STOCKHOLM UNIVERSITY
Department of Statistics
Ellinor Fackle-Fornius

EXAM IN MULTIVARIATE METHODS
September 30 2019

Time: 5 hours

Allowed aids: Pocket calculator, language dictionary.

The exam consists of five questions. To score maximum points on a question solutions need to be clear, detailed and well motivated.

Results will be announced no later than October 14.

Question 1. (16 points)

a) Describe the different measurement scales and give one example of each.

b) Describe dependence/interdependence methods and give two examples of each.

Question 2. (16 points)

Forty engineers were given six tests and measurements were taken on the following six variables

x_1 = intelligence $\quad x_4$ = dotting
x_2 = form relations $\quad x_5$ = sensory motor coordination
x_3 = dynamometer $\quad x_6$ = perseveration

A principal components analysis (PCA) was performed on the correlation matrix

$$\mathbf{R} = \begin{pmatrix} 1.00 & 0.19 & 0.10 & 0.12 & -0.05 & 0.39 \\ 0.19 & 1.00 & -0.12 & -0.27 & -0.24 & -0.07 \\ 0.10 & -0.12 & 1.00 & 0.29 & -0.16 & 0.07 \\ 0.12 & -0.27 & 0.29 & 1.00 & -0.19 & 0.33 \\ -0.05 & -0.24 & -0.16 & -0.19 & 1.00 & -0.15 \\ 0.39 & -0.07 & 0.07 & 0.33 & -0.15 & 1.00 \end{pmatrix}.$$

The eigenvalues of $\mathbf{R}$ were found to be

$$\lambda = \begin{pmatrix} 1.775 & 1.354 & 1.073 & 0.815 & 0.531 & x \end{pmatrix}$$

and the correlations between the first two PCs and the variables are given in the following table.

	rx_1	x_2	x_3	x_4	x_5	x_6
PC1	0.54	-0.13	0.51	0.72	-0.42	0.71
PC2	-0.46	-0.87	0.25	0.37	0.41	-0.12

a) Calculate the last eigenvalue (marked with an x).

b) Calculate the weights (w) for the first PC.

c) Which variable(s) are influential when forming the first two PCs.

d) When is PCA most useful, when the original variables are nearly uncorrelated or highly correlated? Explain why.

Question 3. (16 points)

Consider again the Engineer data described in Question 1. An exploratory factor analysis was now performed. The analysis resulted in a two factor solution with the rotated (rounded) pattern loadings presented in the table below. The two factors are assumed to be orthogonal.

	F_1	F_2
x_1	0.14	0.69
x_2	-0.64	0.61
x_3	0.56	0.12
x_4	0.80	0.16
x_5	-0.07	-0.58
x_6	0.49	0.54

Based on the reported results compute:

a) The communalities.

b) The proportion of variance explained by each factor.

c) The missing values (x's) in the following residual correlation matrix.

$$\begin{pmatrix} 0.5043 & -0.1413 & -0.0612 & -0.1024 & 0.3600 & -0.0512 \\ -0.1413 & 0.2183 & 0.1652 & 0.1444 & 0.0690 & -0.0858 \\ -0.0612 & 0.1652 & 0.6720 & -0.1772 & -0.0512 & -0.2692 \\ -0.1024 & 0.1444 & -0.1772 & 0.3344 & -0.0412 & -0.1484 \\ x & x & -0.0512 & -0.0412 & 0.6587 & 0.1975 \\ x & -0.0858 & -0.2692 & -0.1484 & 0.1975 & 0.4683 \end{pmatrix}.$$

d) RMSR.

Question 4. (16 points)

Prices (in dollars per 12 pack of cans) and Quality ratings on a 10-point scale (where 1 is worst and 10 is best quality) for six brands of beer are presented in the following table.

Brand	Price	Quality
A	7.89	10
B	4.79	4
C	7.65	9
D	6.39	7
E	4.50	3
F	6.25	6

a) Compute a distance matrix consisting of squared Euclidean distances.

b) Use the Centroid method to perform a hierarchical clustering of the brands. Stop when you have obtained 3 clusters.

c) Describe if there are any disadvantages of hierarchical clustering methods.

