



# Strathmore

---

## UNIVERSITY

STRATHMORE INSTITUTE OF MATHEMATICAL SCIENCES  
MASTER OF SCIENCE IN STATISTICAL SCIENCE

END OF SEMESTER EXAMINATION  
STA 8301: MULTIVARIATE STATISTICAL ANALYSIS  
AUGUST 17, 2023 17:00 - 20:00

Answer Question ONE and TWO other questions. You must show *all* work to receive *any credit*.

### Question ONE (30 marks)

You are working as a statistician doing oversight for a food-safety and nutrition board, as part of a governmental investigation. The agency has selected 43 types of cereal produced by three companies (General Mills (G), Kellogg's (K), and Quaker (Q)). Eight numerical nutritional characteristics have been measured for each cereal. These eight variables are *calories*, *protein*, *fat*, *sodium*, *fiber*, *carbohydrates*, and *potassium*. In addition, the data file gives the name of each cereal and the company that produced the cereal. Use the output from the analysis attached to answer some of the following questions:

- Describe any notable associations/relationships among the variables (**8 marks**).
- Write down the expressions for the first few principal components that capture a reasonable variation in the data set. Interpret these principal components in the context of the data set (**8 marks**).
- Briefly describe model-based clustering of multivariate data. What is the best clustering model, based on this analysis? How would you characterize the groups? What proportion of the cereals were produced by Kellogg's (**8 marks**).
- Describe the Wilks' Lambda statistic in MANOVA. Based on this statistic, would you say that the three companies are the same in terms of their average values of the eight variables? Why? (**6 marks**).

### Question TWO (15 marks)

The Bumpus Bird dataset (Bumpus, 1898) consists of five body measurements on 49 female sparrows:  $X_1$  = total length,  $X_2$  = alar length,  $X_3$  = length of beak and head,  $X_4$  = length of humerus, and  $X_5$  = length of keel and sternum (all in millimetres). Use the attached analysis results to answer the following questions:

- (a) Write down the first 2 principal components (PCs) and interpret them (**5 marks**).
- (b) What seems to be a reasonable number of PCs to use? Why? (**5 marks**).
- (c) Identify any potential outlier sparrows by their observation number (**5 marks**).

**Question THREE (15 marks)**

The matrix below shows the correlations between ratings on nine statements about pain made by 123 people suffering from extreme pain. Each statement was scored on a scale from 1 to 6, ranging from agreement to disagreement.

$$Pain.Corr = \begin{bmatrix} 1 & -0.04 & 0.61 & 0.45 & 0.03 & -0.29 & -0.3 & 0.45 & 0.3 \\ -0.04 & 1 & -0.07 & -0.12 & 0.49 & 0.43 & 0.3 & -0.31 & -0.17 \\ 0.61 & -0.07 & 1 & 0.59 & 0.03 & -0.13 & -0.24 & 0.59 & 0.32 \\ 0.45 & -0.12 & 0.59 & 1 & -0.08 & -0.21 & -0.19 & 0.63 & 0.37 \\ 0.03 & 0.49 & 0.03 & -0.08 & 1 & 0.47 & 0.41 & -0.14 & -0.24 \\ -0.29 & 0.43 & -0.13 & -0.21 & 0.47 & 1 & 0.63 & -0.13 & -0.15 \\ -0.3 & 0.3 & -0.24 & -0.19 & 0.41 & 0.63 & 1 & -0.26 & -0.29 \\ 0.45 & -0.31 & 0.59 & 0.63 & -0.14 & -0.13 & -0.26 & 1 & 0.4 \\ 0.3 & -0.17 & 0.32 & 0.37 & -0.24 & -0.15 & -0.29 & 0.4 & 1 \end{bmatrix}$$

The nine pain statements were as follows:

1. Whether or not I am in pain in the future depends on the skills of the doctors.
2. Whenever I am in pain, it is usually because of something I have done or not done.
3. Whether or not I am in pain depends on what the doctors do for me.
4. I cannot get any help for my pain unless I go to seek medical advice.
5. When I am in pain I know that it is because I have not been taking proper exercise or eating the right food.
6. People's pain results from their own carelessness.
7. I am directly responsible for my pain.
8. Relief from pain is chiefly controlled by the doctors.
9. People who are never in pain are just plain lucky.

Use the results of the maximum likelihood factor analysis provided to:

- (a) select the necessary number of common factors (**7 marks**).

