UNIVERSITY OF CAPE TOWN

DEPARTMENT OF STATISTICAL SCIENCES

STA3022F

TEST 2

Question 1 [5 marks]

(a) What is test-retest reliability? (1)

(b) What is internal consistency reliability? (1)

(c) How do you measure internal consistency? Provide three formula's or explanations, not just the names of the methods. (3)

Answer to Q1

(a) One of

A reliable measuring instrument in this context is one that gives consistent scores when used repeatedly.

Or

There should be high correlations between test scores taken over multiple trials.

(b) The group of questions is internally-consistent or reliable if they are able to measure the same underlying construct.

(c) Chronbach's alpha $= \frac{k}{k-1} \times \frac{var(Q_1+\cdots Q_k)-\{var(Q_1)+\cdots+var(Q_k)\}}{var(Q_1+\cdots Q_k)}$

α-if-deleted by calculating Chronbach's alpha without each questions

Item total correlation by calculating the correlation between each question and the sum of all the other questions.

Half mark for each name and half mark for formula/description

Question 2 [16 marks]

(a) In the painters data set in the R package MASS the subjective assessment, on a 0 to 20 integer scale, of 54 classical painters is given. The painters were assessed on four characteristics: composition, drawing, colour and expression. Calculate the Euclidean distance between the following two samples:

```
> painters[1:2,]
          Composition Drawing Colour Expression
Da Udine           10       8     16          3
Da Vinci           15      16      4         14
```

(3)

(b) Why is there no need to scale the data set before calculating the Euclidean distance? (1)

(c) Define $stress$ and explain how it is used. (5)

(d) Explain step by step how to perform hierarchical clustering with the centroid method. (7)

Answer to Q2

(a) $d_{12} = \sqrt{\sum_{j=1}^{p}(x_{1j} - x_{2j})^2} = \sqrt{(10-15)^2 + (8-16)^2 + (16-4)^2 + (3-14)^2} =$

$\sqrt{(-5)^2 + (-8)^2 + (12)^2 + (-11)^2} = \sqrt{25 + 64 + 144 + 121} = \sqrt{354} = 18.8$

(b) All variables are measured on the same 0 to 20 integer scale.

(c) $stress = \sum_{i=1}^{n-1}\sum_{j=i+1}^{n}(d_{ij} - \delta_{ij})^2$

The aim of MDS is to find a representation of the samples so that the dissimilarities between them in the plot, given by δ_{ij}, match the given dissimilarities d_{ij} as closely as possible (optimally).

If the symbols are reversed, no marks are deducted as long as die descriptions are correct.

(d) Start with all objects each in its own cluster.

Merge the two closest clusters

Repeat

Calculate the dissimilarity between the newly merged cluster and each other cluster

By calculating the distance between the cluster means

Merge the two closest clusters

Until all objects are merged into the same cluster.

Use the clustering tree to cut the tree at a specific height or into a specific number of clusters.

QUESTION 3 [17 marks]

The current study aims to identify what factors make some people believe that they are lucky and others believe that they are unlucky. The study is based on a survey of 62 STA3022F students who answered the following questions in an online questionnaire (possible responses for categorical variables are given in brackets).

1. Do you consider yourself to be a lucky person? (*Yes/No*)
2. What is your age?
3. What is your gender? (1 = *Male;* 0 = *Female*)
4. Have you ever won a competition before? (1 = *Yes;* 0 = *No*)
5. How many economic courses have you completed?

A discriminant analysis model has been constructed with the aim of identify which, if any, of the four independent variables are able to distinguish between the two groups (groups labelled as "Yes", and "No").

Questions:

a) Write down the discriminant function. (2)

b) Which groups is the discriminant model able to significantly discriminate between? Provide statistical evidence at the 5% level to support your answer. Clearly state all null and alternate hypotheses. (4)

c) Use the cut-off value rule to classify **Respondent 4**. Clearly indicate the classification rule. Is this a correct classification? (5.5)

d) Compare the overall hit rate with two chance criteria and use these comparisons to evaluate the overall quality of the discriminant model (4)

e) Evaluate whether the discriminant model is better at predicting some groups than others. (Hint: Calculate the correct classification rate for each group) (1.5)

Q3-a) Write down the discriminant function.

$$Z_1 = 0.254 - 2.948 * Q2 + 0.085 * Q3d + 1.383 * Q4d - 0.011 * Q5$$

1✓2 1✓2 1✓2 1✓2

Q3-b)

H_0: There is no difference between the yes and no categories' centroids. ✓
H_1: There is difference between the yes and no categories'centroids.