Question 5. (16 points)

To construct a procedure for detecting potential hemophilia A carriers, blood samples were assayed for two groups of women and measurements on the two variables,

$$X_1 = \log_{10}(\text{AHF Activity})$$
$$X_2 = \log_{10}(\text{AHF-like antigen}),$$

were recorded. "AHF" denotes antihemophilic factor. The first group consists of women who did not carry the hemophilia gene, the *noncarriers* group. The second group consists of known hemophilia A carriers, the so called *obligatory carriers*. The data set contains $n_1 = 30$ women in the *noncarrier* group and $n_2 = 45$ women in the *obligatory carriers* group. We have the sample means

$$\bar{\mathbf{x}}_1 = \begin{pmatrix} -0.1349 \\ -0.3079 \end{pmatrix}, \quad \bar{\mathbf{x}}_2 = \begin{pmatrix} -0.0779 \\ -0.0060 \end{pmatrix}$$

and the inverse of the pooled sample covariance matrix is given by

$$\mathbf{S}_{\text{pooled}}^{-1} = \begin{pmatrix} 86.09 & -61.49 \\ -61.49 & 90.20 \end{pmatrix}.$$

a) Calculate Fisher's linear discriminant function for this data set.

b) Classification was performed with the derived linear discriminant function and resulted in the following confusion matrix

		Predicted group	
		Noncarrier	*Obligatory carrier*
Actual group	*Noncarrier*	26	4
	Obligatory carrier	7	38

Calculate the accuracy, specificity and sensitivity of this classification, assuming the event is being a *"noncarrier"*.

c) The means of the discriminant scores are -1.2723 for the *noncarriers* group and -5.8466 for the *obligatory carriers* group. What would the classification be of a woman with measurements $x_1 = -0.0056$ and $x_2 = -0.1657$ assuming the prior probablilities of *noncarriers* is $\frac{3}{4}$ and that of *obligatory carriers* is $\frac{1}{4}$?

Formula Sheet for the Exam in Multivariate Methods

Vectors and matrices

- Length of a vector $\mathbf{a} = (a_1, a_2, \ldots, a_p)$

$$\|\mathbf{a}\| = \sqrt{a_1^2 + a_2^2 + \cdots + a_p^2}$$

- Determinant of a 2×2 matrix $\mathbf{A}$

$$\det(\mathbf{A}) = |\mathbf{A}| = a_{11}a_{22} - a_{12}a_{21}$$

- Inverse of a 2×2 matrix $\mathbf{A}$

$$\mathbf{A}^{-1} = \frac{1}{a_{11}a_{22} - a_{12}a_{21}} \begin{pmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{pmatrix}$$

- Eigenvalues are the roots of the characteristic equation

$$\det(\mathbf{A} - \lambda\mathbf{I}) = \mathbf{0}$$

For each eigenvalue the solution to

$$(\mathbf{A} - \lambda\mathbf{I})\,\mathbf{x} = \mathbf{0}$$

gives the associated eigenvector $\mathbf{x}$

Distances

- Euclidean

$$D_{ik} = \sqrt{\sum_{j=1}^{p} (x_{ij} - x_{kj})^2}$$

- Statistical

$$SD_{ik} = \sqrt{\sum_{j=1}^{p} \left(\frac{x_{ij} - x_{kj}}{s_j} \right)^2}$$

- Mahalanobis

$$MD_{ik} = \sqrt{(\mathbf{x}_i - \mathbf{x}_k)^T \mathbf{S}^{-1} (\mathbf{x}_i - \mathbf{x}_k)}$$

For $p = 2$

$$MD_{ik} = \sqrt{\frac{1}{1 - r^2} \left[\frac{(x_{i1} - x_{k1})^2}{s_1^2} + \frac{(x_{i2} - x_{k2})^2}{s_2^2} - \frac{2r (x_{i1} - x_{k1}) (x_{i2} - x_{k2})}{s_1 s_2} \right]}$$

Mean-correction and covariance

- Mean-corrected data

$$\underset{(n \times p)}{\mathbf{X}_m} = \{x_{ij}\} = \{X_{ij} - \bar{X}_j\}$$

- Covariance

$$\underset{(p \times p)}{\mathbf{S}} = \{s_{ij}\} = \left\{ \frac{\sum_{i=1}^{n} x_{ij} x_{ik}}{n - 1} \right\} = \frac{\mathbf{SSCP}}{df} = \frac{1}{n - 1} \mathbf{X}_m^T \mathbf{X}_m$$

