PhD Qualifying Examination
Theory Part
Summer 2025

1. (10 points) A factory produces n bulbs using a machine. Assume the observed data $X_1, ..., X_n$ record the failure times of the bulbs and $X_i \sim \text{Exp}(\lambda)$ with constant failure rate λ. Note the pdf for X_i is $f(x; \lambda) = \lambda e^{-\lambda x}$ with mean $1/\lambda$ and variance $(1/\lambda)^2$.

(a) Consider $\beta = 1/\lambda$. We would like to make asymptotic inference on the variance β^2.

(i) (1 point) Derive the MLE $\hat{\beta}_n$ for β. What is the asymptotic distribution? Identify clearly the theorem and conditions used.

(ii) (2 points) From the previous result, derive the asymptotic normality of $\hat{\beta}_n^2$. Identify clearly the theorem and conditions used.

(b) Now we find that the factory has used two machines: Machine A and Machine B. For each bulb, the machine used is chosen at random: with probability π, the bulb is made by Machine A, and with probability $1 - \pi$, it is made by Machine B. However, the machine used for each bulb is not recorded. Suppose Machine A produces bulbs with failure rate λ_A and Machine B produces bulbs with failure rate λ_B.

You are given failure time data of the n bulbs, $x_1, ..., x_n$. However, the latent variables $z_1, ..., z_n$ are not observed, where $z_i = 1$ if the i-th bulb is from Machine A and $z_i = 0$ if the i-th bulb is from Machine B.

(i) (1 point) Write down the complete data log-likelihood, assuming latent variables $z_1, ..., z_n$ are known: $\ell_C(\theta \mid X, Z)$ with data $X = (x_1, ..., x_n)$ and $Z = (z_1, ..., z_n)$ and parameters $\theta = (\pi, \lambda_A, \lambda_B)$.

(ii) (2 points) Calculate $\gamma_i^{(k)} = \mathbb{E}(z_i \mid x_i, \theta^{(k)}) = \Pr(z_i = 1 \mid x_i, \theta^{(k)})$, where $\theta^{(k)}$ is the estimate of θ in the kth (last) iteration. Calculate $\Pr(Z \mid X, \theta^{(k)})$.
Hint: Use Bayes Rule to calculate $\Pr(z_i = 1 \mid x_i, \theta^{(k)})$. *Note that* $(x_1, z_1), (x_2, z_2), ..., (x_n, z_n)$ *are mutually independent.*

(iii) (2 points) Derive the E-step of the EM algorithm. Compute

$$Q(\theta, \theta^{(k)}, X) = \mathbb{E}[\ell_C(\theta \mid X, Z) \mid X, \theta^{(k)}]$$

Hint: Note that $(x_1, z_1), (x_2, z_2), ..., (x_n, z_n)$ *are mutually independent.*

(iv) (2 points) Derive the M-step update equations for $\theta = (\pi, \lambda_A, \lambda_B)$. Use the following optimization to update $(\pi^{(k+1)}, \lambda_A^{(k+1)}, \lambda_B^{(k+1)})$:

$$\theta^{(k+1)} = \arg\max_{\theta} Q(\theta, \theta^{(k)}, X)$$

2. (10 points) Consider random variables Y_i $(i = 1, ..., n)$, the number of hospitalizations of patient i in the past five years, and X_i $(i = 1, ..., n)$ is the treatment covariate of patient i, where $X_i = 0$ means patient i is a control and $X_i = 1$ means patient i is treated. Assume that $Y_1, ..., Y_n$ are independent. For $i = 1, ..., n$, the random variable Y_i follows Poisson(μ_i) with pmf $f(Y_i; \mu_i) = \mu_i^{Y_i} e^{-\mu_i}/Y_i!$. To test for treatment effect, we can consider the following two hypothesis testing setups:

Setup (A): Assume that $\mu_i = \eta_0$ if $X_i = 0$ and $\mu_i = \eta_1$ if $X_i = 1$. We will test whether $\eta_0 = \eta_1$.

Setup (B): Assume that $\mu_i = a + bX_i$. We will test whether $b = 0$.

For simplicity of your calculation, use the notation: $C = \{i : X_i = 0\}$ and $D = \{i : X_i = 1\}$ with sizes n_0 and n_1, respectively. Also denote $\bar{Y}_C = \left(\sum_{i \in C} Y_i\right)/n_0$, $\bar{Y}_D = \left(\sum_{i \in D} Y_i\right)/n_1$, and $\bar{Y} = \left(\sum_{i=1}^n Y_i\right)/n$.

