# EXERCISES SEMINAR 4

Laura Julià Melis
laujul03
12/06/2019

## EXERCISE 6.5

A researcher considered 3 indices measuring the severity of heart attacks. The values of these indices for $n = 40$ heart-attack patients produced the summary statistics:

$$\bar{x} = \begin{bmatrix} 46'1 \\ 57'3 \\ 50'4 \end{bmatrix} \quad \text{and} \quad S = \begin{bmatrix} 101.3 & 63'0 & 71'0 \\ 63'0 & 80'2 & 55'6 \\ 71'0 & 55'6 & 97'4 \end{bmatrix}$$

a) All 3 indices are evaluated for each patient. Test for the equality of mean indices using (6-16) with $\alpha = 0'05$.

(i) Hypothesis to test.

$H_0: C\mu = 0$

$H_1: C\mu \neq 0$ where C is a contrast matrix (rows sum 0)

We will reject $H_0$ if $T^2 = n(C\bar{x})'(CSC')^{-1}C\bar{x} > \frac{(n-1)(q-1)}{(n-q+1)} F_{q-1,n-q+1}(\alpha)$

(ii) Calculations.

- $n = 40$
- $q = 3$
- $C = \begin{bmatrix} \mu_1 - \mu_2 \\ \mu_1 - \mu_3 \end{bmatrix} = \begin{bmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \end{bmatrix}$
- $C\bar{x} = \begin{bmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \end{bmatrix} \begin{bmatrix} 46'1 \\ 57'3 \\ 50'4 \end{bmatrix} = \begin{bmatrix} 46'1 - 57'3 \\ 46'1 - 50'4 \end{bmatrix} = \begin{bmatrix} -11'2 \\ -4'3 \end{bmatrix}$ $\quad \therefore (C\bar{x})' = \begin{bmatrix} -11'2 & -4'3 \end{bmatrix}$
- $CSC' = \begin{bmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \end{bmatrix} \begin{bmatrix} 101'3 & 63'0 & 71'0 \\ 63'0 & 80'2 & 55'6 \\ 71'0 & 55'6 & 97'4 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ -1 & 0 \\ 0 & -1 \end{bmatrix} = \begin{bmatrix} 55'5 & 22'9 \\ 22'9 & 56'7 \end{bmatrix}$
- $(CSC')^{-1} = \begin{bmatrix} 0'0216 & -0'0087 \\ -0'0087 & 0'0212 \end{bmatrix}$
- $T^2 = 40 \cdot \begin{bmatrix} -11'2 & -4'3 \end{bmatrix} \begin{bmatrix} 0'0216 & -0'0087 \\ -0'0087 & 0'0212 \end{bmatrix} \begin{bmatrix} -11'2 \\ -4'3 \end{bmatrix} = 90'49458.$
- $\frac{(n-1)(q-1)}{(n-q+1)} \cdot F_{q-1,n-q+1}(\alpha) = \frac{39 \cdot 2}{38} \cdot F_{2,38}(0'05) = 2'053 \cdot 3'245 = 6'66.$

(iii) Conclusion.

$90'49 > 6'66 \Rightarrow$ We reject $H_0$

b) Judge the differences in pairs of mean indices using 95% simultaneous confidence intervals.

- Simultaneous confidence intervals for single contrasts $c'\mu$

$$c'\mu: \quad c'\bar{x} \pm \sqrt{\frac{(n-1)(q-1)}{(n-q+1)} F_{q-1,n-q+1}(\alpha)} \cdot \sqrt{\frac{c'Sc}{n}} \qquad \rightarrow (6\text{-}18) \text{ page } 281$$

(i) $\mu_1 - \mu_2$: $(46'1 - 57'3) \pm \sqrt{6'66} \cdot \sqrt{\frac{55'5}{40}}$ ($55'5 \rightarrow c_1' S c_1$) $= -11'2 \pm 3'04 = \boxed{[-14'24, -8'16]}$

(ii) $\mu_1 - \mu_3$: $(46'1 - 50'4) \pm \sqrt{6'66} \cdot \sqrt{\frac{56'7}{40}} = -4'3 \pm 3'07 = \boxed{[-7'37, -1'23]}$

- The intervals don't include the 0.

EXERCISE 6.8.

Observations on two responses are collected for three treatments. The observation vectors $\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ are:

- Treatment 1 → $\begin{bmatrix}6\\7\end{bmatrix}, \begin{bmatrix}5\\9\end{bmatrix}, \begin{bmatrix}8\\6\end{bmatrix}, \begin{bmatrix}4\\9\end{bmatrix}, \begin{bmatrix}7\\9\end{bmatrix}$
- Treatment 2 → $\begin{bmatrix}3\\3\end{bmatrix}, \begin{bmatrix}1\\6\end{bmatrix}, \begin{bmatrix}2\\3\end{bmatrix}$
- Treatment 3 → $\begin{bmatrix}2\\3\end{bmatrix}, \begin{bmatrix}5\\1\end{bmatrix}, \begin{bmatrix}3\\1\end{bmatrix}, \begin{bmatrix}2\\3\end{bmatrix}$

a) Break up the observations into mean, treatment, and residual components, as in (6-39) → page 301. Construct the corresponding anovas for each variable (see example 6.9) → page 304

- A vector of observations may be decomposed as suggested by the model ($x_{ej} = \mu + \tau_e + e_{ej}$). So:

$$x_{ej} = \bar{x} + (\bar{x}_e - \bar{x}) + (x_{ej} - \bar{x}_e)$$

observation | overall sample mean ($\hat{\mu}$) | estimated treatment effect ($\hat{\tau}_e$) | residuals ($\hat{e}_{ej}$)

($\tau_e$: treatment effect)

