西安电子科技大学

网络与信息安全学院

《概率论与数理统计》 课程实验报告

实验名称 生日悖论分析及应用

姓名_____ 学号_____ 专业 信息安全

实验报告内容基本要求

一、问题概述和分析

（1）问题描述

在一个 23 人组成的集体中有两个人在同一天出生的概率是多少呢？从直觉上来说，这个概率是很小的，但是，通过计算机模拟可以得知，这个概率已经超过了50%，与我们的直觉中的概率相差很大，因此这被称为“生日悖论”。那么如何运用概率学知识来准确求出准确概率进而理解生日悖论呢？生日悖论又有何应用？

（2）问题分析

通过具体的概率计算及统计图形绘制，可以帮助理解生日悖论。

（3）实验目的

1. 理解并掌握生日悖论及其概率学运算，理解生日悖论及背后思想。
2. 了解生日悖论的具体应用——生日攻击。

二、实验设计总体思路

2.1、引论

任意两人生日相同的概率的计算，此处采用**古典概率模型**。

设随机试验的样本空间 $\Omega=\{\omega_1, \omega_2, \cdots, \omega_n\}$，由于每个基本事件发生的可能性相同，即

$$P(\{\omega_1\})=P(\{\omega_2\})=\cdots=P(\{\omega_n\})$$

又由于基本事件是两两互不相容的，因此

$$\begin{aligned}1 &= P(\Omega) = P(\{\omega_1\} \cup \{\omega_2\} \cup \cdots \cup \{\omega_n\}) \\ &= P(\{\omega_1\}) + P(\{\omega_2\}) + \cdots + P(\{\omega_n\}) \\ &= nP(\{\omega_i\})\end{aligned}$$

$$P(\{\omega_i\}) = \frac{1}{n},\ i = 1, 2, \cdots, n$$

若事件 $A=\{\omega_{i_1}, \omega_{i_2}, \cdots, \omega_{i_k}\}(1\leqslant i_1, i_2, \cdots, i_k\leqslant n)$，则 A 包含 k 个基本事件，即

$$A = \{\omega_{i_1}, \omega_{i_2}, \cdots, \omega_{i_k}\} = \{\omega_{i_1}\} \cup \{\omega_{i_2}\} \cup \cdots \cup \{\omega_{i_k}\}$$

所以事件 A 的概率为

$$P(A) = \sum_{j=1}^{k} P(\{\omega_{i_j}\}) = \frac{k}{n}$$

从而得到古典概率的计算公式为：

$$P(A)=\frac{k}{n}=\frac{\text{有利于事件 } A \text{ 发生的基本事件数}}{\Omega \text{ 中基本事件的总数}}$$

对于一般的概率计算，还会涉及到概率计算的**加法原理**和**乘法原理**。

即对于一件事，若完成它有k类方法，每类方法中又有$m_1, m_2, \ldots, m_k$种方法，而完成这件事只需其中一种方法，则完成这件事共有$m_1 + m_2 + \cdots + m_k$种方法。

对于一件事，若完成它有n个步骤，每个步骤又有$m_1, m_2, \ldots, m_n$种方法，则完成这件事共有$m_1 * m_2 * \ldots * m_n$种方法。

2.2、实验主题部分

2.2.1、实验设计思路

1） 理论分析

首先，我们假设**一年的长度恒定为 365 天**，排除闰年的影响。

其次，古典概率模型要求样本空间有限，且样本空间内的每个基本事件发生概率相同。但事实上，人口出生随时间的分布并不均匀，故本实验对生日悖论进行论证时，采用理想化的随机样本，不考虑现实生活中不同时间段出生人口数量不同的情况，假设人口的出生和具体的季节、时间无关，是均匀分布的，即**任意一个人出生在 365 天中的任何一天的概率都是相同的**。