- (b) interpret the results of the factor solution after the varimax rotation (**8 marks**).

**Question FOUR (15 marks)**

The High School and Beyond (HSB) dataset contains demographic information and standardized test scores for 200 high school students. Suppose our goal is to compare mean vectors (where the variables are the scores on *read*, *write*, *math*, *science*, and *socst*) among the different levels of *ses* (high, middle, and low socioeconomic classes). Using the analysis of the dataset attached, answer the following questions:

- (a) Conduct the MANOVA F-test using Wilks' lambda to test for a difference in (*read*, *write*, *math*, *science*, *socst*) mean vectors across the three *ses* classes. Use a 0.05 significance level and give the *p*-value of the test (**5 marks**).
- (b) Check to see whether the multivariate normality assumption of your test is met. Do you believe your inference is valid? (**5 marks**).
- (c) Examine the sample mean vectors for each group and comment on the differences among the groups in terms of the specific variables. Which approach would help determine pairwise significant differences in the sample mean vectors (**5 marks**).

**Question FIVE (15 marks)**

The SIDS dataset was collected in an investigation of sudden infant death syndrome (SIDS). The two groups here consist of 16 SIDS victims and 49 controls. The *Factor68* variable arises from spectral analysis of 24 hour recordings of electrocardiograms and respiratory movements made on each child. All the infants have a gestational age of 37 weeks or more and were regarded as full term. Use the results of the analysis provided to answer the following questions:

- (a) Use Fisher's linear discriminant analysis (LDA) to classify the children into two groups, based on all four variables. Write down the discriminant function and obtain a plug-in estimate of the misclassification rate (**5 marks**).
- (b) Perform the classification of the children using the logistic regression approach and obtain a plug-in estimate of the misclassification rate. Write down the logistic regression model used (**5 marks**).
- (c) Which of the two approaches is preferable, based on the estimated misclassification rates? Discuss an approach for estimating the misclassification rate that may improve this comparison (**5 marks**).

## Results of Data Analysis (*R* output) for Answering Questions

### Question ONE:

```
>cor.mat <- cor(cereal.new2)
```

```
>round(cor.mat, 3)
```

	Calories	Protein	Fat	Sodium	Fiber	Carbohydrates	Sugar	Potassium
Calories	1.000	0.033	0.388	0.337	-0.019	0.256	0.580	0.141
Protein	0.033	1.000	0.206	0.094	0.513	-0.076	-0.400	0.500
Fat	0.388	0.206	1.000	0.010	0.164	-0.323	0.187	0.312
Sodium	0.337	0.094	0.010	1.000	0.043	0.567	-0.049	0.114
Fiber	-0.019	0.513	0.164	0.043	1.000	-0.241	-0.034	0.929
Carbohydrates	0.256	-0.076	-0.323	0.567	-0.241	1.000	-0.315	-0.223
Sugar	0.580	-0.400	0.187	-0.049	-0.034	-0.315	1.000	0.081
Potassium	0.141	0.500	0.312	0.114	0.929	-0.223	0.081	1.000

```
>summary(cereal.pca, loadings=TRUE)
```

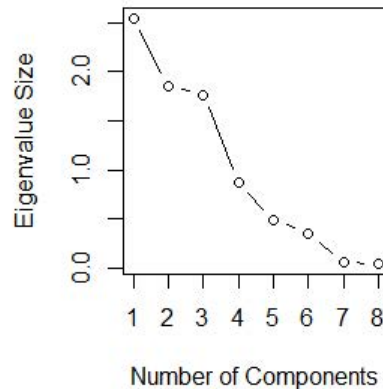
Importance of components:

	Comp.1	Comp.2	Comp.3	Comp.4
Standard deviation	1.5961107	1.3618964	1.3299737	0.9318172
Proportion of Variance	0.3184462	0.2318452	0.2211037	0.1085354
Cumulative Proportion	0.3184462	0.5502914	0.7713951	0.8799306
	Comp.5	Comp.6	Comp.7	
Standard deviation	0.7051956	0.59785320	0.245974321	
Proportion of Variance	0.0621626	0.04467856	0.007562921	
Cumulative Proportion	0.9420932	0.98677173	0.994334648	
	Comp.8			
Standard deviation	0.212891551			
Proportion of Variance	0.005665352			
Cumulative Proportion	1.000000000			

Loadings:

