Stockholm University

Spring 2021
20210319

Department of Statistics

Exam: Multivariate Analysis, Advanced level, 7.5 ECTS credits

For questions about the content of the exam, contact the course coordinator on email tatjana.vonrosen@stat.su.se. Incoming e-mail questions are answered between 13.00 and 14.00.

If the course coordinator needs to send out information to all students during the exam, this is done to your registered email address. Therefore, check your email during the exam.

NOTE! The exam shall be submitted electronically via the department's web site **no later than 19.00 (7PM)**. **The system does not allow submission after deadline** which is a new setup for this semester. Therefore, start the submission well in advance. The last hour of the exam time is intended for arranging the electronic submission.

Please note that practical help is only available during the first hour of the exam by email expedition@stat.su.se. Carefully read the enclosed instructions for exam submission. There you find all the necessary information about submission, anonymous code, extended writing time etc. If you, despite the instructions have problems submitting the exam, email the exam to tenta@stat.su.se. However, this is only done in exceptional cases. Exams sent in by email after deadline will not be corrected.

NOTE! All forms of cooperation and plagiarism are prohibited. We go over all exams carefully to detect cheating. Suspected cheating is reported to the Disciplinary Board and can lead to suspension.

The exam consists of 4 exercises giving a total of 50 points. In order to get full score for an exercise provide detailed and well motivated solutions. In order to pass the exam at least 25 points are needed.

Exercise 1. (10p)

(a) Let $\boldsymbol{X}^{\mathrm{T}} = (X_1, X_2, X_3, X_4) \sim N_4(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu} = (1, 2, 3, 4)^{\mathrm{T}}$ and

$$\boldsymbol{\Sigma} = \begin{pmatrix} 3 & 1 & 0 & 0 \\ 1 & 4 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 3 \end{pmatrix}.$$

(a) What is the conditional distribution of $(X_3, X_4)^{\mathrm{T}}$ given that $X_1 = x_1$ and $X_2 = x_2$?

(b) For the data in Table 1 (the data is also stored in file Exercise1.txt), determine whether the data for each group, Experimental and Control, are multivariate normal. Use formal statistical tests and graphical tools. If either group is nonnormal, apply an appropriate transformation to ensure normality.
Construct plots to detect possible outliers in the transformed data.

Exercise 2. (10p)

Teacher designed a pilot study to compare a new teaching approach (experimental) with the current standard (control). Two independent groups of students finished the course in statistics were compared on the three variables related to their compulsory home assignments: results' interpretation, **IR**; software usage, **SU**; and results' presentation, **RP**. The data are provided in Table 2.

Table 2. Students' scores.

	Experimental group				Control group		
Subject	IR	SU	RP	Subject	IR	SU	RP
1	31	12	24	1	31	50	20
2	52	64	32	2	60	40	15
3	57	42	21	3	65	36	12
4	63	19	54	4	70	29	18
5	42	12	41	5	78	48	24
6	71	79	64	6	90	47	26
7	65	38	52	7	98	18	40
8	60	14	57	8	95	10	10
9	54	75	58				
10	67	22	69				
11	70	34	24				

(a) For $\alpha = 0.05$, test the hypothesis than the mean performance on the three response variables is the same for both groups $H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ vs $H_1 : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$. Check the validity of the assumptions used in your analysis.

(b) Construct and plot a 95% joint confidence ellipse for the population mean vector $\boldsymbol{\mu}^{\mathrm{T}} = (\mu_1, \mu_2) = (E(X_1), E(X_2))$, where X_1 and X_2 are software usage score, **SU**, and results' presentation score, **RP**, respectively.

Exercise 3. (15p)

The data in Table 3 (also stored in the file Exercise3.txt) were collected from an experiment concerning a short-term memory capacity. The participants were divided into two groups: **Group I** consisted of subjects with low short-term memory capacity and **Group II** comprised subjects with high short-term memory capacity. The participants were listening to tape-recorded sentences. Each sentence was followed by a "probe" taken from one of five positions in the sentence. The participant had to respond with the word that came immediately after the probe word in the sentence and the speed of the reaction time was recorded (response variable).

(a) Plot the data.

(b) Conduct the profiles analysis for two groups.

(c) Discuss your findings.

Table 3. Memory capacity data.

Group I						Group II					
	Probe-word positions						Probe-word positions				
subject	1	2	3	4	5	subject	1	2	3	4	5
1	20	21	42	32	32	1	47	25	36	21	27
2	67	29	56	39	41	2	53	32	48	46	54
3	37	25	28	31	34	3	38	33	42	48	49
4	42	38	36	19	35	4	60	41	67	53	50
5	57	32	21	30	29	5	37	35	45	34	46
6	39	38	54	31	28	6	59	37	52	36	52
7	43	20	46	42	31	7	67	33	61	31	50
8	35	34	43	35	42	8	43	27	36	33	32
9	41	23	51	27	30	9	64	53	62	40	43
10	39	24	35	26	32	10	41	34	47	37	46

Exercise 4. (15p)

Gittins (1985) presented an interesting ecological study on the dynamic status of a lowland tropical rain forest in Guyana. The objective of this study was to discover whether the field observations were consistent with the view that the forest was likely to retain the same type of vegetation as the trees composing the forest died and were replaced. The weaker tree-seedling relationships would imply that the composition was changing with time, while the stronger relationships would indicate stability in the vegetation dynamics. Two sets of variables each representing the contribution of trees and seedling communities to the total vegetation, respectively, were observed. These sets are assumed to constitute a random sample. These characteristics are respectively denoted as $x_1, \ldots, x_6$ (for trees) and $y_1, \ldots, y_6$ (for seedlings).