Group Analysis

- Total sum of squares and cross products

$$\mathbf{SSCP}_{\text{total}} = \mathbf{SSCP}_{\text{within}} + \mathbf{SSCP}_{\text{between}}$$

- Pooled within-group sum of squares and cross products

$$\mathbf{SSCP}_{\text{within}} = \sum_{\ell=1}^{g} \mathbf{SSCP}_{\ell}$$

- Pooled covariance matrix

$$\mathbf{S}_{\text{pooled}} = \frac{\mathbf{SSCP}_{\text{within}}}{n-g}$$

- Between-group sum of squares and cross products

$$\mathbf{SSCP}_{\text{between}} = \mathbf{SSCP}_{\text{total}} - \mathbf{SSCP}_{\text{within}}$$

For $g = 2$ groups

$$\mathbf{SSCP}_{\text{between}} = \frac{n_1 n_2}{n_1 + n_2} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T$$

Factor Analysis

- For the two-factor model

$$Var(x) = \lambda_1^2 + \lambda_2^2 + Var(\epsilon) + 2\lambda_1\lambda_2\phi$$

$$Cor(x, \xi_1) = \lambda_1 + \lambda_2\phi$$

$$Cor(x_j, x_k) = \lambda_{j1}\lambda_{k1} + \lambda_{j2}\lambda_{k2} + (\lambda_{j1}\lambda_{k2} + \lambda_{j2}\lambda_{k1})\,\phi$$

- RMSR for EFA

$$RMSR = \sqrt{\frac{\sum_{i=1}^{p}\sum_{j=i+1}^{p} res_{ij}^2}{p(p-1)/2}}$$

- RMSR for CFA

$$RMSR = \sqrt{\frac{\sum_{i=1}^{p}\sum_{j=i}^{p} (s_{ij} - \hat{\sigma}_{ij})^2}{p(p+1)/2}}$$

Two-Group Discriminant Analysis

- Maximize

$$\lambda = \frac{\gamma^T \mathbf{B} \gamma}{\gamma^T \mathbf{W} \gamma}$$

- Fisher's linear discriminant function

$$\gamma^T = (\mu_1 - \mu_2)^T \mathbf{\Sigma}^{-1}$$

- Wilks' Λ

$$\Lambda = \frac{|\mathbf{SSCP}_w|}{|\mathbf{SSCP}_t|}$$

$$F = \left(\frac{1-\Lambda}{\Lambda}\right)\left(\frac{n_1 + n_2 - p - 1}{p}\right) \sim F\left(p, n_1 + n_2 - p - 1\right)$$

- Classification based on decision theory: assign the observation to group 1 if

$$Z \geq \frac{\bar{Z}_1 + \bar{Z}_2}{2} + \ln\left[\frac{p_2 C(1|2)}{p_1 C(2|1)}\right]$$

Logistic regression

- Odds of the event $Y = 1$

$$odds = \frac{p}{1-p}$$

where

$$p = P(Y = 1)$$

- Probability of the event $Y = 1$ as a function of the explanatory variables

$$P(Y = 1|X_1, X_2, \ldots, X_k) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k)}}$$

Quadratic equation

- The roots of the quadratic equation $ax^2 + bx + c$

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

Department of Statistics

Stockholms universitet

Correction sheet

Date: 30/9 - 2019

Room: Uggleviksalen

Exam: Multivariate Methods

Course: Multivariate Methods

Anonymous code: 0001-DNW

☒ I authorise the anonymous posting of my exam, in whole or in part, on the department homepage as a sample student answer.

NOTE! ALSO WRITE ON THE BACK OF THE ANSWER SHEET

Mark answered questions

1	2	3	4	5	6	7	8	9	Total number of pages
X	X	X	X	X					5
Teacher's notes 16	16	16	15	16					

Points	Grade	Teacher's sign.
79		

SU, DEPARTMENT OF STATISTICS

Room: Vaggeviksalen Anonymous code: 0001-DNW Sheet number: 1

1. a) The nominal scale is a non-metric scales where an observation belongs to one "group" and the different "groups" cannot be ranked in someway. An example of a nominal scale variable would be gender (i.e. male/female).
The ordinal scale is a non-metric scale where a variable can be ranked in someway. An example of an ordinal scale variable would be opinion which can be ranked by how much one agrees about a certain statement ('strongly disagree/disagree/agree/strongly agree').
0 1 2 3
The interval scale is a metric scale without an absolute 0 value. An example of a interval scale variable would be degre celcius (°C). It's metric but without an absolute 0.
The ratio scale is a metric scale with an absolute 0 value. An example of a ratio scale variable would be degrees Kelvin (K). It is metric and has an absolute 0.