在具体计算概率时，可以采用直接法和间接法两种方法。

- **直接法**

  计算此概率的一个方法是直接对两人非同一天生日的结果进行计数。考虑每个人有各不相同的生日的结构要比考虑某两个人不是同一生日的结

构更容易。为此从 365 天中选取 n 天，那么则有 C_{365}^{n} 种情况，可以用 $n!$ 种可能次序中的任何一种将这 n 天分配给这些人。于是在 365^n 种可能出现的生日中，存在$C_{365}^{n} \cdot n!$ 种结构，使得没有两个人具有相同的生日。因此概率是

$$\frac{C_{365}^{n} \cdot n!}{365^n}$$

- **间接法**

  计算此概率的另一个方法是首先算出每个人生日都不相同概率，该事件的对立事件的概率即为所求概率。首先选取 n 个人，那么这 n 个人生日均不相同的情况有 $365 * 364 * \ldots * (366-n)$ 种，即 $\frac{365!}{(365-n)!}$ 种，又总情况有 365^n 种，故这 n 个人生日均不相同的概率为 $\frac{365!}{(365-n)! \cdot 365^n}$，则所求生日悖论的概率为

$$1-\frac{365!}{(365-n)! \cdot 365^n}$$

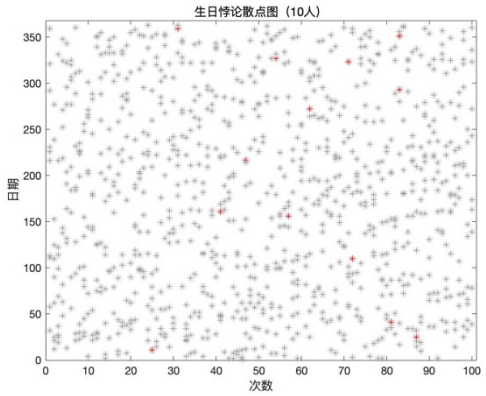
2) 实现方法

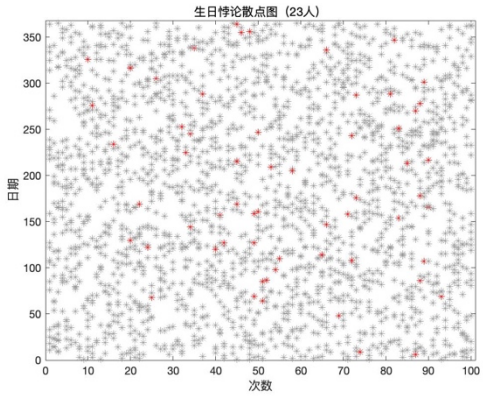
1. 首先做出几个特定人数的概率散点图进行观察，来直观感受
2. 算出人数在 1-365 区间内所有的概率，并做出概率随人数变化的概率图。

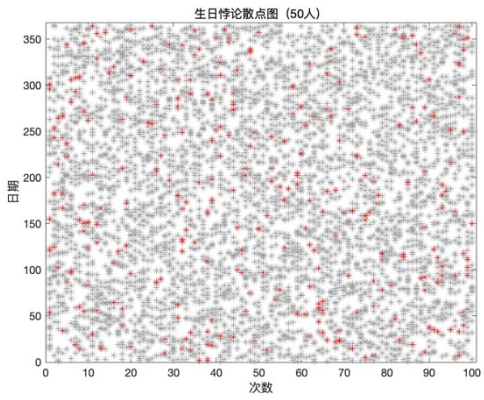
2.2.2、实验结果及分析

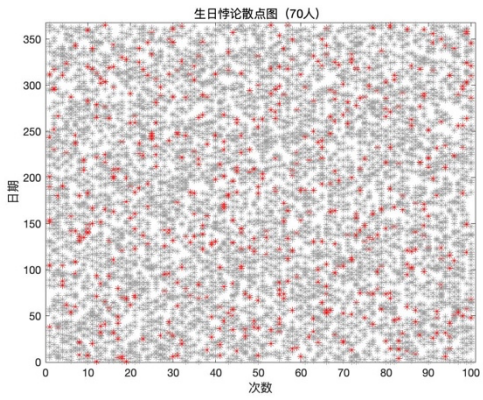
• 散点图：

此处的散点图，横坐标为实验次数，每次都进行 100 次仿真；纵坐标为日期，即 365 天中的具体生日日期。每个横坐标会对应人数个点数，其中红点表示有多个（两个以上）的人的生日在这一天。因此可以大致看红点占所有点的比例来对此人数下至少有两人在生日在同一天的概率进行估计。



