赠券收集问题

温馨提示：读懂本文需要学习过浙大版《概率论与数理统计》或类似课程。

一、问题定义

假设有 n 种赠券，每种赠券获取机率相同，而且赠券亦无限供应。若取赠券 t 张，求能集齐 n 种赠券的机率。

二、数学期望

对于赠券收集问题，所有赠券都集齐的最小抽取次数的数学期望为

$$\mathrm{E[X]} = \mathrm{nH_n}$$

其中 Hn 为第 n 个调和数，即 $\mathrm{Hn} = \sum_{k=1}^{n}\frac{1}{k}$ 。

证明：

设 X 为表示 n 种赠券都至少被抽中一次时的抽取次数的离散随机变量。

设 $\mathrm{X_k}$ 为表示 k – 1 种不同赠券被抽中后，为抽中第 k 种赠券所需的抽取次数的离散随机变量。显然 $\mathrm{X_1} = 1$ ，因为第一个被抽中的赠券与所有已抽中赠券相比，一定是独一无二。

由数学期望的线性可加性[1]，得

$$\mathrm{E[X]} = \sum_{k=1}^{n} E[X_k]$$

当 k - 1 种不同的赠券被抽中后，还有 n – k + 1 种赠券还没被抽中。设 $\mathrm{A_k}$ 为表示在下一次抽奖中，抽中新品种赠券的随机事件。根据题设，每次抽取每种赠券获取机率相同，可得

$$\mathrm{P(A_k)} = \frac{n-k+1}{n}$$

且 $\mathrm{X_k}$ 服从几何分布，即 $\mathrm{X_k} \sim GE(\frac{n-k+1}{n})$ 。

所以 $\mathrm{E[X_k]} = \frac{\mathrm{n}}{n-k+1}$ 。

所以 $\mathrm{E[X]} = \sum_{k=1}^{n} E[X_k] = \sum_{k=1}^{n}\frac{\mathrm{n}}{n-k+1} = n\sum_{k=1}^{n}\frac{1}{k} = \mathrm{nH_n}$ 。

数学期望的估计公式：

调和数没有解析解，但是有估计公式。

当 n 特别大时，第 n 个调和数的估计值为

$$\mathrm{H_n} \approx \ln n + \gamma + \frac{1}{2n}$$

其中 γ 为欧拉-马斯克若尼常数，

$$\gamma \approx 0.57721\ 56649\ 01532\ 86060\ 65120\ 90082\ 40243\ 10421\ 59335$$

因此，赠券收集问题的数学期望的估计公式为

[1] 见浙江大学《概率论与数理统计》（第四版） 高等教育出版社，第 98 页，数学期望的性质 3

$$\mathrm{E}[\mathrm{X}] \approx \mathrm{n}(\ln n+\gamma)+\frac{1}{2}$$

当 $n \geq 5$ 时，上述估计公式精确到小数点后一位。

三、方差

对于赠券收集问题，所有赠券都集齐的最小抽取次数的方差为

$$\mathrm{D}(\mathrm{X})=\mathrm{n}^2 \sum_{k=1}^n \frac{1}{k^2}-n H_n$$

其中 Hn 为第 n 个调和数，即 $\mathrm{Hn}=\sum_{k=1}^n \frac{1}{k}$ 。

证明：

设 X 为表示 n 种赠券都至少被抽中一次时的抽取次数的离散随机变量。

设 $\mathrm{X_k}$ 为表示 k – 1 种不同赠券被抽中后，为抽中第 k 种赠券所需的抽取次数的离散随机变量。

因为 $\mathrm{X_k} \sim GE(\frac{n-k+1}{n})$ ，所以

$$\mathrm{D}(\mathrm{X_k})=\frac{1-\frac{n-k+1}{n}}{\left(\frac{n-k+1}{n}\right)^2}=\frac{\mathrm{n}(\mathrm{k}-1)}{(n-k+1)^2}$$

因为 $\mathrm{X_k}$ 之间相互独立，所以

$$\mathrm{D}(\mathrm{X})=\sum_{k=1}^n D(X_k)$$

$$=\sum_{k=1}^n \frac{\mathrm{n}(\mathrm{k}-1)}{(n-k+1)^2}$$

$$=\mathrm{n} \sum_{k=1}^n \frac{n-k}{k^2}$$

$$=\mathrm{n}^2 \sum_{k=1}^n \frac{1}{k^2}-n \sum_{k=1}^n \frac{1}{k}$$

$$=\mathrm{n}^2 \sum_{k=1}^n \frac{1}{k^2}-n H_n$$

得证。

方差的估计公式：

由巴塞尔问题可得$\sum_{k=1}^{\infty} \frac{1}{k^2}=\frac{\pi^2}{6}$ 。

所以当 n 特别大时，$\sum_{k=1}^n \frac{1}{k^2} \approx \frac{\pi^2}{6}$ 。

所以方差的估计公式为

$$\mathrm{D}(\mathrm{X}) \approx \frac{\pi^2}{6} n^2-\mathrm{n}(\ln n+\gamma)-\frac{1}{2}$$

