陶建容 浙江大学 计算机科学与技术 硕士

1. 逻辑题

abc每人带12天粮食，第三天的时候a把剩下9天的粮食给b3天c3天，自己留3天回岛。bc各有12天粮食。再过三天后，b把自己的9天粮食留给a3天，自己留6天回岛。c有12天粮食，可以走出去。

2. 概率题

从题目中已知：

p(色|男) = 0.05

p(色|女) = 0.0025

p(男) = 0.5

p(女) = 0.5

我们可以求出：

p(色) = p(色|男)p(男) + p(色|女)p(女) = 0.02625

p(男|色) = p(色|男)p(男) / p(色) = 0.95

3. 简答题

k均值：

优点：实现简单，当样本比较密集的时候，类与类之间的区分效果较好。

缺点：k值的选定是个问题，对开始的k个任意中心点比较敏感，结果不稳定，对异常点和孤立点比较敏感，不能像密度聚类一样发现任意形状的簇。

k中心：

优点：它是基于k均值方法进行改进的，主要是为了缓解k均值中噪声敏感的问题，每次迭代后的质点都是从样本点中选取，提高聚类效果。

缺点：其他的问题与k均值相似。

它们相比层次聚类：层次聚类对于不同k的簇可以从层次树中获取，不需要重新跑，但是每个样本不具备再分配能力

4. SQL

（1）

select (tend.the\_time\_sec - tbegin.the\_time\_sec) as cost from

(select sum(datediff(the\_time, '19700101')\*60\*60\*24 + hour(the\_time)\*60\*60 + minute(the\_time)\*60 + second(the\_time)) as the\_time\_sec from phone\_call\_log where type = 1) as tbegin,

(select sum(datediff(the\_time, '19700101')\*60\*60\*24 + hour(the\_time)\*60\*60 + minute(the\_time)\*60 + second(the\_time)) as the\_time\_sec from phone\_call\_log where type = 2) as tend;

（2）

set @t1\_row=0;

set @t2\_row=0;

select sum(ceil(timestampdiff(second,t2.the\_time,t1.the\_time)/60))\*0.15-sum(timestampdiff(second,t2.the\_time,t1.the\_time))\*0.0025 from

(select @t1\_row:=@t1\_row+1 as row\_id, the\_time from phone\_call\_log where phone\_call\_log.type=2 order by call\_id,the\_time) as t1,

(select @t2\_row:=@t2\_row+1 as row\_id, the\_time from phone\_call\_log where phone\_call\_log.type=1 order by call\_id,the\_time) as t2

where t1.row\_id=t2.row\_id;

5. 分析题

从题目中已知：

（1）iOS游戏收入／整个手游市场收入 ＝ 80%

（2）开发者分成 : 苹果分成 ＝ 7:3

（3）中国区畅销榜前30的游戏应用分成收入（我理解为苹果分成收入？）

我们需要得到整个手游市场收入，求出iOS游戏总收入即可：

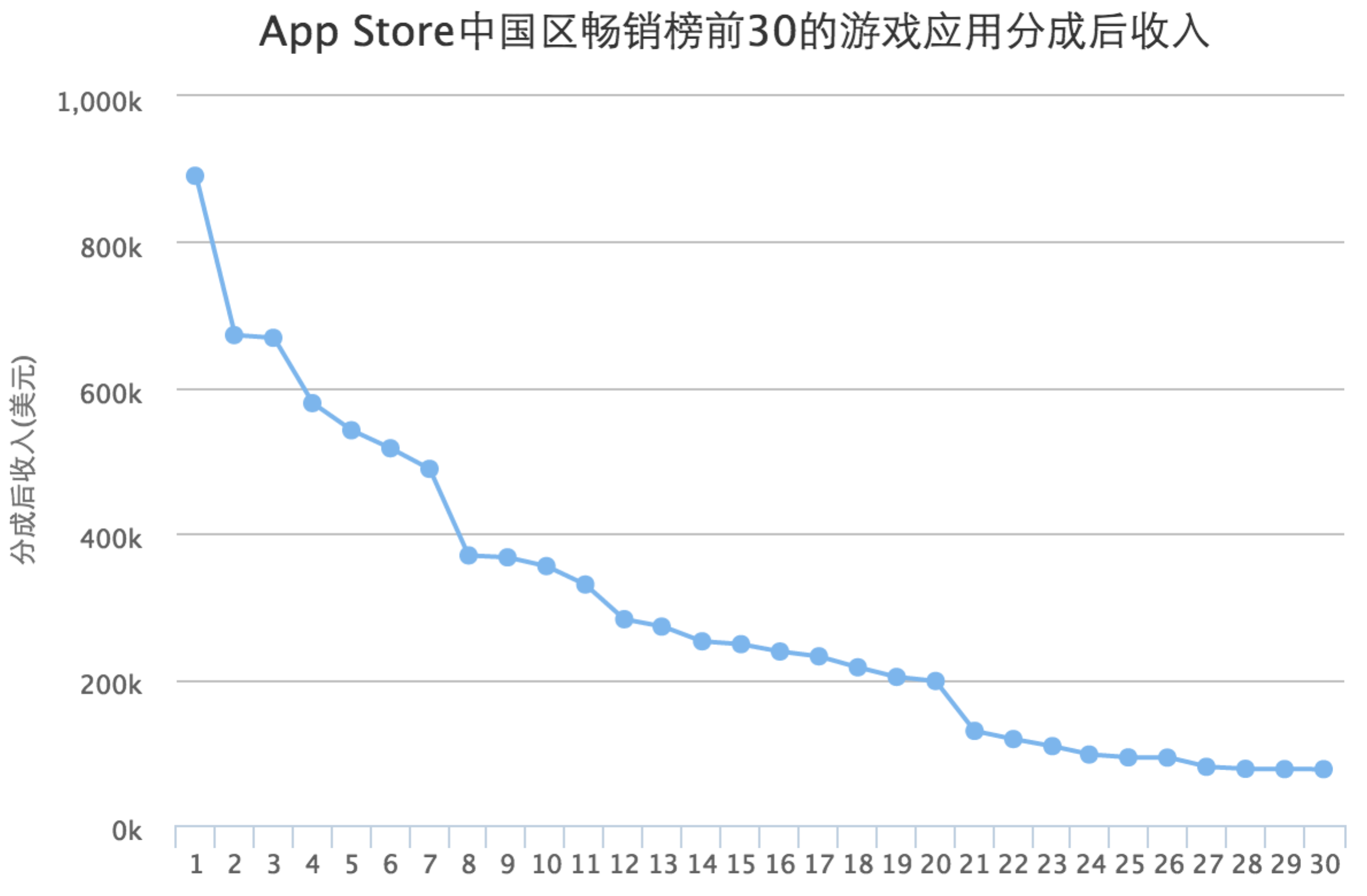
整个手游市场收入 ＝ iOS游戏总收入／80%

iOS游戏总收入 ＝ 苹果分成总收入 ／ （3/10）

所以只要求出苹果分成总收入，即可由下述公式得到整个手游市场收入：

整个手游市场收入 ＝苹果分成总收入 ／ 30% ／ 80%

我们现在只有top30的游戏应用对应的苹果分成收入，如果只求和这30个游戏应用的苹果分成收入，显然并不等于苹果分成总收入，我们需要预测后面排名的游戏应用的收入情况。我们可以先看一些统计的图，大概了解以下数据的分布：