8

b) In interdependence methods all variables are treated equally, you don't divide variables up in dependent and independent variables. Two examples of interdependence methods are principal component analysis (PCA) and cluster analysis (CA). (How all variables interact with each other.)

In dependence methods you have certain dependent variables (y in regression) and certain independent variables (x in regression). Two examples of dependence methods are Logistic regression (LR) and Discriminant analysis (DA). (How certain variables interact with certain other variables to yield values (y) that can be used for further analysis.)

8

16

SU, DEPARTMENT OF STATISTICS

Room: Ugglevikssalen Anonymous code: 0001-DVW Sheet number: 2

2. a) The variance of a standardized variable is 1. As we use correlation matrix, the variables are standardized, ie. the variance in total is equal to p=6. This can also be seen if we sum the diagonal in the R-matrix (1+1+1+1+1+1=6). R

As the sum of the eigenvalues are equal to total variance in the original variables we get:

$1,775+1,354+1,073+0,815+0,531+x=6$ R

$\Rightarrow \quad 5,548+x=6$

$\Rightarrow \quad x=0,452$

The last eigenvalue is 0,452. R

4

b) The last table gives us correlations between the first two PC's and the variables. This is also known as loading. It's given by:

$l_{ij} = \frac{w_{ij}\sqrt{\pi_i}}{s_j}$, where w_{ij} is the weight for variable j on PC i. π_i is the eigenvalue for PC i and s_j is the s.d. for variable j. R

As stated in a), the s.d. is in this case 1 for each variable since we use standardized variables. The loadings formula then becomes: R

$l_{ij} = w_{ij}\sqrt{\pi_i}$ Using this we get:

Weight of Var X_1 on PC 1:

$0,54 = w_{11}\sqrt{1,775} \Rightarrow w_{11} = \frac{0,54}{\sqrt{1,775}} \approx 0,4053$

Weight of Var X_2 on PC 1:

$-0,13 = w_{12}\sqrt{1,775} \Rightarrow w_{12} = \frac{-0,13}{\sqrt{1,775}} \approx -0,0976$

Weight of Var X_3 on PC 1:

$0,51 = w_{13}\sqrt{1,775} \Rightarrow w_{13} = \frac{0,51}{\sqrt{1,775}} \approx 0,3828$

→ Continue

Weight of Var X_4 on PC 1:

$$0,72 = w_{1,4}\sqrt{1,775} \Rightarrow w_{1,4} = \frac{0,72}{\sqrt{1,775}} \approx 0,5404$$

Weight of Var X_5 on PC 1:

$$-0,42 = w_{1,5}\sqrt{1,775} \Rightarrow w_{1,5} = \frac{-0,42}{\sqrt{1,775}} \approx -0,3152$$

Weight of Var X_6 on PC 1:

$$0,71 = w_{1,6}\sqrt{1,775} \Rightarrow w_{1,6} = \frac{0,71}{\sqrt{1,775}} \approx 0,5329$$

R

5

c) One decides this using the loadings given in the last table. For loadings with an absolute value $> 0,5$, one can say that those variables are influential when creating that PC.

For PC 1; variables X_1 (0,54), X_3 (0,51), X_4 (0,72) and X_6 (0,71) are all loading high. These are therefore influential when forming the first PC.

For PC 2; variable X_2 (-0,87) is loading high. This variable is therefore influential when forming the second PC.

3

d) PCA is most useful when the original variables are highly correlated. This is because PCA is considered to be a data reduction technique in which we form PC's that are linear combinations of the original variables. If there is no correlation between original variables it is not possible to create linear combinations of variable (since correlation is a measure of linear relationship between variables). However, if we have high correlation instead, it is easy to create new PC's and the data-reduction will be more clear.

