812

Department of Statistics
University of Wisconsin, Madison
PhD Qualifying Exam Option B
Tuesday, August 30, 2022
12:30-4:30pm, Room 331 SMI

- There are a total of FOUR (4) problems in this exam. Please do all FOUR (4) problems.
- Each problem must be done in a separate exam book.
- Please turn in FOUR (4) exam books.
- Please write your code name and **NOT** your real name on each exam book.

1. Read the entire question carefully before starting.

Definition. Let $\{V_n\}$ be a sequence of estimators of θ. Suppose for all θ there exists $\alpha > 0$ such that

$$n^{\alpha}(V_n - \theta) \xrightarrow{d} Y_\theta,$$

where Y_θ is some random variable, and $\xrightarrow{d}$ stands for convergence in distribution. The *asymptotic mean square error* is the quantity:

$$\text{AMSE}_\theta(V_n) = \mathbb{E}\left[\left(\frac{Y_\theta}{n^\alpha}\right)^2\right].$$

Let $X_1, \ldots, X_n$ be a random sample from a distribution with probability density function

$$f(x;\theta) = \frac{2x}{\theta^2}, \quad 0 < x < \theta,$$

where $\theta > 0$ is an unknown parameter of interest. Solve parts (a) through (d):

(a) Let $X_{(n)} = \max_i X_i$. Find the value of $\alpha > 0$ such that:

$$n^{\alpha}\left(\theta - X_{(n)}\right) \xrightarrow{d} Y,$$

where Y is an exponential random variable with mean being $\theta/2$.

HINT: Derive an expression for $\mathbb{P}\left[n^{\alpha}\left(\theta - X_{(n)}\right) \le x\right]$ and find its limit as $n \to \infty$.

(b) Let $\widehat{\theta} = \frac{2n+1}{2n} X_{(n)}$. Prove $\widehat{\theta}$ is the uniformly minimum variance unbiased estimator of θ.

(c) Show that

$$n^{\alpha}\left(\theta - \widehat{\theta}\right) \xrightarrow{d} Y - \theta/2,$$

where Y is the same as in (a).

(d) Calculate $\frac{\text{AMSE}_\theta(X_{(n)})}{\text{AMSE}_\theta(\widehat{\theta})}$. Provide a meaningful interpretation of this quantity.

REMINDER: let a be a constant. As $n \to \infty$, we have $\left(1 + \frac{a}{n}\right)^n \to e^a$.

2. This question consists of two parts.

   (a) Assume that a random variable X has the probability density function (p.d.f.),

$$f(x;\lambda) = 2\phi(x)\Phi(\lambda x), \quad x \in \mathbb{R}, \tag{1}$$

   with a parameter $\lambda \in \mathbb{R}$, where $\phi(x)$ and $\Phi(x)$ denote the p.d.f. and cumulative distribution function of the standard Gaussian distribution. Find the p.d.f. of $Y = X^2$.

   (b) Assume that (X_1, X_2) has a joint bivariate p.d.f.,

$$f(x_1, x_2; \alpha) = 2\phi(x_1)\phi(x_2)\Phi(\alpha x_1 x_2), \quad x_1, x_2 \in \mathbb{R}, \tag{2}$$

   with a parameter $\alpha \in \mathbb{R}$.

      i. Find the marginal p.d.f. of X_1.
      (Hint: You may use without proving $\mathrm{E}\{\Phi(hZ)\} = \Phi(0) = 1/2$ *for a constant* h *and a standard Gaussian variable* Z*.)*

      ii. Discuss whether X_1 and X_2 are independent or not.

      iii. Consider n independent pairs $\{(X_{i,1}, X_{i,2})\}_{i=1}^n$ following the distribution in (2), and we wish to test for the null hypothesis $H_0 : \alpha = 0$. Consider the sample correlation coefficient $\widehat{\rho} = \frac{\sum_{i=1}^n (X_{i,1}-\overline{X}_{\cdot,1})(X_{i,2}-\overline{X}_{\cdot,2})}{\sqrt{\sum_{i=1}^n (X_{i,1}-\overline{X}_{\cdot,1})^2}\sqrt{\sum_{i=1}^n (X_{i,2}-\overline{X}_{\cdot,2})^2}}$, where $\overline{X}_{\cdot,1} = n^{-1}\sum_{i=1}^n X_{i,1}$ and $\overline{X}_{\cdot,2} = n^{-1}\sum_{i=1}^n X_{i,2}$. Derive the asymptotic distribution of $\widehat{\rho}$, under the null H_0, as $n \to \infty$.

