Department of Statistics
University of Wisconsin-Madison
PhD Qualifying Exam Option B
August 27, 2024
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.

Read all parts of each question carefully before starting.

1. Let R be a binary random variable with $p = P(R = 1)$ and $1 - p = P(R = 0)$. Let X be another random variable such that

$$\text{given } R = 1,\ X \text{ has the normal density } \frac{1}{\sqrt{2\pi}\theta}\exp(-\frac{x^2}{2\theta^2}),$$
$$\text{given } R = 0,\ X \text{ has the double exponential density } \frac{1}{2\theta}\exp(-\frac{|x|}{\theta}),$$

where $\theta > 0$. Suppose that θ is unknown, p is known, and $0 < p < 1$. Let (X_i, R_i), $i = 1, ..., n$, be a random sample from the population of (X, R). Define

$$A = \frac{1}{n}\sum_{i=1}^{n} R_i X_i^2 \qquad \text{and} \qquad B = \frac{1}{n}\sum_{i=1}^{n}(1 - R_i)|X_i|.$$

(a) (i) Obtain the likelihood function of θ given observed (X_i, R_i), $i = 1, ..., n$.
(ii) Show that (A, B) is a minimal sufficient statistic for θ.
(iii) Show that (A, B) is not complete.

(b) Obtain the maximum likelihood estimator (MLE) of θ.

(c) Using the law of large numbers, show that the MLE of θ derived in (b) is consistent as $n \to \infty$.

(d) Obtain the non-degenerate asymptotic distribution of the MLE of θ in (b), as $n \to \infty$.

(e) Show that asymptotically as $n \to \infty$, $\sqrt{A/p}$ is a normally distributed estimator of θ and obtain its asymptotic relative efficiency with respect to the MLE of θ in (b). Which estimator is asymptotically more efficient?

(f) Show that the MLE of θ in (b) divided by θ is a pivotal quantity, and discuss how to obtain a confidence interval for θ with confidence coefficient $1 - \alpha$, using this pivotal quantity, where $\alpha \in (0, 1)$ is a known constant.

(g) Consider testing $H_0 : \theta = 1$ versus $H_1 : \theta = 2$. Obtain the rejection region of uniformly most powerful test with size $\alpha \in (0, 1)$ in terms of A and B and a known constant c. Discuss how to determine the value of c (you do not need to get an explicit form of c).

2. Let $X_1, \ldots, X_n$ ($n \geq 3$) be a random sample from a Log-Normal(μ, σ) population, with probability density function (pdf) given by

$$f(x; \mu, \sigma) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left\{-\frac{(\log x - \mu)^2}{2\sigma^2}\right\}, \quad x > 0, \quad -\infty < \mu < \infty, \quad 0 < \sigma < \infty.$$

Assume μ and σ are unknown. Of interest is the population mean $\theta = E(X_1)$.

(a) Find $\widehat{\theta}$, the maximum likelihood estimator (MLE) of θ. You may use well-known formulas for $\widehat{\mu}$ and $\widehat{\sigma}$, the MLEs of μ and σ respectively, without proving them.

(b) Find the non-degenerate asymptotic distribution of the MLE $\widehat{\theta}$.

(c) Calculate the asymptotic relative efficiency between $\widehat{\theta}$ and $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Which estimator is more efficient asymptotically?

(d) Let α be a given constant between 0 and 1.

(i) Obtain an approximate $100(1-\alpha)\%$ Wald interval for θ.

(ii) Construct another approximate $100(1-\alpha)\%$ confidence interval for θ by constructing a Wald interval for $\log\theta$ and taking an appropriate transformation.

(iii) Show that the ratio of the lengths of the confidence intervals you derived in parts d(i) and d(ii) converges to 1 in probability as $n \to \infty$.

(e) Consider the uniformly minimum variance unbiased estimator (UMVUE) of θ.

(i) Let $U_k = \frac{Z}{\sqrt{Z^2+\chi_k^2}}$, where Z and χ_k^2 are independent random variables, such that Z is standard normal and χ_k^2 follows a chi-square distribution on $k > 0$ degrees of freedom. Find the pdf of U_k. *Hint:* What's the distribution of U_k^2?

(ii) Show that the UMVUE of θ can be expressed as

$$\widetilde{\theta} = \exp(\widehat{\mu})\, M_{U_{n-2}}\left(\sqrt{n-1}\,\widehat{\sigma}\right),$$

where $M_{U_k}(t) = E[\exp(t\,U_k)]$ is the moment generating function of U_k. *Hint:* Let $Y_1 = \log X_1$, and write $Y_1 = \widehat{\mu} + \widehat{\sigma}\left(\frac{Y_1-\widehat{\mu}}{\widehat{\sigma}}\right)$.

Useful facts. You may use the following facts without proving them:

(a) The moment generating function of $Z \sim$ N(0,1) is given by $M_Z(t) = e^{t^2/2}$.