- Treatment means ($\bar{x}_e$, for $e = 1,2,3$)

$$\bar{x}_1 = \begin{bmatrix} \frac{6+5+8+4+7}{5} \\ \frac{7+9+6+9+9}{5} \end{bmatrix} = \begin{bmatrix} 6 \\ 8 \end{bmatrix} \begin{matrix} = \bar{x}_{e1} \\ = \bar{x}_{e2} \end{matrix} \quad \bar{x}_2 = \begin{bmatrix} \frac{3+1+2}{3} \\ \frac{3+6+3}{3} \end{bmatrix} = \begin{bmatrix} 2 \\ 4 \end{bmatrix} \quad \bar{x}_3 = \begin{bmatrix} \frac{2+5+3+2}{4} \\ \frac{3+1+1+3}{4} \end{bmatrix} = \begin{bmatrix} 3 \\ 2 \end{bmatrix}$$

- Overall mean → $\bar{x} = \begin{bmatrix} \frac{48}{12} \\ \frac{60}{12} \end{bmatrix} = \begin{bmatrix} 4 \\ 5 \end{bmatrix} = \begin{bmatrix} \bar{x}_{j=1} \\ \bar{x}_{j=2} \end{bmatrix}$

(i) For variable 1 ($j=1$)

$$\begin{pmatrix} 6 & 5 & 8 & 4 & 7 \\ 3 & 1 & 2 & & \\ 2 & 5 & 3 & 2 & \end{pmatrix} = \begin{pmatrix} 4 & 4 & 4 & 4 & 4 \\ 4 & 4 & 4 & & \\ 4 & 4 & 4 & 4 & \end{pmatrix} + \begin{pmatrix} 2 & 2 & 2 & 2 & 2 \\ -2 & -2 & -2 & & \\ -1 & -1 & -1 & -1 & \end{pmatrix} + \begin{pmatrix} 0 & -1 & 2 & -2 & 1 \\ 1 & -1 & 0 & & \\ -1 & 2 & 0 & -1 & \end{pmatrix}$$

(6-6, 5-6, 8-6, 4-6, 7-6)

$\bar{x}_{j=1} = 4$

row 1 = 6-4
row 2 = 2-4
row 3 = 3-4

2-3, 5-3, 3-3, 2-3

(ii) for variable 2 ($j=2$). $\bar{x}_{j=2} = 5$

row 1 = 8 − 5
row 2 = 4 − 5
row 3 = 2 − 5

$$\begin{pmatrix} 7 & 9 & 6 & 9 & 9 \\ 3 & 6 & 3 & & \\ 3 & 1 & 1 & 3 & \end{pmatrix} = \begin{pmatrix} 5 & 5 & 5 & 5 & 5 \\ 5 & 5 & 5 & & \\ 5 & 5 & 5 & 5 & \end{pmatrix} + \begin{pmatrix} 3 & 3 & 3 & 3 & 3 \\ -1 & -1 & -1 & & \\ -3 & -3 & -3 & -3 & \end{pmatrix} + \begin{pmatrix} 7-6 & 9-6 & 6-8 & 9-6 & 9-6 \\ -1 & 1 & -2 & 1 & 1 \\ -1 & 2 & -1 & & \\ 1 & -1 & -1 & 1 & \end{pmatrix}$$

b) Using the information in part (a), construct the one-way MANOVA table.

- GENERAL FORM OF THE TABLE:

| Source of variation | Matrix of sum of squares and cross products (SSP) | Degrees of freedom (df) |
|---|---|---|
| Treatment | $B = \sum_{\ell=1}^{g} n_\ell \underbrace{(\bar{x}_\ell - \bar{x})(\bar{x}_\ell - \bar{x})'}_{(\bar{x}_\ell - \bar{x})^2} = \begin{bmatrix} SS_{tr(1)} & SS_{tr(12)} \\ SS_{tr(21)} & SS_{tr(2)} \end{bmatrix}$ | $g-1$ |
| Residual (error) | $W = \sum_{\ell=1}^{g} \sum_{j=1}^{n_\ell} \underbrace{(x_{\ell j} - \bar{x}_\ell)(x_{\ell j} - \bar{x}_\ell)'}_{(x_{\ell j} - \bar{x}_\ell)^2} = \begin{bmatrix} SS_{res(1)} & SS_{res(12)} \\ SS_{res(21)} & SS_{res(2)} \end{bmatrix}$ | $\left(\sum_{\ell=1}^{g} n_\ell\right) - g$ |
| Total | $B+W$ | $\left(\sum_{\ell=1}^{g} n_\ell\right) - 1$ |

- Calculations:
  - $g = 3$
  - $n_1 = 5$
  - $n_2 = 3$
  - $n_3 = 4$
  - $g-1 = 3-1 = 2$
  - $\left(\sum_{\ell=1}^{g} n_\ell\right) - g = 5+3+4-3 = 9$
  - $\left(\sum_{\ell=1}^{g} n_\ell\right) - 1 = 5+3+4-1 = 11$

$x_1$

- $SS_{tr(1)} = 5 \cdot 2^2 + 3 \cdot (-2)^2 + 4 \cdot (-1)^2 = 36.$
- $SS_{res(1)} = 0 + (-1)^2 + 2^2 + (-2)^2 + 1^2 + 1^2 + (-1)^2 + 0 + (-1)^2 + 2^2 + 0 + (-1)^2 = 18.$

$x_2$

- $SS_{tr(2)} = 5 \cdot 3^2 + 3(-1)^2 + 4 \cdot (-3)^2 = 84.$
- $SS_{res(2)} = (-1)^2 + 1^2 + (-2)^2 + 1^2 + 1^2 + (-1)^2 + 2^2 + (-1)^2 + 1^2 + (-1)^2 + (-1)^2 + 1^2 = 18.$