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8
Calories	0.114	0.657	0.135	0.100	0.454	0.136	0.469	0.289
Protein	0.421	-0.228	0.253	0.372	0.551	-0.420	-0.242	-0.171
Fat	0.316	0.302	-0.166	0.687	-0.432	0.251	-0.125	-0.209
Sodium	0.273	0.592		-0.502	-0.556			
Fiber	0.558	-0.135		-0.376	0.194	0.453	-0.525	
Carbohydrates	-0.238	0.113	0.632		0.132	0.535	-0.370	-0.297
Sugar	0.566	-0.362	-0.375	0.124	-0.254	-0.435	-0.371	
Potassium	0.583		-0.296	-0.116	0.211	-0.409	0.585	

**Scree Plot for Cereals Data**



```
>summary(best, parameters=TRUE)
```

```
-----  
Gaussian finite mixture model fitted by EM algorithm  
-----
```

Mclust VEV (ellipsoidal, equal shape) model with 2 components:

log-likelihood	n	df	BIC	ICL
-886.6437	43	82	-2081.706	-2081.802

Clustering table:

```
1 2  
29 14
```

Mixing probabilities:

	1	2
0.6733348	0.3266652	

Means:

	[,1]	[,2]
Calories	107.9439974	107.830668
Protein	1.9638542	3.498337
Fat	0.8261418	1.287171
Sodium	190.0554997	160.697055
Fiber	1.0491944	3.084180
Carbohydrates	14.9181546	12.890572
Sugar	8.1402648	6.500624
Potassium	56.7758155	141.396969

```
>table(cereal$Company, best$classification)
```

```

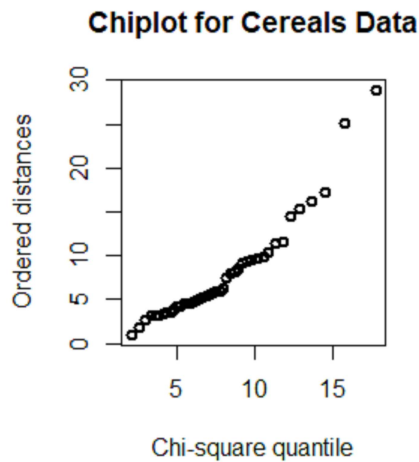
1 2
G 14 3
K 13 7
Q 2 4

```

```

>summary(cereal.manova, test="Wilks")
      Df  Wilks approx F num Df den Df  Pr(>F)
Company  2   0.2684  3.8372    16   66 5.368e-05 ***
Residuals 40
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```



**Question TWO:**

```
> summary(bump.pc, loadings=TRUE)
```

Importance of components:

	Comp.1	Comp.2	Comp.3	Comp.4
Standard deviation	1.9015726	0.7290433	0.62163056	0.5491498

Proportion of Variance	0.7231957	0.1063008	0.07728491	0.0603131
------------------------	-----------	-----------	------------	-----------

Cumulative Proportion	0.7231957	0.8294965	0.90678139	0.9670945
-----------------------	-----------	-----------	------------	-----------

Comp.5

Standard deviation	0.4056199
--------------------	-----------

Proportion of Variance	0.0329055
------------------------	-----------

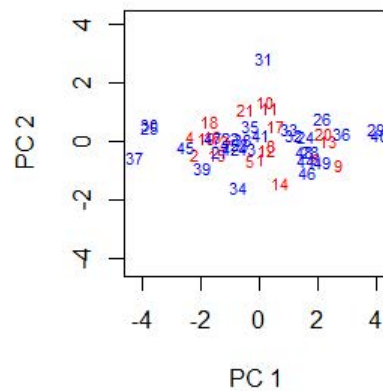
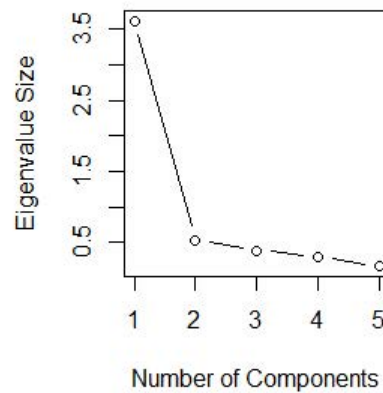
Cumulative Proportion	1.0000000
-----------------------	-----------

Loadings:

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5
--	--------	--------	--------	--------	--------

X1	0.452	0.690	0.420	0.374	
X2	0.462	-0.300	0.341	-0.548	-0.530
X3	0.451	-0.325	-0.454	0.606	-0.343
X4	0.471	-0.185	-0.411	-0.388	0.652
X5	0.398	0.876	-0.178		-0.192

**Scree Plot for Bumpbird Data**



**Question THREE:**

