集卡问题的建模与通解

摘要

本文对于生活中常见的卡片收集问题进行分析，建立了集满卡片所需总卡牌的数量与卡片的种类以及每种卡片出现的概率的分布模型。

针对问题一：首先将问题简化为可放回的摸球问题，第一问所求即为求第一次摸出四种球的次数的期望值。由于 4 这个数字不大，因此可以采用贝叶斯公式和全概率公式的思想，采用递归的方法来计算。最终得到一个人须至少买 22 袋方便面才可以获得抽奖资格。

针对问题二：通过问题一的计算过程，猜想当某一种卡片的概率很低时，即投放比例的方差增大时，所求解的期望值会大大增加。因此通过调整卡片中各人物的投放比例，使用 c 语言和 matlab 模拟蒙特卡罗法模拟实际抽卡情况，观测集齐一套卡片的抽卡次数的频率；并由期望值与方差之间的拟合关系可知，方差越大，期望值越大。

针对问题三：对于有 N 种卡片人物、每种卡片人物投放比例不同的情况，采用概率论与数理统计中的方法，使用 *Maximum-Minimum identity*，直接计算其期望值：

$$\int_0^{+\infty}\{1-\prod_{i=1}^{N}(1-e^{-n_i x})\}dx$$

该期望即为至少买多少袋方便面。

关键字：贝叶斯公式；概率统计模型；蒙特卡罗法；Min-Max 容斥

1. 问题重述

1.1 问题的背景

在日常生活中，商家为了促销会推出各种各样的集卡活动。例如去餐厅就餐时会送给你一个十二生肖的小玩具，当集其十二生肖时，店家会免单一次。但十二生肖的投放比例可能相同，也可能不同。因此考虑去就餐多少次能够获得一次免单机会可以看出店家的活动是否实惠，本题与之类似。

1.2 问题的重述

问题一：在给定了师徒四人卡片的投放比例后，求需要集卡多少次才能把师徒四人全部收集，获得抽奖机会。

问题二：在第一问的基础上，考虑改变卡片的投入比例，会对集卡次数的期望值产生怎样的影响。

问题三：将卡片的种类数进行拓展，在前两文的基础上，考虑在有 N 种卡片，每个卡片的投放比例不同的情况下，集卡次数的期望值为何值。

2. 问题分析

问题一：由于各卡片的投放比例已知，可直接使用全概率公式，并递归调用，求解购买方便面袋数的期望值。

问题二：本题的模型与问题一相同，但是由于要考察问题一的答案与不同人物卡片投入比例的变化关系，所以若再次采用全概率公式，则结果会异常的复杂。因此采用采用 c 语言和 matlab 进行模拟集卡行为，画出频率分布图。并从频率分布图中得到期望值的变化方向。

问题三：基本模型仍然不变，但是卡片的种类拓展到 n 种，由于 n 不确定。使用代码模拟可能会有特殊情况出现，因此我们采用 *Maximum-Minimum identity* 直接进行推导计算，得出该类问题的一个通解。

3. 模型假设

1. 在集卡时，假设方便面货量充足，各种卡片的比例不会发生改变。
2. 用各种卡片的投放比例作为其概率
3. 认为可以采用计算机生成随机数的形式模拟集卡行为。
4. 在进行模拟时，认为计算机产生的随机数的精度可以满足实验的要求。

4.符号说明

符号	含义
A，B，C，D	分别代表八戒、沙僧、唐僧、悟空卡片
E(A，B)	当首次抽到 A 和 B 时抽奖次数的期望值
E(i)	事件 i 的期望
P(i)	抽到卡片种类 i 的概率

n_i	卡片 i 的投放比例，即概率
$S^2(i)$	Case i 的方差
X_i	获得第 i 个类型的卡片需要购买的方便面袋数

5. 模型的建立与求解

5.1 问题一的模型建立与求解

5.1.1 问题一的模型建立及求解

(1)E(A)=P(A)*1+(1-P(A))*(E(A)+1)⇒E(A)=1/P(A)

公式解释：使用全概率公式与贝叶斯公式的思想：如果第一次即抽到 A，概率为 P(A);若第一次抽到的不是 A,概率为 1-P(A),接下来遇到再次抽到 A 需要的抽卡次数为 1+E(A).

