Department of Statistics
University of Wisconsin, Madison
PhD Qualifying Exam Option B
August 27, 2019
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. Consider a multinomial experiment in which there are n trials and four categories. The categories have probabilities θ, 2θ, 3θ and $1-6\theta$, where θ is an unknown parameter in $(0, 1/6)$. Let X_1, X_2, X_3, X_4 denote the counts for these categories. Note that (X_1, X_2, X_3, X_4) has the following multinomial distribution

$$(X_1, X_2, X_3, X_4) \sim M(\theta, 2\theta, 3\theta, 1-6\theta; n). \tag{1}$$

   (a) Show that the distributions of the data form a one-parameter exponential family and identify a minimal complete sufficient statistic.

   (b) Find the maximum likelihood estimator (MLE) $\widehat{\theta}$ of θ.

   (c) Find the Fisher information for θ.

   (d) Find the uniformly minimum-variance unbiased estimator (UMVUE) for θ^2.

   (e) Compute the asymptotic relative efficiency of the estimator $\widetilde{\theta} = X_1/n$ with respect to the MLE $\widehat{\theta}$ as $n \to \infty$.

   (f) Find the limiting distribution of $\sqrt{n}[\widehat{(\theta)}^2 - \theta^2]$.

   (g) Consider a Bayesian model for (X_1, X_2, X_3, X_4) in (1). Assume that $\eta = 6\theta$ has the following prior distribution: $\eta \sim \text{Beta}(\alpha, \beta)$. Derive the posterior distribution for η given (X_1, X_2, X_3, X_4).

Hints:

1 Consider a multinomial distribution of n trials and four categories, where the categories have probabilities θ_1, θ_2, θ_3 and θ_4 and X_1, X_2, X_3, X_4 denote the counts for these categories. This distribution has the following joint mass function:

$$\binom{n}{x_1, x_2, x_3, x_4} (\theta_1)^{x_1} (\theta_2)^{x_2} (\theta_3)^{x_3} (\theta_4)^{x_4}.$$

2 The Beta distribution Beta (α, β) with positive parameters α and β has the following pdf:

$$p(\theta) = (\theta)^{(\alpha-1)}(1-\theta)^{(\beta-1)}.$$

2. Suppose that $Y_1, \ldots, Y_n$ are independent observations following the model:

$$Y_i = \mu + \sigma\,\epsilon_i, \quad i = 1, \ldots, n,$$

with unknown parameters $\mu \in (-\infty, \infty)$ and $\sigma \in (0, \infty)$. The random error terms $\{\epsilon_1, \ldots, \epsilon_n\}$ are i.i.d. from a continuous uniform distribution on the interval $(-\sqrt{3}, \sqrt{3})$. Define

$$T_n = \frac{1}{n}\sum_{i=1}^{n} \frac{(Y_i - \mu)^3}{\sigma^3}, \quad \widetilde{T}_n = \frac{1}{n}\sum_{i=1}^{n} \frac{(Y_i - \widehat{\mu})^3}{\sigma^3}, \quad \widehat{T}_n = \frac{1}{n}\sum_{i=1}^{n} \frac{(Y_i - \widehat{\mu})^3}{\widehat{\sigma}^3},$$

where $\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ and $\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$.

(a) Derive a non-degenerate limiting distribution of $\sqrt{n}(T_n - a_n)$ for some suitable a_n as $n \to \infty$.

(b) Derive a non-degenerate limiting distribution of $\sqrt{n}(\widetilde{T}_n - b_n)$ for some suitable b_n as $n \to \infty$.

(c) Derive a non-degenerate limiting distribution of $\sqrt{n}(\widehat{T}_n - c_n)$ for some suitable c_n as $n \to \infty$.

(d) Compare the variances of the limiting distributions for T_n in part (a), $\widetilde{T}_n$ in part (b), and $\widehat{T}_n$ in part (c).

3. Consider a linear regression model with $p+1$ predictors (including intercept):

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon,$$

with unknown parameters $\beta_0, \beta_1, \cdots, \beta_p \in (-\infty, \infty)$ and $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ with unknown $\sigma^2 > 0$. Consider n independent observations from this model. Let $\mathbf{y} = (y_1, \ldots, y_n)^T$ denote the observed responses; let $\mathbf{X} = [\mathbf{x}_0, \ldots, \mathbf{x}_p]$ denote the $n \times (p+1)$ design matrix, where $\mathbf{x}_i$ corresponds to the $(i+1)^{\text{th}}$ column of $\mathbf{X}$; and let $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)^T$ denote the vector of regression coefficients. Assume that $\text{rank}(\mathbf{X}) = p+1$. Let $\widehat{\beta}_j$ denote the least squares estimate of β_j, and define $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_0, \ldots, \widehat{\beta}_p)^T$.