Therefore, PCA is most useful when dealing with original variables that are highly correlated.

4

16

SU, DEPARTMENT OF STATISTICS

Room: Ugglevikssalen Anonymous code: 0001-DNW Sheet number: 3

$2 r_{1,1} r_{2,1} \phi_{1,2} = 0$

3. a) For the orthogonal case ($\phi_{1,2} = 0$) communalities are as follows:

var	Communality F_1	Communality F_2	Total Communality
X_1	$r_{1,1}^2 = 0{,}14^2 = 0{,}0196$	$r_{1,2}^2 = 0{,}69^2 = 0{,}4761$	$r_{1,1}^2 + r_{1,2}^2 = 0{,}4957$
X_2	$(-0{,}64)^2 = 0{,}4096$	$0{,}61^2 = 0{,}3721$	$0{,}4096 + 0{,}3721 = 0{,}7817$
X_3	$0{,}56^2 = 0{,}3136$	$0{,}12^2 = 0{,}0144$	$0{,}3280$
X_4	$0{,}8^2 = 0{,}64$	$0{,}16^2 = 0{,}0256$	$0{,}6656$
X_5	$(-0{,}07)^2 = 0{,}0049$	$(-0{,}58)^2 = 0{,}3364$	$0{,}3413$
X_6	$0{,}49^2 = 0{,}2401$	$0{,}54^2 = 0{,}2916$	$0{,}5317$
Σ	$1{,}6278$	$1{,}5162$	$3{,}1440$

R

3

b) For the orthogonal case ($\phi_{1,2} = 0$); $(\mathrm{Cor}(X_i, F_1))^2 = (r_{i,1} + r_{i,2}\phi_{1,2})^2 = r_{i,1}^2$

i.e. from a) we get that For F_2: $r_{i,2}^2$

F_1 communality sums to $1{,}6278$ R

and F_2 communality sums to $1{,}5162$

for a total communality of $3{,}1440$ ($1{,}6278 + 1{,}5162 = 3{,}1440$).

In the orthogonal case we have, as I showed, the communalities from a) is the same as shared variance: R

We get that the proportion of variance explained by factor 1 (F_1) is:

$\frac{1{,}6278}{3{,}1440} \approx 0{,}5177$ i.e. $51{,}77\%$ R

The proportion of variance explained by factor 2 (F_2) is then:

$1 - 0{,}5177 = 0{,}4823$ i.e. $48{,}23\%$ R

3

c) We need: $\mathrm{Cor}(X_1, X_5) = 0{,}14 \cdot (-0{,}07) + 0{,}69 \cdot (-0{,}58) = -0{,}4100$

$\mathrm{Cor}(X_1, X_6) = 0{,}14 \cdot 0{,}49 + 0{,}69 \cdot 0{,}54 = 0{,}4412$

$\mathrm{Cor}(X_2, X_5) = (0{,}64) \cdot (-0{,}07) + 0{,}61 \cdot (-0{,}58) = -0{,}3090$ R

For the orthogonal case:

$\mathrm{Cor}(X_j, X_k) = r_{j,1} r_{k,1} + r_{j,2} r_{k,2} + (r_{j,1} r_{k,2} + r_{j,2} r_{k,1}) \phi_{1,2}^{0} = r_{j,1} r_{k,1} + r_{j,2} r_{k,2}$ Correlation

R

$R - \hat{R} = res$, for our three cases we get

$res_{x_1,x_5} = +0,05 - (-0,41) = 0,36$

$res_{x_1,x_6} = 0,39 - 0,4412 = -0,0512$

$res_{x_2,x_5} = -0,24 - (-0,309) = 0,069$

} These are the missing R values in the residual correlation matrix. R 5

d) EFA $RMSR = \sqrt{\frac{\sum_{i=1}^{p}\sum_{j=i+1}^{p} res_{ij}^2}{p(p-1)/2}} = \sqrt{\frac{0,39589954}{15}} \approx 0,1625$ R