由此可得：E(A)=2, E(B)=10/3, E(C)=20/3, E(D)=20.

(2)E(A,B)=P(A)(1+E(B))+P(B)(1+E(A))+(1-P(A)-P(B))(1+E(A,B))

公式解释：若第一次抽到 A，概率为 P(A),则首次集齐 A、B 的抽卡次数为 1+E(B)；若第一次抽到 B，概率为 P(B),则首次集齐 A、B 的抽卡次数为 1+E(A).若第一次既不是 A，也不是 B，概率为 1-P(A)-P(B),则首次集齐 A、B 的抽卡次数为 E(A,B).

解得：$E(A,B)=\frac{P^2(A)+P^2(B)+P(A)P(B)}{P(A)P(B)(P(A)+P(B))}$ =49/12

同理可得：

E(A,C)=278/39　　E(A,D)=222/11

E(B,C)=70/9　　E(B,D)=430/21　　E(C,D)=65/3

(3)同理可得：

E(A,B,C)=P(A)(1+E(B,C))+P(B)(1+E(A,C))+P(C)(1+E(A,B))+…

P(D)(1+E(A,B,C))

E(A,B,C)=3749/468　　E(A,B,D)=20.5845

E(B,C,D)=1381/63　　E(A,C,D)=21.739

(4)根据上述公式解释，同理可得：

E(A,B,C,D)=P(A)(1+E(B,C,D))+P(B)(1+E(A,C,D))+…

P(C)(1+E(A,B,D))+P(D)(1+E(A,B,C))