3. A company supplies a customer with a large number of batches of raw materials. The customer makes two sample determinations from each of five *randomly selected* batches to control the quality of the incoming material. The data are shown below.

Batch				
B1	B2	B3	B4	B5
74	68	75	74	81
76	72	77	74	79

We consider the random-effects model

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij}, \quad \text{for all } i = 1, \ldots, 5 \text{ and } j = 1, 2, \tag{3}$$

where both τ_i and ε_{ij} are *random variables* following the assumptions

$$\tau_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma_b^2), \quad \varepsilon_{ij} \overset{\text{i.i.d.}}{\sim} N(0, \sigma_e^2), \quad \text{and} \quad \tau_i \perp \varepsilon_{ij}, \tag{4}$$

for all $i = 1, \ldots, 5$ and $j = 1, 2$. Here, i.i.d. means "independent and identically distributed", and $\perp$ means "independently distributed".

For notational convenience, we may also consider reparameterization

$$\sigma_t^2 := \sigma_b^2 + \sigma_e^2, \quad \text{and} \quad \eta := \frac{\sigma_b^2}{\sigma_b^2 + \sigma_e^2},$$

where σ_t^2 represents the total variation, and η represents the proportion of the total variation attributable to batches.

(i) Assume η is known but σ_t^2 is unknown. For parts (b)-(c), please express your answer in terms of (Y_{ij}, η); there is no need to plug in the data.

  (a) Compute $\mathbb{E}(Y_{ij})$ and $\text{Cov}(Y_{ij}, Y_{i',j'})$ under model assumptions (3)-(4).

  (b) Derive the unbiased estimates $(\widehat{\mu}, \widehat{\sigma}_b^2, \widehat{\sigma}_e^2)$.

  (c) Perform the following hypothesis testing

$$H_0 : \mu = 78, \quad \text{vs.} \quad H_a : \mu \neq 78.$$

  Describe the test statistic, null distribution, degree of freedom, and p-value.
  Hint: It is helpful to first calculate $\text{Var}(\widehat{\mu})$ under model assumptions (3)-(4).

(ii) Assume both η and σ_t^2 are unknown. Please plug in the data in this part.

  (a) Complete the following one-way ANOVA table. Here, SS denotes the sums of squares, df denotes the degree of freedom, and MS denotes the mean squares. The entries to be filled in are marked by "?".

ANOVA Table

Source	SS	df	MS	F	p-value
Batch	?	?	?	9.29	0.016
Error	?	?	?	-	-
Total	?	?	-	-	-

(b) Let MS_b denote the MS for batch effect, and MS_e denote the MS for measurement error. Prove that

$$\mathbb{E}(\text{MS}_b) = 2\sigma_b^2 + \sigma_e^2, \quad \text{and} \quad \mathbb{E}(\text{MS}_e) = \sigma_e^2.$$

(c) Find the estimates $(\widehat{\sigma}_e^2, \widehat{\sigma}_b^2)$ using the method of moment and your results in previous two parts.

(d) The customer is concerned about significant variation among batches. Please perform the hypothesis testing

$$H_0 : \sigma_b^2 = 0, \quad \text{vs.} \quad H_a : \sigma_b^2 \neq 0.$$

Describe the test statistics, null distribution, degree of freedom, p-value, and your conclusion.

