Exercise02

Dennis Perrett

1/13/2022

Exercise 9. A psychologist administers three subtests of an intelligence test. The first subtest covers verbal skills, while the other two subtests cover mathematical skills. The results of the three subtests are represented by a 3-vector $\mathbf{x}$. The psychologist is interested in the overall score, which is the sum of the three individual results, and a second value indicating whether the respondant's aptitude is more in the area of language or arithmetic. This value is calculated by subtracting the sum of the results in the two arithmetical subtests from the result in the verbal subtest. To avoid negative values, 20 is added to this difference. These two numbers are combined into a 2-vector $\mathbf{y}$.

a. Show that $\mathbf{y}$ can be obtained through an affine transformation $\mathbf{x} \to \mathbf{A}\mathbf{x} + \mathbf{b}$ by determining the matrix $\mathbf{A}$ and the vector $\mathbf{b}$.

$$\mathbf{y} = \mathbf{A} \begin{pmatrix} 7 \\ 6 \\ 2 \end{pmatrix} + \begin{pmatrix} 0 \\ 20 \end{pmatrix} \mathbf{y} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & -1 & -1 \end{pmatrix} \begin{pmatrix} 7 \\ 6 \\ 2 \end{pmatrix} + \begin{pmatrix} 0 \\ 20 \end{pmatrix}$$

b. Compute the $\mathbf{y}$ values for a subject with results

$$\mathbf{x}' = (7, 6, 2)$$

in the three subtests.

$$\mathbf{y} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & -1 & -1 \end{pmatrix} \begin{pmatrix} 7 \\ 6 \\ 2 \end{pmatrix} + \begin{pmatrix} 0 \\ 20 \end{pmatrix} = \begin{pmatrix} 15 \\ 19 \end{pmatrix}$$

Exercise 10. Which of the matrices $A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$ and $B = \begin{pmatrix} -4 & 2 \\ 2 & 4 \end{pmatrix}$ cannot be a covariance matrix? A. Cannot. Must be symmetrical along the diagonal. B. Cannot. Is symmetrical, but variance cannot be negative.

Exercise 11. Let the (5×3)-matrix

$$\mathbf{X} = \begin{pmatrix} 2 & 1 & 8 \\ 3 & 0 & 5 \\ 3 & 1 & 4 \\ 4 & 1 & 1 \\ 3 & 2 & 2 \end{pmatrix}$$

represent the data of 5 subjects in 3 variables. Calculate by hand and using R:

a. the mean-vector $\bar{x}$ and the variance-covariance matrix $\mathbf{S}$;

- $\bar{x} = (3, 1, 4)$

- $$\mathbf{S} = \begin{pmatrix} E[(x_{i1}-\bar{x_1})(x_{i1}-\bar{x_1})] & E[(x_1-\bar{x_1})(x_{i2}-\bar{x_2})] & E[(xi_{i1}-\bar{x_1})(x_{i3}-\bar{x_3})] \\ E[(x_{i2}-\bar{x_2})(x_{i1}-\bar{x_1})] & E[(x_{i2}-\bar{x_2})(x_{i2}-\bar{x_2})] & E[(x_{i2}-\bar{x_2})(x_{i3}-\bar{x_3})] \\ E[(x_{i3}-\bar{x_3})(x_{i1}-\bar{x_1})] & E[(x_{i3}-\bar{x_3})(x_{i2}-\bar{x_2})] & E[(x_{i3}-\bar{x_3})(x_{i3}-\bar{x_3})] \end{pmatrix}$$

$$E[(x_{i1}-\bar{x_1})(x_{i1}-\bar{x_1})] =$$
$$\frac{1}{5-1}((2-3)(2-3)+(3-3)(3-3)+(3-3)(3-3)+(4-3)(4-3)+(3-3)(3-3)) =$$
$$\frac{1}{5-1}(1+0+0+1+0) = \frac{2}{4}$$
$$E[(x_{i2}-\bar{x_2})(x_{i2}-\bar{x_2})] =$$
$$\frac{1}{5}((1-1)(1-1)+(0-1)(0-1)+(1-1)(1-1)+(1-1)(1-1)+(2-1)(2-1)) =$$
$$\frac{1}{5-1}(0+1+0+0+1) = \frac{2}{4}$$
$$E[(x_{i3}-\bar{x_3})(x_{i3}-\bar{x_3})] =$$
$$\frac{1}{5}((8-4)(8-4)+(5-4)(5-4)+(4-4)(4-4)+(1-4)(1-4)+(2-4)(2-4)) =$$
$$\frac{1}{5-1}(16+1+0+9+4) = \frac{30}{4} = 7.5$$
$$E[(x_{i1}-\bar{x_1})(x_{i2}-\bar{x_2})] =$$
$$\frac{1}{5}((2-3)(1-1)+(3-3)(0-1)+(3-3)(1-1)+(4-3)(1-1)+(3-3)(2-1)) =$$
$$\frac{1}{5-1}(0+0+0+0+0) = \frac{0}{4} = 0$$
$$E[(x_{i1}-\bar{x_1})(x_{i3}-\bar{x_3})] =$$
$$\frac{1}{5}((2-3)(8-4)+(3-3)(5-4)+(3-3)(4-4)+(4-3)(1-4)+(3-3)(2-4)) =$$
$$\frac{1}{5-1}(-4+0+0+-3+0) = \frac{-7}{4}$$
$$E[(x_{i2}-\bar{x_2})(x_{i3}-\bar{x_3})] =$$
$$\frac{1}{5}((1-1)(8-4)+(0-1)(5-4)+(1-1)(4-4)+(1-1)(1-4)+(2-1)(2-4)) =$$
$$\frac{1}{5-1}(0+-1+0+0+-2) = \frac{-3}{4} = 0$$

