰有两位参议员出生在同一天 j 彼此独立，那么数量将服从均值为 10.415 的泊松分布。

*标题的第二行是塞缪尔·克莱门斯关于统计学的一句名言。

†评论、勘误和辱骂请发送至 rtd@math.duke.edu

- 2018年参议院的答案：一组三人同日生日，5月3日：Jim Risch（爱达荷州）、David Vitter（路易斯安那州）、Ron Wyder（俄勒冈州）。十组双重生日：1月7日、3月31日、6月22日、8月24日、9月29日、10月20日和24日、11月17日、12月7日和10日

我很好奇 ChatGPT 会怎么回答，于是输入了：
参议院中至少有三名成员生日相同的概率是多少?
它给出的答案是：

> 一个有 n 名成员的参议院中，至少有三人生日相同的概率约为
>
> $$1-(364/365)^{\binom{n}{3}}-(364/365)^{n}+(364/365)^{\binom{n}{2}}.$$
>
> 这是因为一年有 365 个可能的生日，而任意一名参议员的生日与其他任何人都不同的概率是 364/365。没有两名参议员生日相同的概率是 $(364/365)^{\binom{n}{2}}$，没有三名参议员生日相同的概率是 $(364/365)^{\binom{n}{3}}$。因此，要计算至少三名参议员生日相同的概率，我们需要减去没有两人生日相同的概率和没有三人生日相同的概率，再加上所有人生日都相同的概率。例如，如果参议院有 100 名成员，那么至少有三名成员生日相同的概率约为 0.527，即 52.7%。

为了开始理解这个答案，我们考虑以下说法：“没有两位参议员的生日相同的概率是 $(364/365)^{\binom{n}{2}}$。”如果事件 $A_{i,j}$ 即第 i 位和第 j 位参议员的生日相同，是相互独立的，那么这将是正确的；但它们并不独立。如果 $\{i,j\}\neq\{k,\ell\}$ 是互不相同的整数，那么

$$P(A_{i,j}\cap A_{k,\ell})=\frac{1}{365^2}=P(A_{i,j})P(A_{k,\ell})$$

这表明这些事件是**两两独立**的，但若要实现真正的独立性，我们需要

$$P(A_{1,2}\cap A_{2,3}\cap A_{3,1})=P(A_{1,2})P(A_{2,3})P(A_{3,1})$$

遗憾的是，左边是 $1/(365)^2$，而右边是 $1/(365)^3$。ChatGPT 的答案在这一部分存在的第二个问题是：众所周知，没有两位参议员的生日相同的概率是

$$\frac{365\cdot 364\cdots 266}{(365)^{100}}=3.072\times 10^{-7} \tag{1}$$ ans1

而不是 $(354/365)^{\binom{100}{2}}=1.265\times 10^{-6}$

第二个说法“没有三位参议员生日相同的概率是 $(364/365)^{\binom{n}{3}}$”，则更加令人费解。为了解释这一点，我们来看看 Suraj Regmi 于 2019 年 1 月 26 日发表的博客文章；本文撰写时，这篇文章是 Google 搜索

“生日三元组数学公式”的首条结果。Regmi 推理如下：共有 $C(n,3)$ 个三元组，而要使所有三元组的出生日期都不相同，概率就变为

$$(1-1/(365)^2)^{C(n,3)} \tag{2}$$ ans2

因为这些三元组是独立事件。

虽然双重生日的情况比较微妙，涉及成对独立事件与完全独立事件之间的区别，但三元组的情况并非如此。如果$A_{i,j,k}$ 是参议员 i,j,k 生日相同的概率，那么

$$P(A_{1,2,3}\cap A_{2,3,4}) = \frac{1}{(365)^3} < \frac{1}{(365)^4} = P(A_{1,2,3})\cdot P(A_{2,3,4})$$