$p=6 \Rightarrow p(p-1)/2 = 6 \cdot 5/2 = 15$

$\sum_{i=1}^{p}\sum_{j=i+1}^{p} res_{ij}^2 = (-0,1413)^2 + (-0,0612)^2 + 0,36^2 + (-0,0512)^2 +$
$+ 0,1652^2 + 0,1444^2 + 0,069^2 + (-0,0858)^2 +$
$+ (-0,1772)^2 + (-0,0512)^2 + (-0,2692)^2 +$
$+ (-0,0412)^2 + (-0,1484)^2 + 0,1975^2 = 0,39589954$ R

5

16

4. a)

$$D = \begin{array}{c|cccccc} & A & B & C & D & E & F \\ A & 0 & & & & & \\ B & 6{,}75 & 0 & & & & \\ C & 1{,}03 & 5{,}76 & 0 & & & \\ D & 3{,}55 & 3{,}40 & 2{,}36 & 0 & & \\ E & 7{,}78 & 1{,}04 & 6{,}78 & 4{,}42 & 0 & \\ F & 4{,}32 & 2{,}48 & 3{,}31 & 1{,}01 & 3{,}47 & 0 \end{array}$$

symmetry

$$D_{AB} = \sqrt{\sum_{j=1}^{p} (x_{ij} - x_{kj})^2} = \sqrt{(4{,}79 - 7{,}89)^2 + (4-10)^2} \approx 6{,}75$$

$$D_{AC} = \sqrt{(7{,}65 - 7{,}89)^2 + (9-10)^2} \approx 1{,}03$$

$$D_{AD} = \sqrt{(6{,}39 - 7{,}89)^2 + (7-10)^2} \approx 3{,}35$$

$$D_{AE} = \sqrt{(4{,}50 - 7{,}89)^2 + (3-10)^2} \approx 7{,}78$$

$$D_{AF} = \sqrt{(6{,}25 - 7{,}89)^2 + (6-10)^2} \approx 4{,}32$$

squared Euclidean

$$D_{BC} = \sqrt{(7{,}65 - 4{,}79)^2 + (9-4)^2} \approx 5{,}76$$

$$D_{BD} = \sqrt{(6{,}39 - 4{,}79)^2 + (7-4)^2} = 3{,}40$$

$$D_{BE} = \sqrt{(4{,}50 - 4{,}79)^2 + (3-4)^2} \approx 1{,}04$$

$$D_{BF} = \sqrt{(6{,}25 - 4{,}79)^2 + (6-4)^2} \approx 2{,}48$$

$$D_{CD} = \sqrt{(6{,}39 - 7{,}65)^2 + (7-9)^2} \approx 2{,}36$$

$$D_{CE} = \sqrt{(4{,}50 - 7{,}65)^2 + (3-9)^2} \approx 6{,}78$$

$$D_{CF} = \sqrt{(6{,}25 - 7{,}65)^2 + (6-9)^2} \approx 3{,}31$$

$$D_{DE} = \sqrt{(4{,}50 - 6{,}39)^2 + (3-7)^2} \approx 4{,}42$$

$$D_{DF} = \sqrt{(6{,}25 - 6{,}39)^2 + (6-7)^2} \approx 1{,}01$$ - shortest

$$D_{EF} = \sqrt{(6{,}25 - 4{,}50)^2 + (6-3)^2} \approx 3{,}47$$

4

c) There are disadvantages of hierarchical clustering methods. One is that there is no possibility to reassign a subject even if it is wrongly placed in a cluster. Another is that ties can lead to multiple solution. A third is that outliers can have a big effect. Those are 3 disadvantages with hierarchical clustering methods.

3

for d)

b) Step 1: Merge brand D and F, as these brands yields the shortest Euclidean distance (1,01), to form cluster (DF) R

Centroid: $m_{(DF)} = \left(\frac{6,39+6,25}{2}, \frac{7+6}{2}\right) = (6,32, 6,5)$ R

New distance matrix

	(DF)	A	B	C	E
(DF)	0				
A	3,84	0			
B	2,93	6,75	0		
C	2,83	1,03	5,76	0	
E	3,94	7,78	1,04	6,78	0

$D_{(DF)A} = \sqrt{(6,32-7,89)^2+(6,5-10)^2} \approx 3,84$

$D_{(DF)B} = \sqrt{(6,32-4,79)^2+(6,5-4)^2} \approx 2,93$

$D_{(DF)C} = \sqrt{(6,32-7,65)^2+(6,5-9)^2} \approx 2,83$

$D_{(DF)E} = \sqrt{(6,32-4,50)^2+(6,5-3)^2} \approx 3,94$ OK

Step 2: Merge brand A and C, as this is the shortest Euclidean distance, to form cluster (AC). R