(b) The Gamma(α, β) distribution has pdf

$$f(x; \alpha, \beta) = \frac{1}{\Gamma(\alpha)\beta^\alpha}\, x^{\alpha-1}\, e^{-\frac{x}{\beta}}, \quad x > 0, \quad \alpha > 0, \quad \beta > 0.$$

(i) Suppose $X \sim$ Gamma(α, β). Then $E(X) = \alpha\beta$ and $\text{var}(X) = \alpha\beta^2$.
(ii) The chi-square distribution on $k > 0$ degrees of freedom is a special case of the Gamma(α, β) distribution, with $\alpha = k/2$ and $\beta = 2$.

(c) The Beta(α, β) distribution has pdf

$$f(x; \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\, x^{\alpha-1}\,(1-x)^{\beta-1}, \quad 0 < x < 1, \quad \alpha > 0, \quad \beta > 0.$$

Let $G_\alpha \sim$ Gamma(α, 1) and $G_\beta \sim$ Gamma(β, 1) be independent. Then $\frac{G_\alpha}{G_\alpha+G_\beta} \sim$ Beta(α, β).

3. Consider the linear model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon, \tag{1}$$

where x_1, x_2, x_3 are deterministic design points, ϵ is distributed as $N(0, \sigma^2)$, and β_0, β_1, β_2, β_3, and σ^2 are unknown parameters. Suppose we independently collect n observations $(Y_i, x_{i1}, x_{i2}, x_{i3})$, $i = 1, \ldots, n$, from linear model (1). Assume also that

$$\mathbf{X}^\top \mathbf{X}/n = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & r & 0 \\ 0 & r & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{2}$$

with $r \in (-1, 1)$, where $\mathbf{X}$ is the $n \times 4$ matrix whose ith row is $(1, x_{i1}, x_{i2}, x_{i3})$, and $\mathbf{X}^\top$ is the transpose of $\mathbf{X}$.

Hint: In the following problems, you may find the following formula on matrix inverse useful: When $ad - bc \neq 0$,

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \frac{1}{ad - bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}.$$

(a) We aim to estimate the contrast $\eta = \cos(\phi)\beta_1 + \sin(\phi)\beta_2$, for a fixed, given constant $\phi \in [0, 2\pi]$.

(i) Specify the standard deviation of the maximum likelihood estimator of the contrast η.

(ii) If $\phi = 0$, how large does the sample size n need to be so that the standard deviation in part (i) is no larger than 0.2?

(iii) For which values of ϕ will the standard deviation in part (i) be maximized and minimized, respectively?

(b) Suppose Researcher A fitted linear model (1) and obtained R outcome below.

```
lm(formula = Y ~ X)

Residuals:
Min        1Q    Median        3Q       Max
-2.58201 -0.52738  0.00281  0.65058  1.82364

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.0516     0.0998  10.537  < 2e-16 ***
X1            0.9598     0.1152   8.328 5.71e-13 ***
X2            0.9538     0.1152   8.277 7.35e-13 ***
X3            1.1084     0.0998  11.106  < 2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

```
Residual standard error: 0.998 on 96 degrees of freedom
Multiple R-squared:  0.8061,Adjusted R-squared:    0.8
F-statistic:   133 on 3 and 96 DF,  p-value: < 2.2e-16
```

Researcher A wants to examine whether $\beta_1 = \beta_2$ or not. Write down a null hypothesis explicitly, construct a test statistic and provide its distribution, and provide numerical values of all the related quantities based on the R outcome above.

(c) Researcher B has devised a new way to reduce correlation such that r becomes $r/2$ in the matrix (2). Let $\text{Length}_{\text{Old}}$ be the length of 95% symmetric two-sided confidence interval for η with $\mathbf{X}^\top\mathbf{X}$ given by (2), and let $\text{Length}_{\text{New}}$ be the length of 95% symmetric two-sided confidence interval for η with r in (2) changed to $r/2$. Obtain the ratio $(\text{Length}_{\text{Old}} - \text{Length}_{\text{New}})/\text{Length}_{\text{Old}}$. Which confidence interval would you prefer and why?

(d) Researcher C did not know the true data-generation model (1) and mistakenly fitted the following wrong model

$$Y = \gamma_0 + \gamma_1 x_1 + \epsilon,$$

using independent (Y_i, x_{i1}), $i = 1, ..., n$. Let $\hat{\gamma}_0$ and $\hat{\gamma}_1$ denote the ordinary least square estimators for γ_0 and γ_1, respectively, under the wrong model.

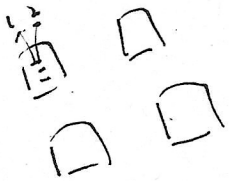
(i) Provide an expression for the bias of $\hat{\gamma}_1$ for estimating β_1 under model (1).

(ii) Derive a condition under which the squared bias of $\hat{\gamma}_1$ in part (i) is no more than 1/4 of the variance of the MLE of β_1 under model (1). Express your answer in terms of a condition on the "signal size" $\sqrt{n}\beta_2/\sigma$.

(iii) Suppose there is a new observation Y^* generated under the true model (1), i.e.,

$$Y^* = \beta_0 + \beta_1 x_1^* + \beta_2 x_2^* + \beta_3 x_3^* + \epsilon^*,$$

where ϵ^* follows $N(0, \sigma^2)$ and is independent from Y_i, $i = 1, ..., n$. Researcher C didn't observe Y^* and wanted to predict it as $\hat{Y}^* = \hat{\gamma}_0 + \hat{\gamma}_1 x_1^*$. Derive the expectation of the squared prediction error $(Y^* - \hat{Y}^*)^2$.



4. *A study of of vitamin C values in bell peppers.* Four ($a = 4$) plants were randomly selected, three peppers ($b = 3$) were randomly taken from each plant, and two ($n = 2$) random 100-mg samples were taken from each pepper. The measured response for each sample is vitamin C concentration.

   The data was analyzed with the following code where plant is denoted as factor A ($i = 1, 2, 3, 4$) and pepper is denoted as factor B ($j = 1, 2, 3$).