(a) (3 points) Under Setup (A), we will test

$$H_0^A : \eta_0 = \eta_1 \quad \text{vs.} \quad H_1^A : \eta_0 \neq \eta_1.$$

Derive the likelihood ratio test statistic T_{LR} to test H_0^A. What is the asymptotic distribution of T_{LR} under H_0^A?

(b) Under Setup (B), we will test

$$H_0^B : b = 0 \quad \text{vs.} \quad H_1^B : b \neq 0.$$

This model has two parameters $\theta = (b, a)$. Note that we put b in front of a in θ so that the first parameter is b to be tested, while the second parameter a is a nuisance parameter. This allows you to use the hint of the inverse of partitioned matrix below.

(i) (2 points) Calculate the constrained MLE under H_0^B for θ (denoted as $\tilde{\theta} = (\tilde{b}, \tilde{a})$).

(ii) (3 points) Calculate the total information matrix $I_T(\theta)$ as a function of θ.

(iii) (2 points) Find the score test statistic T_S to test H_0^B. What is the asymptotic distribution of T_S under H_0^B?

Hint: Suppose

$$I = \begin{pmatrix} I_{11} & I_{12} \\ I_{21} & I_{22} \end{pmatrix}, \quad I^{-1} = \begin{pmatrix} I^{11} & I^{12} \\ I^{21} & I^{22} \end{pmatrix}$$

Then we have $I^{11} = (I_{11} - I_{12} I_{22}^{-1} I_{21})^{-1}$.

3. (10 points) Consider a pair of random variables (X, Y) where X is discrete with $Pr[X = 1] = Pr[X = 4] = \frac{1}{2}$ and $[Y|X = x] \sim N(\theta, x)$.

(i) (3 points) Show that the joint density of (X, Y) is

$$f(x, y; \theta) = \frac{1}{2\sqrt{2\pi x}} exp\{-\frac{(y-\theta)^2}{2x}\}.$$

Verify the identifiability condition required for the maximum likelihood method.

(ii) (4 points) Consider an iid random sample $\{(x_i, y_i) : i = 1, 2, \ldots, n\}$, find the maximum likelihood estimate of θ, denoted as $\widehat{\theta}_n$, and determine the limiting distribution of $\sqrt{n}(\widehat{\theta}_n - \theta)$. No need to verify regularity conditions.

(iii) (3 points) Find the influence function of $\widehat{\theta}_n$.

4. (10 points) Let $X_1, \ldots, X_n$ be a sample from the beta distribution with density,

$$f(x;\theta) = \theta x^{\theta-1}, \quad \text{where } 0 < x < 1 \text{ and } \theta > 0.$$

(i) (2 points) Let $\widehat{\eta}_n$ be the sample median, and $\eta = m(\theta)$ the population median as a function of θ. Find the limiting distribution of $\sqrt{n}(\widehat{\eta}_n - \eta)$.
Hint: the influence function for the sample median $\widehat{\eta}_n$ is:

$$IC(x, F, \eta) = \frac{0.5 - I\{F(x) \leq 0.5\}}{f(\eta)}.$$

You may use this information without verifying any regularity condition for the consistency of the sample median.

(ii) (1 point) Let $\widehat{\theta}_n = log(\frac{1}{2})/log(\widehat{\eta}_n)$, show that $\widehat{\theta}_n \to^p \theta$, as $n \to \infty$.

(iii) (3 points) Provide the influence function of $\widehat{\theta}_n$ and find the limiting distribution of $\sqrt{n}(\widehat{\theta}_n - \theta)$.

(iv) (2 points) From $\widehat{\theta}_n$, derive the one-step Newton-Raphson estimator $\tilde{\theta}_n$ that is as efficient as the maximum likelihood estimate.

(v) (2 points) Find the limiting distribution of $\sqrt{n}(\tilde{\theta}_n - \theta)$.
Hint: use the fact that $\tilde{\theta}_n$ is an efficient estimate of θ.

5. (10 points) Consider the following generalized linear model.
For $i = 1, \ldots, n$, we assume that

- $\mathbb{E}(y_i) = \mu_i$
- $\log(\mu_i) = \alpha + x_i\beta$
- $\text{Var}(y_i) = \sigma^2\mu_i$
- $\beta \in \mathbb{R}$ (scalar)

(i) (4 points) Provide the equation for the **maximum quasi-likelihood estimator** of β.