Centroid: $m_{AC} = \left(\frac{7,89+7,65}{2}, \frac{10+9}{2}\right) = (7,77, 9,5)$ R

New distance matrix

	(DF)	(AC)	B	E
(DF)	0			
(AC)	3,33	0		
B	2,93	6,26	0	
E	3,94	7,28	1,04	0

$D_{(AC)(DF)} = \sqrt{(7,77-6,32)^2+(9,5-6,5)^2} \approx 3,33$

$D_{(AC)B} = \sqrt{(7,77-4,79)^2+(9,5-4)^2} \approx 6,26$

$D_{(AC)E} = \sqrt{(7,77-4,50)^2+(9,5-3)^2} \approx 7,28$

Step 3: Merge brand B and E, as this is the shortest Euclidean distance, to form cluster (BE). R

Now, I have obtained 3 clusters; (DF), (AC) and (BE), therefore this is the end of the clustering process as described in the question.

8

15

SU, DEPARTMENT OF STATISTICS

Room: Ugglerikssalen Anonymous code: 0001-DNW Sheet number: 5

5. Group 1: non-carriers, $n_1 = 30$ Group 2: obligatory carriers, $n_2 = 45$

a) Fisher's linear discriminant function is given by:

$\delta^T = (\mu_1 - \mu_2)^T \Sigma^{-1}$, and for a sample it is given by:

$$\hat{\delta}^T = (\bar{x}_1 - \bar{x}_2)^T S_{pooled}^{-1} = \left(\begin{pmatrix} -0,1349 \\ -0,3019 \end{pmatrix} - \begin{pmatrix} -0,0779 \\ -0,0060 \end{pmatrix}\right)^T \begin{pmatrix} 86,09 & -61,49 \\ -61,49 & 90,20 \end{pmatrix} =$$

$$= \begin{pmatrix} -0,0570 \\ -0,3019 \end{pmatrix}^T \begin{pmatrix} 86,09 & -61,49 \\ -61,49 & 90,20 \end{pmatrix} = (-0,0570 \quad -0,3019) \begin{pmatrix} 86,09 & -61,49 \\ -61,49 & 90,20 \end{pmatrix} =$$

$$= ((-0,0570) \cdot 86,09 + (-0,3019) \cdot (-61,49) \quad (-0,0570) \cdot (-61,49) + (-0,3019) \cdot 90,20) =$$

$$= (13,6567\,01 \quad -23,726450) \quad R$$

Which gives $\hat{\delta} = \begin{pmatrix} 13,656701 \\ -23,726450 \end{pmatrix}$ 5

b) Assuming the event is non-carrier:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} = \frac{26+38}{75} = \frac{64}{75} \approx 0,8533 \quad R$$

$$\text{Sensitivity} = \frac{TP}{TP+FN} = \frac{26}{26+4} = \frac{26}{30} \approx 0,8667 \quad R$$

$$\text{Specificity} = \frac{TN}{TN+FP} = \frac{38}{38+7} = \frac{38}{45} \approx 0,8444 \quad R \quad 6$$

c) $\bar{z}_1 = -1,2723$ $P_1 = 0,75$

$\bar{z}_2 = -5,8466$ $P_2 = 1-0,75 = 0,25$

Cut-off value: $z_c = \frac{\bar{z}_1 + \bar{z}_2}{2} + \ln\left(\frac{P_2}{P_1}\right) =$

$$= \frac{-1,2723 + (-5,8466)}{2} + \ln\left(\frac{0,25}{0,75}\right) =$$

$$\approx \frac{-7,1189}{2} - 1,0986 = -8,3175$$

$x_1 = -0,0056$

$x_2 = -0,1657$

$\hat{z}_{x_1=-0,0056, x_2=-0,1657} = 13,656701 \cdot (-0,0056) + (-23,726450) \cdot (-0,1657) \approx 3,8550$ R

As $z = 3,8550 > -8,3175 = z_c$, a women with $x_1 = -0,0056$ and $x_2 = -0,1657$ would be assigned to group 1, i.e she would be classified as a non-carrier. R

5

16