$x_1, x_2$

- $SS_{tr(12)} = 5 \cdot 2 \cdot 3 + 3(-2)(-1) + 4(-1)(-3) = 48.$
- $SS_{res(12)} = 0 \cdot (-1) + (-1) \cdot 1 + 2 \cdot (-2) + (-2) \cdot 1 + \dots + 0 \cdot (-1) + (-1) \cdot 1 = -13$

- Results:

| | | df |
|---|---|---|
| Treatment | $B = \begin{bmatrix} 36 & 48 \\ 48 & 84 \end{bmatrix}$ | 2 |
| Residual | $W = \begin{bmatrix} 18 & -13 \\ -13 & 16 \end{bmatrix}$ | 9 |
| Total | $B+W = \begin{bmatrix} 54 & 35 \\ 35 & 102 \end{bmatrix}$ | 11 |

c) Evaluate Wilks' lambda $\Lambda^*$, and use table 6.3 to test for treatment effects. Set $\alpha = 0'01$. Repeat the test using the chi-square approximation with Barlett's correction (6-43) Compare results.

- Wilk's lambda: $\Lambda^* = \frac{|W|}{|W+B|} = \frac{18 \cdot 18 - (-13)^2}{54 \cdot 102 - (35)^2} = \frac{155}{4283} = \boxed{0'03619}$

- Hypothesis of the test (treatment effects)? → page 302.

$H_0$: $\tau_1 = \tau_2 = \tau_3 = 0$

$H_1$: at least one $\tau \neq 0$

- statistic (from table 6.3 in page 303).

$$p=2,\ g>2 \Rightarrow \left(\frac{(\sum n_e)-g-1}{g-1}\right)\left(\frac{1-\sqrt{\Lambda^*}}{\sqrt{\Lambda^*}}\right) \sim F_{2(g-1),\,2(\sum n_e - g-1)}(\alpha)$$

* $\left(\frac{5+3+4-3-1}{3-1}\right)\cdot\left(\frac{1-\sqrt{0'03619}}{\sqrt{0'03619}}\right) = 4 \cdot 4'2566 = 17'0264$

* $F_{2(3-1),\,2(5+3+4-3-1)}(0'01) = F_{4,16}(0'01) = 4'7726$
  ↓ qf(1-0'01, df1=4, df2=16)

- Conclusion $17'0264 > 4'7726 \Rightarrow$ We reject $H_0$.

$\boxed{\chi^2 \text{ approximation}}$. → for long sample size

$$-\left(n+1-\frac{(p+g)}{2}\right)\ln\Lambda^* \sim \chi^2_{p(g-1)}(\alpha)$$

* $-\left(12+1-\frac{2+3}{2}\right)\cdot\ln(0'03619) = -8'5 \cdot -3'31897 = 28'2113$

* $\chi^2_{2(3-1)}(0'01) = \chi^2_4(0'01) = 13'2767$
  ↑ qchisq(1-0'01, df=4)

Conclusion $28'2113 > 13'2767 \Rightarrow$ We reject $H_0$

- With both measurements we conclude that there are differences among the treatments.

EXERCISE 6.19

Cost data on $X_1$ = fuel, $X_2$ = repair, $X_3$ = capital, presented in table 6.10 (page 345) for $n_1$ = gasoline, $n_2$ = 23 diesel trucks. (Chapter 6.3 page 284.)

a) Test for differences in the mean cost vectors. Set $\alpha = 0'01$.

(i) Hypothesis: $H_0$: $\mu_1 - \mu_2 = \delta_0 = 0$ → "there aren't differences"

$H_1$: $\mu_1 - \mu_2 \neq 0$

We will reject $H_0$ if: $T^2 > c^2$, where:

$$T^2 = (\bar{x}_1 - \bar{x}_2 - \delta_0)' \left[ \left( \frac{1}{n_1} + \frac{1}{n_2} \right) S_{pooled} \right]^{-1} (\bar{x}_1 - \bar{x}_2 - \delta_0) \quad \text{with } \delta_0 = 0 \text{ (in our case)}$$

$$c^2 = \frac{(n_1 + n_2 - 2)p}{n_1 + n_2 - p - 1} F_{p, n_1+n_2-p-1}(\alpha)$$

- Calculations $\begin{cases} 1 = \text{"Gasoline"} \\ 2 = \text{"Diesel"} \end{cases}$

$$\bar{x}_1 = \begin{bmatrix} 12'219 \\ 8'113 \\ 9'590 \end{bmatrix} \; ; \; \bar{x}_2 = \begin{bmatrix} 10'106 \\ 10'762 \\ 18'168 \end{bmatrix} \; ; \; \bar{x}_1 - \bar{x}_2 = \begin{bmatrix} 2'113 \\ -2'650 \\ -8'578 \end{bmatrix} \; ; \; \begin{matrix} n_1 = 36 \\ n_2 = 23 \end{matrix}$$

colMeans()

$$S_1 = \begin{bmatrix} 23'013 & 12'366 & 2'907 \\ 12'366 & 17'544 & 4'773 \\ 2'907 & 4'773 & 13'963 \end{bmatrix} \; ; \; S_2 = \begin{bmatrix} 4'362 & 0'760 & 2'362 \\ 0'760 & 25'851 & 7'686 \\ 2'362 & 7'686 & 46'654 \end{bmatrix}$$

cov()

$$S_{pooled} = \frac{n_1 - 1}{n_1 + n_2 - 2} S_1 + \frac{n_2 - 1}{n_1 + n_2 - 2} S_2 = \frac{35}{57} S_1 + \frac{22}{57} S_2 = \begin{bmatrix} 15'814 & 7'886 & 2'697 \\ 7'886 & 20'750 & 5'897 \\ 2'697 & 5'897 & 26'581 \end{bmatrix}$$