当 $n=100$ 式 (2) 的计算结果为 0.29708 时，得出的答案是

$$1-0.29708 = 0.70292 \tag{3}$$ ans3

公平地说，对这个机器人和这位博主，我应该承认自己在圣母大学犯了错：我的事件并非独立事件

$$\begin{aligned} P(B_1^2) &= 0.037155 \qquad & P(B_2^2|B_1^2) &= 0.035676 \\ P(B_1^3) &= 0.003325 \qquad & P(B_2^3|B_1^3) &= 0.003032 \end{aligned}$$

不过，我并没有声称它们是独立事件。我的目标是给出一个简单的近似值，用来计算恰有三位参议员出生在同一天的概率，而这个近似值相当准确。在 Wolfram 生日问题计算器中输入 100 [4]，就能得到参议院生日问题的答案。

$$\begin{array}{ll} \text{at least 2 the same} & 1-3.072\times10^{-7} \\ \text{at least 3 the same} & 0.6459 \end{array}$$

第一个结果是 1 减去 (1)。第二个答案表明 (2) 和 (3) 是错误的。

请注意，此处的 0.6459 是至少有 3 位参议员生日相同的概率，而此前对恰好 3 位参议员生日相同的估计值为 0.6137。在第 2.1 节末尾，读者将看到，使用泊松近似得到的结果，与使用容斥公式的前六项计算出的恰好有 3 位参议员生日相同的概率 0.6140 非常接近。我们对双人同日生日的期望数量的近似值 10.415，与真实期望值 10.3645 非常吻合；这并不令人意外，因为期望值对依赖关系并不敏感。幸运的是，这一性质对我们有利，因为我们的论证意味着双人同日生日的数量服从泊松分布，但如下文的结果所示（见图 1），这一说法并不十分准确。

2 计算

2.1 三重生日的概率

令 T 表示三人同日生日的数量，即一年中恰好有三位参议员在这一天过生日的天数。若用 i 简写为 $(i_1,\ i_2,\ i_3)$，其中 $1 \le i_1 < i_2 < i_3 \le 100$，并令 A_i 表示参议员 $i_1,\ i_2,\ i_3$ 生日相同且没有其他参议员在这一天过生日这一事件，则

$$\begin{aligned} q_1 = \sum_i P(A_i) &= \binom{100}{3}\left(\frac{1}{365}\right)^2\left(\frac{364}{365}\right)^{97}\cdot\frac{365}{365} \\ &= \binom{100}{3}\frac{365(364)^{97}}{(365)^{100}} = 0.9301 \end{aligned}$$

这给出了 ET，以及 $\cup_i A_i$ 概率的一个上界。要利用博纳费罗不等式得到一个下界，我们需要减去

$$\begin{aligned} q_2 = \sum_{i<j} P(A_i \cap A_j) &= \frac{1}{2}\cdot\binom{100}{3}\left(\frac{1}{365}\right)^2\binom{97}{3}\frac{364}{365}\left(\frac{1}{365}\right)^2\left(\frac{363}{365}\right)^{94} \\ &= \frac{1}{2}\cdot\binom{100}{3}\binom{97}{3}\frac{365\cdot 364\cdot(363)^{94}}{(365)^{100}} = 0.3996 \end{aligned}$$

（这里，$<$ 是三元组 (i_1, i_2, i_3) 上的字典序。）

要得到第二个上界，我们需要加上

$$\begin{aligned} q_3 &= \sum_{i<j<k} P(A_i \cap A_j \cap A_k) \\ &= \frac{1}{3!}\cdot\binom{100}{3}\binom{97}{3}\binom{94}{3}\left(\frac{1}{365}\right)^6\frac{364}{365}\cdot\frac{363}{365}\left(\frac{362}{365}\right)^{91} \\ &= \frac{1}{3!}\cdot\binom{100}{3}\binom{97}{3}\binom{94}{3}\frac{P_{365,3}\cdot(362)^{100-9}}{(365)^{100}} = 0.1054 \end{aligned}$$

要得到第二个下界，我们需要减去

$$q_4 = \sum_{i<j<k<\ell} P(A_i \cap A_j \cap A_k \cap A_\ell)$$

为此，我们使用通用公式

$$q_k = \frac{1}{k!}\prod_{k=0}^{k-1}\binom{100-3j}{3}\cdot\frac{P_{365,k}\cdot(365-k)^{100-3k}}{(365)^{100}}$$

