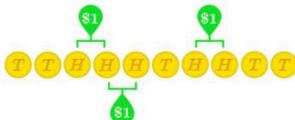


作业 2

1. (10 分) 有这样一个游戏: 你抛一枚均匀的硬币 10 次, 按次序记下正反面的情况, 每连续出现一对正面, 你就能获得 1 枚金币。例如, 如果一次游戏的结果是 TTHHHTHHTT(H 表示正面, T 表示反面), 你将获得 3 枚金币, 如下图所示。请问这个游戏的期望奖励是多少? 请给出具体计算细节。



在任意两枚连续硬币中, 出现“H-H”的概率为 $P(HH) = P(H) \times P(H) = \frac{1}{4}$

10 次抛掷中, 可形成 9 个这样的单元, 且每次抛掷均相互独立

$$\therefore \text{期望 } E = P(HH) \times 9 = 2.25$$

2. (10 分) 随机变量 X 服从泊松分布 $P(\lambda)$, 请给出其期望和方差的具体计算步骤。

$$P(X=k) = e^{-\lambda} \frac{\lambda^k}{k!}, k=0, 1, \dots, \lambda$$

$$\therefore E(X) = \sum_{k=0}^{\infty} k \cdot P(X=k) = \sum_{k=0}^{\infty} k \cdot \frac{\lambda^k}{k!} e^{-\lambda} = \lambda e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} \stackrel{\text{Taylor}}{=} \lambda e^{-\lambda} \cdot e^{\lambda} = \lambda$$

$$D(X) = E(X^2) - [E(X)]^2$$

$$\begin{aligned} E(X^2) &= E[X(X-1)+X] = \sum_{k=0}^{\infty} k(k-1) \frac{\lambda^k}{k!} e^{-\lambda} + E(X) = \lambda^2 e^{-\lambda} \sum_{k=2}^{\infty} \frac{\lambda^{k-2}}{(k-2)!} + \lambda \\ &= \lambda^2 \cdot e^{-\lambda} \cdot e^{\lambda} + \lambda = \lambda^2 + \lambda \end{aligned}$$

$$\therefore D(X) = \lambda^2 + \lambda - \lambda^2 = \lambda$$

3. (10分)第一章中提到的Francis Galton做过一个实验叫高尔顿钉板实验，请你自行上网搜索学习，解释这是一个什么实验，它和中心极限定理有什么关系？

实验：让一些小球垂直下落，中途有序摆放一些钉子，使得小球每次碰到后向左向右的概率相等，最终落到下方区域，下方区域聚集的数量即代表概率分布。

关系：中心极限定理指随机变量的算术平均渐近服从正态分布，当实验的小球数足够多时，下方的累积效果即呈现正态分布，符合中心极限定理。
(各位置小球数)

4. (10分)在第四章数据的收据中，我们讲述了几个由于样本不具有代表性而得出错误结论的例子，包括美国大选民意调查、长相与智商、战斗机加固。除了这几个例子，请你再讲述一个类似的例子，介绍故事的背景、样本有什么问题、得出了什么错误结论等。

1985年可口可乐想要创新，将新老口味可乐让大众进行对比，消费者样本有20多万份，调查结果显示53%的人喜欢新口味，47%的人喜欢老口味，总体上新口味可乐更受欢迎。但新可乐上市后却遭到大量投诉和反对。