$$\mathbf{S} = \begin{pmatrix} \frac{2}{4} & 0 & \frac{-7}{4} \\ 0 & \frac{2}{4} & 0 \\ \frac{-7}{4} & 0 & 7.5 \end{pmatrix}$$

```
X <-matrix(c(2,1,8,3,0,5,3,1,4,4,1,1,3,2,2),5,3,byrow=T)
cov(X)
```

```
##       [,1]  [,2]  [,3]
## [1,]  0.50  0.00 -1.75
## [2,]  0.00  0.50 -0.75
## [3,] -1.75 -0.75  7.50
```

b. the inverse $\mathbf{S}^{-1}$ of the variance-covariance matrix;

- $det(X) = ((0.5*0.5*7.5)+(0)+(0)) - ((-1.75*0.5*-1.75)+(-0.75*-0.75*0.5)+(0)) = 1.875 - 1.53125 - 0.28125 = 0.0625$ $\frac{1}{0.625}\begin{pmatrix} 3.1875 & 1.3125 & 0.875 \\ 1.3125 & 0.6875 & 0.375 \\ 0.875 & 0.375 & 0.25 \end{pmatrix} = \begin{pmatrix} 51 & 21 & 14 \\ 21 & 11 & 6 \\ 14 & 6 & 4 \end{pmatrix}$

c. the Mahalanobis distance from the data vector of subject 1 (column vector formed by 1st row of $\mathbf{X}$) to the mean vector $\bar{x}$. $\sqrt{(x-\bar{x})^\top \mathbf{S}^{-1}(x-\bar{x})} = \sqrt{(-1,0,4)\mathbf{S}^{-1}\begin{pmatrix} -1 \\ 0 \\ 4 \end{pmatrix}} = \sqrt{3} = 1.732$

```
sqrt(mahalanobis(X,colMeans(X),cov(X)))
```

```
## [1] 1.732051 1.732051 0.000000 1.732051 1.732051
```

```
sqrt( t(X[1,]-colMeans(X)) %*% solve(cov(X)) %*% (X[1,]-colMeans(X)) )
```

```
##          [,1]
## [1,] 1.732051
```

Exercise 12. The file "cheddar.txt" contains concentrations of various chemicals in 30 samples of mature cheddar cheese, and a subjective measure of taste for each sample. The columns, "Acetic", "H2S", and "Lactic" provide the concentration of acetic acid, hydrogen sulfide (both on a log scale) and lactic acid. Input the data into R and define a data matrix $\mathbf{X}$ which stores the respective concentrations in three columns. Perform the following caluclations in R.

```
dt <- as.matrix(read.table("../cheddar.txt",header=T))
```

a. Compute the mean vector $\bar{x}$ and the covariance matrix $\mathbf{S}$

```
x_bar <- colMeans(X)
S <- var(X)
```

b. Determine the diagonal Matrix Λ of eigenvalues and the matrix G of eigenvectors for S

```
lambda <- eigen(S)$values
G <- eigen(S)$vectors
```

c. Show by calculation that **G'S** $= \Lambda$

```
zapsmall(t(G) %*% S %*%G)
```

```
##          [,1] [,2]     [,3]
## [1,] 7.984344  0.0 0.000000
## [2,] 0.000000  0.5 0.000000
## [3,] 0.000000  0.0 0.015656
```

d. Compute the total sample variance based on S and Λ. What do you find? What is the reason for this observation?

Total sample variance is the sum of the diagonal (trace) of the covariance matrix.

```
sum(diag(S))
```

```
## [1] 8.5
```

```
sum(lambda)
```

```
## [1] 8.5
```

The reason is, that the eigenvalues are a type of variance decomposition.

e. Compute the generalised sample variance for S and compare the result with the product of the eigenvalues. What od you find? What is the reason for this observation?

