

作业 1

1. (10 分) 继课程第一章中提到的 Francis Galton, Karl Pearson, Ronald Fisher 之后, 20 世纪后半期有哪些著名的统计学家? 请大家自行上网搜索, 写出三位统计学家的名字, 并对每位统计学家使用至少 2~3 句话介绍他们的贡献.

- ① John W. Tukey: 数据科学先驱, 发明了快速傅里叶变换和箱线图, 提出并推广了探索性数据分析, 阐明了探索性分析和验证性数据分析间的重要区别
- ② Donald B. Rubin: 在缺失数据、因果推断、抽样调查、贝叶斯推断等统计学方法上作出基础性贡献, 将统计与心理学、经济学、法律、流行病学、生物医学等领域结合.
- ③ Bradley Efron: 首次系统性提出 Bootstrap 方法, 把构建统计量置信区间变得简单, 将经验贝叶斯、生存分析、指数族、Bootstrap 等理论与应用主题结合.

2. (10 分) 在课程第一章大数据的发展历程中, 我们提到了在大数据成熟阶段, MapReduce 等大数据技术受到追捧, 请大家自行上网搜索, 了解 MapReduce 是一项怎样的大数据技术, 并用一段话描述你理解的 MapReduce 的思想原理。

MapReduce 是一个基于集群的计算平台, 是一个简化分布式编程的计算框架. 它的核心是分治思想, 将分布式计算抽象为 Map 和 Reduce 两个阶段, Mapper 负责“分”, 将数据或计算规模缩小, 进行就地计算和并行计算; Reducer 负责对上面处理的结果进行汇总, 从而进行大数据编程和计算处理。

3. (15 分) 在课程第二章 PPT 第 19 页中, 我们定义了无重复地随机抽取情况下的排列数与组合数, 请推导出从 n 个不同元素中, 有重复地随机抽取(sampling with replacement) k ($k \leq n$) 个, 不同排列方式个数和不同组合方式个数分别是多少?

每次均可抽取 n 个元素 ⋯ 排列种数为 n^k
组合种数为 $\frac{n^k}{k!}$

4. (10分)假设一个房子里有1个客厅和4个房间，你进入客厅后发现了4把钥匙，已知每把钥匙只能打开其中一个房间的门，每个房间的门只能由其中一把钥匙打开(即钥匙和门是一一对应的)。现在每个房间门口站了1个人，你把这4把钥匙随机发给这4个人，让他们尝试打开各自对应的房间。请问至少有1个人能打开房间门的概率是多少？请给出具体计算步骤。

设依次给1号，…，4号房间的钥匙是a号、b号、c号、d号房的，有 X 个是对应的
总种数为 $A_4^4 = 24$ (种)
仅考虑全错排之情形，即 $X=0$ 。

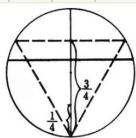
$$(2, 1, 4, 3), (2, 3, 4, 1), (2, 4, 1, 3), (3, 1, 4, 2), (3, 4, 1, 2), (3, 4, 2, 1)$$

$$(4, 1, 2, 3), (4, 3, 1, 2), (4, 3, 2, 1) \quad \text{共9种} \quad \therefore P(X=0) = \frac{9}{24} = \frac{3}{8}$$

$$\therefore P(X \geq 1) = 1 - P(X=0) = \frac{5}{8}$$

5. (15分)在课程第二章PPT第24页中，我们提到了贝特朗悖论(Bertrand's Paradox)，请大家自行上网搜索这个悖论的相关内容，如果你是老师，现在想要向学生讲述这个悖论的内容，你会怎么讲？请把你的讲述内容详细呈现出来(可以是文字+公式描述，也可以是你制作的PPT截图)。

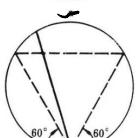
提出问题：利用所学，计算：在圆内任选一条弦，其弦长大于圆内接正三角形边长的概率为？
教授解法：