(a) Find the sample canonical variates corresponding to significant (at the $\alpha = 0.05$) canonical correlations. Interpret these canonical variates.

(c) What proportion qf the total sample variance of the first set $\boldsymbol{X} = (x_1, \ldots, x_6)$ is explained by the first canonical variate $\hat{U}_1$? What proportion of the total sample variance of the second set $\boldsymbol{Y} = (y_1, \ldots, y_6)$ is explained by the canonical variate $\hat{V}_1$? Discuss your answers.

Remark. Variables $x_1, \ldots, x_6$ represent the contribution of trees communities *Greenheart, Wallaba, Pentaclethra, Morabukea, Mora/ Eschweilera, Jessenia* to the total vegetation.

Variables $y_1, \ldots, y_6$ represent the contribution of seedling communities: *Greenheart, Wallaba, Pentaclethra, Morabukea, Mora, Eschweilera* to the total vegetation.

Table 4.

Site	x_1	x_2	x_3	x_4	x_5	x_6	y_1	y_2	y_3	y_4	y_5	y_6
1	0.43	0.10	0.22	0.45	0.42	0.59	0.84	0.04	0.20	0.18	0.40	0.08
2	0.09	0.05	0.49	0.11	0.74	0.23	0.23	0.10	0.04	0.23	0.17	0.88
3	0.43	0.12	0.37	0.17	0.63	0.46	0.78	0.06	0.10	0.21	0.30	0.48
4	0.24	0.04	0.86	0.07	0.28	0.25	0.36	0.10	0.05	0.77	0.31	0.15
5	0.29	0.02	0.71	0.09	0.32	0.48	0.18	0.11	0.07	0.68	0.16	0.33
6	0.53	0.12	0.28	0.34	0.29	0.55	0.51	0.07	0.48	0.15	0.61	0.20
7	0.55	0.01	0.06	0.23	0.72	0.06	0.94	0.02	0.13	0.09	0.21	0.07
8	0.62	0.58	0.14	0.22	0.34	0.01	0.08	0.34	0.07	0.07	0.84	0.28
9	0.10	0.95	0.07	0.04	0.09	0.06	0.15	0.67	0.03	0.21	0.43	0.23
10	0.09	0.08	0.91	0.24	0.16	-0.11	0.06	0.06	0.02	0.87	0.01	-0.03
11	0.67	0.08	0.36	0.15	0.42	0.30	0.67	0.06	0.11	0.06	0.63	-0.04
12	0.62	0.06	0.20	0.23	0.69	0.00	0.40	0.15	0.08	0.14	0.69	0.02
13	0.10	0.94	0.04	0.02	0.01	0.02	0.06	0.93	0.03	-0.04	0.20	-0.07
14	0.02	0.97	-0.01	0.01	0.02	0.02	0.03	0.99	0.02	0.04	-0.04	0.04
15	0.64	0.06	0.20	0.37	0.50	0.30	0.40	0.11	0.46	0.05	0.73	0.01
16	0.67	0.11	0.30	0.20	0.57	-0.02	0.62	0.18	0.09	0.10	0.51	-0.04
17	0.90	0.21	0.08	0.18	0.11	-0.01	0.16	0.17	0.07	0.10	0.94	0.05
18	0.02	0.98	0.01	0.01	0.01	0.03	0.05	0.95	0.03	0.01	0.14	-0.01
19	0.33	0.08	0.20	0.81	0.39	0.14	0.14	0.07	0.96	0.06	0.18	0.05
20	0.25	0.02	0.08	0.96	0.09	0.02	0.08	0.00	0.99	0.01	0.09	-0.01
21	0.28	0.87	0.05	0.06	0.03	0.02	0.06	0.82	0.04	0.22	0.27	0.16
22	0.79	0.14	0.24	0.17	0.18	0.38	0.46	0.12	0.09	0.13	0.84	0.07
23	0.91	0.15	0.04	0.17	0.17	0.16	0.44	0.21	0.05	0.16	0.74	0.12
24	0.18	0.08	0.28	0.68	0.49	0.38	0.24	0.02	0.96	0.04	0.06	0.03
25	0.27	0.05	0.32	0.45	0.72	0.21	0.84	0.07	0.31	0.23	0.26	0.19

Note: The data in Table 4 is available in the file exercise4.txt.

Table1.

	Experimental Group				Control Group			
Subject	L	S	R	W	L	S	R	W
1	34	66	39	97	33	56	36	81
2	35	60	39	95	21	39	33	74
3	32	57	39	94	29	47	35	89
4	29	53	39	97	22	42	34	85
5	37	58	40	96	39	61	40	97
6	35	57	34	90	34	58	38	94
7	34	51	37	84	29	38	34	76
8	25	42	37	80	31	42	38	83
9	29	52	37	85	18	35	28	58
10	25	47	37	94	36	51	36	83
11	34	55	35	88	25	45	36	67
12	24	42	35	88	33	43	36	86
13	25	59	32	82	29	50	37	94
14	34	57	35	89	30	50	34	84
15	35	57	39	97	34	49	38	94
16	29	41	36	82	30	42	34	77
17	25	44	30	65	25	47	36	66
18	28	51	39	96	32	37	38	88
19	25	42	38	86	22	44	22	85
20	30	43	38	91	30	35	35	77
21	27	50	39	96	34	45	38	95
22	25	46	38	85	31	50	37	96
23	22	33	27	72	21	36	19	43
24	19	30	35	77	26	42	33	73
25	26	45	37	90	30	49	36	88
26	27	38	33	77	23	37	36	82
27	30	36	22	62	21	43	30	85
28	36	50	39	92	30	45	34	70