```
pepperD <- read.table("pepper.txt", header=T)

plant <- factor(pepperD$plant)
contrasts(plant) <- contr.sum
pepper <- factor(pepperD$pepper)
contrasts(pepper) <- contr.sum
vitaminC <- pepperD$vitaminC

result <- lm( vitaminC ~ plant/pepper )
```

   1|plant + 1|plant:pepper.

   The corresponding ANOVA table is given by:

Source	df	E[Mean Squares]
Plant (MS_A)	$a-1$	$\sigma^2_\epsilon + n\sigma^2_\beta + nb\sigma^2_\alpha$
Pepper(Plant) ($MS_{B(A)}$)	$a(b-1)$	$\sigma^2_\epsilon + n\sigma^2_\beta$
Error (MS_{Err})	$ab(n-1)$	σ^2_ϵ

   where σ_ϵ, σ^2_β, and σ^2_α denote the variance components due to error, pepper, and plant.

   (a) Specify the random effects model fitted in `result` and the underlying assumptions. Be as specific as possible.

   (b) The mean-squares associated with plant, pepper, and error terms are given as

```
> tmp <- anova(result)
> tmp$"Mean Sq"
[1] 2.520115278 0.328775000 0.006654167
```

   Find the Method-of-Moments estimates of the variance components for the model you specified in (a).

   (c) Test if there is a plant effect at significance level of 0.01. Specify your test statistic, its distribution, and the conclusion from the test.

   (d) Test if there is a pepper effect at significance level of 0.01. Specify your test statistic, its distribution, and the conclusion from the test.

   (e) Estimate the correlation between (i) measurements taken from the same pepper, and (ii) measurements taken from the same plant but different peppers.

   (f) The analyst decides to fit the same model (model fitted in `result` in the above R output) as:

```
result1 <- lmer(vitaminC ~ 1 + (1|plant) + (1|plant:pepper), REML = F)
```

   and then also considers an alternative model

```
result2 <- lmer(vitaminC ~ 1 + (1|plant:pepper), REML = F)

> summary(result1)
Linear mixed model fit by maximum likelihood  ['lmerMod']
Formula: vitaminC ~ 1 + (1 | plant) + (1 | plant:pepper)

AIC      BIC   logLik deviance df.resid
9.6     14.3    -0.8      1.6       20

Scaled residuals:
Min       1Q   Median       3Q      Max
-1.68307 -0.24393 -0.07596  0.35320  1.99461

Random effects:
Groups       Name        Variance Std.Dev.
plant:pepper (Intercept) 0.161060 0.40132
plant        (Intercept) 0.260220 0.51012
Residual                 0.006654 0.08157
Number of obs: 24, groups:  plant:pepper, 12; plant, 4

Fixed effects:
Estimate Std. Error t value
(Intercept)   3.0121     0.2806   10.73

> summary(result2)
Linear mixed model fit by maximum likelihood  ['lmerMod']
Formula: vitaminC ~ 1 + (1 | plant:pepper)

AIC      BIC   logLik deviance df.resid
12.0     15.5    -3.0      6.0       21

Scaled residuals:
Min       1Q   Median       3Q      Max
-1.78333 -0.25528 -0.04439  0.32166  1.89436

Random effects:
Groups       Name        Variance Std.Dev.
plant:pepper (Intercept) 0.421279 0.64906
Residual                 0.006654 0.08157
Number of obs: 24, groups:  plant:pepper, 12

Fixed effects:
Estimate Std. Error t value
(Intercept)   3.0121     0.1881   16.01
```

H_0:

Test if there is a pepper effect using a likelihood ratio test based on the above fits (result1 and result2) at significance level of 0.01. Discuss any potential discrepancy between the test based on these fits versus your result in part (d).