北京大学数学科学学院期末试题

2025–2026 年第一学期

考试科目 数理统计　　考试时间 2026 年 1 月 5 日

姓名________　　学号________

本试题共 6 道大题，满分 100 分。

1.(12 points) 设 $X_1, \ldots, X_n$ i.i.d, 分布为 $N(\mu, \sigma^2)$, 其中 σ^2 已知。

- 利用检验反转法构造 μ 的一个 $1-\alpha$ 置信区间。
- 如果 σ^2 未知，利用枢轴量构造 μ 的一个 $1-\alpha$ 置信区间。

1.(12 points) Suppose $X_1, \ldots, X_n$ i.i.d, $X_1 \sim N(\mu, \sigma^2)$, where σ^2 is known.

- Construct a $1-\alpha$ confidence interval of μ by inverting a hypothesis test.
- If σ^2 is unknown, construct a $1-\alpha$ confidence interval of μ using the pivotal quantity.

2.(28 points) 设 $X_1, \ldots, X_n$ i.i.d., 分布为 Bernoulli(p).

- 给出 p 的一个近似的 $1-\alpha$ 置信区间 (例如 Wald type 区间)。
- 考虑 square error loss, p 的先验分布为 Beta 分布，即 $p \sim \text{Beta}(\alpha, \beta)$, 其密度函数为 $f(p) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}; p \in (0,1)$, 求 p 的 Bayes estimator. 注：$\text{Beta}(\alpha, \beta)$ 的期望为 $\frac{\alpha}{\alpha+\beta}$
- 计算 Bayes estimator 的风险.
- 求 p 的最小最大估计量。

2.(28 points) Suppose $X_1, \ldots, X_n$ i.i.d. and $X_1 \sim Bernoulli(p)$.

- Construct an approximate $1-\alpha$ confidence interval for p (e.g., the Wald type interval).
- Consider the square error loss and p has a prior of Beta distribution, $p \sim \text{Beta}(\alpha, \beta)$ with distribution function $f(p) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}$. Find the Bayes estimator of p. Notice: If $X \sim \text{Beta}(\alpha, \beta)$, then $EX = \frac{\alpha}{\alpha+\beta}$.

- Compute the risk of the Bayes estimator.

- Find the minimax estimator of p.

3.(20 points) 设 ANOVA 模型 $Y_{ij} = \mu + \tau_i + \epsilon_{ij}$, $i = 1, 2, ..., I$, $j = 1, ..., n_i$, 其中 $\sum_{i=1}^{I} \tau_i = 0$, 且 ϵ_{ij} 独立, $E(\epsilon_{ij}) = 0$, $Var(\epsilon_{ij}) = \sigma^2$。定义

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

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

$$SS_B = \sum_{i=1}^{I} n_i (\overline{Y}_{i.} - \overline{Y}_{..})^2.$$

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

- 证明 $SS_{TOT} = SS_W + SS_B$。

- 若 $n_i = J$, 证明 $E(SS_W) = I(J-1)\sigma^2$ 以及 $E(SS_B) = J \sum_{i=1}^{I} \tau_i^2 + (I-1)\sigma^2$。

3.(20 points) Suppose an ANOVA model: $Y_{ij} = \mu + \tau_i + \epsilon_{ij}$, $i = 1, 2, ..., I$ and $j = 1, ..., n_i$, where $\sum_{i=1}^{I} \tau_i = 0$. The ϵ_{ij} are independent and $E(\epsilon_{ij}) = 0$, $Var(\epsilon_{ij}) = \sigma^2$. Define

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

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

$$SS_B = \sum_{i=1}^{I} n_i (\overline{Y}_{i.} - \overline{Y}_{..})^2.$$

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

- Show that $SS_{TOT} = SS_W + SS_B$.

- If $n_i = J$ for $i = 1, 2, ..., I$. Prove that $E(SS_W) = I(J-1)\sigma^2$ and $E(SS_B) = J \sum_{i=1}^{I} \tau_i^2 + (I-1)\sigma^2$.