$$\left[ \left( \frac{1}{n_1} + \frac{1}{n_2} \right) S_{pooled} \right]^{-1} = \left[ \frac{59}{828} S_{pooled} \right]^{-1} = \begin{bmatrix} 1'096 & -0'411 & -0'020 \\ -0'411 & 0'876 & -0'153 \\ -0'020 & -0'153 & 0'564 \end{bmatrix}$$

$$\Rightarrow T^2 = 50'91435.$$

$$\Rightarrow \frac{(36+23-2)\cdot 3}{36+23-3-1} F_{3, 36+23-3-1}(0'01) = 3'109 \cdot F_{3,55}(0'01) = 3'109 \cdot 4'159 = 12'93$$

As $T^2 = 50'9 > c^2 = 12'9 \Rightarrow$ we reject $H_0$.

b) If the hypothesis of equal cost vector is rejected in (a), find the linear combination of mean components most responsible for the rejection.

(From page 291) the coefficient vector for the linear combination most responsible for rejection is proportional to $S_{pooled}^{-1} \cdot (\bar{x}_1 - \bar{x}_2)$.

$$(S_{pooled})^{-1} = \begin{bmatrix} 0'078 & -0'029 & -0'001 \\ -0'029 & 0'062 & -0'011 \\ -0'001 & -0'011 & 0'040 \end{bmatrix} \Rightarrow S_{pooled}^{-1}(\bar{x}_1 - \bar{x}_2) = \begin{bmatrix} 0'255 \\ -0'134 \\ -0'319 \end{bmatrix}$$

c) Construct 99% simultaneous CI for the pairs of mean components. Which appear to be different?

(From page 288) $(\bar{x}_{1i} - \bar{x}_{2i}) \pm c \sqrt{\left( \frac{1}{n_1} + \frac{1}{n_2} \right) s_{ii,pooled}}$ for $i = 1, 2 \dots p$.

- For $\mu_{11} - \mu_{21} \rightarrow (\bar{x}_{11} - \bar{x}_{21}) \pm \sqrt{c^2} \cdot \sqrt{\frac{59}{828} \cdot s_{11,pooled}} = 2'113 \pm \sqrt{12'93} \cdot \sqrt{1'127} = 2'113 \pm 3'817$
$= [-1'704 \; , \; 5'93]$

- For $\mu_{12} - \mu_{22} \rightarrow -2'650 \pm \sqrt{12'93} \cdot \sqrt{1'4786} = -2'650 \pm 4'372 = [-7'022, 1'722]$

- for $\mu_{13} - \mu_{23} : -8'578 \pm \sqrt{12'93} \cdot \sqrt{1'894} = -8'578 \pm 4'95 = [-13'528, -3'628]$

These appear to be different.

d) Comment on the validity of the assumptions used in your analysis. Note in particular that observations 9 and 21 for "gasoline" have been identified as multivariate outliers (Exercise 5.22). Repeat c without them.

(A1) Assumptions in pages 284-285.

(*) Concerning the structure of the data ⇒ samples $X_1$ and $X_2$ are random and independent ✓

(*) When $n_1$ and $n_2$ are small
- 1 ⇒ Both populations are multivariate normal (could be more or less) ↓ mvnTest()
- 2 ⇒ Have same covariance matrix → $\Sigma_1 = \Sigma_2$ ↓ $S_1 \neq S_2$ ⇒ this assuption is not fulfilled

When $\Sigma_1 \neq \Sigma_2$ (and large samples) we can use → $(\bar{x}_1 - \bar{x}_2)' \left( \frac{1}{n_1} S_1 + \frac{1}{n_2} S_2 \right)^{-1} (\bar{x}_1 - \bar{x}_2) \sim \chi^2_p(\alpha)$

we are not pooling here.

- Calculations without obs. 9 and 21.

$$(\bar{x}_1 - \bar{x}_2) = \begin{bmatrix} 1'206 \\ -3'129 \\ -8'607 \end{bmatrix} \quad ; \quad S_1 = \begin{bmatrix} 9'025 & 5'156 & 3'202 \\ 5'156 & 14'259 & 4'389 \\ 3'202 & 4'319 & 11'987 \end{bmatrix} \quad ; \quad n_1 = 34 \quad n_2 = 23$$

So the statistic now is → $42'63831$ and $\chi^2_p(0'01) = qchisq(0'99, 3) = 11'345$

So we still rejecting $H_0$.

## EXERCISE 6.22.

a) Look for gender differences by testing for equality of group means. Use $\alpha = 0'05$. If you reject $H_0: \mu_1 - \mu_2 = 0$, find the linear combination most responsible.

(i) Hypothesis:

