# Assignment 1

This assignment contributes to 15 % of your final grade.

**Due Date: Sunday 14 October by 11:59 PM**

Late submissions will receive a 10 % penalty for each additional 24 hours over the due date.

## Instructions

Complete the questions below either by hand, or using SAS code or a combination of the two.

- When completing by hand ensure you show all working and formulas used to receive full marks.
- When completing by SAS code ensure you include the SAS output and your full SAS code file (or copy of code text in an appendix) with your assignment.

Submissions are to be made through Canvas. All work is to be your own as in accordance with the Assessment declaration.

## *Question 1*

The data file pollution.csv (as on Canvas and SAS Studio) contains information on air pollution measurements. Using the file examine the pair of measurements $X_5$=Nitrious Oxide and $X_6$=Ozone for bivariate normality by completing the following:

a) Calculate the distances of these observations from their means

(2 marks)

b) Determine the proportion of the observations falling within the estimated 50% probability contour of a bivariate normal distribution

(1 marks)

c) Construct a chi-square plot of your distances from part a) above

(2 marks)

d) Given your results in part b) and part c) are these data approximately bivariate normal? Explain.

(3 marks)

## *Question 2*

The datafile results.csv, contains three test results assessing different types of intelligence. Test the following hypothesis at α=0.02.

$$H_0 : \mu' = \begin{bmatrix} 85 & 75 & 55 \end{bmatrix}$$

a) Conduct the hypothesis test showing all steps required. If the values given above are the average score for all college students over the last ten years, is there reason to believe the group in the datafile are scoring differently? Explain.

(5 marks)

b) Determine the lengths and directions for the axes of the 90% confidence ellipsoid for μ

(2 marks)

c) Construct the three possible scatter diagrams from the pairs of variables. Do these data appear to be normally distributed? Discuss

(3 marks)

## *Question 3*

Using the data matrix

$$\mathbf{X} = \begin{bmatrix} 3 & 4 & 15 & -6 \\ 2 & 4 & 14 & -7 \\ 3 & 4 & 15 & -5 \\ 3 & 3 & 16 & -6 \\ 2 & 5 & 15 & -7 \\ 1 & 4 & 14 & -4 \end{bmatrix}$$

Calculate the following:

a) The independent 95 % confidence intervals for each variable

(1.5 marks)

b) The Bonferroni 95 % confidence intervals for each variable

(1.5 marks)

c) The simultaneous 95 % confidence intervals for each variable

(1.5 marks)

d) The 95 % confidence interval for the difference between $\mu_2$ and $\mu_4$. Are these means different?

(1.5 marks)

e) Discuss the results from part a) to part d) above and explain any differences in the observed estimates.

(2 marks)

## *Question 4*

The data file track.csv (as on Canvas and SAS Studio) contains information on female national track records. Using the file completing the following:

a) Read the datafile in and obtain the sample correlation matrix R, and determine the eigenvalue/eigenvector pairs.

(3 marks)

b) State the first two principal components for the standardized variables and calculate the cumulative percentages of the total (standardized) sample variance explained.

(2 marks)

c) Prepare a table showing the correlation of the standardized variables with the first two components.

(2 marks)

d) Interpret the two principal components from Part b).

(3 marks)

e) Rank the nations based on their score on the first principal component. Discuss whether this meets your expectations.

(2 marks)

## *Question 5*

The correlation matrix below is from the measurement of skeletal features of white leghorn fowl (Dunn, *Storrs Agricultural Experimental Station Bulletin,* **52**, 1928). Where

$X_1$ = Skull length
$X_2$ = Skull breadth
$X_3$ = Femur length
$X_4$ = Tibia length
$X_5$ = Humerus length
$X_6$ = Ulna length

$$\mathbf{R} = \begin{bmatrix} \mathbf{1} & \mathbf{0.505} & \mathbf{0.569} & \mathbf{0.602} & \mathbf{0.621} & \mathbf{0.603} \\ \mathbf{0.505} & \mathbf{1} & \mathbf{0.422} & \mathbf{0.467} & \mathbf{0.482} & \mathbf{0.450} \\ \mathbf{0.569} & \mathbf{0.422} & \mathbf{1} & \mathbf{0.926} & \mathbf{0.877} & \mathbf{0.878} \\ \mathbf{0.602} & \mathbf{0.467} & \mathbf{0.926} & \mathbf{1} & \mathbf{0.874} & \mathbf{0.894} \\ \mathbf{0.621} & \mathbf{0.482} & \mathbf{0.877} & \mathbf{0.874} & \mathbf{1} & \mathbf{0.937} \\ \mathbf{0.603} & \mathbf{0.450} & \mathbf{0.878} & \mathbf{0.894} & \mathbf{0.937} & \mathbf{1} \end{bmatrix}$$

Using the maximum likelihood procedure the following estimated factor loadings were extracted:

| **Variable** | **Estimated Loadings** | | **Varimax rotated loadings** | |
|---|---|---|---|---|
| | **F1** | **F2** | **F1*** | **F2*** |
| 1 | 0.602 | 0.200 | 0.484 | 0.411 |
| 2 | 0.467 | 0.154 | 0.375 | 0.319 |
| 3 | 0.926 | 0.143 | 0.603 | 0.717 |
| 4 | 1.000 | 0.000 | 0.519 | 0.855 |
| 5 | 0.874 | 0.476 | 0.861 | 0.499 |
| 6 | 0.894 | 0.327 | 0.744 | 0.594 |

d) Using the unrotated estimated factor loadings, obtain the maximum likelihood estimates of the following:
   i. The specific variances.
   ii. The communalities.
   iii. The proportion of variance explained by each factor.
   iv. The residual matrix $\mathbf{R} - \hat{\mathbf{L}}\hat{\mathbf{L}}' - \hat{\boldsymbol{\Psi}}$

(2.5 marks)

e) Using the varimax rotated estimated factor loadings, obtain the maximum likelihood estimates of the following:
   i. The specific variances.
   ii. The communalities.
   iii. The proportion of variance explained by each factor.
   iv. The residual matrix $\mathbf{R} - \hat{\mathbf{L}}\hat{\mathbf{L}}' - \hat{\boldsymbol{\Psi}}$

(2.5 marks)

f) Comment on the results using the two loading methods by comparing your results from part a) and part b) above.

(2 marks)