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$\theta_1$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) $V(\hat{\theta}_1) = V(a\bar{X}+b\bar{Y}) = V(a\bar{X}) + V(b\bar{Y}) + 2\,cov(a\bar{X}, b\bar{Y}) = a^2V(\bar{X}) + b^2V(\bar{Y}) + 2ab\,cov(\bar{X}, \bar{Y})$

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

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

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

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)

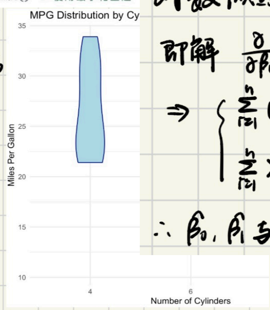
8. (15 分)美国劳工统计局发现，一个容量$n = 6000$人的样本中有 516 个人失业。

(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)

```r
library(ggplot2)

# 绘制气缸数量 (cyl) 对每加仑行驶英里
ggplot(mtcars, aes(factor(cyl), mpg))
  geom_violin(fill = "lightblue", co
  labs(title = "MPG Distribution by (
       x = "Number of Cylinders",
       y = "Miles Per Gallon") +
  theme_minimal()  # 使用最小化主题
```



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

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

上界 $\theta_1 = 0.086 + Z\times0.0036$ 下界 $\theta_2 = 0.086 - Z\times0.0036$

95%的置信度 $Z = 1.96$ $\Rightarrow \begin{cases} \theta_1 \approx 0.093 \\ \theta_2 \approx 0.079 \end{cases}$

$\therefore$ 95%置信区间为 0.079 ~ 0.093

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

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 $\therefore \alpha = 0.05$时，能得出总体失业率高于8%的结论.

1. (10 分) 在线性回归模型$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$下，假设$\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$独立并都服从$N(0, \sigma^2)$，请使用极大似然法估计$\beta_0$和$\beta_1$，得到$\hat{\beta}_0$和$\hat{\beta}_1$的形式，并与最小二乘法的估计量比较，它们是否有区别?

似然函数：$L(\beta_0, \beta_1, \delta^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\delta^2}} e^{-\frac{\varepsilon_i^2}{2\delta^2}} = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\delta^2}} e^{-\frac{(Y_i - \beta_0 - \beta_1 X_i)^2}{2\delta^2}}$

对数似然函数：$\ln L(\beta_0, \beta_1, \delta^2) = -\frac{n}{2}\ln(2\pi) - n\ln\delta - \sum_{i=1}^{n} \frac{(Y_i - \beta_0 - \beta_1 X_i)^2}{2\delta^2}$

即解 $\frac{\partial}{\partial\beta_0} \ln L(\beta_0, \beta_1, \delta^2) = 0$ 和 $\frac{\partial}{\partial\beta_1} \ln L(\beta_0, \beta_1, \delta^2) = 0$

$\Rightarrow \begin{cases} \sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0 \\ \sum_{i=1}^{n} X_i(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0 \end{cases} \Rightarrow \begin{cases} \hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \\ \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \end{cases}$

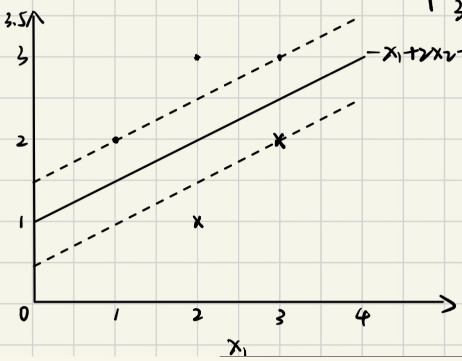
$\therefore$ $\hat{\beta}_0, \hat{\beta}_1$与最小二乘法的形式相同

7. (10 分) 假设有观测数据$\{(y_i, x_{i1}, x_{i2}),\ i = 1, 2, \dots, 5\} = \{(1,1,2), (1,2,3), (1,3,3), (-1,2,1), (-1,3,2)\}$. 试求出最大间隔分离超平面的表达式，并手动画一个二维散点图，在图上画出该分离超平面及支持向量.