$H_0: \mu_F - \mu_M = 0$
$H_1: \mu_F - \mu_M \neq 0$

1 = F
2 = M

(ii) $T^2$-statistic

$$T^2 = (\bar{x}_1 - \bar{x}_2)' \left( \left( \frac{1}{n_1} + \frac{1}{n_2} \right) \cdot S_{pooled} \right)^{-1} (\bar{x}_1 - \bar{x}_2)$$

$$\bar{x}_1 = \begin{bmatrix} 0'3136 \\ 5'1788 \\ 2'3152 \\ 36'1548 \end{bmatrix} \quad ; \quad \bar{x}_2 = \begin{bmatrix} 0'3972 \\ 5'3296 \\ 3'6876 \\ 49'4104 \end{bmatrix} \quad ; \quad n_1 = n_2 = 25 \quad \therefore \frac{1}{n_1} + \frac{1}{n_2} = \frac{1}{25} + \frac{1}{25} = 0'08 \quad ; \quad (\bar{x}_1 - \bar{x}_2) = \begin{bmatrix} -0'0836 \\ -0'1508 \\ -1'3724 \\ -11'2656 \end{bmatrix}$$

$$S_{pooled} = \frac{n_1 - 1}{n_1 + n_2 - 2} S_1 + \frac{n_2 - 1}{n_1 + n_2 - 2} S_2 = \frac{1}{2} S_1 + \frac{1}{2} S_2 = \begin{bmatrix} 0'008 & 0'112 & 0'018 & 0'090 \\ 0'112 & 1'962 & 0'054 & 2'356 \\ 0'018 & 0'054 & 0'248 & 2'203 \\ 0'090 & 2'356 & 2'203 & 39'256 \end{bmatrix}$$

(cov(fem) and cov(male))

$$(0'08\, S_{pooled})^{-1} = \begin{bmatrix} 109228'206 & -6899'537 & -11736'028 & 821'790 \\ & 442'799 & 744'625 & -52'515 \\ & & 1338'521 & -92'849 \\ & & & 6'793 \end{bmatrix}$$

So $T^2 = 96'37322$.

- Critical value $c^2$

$$c^2 = \frac{(n_1+n_2-2)p}{n_1+n_2-p-1} \cdot F_{p,n_1+n_2-p-1}(\alpha) = \frac{48 \cdot 4}{45} \cdot F_{4,45}(\alpha = 0'05) = 4'26 \cdot 2'57874 = 11$$

- Conclusion

As $T^2 = 96'37 > 11 = c^2$ $\Rightarrow$ we reject $H_0$

(iii) Coefficient vector for the linear combination most responsible for rejection $\Rightarrow S^{-1}_{pooled} \cdot (\bar{x}_1 - \bar{x}_2)$

This is $\rightarrow$ $\begin{bmatrix} -99'399 \\ 6'376 \\ 6'228 \\ -0'791 \end{bmatrix}$

b) Construct the 95% simultaneous CI for each $\mu_{1i} - \mu_{2i}$, $i = 1:4$. Compare with Bonferroni.

Simultaneous CI

Formula $\Rightarrow$ $(\bar{x}_{1i} - \bar{x}_{2i}) \pm c\sqrt{\left(\frac{1}{n_1}+\frac{1}{n_2}\right) s_{ii,pooled}}$ for $i = 1, \dots, p$

$c = \sqrt{11} = 3'32$

- For $\mu_{11} - \mu_{21}$ : $-0'0836 \pm 3'32 \cdot \sqrt{0'008 \cdot 0'008} = -0'0836 \pm 0'0839 = [-0'92, -0'75]$
- For $\mu_{12} - \mu_{22}$ : $-0'1508 \pm 3'32 \cdot \sqrt{0'008 \cdot 1'962} = -0'1508 \pm 1'314 = [-1'465, 1'16]$
- For $\mu_{13} - \mu_{23}$ : $-1'3724 \pm 3'32 \cdot \sqrt{0'008 \cdot 0'268} = -1'3724 \pm 0'5 = [-1'87, -0'87]$
- For $\mu_{14} - \mu_{24}$ : $-11'2656 \pm 3'32 \cdot \sqrt{0'008 \cdot 39'256} = -11'2656 \pm 5'878 = [-17'14, -5'388]$

Bonferroni

Formula $\Rightarrow$ $(\bar{x}_{1i} - \bar{x}_{2i} \pm t_{n_1+n_2-2}\left(\frac{\alpha}{2p}\right)\sqrt{\left(\frac{1}{n_1}+\frac{1}{n_2}\right) s_{ii,pooled}}$ for $i = 1, \dots, p$

* $t_{n_1+n_2-2}\left(\frac{\alpha}{2p}\right) = t_{48}\left(\frac{0'05}{8}\right) = t_{48}(0'00625) = 2'595323$.

- For $\mu_{11} - \mu_{21}$ : $-0'0836 \pm 2'595 \cdot \sqrt{0'008 \cdot 0'008} = -0'0836 \pm 0'066 = [-0'1496, -0'0176]$
- For $\mu_{12} - \mu_{22}$ : $-0'1508 \pm 2'595 \cdot \sqrt{0'008 \cdot 1'962} = -0'1508 \pm 1'028 = [-1'1788, 0'8772]$
- For $\mu_{13} - \mu_{23}$ : $-1'3724 \pm 2'595 \cdot \sqrt{0'008 \cdot 0'268} = -1'3724 \pm 0'394 = [-1'7664, -0'9784]$
- For $\mu_{14} - \mu_{24}$ : $-11'2656 \pm 2'595 \cdot \sqrt{0'008 \cdot 39'256} = -11'2656 \pm 4'5 = [-15'6656, -6'6656]$.

c) The data doesn't represent a random sample. Comment on the possible implications of this info.

If the data was obtained from volunteers, is not random so the assumptions concerning the structure of the data is not fulfilled $\Rightarrow$ we can't make inferences about the population.

## EXERCISE 8.4.

Find the PC's and the proportion of the total population variance explained by each when the covariance matrix is:

$$\Sigma_1 = \begin{bmatrix} \sigma^2 & \sigma^2\rho & 0 \\ \sigma^2\rho & \sigma^2 & \sigma^2\rho \\ 0 & \sigma^2\rho & \sigma^2 \end{bmatrix} \quad , \quad -\frac{1}{\sqrt{2}} < \rho < \frac{1}{\sqrt{2}}$$

(i) From "Result 8.1" in page 432.