```
> factanal(covmat=pain.corr,factors=2, rotation="none", n.obs=123)
```

Call:

```
factanal(factors = 2, covmat = pain.corr, n.obs = 123, rotation = "none")
```

Uniquenesses:

```
[1] 0.539 0.695 0.318 0.448 0.579 0.356 0.464 0.436 0.766
```

Loadings:

	Factor1	Factor2
1	0.648	0.202
2	-0.358	0.421
3	0.724	0.396
4	0.691	0.273
5	-0.304	0.573
6	-0.511	0.619
7	-0.556	0.477
8	0.710	0.244
9	0.483	

	Factor1	Factor2
SS loadings	2.950	1.447
Proportion Var	0.328	0.161
Cumulative Var	0.328	0.489

Test of the hypothesis that 2 factors are sufficient.  
The chi square statistic is 60.41 on 19 degrees of freedom.  
The p-value is 3.33e-06

```
> factanal(covmat=pain.corr,factors=3, rotation="none", n.obs=123)
```

Call:

```
factanal(factors = 3, covmat = pain.corr, n.obs = 123, rotation = "none")
```

Uniquenesses:

```
[1] 0.404 0.518 0.336 0.455 0.499 0.171 0.496 0.239 0.754
```

Loadings:

	Factor1	Factor2	Factor3
1	0.607	0.297	0.374
2	-0.458	0.288	0.435
3	0.610	0.506	0.189
4	0.621	0.400	
5	-0.408	0.441	0.375
6	-0.677	0.591	-0.145
7	-0.626	0.331	
8	0.674	0.481	-0.275
9	0.446	0.169	-0.135

	Factor1	Factor2	Factor3
SS loadings	3.004	1.501	0.624
Proportion Var	0.334	0.167	0.069
Cumulative Var	0.334	0.501	0.570

Test of the hypothesis that 3 factors are sufficient.  
The chi square statistic is 19.2 on 12 degrees of freedom.  
The p-value is 0.0838

```
> factanal(covmat=pain.corr,factors=4, rotation="none", n.obs=123)
```

Call:

```
factanal(factors = 4, covmat = pain.corr, n.obs = 123, rotation = "none")
```

Uniquenesses:

```
[1] 0.433 0.600 0.297 0.482 0.169 0.005 0.536 0.005 0.731
```

Loadings:

	Factor1	Factor2	Factor3	Factor4
1	0.496	0.125	0.391	0.391
2	-0.494	0.375		
3	0.485	0.353	0.333	0.483
4	0.562	0.321	0.177	0.260
5	-0.412	0.259	0.738	-0.223
6	-0.748	0.660		
7	-0.594	0.284	-0.157	
8	0.752	0.655		
9	0.371	0.188	-0.100	0.293

	Factor1	Factor2	Factor3	Factor4
SS loadings	2.826	1.301	0.996	0.620
Proportion Var	0.314	0.145	0.111	0.069
Cumulative Var	0.314	0.459	0.569	0.638

Test of the hypothesis that 4 factors are sufficient.  
The chi square statistic is 9.72 on 6 degrees of freedom.  
The p-value is 0.137

```
> fact(r=pain.corr,method="norm",rotation="varimax",maxfactors=3)
```

\$'eigen.values'

```
[1] 3.442 1.930 0.942 0.754 0.515 0.444 0.422 0.326 0.225
```

\$method

```
[1] "multivariate normal - varimax rotation"
```

\$loadings

	Factor1	Factor2	Factor3
[1,]	0.649	-0.372	0.190
[2,]	-0.126	0.194	0.655

```
[3,] 0.794 -0.144 0.116
[4,] 0.725 -0.106 -0.092
[5,] 0.015 0.292 0.645
[6,] -0.083 0.825 0.377
[7,] -0.225 0.590 0.325
[8,] 0.815 0.062 -0.304
[9,] 0.437 -0.076 -0.221
```

\$communalities

```
[1] 0.596 0.482 0.664 0.545 0.501 0.829 0.504 0.761 0.246
```

\$importance

```
                Factor1 Factor2 Factor3
variance explained 2.507 1.331 1.291
percent explained 0.279 0.148 0.143
```

\$residuals