(a) Suppose $\mathbf{c} = (c_0, \ldots, c_p)^T$ is a vector of known constants. Consider testing the hypothesis:

$$\mathcal{H}_0 : \sum_{j=0}^{p} c_i \beta_j = h \quad \text{vs.} \quad \mathcal{H}_1 : \sum_{j=0}^{p} c_j \beta_j \neq h,$$

where h is a given constant. Derive a suitable test statistic for testing this hypothesis. Find its distribution and specify the rejection region at significance level α.

(b) Suppose we wish to predict the value of a future observation Y_{n+1}. Let $\mathbf{z} = (1, z_1, \ldots, z_p)^T$ denote its corresponding vector of predictor variables (i.e., $X_i = z_i$, for $i = 1, \ldots, p$), and consider the prediction $\widehat{Y}_{n+1} = \mathbf{z}^T \widehat{\boldsymbol{\beta}}$. Find the distribution of $Y_{n+1} - \widehat{Y}_{n+1}$.

(c) Given the vector $\mathbf{z}^T$ of predictor variables for the future observation Y_{n+1}, find an interval I such that $\mathbb{P}(Y_{n+1} \in I) = 1 - \alpha$.

(d) Suppose an additional predictor variable X_{p+1} is added to the model to obtain

$$Y = \gamma_0 + \gamma_1 X_1 + \cdots + \gamma_{p+1} X_{p+1} + \varepsilon.$$

Suppose that the augmented design matrix $\widetilde{\mathbf{X}} = [\mathbf{X}, \mathbf{X}_{p+1}]$ has rank $p+2$. Let $\widehat{\mathbf{x}}_{p+1} = a_0 \mathbf{x}_0 + \cdots + a_p \mathbf{x}_p$ denote the least-squares projection of $\mathbf{x}_{p+1}$ onto the subspace of $\mathbb{R}^n$ spanned by the columns of $\mathbf{X}$. Find the residual vector $\mathbf{r}_{p+1} = \mathbf{x}_{p+1} - \widehat{\mathbf{x}}_{p+1}$ in terms of $\mathbf{X}$ and $\mathbf{x}_{p+1}$.

(e) Express the least squares estimates $\widehat{\gamma}_0, \ldots, \widehat{\gamma}_{p+1}$ in terms of only $\widehat{\beta}_0, \ldots, \widehat{\beta}_p, a_0, \ldots, a_p, \mathbf{r}_{p+1}$ and $\mathbf{y}$.

4. Genome editing technologies became one of the most important genetic tools in the implementation of pathogen resistance in plants. In a recent experiment, plant pathologists targeted a virulent bacterial pathogen Pseudomonas syringae susceptibility gene in Arabidopsis thaliana (*A. thaliana*) for mutation by the CRISPR-Cas9 genome editing system and produced Pseudomonas resistant plants. Then, they carried out an experiment in a green house to compare overall health of the two plant types that are labeled as wild type (WT) and mutant (MU). In the green house, one table with eight trays, and two pots in each tray were used to design the experiment. Two wild type plant seedlings were planted in one pot, and two mutant plant seedlings were planted in the other pot. Plant type assignments to two pots within each tray were carried out by tossing a fair coin. A quantitative measurement of overall plant health as measured by an aggregate score of overall plant height, numbers of leaves on the flowering stem, numbers of flowers, and seeds per flower were recorded for each plant after 6 weeks. Table 1 below presents these scores.

Table 1: Scores of plant healths after 6 weeks.

	Wild Type Plant Pot		Mutant Plant Pot	
Tray	Plant 1	Plant 2	Plant 1	Plant 2
1	6	5	3	4
2	7	8	6	7
3	8	9	4	5
4	4	6	5	5
5	8	8	7	7
6	5	7	2	4
7	10	9	9	7
8	5	7	1	4

This question has two main parts: Part I and Part II. Please answer both parts.

Part I.

Let i index the plant type ($i \in \{\text{WT, MU}\}$), j index trays ($j = 1, ..., 8$), and k index plants within pots ($k = 1, 2$). Let y_{ijk} denote the response corresponding to type i, tray j, and plant k.