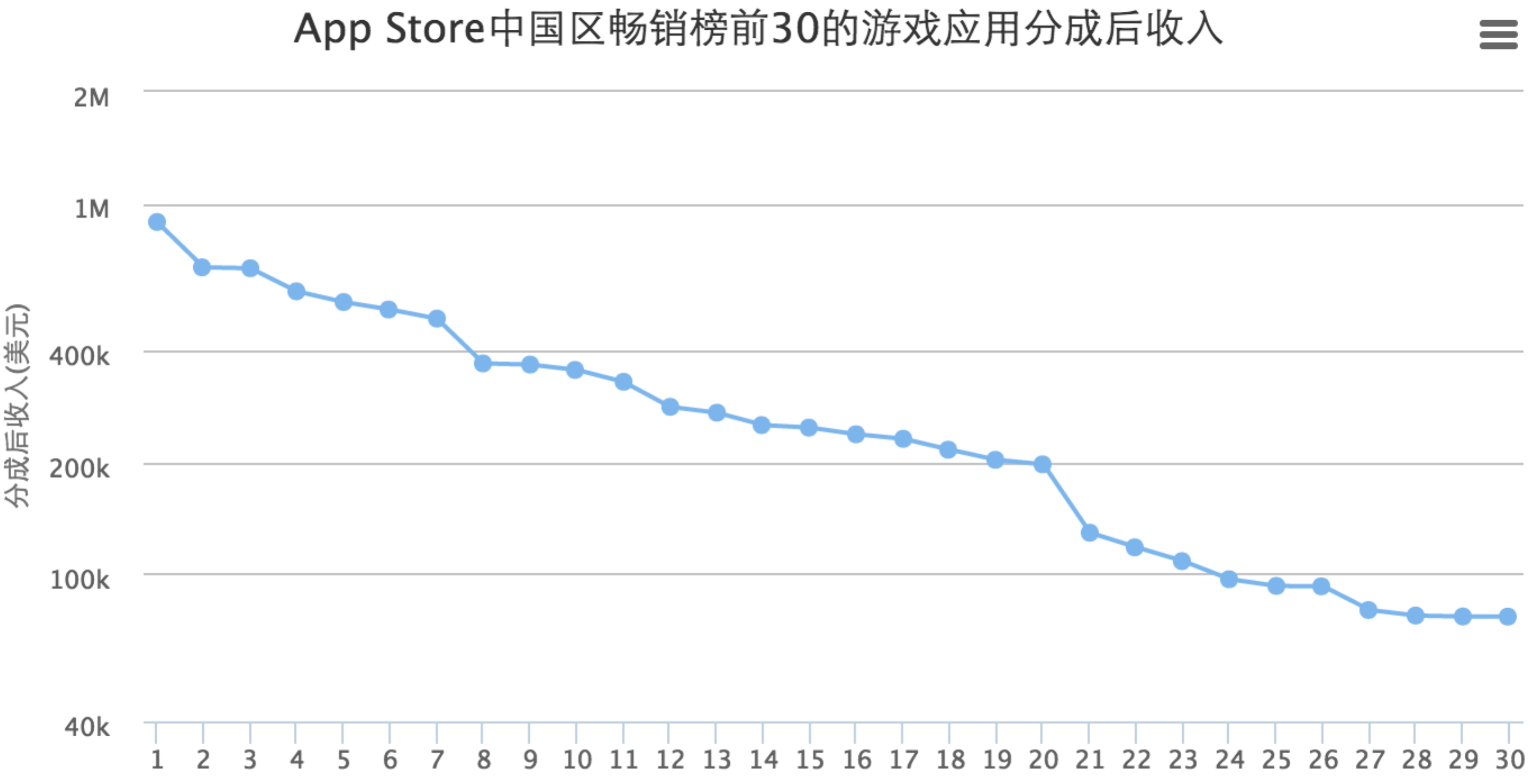
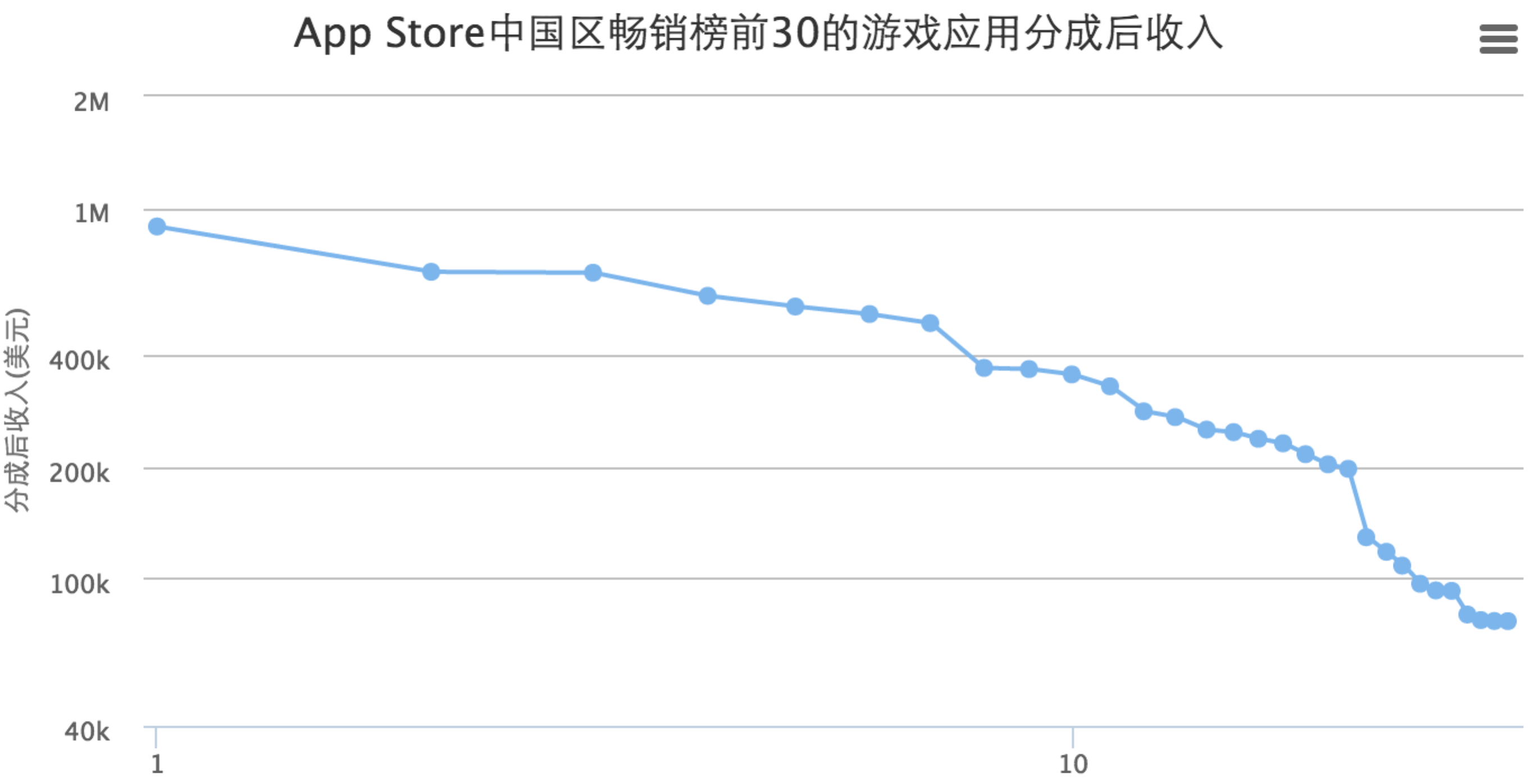
从这张数据分布图，我们可以得到两个结论：

1. 排名30之后的游戏收入比较均匀，下降比较平滑，存在一个long tail
2. 曲线逼近与y=a(x-b)^2的左半段

根据上述结论，我们可以拟合出一条曲线，求出a和b，由此可以求出b对应于收入为0的游戏应用，b排名之后的游戏应用可以忽略，我们只需要计算出b排名之前的游戏应用的总收入即可。