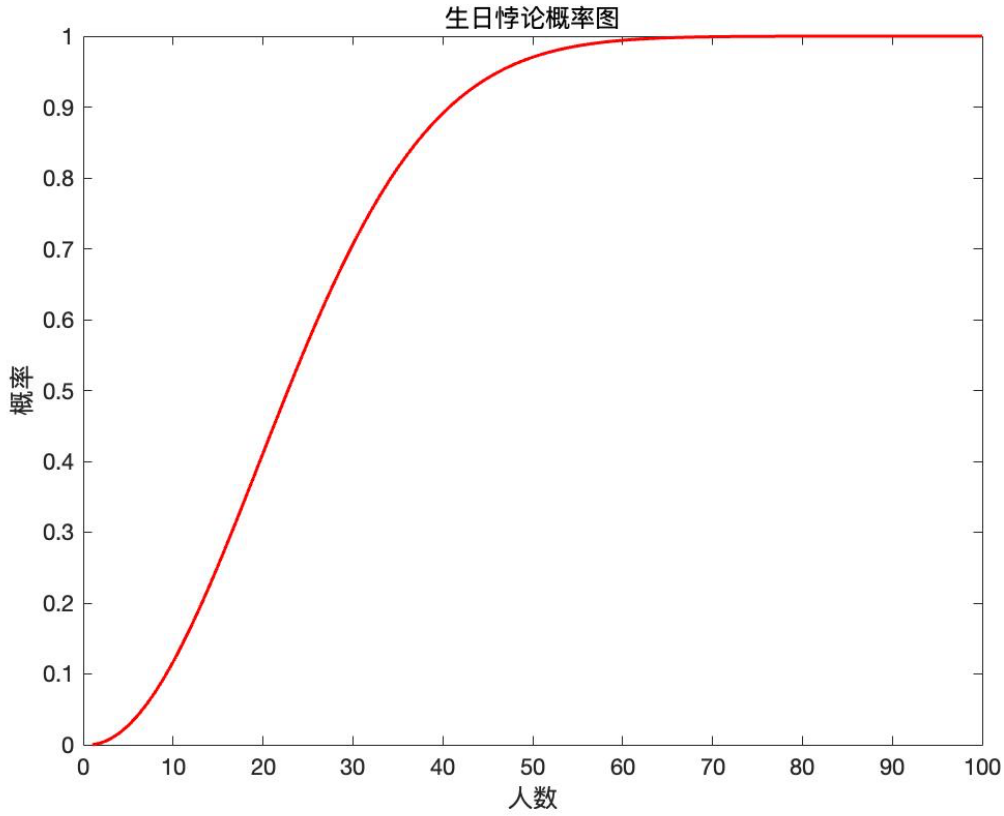






由此可以看到，当人数为 10 时，此概率还比较小，仿真实验中只有个别实验出现了红点，但当人数为 23 时，就已经有接近 50%的仿真实验中出现红点，当人数为 50 时，就几乎所有的仿真实验中都出现了红点，而当人数达到 70 时，只有极个别仿真实验中没有出现红点。

• 概率图



通过概率图，可以更加直观的发现概率随人数的增长的变化情况。

n	$p(n)$
1	0.0%
5	2.7%
10	11.7%

20	41.1%
23	50.7%
30	70.6%
40	89.1%
50	97.0%
60	99.4%
70	99.9%
75	99.97%
100	99.99997%
200	99.99999999999999999999999998%
300	$(100-6*10^{-80})\%$
350	$(100-3*10^{-129})\%$
365	$(100-1.45*10^{-155})\%$
366	100%