四、累积分布函数

为了计算双侧置信区间和单侧置信区间，需要计算累积分布函数。

在计算累积分布函数前，先介绍第二类 Stirling 数。第二类 Stirling 数实际上是集合的一个拆分，表示将 n 个不同的元素拆分成 m 个集合的方案数，记为 $\mathrm{S(n,m)}$ 或 $\left\{ n \atop m \right\}$ 。

第二类 Stirling 数的**递推公式**为

$$\mathrm{S(n+1,m)=S(n,m-1)+m\cdot S(n,m)}$$

证明：

为求得递推公式，可从定义出发，考虑第 n + 1 个元素的情况，即假设要把 n + 1 个元素分成 m 个集合。

情况 1：如果 n 个元素构成了 m - 1 个集合，那么第 n + 1 个元素需要单独构成一个集合，方案数为 $\mathrm{S(n,m-1)}$。

情况 2：如果 n 个元素已经构成了 m 个集合，则需将第 n + 1 个元素插入到任意一个集合，方案数 $\mathrm{m\cdot S(n,m)}$。

综合两种情况，可得 $\mathrm{S(n+1,m)=S(n,m-1)+m\cdot S(n,m)}$ 。

第二类 Stirling 数的**通项公式**为

$$\mathrm{S(n,m)}=\frac{1}{\mathrm{m!}}\sum_{i=0}^{m}(-1)^i\cdot C_m^i\cdot(m-i)^n$$

证明：

第二类 Stirling 数$\mathrm{S(n,m)}$为将 n 个不同的小球放入 m 个相同的盒子里（n≥m）且不允许有空盒的方案数。

现假设将 n 个不同的小球放入 m 个不同的盒子里，允许有空盒的方案总数为 $\mathrm{m^n}$ 。

设 $\mathrm{A_i}$ 表示第 i 个盒子为空时的方案总数。根据假设情形的定义，可得下列三个公式：

$$|\mathrm{A_i}|=(m-1)^n$$

$$\sum_{1\le i<j\le m}\left|A_i\cap A_j\right|=C_m^2(m-2)^n$$

$$\sum_{1\le i_1<i_2<\cdots<i_k\le m}\left|A_{i_1}\cap A_{i_2}\cap\cdots\cap A_{i_k}\right|=C_m^k(m-k)^n$$

根据容斥原理[2]，可得

$$|\mathrm{A_1\cup A_2\cup\cdots\cup A_m}|$$

$$=\sum_{k=1}^{m}(-1)^{k-1}\sum_{1\le i_1<i_2<\cdots<i_k\le m}\left|A_{i_1}\cap A_{i_2}\cap\cdots\cap A_{i_k}\right|$$

$$=\sum_{k=1}^{m}(-1)^{k-1}C_m^k(m-k)^n$$

所以没有空盒的方案数为

$$\left|\overline{A_1}\cup\overline{A_2}\cup\cdots\cup\overline{A_m}\right|$$

$$=m^n-|\mathrm{A_1\cup A_2\cup\cdots\cup A_m}|$$

[2] 容斥原理为$|\mathrm{A_1\cup A_2\cup\cdots\cup A_m}|=\sum_{1\le i\le m}|A_i|$-$\sum_{1\le i<j\le m}\left|A_i\cap A_j\right|+\sum_{1\le i<j<k\le m}\left|A_i\cap A_j\cap A_k\right|-\cdots+(-1)^{m-1}\left|\mathrm{A_1\cap A_2\cap\cdots\cap A_m}\right|$。容斥原理可用数学归纳法证明，具体证明略。

$$= C_m^0(m-0)^n - \sum_{k=1}^{m}(-1)^{k-1} C_m^k (m-k)^n$$

$$= \sum_{k=0}^{m}(-1)^k C_m^k (m-k)^n$$

当盒子相同时，方案数为上式处以盒子的全排列数 m! 。

所以$\mathrm{S(n,m)} = \frac{1}{\mathrm{m!}}\sum_{i=0}^{m}(-1)^i \cdot C_m^i \cdot (m-i)^n$ 。

