Applied Statistics Qualifier Examination
(Part II of the STAT AREA EXAM)

May 22, 2024; 11:00AM-1:00PM EDT

General Instructions:

(1) The examination contains 4 Questions. You are to **answer 3 out of 4** of them. *** Please only turn in solutions to 3 questions *** (Please note that if you choose to do Question 4, you will have the flexibility to choose to do one of the two problems (4a, 4b) in that category, in this special transition period.)

(2) You may use up to 4+2 books and 4+2 class notes, plus your calculator and the statistical tables.

(3) NO computer, internet, cell phone, or smart watch is allowed.

(4) ***This is a 2-hour exam*** **11:00am- 1:00 PM – Please turn in by 1:00pm.**

Please be sure to fill in the appropriate information below:

I am submitting solutions to QUESTIONS _____, _____, and _____ of the applied statistics qualifier examination. Please put your name on every page of your exam solutions, and add page number for solutions to each question individually.

There are _____ pages of written solutions.

Please read the following statement and sign below:

Academic integrity is expected of all students at all times, whether in the presence or absence of members of the faculty. Understanding this, I declare that I shall not give, use, or receive unauthorized aid in this examination.

_____
(Signature)

_____
(Name)

_____
(SBU ID)

Name: _____ Signature: _____

1. The salinity of water samples taken from four separate sites, namely sites A-D, in Long Island was measured. Below is the data collected.

Site	Sample Size	Data	Sample mean	Sample variance
A	10	38.3, 39.3, 44.7, 40.2, 40.4, 45.1, 41.4, 36.2, 37.9, 38.7	40.22	8.19
B	15	45.7, 43.1, 43.2, 42.3, 40.3, 47.4, 43.5, 36.1, 44.1, 40.6, 38.8, 41.3, 38.9, 39.8, 40.1	41.68	8.51
C	10	27.8, 38.6, 35.7, 30.2, 40.3, 36.8, 33.7, 38.8, 38.7, 38.5	35.91	17.05
D	15	38.9, 38.4, 35.7, 34.7, 34.4, 33.1, 35.1, 30.6, 45.2, 41.1, 31.2, 34.3, 34, 39.3, 35.6	36.11	14.88

Let A_i, B_i, C_i and D_i denote random samples from site A, B, C and D, respectively. Based on historical data, it can be assumed that $A_i \sim N(\mu_A, \sigma_A^2)$, $B_i \sim N(\mu_B, \sigma_B^2)$, $C_i \sim N(\mu_C, \sigma_C^2)$ and $D_i \sim N(\mu_D, \sigma_D^2)$, where $\sigma_A^2 = \sigma_B^2 = \frac{\sigma_C^2}{2} = \frac{\sigma_D^2}{2}$, but the actual values of μ_i's and σ_i^2 are unknown.

A research group is interested in testing the following hypotheses simultaneously at overall $\alpha = 0.05$

$H_0: 4\mu_A - 2\mu_B - 2\mu_C = 1$ vs $H_1: 4\mu_A - 2\mu_B - 2\mu_C > 1$
and
$H_0: \mu_A + \mu_C - 2\mu_D = 2$ vs $H_1: \mu_A + \mu_C - 2\mu_D \neq 2$

What is the conclusion of the hypotheses tests? Show your work to get full credit.

Name: _____ Signature: _____

2. Suppose that a population of individuals can be partitioned into k sub-populations or groups, $G_1, \dots, G_k$, with relative frequencies $\pi_1, \dots, \pi_k$. Multivariate measurements Z made on individuals have the following distributions for the k groups:

$$G_j: Z \sim N_p(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}), \qquad j = 1, \dots, k.$$

Let $\boldsymbol{z}^*$ be an observation made on an individual drawn at random from the combined population. The prior odds that the individual belongs to G_j are $\pi_j/(1-\pi_j)$. Show that the posterior odds for G_j given $\boldsymbol{z}^*$ are

$$dds(Y = j|\boldsymbol{z}^*) = \frac{\pi_j}{1-\pi_j} \times e^{\alpha_j + \boldsymbol{\beta}_j^T \boldsymbol{z}^*}.$$

(a) Find the expression for α_j and $\boldsymbol{\beta}_j$ in terms of $\boldsymbol{\mu}_j$ and $\boldsymbol{\Sigma}$.

(b) What simplifications can be made if the k Normal means $\boldsymbol{\mu}_j$ lie on a straight line in R^p?

(c) Comment on the difference between the maximum likelihood estimation of α_j and $\boldsymbol{\beta}_j$ via the normal-theory likelihood and estimation via logistic regression.

Name: _____ Signature: _____

3. Consider the model $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{V})$ and $\boldsymbol{X}$ is $n \times p$ of rank p. Assume $\boldsymbol{V}$ is known but not σ^2.

a) Find an unbiased estimator of σ^2.

b) Find an unbiased estimator of $\frac{1}{\sigma^2}$.

c) For the hypotheses: $H_0: K'\beta = 0 \; vs \; H_1: K'\beta \neq 0$, where K' is a known matrix of order $s \times p$ and rank s, show how to test this hypothesis.

Name: _____ Signature: _____

4a. Consider the VAR (1) model:

$$\begin{bmatrix} 1 & -2 \\ 0 & 1 \end{bmatrix} \begin{pmatrix} Y_{1,t} \\ Y_{2,t} \end{pmatrix} = \begin{bmatrix} 0.9 & 0.2 \\ -0.3 & 0.7 \end{bmatrix} \begin{pmatrix} Y_{1,t-1} \\ Y_{2,t-1} \end{pmatrix} + \begin{pmatrix} \epsilon_{1,t} \\ \epsilon_{2,t} \end{pmatrix},$$

$$\text{where } \begin{pmatrix} \epsilon_{1,t} \\ \epsilon_{2,t} \end{pmatrix} \overset{i.i.d.}{\sim} \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 4 \end{bmatrix} \right)$$

(a) Which VAR form, structural or reduced form, is the above model? Please transform this model to the other form. That is, if the above is a structural form, please transform it to the reduced form, and vice versa. What is the variance-covariance matrix, Σ, of the error term of the reduced form?
(b) Is this VAR(1) series stationary? Please show the entire derivation.
(c) Please derive the mean and variance-covariance matrix of the time series vector $(Y_{1,t} \;\; Y_{2,t})$ from the reduced form of this VAR(1).
(d) Please derive the marginal series of this VAR(1).

Name: _____ Signature: _____

4b. Using the following data set, we aim to predict COVID infection using the predictors 'Loss of taste or smell (LOTOS)' and 'Fever'.

LOTOS	Fever	COVID
Yes	Yes	Yes
No	No	No
No	Yes	Yes
Yes	No	Yes
Yes	No	Yes
No	No	No
Yes	Yes	Yes
Yes	No	No
No	No	No
Yes	Yes	Yes

(a) Please build a decision tree (in the format of ID4) using the Gini Index as the splitting criterion.
(b) Please write down a suitable logistic regression model for this predictive work. Please show steps of how you will fit this model. (Surely, it will be excellent if you can finish fitting this model if time permits.)
(c) Do you think the decision tree and the logistic regression model above would yield identical results or not? Please discuss your reasoning in detail using all means necessary such as derivations and/or figure illustrations.