解得：E(A,B,C,D)=21.97

5.1.2 最终结果

又购买的方便面袋数为整数，所以问题一的答案为：**22**。即一个人须至少购买购买 22 袋方便面才能获得抽奖资格。

5.2 问题二的模型建立与求解：

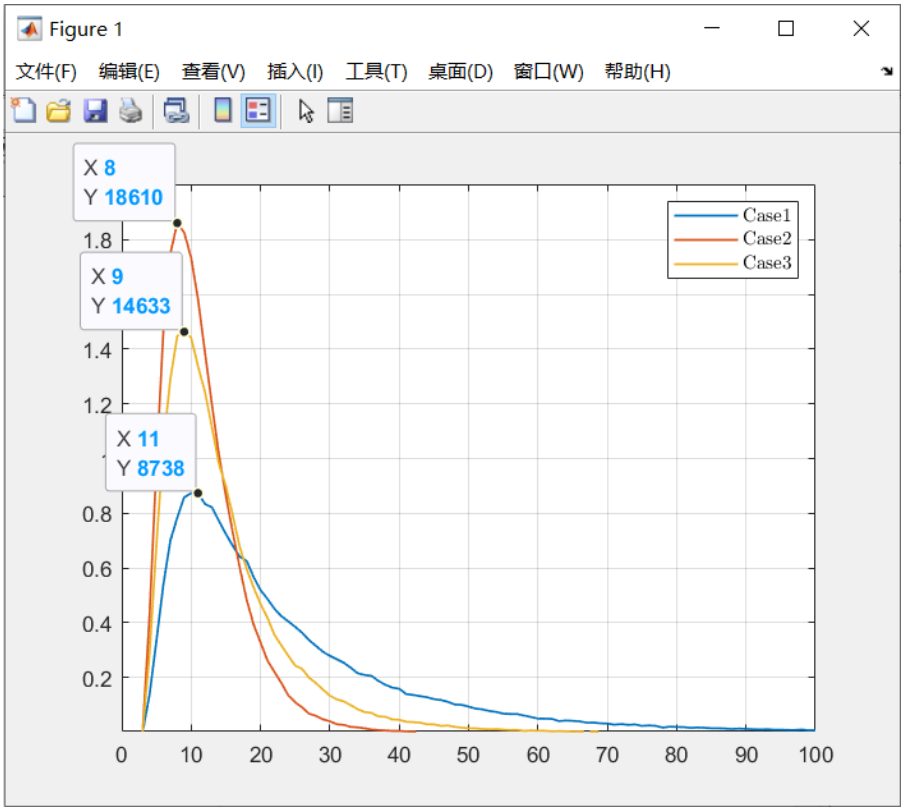
5.2.1 建立模型与求解

猜想抽卡的次数与卡片中最稀有的卡片的投放比例有关，即与方差有关。因此我们主要改变四张卡片的离散程度：

	A	B	C	D
Case1	0.5	0.3	0.15	0.05
Case2	0.3	0.3	0.2	0.2
Case3	0.4	0.3	0.2	0.1

做出三者的频率分布图如下所示：

曲线从上至下依次为：Case2，Case3，Case1



5.2.2 对于问题猜想的验证。

由图，随着各种情况方差的增大，曲线逐渐右移，且峰值降低，这意味着集齐一套卡牌的期望值增加。为了观察方差与期望之间的关系，添加两组数据：

1.P(A)=P(B)=P(C)=P(D)=0.25

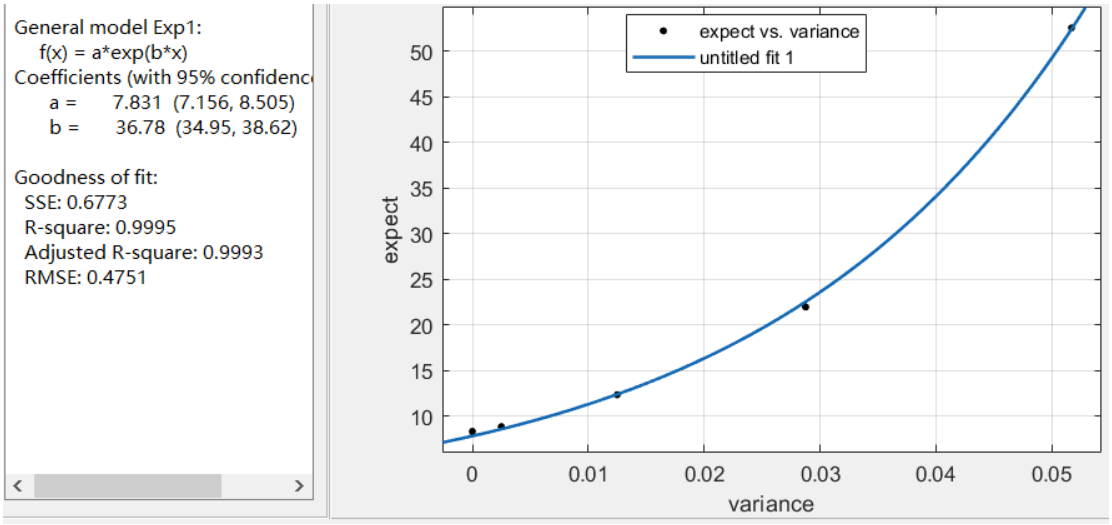
2.P(A)=0.6,P(B)=0.3,P(C)=0.08,P(D)=0.02

计算后可得知，5 种情况的期望值及方差如下：

	期望	方差
Case1	21.97	0.02875
Case2	8.86	0.0025
Case3	12.36	0.0125
Case4	8.33	0
Case5	52.58	0.0517

