

关于对最优停止问题的研究

赵 水

(浙江工业大学 计算机学院, 浙江 杭州 310014)

37 法则是指, 在统计学中, 对样本数量固定 (每个样本只出现一次) 的情况下, 面对选择其中最优的那个样本这个问题时, 如何保证选择的样本尽可能最优的一种策略。就是把样本总量的前 37% 的样本做为参考, 其中的最优的那个作为对照点, 如果在剩下 63% 的样本中, 出现比对照点好的样本, 就果断选择它。^①

1 实 验

下面先用理论证明

假设有 n 个男生排队找女生, 女生拒绝掉前 k 个, 利用前 k 个男生估计最佳男生的水平, 之后的男生一旦达到这个水平, 就接受之。现须确定其中的 k 与 n 的关系。

对每一个 k , 如果最佳男生在第 i 个位置 (i 在 k 到 n 之间), 那么女生选中他当且仅当前 $i-1$ 个男生中最优秀者在前 k 个中, 可能性为 $k/(i-1)$, 则

$$P(k) = \sum_{i=k+1}^n \frac{k}{(i-1)n} = \frac{k}{n} \sum_{i=k+1}^n \frac{1}{i-1}$$

当 $n \rightarrow \infty$ 时, 令 $x = \frac{k}{n}$, 得

$$p(k) = x \int_x^1 \frac{dt}{t} = -x \ln x$$

对 $y = -x \ln x$, 令 $\frac{dy}{dx} = 0$, 得 $x = \frac{1}{e}$ ^②

可得 x 取 $1/e$ 时即大约 37% 时, 取最优解的概率最大。

下面使用 python 语言用算法来检验 37 法则的正确性与准确性:

以下为代码：

```
import random
import matplotlib.pyplot as plt
from pylab import *
mpl.rcParams['font.sans-serif'] = ['SimHei']
mpl.rcParams['axes.unicode_minus'] = False

def createRandomArray(numOfArray, numRange): #定义函数用来生成大小为 numOfArray,
在 0 到 numRange 范围内的整数的数组
    arr = []
    for i in range(0, numOfArray):
        arr.append(random.randint(0, numRange))
    return arr

Yarray = [] #用于绘制图表，充当图表的 y 轴元素的集合
for i in range(0, 100): #初始化数组，刚开始所有的准确性都为 0
    Yarray.append(0)

for times in range(0, 10000): #测试 10000 次
    arr = createRandomArray(100, 100) #每次重新生成数组
    for x in range(0, 100): #从 0 到 99 移动 x（即逐渐增大抽取样本的数量）
        maxOfX = arr[0] #用于记录 0 到 x 范围内的最大值
        target = arr[0] #用于记录 x 到 99 范围内的第一个大于前 x 个数最大值的数值

        for i in range(0, x): #寻找前 x 个数的最大值
            if arr[i] > maxOfX:
                maxOfX = arr[i]
        for j in range(x, 100): #寻找 x 之后第一个大于最大值的数
            if arr[j] > maxOfX:
                target = arr[j]
                break

        if target == max(arr): #如果这个数确实为整个数组的最大值，则正确，当前下
标的正确个数加一
            Yarray[x] += 1
x = [] #用于充当图表的 x 轴坐标
for i in range(0, 100):
    x.append(i)
    Yarray[i] /= 10000 #测试次数为 10000 次，取其正确的概率

plt.xlabel("只用于测试而不选取的样本数")
plt.ylabel("所选取的目标为所有对象的最大值的概率")
```

```
plt.plot(x, Yarray) #绘制调表  
plt.show()
```

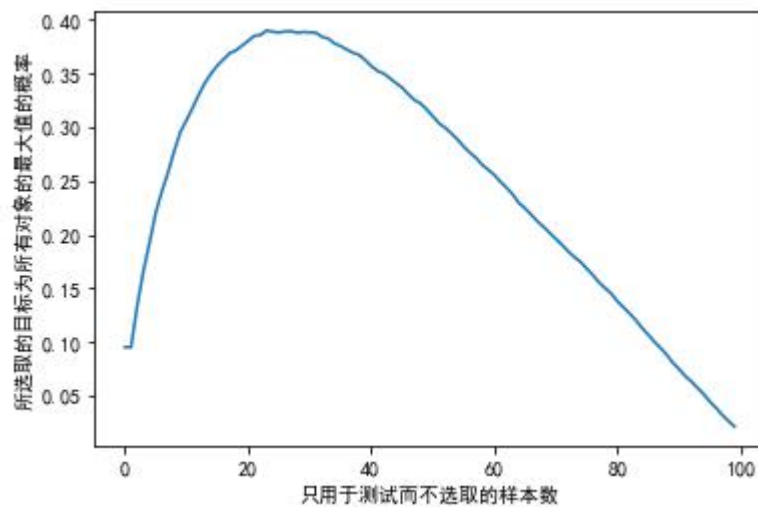


图 1 某次实验通过上述代码所运行出来的结果

2 结 论

多次实验下来，都得到了与图 1 相似的概率曲线，而从图中可以看出，确实在抽取样本数为 37% 时，按照上述算法其能够抽取到整个列表中最大值的概率最大，因此验证了三七定律的正确性。

参考文献：

①小川看世界.37 法则，一种无奈的策略[E B/O L].

[2020-10-14].<https://baijiahao.baidu.com/s?id=1621436361074205070&wfr=spider&for=pc>

②王希. TED 最近的“爱情数学”中提到的 最优停止论 的 37% 是如何计算的？[E B/O L].

[2020-10-17]<https://www.zhihu.com/question/28193509/answer/39925477>