通过计算，实际上，想要概率达到 50%，只需要 23 人；想要概率达到 99.9%，则只需要 70 人；当人数大于 70 时，则概率几乎为 100%。

2.2.3、程序及其说明

两个程序均为 matlab 程序，可以直接运行复现。

```
clc, clear, close all
m=100; %仿真次数
N=30;%学生人数
A = zeros(N);
for j = 1:m
    B = zeros(365);
    for i=1:N
        A(i)=unidrnd(365);%生日的365天
        B(A(i))= B(A(i))+1;
        if (B(A(i))>1)
            plot(j,A(i),'*','color',[1 0 0]);
            hold on;
        else
            plot(j,A(i),'*','color',[0.6 0.6 0.6]);
            hold on;
        end
    end
end


xlabel("次数");
ylabel("日期");
title("生日悖论散点图");
axis([0,101,0,368]);
```

```
clc ,clear, close all
P = ones(365, 1);
x = [1:100];
for i = 1:365
    for j = 1:i
        P(i) = P(i) * (365 - j + 1);
    end
    P(i) = 1 - P(i) / 365^i;
end
plot(x,P(x), "Color",[1, 0, 0], 'LineWidth', 1.5);
xlabel('人数');
ylabel('概率');
title('生日悖论概率图');
```

2.3、对生日悖论的应用——生日攻击的说明

生日攻击是利用概率论中的生日悖论思想，找到冲突的 Hash 值，伪造报文，使身份验证算法失效。由于此处的 Hash 函数需要涉及密码学知识，故此处只简单介绍 Hash 函数，并对简单的生日攻击进行说明。

Hash 函数，又名单向散列函数。单向散列函数有一个输入和一个输出，其中输入称为**消息**，输出称为**散列值**。单向散列函数可以根据消息的内容计算出散列值，由于理论上，不同消息会产生不同的散列值，因此散列值可以用来检查消息完整性和身份。

散列值的长度和消息长度无关。无论消息是 1bit，还是 1Mb，或者是 1Gb，采用相同的函数算法时，输出的散列值长度总是固定的。

但由于固定 Hash 函数算法生成的散列值长度是相同的，因此当消息足够多时，总会产生两个消息产生相同的散列值。以 SHA-3 算法举例，它会产生 256bit 的散列值，因此也就最多可以使 2^{256} 个不同的消息产生不同的散列值，当消息多余这个数时，则一定会产生由不同消息但生成相同散列值的情况，因此，如果对 SHA-3 进行暴力破解，100%破解则需要尝试 2^{256} 次，而 50%的概率破解的话，根据我们的直觉，大概为 2^{255} 次，即实验次数的一半。

因此，如果只是寻找任意两条不同的消息，使其产生相同的散列值来对 SHA-3 算法进行攻击，这种情况就和之前介绍的生日悖论十分类似了，即可类比为寻找任意两人生日在同一天的概率。之前已经论证过，对于生日悖论来说，当人数达到 23 人时，任意两人生日在同一天的概率就已经达到了 50%，因此类比可以推出，任意散列值一致的概率比想象中要高。

因此，生日攻击其实是利用了“任意散列值一致的概率比想象中要高”这一特性。此处的“散列值”就相当于“生日”，而“所有可能出现的散列值的数量”就相当于“一年的天数”。

实际上，当我们将此问题一般化，即“假设一年的天数为 Y 天，那么 N 个人的集合中至少有两个人生日一样的概率大于二分之一时，N 至少是多少”，此时，假设 Y 非常大（散列函数 SHA-3 总的散列值构成的空间已经满足了“非常大”），那么近似的计算结果为

$$N = \sqrt{Y}$$

即人数为一年天数的平方根时，即可达到 50%的概率。回到 SHA-3 散列值碰撞的问题上，由上式可知，只需要需要尝试 2^{128} 次，即有 50%的概率成功破解散列函数。相比我们直觉上认为的 2^{255} 次，运用生日攻击的思想进行破解需要的次数比我们的直觉小得多。

2.4、体会

直觉并不总是可信，要学会以科学的计算方法，树立科学的思想方法和观念，打破直觉带来的错误印象。