THE UNIVERSITY OF MINNESOTA

Statistics 5401/8401 October 18, 2005

Solutions to Sample Midterm Examination

The exhibits that were in a separate booklet are included here. Instead of tables of Bonferronized F-probability points, MacAnova output is used.

Exhibit 1 (for problem 1)

```
Cmd> dogteeth <- read("","dogteeth") # read from file
Dogteeth          35    6 labels
) Tooth measurements on Golden retrievers (dogs) from kennels in
) England, France, United States, and Canada
) Col.  1: Country, 1=England, 2 = France, 3 = USA, 4=Canada
) Col.  2: X1=length of mandible
) Col.  3: X2=breadth of mandible below 1st molar
) Col.  4: X3=breadth of articular condyle
) Col.  5: X4=height of mandlible below 1st molar
) Col.  6: X5=length of 1st molar
Read from file "TP1:Stat5401:Exams:DogData.txt"

Cmd> country <- factor(dogteeth[,1]); list(country)
country         REAL   35    1     FACTOR with 4 levels (labels)

Cmd> y <- dogteeth[,-1] # 35 by 5 data matrix without factor

Cmd> means <- tabs(y,country,means:T); means # rows are group means
(1,1)       127.2      10.138      21.087       21.85      20.225
(2,1)      120.65        9.57       18.34       21.15        19.3
(3,1)      126.68      9.9444      19.956      21.533      19.933
(4,1)      131.86      10.612      22.387      22.675      20.838
              X1          X2          X3          X4          X5

Cmd> n <- tabs(,country); n # sample sizes from the 4 countries
(1)          8          10          9           8

Cmd> manova("y=country")
Model used is y=country
WARNING: summaries are sequential
NOTE: SS/SP matrices suppressed because of size; use 'manova(,sssp:T)'
                          SS and SP Matrices
                DF
CONSTANT         1
                    Type 'SS[1,,]' to see SS/SP matrix
country          3
                    Type 'SS[2,,]' to see SS/SP matrix
ERROR1          31
                    Type 'SS[3,,]' to see SS/SP matrix

Cmd> h <- matrix(SS[2,,])

Cmd> fh <- DF[2]

Cmd> e <- matrix(SS[3,,])

Cmd> fe <- DF[3]

Cmd> p <- ncols(y)
```

Exhibit 1 (continued)

```
Cmd> h # hypothesis matrix
              MandLnth   MandBrdth CndyleBrdth      MandHt   MolarLnth
MandLnth        574.46        52.4      208.14      75.056      78.248
MandBrdth         52.4      4.9881      19.685      7.3223      7.3637
CndyleBrdth     208.14      19.685      79.304      28.507      29.323
MandHt          75.056      7.3223      28.507      10.946      10.701
MolarLnth       78.248      7.3637      29.323      10.701      10.933

Cmd> e # error matrix
              MandLnth   MandBrdth CndyleBrdth      MandHt   MolarLnth
MandLnth        1686.5      108.29      438.69      234.21      70.768
MandBrdth       108.29      13.031      36.175      17.569      4.0754
CndyleBrdth     438.69      36.175      202.94      77.266       11.66
MandHt          234.21      17.569      77.266        61.3      1.0675
MolarLnth       70.768      4.0754       11.66      1.0675      20.994

Cmd> s <- e/fe ; s # pooled variance matrix
              MandLnth   MandBrdth CndyleBrdth      MandHt   MolarLnth
MandLnth        54.404      3.4931      14.151      7.5553      2.2828
MandBrdth       3.4931     0.42035      1.1669     0.56675     0.13147
CndyleBrdth     14.151      1.1669      6.5466      2.4924     0.37612
MandHt          7.5553     0.56675      2.4924      1.9774    0.034435
MolarLnth       2.2828     0.13147     0.37612    0.034435     0.67722

Cmd> # Compute T^2 for all pairs of means

Cmd> names <- vector("England","France","USA","Canada")

Cmd> tsq <- matrix(dmat(4,0),labels:structure(names,names)) #EMPTY

Cmd> for(i,run(4)){for(j,run(4)){ # Fill matrix tsq
        dij <- vector(means[i,]-means[j,]) # difference of mean vectors
        vhatij <- (1/n[i]+1/n[j])*s
        tsq[i,j] <- dij' %*% solve(vhatij) %*% dij};;}

Cmd> tsq
           England      France         USA      Canada
England          0      9.4974      2.7166      4.0353
France      9.4974           0      4.9819      24.137
USA         2.7166      4.9819           0      11.069
Canada      4.0353      24.137      11.069           0

Cmd> releigenvals(h,e) # relative eigenvalues
(1)     0.82745    0.069417    0.023853  2.5498e-16  -3.432e-16
```

1. In **Exhibit 1** are analyses on 5 measurements on the teeth of Golden retriever dogs from breeders in 4 countries, England, France, United States and Canada. **Tables 1a**, **1b**, and **1c** on p. 5, 6 and 7 of the exhibit booklet are tables of Bonferronized F probability points, that is $F_{f_1,f_2}(\alpha/k)$, for Bonferronizing factor k = 1, 2, 3, 4, 5, 6, denominator degrees of freedom f_2 = 26, 27, ..., 35, and numerator degrees of freedom f_1 =3, 4 and 5, respectively.