由此得到 $q_4 = 0.019153181$、$q_5 = 2.548039\times10^{-3}$ 和 $q_6 = 2.57641\times10^{-4}$。要计算答案

$$
\begin{array}{ll}
\text{upper bound } u_1 & q_1 = 0.931045 \\
\text{lower bound } v_1 & q_1 - q_2 = 0.530545 \\
\text{upper bound } u_2 & v_1 + q_3 = 0.635962 \\
\text{lower bound } v_2 & u_2 - q_4 = 0.616809 \\
\text{upper bound } u_3 & v_2 + q_5 = 0.614261 \\
\text{lower bound } v_3 & u_3 - q_6 = 0.614004
\end{array}
$$

之后，我们需要对 n 个人和一个有 d 天的日历进行这项计算。在这种情况下，容斥中的第 k 项为

$$
q_k(n,d) = \frac{1}{k!}\prod_{k=0}^{k-1}\binom{n-3j}{3} \cdot \frac{P_{d,k}\cdot(d-k)^{100-3k}}{(365)^{100}}
$$

2.2 双重生日的数量

将 k 记作 (k_1, k_2) 的简写，其中 $1 \le k_1 < k_2 \le 100$。令 C_k 表示参议员k_1 和 k_2 的生日相同，且参议院其他成员的生日均与之不同这一事件，并令 D 表示双重生日的数量。利用 $\binom{100}{2} = 4950$

$$
ED = \sum_k P(C_k) = \binom{100}{2}\left(\frac{1}{365}\right)\left(\frac{364}{365}\right)^{98} = 10.3645 \tag{4}
$$

Edouble

与近似值 10.415 相比。误差来自对二项分布进行的泊松分布近似。

$$
365 \cdot P(\text{binomial}(100, 1/365) = 2)365 \cdot 0.028396 = 10.3645
$$

如果 D 服从参数为 λ 的泊松分布，那么 $ED(D-1) = \lambda^2$。用 $j < k$ 表示 $\mathbb{Z}^2$ 上的字典序，并注意到生日巧合是两两独立的

$$
\begin{aligned}
ED(D-1) &= \sum_{j<k} P(C_j \cap C_k) \\
&= \binom{100}{2}\left(\frac{1}{365}\right)\binom{98}{2}\left(\frac{364}{365}\cdot\frac{1}{365}\right)\left(\frac{363}{365}\right)^{96} \\
&= ED \cdot \binom{98}{2}\left(\frac{1}{364}\right)\left(\frac{363}{364}\right)^{96} = ED \cdot 10.027 < (ED)^2
\end{aligned}
$$

其中在最后一步，我们乘以并除以了 $(364/365)^{98}$

Hocking 和 Schweterman [1] 推导出了一个公式，用于计算出现 k 双重生日且没有三人（或更多人）生日相同的 p_k 概率。正如我们在（1）中已经指出的

$$
p_0 = \frac{P_{365,100}}{(365)^{100}} = 3.072 \times 10^{-7}
$$

按照我们讨论三重生日时的论证方式

$$
\begin{aligned}
p_1 &= \binom{100}{2}\frac{1}{365}\cdot\frac{P_{364,98}}{(365)^{98}}\cdot\frac{365}{365} = \binom{100}{2}\cdot\frac{P_{365,99}}{(365)^{100}} \\
p_2 &= \frac{1}{2}\binom{100}{2}\frac{1}{365}\cdot\binom{98}{2}\frac{364}{365}\cdot\frac{1}{365}\cdot\frac{P_{363,96}}{(365)^{96}} \\
&= \frac{1}{2}\binom{100}{2}\binom{98}{2}\cdot\frac{P_{365,98}}{(365)^{100}} \\
p_3 &= \frac{1}{3!}\binom{100}{2}\frac{1}{365}\cdot\binom{98}{2}\frac{364}{365}\cdot\frac{1}{365}\cdot\binom{96}{2}\frac{363}{365}\cdot\frac{1}{365}\cdot\frac{P_{362,94}}{(365)^{94}} \\
&= \frac{1}{3!}\binom{100}{2}\binom{98}{2}\binom{96}{2}\cdot\frac{P_{365,97}}{(365)^{100}}
\end{aligned}
$$