```
  Min. 1st Qu. Median  Mean 3rd Qu.  Max.
-0.082 -0.014  0.001 -0.001  0.018 0.066
```

**Question FOUR:**

```
> hsb.manova <- manova(cbind(read, write, math, science, socst) ~ ses)
```

```
> hsb.manova
```

Call:

```
manova(cbind(read, write, math, science, socst) ~ ses)
```

Terms:

```
                ses Residuals
resp 1 1832.358 19087.062
resp 2  858.715  17020.160
resp 3 1307.091 16158.704
resp 4 1561.578 17945.922
resp 5  2528.18  20408.01
Deg. of Freedom  2          197
```

Residual standard errors: 9.843203 9.294985 9.056703 9.544425 10.17811

Estimated effects may be unbalanced

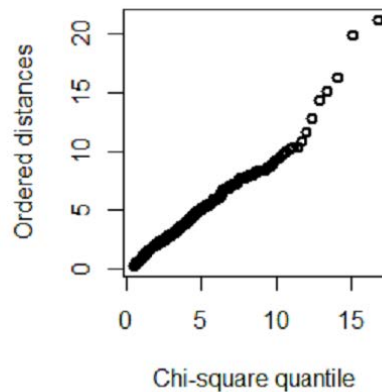
```
> summary(hsb.manova, test="Wilks")
```

```
                Df Wilks approx F num Df den Df Pr(>F)
ses                2  0.85105  3.2418    10  386 0.0004946 ***
Residuals 197
```

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

```
> means.by.grps <- cbind(tapply(read,ses,mean), tapply(write,ses,mean),
                          tapply(math,ses,mean), tapply(science,ses,mean), tapply(socst,
> means.by.grps
      [,1]      [,2]      [,3]      [,4]      [,5]
high 56.50000 55.91379 56.17241 55.44828 57.13793
low   48.27660 50.61702 49.17021 47.70213 47.31915
middle 51.57895 51.92632 52.21053 51.70526 52.03158
```



### Question FIVE:

```
> dis.sid.new <- lda(Group ~ BW + Factor68 + HR + Gesage, data=sids.new, prior=c(0.5,0.5)
> dis.sid.new
```

Call:

```
lda(Group ~ BW + Factor68 + HR + Gesage, data = sids.new, prior = c(0.5,0.5))
```

Prior probabilities of groups:

```
  0  1
0.5 0.5
```

Group means:

```
      BW Factor68      HR Gesage
0 3437.857 0.3108163 129.2408 40.00
1 2964.688 0.4018125 132.9500 39.25
```

Coefficients of linear discriminants:

```
          LD1
BW      -0.001148097
```

```
Factor68 9.966188556
HR        0.001144709
Gesage    -0.137896906
```

```
> group<-predict(dis.sid.new, sids.new, method='plug-in')$class
> table(group,Group)
```

```
      Group
group 0 1
      0 41 3
      1  8 13
```

```
> glm.fit.sids <- glm(Group ~ BW + Factor68 + HR + Gesage, data = sids.new, family=binom)
> summary(glm.fit.sids)
```

Call:

```
glm(formula = Group ~ BW + Factor68 + HR + Gesage, family = binomial, data = sids.new)
```

Deviance Residuals:

```
      Min       1Q   Median       3Q      Max
-1.5976 -0.6425 -0.3204 -0.1153  2.6181
```

Coefficients:

```
      Estimate Std. Error z-value Pr(>|z|)
(Intercept)  2.3376266 13.7711841  0.170  0.86521
BW           -0.0017276  0.0007537 -2.292  0.02190 *
Factor68     16.2727269  5.9595617  2.731  0.00632 **
HR           -0.0054844  0.0261659 -0.210  0.83398
Gesage       -0.0742118  0.3589607 -0.207  0.83621
```

---

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(Dispersion parameter for binomial family taken to be 1)

```
Null deviance: 72.549 on 64 degrees of freedom
Residual deviance: 49.713 on 60 degrees of freedom
AIC: 59.713
```

Number of Fisher Scoring iterations: 5

```
> group.probs<-predict(glm.fit.sids, sids.new, type='response')
> pred.group <- rep(0,times=nrow(sids.new))
> pred.group[group.probs > 0.5] <- 1
> table(pred.group,Group)
```

```
      Group
```

```
pred.group 0 1
           0 46 9
           1 3 7
```