Final 2021. 06. 21

I. Statistics (120%)

1. (50%) For the following data from 3 populations:

	Observations
Population 1	$(0.37, 1.06, 0.36)$,$(1.06, 1.65, 2.06)$, $(0.81, 0.92, 0.80)$ $(-0.98, 1.75, 1.08)$, $(0.92, -0.85, 0.73)$
Population 2	$(1.13, 2.53, 1.42)$, $(0.71, 3.19, 2.01)$, $(2.28, 1.00, 2.16)$ $(1.54, 1.99, 0.21)$, $(2.11, 1.45, 1.46)$
Population 3	$(1.63, 2.62, 1.15)$, $(2.66, 4.23, 0.23)$, $(1.55, 1.36, 2.05)$ $(-0.82, 1.22, 0.75)$, $(2.15, 3.10, 1.00)$

Please do the following:

(a) Find the variance-covariance matrix and the correlation matrix using all data

(b) Find the principal components by both 75% criterion and mean criterion using all data.

(c) For the data in all populations, please use Fisher's discrimination method to find $\hat{a}_1$ and $\hat{a}_2$, the coefficient vectors of the discriminant functions.

(d) Find the error rate for the 15 observations based on $\hat{a}_1$ and $\hat{a}_2$ in (c).

(e) Please find the smallest error rate for the above data as using K-means method with the number of clusters equal to 3.

2. (50%) Here is a set of data with the model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i, i = 1, \cdots, 5 \ \ \epsilon_i \sim N(0, \sigma^2),$$

y_i	72	81	98	123	129
x_{i1}	-1	-1	0	1	1
x_{i2}	-1	0	0	0	1

Compute the following.

(a) Least squares estimate and R^2.

(b) Find F statistic and p-value to test $H_0: \beta_1 = \beta_2 = 0$.

(c) Find the t statistic and p-value to test $H_0: \beta_1 \le 22$ vs. $H_1: \beta_1 > 22$.

(d) Find F statistic and p-value to test $H_0: \beta_1 - 3\beta_2 = 0$.

(e) Find AIC and BIC (or SBC) and select the best model(s) based on the two criteria.

3. (20%)

(a) Let a random sample of $X_1, \cdots, X_n \sim Poisson(\lambda)$. Then, two estimators of of $(1+\lambda)e^{-\lambda}$ are

$$\delta_1 = (1+\bar{X})e^{-\bar{X}}; \ \delta_2 = \left(\frac{n-1}{n}\right)^{n\bar{X}} + \bar{X}\left(\frac{n-1}{n}\right)^{n\bar{X}-1}.$$

Please use setClass and setMethod in R to implement the two estimators and the method to compute the averages of the absolute differences between the above two estimates and the true value of the parameter. Then, generate 1000 samples of random Poisson data with size 100 and mean 1 and compute the mean absolute difference for these estimates.

(b) Using logistic regression models to analyze the data in the following table with frequencies for delinquency and two variables, whether being a boy scout and socioeconomic status.

		Delinquent (crime)	
Socioeconomic Status	Boy Scout	Yes	No
Low	Yes	10	40
	No	40	160
Medium	Yes	18	132
	No	18	132
High	Yes	8	192
	No	2	48

II. Computing (120%)

1. (25%) Let

$$z = (y - 1.5x^2)(y - 0.5x^2), x, y = -10, -9.9, \cdots, 0, 0.1, \cdots, 10$$

and

$$y = 3e^x cos(x^3) - 10log(x) + 2x^{-1.2}, x = 1.1, 1.2, \cdots, 3.$$

Please write a program to plot the two functions in two plots and place the two plots within one figure.

2. (25%) Please generate the data from the model

$$y_i = 3x_{i1} + 5x_{i2} + \epsilon_i, \epsilon_i \sim N(0, 3^2), i = 1, \cdots, 100,$$

where both x_{i1} and x_{i2} are generated from a standard normal random variable. Then, the mean residual sum of square s^2 can be obtained. By repeating the above process 1000 times, please find mean absolute difference of the mean residual sum of square and the true variance of the random errors.

3. (25%) Write a program to find the solution to 6 decimal places of accuracy using Newton's method for the following equations

$$\boldsymbol{x^2 - 2x - y + 0.5 = 0}$$
$$\boldsymbol{x^2 + 4y^2 - 4 = 0}$$

with starting point $\begin{bmatrix} 2 \\ 0.25 \end{bmatrix}$.

4. (25%) Please approximate the integral

$$\int_0^1 \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$$

by Simpson's method to do the following.

(a) Please use 10 sub-intervals to approximate the integral. (15%)

(b) Suppose $S(N)$ is the value as using Simpson's method with N sub-intervals. Find the smallest N such that

$$|S(N) - 0.34134474| \le 0.00000001.$$

(10%)

5. (20%) Let a random sample of $X_1, \cdots, X_n \sim Poisson(\lambda)$. Then, the estimators of λ are the sample mean. Based on the central limit theorem,

$$\frac{\bar{X} - \lambda}{\sqrt{Var(\bar{X})}} \approx N(0, 1)$$

Please check if the central limit theorem provides a good approximation as

(a) $\mathrm{n} = 10, \lambda = 1.$

(b) $\mathrm{n} = 50, \lambda = 0.1.$

(c) $\mathrm{n} = 100, \lambda = 0.1.$

(d) $\mathrm{n} = 1000, \lambda = 1.$