北京大学数学科学学院数理统计期末试题

2022–2023 年第一学期

考试科目___数理统计___　　考试时间 2022 年 12 月 19 日

姓名________________　　学号________________

本试题共 7 道大题，满分 100 分。

1. (12 分) 设 $X_1, \ldots, X_n$ 是来自正态分布 $N(\mu, \sigma_0^2)$ 的简单随机样本，σ_0^2 已知。

(1) 求 μ 的最大似然估计；

(2) 求 μ 的矩估计；

(3) 求 Fisher 信息量 $I(\mu)$；

(4) 求 μ 的无偏估计的方差下界；

(5) (1) 中的最大似然估计是否是 μ 的最小方差无偏估计 (需说明理由)？

(6) 试找出 μ^2 的一个无偏估计。

1. (12 points) Suppose $X_1, \ldots, X_n$ are i.i.d random samples from normal distribution $N(\mu, \sigma_0^2)$, with σ_0^2 known.

(1) Find the MLE of μ;

(2) Find the moment estimate of μ;

(3) Calculate the Fisher information $I(\mu)$ of μ;

(4) Find the variance lower bound of unbiased estimator for μ;

(5) Is the MLE in (1) the UMVUE of μ (Please give your reason)?

(6) Try to find an unbiased estimator for μ^2.

2. (14 分) 若随机变量 X 的分布密度可取下面的 $f_0(x)$ 或 $f_1(x)$：

$$f_0(x) = \begin{cases} \sqrt{\frac{2}{\pi}} e^{-\frac{x^2}{2}} & \text{当} x \geq 0 \\ 0 & \text{其他} \end{cases} ; \quad f_1(x) = \begin{cases} x e^{-\frac{x^2}{2}} & \text{当} x \geq 0 \\ 0 & \text{其他} \end{cases}$$

基于 X 的一个观测值，对检验问题 $H_0 : f(x) = f_0(x) \leftrightarrow H_1 : f(x) = f_1(x)$, 利用 N-P 引理求检验水平为 α 的 UMP 检验 ϕ，并求其第二类错误的概率。

2. (14 points) Suppose random variable X has density $f_0(x)$ or $f_1(x)$:

$$f_0(x) = \begin{cases} \sqrt{\frac{2}{\pi}} e^{-\frac{x^2}{2}} & \text{when } x \geq 0 \\ 0 & \text{otherwise} \end{cases} ; \quad f_1(x) = \begin{cases} x e^{-\frac{x^2}{2}} & \text{when } x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

Based on an observation of X, for the hypothesis testing $H_0 : f(x) = f_0(x) \leftrightarrow H_1 : f(x) = f_1(x)$, use N-P lemma to get a UMP test ϕ of significance level α, and calculate the probability of Type 2 error.

3. (14 分) 设 $X_1, ..., X_n$ 是 i.i.d. 随机变量, 其密度函数为:

$$f(x) = \begin{cases} \frac{1}{\theta} e^{-(x-\theta)/\theta} & \text{当} x \geq \theta \\ 0 & \text{其他} \end{cases}$$

其中 $\theta \geq 0$ 未知。

(1) 证明 $X_{(1)}/\theta$ 是枢轴量，其中 $X_{(1)}$ 是最小次序统计量。

(2) 基于 (1) 中的枢轴量，求 θ 的置信区间 (置信度为 $1-\alpha$)。

3. (14 points) Suppose $X_1, \ldots, X_n$ are iid ramdom variables having following p.d.f.

$$f(x) = \begin{cases} \frac{1}{\theta} e^{-(x-\theta)/\theta} & \text{when } x \geq \theta \\ 0 & \text{otherwise} \end{cases}$$

where $\theta \geq 0$ is unknown.

(1) Show that $X_{(1)}/\theta$ is a pivotal quantity, where $X_{(1)}$ is the smallest order statistic.

(2) Obtain a confidence interval (with confidence level $1-\alpha$) for θ based on the pivotal quantity in (1).

4. (16 分) 设 ANOVA 模型 $Y_{ij}=\mu+\tau_i+\epsilon_{ij}$, $i=1,2,...,I$，$j=1,...,J$，其中 $\sum_{i=1}^{I}\tau_i=0$，ϵ_{ij} 相互独立且 $\epsilon_{ij}\sim N(0,\sigma^2)$，其中 σ^2 未知。定义

$$SS_{TOT}=\sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{ij}-\overline{Y}_{..})^2;$$
$$SS_W=\sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{ij}-\overline{Y}_{i.})^2;$$
$$SS_B=\sum_{i=1}^{I}J(\overline{Y}_{i.}-\overline{Y}_{..})^2.$$

其中 $\overline{Y}_{i.}=\frac{1}{J}\sum_{j=1}^{J}Y_{ij}$, $\overline{Y}_{..}=\frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J}Y_{ij}$.