经优化，即求 $\min \frac{1}{2}\|W\|^2$，使得 $i = 1, \dots, 5$，$y_i(W^T x_i + b) \ge 1$

即 $y_i(W_1 x_{i1} + W_2 x_{i2} + b) \ge 1$. $\Rightarrow \begin{cases} W_1 + 2W_2 + b \ge 1 \\ 2W_1 + 3W_2 + b \ge 1 \\ 3W_1 + 3W_2 + b \ge 1 \\ 2W_1 + W_2 + b \le -1 \\ 3W_1 + 2W_2 + b \le -1 \end{cases}$ $\Rightarrow W_1 = -1, W_2 = 2, b = -2$

此时 $\min \frac{1}{2}\|W\|^2 = \frac{5}{2}$



$-x_1 + 2x_2 - 2 = 0$ 为分离超平面

(1,2), (3,2), (3,3) 为支持向量

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) $P(X) = 0.05$. $P(Y) = 0.98\times0.05 + 0.04\times0.95 = 0.087$

$P(Y|X) = 0.98$

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

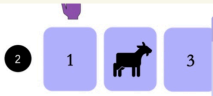
(2) 设两次均为阳性为事件Y^2. $P(Y^2) = 0.98^2\times0.05 + 0.04^2\times0.95 = 0.04954$

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

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

例2.7

请使用条件概率给出Monty Hall问题(即三门问题)的解答。



例2.7解答

- 首先用数学语言描述问题(假设你选择的是1号门，主持人打开的是2号门)。
- 令事件A_1, A_2, A_3分别表示主持人未开门时1、2、3号门后面有车，事件B表示主持人打开的是2号门，那么有

$$P(A_1) = P(A_2) = P(A_3) = \frac{1}{3}, P(B|A_1) = \frac{1}{2}, P(B|A_2) = 0, P(B|A_3) = 1.$$

- 我们想求的概率是$P(A_1|B)$与$P(A_3|B)$. 由全概率公式及乘法定律，有

$$P(B) = P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + P(B|A_3)P(A_3) = \frac{1}{2}\times\frac{1}{3} + 1\times\frac{1}{3} = \frac{1}{2},$$

$$\Rightarrow P(A_1|B) = \frac{P(B|A_1)P(A_1)}{P(B)} = \frac{1/2 * 1/3}{1/2} = \frac{1}{3},\ P(A_3|B) = \frac{P(B|A_3)P(A_3)}{P(B)} = \frac{1 * 1/3}{1/2} = \frac{2}{3}.$$

- 所以应该改选3号门！

例5.2

设$X_1, X_2, \dots, X_n$为来自总体X的一个简单随机样本, 总体的均值为$E(X) = \theta_1$, 方差为$\mathrm{Var}(X) = \theta_2$, θ_1, θ_2为未知参数, 试证明样本均值$\bar{X}$和样本方差S^2分别是θ_1和θ_2的无偏估计量.

例5.2 解答

$$E(\bar{X}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \theta_1. \quad \mathrm{Var}(\bar{X}) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) = \frac{\theta_2}{n}.$$

$$\because (n-1)S^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n}[(X_i - \theta_1) + (\theta_1 - \bar{X})]^2 = \sum_{i=1}^{n}(X_i - \theta_1)^2 - n(\bar{X} - \theta_1)^2$$

$$\therefore (n-1)E(S^2) = \sum_{i=1}^{n} E(X_i - \theta_1)^2 - nE(\bar{X} - \theta_1)^2 = \sum_{i=1}^{n}\theta_2 - n\frac{\theta_2}{n} = (n-1)\theta_2$$

$$\therefore E(S^2) = \theta_2.$$

例5.4

设总体 X 的概率分布为

X	0	1	2	3
P	θ^2	$2\theta(1-\theta)$	θ^2	$1-2\theta$

其中 $0 < \theta < 0.5$是未知参数。现有来自该总体的一组简单随机样本的观测值: 3, 1, 3, 0, 3, 1, 2, 3, 求参数θ的矩估计值与极大似然估计值。

例5.4 解答

为求矩估计量, 考虑总体的期望:

$$E(X) = 0\times\theta^2 + 1\times 2\theta(1-\theta) + 2\times\theta^2 + 3\times(1-2\theta) = 3 - 4\theta.$$

由此可得 θ 的矩估计量为

$$\hat{\theta} = \frac{3 - \bar{X}}{4}.$$

由样本观测值算得

$$\bar{x} = \frac{3+1+3+0+3+1+2+3}{8} = 2 \Longrightarrow \theta \text{ 的矩估计值为 } \hat{\theta} = 1/4.$$

例5.4 解答

对于给定的样本观测值, 似然函数为:

$$L(\theta) = [\theta^2]^1\times[2\theta(1-\theta)]^2\times[\theta^2]^1\times[1-2\theta]^4 = 4\theta^6(1-\theta)^2(1-2\theta)^4.$$

对似然函数取对数得到对数似然函数(log-likelihood function)

$$\ell(\theta) = \ln L(\theta) = \ln 4 + 6\ln\theta + 2\ln(1-\theta) + 4\ln(1-2\theta).$$