Let $\Sigma_1$ be the covariance matrix associated with the random vector $X' = [X_1, \dots, X_p]$. Let $\Sigma_1$ have the eigenvalue-eigenvector pairs $(\lambda_1, e_1), \dots, (\lambda_p, e_p)$ where $\lambda_i \geq 0$. Then the ith principal component is given by

$$Y_i = e_i' X = e_{i1}X_1 + e_{i2}X_2 + \dots + e_{ip}X_p \quad , \quad \text{with } i = 1, \dots, p.$$

- Calculations

  - $p = 3$

  - Eigenvalues $\Rightarrow |\Sigma_1 - \lambda I| = 0$

$$\det\left(\begin{bmatrix} \sigma^2-\lambda & \sigma^2\rho & 0 \\ \sigma^2\rho & \sigma^2-\lambda & \sigma^2\rho \\ 0 & \sigma^2\rho & \sigma^2-\lambda \end{bmatrix}\right) = 0$$

$$(\sigma^2-\lambda)^3 + 0 + 0 - 0 - (\sigma^2\rho)^2(\sigma^2-\lambda) - (\sigma^2\rho)^2(\sigma^2-\lambda) = 0$$

$$(\sigma^2-\lambda)^3 - 2\left((\sigma^2\rho)^2(\sigma^2-\lambda)\right) = 0$$

$$(\sigma^2-\lambda)\cdot\left[(\sigma^2-\lambda)^2 - 2(\sigma^2\rho)^2\right] = 0$$

$\sigma^2 - \lambda = 0 \;\Rightarrow\; \boxed{\lambda_1 = \sigma^2}$

$(\sigma^2-\lambda)^2 - 2\sigma^4\rho^2 = 0$ $\quad a = 1,\ b = 2\sigma^2,\ c = \sigma^4 - 2\sigma^4\rho^2$

$\lambda^2 - 2\lambda\sigma^2 + \sigma^4 - 2\sigma^4\rho^2 = 0$

$$\lambda = \frac{2\sigma^2 \pm \sqrt{(-2\sigma^2)^2 - 4\cdot 1\cdot(\sigma^4 - 2\sigma^4\rho^2)}}{2\cdot 1} = \frac{2\sigma^2 \pm \sqrt{8\sigma^4\rho^2}}{2} =$$

$$= \frac{2\sigma^2 \pm 2\sqrt{2}\,\sigma^2\rho}{2} \;\Rightarrow\; \boxed{\lambda_2 = \sigma^2(1+\sqrt{2}\rho)} \quad \boxed{\lambda_3 = \sigma^2(1-\sqrt{2}\rho)}$$

- Eigenvectors $\Rightarrow (\Sigma_1 - \lambda_i I)v = 0$

**For $\lambda_1$** $\quad \sigma^2 - \sigma^2 v_1 + \sigma^2\rho v_2 + 0 v_3 = 0 \;\Rightarrow\; \sigma^2\rho\, v_2 = 0 \;\Rightarrow\; v_2 = \rho\sigma^2 \;\Rightarrow\; v_1 = [0, \rho\sigma^2, 0]$

**For $\lambda_2$** $\quad [\sigma^2 - \sigma^2(1+\rho\sqrt{2})]v_1 + \sigma^2\rho v_2 + 0 v_3 = 0 \;\Rightarrow\; \sigma^2\rho\sqrt{2}\,v_1 + \sigma^2\rho v_2 = 0 \;\Rightarrow\; v_2 = [-1, \sqrt{2}, 0]$

$v_1 = \frac{-\sigma^2\rho v_2}{\sigma^2\rho\sqrt{2}}$

**For $\lambda_3$** $\quad [\sigma^2 - \sigma^2(1-\rho\sqrt{2})]v_1 + \sigma^2\rho v_2 + 0 v_3 = 0 \;\Rightarrow\; -\sigma^2\rho\sqrt{2}\,v_1 + \sigma^2\rho v_2 = 0 \;\Rightarrow\; v_3 = [1, \sqrt{2}, 0]$

$$v_1 = \frac{-\sigma^2\rho\cdot v_2}{-\sigma^2\rho\cdot\sqrt{2}}.$$

Normalized.

$$\begin{cases} e_1' = [0, 1, 0] \\ e_2' = [-\frac{1}{\sqrt{3}}, \frac{\sqrt{2}}{\sqrt{3}}, 0] \\ e_3 = [\frac{1}{\sqrt{3}}, \frac{\sqrt{2}}{\sqrt{3}}, 0] \end{cases}$$

PC's

$Y_1 = X_2$

$Y_2 = -\frac{1}{\sqrt{3}} X_1 + \frac{\sqrt{2}}{\sqrt{3}} X_2$

$Y_3 = \frac{1}{\sqrt{3}} X_1 + \frac{\sqrt{2}}{\sqrt{3}} X_2$

(ii) From (8-4) in page 433, the proportion of total population variance due to Kth principal component is:

$$\frac{\lambda_K}{\lambda_1 + \lambda_2 + \dots + \lambda_p} \quad \text{for } K = 1, 2, \dots, p.$$

- For $K=1$ $\Rightarrow \frac{\sigma^2}{\sigma^2+\sigma^2+\sigma^2\sqrt{2}\rho+\sigma^2-\sigma^2\sqrt{2}\rho} = \frac{\sigma^2}{3\sigma^2} = \boxed{\frac{1}{3}}$

- For $K=2$ $\Rightarrow \frac{\sigma^2+\sigma^2\sqrt{2}\rho}{\sigma^2+\sigma^2+\sigma^2\sqrt{2}\rho+\sigma^2-\sigma^2\sqrt{2}\rho} = \frac{\sigma^2(1+\sqrt{2}\rho)}{3\sigma^2} = \boxed{\frac{1}{3}(1+\sqrt{2}\rho)}$

- For $K=3$ $\Rightarrow \frac{\sigma^2-\sigma^2\sqrt{2}\rho}{\sigma^2+\sigma^2+\sigma^2\sqrt{2}\rho+\sigma^2-\sigma^2\sqrt{2}\rho} = \frac{\sigma^2(1-\sqrt{2}\rho)}{3\sigma^2} = \boxed{\frac{1}{3}(1-\sqrt{2}\rho)}$

EXERCISE 8.6.

