Department of Statistics
University of Wisconsin, Madison
PhD Qualifying Exam Option B
August 31, 2021
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 the scaled uniform distribution, with probability density function (pdf) indexed by parameter $\theta > 0$ given below:

$$f(x;\theta) = \frac{1}{\theta}, \quad 0 < x < \theta < \infty.$$

Let $X_1, \ldots, X_n$ be a random sample from this distribution. Suppose we endow θ with a Pareto(a, b) prior, with prior pdf $\pi(\theta)$ given by:

$$\pi(\theta) = \frac{ab^a}{\theta^{a+1}}, \quad 0 < b < \theta < \infty, \quad a > 0.$$

(a) Show that the posterior distribution for θ is Pareto(a', b'), where $a' = a + n$ and $b' = \max\left(b, X_{(n)}\right)$. (*Note:* recall that $X_{(n)} = \max_{i=1,\ldots,n} X_i$.)

(b) Find $\delta(X_1, \ldots X_n)$, the Bayes estimator for θ with respect to

i. squared error loss: $L(\theta, \delta) = (\delta - \theta)^2$.

ii. absolute error loss: $L(\theta, \delta) = \left|\delta - \theta\right|$.

(c) Suppose we test the hypotheses $H_0 : \theta \geq \theta_0$ vs. $H_A : \theta < \theta_0$ for some $\theta_0 > 0$.

i. Consider the test that rejects the null hypothesis when the posterior probability of H_0 is less than $\alpha \in (0, 1)$. Show that the rejection region is given by $\mathcal{R} = \left\{b' < \theta_0 \alpha^{\frac{1}{a'}}\right\}$.

ii. Calculate the power function of the test in part (i) above; that is, calculate $\beta(\theta) = \mathbb{P}\left(\mathcal{R} \middle| \theta\right)$.

iii. Calculate $\beta_{UMP}(\theta)$, the power function of the uniformly most powerful (UMP) level α test of H_0 vs H_A.

iv. Let $a \to 0$ and $b \to 0$.

A. Show that $\beta(\theta) \to \beta_{UMP}(\theta)$ pointwise.

B. Briefly explain, in qualitative terms, why the performance of the Bayesian test is approaching the performance of a *frequentist test* (the UMP test is a frequentist test in the sense that it can be defined without reference to a prior distribution). Your answer will be scored based on the soundness of your explanation. *Hint:* think about what is happening to the prior distribution as a and b go to zero.

2. Let $X_1, X_2, \ldots, X_n$ be independent identically distributed random variables having the following probability density function:

$$f_{a,b,c,\theta_1,\theta_2}(x) = \begin{cases} ae^{-\frac{x}{\theta_1}+1}, & x > \theta_1; \\ b, & -\theta_2 < x \leq \theta_1; \\ ce^{\frac{x}{\theta_2}+1}, & x \leq -\theta_2 \end{cases} \tag{1}$$

where $a \geq 0$, $b \geq 0$, $c \geq 0$, $\theta_1 > 0$ and $\theta_2 > 0$ are unknown.

(a) Identify a general formula satisfied by a, b, c, θ_1, and θ_2 such that (1) is a density function.

(b) Explicitly express the first two moments $\mu_1 = E(X_1)$ and $\mu_2 = E(X_1^2)$ as a function of a, b, c, θ_1, and θ_2. For $\mu_3 = E(X_1^3)$ and $\mu_4 = E(X_1^4)$, the following formulas can be used in later parts if needed.

$$\begin{aligned} E(X_1^3) &= -\left(16c + \frac{b}{4}\right)\theta_2^4 + \left(16a + \frac{b}{4}\right)\theta_1^4, \\ E(X_1^4) &= \left(65c + \frac{b}{5}\right)\theta_2^5 + \left(65a + \frac{b}{5}\right)\theta_1^5. \end{aligned}$$

(c) Suppose θ_1 and θ_2 are now fixed, but $a > 0$, $b > 0$, and $c > 0$ are unknown.

i) Find a condition on a, b, and c such that $E(X_1)$ is positive.

ii) Assuming $\theta_1 = \theta_2 = 1$. Prove that there does not exist a set of a, b, and c such that $Var(X_1)$ is minimized.

(d) Suppose now $a = b = c$.

i) Obtain estimators $\widehat{\theta_1}$ of θ_1, $\widehat{\theta_2}$ of θ_2 respectively using the method of moments.

ii) Suppose $\theta = \theta_1 = \theta_2$. Obtain a method of moment estimator $\widehat{\theta}$ of θ. Obtain the asymptotic distribution of $\widehat{\theta}$.