得证。

第二类 Stirling 数的性质：

1. $\mathrm{S(n,0)} = 0^{\mathrm{n}}$
2. $\mathrm{S(n,1)} = 1$
3. $\mathrm{S(n,2)} = 2^{n-1} - 1$
4. $\mathrm{S(n,n)} = 1$
5. $\mathrm{S(n,n-1)} = C_n^2$
6. $\mathrm{S(n,n-2)} = C_n^3 + 3C_n^4$

上述性质均可用小球盒子模型证明，证略。

对于赠券收集问题而言，可将 t 次抽取看作 t 个小球，n 种赠券看作 n 个不同盒子，则有 $\mathrm{n!\,S(t,n)}$ 种方案使得每个盒子不为空，即在 t 次抽取后，每种赠券至少抽得一张。又因为 t 次抽取的方案总数为 $\mathrm{n^t}$，所以赠券收集问题的累积分布函数为

$$F(t,n) = \frac{n!\,S(t,n)}{n^t} = \frac{\sum_{i=0}^{n}(-1)^i C_n^i (n-i)^t}{n^t}$$

上式的具体含义为 t 次抽取后，n 种赠券中的每种赠券至少已经抽得一张的概率。

五、概率密度函数

记抽取次数为 T，赠券种类为 N，则赠券收集问题的概率密度函数为

$$P\{T=t, N=n\} = \frac{n!\,S(t-1,n-1)}{n^t}$$

上式的具体含义为在第 t 次抽取时，恰好抽中第 n 种赠券的概率。

证明：

因为在第 t 次抽取时，恰好抽中第 n 种赠券，所以前面的 t – 1 次抽取一共抽中 n - 1 种赠券，方案总数为 $(n-1)!\,S(t-1,n-1)$ 。因为第 t 次抽取抽中的可能是 n 种赠券中的任意一种，所以在第 t 次抽取时，恰好抽中第 n 种赠券的方案总数为

$$\mathrm{n} \times (n-1)!\,S(t-1,n-1) = \mathrm{n!}\,S(t-1,n-1)$$

又因为 t 次抽取的方案总数为 $\mathrm{n^t}$，所以赠券收集问题的概率密度函数为

$$P\{T=t, N=n\} = \frac{n!\,S(t-1,n-1)}{n^t}$$

得证。

赠券收集问题的概率密度函数同样可由累积分布函数推导得到：

$$P\{T=t, N=n\} = F(t,n) - F(t-1,n)$$

$$= \frac{\sum_{i=0}^{n}(-1)^i C_n^i (n-i)^t}{n^t} - \frac{\sum_{i=0}^{n}(-1)^i C_n^i (n-i)^{t-1}}{n^{t-1}}$$

$$
\begin{aligned}
&= \frac{\sum_{i=0}^{n}(-1)^i C_n^i (n-i)^t}{n^t} - \frac{\sum_{i=0}^{n}(-1)^i C_n^i (n-i)^{t-1}\cdot n}{n^t} \\
&= \frac{\sum_{i=0}^{n}(-1)^i C_n^i [(n-i)^t - n(n-i)^{t-1}]}{n^t} \\
&= \frac{\sum_{i=0}^{n}(-1)^{i+1} C_n^i (n-i)^{t-1}\cdot i}{n^t} \\
&= \frac{(-1)^{0+1} C_n^0 (n-0)^{t-1}\cdot 0 + \sum_{i=1}^{n}(-1)^{i+1} C_n^i (n-i)^{t-1}\cdot i}{n^t} \\
&= \frac{\sum_{i=1}^{n}(-1)^{i+1} C_n^i (n-i)^{t-1}\cdot i}{n^t} \\
&= \frac{\sum_{i=0}^{n-1}(-1)^{i} C_n^{i+1} (n-i-1)^{t-1}\cdot (i+1)}{n^t} \\
&= \frac{\sum_{i=0}^{n-1}(-1)^{i}\ \frac{n!}{(i+1)!\,(n-i-1)!}(n-i-1)^{t-1}\cdot (i+1)}{n^t} \\
&= \frac{\sum_{i=0}^{n-1}(-1)^{i}\ \frac{(n-1)!}{i!\,(n-1-i)!}(n-i-1)^{t-1}\cdot n}{n^t} \\
&= \frac{\sum_{i=0}^{n-1}(-1)^{i}\ C_{n-1}^i (n-1-i)^{t-1}\cdot n}{n^t} \\
&= \frac{n!\cdot \frac{1}{(n-1)!}\sum_{i=0}^{n-1}(-1)^{i}\ C_{n-1}^i (n-1-i)^{t-1}}{n^t} \\
&= \frac{n!\,S(t-1,n-1)}{n^t}
\end{aligned}
$$