备注：下面两张图分别对纵坐标取log，对横纵坐标均取log。

对于前一张图，似乎也可以尝试拟合该曲线，求出拟合曲线与x轴的交点，和上述b含义相同。

6. 问题：

1. n是long类型，但循环变量i是int类型。二者在不同机器上表示的位数可能不同，如果某个机器表示long的位数多于表示int的位数，则当n大于int的最大值时，i出现上溢出，程序进入死循环，且结果temp的值无意义。
2. 即使循环变量改为long，当n是long的最大值时，程序依然无法退出循环，原因同上。
3. 如果编译器没有优化，计算机的乘法运算的复杂度很高，对程序效率会造成影响。
4. 根据题意的描述（1-2+3-4+5-6+7-8..........+n），n前面的符号是正，且n是一个很大的数，则隐藏条件是n可能是一个奇数，代码中没有指出这一点。当然此处也有过度解读题意的可能。

改进for循环：

long pairs = n >> 1; // 一个pair表示(-2+3)，(-4+5)这样的组合，pairs记录有多少个组合

temp = 1;

for (long i = 0; i < pairs; ++i) {

temp += 1; // 每个pair组合之和都为1

}

if ((n & 1) != 0) temp -= n; // 若n为偶数，则额外减去多余的一项

7. 使用a^x = b来计算x，用二分法逼近解。

long double myLog(int a, int b) {

assert(a > 1 && b > 0);

long double lb = b;

long double la = a;

long double left, right, mid, eps = 1e-9, currB;

if (lb > 1) { // 确定答案的上下界

left = 0;

right = 1;

while (pow(la, right) < b) {

left = right;

right \*= 2;

}

}

else {

left = -1;

right = 0;

while (pow(la, left) > b) {

right = left;

left \*= 2;

}

}

mid = left + (right - left) / 2;

currB = pow(la, mid);

while (fabs(currB - lb) > eps) {

if (currB > lb) right = mid;

else left = mid;

mid = left + (right - left) / 2;

currB = pow(la, mid);

}

return mid;

}

8. python实现

def countKangaroo(n):

dp = [0] \* 10

dp[5] = 1

for year in range(2, n+1):

newKangaroo = 0

for i in range(4, 10):

newKangaroo += dp[i]

for i in range(9, 1, -1):

dp[i] = dp[i-1]

dp[1] = newKangaroo

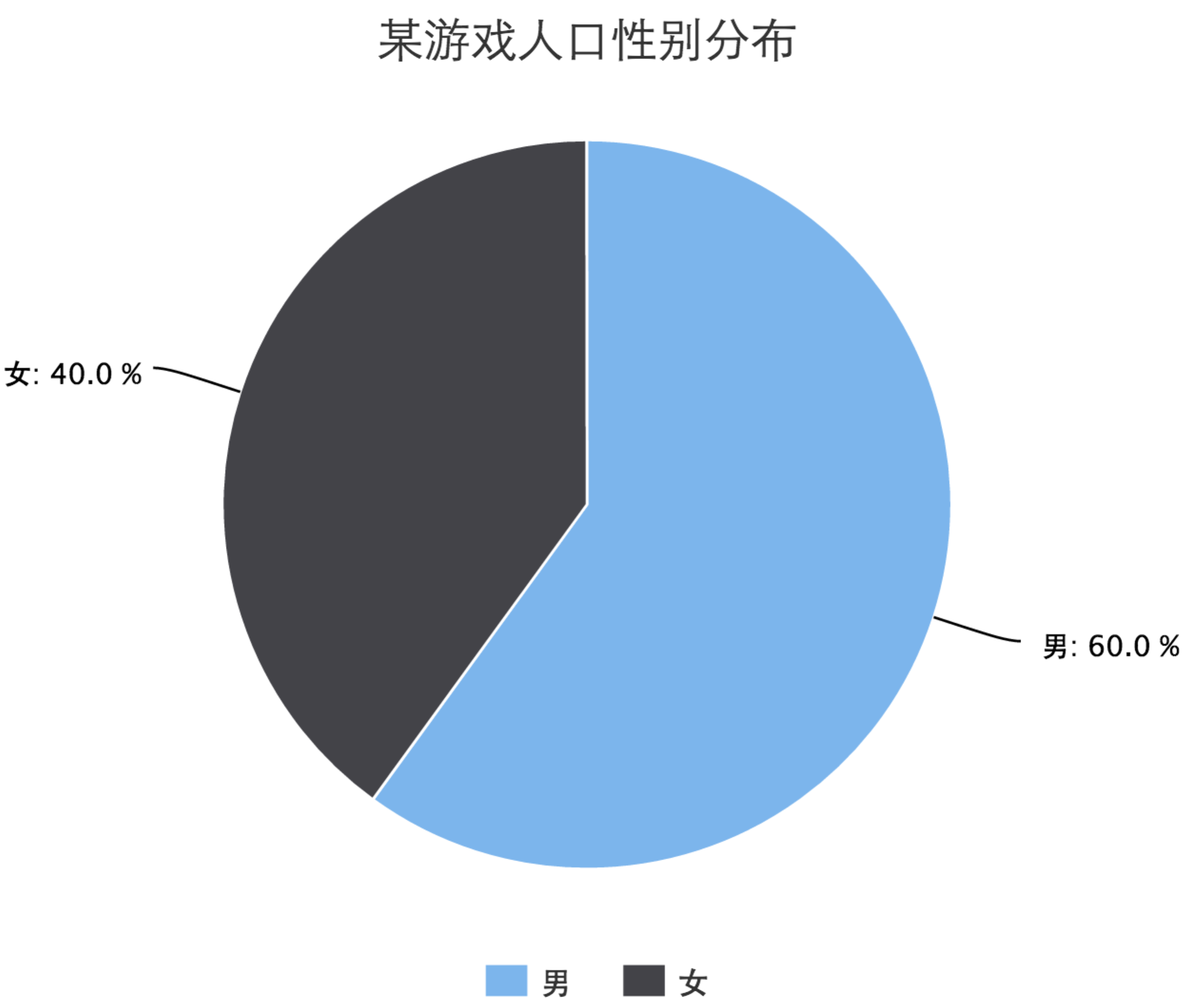
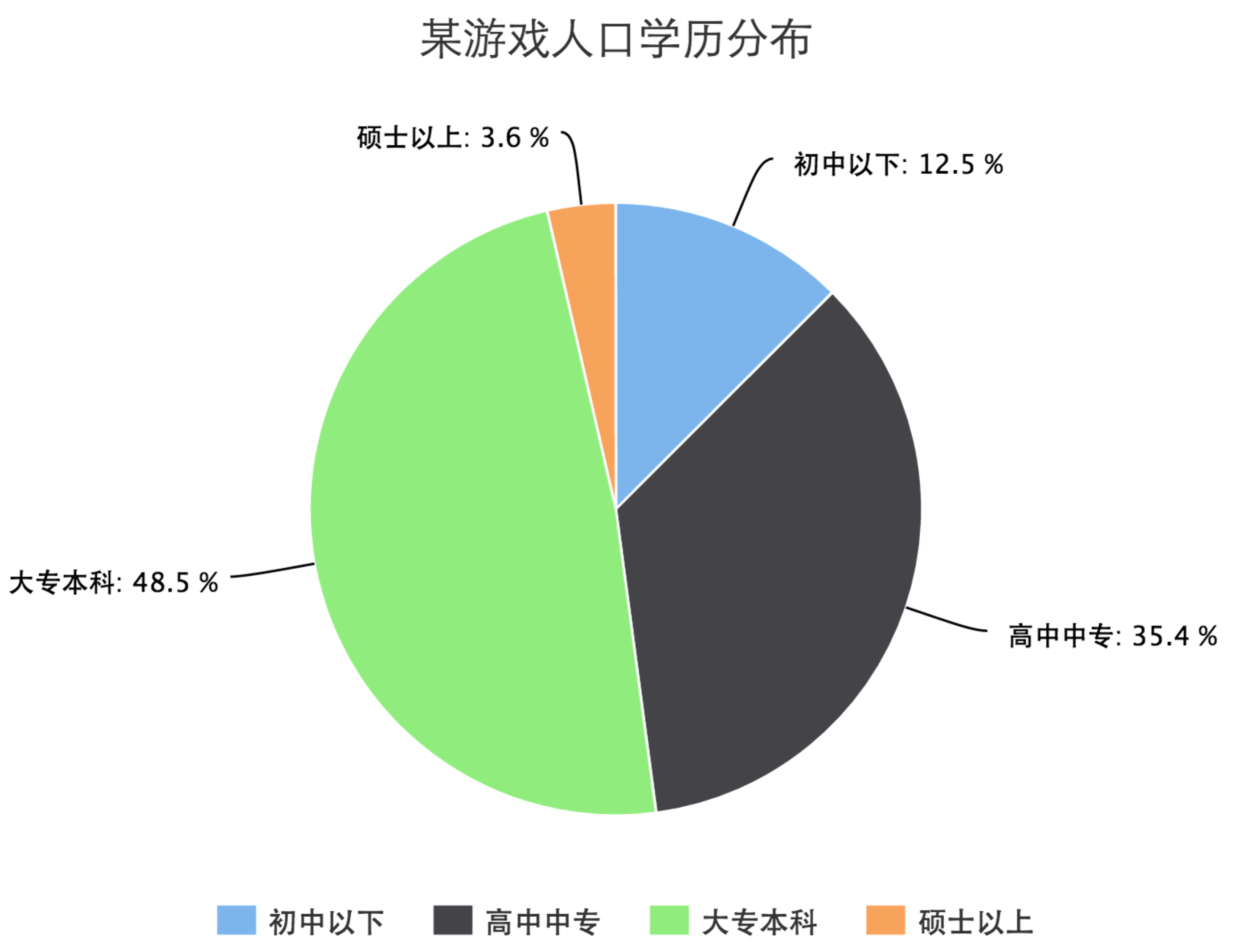
return sum(dp)