4. In an "Introductory Experimental Psychology" class, the students ran the following experiment to assess the effects of visualization and grouping on the memory. Twelve subjects were randomly divided into two groups of six, and each group was given a set of cards with words on them. While the words for both groups were the same, cards for group #1 had only words written on them where as the cards for group # 2 also had a picture depicting the word. Then, the cards were grouped into two as cards for related words (related) and cards for unrelated words (unrelated). Each subject was given two sets of cards to remember (related and unrelated) in random order. The ability to remember was recorded as the response.

		Words	
	Subject	Unrelated	Related
Without pictures (woP)	1	10	18
	2	14	19
	3	17	18
	4	8	12
	5	12	14
	6	15	20
With pictures (wP)	1	16	35
	2	19	32
	3	22	37
	4	20	33
	5	24	39
	6	21	32

Below is a snapshot of the file used in the analysis.

```
picture subject related y
1 1 1 10
1 1 2 18
1 2 1 14
1 2 2 19
1 3 1 17
1 3 2 18
........
2 5 1 24
2 5 2 39
2 6 1 21
2 6 2 32
```

The data is analyzed using the following R code.

```
Memory <- read.table("Memory.txt", header=T)
picture <- factor( Memory$picture )
subject <- factor( Memory$subject)
related <- factor( Memory$related )
y <- Memory$y
contrasts(picture) <- contr.sum
contrasts(related) <- contr.sum
contrasts(subject) <- contr.sum

result <- lmer( y ~ picture*related + (1|picture:subject))

summary(result)
Linear mixed model fit by REML ['lmerMod']
Formula: y ~ picture * related + (1 | picture:subject)

REML criterion at convergence: 108.7

Scaled residuals:
Min      1Q  Median      3Q     Max
-1.5121 -0.5606  0.1143  0.4739  1.1989

Random effects:
Groups           Name        Variance Std.Dev.
picture:subject  (Intercept) 5.733    2.394
Residual                     3.408    1.846
Number of obs: 24, groups:  picture:subject, 12

Fixed effects:
                    Estimate Std. Error t value
(Intercept)          21.1250     0.7873  26.833
picture1             -6.3750     0.7873  -8.098
related1             -4.6250     0.3768 -12.273
picture1:related1     2.5417     0.3768   6.745

Correlation of Fixed Effects:
             (Intr) pictr1 reltd1
picture1     0.000
related1     0.000  0.000
pctr1:rltd1 0.000  0.000  0.000


################################

logLik( lm(y ~ picture*related))
'log Lik.' -58.42078 (df=5)
```

```
logLik(result)
'log Lik.' -54.36441 (df=6)
pchisq(4.05637, df=1, lower.tail=FALSE)
[1] 0.044005

################################

result1 <- lmer( y ~ picture + related + (1|picture:subject))

################################

anova(result1, result)
refitting model(s) with ML (instead of REML)
Data: NULL
Models:
result1: y ~ picture + related + (1 | picture:subject)
result: y ~ picture * related + (1 | picture:subject)
        npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
result1    5 141.58 147.47 -65.790   131.58
result     6 122.84 129.91 -55.422   110.84 20.736  1  5.272e-06 ***

################################

result2 <- lmer( y ~ picture + (1|picture:subject))

################################

anova(result2, result1)
refitting model(s) with ML (instead of REML)
Data: NULL
Models:
result2: y ~ picture + (1 | picture:subject)
result1: y ~ picture + related + (1 | picture:subject)
        npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
result2    4 161.76 166.47 -76.878   153.76
result1    5 141.58 147.47 -65.790   131.58 22.177  1  2.487e-06 ***
```

(a) Write down the model fitted in the `result` object in the above R code. Specify all the assumptions and the estimated parameters.

(b) Is there significant variability in the ability to remember among the subjects? Justify your answer by formulating a test and reporting the result of the test.

(c) Do you think the full model with the interaction term can be reduced to a simpler model based on the analysis above? If so, what would be the simpler model.

(d) What is the estimated average ability to remember a set of unrelated words without their pictures?

(e) Is there any evidence that the average ability to remember is different for unrelated and related sets of words? Justify your answer.