与假设情况一致，若方差增大，则期望值增大，下面我们对其进行检验。

对其进行拟合，可观察到使用指数函数拟合优度极高



5.2.3 最终结果

改变卡片的投放比例会对抽卡次数的期望产生影响；并且投放比例的方差越大，集卡次数的期望值越大。

5.3 问题三的模型建立与求解

5.3.1min-max 容斥

首先我们给出引理：给定集合 G，设 max(G)为 G 中的最大值，min(G)为 G 中的最小值，则：

$\max(G)=\sum_{T\in G}(-1)^{|T|-1}\min(T)$

证明：

构造容斥系数 f(x),使得 $\max(G)=\sum_{T\in G}f(|T|)\min(T)$

考虑第 x+1 大的元素会被统计到的贡献为$\sum_{i=0}^{x} C_x^i f(i+1)$=[x=0]

进行二项式反演:$f(x+1)=\sum_{i=0}^{x}(-1)^{x-i}C_x^i$[i=0]= $(-1)^x$

故$f(x)=(-1)^{x-1}$

综上，max(G)=$\sum_{T\in G}(-1)^{|T|-1}$min(T)

因此我们得到，max(a,b)=a+b-min(a,b)

max(a,b,c)=a+b+c-min(a,b)-min(b,c)-min(a,c)+min(a,b,c)

5.3.2 问题三的模型建立与解

收集完成一整套卡片人物的抽卡次数$x=\max\{x_1,x_2,\cdots,x_N\}$,，因此我们采用 min-max 容斥：

$$E(x)=E\{(x_1,x_2,\cdots,x_N)\}$$

$$=\sum_i E(x_i)-\sum_{i<j}E\{min(x_i,x_j)\}+\sum_{i<j<k}E\{min(x_i,x_j,x_k)\}-\cdots$$

$$\cdots+(-1)^{N+1}E\{\min(x_1,x_2,\cdots,x_N)\}$$

$$=\sum_i\frac{1}{n_i}-\sum_{i<j}\frac{1}{n_i+n_j}+\sum_{i<j<k}\frac{1}{n_i+n_j+n_k}-\cdots$$

$$\cdots+(-1)^{N+1}\frac{1}{n_1+n_2+\cdots+n_N}$$

又$\int_0^{+\infty}e^{-nx}dx=\frac{1}{n}$

所以 1-$\prod_{1=1}^{N}(1-e^{-n_ix})$=$\sum_i e^{-n_ix}$-$\sum_{i<j}e^{-(n_i+n_j)x}+\cdots$

$$\cdots+(-1)^{N+1}e^{-(n_1+\cdots+n_N)x}$$

因此我们得到最终表达式：$E(X)=\int_0^{+\infty}\{1-\prod_{1=1}^{N}(1-e^{-n_ix})\}dx$

5.3.3 最终结果

若卡片人物共 N 个，每个人物投放比例为n_i,i=1,…，N，则一个人须购买

$$\int_0^{+\infty}\{1-\prod_{i=1}^{N}(1-e^{-n_ix})\}dx$$

袋方便面才能获得抽奖资格。

6. 模型评价与推广

6.1 模型优点：问题一和问题三的模型，分别使用了基于贝叶斯公式的概率模型、和基于 min-max 容斥的期望模型，过程严谨，计算结果准确，并且给出了此类问题的一个确定解。问题二的模型采取蒙特卡洛法进行模拟，以图表形式给出结果，较为直观。

6.2 模型缺点：问题一的概率模型求解过程复杂，需要进行大量计算，当卡片种类增多时，使用意义不大。模型二的数据由于模拟次数只有 20 万次，因此频率分布曲线仍有折线部分，存在误差；且只考虑了方差对其的影响，模型过于简化。

6.3 模型推广改进：使。用问题一的模型计算问题三时，由于是递归算法，所以可以类似模型三进行公式的推导。模型二在进行蒙特卡罗法进行模拟时，需要使用更密集的数据进行拟合，减小误差。

参考文献

[1] 姜启源，谢金星，叶俊，数学模型.第五版[M]，北京：高等教育出版社，2018

[2] 盛骤, 谢式千, 潘承毅. 概率论与数理统计.第四版[M]，北京：高等教育出版社, 2008.

[3] Ross.Sheldon, *A first course in probability*[M], 北京：人民邮电出版社, 2007.