数据科学导论 Presentation Q&A 以及总结

Q1.1: 关于本次抽卡概率的算法是如何计算的?
A1.1: 首先我们在代码中定义了每个角色(item),并且做好定义数值等工作.
在调用数据集的时候开始统计的总数量,并进行索引统计(每个 5 星,4 星甚至 3 星都有对应的索引),接下来将数据绘成频率直方图.

Q1.2: 你们是如何得知机制的?
A1.2: 我们根据游戏官网的相关规则,并且参考在某一抽(73 抽为一个水平线)之后概率呈现线性上升的过程.因为按照最简单的理解,认为前 89 抽每抽抽到五星的概率为 0.6%, 第九十抽抽到五星的概率为100%来算, 抽到五星的综合概率是1.43%, 这和米哈游公布的1.600%的综合概率差距很大, 因此这个模型应该不对.所以我们选用的是前者讲的概率模型.

$$P_{\text{常驻和角色祈愿五星}}[i] = \begin{cases} 0.006 & (i \leq 73) \\ 0.006 + 0.053 \cdot (i - 73) & (74 \leq i \leq 89) \\ 1 & (i = 90) \end{cases}$$

Q1.3: 会不会出现非常欧/非常非(即在抽奖经常能中头奖/抽了好多次触及到了保底机制才中奖)?
A1.3: 客观事实上存在该情况,但是这两类极端情况算为极小概率事件,网上所传的多个十连多金抽卡毕竟也就那么几个人发视频,玩家整体的数量是比投稿人多的.
越抽越非的情况解释如下图: 随着你抽卡次数的增加, 平均抽出的次数会向下图的数据靠拢.

设有相互独立的随机变量序列 $\{X_i\}$, 期望存在, $E(X_i) = \mu_i$, 方差存在且有共同上界, $D(X_i) = \sigma_i^2 < M$, 则 $\forall\, \varepsilon > 0$, 有 $\lim\limits_{n\to\infty} P\left\{\left|\frac{1}{n}\sum\limits_{i=1}^{n} X_i - \frac{1}{n}\sum\limits_{i=1}^{n}\mu_i\right| < \varepsilon\right\} = 1$

切比雪夫大数定律揭示了随着样本容量的增加, 样本平均数将接近于总体平均数的规律. 因此越抽越非的问题, 实质上是随着抽数的不断增加, 样本平均数向总体平均数回归的过程. 而根据切比雪夫大数定律这里的总体平均数可以由实验次数趋于无穷时的样本平均数来计算出.

基于《原神》角色池的抽卡分析:https://zhuanlan.zhihu.com/p/406958011
这一篇知乎的文章能更好解释极端情况.

Q2: 极端情况为什么不影响到最终得到的相关结果(即出现极端情况和最终相关概率出入不大)?
A2: 数据集样本够大,更多的抽卡结果越靠近真实情况.除非出现大量 Dirty Data(什么前面几抽好几个金的),并且我们参考了多个统计数据的官网(paimon.moe , 非小酋等),结果表明我们得到的结果与统计网站的出入非常接近.

Q2.2:是否存在你运气好导致他人运气不好的情况?
A2.2:每个人的抽卡卡池均为独立抽奖而不是共同奖池,不存在你抽的好人家抽不好.

PS: 抽卡概率在后续部份的区间虽然存在高斯分布,但是曲线是稍微往右偏移的.

Q2.3:大量出现欧皇的情况下怎么办
A2.3:这数据基本能算噪声处理了.尽管不处理,那也只能说你运气是真的好呗.世事百态,万事皆可发生^_^

Q3.1: 抽卡是真随机数抽卡还是伪随机数抽卡
A3.1: 计算机层面上大部分情况产生的随机数均为伪随机数(伪随机数均为算法可得,真随机数得从 cpu 等物理硬件中产生),
~~根据程序员第一定律:能省尽省.~~ 游戏厂商为了选择降低成本,大部分情况下都是选用了 PRNG 的形式去处理抽卡算法.游戏往往需要在短时间内处理极大量的抽卡请求，能满足这样需求的 TRNG(真随机数发生器 True Random Number Generator)成本太高，调试和维护也比较麻烦，而现有的高端商用 PRNG 加上高质量的种子生成模式足以应付需求。
补充:几乎所有抽卡游戏都不会使用 PRD(war3 中的“概率修正”性质的)伪随机分布 Pseudo Random Distribution
PRD 在减少“一直抽不出五星卡”的情况的同时更大幅度的减少了“连续抽到五星卡”的情况(此处“更大幅度”的前提是“抽到五星卡”的初始概率比较低，联系一下上面关于暴击的概率计算就明白了)，而这种“欧洲人”的情况对于游戏体验的提升是巨大的，很多玩家持续玩下去的动力都是获得了这种体验后的喜悦或者对于获得这种体验的期待，更重要的是在“出五星卡概率恒定”和“有足够多的玩家数量”两个条件下，极少数“连续抽到五星卡”的情况是不影响运营利益的(总期望不变且更加接近理论期望)。

Q3.2: 如果没抽到的情况下,你愿意会充钱去抽奖吗
A3.2: 这是基于玩家的个人选择,抽奖与否都是看我们玩家自己的选择.

