

Multivariate Data Analysis

Notes & Solutions to Exercises 3

1)

i) *Measurements of cranial length x_{11} and cranial breadth x_{12} on 35 female frogs*

*gave $\bar{x}'_1 = (22.860, 24.397)$ and $S_1 = \begin{pmatrix} 17.683 & 20.290 \\ * & 24.407 \end{pmatrix}$. Test the*

hypothesis that $\mu_{11} = \mu_{12}$.

Using the result from Task Sheet for week 8, Q2, illustrated Q3, we test

$H_0 : C'\mu_1 = 0$ where $C' = (1, -1)'$ by comparing

$35(22.860, 24.397)(1, -1)'[(1, -1)S_1(1, -1)']^{-1}(1, -1)(22.860, 24.397)'$ with $T^2(1, 34)$, i.e. $35 \times (-1.537) \times 1.51^{-1} \times (-1.537) = 54.75$ and compare with

$(34 - 1 + 1)/34 \times 1 \times 54.75$ with $F_{1, 34-1+1}$ i.e. 7.4 with t_{34} and conclude that there is very strong evidence that the cranial lengths and breadths of female frogs are different.

ii) *Similar measurements on 14 male frogs gave*

*$\bar{x}'_2 = (21.821, 22.843)$ and $S_2 = \begin{pmatrix} 18.479 & 19.095 \\ * & 20.756 \end{pmatrix}$.*

Calculate the pooled variance matrix for male & female frogs and test the hypothesis that female & male frogs come from populations with equal mean vectors.

Pooled variance matrix is

$$(34 \times S_1 + 13 \times S_2) / 47 = \begin{pmatrix} 17.903 & 19.959 \\ * & 23.397 \end{pmatrix} = S \text{ (say).}$$

$$\text{Now } S^{-1} = \begin{pmatrix} 1.140 & -0.973 \\ * & 0.873 \end{pmatrix}$$



Hotelling's T^2 is $[35 \times 14 / 49] \times (1.039, 1.554) S^{-1} (1.039, 1.554)' = 1.968$ and we compare this with $T^2(2,47) = 2.0434 F_{2,46}$, i.e. compare 0.963 with $F_{2,46}$ and we conclude that the data give no evidence of a difference between the sizes of skulls of Male and Female frogs.

2) Using your favourite computer package, access the British Museum Mummy Pots data (see task sheet for week 4) and calculate the two shape variables 'taper' and 'point'.

i) Do the two batches of pots differ in overall shape as reflected by the calculated shape measures 'taper' and 'point'?

First in Minitab; look at the following output:

Results for: BRMUSEUM.MTW

```
MTB > let c6=(c3-c4)/c2
MTB > let c7=c3/c4
MTB > name c6 'taper' c7 'point'
MTB > ANOVA 'taper' 'point' = batch;
SUBC> MANOVA;
SUBC> NoUnivariate.
```

ANOVA: taper, point versus batch

Criterion	Test Statistic	F	DF	P
Wilk's	0.89638	1.272 (2, 22)	22	0.300
Lawley-Hotelling	0.11560	1.272 (2, 22)	22	0.300
Pillai's	0.10362	1.272 (2, 22)	22	0.300
Roy's	0.11560			

You can see that the p-values for all of the tests are 0.3 (since only two groups all the tests are functionally related to each other and so equivalent. The value of Hotelling's T^2 is $23 \times 0.1156 = 2.6588$ and if you look at the corresponding p-value it is 0.300 (of course!). Note that there is one missing value and so this entire pot has been excluded from the analysis, leaving only 25 pots. The answer to the questions is no, there is no significant evidence that the pots differ in shape.



ii) Do the two batches of pots differ in overall size?

```
MTB > ANOVA 'length'-'base-circ' = batch;
SUBC> MANOVA;
SUBC> NoUnivariate.
```

ANOVA: length, rim-cir, base-circ versus batch

```
* Warning * Not all response variables have the same missing value
* pattern. You would get different univariate results if
* you ran this command separately for each of these response
* variables. See the Help topic "missing values" for details.
```

```
MANOVA for batch          s = 1    m = 0.5    n = 9.5

Criterion      Test Statistic          F          DF          P
Wilk's         0.56772          5.330    ( 3, 21)  0.007
Lawley-Hotelling 0.76145          5.330    ( 3, 21)  0.007
Pillai's       0.43228          5.330    ( 3, 21)  0.007
Roy's          0.76145
```

Yes, all p-values are 0.007 and so conclude that there is very strong evidence of a difference in overall size between the batches.

In R the corresponding analyses are:

```
> attach(brmuseum)
> library(MASS)
> batch=factor(batch)
> taper=(rim.cir-base.circ)/length
> point=rim.cir/base.circ
> shape.manova=manova(cbind(taper,point)~batch)
> summary(shape.manova)
          Df  Pillai approx F num Df den Df Pr(>F)
batch      1 0.10362  1.27156      2    22 0.3002
Residuals 23
> summary(shape.manova,test="Hotelling-Lawley")
          Df Hotelling-Lawley approx F num Df den Df Pr(>F)
batch      1          0.1156   1.2716      2    22 0.3002
Residuals 23
> size.manova=manova(cbind(length,rim.cir,base.circ)~batch)
> summary(size.manova)
          Df  Pillai approx F num Df den Df  Pr(>F)
batch      1 0.4323   5.3301      3    21 0.006877 **
Residuals 23
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



```

> summary(size.manova, test="Hotelling-Lawley")
      Df Hotelling-Lawley approx F num Df den Df Pr(>F)
batch   1           0.7614    5.3301     3    21 0.006877
**
Residuals 23
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

The answers are of course essentially identical. In this two group case there is no actual advantage in looking at both the Pillai trace (the **R** default) and the Hotelling-Lawley statistics since they have precisely the same significance. In general you should decide which statistics you are going to base your inference on and you should definitely not choose the one with the most significant result. In most cases there should be general agreement between the available statistics (p-values differing only marginally); if there is a substantial difference then it indicates something most unusual and unexpected about the data which is worth investigating — it may mean that you have outliers or some other form of non-normality indicating our model is not appropriate.