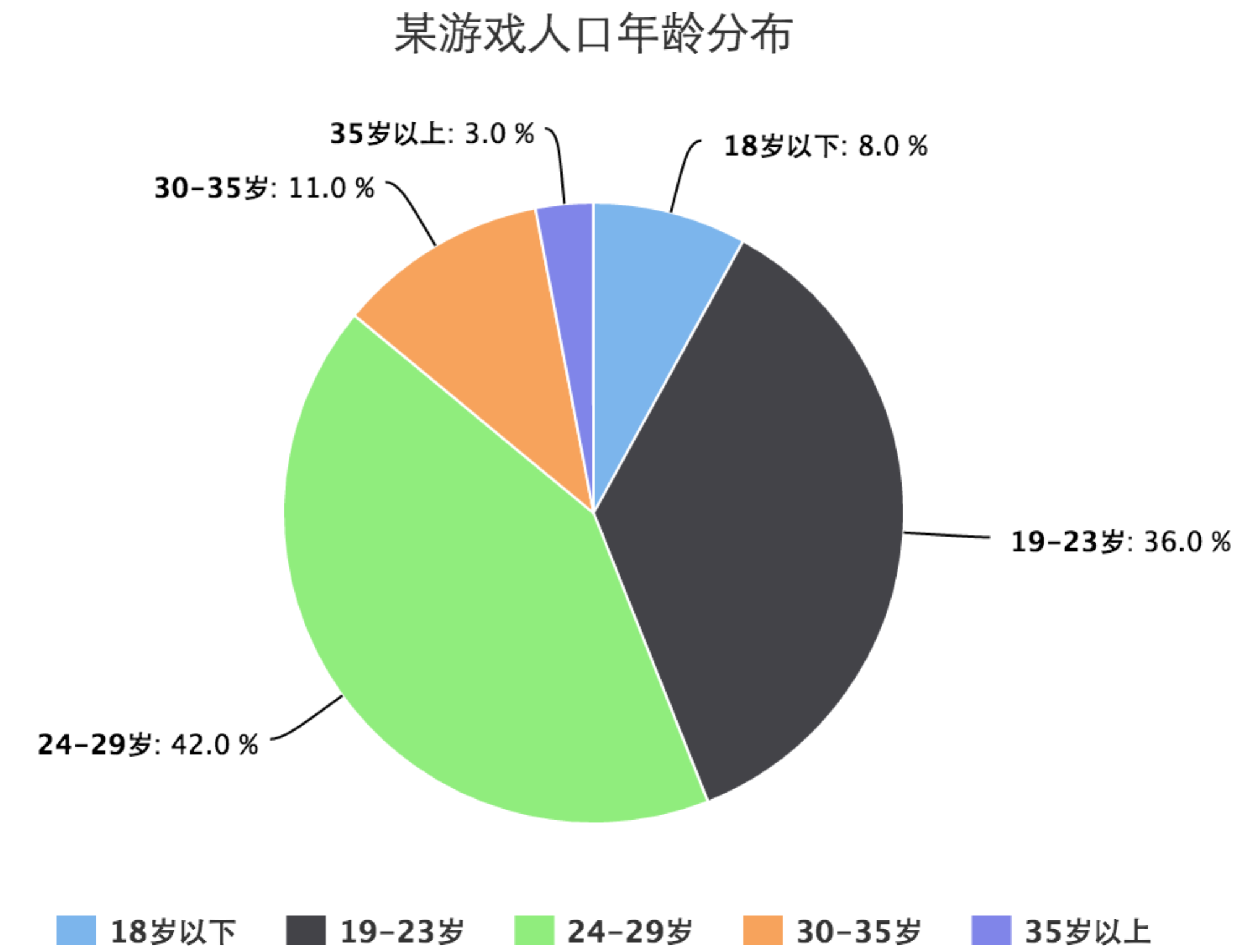
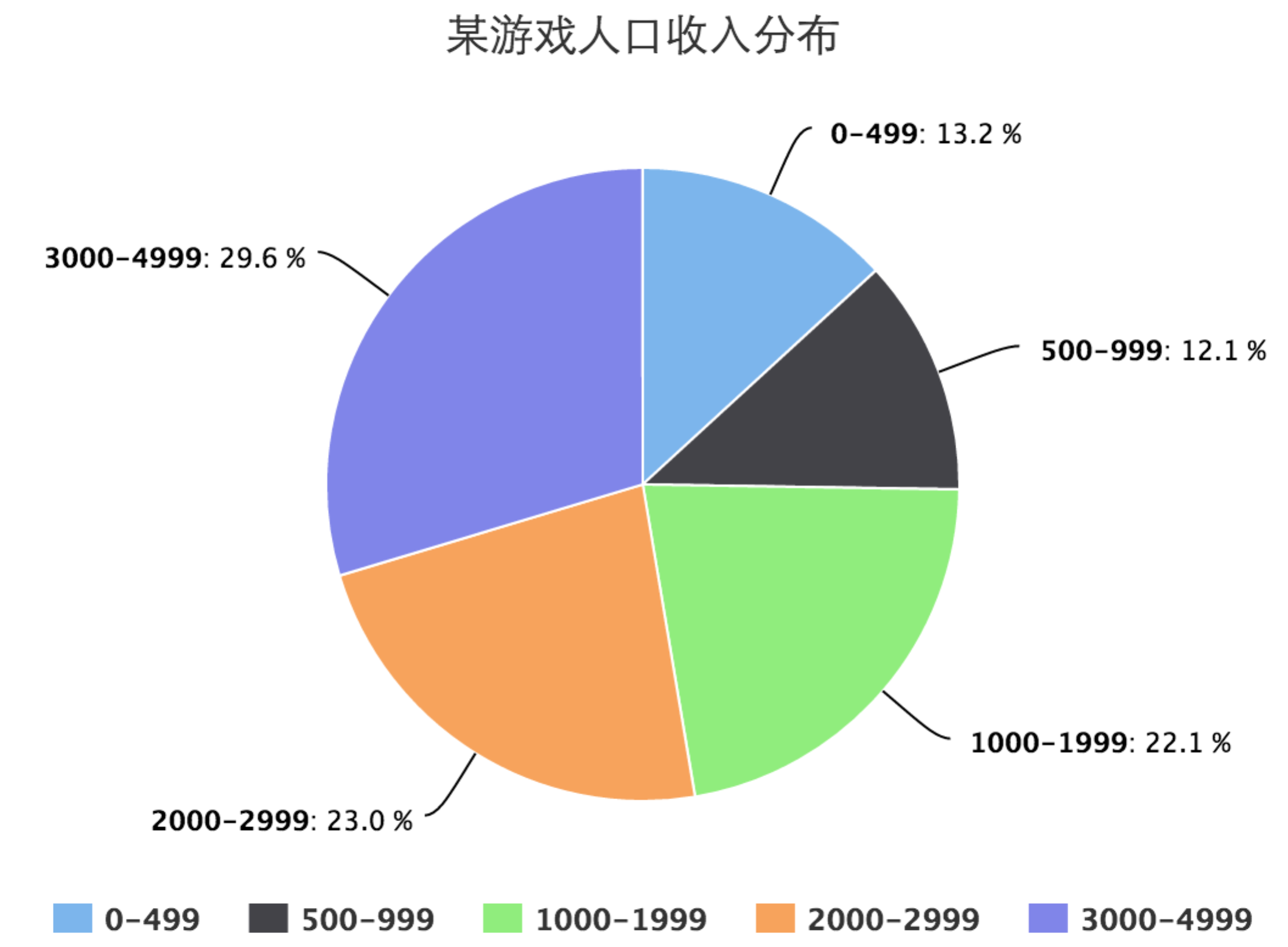
countKangaroo(2000)：