Data on $x_1$ = sales and $x_2$ = profits for the 10 largests companies in the world were listed in Exercise 1.4. (Chap 1).

$$\bar{x} = \begin{bmatrix} 156'6 \\ 14'7 \end{bmatrix} \quad ; \quad S = \begin{bmatrix} 7476,45 & 303'62 \\ 303'62 & 26'19 \end{bmatrix}$$

a) Determine the sample PC's and their variances for these data.

- Eigen values - eigen vectors: $\lambda_1 = 7488'80293$ ; $e_1' = [-0'999, 0'040]$
  $\lambda_2 = 13'83707$ ; $e_2' = [-0'040, -0'999]$

- PC's

$Y_1 = -0'999 X_1 + 0'040 X_2$

$Y_2 = -0'040 X_1 - 0'999 X_2$

- Variances

$\widehat{Var}(Y_1) = \hat{\sigma}_{11} = \hat{\lambda}_1 = 7488'8.$

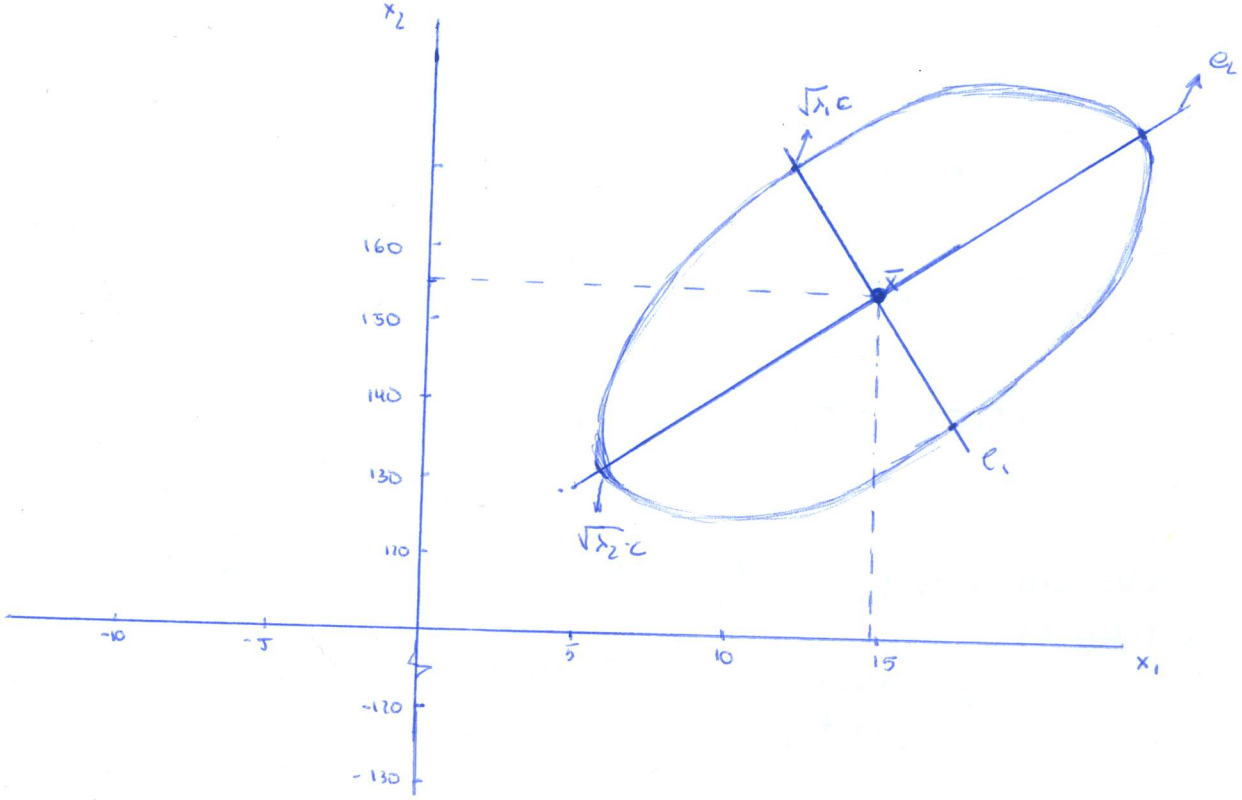
$\widehat{Var}(Y_2) = \hat{\sigma}_{22} = \hat{\lambda}_2 = 13'84$

b) Find the proportion of the total sample variance explained by $\hat{y}_1$.

For $K=1$ $\Rightarrow \frac{\hat{\lambda}_1}{\hat{\lambda}_1 + \hat{\lambda}_2} = \frac{7488'8}{7488'8 + 13'84} = \boxed{0'99816}$

c) Sketch the constant density ellipse $(x-\bar{x})' S^{-1} (x-\bar{x}) = 1'4$, and indicate the PC's.