The sharp-eyed may notice that the data sets provided in Minitab and **R** formats are slightly different; Minitab includes measurements for pot ID-number 6826 which has a damaged rim and so no available measurement, this case was deleted from the **R** version.

- iii) *Without doing any calculations,*
- a) *would your answer to (ii) be different in any respect if you used the scores on the three PCs calculated from the size variables?*
 - b) *Would it make any difference were you to calculate the PCs using the correlation matrix instead of the covariance matrix?*

Since the PCs (provided you take all of them) are just a linear transformation of the data (whether the matrix of eigenvectors is



calculated from the covariance or correlation matrix) there should be no difference in the results on using the PCs. If not convinced then look at the following:

```
> size.pc<-princomp(cbind(length,rim.cir,base.circ))
> sizepc.manova=manova(size.pc$scores~batch)
> summary(sizepc.manova,test="Hotelling-Lawley")
              Df Hotelling-Lawley approx F num Df den Df Pr(>F)
batch          1           0.7614    5.3301     3    21 0.006877
**
Residuals 23
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

3) x_1, \dots, x_n are independent measurements of $N_p(\mu, \sigma^2 I_p)$

i) Shew that the maximum likelihood estimate of μ , subject to $\mu' \mu = r_0^2$ (a known constant) is the same whether σ is known or unknown.

This example is very like example 5.5.3 in the lecture notes:

We have $\ell(\mu; X) = -\frac{1}{2}(n-1)\text{trace}(S\sigma^{-2}) - \frac{1}{2}n(\bar{x} - \mu)'(\bar{x} - \mu)\sigma^{-2} - \frac{1}{2}n\text{plog}(2\pi) - \frac{1}{2}n\text{plog}(\sigma^2)$

Let $\Omega = \ell(\mu) - \lambda(\mu' \mu - r_0^2)$ then $\frac{\partial \Omega}{\partial \mu} = n(\bar{x} - \mu)\sigma^{-2} - 2\lambda\mu$.

So we require $\hat{\mu} = \frac{n\bar{x}}{n+2\lambda\sigma^2}$ then $\mu' \mu = r_0^2$ implies $(n+2\lambda\sigma^2)^2 r_0^2 = n^2 \bar{x}' \bar{x}$ and

so $\hat{\mu} = \frac{\bar{x} r_0}{\sqrt{\bar{x}' \bar{x}}}$ which does not depend on σ^2 .

ii) Find the maximum likelihood estimate of σ when neither μ nor σ are known.

$$\frac{\partial \Omega}{\partial \sigma} = (n-1)\text{tr}(S)\sigma^{-3} + n(\bar{x} - \mu)'(\bar{x} - \mu)\sigma^{-3} - n\text{p}\sigma^{-1}$$

$$\text{so } \hat{\sigma} = \sqrt{\frac{1}{n\text{p}} [(n-1)\text{tr}(S) + n(\bar{x} - \hat{\mu})'(\bar{x} - \hat{\mu})]}$$

$$= \sqrt{\frac{1}{n\text{p}} [(n-1)\text{tr}(S) + n(\sqrt{\bar{x}' \bar{x}} - r_0)^2]} = \sqrt{\frac{1}{n\text{p}} [\sum x_i' x_i - 2nr_0 \sqrt{\bar{x}' \bar{x}} + nr_0^2]}$$



- iii) Hence, in the case when $\sigma = \sigma_0$ (a known constant) construct the likelihood ratio test of $H_0 : \mu' \mu = r_0^2$ vs $H_A : \mu' \mu \neq r_0^2$ based on n independent observations of $N_p(\mu, \sigma_0^2 I_p)$.

Under H_0

$$\ell_{\max} = K - \frac{1}{2}n(\sqrt{\bar{x}'\bar{x}} - r_0)^2 \sigma_0^{-2}$$

Under H_A we have

$$\hat{\mu} = \bar{x} \quad \text{so} \quad \ell_{\max} = K$$

so LRT statistic is $n(\sqrt{\bar{x}'\bar{x}} - r_0)^2 \sigma_0^{-2}$ and under H_0 this $\sim \chi_1^2$

[1 d.f. since p parameters in μ estimated under H_A and p with 1 constraint under H_0]

- iv) In an experiment to test the range of a new ground-to-air missile thirty-nine test firings at a tethered balloon were performed and the three dimensional coordinates of the point of ignition of the missile's warhead measured. These gave a mean result of $(0.76, 0.69, 0.66)'$ relative to the site expressed in terms of the target distance. Presuming that individual measurements are independently normally distributed with unit variance, are the data consistent with the theory that the range of the missile was set correctly?

We have $\sigma_0=1=r_0$ and so

$$n(\sqrt{\bar{x}'\bar{x}} - r_0)^2 \sigma_0^{-2} = 39(\sqrt{(0.76, 0.69, 0.66)'(0.76, 0.69, 0.66)} - 1)^2$$

= 1.894 ($\ll 3.84 = \chi_{1,0.95}^2$) and so yes, the data are consistent with the theory that the range was set correctly