3. Consider the linear regression model

$$Y_{ij} = \alpha_0 + \alpha_i + \varepsilon_{ij}, \quad \text{for } j = 1, 2, \ldots, n_i, \text{ and } i = 1, \ldots, m, \tag{2}$$

where α_i's are unknown parameters, $\varepsilon_{ij} \sim N(0, \sigma_i^2)$ are independent Gaussian noise with unknown variance σ_i^2, n_i is the sample size for group i, and m is the number of groups.

(i) For this part, assume $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_m^2 = \sigma^2$.

(a) Let $\phi = \sum_{i=0}^m \ell_i \alpha_i$ be a linear combination of parameter α_i's with given coefficients ℓ_i's. Prove that ϕ is estimable if and only if $\ell_0 = \sum_{i=1}^m \ell_i$.
Hint: ϕ is called estimable if it can be represented as the expectation of a linear combination of Y_{ij}.

(b) Let $\mu_i = \alpha_0 + \alpha_i$ denote the group mean for $i = 1, \ldots, m$. Write down the least-squares estimates for μ_i's, the MLE $\widehat{\sigma}^2_{\text{MLE}}$, and the unbiased estimator $\widehat{\sigma}^2_{\text{usual}}$ for σ^2.

(c) Suppose the data Y_{ij} are generated from the ground truth model

$$Y_{ij} = \alpha_0 + \varepsilon_{ij}, \quad \text{with} \quad \varepsilon_{ij} \sim_{\text{i.i.d}} N(0, \sigma^2). \tag{3}$$

However, the experimenter uses the overfitted model (2) to fit the data, and reports $\widehat{\sigma}^2_{\text{usual}}$ from question (b) as the analysis result. Show that the $\widehat{\sigma}^2_{\text{usual}}$ from overfitted model is still an unbiased estimate of σ^2 in model (3), despite the model misspecification.

(d) Let $\widehat{\sigma}^2_{\text{red}}$ be the estimate of σ^2 based on reduced (and true) model (3). Consider the 95%-confidence intervals (CI) for σ^2 based on the χ^2 procedure. Show that the expected length of CI from reduced model (i.e., based on $\widehat{\sigma}^2_{\text{red}}$) is smaller than the overfitted model (i.e., based on $\widehat{\sigma}^2_{\text{usual}}$).

(ii) For this part, suppose there is an additional known variable, denoted η_i, associated with each of the group. We return to the original setting, where both data and fitted model are based on (2),

(a) Suppose $\sigma_i^2 = \sigma^2 \eta_i^2$ for all $i = 1, \ldots, m$. Find the best linear unbiased estimator for μ_i's. Could you use a standard R function routine to find the results, or would you need to develop a general regression package? Explain.

(b) Now suppose η_i values are used to model the mean with the assumption $\mu_i = \beta_0 + \beta_1 \eta_i$, but we return to assuming that $\sigma_1^2 = \cdots = \sigma_m^2 = \sigma^2$ as in part (i). The experimenter is interested in testing whether the group means change linearly in η_i, or in an arbitrarily unstructured way over i. Formulate the question into a hypothesis testing problem. State the test statistic, null distribution, and rejection procedure. You could write your answers in matrix or algebraic forms; no need to simply the expressions.

4. Scientists want to study the effect of an anti-bacterial drug in fish lungs. The drug is administered at 5 dose-levels (0, 2, 4, 8, and 16 mg/L) as summarized in the below table to large controlled tanks with 100 fish in each through the filtration system. There are 20 tanks and each dose is randomly assigned to 4 tanks. At the end of the experiment, the fish are sacrificed, and the amount of bacteria in each fish is measured to yield total amount of bacteria per tank.

Dose	1	2	3	4	5
Dose of drug (ml/g)	0	2	4	8	16

Let y_{ij} denote the total amount of bacteria from the jth tank with the ith dose, $i = 1, \cdots, 5$ and $j = 1, 2, \cdots, 4$. Furthermore, suppose

$$y_{ij} = \mu_i + \epsilon_{ij}, \tag{4}$$

where $\mu_1, \cdots, \mu_5 \in \mathbb{R}$ are unknown parameters and ϵ_{ij} are independent and identically distributed $\mathcal{N}(0, \sigma^2)$ random variables for some unknown $\sigma^2 \in \mathbb{R}^+$. Use the `R` code and partial output provided below to answer the following questions.

(a) Provide the best linear unbiased estimator of μ_1.

(b) Provide the best linear unbiased estimator of μ_2.

(c) Determine the standard error of your esimate of μ_2 from part (b).

(d) Conduct a test of $H_0 : \mu_1 = \mu_2$. Provide a test statistics, the distribution of the test statistic (both under the null and the alternative), a p-value, and a conclusion.

(e) Provide an F-statistic for testing $H_0 : \mu_3 = \mu_4$.