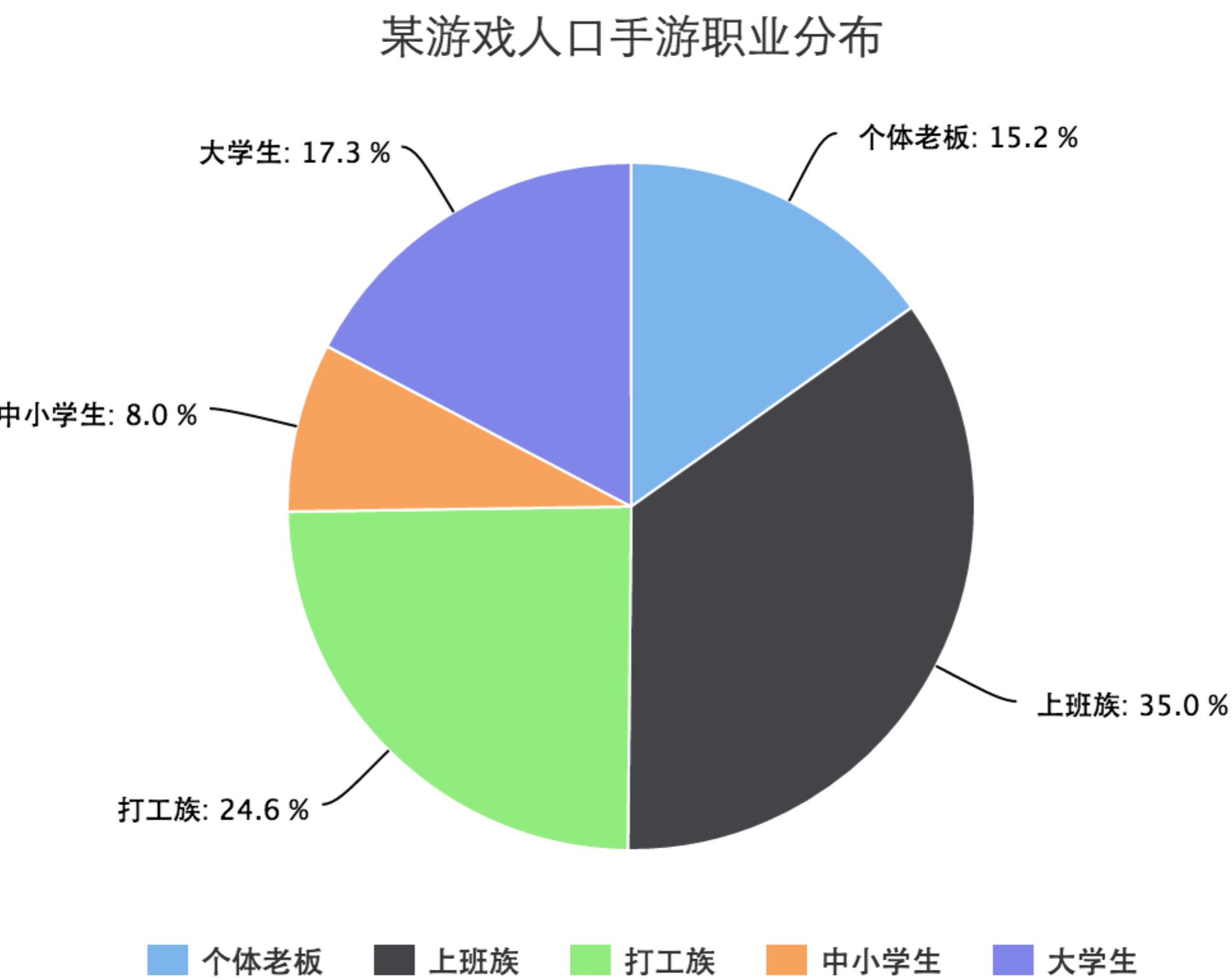
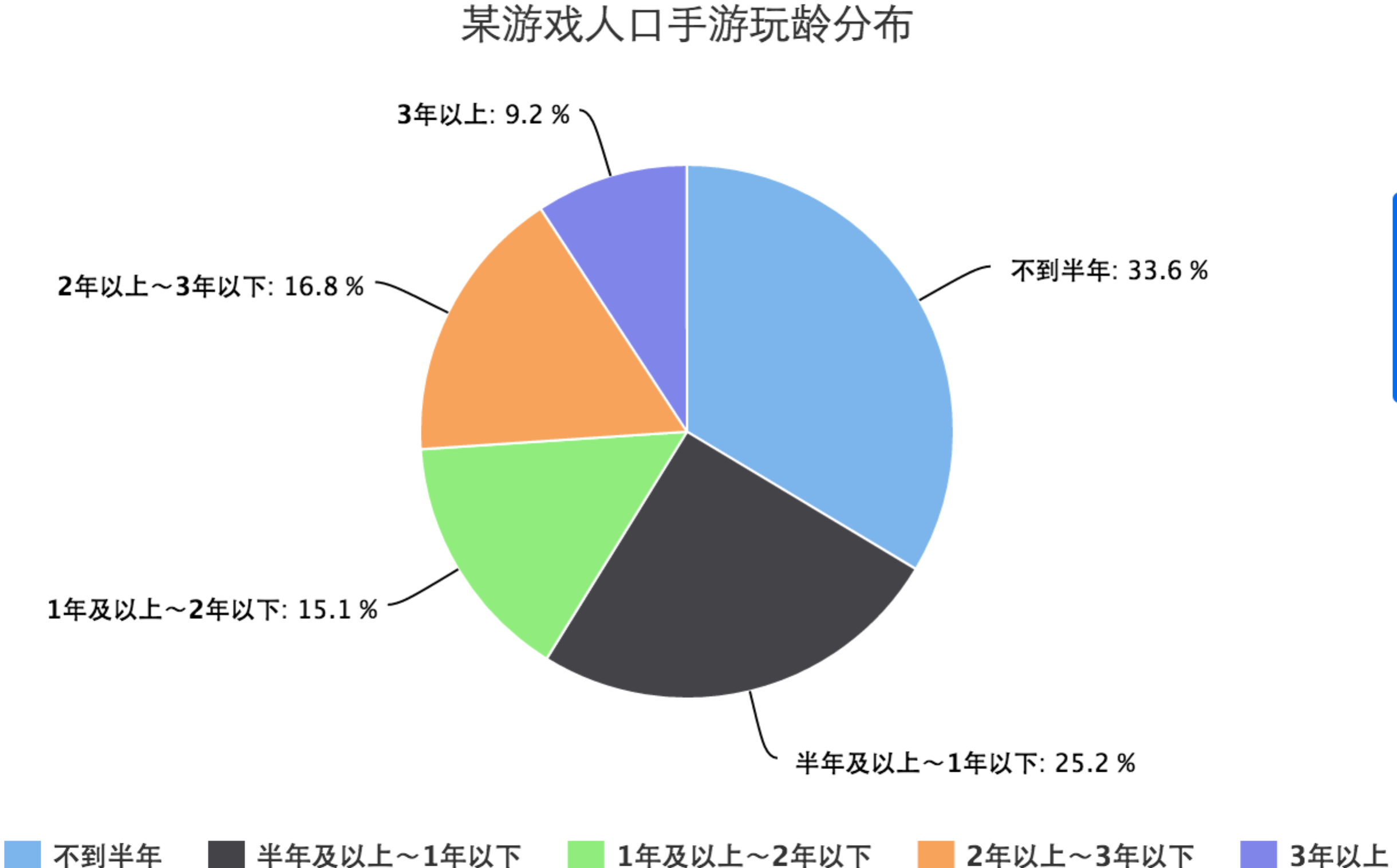
8863424305202371617492840015802000933235622289963477491205586418635305821596023167947160892598019002258056807027698421798702859238393278992295470506881319668533448109014009582297362986948195626431724957596512021002022925641637691823859212351104930723969634L