Q4: 心理学的问题
如果有人抽了 69 抽之后,再抽几抽就出货了,而且这类物品对于游戏属于是关键性物品(抽到与否会大幅度影响你的游戏体验). 但是今天是卡池的最后一天,你目前的选择是只有充钱,而且选择有单抽和十连,选哪一种好?
A4: 十连抽卡在 ppt 中已经介绍过了.就是 10 次的独立抽卡(单抽),把动画缝合在一起罢了.
至于抽不抽还是看玩家选择,玩家的行为是自由的.
有的人会考虑放弃也有人考虑放手一搏去抽,但是是否抽其实影响不算很大.

Q5.1: 样本中如果存在脏数据怎么办.
A5.1: 巧了,我们代码里头还真的有简单的预处理代码(在作业提交过程中我会将其他文件贴上),抽卡的脏数据造谣成本不高,很异常的数据都可以甚至使用人工剔除.至于是否能进一步对数据进行降噪,我们能力有限 QAQ,总结末尾会提供参考文献:

```python
for i in tqdm.tqdm(file_list):  # 五星概率测试

    folder_paths = [total_files, i]
    folder_path = osp.join(*folder_paths)
    for j in range(3):
        file_name = file_c[j]
        processing_file1 = osp.join(total_files, str(i).rjust(4, '0'), file_name)

        if os.path.exists(processing_file1):
            try:
                data = pd.read_csv(processing_file1)
            except:  # 文件为空
                continue
        else:
            continue
        if max(data.index) < least_gacha_time:  # 略去量少数据
            continue
        counter_5s = 0

        for index, row in data.iterrows():

            counter_5s += 1
            this_star = data.iloc[index].values[3]
            if this_star==5:

                per_index=index-now_index



                if per_index>0:
                    now_index=index
                    if per_index in per_index_counts:
                        per_index_counts[per_index] += 1
                    else:
                        per_index_counts[per_index] = 1
```

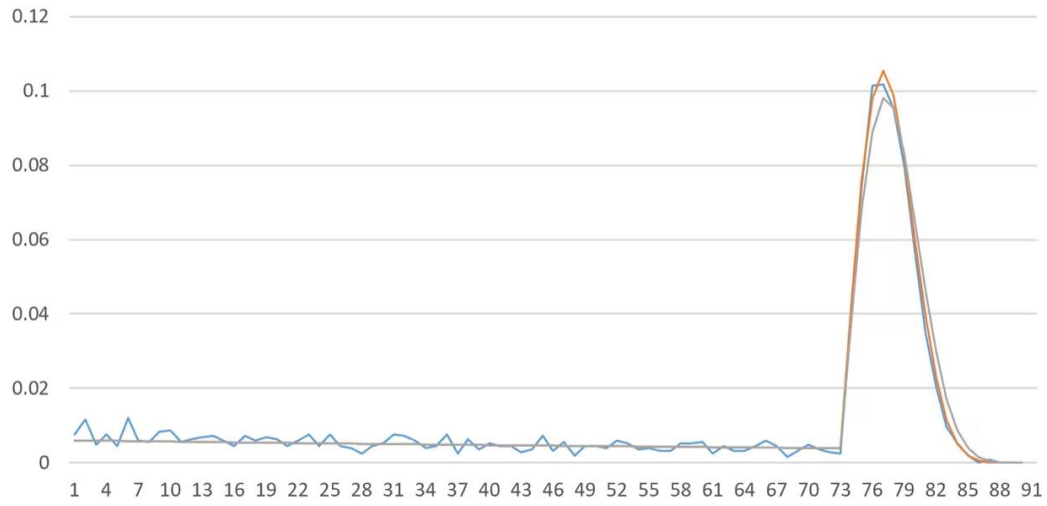
```python
for index, row in data.iterrows():
    all_raw_pull += 1
    counter_4 += 1
    counter_5 += 1
    this_star = data.iloc[index].values[3]
    if this_star == 4:  # 这次是四星
        if been_5 and pure_4_star_model:  # 特殊分析时，中间有五星，就略过本次
            counter_4 = 0
            been_5 = 0
            continue

        # 筛选UP池时发现不是这个池子
        if pool_select:  # 开启了UP池筛选
            check_select_mark = 0
            for pool_num in pool_list:
                if wish_filter(0, data.iloc[index].values[0], pool_num, 'NULL'):
                    check_select_mark = 1
            if check_select_mark == 0:
                counter_4 = 0
                continue
        if counter_4 >= 11:  # 小概率事件
            print(i + ' ' + file_name + ' ' + str(index) + ' 四星间隔合理')
        if counter_4 >= 12:  # 极低概率事件
            print(i)
            print(counter_4)
            print('四星间隔超出阈值，需要检查')
        if data.iloc[index].values[2] == '武器':
            star_4_distribution[counter_4][j][2] += 1
        if data.iloc[index].values[2] == '角色':
            if j == 1:  # 常驻池
                star_4_distribution[counter_4][j][1] += 1
            elif wish_filter(1, data.iloc[index].values[0], 0, data.iloc[index].values[1]):
                # 是UP角色
                star_4_distribution[counter_4][j][0] += 1
            else:  # 非UP四星角色
                star_4_distribution[counter_4][j][1] += 1

        gacha_time_4 += counter_4  # 记录本次所用抽数
        counter_4 = 0
        been_5 = 0
    if this_star == 5:
        max_5_star_pull = max(max_5_star_pull, counter_5)
```