(f) Scientists would like to consider a simple linear regression model with total amont of bacteria as a response and anti-bacterial drug dose as a quantitative variable to fit these data. Does such a model provide a better fit compared to the model in (4)? Provide a test statistic, its null distribution, a p-value, and a conclusion. You may find the upper 5^{th} percentiles of various F distributions in the `R` output useful for drawing your conclusion.

(g) Provide a matrix $\mathbf{A}$ and a vector c such that the null hypothesis of the test in part (f) can be written as $H_0 : \mathbf{A}\mu = c$, where $\mu = (\mu_1, \cdots, \mu_5)^T$.

R code and partial output for question # 4:

```
d <- rep(c(0, 2, 4, 8, 16), each = 4)
# y is the data vector representing total amount of bacteria per tank.
# Its entries are ordered to appropriately match the vector d.

dose <- factor(d)
m1 <- lm(y ~ dose)
```

```
summary(m1)

Call:
lm(formula = y ~ dose)

Residuals:
     Min       1Q   Median       3Q      Max
-15.9814  -5.6505   0.6174   3.3041  14.0753

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  350.713      4.204  83.428  < 2e-16 ***
dose2         -9.060      5.945  -1.524   0.1483
dose4         -1.826      5.945  -0.307   0.7630
dose8        -21.664      5.945  -3.644   0.0024 **
dose16       -65.442      5.945 -11.008 1.39e-08 ***


anova(m1)
Analysis of Variance Table

Response: y
          Df  Sum Sq Mean Sq F value    Pr(>F)
dose         11667.6
Residuals     1060.3


is.numeric(d)
[1] TRUE

m2 <- lm(y~d)

anova(m2)
Analysis of Variance Table

Response: y
          Df  Sum Sq Mean Sq F value    Pr(>F)
d            10819.3
Residuals    1908.6
```

```
## Upper 5th percentiles of various F-distributions

> qf(0.05, 1, 5, lower.tail = F)
[1] 6.607891
> qf(0.05, 2, 5, lower.tail = F)
```

```
[1] 5.786135
> qf(0.05, 3, 5, lower.tail = F)
[1] 5.409451
> qf(0.05, 4, 5, lower.tail = F)
[1] 5.192168
> qf(0.05, 5, 5, lower.tail = F)
[1] 5.050329
> qf(0.05, 1, 10, lower.tail = F)
[1] 4.964603
> qf(0.05, 2, 10, lower.tail = F)
[1] 4.102821
> qf(0.05, 3, 10, lower.tail = F)
[1] 3.708265
> qf(0.05, 4, 10, lower.tail = F)
[1] 3.47805
> qf(0.05, 5, 10, lower.tail = F)
[1] 3.325835
> qf(0.05, 6, 10, lower.tail = F)
[1] 3.217175
> qf(0.05, 7, 10, lower.tail = F)
[1] 3.135465
> qf(0.05, 8, 10, lower.tail = F)
[1] 3.071658
> qf(0.05, 9, 10, lower.tail = F)
[1] 3.020383
> qf(0.05, 10, 10, lower.tail = F)
[1] 2.978237
> qf(0.05, 1, 15, lower.tail = F)
[1] 4.543077
> qf(0.05, 2, 15, lower.tail = F)
[1] 3.68232
> qf(0.05, 3, 15, lower.tail = F)
[1] 3.287382
> qf(0.05, 4, 15, lower.tail = F)
[1] 3.055568
> qf(0.05, 5, 15, lower.tail = F)
[1] 2.901295
> qf(0.05, 6, 15, lower.tail = F)
[1] 2.790465
> qf(0.05, 7, 15, lower.tail = F)
[1] 2.706627
> qf(0.05, 8, 15, lower.tail = F)
[1] 2.640797
> qf(0.05, 9, 15, lower.tail = F)
[1] 2.587626
> qf(0.05, 10, 15, lower.tail = F)
[1] 2.543719
> qf(0.05, 1, 20, lower.tail = F)
[1] 4.351244
> qf(0.05, 2, 20, lower.tail = F)
[1] 3.492828
> qf(0.05, 3, 20, lower.tail = F)
[1] 3.098391
> qf(0.05, 4, 20, lower.tail = F)
[1] 2.866081
> qf(0.05, 5, 20, lower.tail = F)
[1] 2.71089
> qf(0.05, 6, 20, lower.tail = F)
[1] 2.598978
> qf(0.05, 7, 20, lower.tail = F)
[1] 2.514011
> qf(0.05, 8, 20, lower.tail = F)
[1] 2.447064
> qf(0.05, 9, 20, lower.tail = F)
[1] 2.392814
> qf(0.05, 10, 20, lower.tail = F)
[1] 2.347878
```