(ii) (2 points) Provide an estimator for σ^2 based on $\hat{\theta} = \begin{pmatrix} \hat{\alpha} & \hat{\beta} \end{pmatrix}^T$, where $\hat{\alpha}$ and $\hat{\beta}$ are the maximum quasi-likelihood estimators of α and β, respectively.

(iii) (4 points) Find the **asymptotic distribution** of $\hat{\theta}$.

6. (10 points) Consider the following mixed model:

$$y_{ij} = \alpha + u_i + \epsilon_{ij},$$

where $i = 1, \ldots, m,\ j = 1, \ldots, n_i$; $u_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_u^2)$; $\epsilon_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$; u_i's and ϵ_{ij}'s are independent.

(i) (4 points) Suppose it is known that $\sigma_u^2 = \rho\sigma^2$ for a known constant ρ. Find the **maximum likelihood estimator** (MLE) for α **without directly maximizing the likelihood** (you may leave matrix operation there, but clearly specify all the matrices used). Justify why your estimator is the MLE.

(ii) (3 points) The REML estimator for σ^2 is given by:

$$\hat{\sigma}^2 = \frac{1}{\sum_{i=1}^{m}(n_i - 1)} \sum_{i=1}^{m} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i\cdot})^2$$

Show that $\hat{\sigma}^2$ is an **unbiased estimator** for σ^2.

(iii) (3 points) Find the **best predictor** $\text{BP}(u_i)$.
Hint: Suppose that

$$\boldsymbol{y} = \begin{pmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \end{pmatrix} \sim \mathcal{N}\left\{ \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \begin{bmatrix} \boldsymbol{V}_{11} & \boldsymbol{V}_{12} \\ \boldsymbol{V}_{21} & \boldsymbol{V}_{22} \end{bmatrix} \right\}$$

Then, we have

$$\boldsymbol{y}_1 | \boldsymbol{y}_2 \sim \mathcal{N}\big(\boldsymbol{\mu}_1 + \boldsymbol{V}_{12}\boldsymbol{V}_{22}^{-1}(\boldsymbol{y}_2 - \boldsymbol{\mu}_2), \boldsymbol{V}_{11} - \boldsymbol{V}_{12}\boldsymbol{V}_{22}^{-1}\boldsymbol{V}_{21}\big)$$

7. (10 points) Researchers investigate the effect of **exercise treatment** on **systolic blood pressure (SBP)**. Patients are randomized into **4 groups** with varying **exercise durations**: 0, 30, 60, and 90 minutes per day. For each patient j in group i ($i = 0, 1, 2, 3$; $j = 1, \ldots, 5$), they also measure their **baseline BMI**, denoted x_{ij}. Define d_i as the numerical exercise duration level.

The models and their SSEs are:

Model	Description	Model Equation	SSE
A	Categorical treatment + BMI	$y_{ij} = \mu + \alpha_i + \theta x_{ij} + \epsilon_{ij}$	SSE_A
B	Categorical treatment only	$y_{ij} = \mu + \alpha_i + \epsilon_{ij}$	SSE_B
C	Linear treatment + BMI	$y_{ij} = \mu + \delta d_i + \theta x_{ij} + \epsilon_{ij}$	SSE_C
D	Linear treatment only	$y_{ij} = \mu + \delta d_i + \epsilon_{ij}$	SSE_D
E	Mean model	$y_{ij} = \mu + \epsilon_{ij}$	SSE_E
Total			TSS

(a) (2 points) Demonstrate that Model D is nested within Model B.

(b) (4 points) Using the SSEs, derive expressions for the following sequential sums of squares in terms of $\text{SSE}_A, \ldots, \text{SSE}_E$ and TSS.

(i) $R(\mu)$

(ii) $R(\boldsymbol{\alpha} \mid \mu, \delta)$

(iii) $R(\theta \mid \boldsymbol{\alpha})$

(iv) $R(\boldsymbol{\alpha} \mid \mu, \delta, \theta)$

(c) (4 points) Construct null hypotheses and conduct tests at the $\alpha = 0.05$ level for the following.

(i) **Test 1:** Does BMI explain significant variability in SBP? Use Model B vs Model A.

(ii) **Test 2:** Is the dose-response relationship adequately modeled by a linear effect? Compare Model A vs Model C.