First we need to calculate the distance:

1✓2 1✓2

$$d^2 = (-1.0242 - 1.0974)^2 = 4.501187$$

1✓2 (ratio) 1✓2 (answer)

$$F_{yes,low} = \frac{(n-1-p)n_1n_2}{p(n-2)(n_1+n_2)}d^2 = \frac{(62-1-4)*34*28}{4*(62-2)*(34+28)} * 4.501187 = 16.41481$$

$$F_{critical} = F_{p,n-1-p,\alpha} = F_{4,62-1-4,0.05} = F_{4,57,0.05} = 2.533$$

1✓2 (comparison) 1✓2 (conclusion)

Since F calculated is greater than the critical F value, centroids are significantly different from each other at 5% sig. level.

(or alternatively they can say that the F calculated is very high)

Q3-c) First we need to calculate the cut-off value 1✓2 (ratio)

$$Cut-off = \frac{n_1\bar{Z}^2 + n_2\bar{Z}^1}{n_1 + n_2} = \frac{34 * 1.0974 + 28 * (-1.0242)}{34 + 28} = 0.1392581$$ 1✓2 (answer)

Then we need to specify the rule:

If Z<0.1392581 then classify as "YES" 1✓2

Calculate Z value for the 4th respondent:

$$Z_4 = 0.254 - 2.948 * 20 + 0.085 * 1 + 1.383 * 0 - 0.011 * 2 = -58.643$$ 1✓2 1✓2 1✓2 1✓2 1✓2

Since $Z_4 < 0.1392581$, classify as "YES", hence the centroid for Yes is negative 1✓2 1✓2

Therefore it is a correct classification. 1✓2

Q3-d) Evaluate the hit-rate.

$$Hit-rate = \frac{28 + 24}{62} = 83.87\%$$ 1✓2 1✓2

$$H_{max} = \max(34/62, 28/62) = 54.84\%$$ 1✓2 1✓2

$$H_{prop} = (\frac{34}{62})^2 + (\frac{28}{62})^2 = 50.47\%$$ 1✓2 1✓2

Hit-rate is greater than both H_{max} and H_{prop}, therefore this indicates a good hit-rate. 1✓2 1✓2

Q3-e) Evaluate the hit-rate for each category

$$Hit - rate(yes) = \frac{28}{34} = 82.4\%$$ 1✓2

$$Hit - rate(no) = \frac{24}{28} = 85.7\%$$ 1✓2

Both correct classification rates are similar and very good. 1✓2

Q4-a) Interpret the Classification Tree and define an appropriate decision rule for selecting a positive return.

(1) If CACL<=1.1694 & ROCS<=4.4486 & WCTA<= - 0.3326, then classify as Not Fail 1✓2
(2) If CACL<=1.1694 & ROCS<=4.4486 & WCTA> - 0.3326, classify as Fail 1✓2
(3) If CACL<=1.1694 & ROCS>4.4486, then classify as NotFail 1✓2
(4) If CACL>1.1694 & CLTA<=0.70635 & Sales <=3091.5, then classify as Fail 1✓2
(5) If CACL>1.1694 & CLTA<=0.70635 & Sales >3091.5, then classify as Not Fail 1✓2
(6) If CACL>1.1694 & CLTA>0.70635, then classify as Fail 1✓2

Q4-b)

Firm	SALES	ROCS	CLTA	CACL	WCTA	FAIL
2	16149	-1.07	1.22	0.62	-0.46	0

1✓2 1✓2 1✓2

CACL=0.62 < 1.1694 & ROCS = -1.07 <4.4486 & WCTA = -0.46 < -0.3326

Therefore classify as Not Fail. 1✓2

Q4-c)

1✓2 1✓2 1✓2

$$DI_1 = 1 - ((\frac{29}{60})^2 + (\frac{31}{60})^2) = 0.4994444$$

OR

$$DI_1 = 1 - ((\frac{30}{60})^2 + (\frac{30}{60})^2) = 0.5$$

The variable is chosen according to the reduction in the DI. The variable that creates the maximum reduction in the index is chosen for splitting the node. 1✓2

Q4-d)

Bonsai techniques check the several stopping criteria before letting the tree grow fully. 1✓2

Pruning techniques let the grow fully and then start pruning the tree. 1✓2

Q4-d)

Classification Table

		Predicted Groups		
		Fail	NotF	Total
Observed Groups	Fail	21+2+2=25	29-25=4	29
	NotF	31-28=3	2+4+22=28	31
	Total	28	32	60

1✓2 (totals)

OR

Classification Table

		Predicted Groups		
		Fail	NotF	Total
Observed Groups	Fail	21+2+2=25	1+1+3=5	30
	NotF	2+0+0=2	2+4+22=28	30
	Total	27	33	60

1✓2 (totals)

OR

Classification Table

		Predicted Groups		
		Fail	NotF	Total
Observed Groups	Fail	21+2+2=25	1+1+3=5	29
	NotF	2+0+0=2	2+4+22=28	31
	Total	27	33	60

1✓2 (totals)