

敏感性问卷调查

标签：概率论与数理统计

背景

学生看毛片会严重影响身心健康发展。但这些都是避着教师和家长进行的，属于个人隐私行为。

现在要设计一个调查方案，从调查数据中估计出学生中看毛片的比率 p 。

思考

- 由于担心个人隐私被泄露，被调查者一般不会配合
- 在公开场合进行这样的调查，一般数据都会失真
- 利用互联网问卷是个不错的选择，但需要pc或者移动终端的支持
- 纸质调查如何设计的让被调查者消除隐私泄露的担心并如实回答呢？

方案

被调查者只需要回答以下两个问题的中的一个，而且只需要回答Yes or No即可：

问题A：

你的生日是否在7月1日之前？

问题B：

你是否看过毛片？

规则

1. 被调查者在没有旁人的情况下，独自一人回答问题；
2. 被调查者从一个罐子中随机抽取一只球，看过颜色后即放回，
若抽到白球则回答问题A；若抽到红球则回答问题B；罐子中只有白球和红球；
3. 被调查者回答完后将答卷放入一密封的投票箱内。

实践

n 张答卷（ n 较大，1000以上），其中有 k 张回答“是”，而我们无法知道此 n 张答卷中有多少张是回答问题B的，同样也无法得知 k 张回答“是”的答卷中有多少张是回答问题B的，担忧两个信息我们是预先知道的：

(1) 在参加人数较多的场合，任选一人其生日在7月1日之前的概率为0.5；

(2) 罐子中红球的比率 π 已知.

现在就是要用这4个数据($n, k, 0.5, \pi$)来求出 p .

由全概率公式知道：

$$P(\text{是}) = P(\text{白球})P(\text{是}|\text{白球}) + P(\text{红球})P(\text{是}|\text{红球}) \quad (\text{Eq.1})$$

所以，将 $P(\text{红球}) = \pi, P(\text{白球}) = 1 - \pi, P(\text{是}|\text{白球}) = 0.5, P(\text{是}|\text{红球}) = p$ 代入上式，同时使用频率 k/n 代替概率 $P(\text{是})$ ，得到：

$$\frac{k}{n} = 0.5(1 - \pi) + p \cdot \pi \quad (\text{Eq.2})$$

由此得到：

$$p = \frac{k/n - 0.5(1 - \pi)}{\pi} \quad (\text{Eq.3})$$

Eq.3就是我们要求的概率。

例如，罐子里有红球30个，白球20个，则 $\pi = 0.6$,调查结束后收到1583张有效答卷，其中有389张回答“是”，由**Eq.3**可以计算得到：

$$p = \frac{389/1583 - 0.5(1 - 0.6)}{0.6} = 0.0762$$

这表明大概有7.62%的人看过毛片哦。

这个敏感性调查可以用在多个场景，如调查人群赌博率、偷税漏税率，吸毒率、作弊率等。