Generalised sample variance is the determinant of the covariance matrix.

```
det(S)
```

```
## [1] 0.0625
```

```
prod(lambda)
```

```
## [1] 0.0625
```

Exercise 13. The students participating in a course were randomly divided into two groups that recieved math lessons using different teaching methods. In a final test, overall scores were determined for pure arithmetic tasks (AT) and word problems (TT). The following scores were obtained for the two teaching methods:

```
AT <- matrix(c(180,150,160,180,180,150,200,160),4,2,byrow=T)
TT <- matrix(c(90,120,80,150,80,100,110,130),4,2,byrow=T)
dt <- cbind(AT,TT)
m1 <- cbind(dt[,1],dt[,3])
m2 <- cbind(dt[,2],dt[,4])
```

Perform a multivariate statistical test by hand and using R to decide whether the two teaching methods differe with respect to the results.

a. Which statistical hypothesis is tested?

- mean(x,y|A) == mean(x,y|B) or are the 2 mean vectors the same?

b. Identify the statistical assumptions on which the test is based?

- Multivariate normality, Independence, and equal covariance matrices.

c. Run the test for Type I error set to $\alpha = 0.05$ and thereby answer the question posed above.

$$T^2 = \frac{n*n}{n+n}[(x_1 - x_2)^\top \cdot S^{-1} \cdot (x_1 - x_2)]$$

$$S = \frac{1}{n+n-2}[(n-1) \cdot S_1 + (n-1) \cdot S_2]$$

$$F = \frac{n+n-df_1-df_2}{(n+n-df_1)*2} \cdot T^2)$$

```
n = length(m1)/2
S <- (1/(n+n-2))*((n-1)*var(m1)+(n-1)*var(m2))
t_sq <- (n*n)/(n+n)*t(colMeans(m1)-colMeans(m2))%*%solve(S) %*% (colMeans(m1)-colMeans(m2))
f <- (n+n-2-1)/((n+n-2)*2) * t_sq
p_val <- round(df(F,2,n+n-2-1),5)
```

With R

```
library(Hotelling)
```

```
## Loading required package: corpcor
```

```
ht2 <- hotelling.stat(m1, m1, shrinkage = FALSE, var.equal = TRUE)

F <- (n+n-2-1)/((n+n-2)*2)*ht2$statistic
p_val <- round(df(F,2,n+n-2-1),5)
```

$P - value < 0.05$. Therefore we can conclude that the different methods make a difference.

d. Briefly discuss whether the assumptions are met.

- Probably not. Sample size is too small for robust statistics.

e. Do the teaching methods differ when performing multiple univariate tests in the single variables. Adjust the Type 1 error using bonferroni correction.

```
t.test(m1[,1],m2[,1]) # AT Difference
```

```
##
##  Welch Two Sample t-test
##
## data:  m1[, 1] and m2[, 1]
## t = 1.8516, df = 5.88, p-value = 0.1145
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -6.560968 46.560968
## sample estimates:
## mean of x mean of y
##       180       160
```

```
p.adjust(.1145, method = "bonferroni",
         n = 4)
```

```
## [1] 0.458
```

- No. They are not different on a univariate level.

Exercise 14. Before and after a group dynamic training, four subjects rated their social anxiety and their striving for dominance in group situations on a 7 category Likert scale. The following results were obtained:

```
pre = matrix(c(5,3,
                4,3,
                6,2,
                6,3),4,2,byrow=TRUE)
post = matrix(c(3,3,
                4,4,
                2,3,
                4,3),4,2,byrow=TRUE)
```

a. Test by hand and using R whether the training has changed the attitude of the subjects. Assume multinormally distributed values, and a Type I error of $\alpha = 0.05$.

$$T^2 = n \cdot \vec{d}' S_d^{-1} \cdot \vec{d}$$

```
d <- pre-post
n = length(d)/2
d_means <- colMeans(d)
d_means_mat <-matrix(d_means,4,2,byrow=T)
S = (1/(n-1))*t(d-d_means_mat)%*%(d-d_means_mat)
S_inv = solve(S)

T_sq <- n * t(d_means) %*% S_inv %*% d_means
p = 2 #(number of parameters?)
f = (n-p)/((n-1)*p) * T_sq
f>qf(0.95,2,2)
```

```
##       [,1]
## [1,] FALSE
```

We can not reject the null that the 2 means are different.

b. If the multivariate test is not significant, does it then make sense to perform univariate tests on each of the two dependent variables? Provide reasons for your answer.

- No. It does not make sense. The range (confidence interval) for multivariate significant difference reaches only to the limits of the univariate confidence intervals. That is, if it is insignificant at the multivarite level, the underlying data cannot be significant at the univariate level.
- Inversely, this however is not the case. The univariate data may insignificant, but at the multivariate level, the data may significantly differ between groups.