4.(10 points) 对于 ANOVA 模型

$$Y_{ij} = \theta_i + \varepsilon_{ij}, i = 1, \ldots, I, j = 1, \ldots, J,$$

其中 ε_{ij} 独立同分布, $\varepsilon_{ij} \sim N(0, \sigma^2)$, σ **未知**。定义

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

请写出对 $H_0 : \theta_1 = \theta_2 = ... = \theta_I$ vs $H_1 : \theta_i \neq \theta_j$ for some $i \neq j$ 的真实水平为 α 的假设检验，并且给出确定临界值的方法 (例如：使用 F 检验或似然比检验)。

4.(10 points) For the ANOVA model

$$Y_{ij} = \theta_i + \varepsilon_{ij}, i = 1, \dots, I, j = 1, \dots, J,$$

where ε_{ij} are i.i.d distributed, $\varepsilon_{ij} \sim N(0, \sigma^2)$, and σ **is unknown**. Define

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

Please write the size α hypothesis testing $H_0 : \theta_1 = \theta_2 = ... = \theta_I$ vs $H_1 : \theta_i \neq \theta_j$ for some $i \neq j$, and describe explicitly how to determine the critical value (e.g., use F test or likelihood ratio test).

5.(21 points) 对于简单线性回归模型 $y_i = \beta x_i + \varepsilon_i$，其中 ε_i 相互独立且都服从 $N(0, \sigma^2)$，σ^2 **未知**。现有 n 个观测数据 (x_i, y_i)，$i = 1, \dots, n$.

- 写出最小二乘估计 $\hat{\beta}$，以及 σ^2 的无偏估计。
- 求 β 的 $1 - \alpha$ 置信区间。
- 给定一个新观察 x_0，求 y_0 的 $1 - \alpha$ 预测区间。

5.(21 points) For simple linear model $y_i = \beta x_i + \varepsilon_i$, ε_i i.i.d and $\varepsilon_i \sim N(0, \sigma^2)$, σ^2 **is unknown**. Suppose we have n observations (x_i, y_i), $i = 1, \dots, n$.

- Write the least squares estimate $\hat{\beta}$, and find an unbiased estimate of σ^2.
- Construct a $1 - \alpha$ confidence interval for β.
- For a new observation x_0, construct a $1 - \alpha$ prediction interval for y_0.

6.(9 points) 假设 n 个观测数据 (x_i, y_i)，$i = 1, \ldots, n$ 满足简单线性回归模型 $y_i = \beta x_i + \beta_0 + \varepsilon_i$，其中 ε_i 相互独立且都服从 $N(0, \sigma^2)$。我们可以通过画残差图和散点图判断其是否服从假设以及是否存在异常数据。假设根据数据绘制了图 A(残差图)，B(残差图) 和 C(散点图)，它们分别反映了数据的什么问题？解释你的答案。

6.(9 points) Suppose n observations (x_i, y_i), $i = 1, \ldots, n$ follow simple linear model $y_i = \beta x_i + \beta_0 + \varepsilon_i$, where ε_i i.i.d and $\varepsilon_i \sim N(0, \sigma^2)$. We can assess whether the assumptions are met and whether there are any abnormal data by plotting residual plots and scatter plots. Suppose Figures A (residual plot), B (residual plot), and C (scatter plot) are generated based on the data. What problems do they each reveal about the data? Explain your answer.

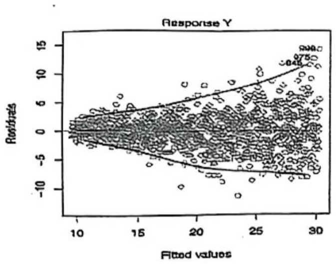

图 1: A

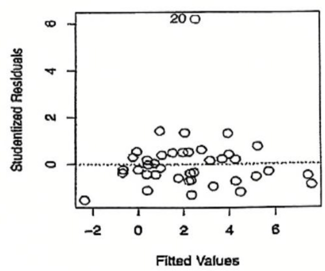

图 2: B

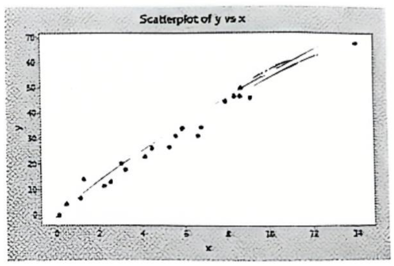

图 3: C