求对数似然函数的一阶导, 并令其等于0:

$$0 = \frac{d\ell(\theta)}{d\theta} = \frac{6}{\theta} - \frac{2}{1-\theta} - \frac{8}{1-2\theta} = \frac{24\theta^2 - 28\theta + 6}{\theta(1-\theta)(1-2\theta)}.$$

解方程 $24\theta^2 - 28\theta + 6 = 0$可得

$$\hat{\theta} = \frac{7 \pm \sqrt{13}}{12},$$

而由于参数 θ需满足 $0 < \theta < 0.5$, 因此θ的极大似然估计值为

$$\hat{\theta} = \frac{7 - \sqrt{13}}{12}.$$

例5.5

设$X_1, X_2, \dots, X_n$为来自总体$X \sim N(\mu, \sigma^2)$的一个简单随机样本, 已知$\sigma^2 = 1$, 求未知参数μ的95%置信区间. 若给定样本观测值4.6, 4.2, 5.0, 3.1, 3.4, 2.4, 4.4, 3.2, 3.9, 4.0, 求该95%置信区间的具体数值.

例5.5 解答

- 构造区间估计的一个自然思路是从点估计出发, 以点估计为中心加减一个量构成一个区间.
- 在本例中, μ的一个很好的点估计是$\bar{X}$, 它是μ的无偏估计. 按置信区间的定义我们希望确定数值$c_1 > 0, c_2 > 0$, 使得

$$P\left(\bar{X} - c_1 < \mu < \bar{X} + c_2\right) = 1 - \alpha.$$

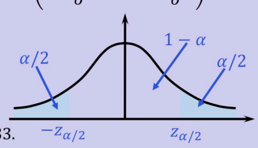
- 上式的等价表示是$P\left(\mu - c_2 < \bar{X} < \mu + c_1\right) = 1 - \alpha$.
- 显然, c_1, c_2的确定依赖于$\bar{X}$的分布. 前面我们提到过 $\bar{X} \sim N(\mu, \sigma^2/n)$, 因此$c_1, c_2$要满足

$$1 - \alpha = P(\mu - c_2 < \bar{X} < \mu + c_1) = P\left(\frac{-c_2}{\sigma/\sqrt{n}} < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < \frac{c_1}{\sigma/\sqrt{n}}\right) = P\left(-\frac{\sqrt{n}c_2}{\sigma} < Z < \frac{\sqrt{n}c_1}{\sigma}\right),$$

- 其中$Z \sim N(0,1)$. 为使得区间宽度$c_1 + c_2$尽可能小, 不难得到

$$c_1 = c_2 = z_{\alpha/2}\frac{\sigma}{\sqrt{n}},$$

- 其中z_α为标准正态分布的上α分位点, 可通过查表得到.
- 常用标准正态分布分位点值: $z_{0.05} = 1.645, z_{0.025} = 1.96, z_{0.01} = 2.33$.



例5.5 解答

- 因此, 参数μ的$100(1-\alpha)\%$置信区间是 (σ^2已知的情况下)

$$\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right).$$

- 根据样本观测值算出$\bar{x} = 3.82$, 代入$\sigma^2 = 1$及$z_{0.025} = 1.96$可得参数μ的95%置信区间的具体数值:

$$3.82 \pm 0.62 = (3.20, 4.44).$$

例6.1解答续

基于给定的数据估计σ^2, 根据$\hat{y}_i = -266.5344 + 6.1376x_i$可计算$\hat{y}_i$的值:

i	1	2	3	4	5	6	7	8	9	10
x_i	63	64	66	69	69	71	71	72	73	75
y_i	127	121	142	157	162	156	169	165	181	208
$\hat{y}_i$	120.12	126.27	138.55	156.96	156.96	169.24	169.24	175.37	181.51	193.79
$(y_i - \hat{y}_i)^2$	47.14	27.79	11.92	0.0016	25.40	175.17	0.055	107.59	0.261	202.05

由此可计算得到$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = 597.386$, 进一步有

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{597.386}{8} = 74.6733 \Longrightarrow \hat{\sigma} = \sqrt{74.6733} \approx 8.6414.$$

$\hat{\sigma}$在R中使用`lm()`函数后的输出里也有给出

```
Residual standard error: 8.641 on 8 degrees of freedom
Multiple R-squared:  0.897,     Adjusted R-squared:  0.8841
F-statistic: 69.67 on 1 and 8 DF,  p-value: 3.214e-05
```