Consider the following `R` output to answer parts **(a)-(f)**. The `R` output may contain more parts than you may need to utilize.

```
> score
[1]  7  8  6  7  8  9  4  5  8  8  7  7  5  7  2  4  5  6  3  4  9 10  7  9
 5  7  1  4  4  6  5  5
> plant.type
[1] WT WT MU MU WT WT MU MU WT WT MU MU WT WT MU MU WT WT MU MU WT WT MU MU
WT WT MU MU WT WT MU MU
```

```
Levels: WT MU
> pot
[1] 1 1 2 2 1 1 2 2 1 1 2 2 1 1 2 2 1 1 2 2 1 1 2 2 1 1 2 2 1 1 2 2
Levels: 1 2
> tray
[1] 1 1 1 1 2 2 2 2 3 3 3 3 4 4 4 4 5 5 5 5 6 6 6 6 7 7 7 7 8 8 8 8
Levels: 1 2 3 4 5 6 7 8

m1 <- lmer(score ~ plant.type + (1|tray) + (1|plant.type:tray), REML = FALSE)
> summary(m1)
Linear mixed model fit by maximum likelihood  ['lmerMod']
Formula: score ~ plant.type + (1 | tray) + (1 | plant.type:tray)

AIC      BIC   logLik deviance df.resid
125.2    132.5    -57.6    115.2       27

Scaled residuals:
Min      1Q  Median      3Q     Max
-2.0748 -0.6335  0.2087  0.5204  1.3972

Random effects:
Groups          Name        Variance Std.Dev.
plant.type:tray (Intercept) 0.2813   0.5303
tray            (Intercept) 2.0313   1.4252
Residual                    1.1250   1.0607
Number of obs: 32, groups:  plant.type:tray, 16; tray, 8

Fixed effects:
Estimate Std. Error t value
(Intercept)    7.0000     0.5995  11.677
plant.typeMU  -2.0000     0.4593  -4.355

Correlation of Fixed Effects:
(Intr)
plant.typMU -0.383
```

(a) Write down the model fitted in the `m1` object in mathematical notation by using the terms μ_{WT}, μ_{MT}, a_j, b_{ij}, and e_{ijk}. State the underlying assumptions of each term explicitly and specify the estimated values of the parameters in this model.

(b) Specify the experimental units and one reason for including the `(1|plant.type:tray)` term in the model.

(c) Let $\bar{y}_{ij.} = \sum_{k=1}^{2} y_{ijk}, \forall i, j$. Find the distribution of $\bar{y}_{WT1.} - \bar{y}_{MU1.}$.

(d) Find an unbiased estimator for the variance of $\bar{y}_{WT1.} - \bar{y}_{MU1.}$ and calculate its value.

(e) Provide a 95% confidence interval for $\mu_{WT} - \mu_{MU}$.

(f) Would you recommend any changes to this experimental design using the same resources (eight trays, two pots per tray, sixteen seedlings of wild type plant, and sixteen seedlings of mutant plant). Explain why or why not. Full credit requires supporting your argument with explicit calculations.

Part II.

A variation of this experiment was carried out by using a protective pesticide. Four of the trays (2, 4, 5, and 7) were randomly assigned to low pesticide level, and the remaining four trays (1, 3, 6, and 8) were assigned to high pesticide level. The pesticides were applied in the middle point of 3 weeks. Let $l \in \{L, H\}$ index pesticide levels. Let μ_{il} denote the expected value of the health score for plant type i and pesticide level l.

Consider the following additional `R` output to answer parts **(g)-(i)**. The `R` output may contain more parts than you may need to utilize.

```
> score
[1]  7  8  6  7  8  9  4  5  8  8  7  7  5  7  2  4  5  6  3  4  9 10  7  9
 5  7  1  4  4  6  5  5
> plant.type
[1] WT WT MU MU WT WT MU MU WT WT MU MU WT WT MU MU WT WT MU MU WT WT MU MU
WT WT MU MU WT WT MU MU
Levels: WT MU
> pot
[1] 1 1 2 2 1 1 2 2 1 1 2 2 1 1 2 2 1 1 2 2 1 1 2 2 1 1 2 2 1 1 2 2
Levels: 1 2
> tray
[1] 1 1 1 1 2 2 2 2 3 3 3 3 4 4 4 4 5 5 5 5 6 6 6 6 7 7 7 7 8 8 8 8
Levels: 1 2 3 4 5 6 7 8
> pesticide
[1] H H H H L L L L H H H H L L L L L L L L H H H H L L L L H H H H
Levels: L H

anova(lm(score~pesticide+tray+plant.type+pesticide:plant.type+tray:plant.type))
Analysis of Variance Table

Response: score
Df Sum Sq Mean Sq F value    Pr(>F)
pesticide            1 36.125  36.125 32.1111 3.504e-05 ***
tray                 6 42.375   7.062  6.2778  0.001520 **
plant.type           1 32.000  32.000 28.4444 6.725e-05 ***
pesticide:plant.type  1 10.125  10.125  9.0000  0.008479 **
tray:plant.type      6  3.375   0.563  0.5000  0.799299
Residuals           16 18.000   1.125
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

(g) Let $\bar{\mu}_{.l} = (\mu_{WTl} + \mu_{MUl})/2$ for $l = L, H$. Compute an F-statistic for testing $H_0 : \bar{\mu}_{.L} = \bar{\mu}_{.H}$ and specify the degrees of freedom associated with it.

(h) State the null hypothesis of no interaction between the factors plant type and pesticide level.

(i) Compute a t-statistic for testing the null hypothesis of no interaction between the factors plant type and pesticide level and specify its degrees of freedom.