- Center of the ellipse → $\bar{x}' = [155'6 \;,\; 14'4]$
- Axes of the ellipse
  - major axis → $e_1' = [-0'999, 0'040]$
  - minor axis → $e_2' = [-0'040, 0'999]$
- Axes half length
  - major axis → $\sqrt{\lambda_1} \cdot \sqrt{c^2} = \sqrt{7488'8} \cdot \sqrt{1'4} = 102'39$
  - minor axis → $\sqrt{\lambda_2} \cdot \sqrt{c^2} = \sqrt{13'83} \cdot \sqrt{1'4} = 4'4$




d) Compute the correlation coefficients $r_{\hat{y}_i, x_k}$, $k = 1, 2$. What interpretation, if any, can you give to the first PC?

(From page 442) → $r_{\hat{y}_i, x_k} = \dfrac{\hat{e}_{ik} \sqrt{\hat{\lambda}_i}}{\sqrt{s_{kk}}}$ for $i, k = 1, 2, \dots, p$ (corre. coeff. between variable $x_k$ and the prine. component $\hat{y}_i$)

$$r_{\hat{y}_1, x_1} = \frac{\hat{e}_{11} \sqrt{\hat{\lambda}_1}}{\sqrt{s_{11}}} = \frac{-0'999 \cdot \sqrt{7488'8}}{\sqrt{7476'45}} = -0'9998 \approx -1$$

$$r_{\hat{y}_1, x_2} = \frac{\hat{e}_{12} \sqrt{\hat{\lambda}_1}}{\sqrt{s_{22}}} = \frac{0'040 \cdot \sqrt{7488'8}}{\sqrt{26'19}} = 0'688$$

Interpretation => Most of the variability (total) can be find in the $x_1$ (sales) direction, and, also, this variable is the main one (that gives more info) in the first Principal component.

## EXERCISE 8.10

Weekly rates of return for five stocks listed on the NY Stock Exange are given in Table 8.4.

a) Construct the sample covariance matrix S, and find the sample PC's in (8-20).

$$S = \begin{bmatrix} 0'00043 & 0'00028 & 0'00016 & 0'00006 & 0'00009 \\ 0'00028 & 0'00044 & 0'00018 & 0'00018 & 0'00012 \\ 0'00016 & 0'00018 & 0'00022 & 0'00007 & 0'00006 \\ 0'00006 & 0'00018 & 0'00007 & 0'00072 & 0'00051 \\ 0'00009 & 0'00012 & 0'00006 & 0'00051 & 0'00077 \end{bmatrix}$$

$$\text{PC's.} \begin{cases} Y_1 = 0'223X_1 + 0'307X_2 + 0'155X_3 + 0'639X_4 + 0'651X_5 \\ Y_2 = 0'625X_1 + 0'570X_2 + 0'345X_3 - 0'240X_4 - 0'322X_5 \\ Y_3 = 0'326X_1 - 0'250X_2 - 0'038X_3 - 0'642X_4 + 0'646X_5 \\ Y_4 = 0'663X_1 - 0'414X_2 - 0'497X_3 + 0'309X_4 - 0'216X_5 \\ Y_5 = 0'118X_1 - 0'509X_2 + 0'780X_3 + 0'146X_4 - 0'094X_5 \end{cases}$$

b) Determine the proportion of total sample variance explained by the first three PC's. Interpret.

- For $K=1 \Rightarrow \frac{0'0014}{0'0014+0'0007+0'0003+0'0001+0'0001} = \frac{7}{13} = 0'5385$
- For $K=2 \Rightarrow \frac{0'0008}{0'0014+0'0007+0'0003+0'0001+0'0001} = \frac{7}{26} = 0'269$
- For $K=3 \Rightarrow \frac{0'0003}{0'0014+0'0008+0'0003+0'0001+0'0001} = \frac{3}{20} = 0'1154$

$\Big\} 0'9229$

* Eigen values $\longrightarrow \lambda_1 = 0'0014$, $\lambda_2 = 0'0008$, $\lambda_3 = 0'0003$, $\lambda_4 = 0'0001$, $\lambda_5 = 0'0001$

- $K = 1:3 \Rightarrow \frac{0'0014+0'0007+0'0003}{\sum_{i=1}^{5}\lambda_i} = 0'92 \Rightarrow$ If we "remove" variables 4 and 5, we will keep having 90% of the total variance, so we won't lose much information.

c) Construct Bonferroni simultaneous 90% CI for the variances $\lambda_i$, $i=1,\dots,3$ of $Y_1, Y_2, Y_3$.

- (From page 457) → Bonferroni-type simultaneous $100(1-\alpha)\%$ intervals are obtained by:

$$\frac{\hat{\lambda}_i}{\left(1 + z\left(\frac{\alpha}{2m}\right)\sqrt{\frac{2}{n}}\right)} \leq \lambda_i \leq \frac{\hat{\lambda}_i}{\left(1 - z\left(\frac{\alpha}{2m}\right)\sqrt{\frac{2}{n}}\right)}$$

- We have that $m=3$, $n=103$, so $z\left(\frac{0'1}{5}\right) = 2'128$, $\sqrt{\frac{2}{103}} = 0'1393$

$2'128 \cdot 0'1393 = 0'2965$

- For $\lambda_1 \Rightarrow \frac{0'0014}{1+0'2965} \leq \lambda_1 \leq \frac{0'0014}{1-0'2965} \Rightarrow$ $[0'00105, 0'00194]$

- For $\lambda_2 \Rightarrow \frac{0'0004}{1+0'2965} \leq \lambda_2 \leq \frac{0'0004}{1-0'2965} \Rightarrow$ $[0'00054, 0'001]$

- For $\lambda_3 \Rightarrow \frac{0'0003}{1+0'2965} \leq \lambda_3 \leq \frac{0'0003}{1-0'2965} \Rightarrow$ $[0'0002, 0'00036]$.

d) Given the results in parts a-c, do you feel that the stock rates-of-return data can be summarized in fewer than 5 dim?

Yes → answer given in question b.