9. 作图题

这类用户Ground Truth相关的数据，利用饼图来展示比较的人性化。我平时在论文中绘制此类图时，利用R的ggplot2和python的matplotlib绘制为主，这里出于效率和简洁的考虑，利用hightcharts绘制。

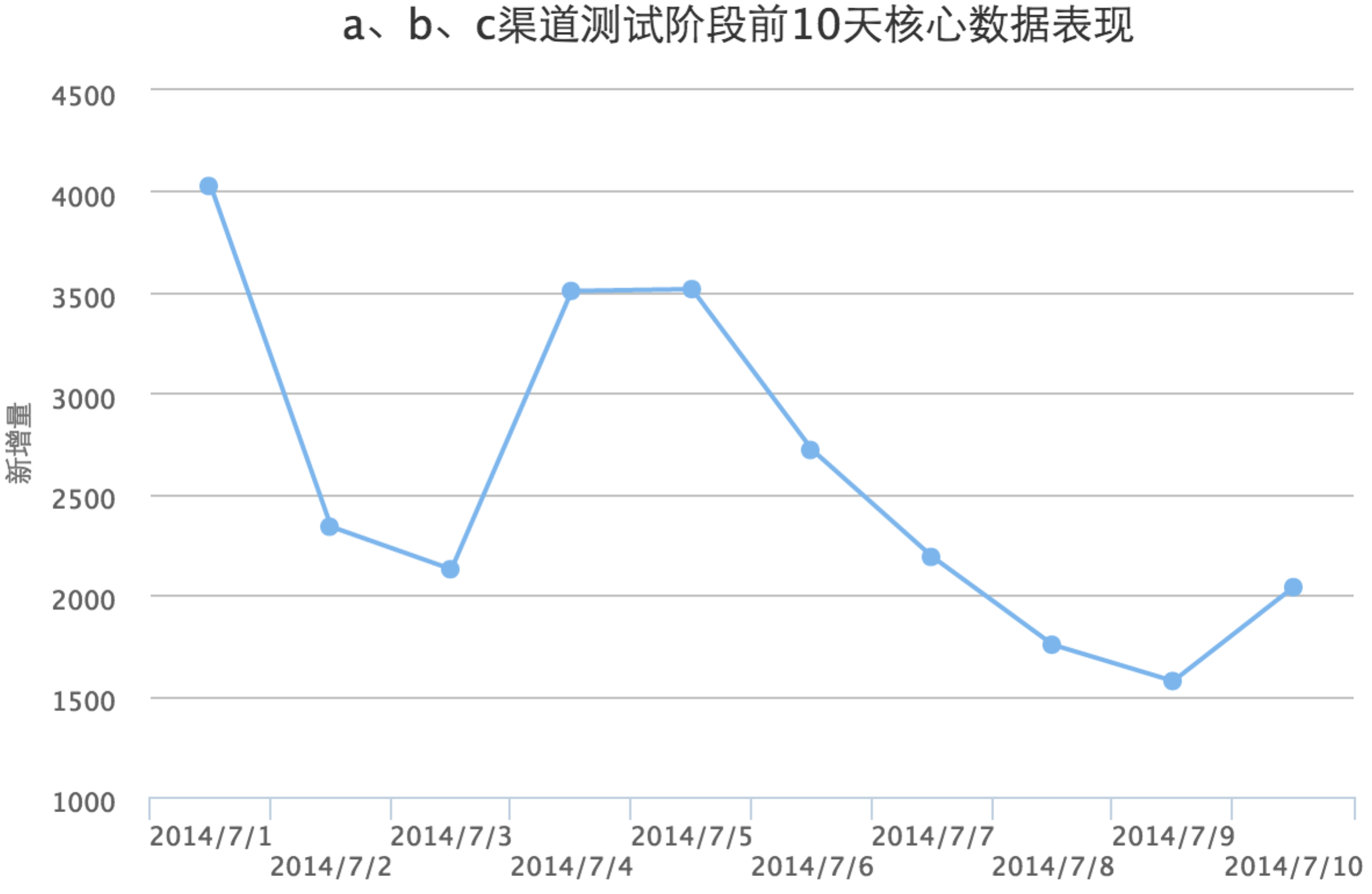
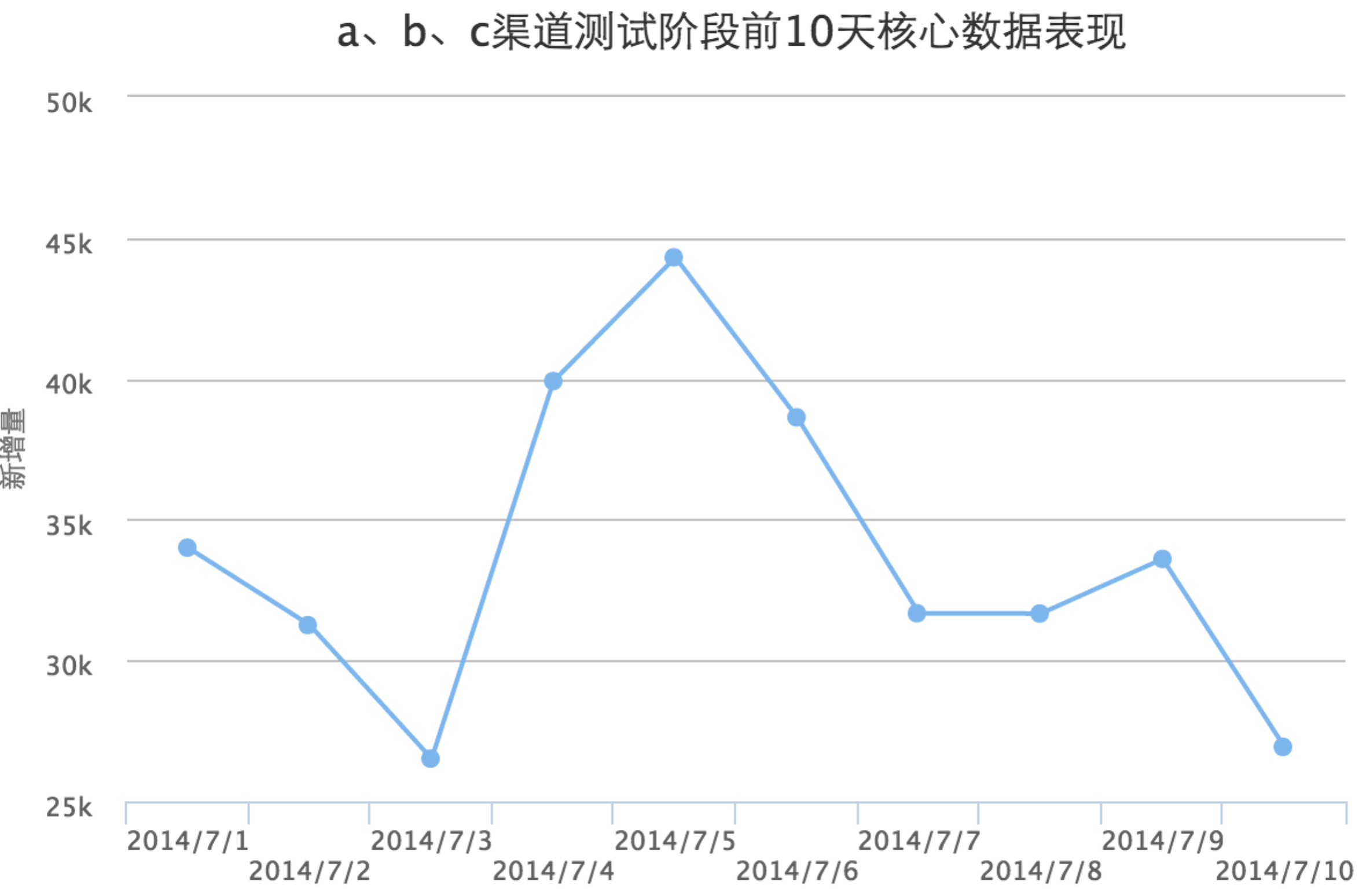
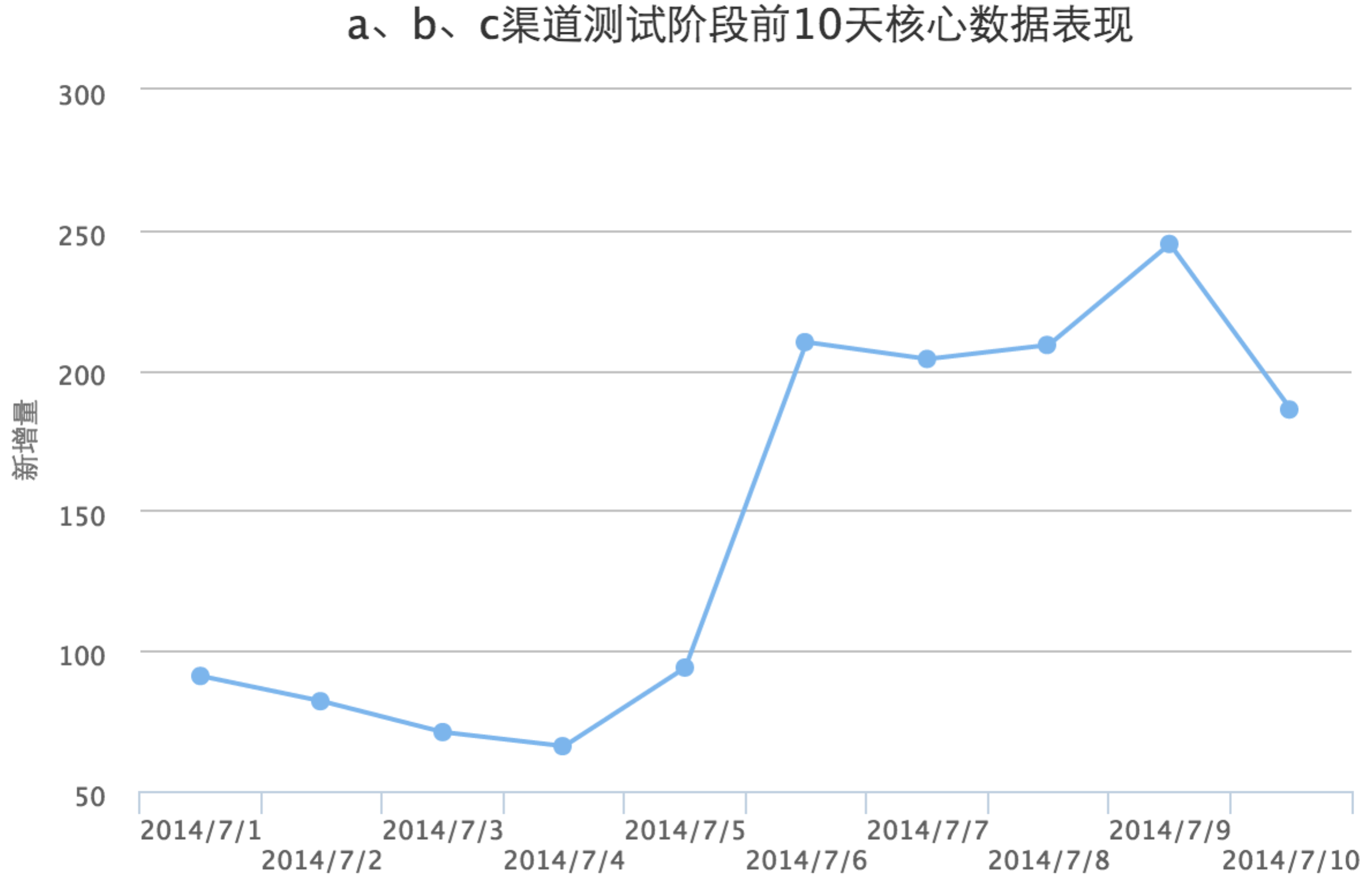
 