Q5.2: 免费获得的游戏代币和付费代币投入进去的数据是否分开?(原神中的原石可通过游戏内活动获得也可以通过充值转化) 付费所抽的物品和免费所抽的物品是否分开统计?
A5.2: 抽奖反正用的代币都是一样的(对于原神来讲),所以充不充钱你反正抽出来的东西的概率大家人人平等.而且不存在你充值与否会影响你的概率,因为抽奖算法就固定在那不动了,除非游戏厂商真的敢违约干这种事.
是否对比过免费抽奖和付费抽奖这个后面,我们所收集的数据 Only 物品,并且只能得到概率,因为人家抽奖的记录不会区分你是否充值,只会在我们抓包的过程中给我们 feedback 抽卡结果.
而且玩抽卡游戏,真的极少数玩家是完全不抽钱的...

Q5.3: 是不是真的都是 90 抽才出:
A5.3: 在[73,90]抽时候,我们通过计算得到的数据,77 抽还没出金的概率是最大的,但是在后面你咋抽基本也都快出货了,90 抽还不出货算是极小概率事件.而且游戏存在保底机制,前面 89 抽再不出 90 抽也是 100%出货.



Q6: 90 次抽之后能立刻获得第二个五星嘛?
A6: 抽到五星角色之后你的概率直接回归到[0,73]的概率:均为 0.6%,抽不抽得到,看你运气

PS:一个玩家连续进行十连，在每次十连中不同位置出现频率是否趋于均等，那结果是会。可以把第 n 次出金位置设为 X_n，X 到 X_{n+1} 则有一个概率转移矩阵。显然 uniform distribution 是这个概率转移矩阵对应的马尔可夫链的 stationary distribution，又因这个 Markov Chain 是 irreducible 的,所以这个 stationary distribution unique。所以出现频率会趋于均等

Q7:十抽是真的十次单抽吗?
A7:每一抽单独计算,并且抽一次转移一次...你怎么抽综合概率就摆在那里了.

总结:上面的问题有的确实是我们不太好考虑到的问题,但是抽卡数据集也是有限的信息.所以我们是尽可能的回答.

反思:对于本次の汇报我们只是简单的做了一份抽取人物卡池的(因为还有一种卡池叫武器池,机制更为复杂).但是光对人物的简单抽奖概率模型分析,,我们也能大概了解到更具体更细致的一个样式.可能 pre 和本篇总结仍均有一定的漏洞,但是也尽力解释了相关内容.希望在下一次汇报的时候能更清楚的介绍.

参考文献:

[概率贴外传系列 关于“时间玄学”，单独开个帖子讲一下吧……](https://ngabbs.com/read.php?tid=14237611)

[原神抽卡保底机制是什么？是否涉及概率诈骗？](https://www.zhihu.com/question/423181809)

[一文说清楚原神抽卡机制，概率，期望等问题](https://zhuanlan.zhihu.com/p/522246996)

[原神抽卡机制研究（一）——五星的保底机制](https://www.bilibili.com/read/cv8772558/)

[原神抽卡全机制总结](https://www.bilibili.com/read/cv10468091/)

[基于千万级抽卡数据的补充统计](https://www.bilibili.com/read/cv14841352/)

[解析游戏中的抽卡机制](https://zhuanlan.zhihu.com/p/595041406)

[原神概率公示]https://ys.mihoyo.com/main/news/public

[闲聊杂谈]半夜想一个数学问题睡不着

https://nga.178.com/read.php?tid=35071531&forder_by=postdatedesc&page=2

游戏中的概率和随机:https://zhuanlan.zhihu.com/p/669348838

抽卡游戏机制研究——以《原神》、《明日方舟》为例

https://www.bilibili.com/read/cv19423852/

[马王堆 1 号坑]wjndante 的“概率与随机”三部曲之其一——如何用计算机生成“真随机”(仍在施工中……)

https://ngabbs.com/read.php?tid=8477978