样本问题：和可口可乐受众相比有偏差，样本量不够，新老口味喜欢的人比例相差不大，无法下结论。

结论：错误地认为新口味更受欢迎，实际上新老口味均有广泛群体。

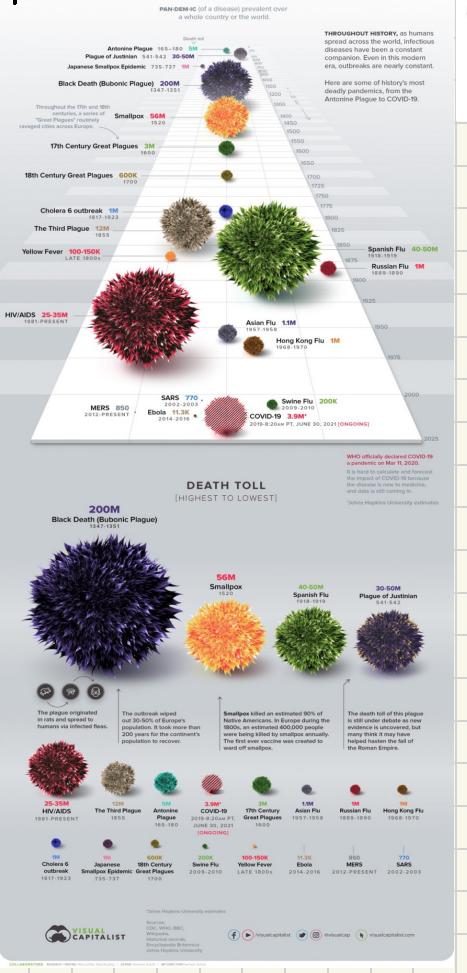
5. (10分)除课堂上讲过的数据可视化例子外，请你讲述一个令你印象深刻的的数据可视化例子，介绍数据的背景、让你印象深刻的原因、数据可视化有什么特色等。

背景：Nicholas Lepan制作的大流行病历史的信息图，讲述了人类历史上所有已知的大流行病，包括病名、死亡人数和大流行病发生的年份。

原因：将刚经历的疫情也算计在内，并和以前的病毒作对比，在人类进程中并不很起眼。

特色：鲜明地展示了病毒的危害性，死亡人数多少和时间先后一目了然。

T5. HISTORY OF PANDEMICS



6. (15分)在第四章数据的可视化中，我们用R的ggplot2软件包画了包括柱状图、直方图、箱线图、相关热图在内的四种常见统计图形。请你自行上网搜索，挑选另外两种统计图形，并选择一个ggplot2软件包中自带的数据集，用你喜欢的颜色画出这两种统计图形。请在答题区给出对这两种统计图形的介绍、所画图形的截图、作图使用的R语言代码。

条形图 (Bar Plot)

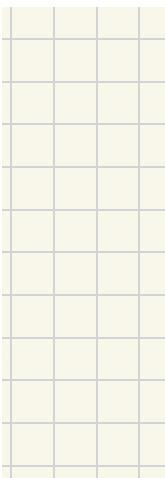
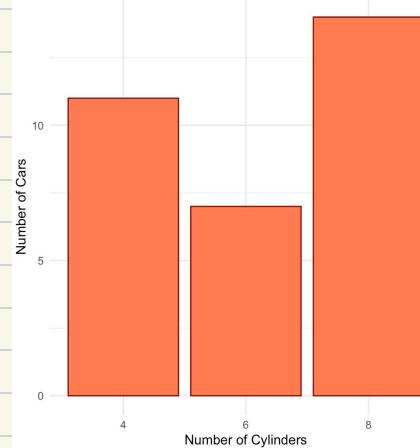
条形图是表示类别数据与数值大小关系的图形。这种图形特别适用于比较不同类别或组的数量或频率。

```
# 计算每种气缸数量的汽车数量
car_count_by_cyl <- table(mtcars$cyl)

# 将表格转换为数据框，以便用ggplot绘图
car_count_df <- as.data.frame(car_count_by_cyl)

# 绘制气缸数量的条形图
ggplot(car_count_df, aes(x = Var1, y = Freq)) +
  geom_bar(stat = "identity", fill = "coral", color = "darkred") +
  labs(title = "Number of Cars by Cylinder Count",
       x = "Number of Cylinders",
       y = "Number of Cars") +
  theme_minimal() # 使用最小化主题
```