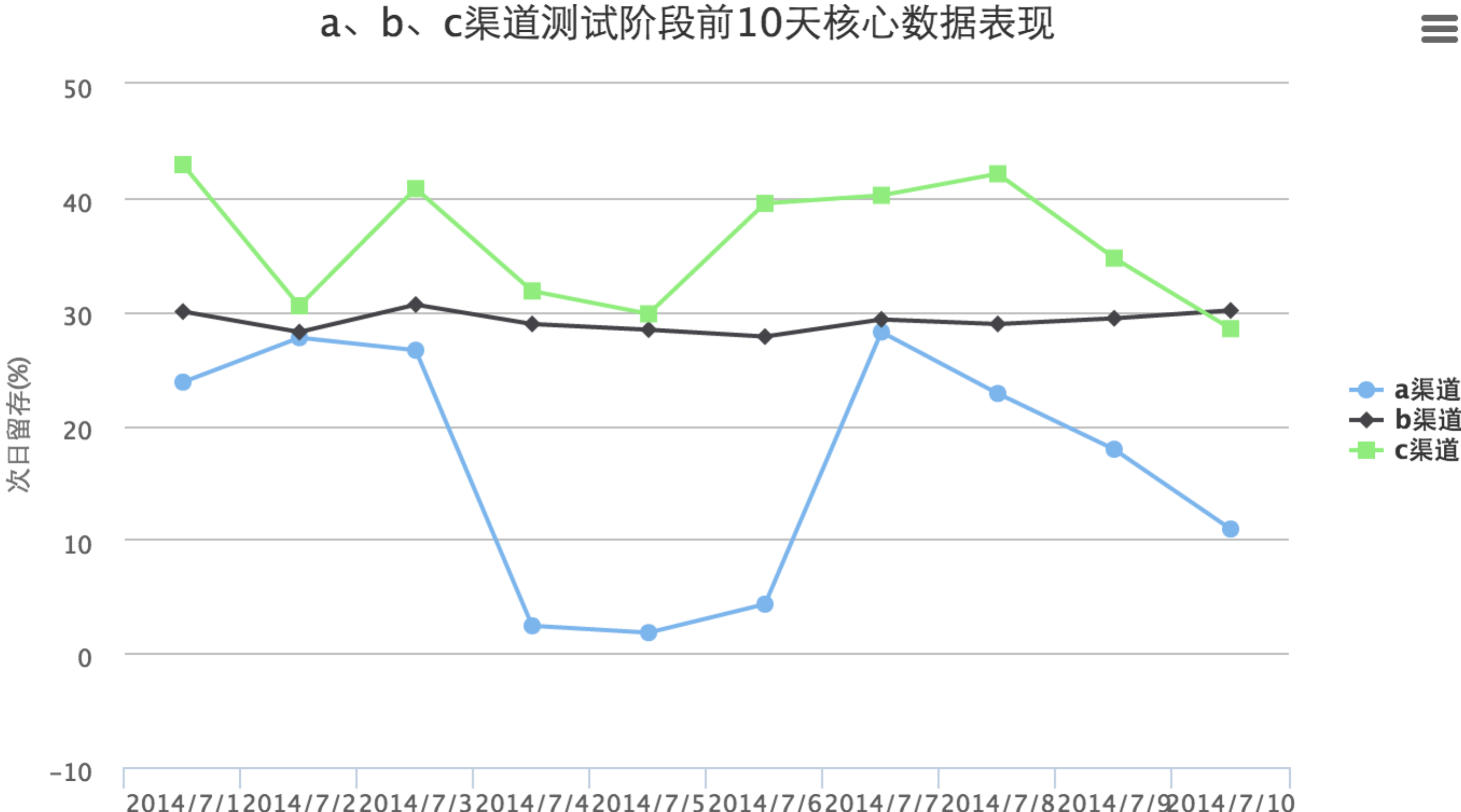
10. 依据题目内容详细作答

首先单独分析四张表格：

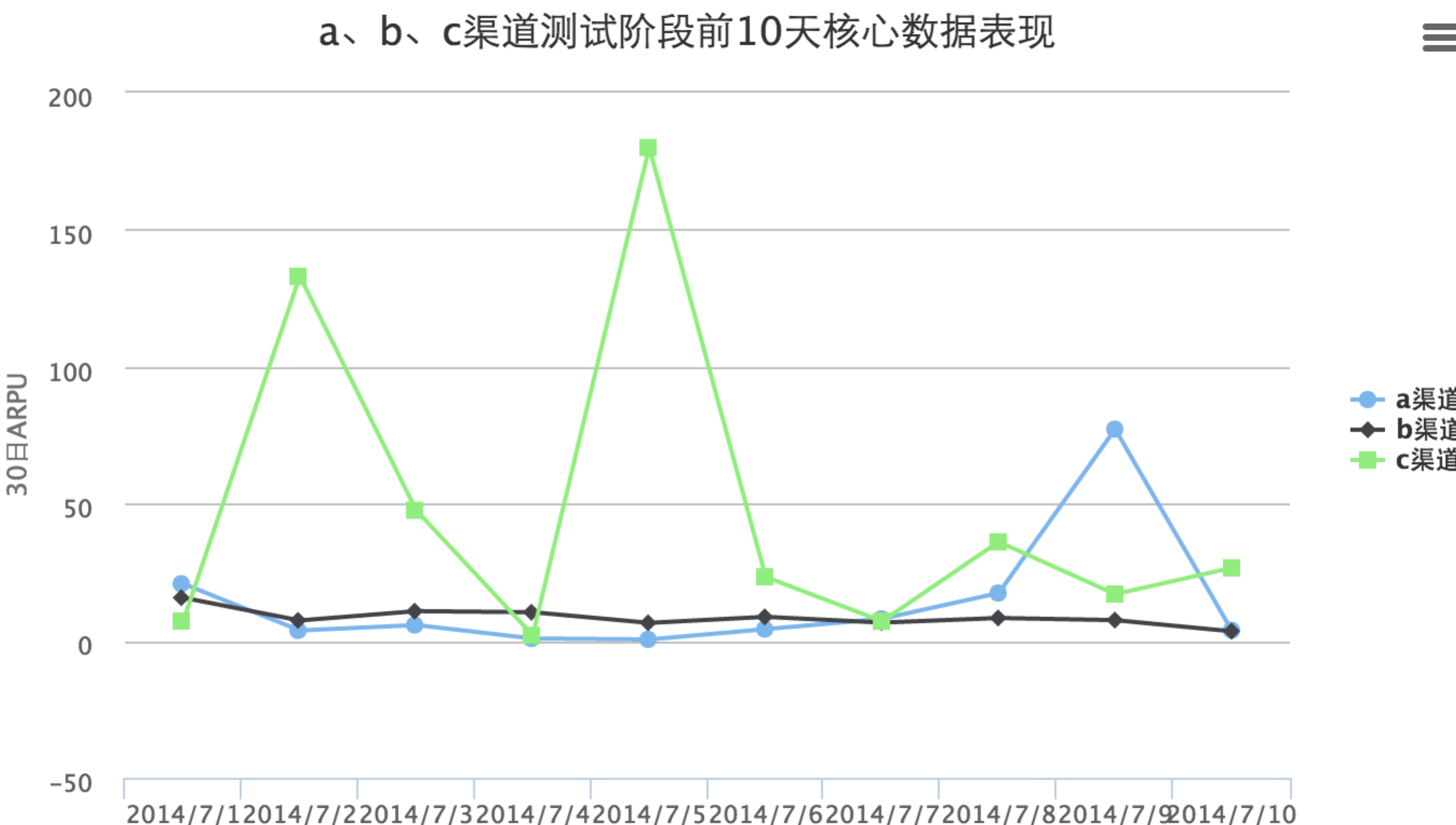
1. 表格1: 针对不同类型的渠道，尽管现有渠道数量相对较少，但是Top5市场份额总和却相对较大，现有渠道数量相对较多，但是Top5市场份额总和却相对较少。
2. 表格2: 三种典型渠道的推广费用对比。
3. 表格3: 从新增量来看，推广前段新增量下降，中段出现大幅上升，后端开始下滑。显然a渠道每天的新增量最大，b渠道其次，c渠道最后。从10天的变化趋势来看，c渠道的后段表现要好很多。

从次日留存来看，c渠道的表现比较好，b渠道比较稳定，a渠道中段出现大量用户流失。



从30日ARPU来看，c渠道会有爆发性的增长出现，a渠道在后段出现一次爆发性增长，b渠道比较平滑。



1. 表格4:

结合表格2来看，在相同的推广费用投入下（假设均投入230万，即分别使用1条a渠道，使用23条b渠道，使用153条c渠道（但是c渠道只有86条，所以我们必须修改方案如下））

（使用1条a渠道，使用23条b渠道）

|  |  |  |  |
| --- | --- | --- | --- |
|  | a渠道 | b渠道 | c渠道 |
| 渠道条数 | 1 | 23 | 86 |
| 推广费用投入 | 230万 | 230万 | 129万 |
| 累计新增量 | 338508 | 592986 | 125388 |
| 累计30日消费 | 4693744 | 5388785 | 4728796 |

综合上述四个表格的分析，我们可知c渠道的top5市场份额总和较大，c渠道单渠道的推广投入费用较低，使用c渠道的用户新增量在后期会相对平稳，使用c渠道的用户流失较少，使用c渠道会出发爆发性的ARPU增长。在相同的推广费用投入下，使用c渠道可以产生更高的收益（累计30日消费），但是新增用户量先对较少（不过对于c渠道来说，这些用户的流失率比较低，属于比较忠诚的用户）。

结论：如果公司希望在最低的推广投入下获得更多的收入，我建议选择使用C类型的渠道来进行推广，但是如果公司希望可以既获得较多的收入，又能积累大量忠实的用户量，我建议选择使用B类型的渠道。