(1) 证明 $SS_{TOT}=SS_W+SS_B$。

(2) 计算 $E(SS_W)$ 和 $E(SS_B)$。

(3) 基于 (2) 的结果, 给出 σ^2 的一个无偏估计。

4. (16 points) Suppose an ANOVA model: $Y_{ij}=\mu+\tau_i+\epsilon_{ij}$, $i=1,2,...,I$ and $j=1,...,J$, where $\sum_{i=1}^{I}\tau_i=0$. The $\epsilon_{ij}\sim N(0,\sigma^2)$ are independent, where σ^2 is unknown. Define

$$SS_{TOT}=\sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{ij}-\overline{Y}_{..})^2;$$
$$SS_W=\sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{ij}-\overline{Y}_{i.})^2;$$
$$SS_B=\sum_{i=1}^{I}J(\overline{Y}_{i.}-\overline{Y}_{..})^2.$$

where $\overline{Y}_{i.}=\frac{1}{J}\sum_{j=1}^{J}Y_{ij}$ and $\overline{Y}_{..}=\frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J}Y_{ij}$.

(1) Show that $SS_{TOT}=SS_W+SS_B$.

(2) Calculate $E(SS_W)$ and $E(SS_B)$.

(3) Based on the result of (2), give an unbiased estimator of σ^2.

5. (14 分) 设 $X_1,\dots,X_n$ 为参数为 p 的伯努利分布的独立同分布随机样本, 即 $p(X_i=1)=p$ 并且 $P(X_i=0)=1-p$。考虑先验分布为 $[0,1]$ 区间上的均匀分布, $U(0,1)$。

(1) 在损失函数 $L(\hat{p}, p) = \frac{(\hat{p}-p)^2}{p(1-p)}$ 下，求 p 的贝叶斯估计量。

(2) 证明 (1) 中得到的贝叶斯估计量也是最大最小估计量。

你可以不加证明地使用:

$$\int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

$$\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$$

5. (14 points) Let $X_1, \ldots, X_n$ be iid random sample from $Bernoulli(p)$, which means that $p(X_i = 1) = p$ and $P(X_i = 0) = 1 - p$. Consider a prior of p with uniform distribution $U(0, 1)$.

(1) Under loss function $L(\hat{p}, p) = \frac{(\hat{p}-p)^2}{p(1-p)}$, find the bayes estimator of p.

(2) Prove that the bayes estimator you derive in (1) is also a minimax estimator.

You can use the fact that:

$$\int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

$$\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$$

6. (16 分) 回归方法可以用来预测葡萄的产量。在七月，葡萄会结出浆果，浆果丛的面积可以用来预测收获时葡萄的最终产量。假如我们有来自于一项关于浆果丛面积 (x) 和收获产量 (y) 之间关系的研究。假设 $x_1, \ldots, x_n$ 和 $y_1, \ldots, y_n$ 是 n 个不同年份记录的浆果丛面积以及收获产量。我们对浆果丛面积和收获产量之间的关系假设一个线性模型 $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$，其中 ϵ_i 相互独立且服从正态分布 $N(0, \sigma^2)$, 其中 σ^2 未知，并用最小二乘法来对 β_0 和 β_1 进行估计。

(1) 说明如何检验:$H_0 : \beta_1 = 0$ versus $H_a : \beta_1 \neq 0$. (检验的显著水平为 $\alpha = 0.05$)。

(2) 给出 $\mu_0 = \beta_0 + \beta_1 x_0$ 的一个 95% 置信区间。

(3) 给出七月份浆果丛面积为 x_0 英亩 ($x = x_0$) 的葡萄田在收获时的收获产量的 95% 预测区间 (prediction interval)。

(4) 给出 R^2 统计量的计算公式。

6. (16 分) The regression method can be used to predict crop yield of grapes. In July, the grape vines produce clusters of berries, and a count of these clusters can be used to predict the final crop yield at harvest time. Suppose we have a portion of data taken from a study on the cluster counts (x) and yields (y). Suppose $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$ are cluster counts and yields collected in n different years. Suppose that a linear model $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ (suppose ϵ_i are independent and normal-distributed with mean zero and unkonwn variance σ^2) is fit by the method of least squares to the data.

(1) Show how to test whether $H_0 : \beta_1 = 0$ versus $H_a : \beta_1 \neq 0$ (significant level is set to be $\alpha = 0.05$).

(2) Derive a 95% confidence interval for $\mu_0 = \beta_0 + \beta_1 x_0$.

(3) Derive a 95% prediction interval for the crop yields of grapes whose cluster count is x_0 in July ($x = x_0$).

(4) Give the formula of R^2 statistic.