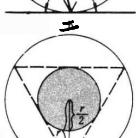
法一：在垂直直径上任取一点，作垂直于该直径的弦，由几何关系知，

取中间 $\frac{1}{2}$ 长度时(即本处至 $\frac{3}{4}$ 处)，其长度大于正△边长。

$$\therefore P_1 = \frac{1}{2}$$



法二：过内接正△的一个顶点作弦，在所夹△两边时，其弦长大于正△边长，而夹角为 60° 。 $\therefore P_2 = \frac{60^\circ}{180^\circ} = \frac{1}{3}$



法三：当弦的中点在圆中阴影中时，弦长大于正△边长，而大圆面积为

$$S_1 = \pi r^2, \text{ 阴影为 } S_2 = \pi (\frac{r}{2})^2 \quad \therefore P = \frac{S_2}{S_1} = \frac{1}{4}$$

列出主题：三种解法三个结果，上述概率问题即为贝特朗悖论
追究原因：导致概率各不相同的原因？

↓
法一样本用长度，法二用角度，法三用面积。

↓
本质：样本空间不同导致！

6. (15分)假定人群中某种疾病的患病率为5%。在检查时，患病者和非患病者被检查出阳性的概率分别为0.98和0.04。

(1) 现从人群中随机抽出一人进行检查，发现其呈阳性，则此人为患病者的概率是多少？

(5分)

(2) 此人又独立地做了一次检查，结果仍然是阳性，请问在两次检查均呈阳性的情况下，此人为患病者的概率是多少？(10分)

设患病为事件X，检测出阳性为事件Y

$$(1) P(X) = 0.05, P(Y|X) = 0.98 \times 0.05 + 0.04 \times 0.95 = 0.087$$

$$P(Y|X) = 0.98$$

$$\therefore P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)} \approx 0.56$$

$$(2) \text{设两次均为阳性为事件} Y^2, P(Y^2) = 0.98^2 \times 0.05 + 0.04^2 \times 0.95 = 0.04954$$

$$P(Y^2|X) = 0.98^2 = 0.9604$$

$$\therefore P(X|Y^2) = \frac{P(Y^2|X) \cdot P(X)}{P(Y^2)} \approx 0.97$$

7. (10分)如果事件A与事件B相互独立，事件B与事件C也相互独立，请问事件A与事件C是否一定相互独立？如果是，请给出证明；如果不是，请举一个反例。

不一定。反例：A = “抛硬币第一次为正”，B = “抛硬币第三次为正”，C = “抛硬币前两次为正”

8. (10分)有这样一个虚拟的国家，这个国家的人重男轻女思想都非常严重，因此每对夫妻生孩子都会遵守这样一个规则：如果第一个孩子是男孩，则不再继续生了，否则一直生到生出一个男孩为止。假设不考虑多胞胎的情况，且每次孕育男孩和女孩的概率是相等的，请问这个国家下一代的男女比例大概是多少？

$$1\text{男0女}: P = \frac{1}{2}, 1\text{男1女}: P = \frac{1}{4}, 1\text{男2女}: P = \frac{1}{8}, \dots$$

则平均家庭男孩比例为 $P = 1 \times \frac{1}{2} + \frac{1}{2} \times \frac{1}{2^2} + \frac{1}{3} \times \frac{1}{2^3} + \dots + \frac{1}{n} \times \frac{1}{2^n} + \dots$

$$\begin{aligned} P &= \sum_{i=1}^{\infty} \frac{1}{i} \cdot \frac{1}{2^i} = \ln 2 \sum_{i=1}^{\infty} \frac{x^i}{i} = \ln 2 \int_0^x \sum_{i=1}^{\infty} x^{i-1} dx = \int_0^x \frac{1}{1-x} dx \\ &= \left[\ln(1-x) \right]_0^x = -\ln(1-x) \Big|_0^{\frac{1}{2}} = \ln 2 \quad \text{即男女比为 } \ln 2 : (1 - \ln 2) \end{aligned}$$

9. (5分)期末报告将以小组的形式展开(自由组队,3人一组)，在期中考试周后(第10周开始)，每节课会有一至两个小组进行报告，最后一节课(2024年6月3日)会有一次集中报告。请你寻找另外两名同学组成一组，把你们三人的名字写下来。

李衍熙 12312110

陈鹏鹏如 12112101

陈炫伊 12311451