Number of Cars by Cylinder Count



7. (15分) X_1, X_2, \dots, X_n 和 Y_1, Y_2, \dots, Y_m 为来自同一个总体的两个独立的简单随机样本，样本容量分别为 n, m 。令总体的均值为 θ_1 ，方差为 θ_2 ， \bar{X}, \bar{Y} 分别为两个样本的样本均值。

(1) 证明对于任意满足 $a + b = 1$ 的常数 a, b ， $\hat{\theta}_1 = a\bar{X} + b\bar{Y}$ 都是 θ_1 的无偏估计量。(5分)

(2) 确定使得(1)中定义的 $\hat{\theta}_1$ 的方差达到最小的常数 a, b 。(10分)

$$(1) E(\bar{X}) = E(\bar{Y}) = \theta_1, V(\bar{X}) = \frac{\theta_2}{n}, V(\bar{Y}) = \frac{\theta_2}{m}$$

$$E(\hat{\theta}_1) = E(a\bar{X} + b\bar{Y}) = aE(\bar{X}) + bE(\bar{Y}) = (a+b)\theta_1 = \theta_1$$

$\therefore \hat{\theta}_1$ 是 θ_1 的无偏估计量

$$(2) V(\hat{\theta}_1) = V(a\bar{X} + b\bar{Y}) = V(a\bar{X}) + V(b\bar{Y}) + 2\text{cov}(a\bar{X}, b\bar{Y}) = a^2 V(\bar{X}) + b^2 V(\bar{Y}) + 2ab\text{cov}(\bar{X}, \bar{Y})$$

$$\text{cov}(\bar{X}, \bar{Y}) = E(\bar{X}\bar{Y}) - E(\bar{X})E(\bar{Y}) = \theta_1^2 - \theta_1^2 = 0$$

$$\therefore V(\hat{\theta}_1) = a^2 \frac{\theta_2}{n} + b^2 \frac{\theta_2}{m} = (a^2 + b^2) \theta_2 \quad \because a+b=1 \quad \therefore V(\hat{\theta}_1) = \left(\frac{a^2}{n} + \frac{(1-a)^2}{m}\right) \theta_2$$

$$\frac{dV(\hat{\theta}_1)}{da} = \left(\frac{2a}{n} - \frac{2(1-a)}{m}\right) \theta_2 = 0 \Rightarrow a = \frac{n}{m+1} \text{时, } V(\hat{\theta}_1) \text{ 最小. } b = \frac{m+1-n}{m+1}$$

小提琴图 (Violin Plot)

小提琴图是用于展示数值数据分布及其概率密度的图形。这种图形特别适用于比较多个组或类别中的数据分布。

(见下页)

8. (15分) 美国劳工统计局发现，一个容量 $n = 6000$ 的样本中有 516 个人失业。

(1) 求美国总体失业率的 95% 置信区间。(5分)

(2) 在显著性水平 $\alpha = 0.05$ 下检验美国总体失业率是否高于 8%。(10分)

```
library(ggplot2)
```

```
# 绘制气缸数量(cyl)对每加仑行驶英里数(mpg)影响的小提琴图  
ggplot(mtcars, aes(factor(cyl), mpg)) +  
  geom_violin(fill = "lightblue", color = "darkblue") + # 设置小提琴图填充和边框颜色  
  labs(title = "MPG Distribution by Cylinder Count",  
       x = "Number of Cylinders",  
       y = "Miles Per Gallon") +  
  theme_minimal() # 使用最小化主题
```

(1) 样本失业率 $P = \frac{516}{6000} = 0.086$

$$\text{标准误差 } SE = \sqrt{\frac{P(1-P)}{n}} = \sqrt{\frac{0.086(1-0.086)}{6000}} \approx 0.0036$$

$$\text{上界 } U_1 = 0.086 + Z \times 0.0036 \quad \text{下界 } L_2 = 0.086 - Z \times 0.0036$$

$$95\% \text{ 的置信度 } Z = 1.96 \Rightarrow \begin{cases} U_1 \approx 0.093 \\ L_2 \approx 0.079 \end{cases}$$