7. (14 分) 设 $Y = X\beta + e$, 其中 X 是 $n \times p$ 矩阵 (秩为 p)，β 是 p 维未知参数向量，$e = (e_1, \ldots, e_n)'$, $e_1, \ldots, e_n$ 相互独立同分布，共同分布为 $N(0, \sigma^2)$ (σ^2 未知) ($n > p \geq 2$)。我们可以将 X 和 β 写作 $X = (X_1, X_2)$，$\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}$ 其中 X_1 是一个 $n \times (p-s)$ 的矩阵并且 X_2 是一个 $n \times s$ 的矩阵, $\beta_1 \in R^{p-s}$ 并且 $\beta_2 \in R^s$。假设我们想要检验：$H_0 : \beta_2 = \beta_2^*$ versus $H_1 : \beta_2 \neq \beta_2^*$，其中 β_2^* 是一个已知的向量。

(1) 在 H_0 假设下, 回归模型变为 $Y - X_2\beta_2^* = X_1\beta_1 + e$。假设 $\tilde{\beta}_1$ 是 β_1 在 H_0 限制下的最小二乘估计量, 证明 $X_1\tilde{\beta}_1 = P_{X_1}(Y - X_2\beta_2^*)$, 其中 $P_{X_1} = X_1(X_1^T X_1)^{-1} X_1^T$ 是 X_1 的列空间的投影矩阵。

(2) 定义 $Z_2 = (I_n - P_{X_1})X_2$, 证明 Z_2 的列空间为 $Col(X_1)^{\perp} \cap Col(X)$, 这一结果说明 $P_X = P_{X_1} + P_{Z_2}$。

(3) 假设 $\hat{\beta} = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix}$ 是 β 没有 H_0 限制时的最小二乘估计量。定义 $\hat{Y} = X_1\hat{\beta}_1 + X_2\hat{\beta}_2$，并且 $\tilde{Y} = X_1\tilde{\beta}_1 + X_2\beta_2^*$. 定义 $R_1^2 = \|Y - \tilde{Y}\|^2$, $R_0^2 = \|Y - \hat{Y}\|^2$。证明 $R_1^2 - R_0^2 = \|\hat{Y} - \tilde{Y}\|^2$。

(4) 证明 $F = \frac{R_1^2 - R_0^2}{R_0^2} * \frac{n-p}{s}$ 在 H_0 假设下有着自由度为 $(s, n-p)$ 的 F 分布，并且利用这一结论给出 $H_0 : \beta_2 = \beta_2^*$ vs $H_1 : \beta_2 \neq \beta_2^*$ 的一个水平为 α 的检验。

7. (14 points) Suppose $Y = X\beta + e$, where X is a $n \times p$ matrix (rank is p), β is an unknown parameter vector of length p, $e = (e_1, \dots, e_n)'$, and $e_1, \dots, e_n$ are independent identical distributed as $N(0, \sigma^2)$ (σ^2 unknown) ($n > p \geq 2$). Write $X = (X_1, X_2)$ and $\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}$ where X_1 is a $n \times (p-s)$ matrix and X_2 is a $n \times s$ matrix, $\beta_1 \in R^{p-s}$ and $\beta_2 \in R^s$. Suppose the hypothesis test of interest is $H_0 : \beta_2 = \beta_2^*$ versus $H_1 : \beta_2 \neq \beta_2^*$, where β_2^* is known.

(1) Under H_0, the restricted model becomes $Y - X_2\beta_2^* = X_1\beta_1 + e$. Suppose $\tilde{\beta}_1$ is the least square estimator of β_1 under H_0, show that $X_1\tilde{\beta}_1 = P_{X_1}(Y - X_2\beta_2^*)$, where $P_{X_1} = X_1(X_1^T X_1)^{-1}X_1^T$ is the projection matrix onto the column space of X_1.

(2) Define $Z_2 = (I_n - P_{X_1})X_2$, show that the column space of Z_2 is $Col(X_1)^\perp \cap Col(X)$, which means that $P_X = P_{X_1} + P_{Z_2}$.

(3) Denote $\hat{\beta} = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix}$ as the least square estimator of β without the H_0 constrian. Denote $\hat{Y} = X_1\hat{\beta}_1 + X_2\hat{\beta}_2$ and $\tilde{Y} = X_1\tilde{\beta}_1 + X_2\beta_2^*$. Define $R_1^2 = \|Y - \tilde{Y}\|^2$, $R_0^2 = \|Y - \hat{Y}\|^2$. Show that $R_1^2 - R_0^2 = \|\hat{Y} - \tilde{Y}\|^2$.

(4) Prove that $F = \frac{R_1^2 - R_0^2}{R_0^2} * \frac{n-p}{s}$ follows an F distribution with degree of freedom $(s, n-p)$ under H_0 and use this result to derive a size α test for $H_0 : \beta_2 = \beta_2^*$ versus $H_1 : \beta_2 \neq \beta_2^*$.