参照 [1] 中的（1）并进行一些代数运算，我们可以看到，一般情况下， k 和 n的公式为

$$
p_k = \frac{1}{k!}\prod_{j=0}^{k-1}\binom{n-2j}{2}\cdot\frac{P_{365,n-k}}{(365)^n}
$$

使用这个公式，我们可以看到

$$
P_k = P_{k-1}\cdot\frac{1}{k}\cdot\binom{n-2(k-1)}{2}\cdot\frac{1}{365-n+k}
$$

如果 p_k 是泊松分布（λ）的 α 倍（回想一下， $\sum_k p_k = P(T=0)$），那么我们就有 $p_k = \lambda p_{k-1}$，因此 $p_k/P(T=0)$不是泊松分布。为了与泊松分布进行比较（见图 1），我们注意到

$$
\sum_k kP_k = 3.87454 \qquad \sum_k p_k = P(T=0) = 0.354135 \tag{5}
$$

notrip

这与 Wolfram Alpha 的结果 $P(T=0) = 1-0.6549$ 一致。因此，我们有 $E(D|T=0) = 10.941$。

2.3 三重生日的数量

McKinney [2]或许是第一个尝试确定以下概率的人：随机选取的 n人中， r 会有相同的生日。为此，他令 X_i、 $1 \le i \le n$在 $\{1,2,\dots M\}$上独立同分布均匀。令 n_i 为样本中出现 i 次的数值个数。他的主要结果是

$$
P_n(n_1, n_2, \dots n_{r-1}) = \frac{n!}{\prod_{i=1}^{r-1} n_i!(j!)^{n_j}}\frac{P_{M,\sum_{i=1}^{r-1} n_i}}{M^n} \tag{6}
$$

McKf

证明。 (6) 中的第二个因子表示 n 个相互独立的均匀分布变量依次出现 n_1个不重复值、 n_2 对成对值和 n_3 组三元组的概率。第一个因子表示这种特定排列顺序中可区分的排列方式数。 □

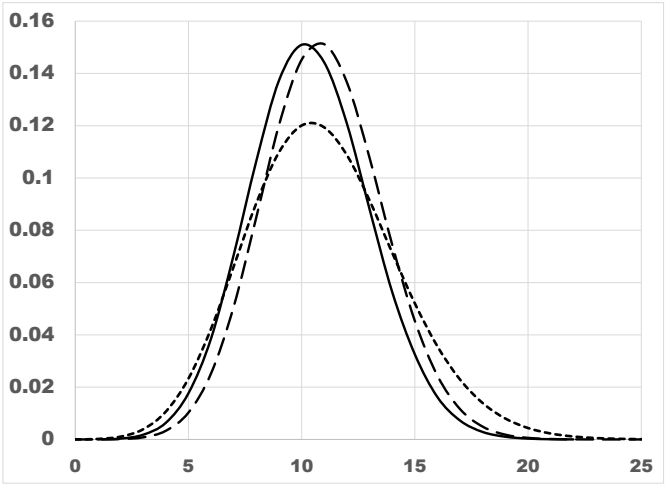

图 1：在不存在三人同日生日的条件下，$p_k =$ 双重生日概率的图表（较长虚线），与均值为 10.941 的泊松分布（较短虚线）的比较。实线给出了对无条件分布进行 100 万次模拟的结果。均值为 10.36，与（4）一致。

如果我们令 $G_r^c =$ 样本中没有任何值重复出现 r 次或更多次，那么

$$P(G_r^c) = \sum \{P_n((n_1, n_2, \ldots n_{r-1}) : \sum_i i n_i = n\}$$

他利用这些结果计算出了以下 $P(G + r)$ 的值。（他的 $E = G_r^c$。）

$r = 2$	$n = 22$	0.4758	$n = 23$	0.5074
$r = 3$	$n = 87$	0.4998	$n = 88$	0.5114
$r = 4$	$n = 186$	0.4758	$n = 187$	0.5033