∴ 95% 置信区间为 0.079 ~ 0.093

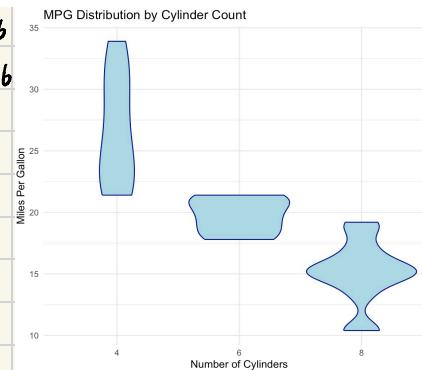
(2) 假设 $H_0: p \leq 0.08$, $H_1: p > 0.08$

$$\text{标准误差 } SE = \sqrt{\frac{p(1-p)}{n}} \approx 0.0035$$

$$Z = (\hat{p} - p) / SE = (0.086 - 0.08) / 0.0035 \approx 1.714$$

由表, $\alpha = 0.05$ 时, $Z^* = 1.645$

而 $Z = 1.714 > 1.645$ 在拒绝域内 ∴ $\alpha = 0.05$ 时, 能得出总体失业率高于 8% 的结论.



9. (5分) 期末报告大家已经分好小组, 老师设置了一个腾讯文档供大家选择期末报告的时间, 最后一节课(2024年6月3日)会有一次集中报告, 可容纳8个小组, 其他小组会分布在10-15周的上课时间, 每次课程最多两个小组。请各个小组商量好报告时间, 点击[这里](#)填写选好的时间, 并将报名截图贴至答题区, 如果你想选择的时间已经被其他小组优先占用, 请不要修改其他小组的选择。期末报告指引见课程 Blackboard/学习评价/期末报告。

21:46 Sat Mar 30

我的作业 - 魅力统计 (2024春)

docs.qq.com

2024春魅力统计期末报告安排

	A	B	C	D	E	F	G	H	I
1	周次	日期	星期	报名队伍成员姓名	备注				
2	10	2024/4/22	星期一	队伍1:肖泽昊, 肖健坤, 廖钰欣 队伍2:肖泽瑞, 乔佳明, 李思婧					
4	11	2024/4/29	星期一	队伍1:黄宇航, 仇昊辰, 李世豪 队伍2:付雨, 陈雨菲, 陈思行					
6	12	2024/5/6	星期一	队伍1:潘浩舟, 马康文, 张子木 队伍2:黄云毅, 徐璇喆, 杨倩羽					
8	2024/5/11		星期六	队伍1:徐锐, 张笑宇, 姚袁明 队伍2:宋子墨, 傅文楷, 魏华坤	劳动节统一调课				
10	13	2024/5/13	星期一	队伍1:马婧, 韦雅琳, 张欣然 队伍2:郑文莉, 柳莺佳, 张雨洋					
12		2024/5/15	星期三	队伍1:张晓雨, 韦娅丽, 肖世泽 队伍2:陈政, 刘逸琪, 张海元					
14	14	2024/5/20	星期一	队伍1:徐欣, 赵孟飞, 姜跃翀 队伍2:袁司机, 刘凡哲, 康琪雪					
16	15	2024/5/27	星期一	队伍1:孙圣凯, 陈梓民, 董子领 队伍2:张皓诚, 江子涵, 任轩锐					
18		2024/5/29	星期三	队伍1:李轩然, 陈鹏如, 陈炫伊 队伍2:					
20				队伍1:					
21				队伍2:					
22				队伍3:					
23				队伍4:					
24				队伍5:					
25				队伍6:					
26				队伍7:					
27				队伍8:					

5.29

周三

队伍一