为了计算三重生日数量的分布，我们写出了一组递推公式，其灵感来自 Wolfram MathWorld [5] 中关于 $Q_i(n,\ d) =$ 的公式，即在一个包含 n 人、拥有 d 个可能的生日的群体中，恰好有 i 人（且不超过此数）共享同一生日的概率。这里，我们令 $\tau_k(n,\ d)$ 表示在一个包含 n 人、拥有 d 个可能的生日的群体中，恰好出现 k 个三人同日生日的概率。

我们已经计算出 $\tau_0(100,\ 365) = 0.386$。

$$\tau_1(100, 365) = \binom{100}{3} \frac{1}{365^2} \left(\frac{364}{365}\right)^{97} \times \tau_0(97, 364)$$

该术语 $(364/365)^{97}$ 表示剩余的97人中，没有人的生日与三人同日生日相同的概率。若以这一事件为条件，则他们的生日在剩余的364种可能性中均匀分布

同样地，在剩余的364种可能性中

$$\begin{aligned}\tau_2(100,365) &= \frac{1}{2!}\binom{100}{3}\frac{1}{365^2}\binom{97}{3}\frac{364}{365}\frac{1}{365^2}\left(\frac{363}{365}\right)^{94}\times\tau_0(94,363)\\ &= \frac{1}{2!}\binom{100}{3}\binom{97}{3}\cdot\frac{P_{365,2}\cdot(363)^{94}}{(365)^{100}}\times\tau_0(94,363)\end{aligned}$$

按照这种模式，我们可以看到

$$\tau_3(100,365) = \frac{1}{3!}\binom{100}{3}\binom{97}{3}\binom{94}{3}\cdot\frac{P_{365,3}\cdot(362)^{91}}{(365)^{100}}\times\tau_0(91,362)$$

$$\tau_4(100,365) = \frac{1}{4!}\prod_{j=0}^{3}\binom{100-3j}{3}\cdot\frac{P_{365,4}\cdot(361)^{88}}{(365)^{100}}\times\tau_0(88,361)$$

这些概率比人们想象的更容易计算。 $\times$ 符号左侧的量是在第 2.1 节中 q_k 计算出的。因此，接下来只需使用博纳费罗不等式计算 $\tau_0(100-3k,\ 365-k)$（并在第四个界处停止）。下表给出了计算结果，并将其与通过 100 万次模拟计算出的值进行比较。

k	q_k	$1-\tau_0(100-3k,\ 365-k)$	计算	模拟
0			0.386	0.380921
1	0.93014	0.58796	0.38325	0.381977
2	0.39960	0.55777	0.17672	0.176321
3	0.10542	0.52719	0.049843	0.049634
4	0.019153	0.49585	0.009656	0.009604
5	2.548×10^{-4}	0.46415	0.001365	0.001375

致谢

本研究部分由美国国家科学基金会概率项目的资助款 DMS 2153429 支持。本文所表达的观点仅代表作者本人，并不一定反映美国国家科学基金会的观点。计算工作由我的学生 Hwai-Ray Tung 完成。他将于 2023年5月 从杜克大学毕业，并于 7月1日 前往犹他大学数学系担任博士后职位。届时，他的导师将成为詹姆斯·B·杜克数学荣休教授。

参考文献

[1] Hocking, R.L., and Schwertman, N.C. (1986) 对生日问题的扩展：恰好出现 k 次匹配。*College Math J.* 17, 315–321[2] McKinney, E.H. (1966) 广义生日问题。*American Math. Monthly* 73, 385– 387[3] Regmi, S. (2019) 生日三胞胎、共同生日与泊松分布 https://surajregmi.medium.com/[4]Wolfram Alpha 生日计算器 https://www.wolframalpha.com/input/?i=birthday+problem+calculator[5] Wolfram MathWorld https://mathworld.wolfram.com/BirthdayProblem.html