(a) (15) Use Bonferronized F-tests to test the null hypothesis that the expected tooth measurements are the same in all four Countries. State the null and alternative hypotheses using $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, $\boldsymbol{\mu}_3$, and $\boldsymbol{\mu}_4$ as notation for the four 5 dimensional mean vectors. Hypothesis and error matrices on on p. 2 of the exhibit booklet.

Solution

The null hypothesis is $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \boldsymbol{\mu}_3 = \boldsymbol{\mu}_4$. The alternative hypothesis is H_1: $\boldsymbol{\mu}_i \neq \boldsymbol{\mu}_j$ some $i \neq j$.

The F statistic for variable j is $F_j = (h_{jj}/f_h)/(e_{jj}/f_e) = (f_e/f_h)\times h_{jj}/e_{jj} = (31/3)h_{jj}/e_{jj}$.
From the output

$F_1 = (31/3)\ 574.46/1686.5 = 3.520$
$F_2 = (31/3)\ 4.988/13.031 = 3.955$
$F_3 = (31/3)\ 79.304/202.94 = 4.038$
$F_4 = (31/3)\ 10.946/61.3 = 1.845$
$F_5 = (31/3)\ 10.933/20.994 = 5.381$

Since p = 5, you Bonferronize by 5.

```
Cmd> invF(.05/5,3,31,upper:T) # or see table 1a
(1)       4.4837       Bonferronized  critical  value
```

Since $F_5 = 5.381 > 4.4837$, you can reject H_0.

(b) (15) Use Bonferronized Hotelling's T^2 to test the same hypothesis as in (a). Values of T^2 are computed on p. 2 of the Exhibit booklet.
Solution
The MacAnova output provides six T^2 statistics, each of which tests one of the hypothesis $\boldsymbol{\mu}_i = \boldsymbol{\mu}_j$. The largest of these is $T_{2,3}^2 = 24.137$ testing whether France and Canada have the same means. Recalling that $(f_e-p+1)T^2/(f_e p) = F_{p,f_e-p+1}$, $(27/155)T^2 = F_{5,27}$. Since there are six T_{ij}^2 tests, you Bonferronize by 6:

```
Cmd> invF(.05/6,5,27,upper:T) # or see table 1c
(1)       3.9287
```

Since (27/155) 24.137 = 4.2045 > 3.9287, again you can reject H_0 at the 5% level. Moreover you can conclude that France and Canada have significantly different mean vectors.

(c) (15) Use a test involving the eigenvalues of **H** relative to **E** to test the same null hypothesis as in (a). Relative eigenvalues are on p. 2 of the exhibit booklet.
Solution

Roy's **maximum root test statistic** is $\hat{\theta}_{max} = \dfrac{\hat{\lambda}_{max}}{1+\hat{\lambda}_{max}}$ = 0.82745/1.82745 = 0.45279. $s = \min(p,f_h)$ $= \min(5,3) = 3$, $m = (|p-f_h| - 1)/2 = (2-1)/2 = 1/2$, $n = (f_e - p - 1)/2 = (31-5-1)/2 = 12.5$.
From the 5% chart for $s = 3$, the critical value is about .45 so it's borderline significant. [simulation with 10,000 replicates shows the P-value is about .0445 and the critical value is about 0.445 < 0.453.]

Wilks likelihood ratio is $\Lambda^* = \dfrac{1}{\prod_{i=1}^{s}(1+\hat{\lambda}_i)}$ and approximately $-m_1\log(\Lambda^*) = m_1\sum_{i=1}^{s}\log(1+\hat{\lambda}_i)$, $m_1 = f_e - (p - f_h + 1)/2 = 31 - (5-3+1)/2 = 29.5$

```
Cmd> lambdahat <- releigen(h,e); 29.5*sum(log(1 + lambdahat))
```

```
(1)        20.461
```

The upper 5% of χ_{15}^2 is 25.00, so Wilks test does not provide significant evidence the means differ. Using MacAnova I found that the χ_{15}^2 P-value is 0.15494 and using macro `cumwilks()`, the exact P-value is 0.15716, both greater than $\alpha = .05$.

Hotelling's trace statistic is $\sum_{i=1}^{s}\hat{\lambda}_i$ = .82745 + .069417 + .023853 = 0.9207. $m_2\sum_{i=1}^{s}\hat{\lambda}_i$ is approximately χ_{15}^2 where $m_2 = f_e - p - 1 = 31 - 5 - 1 = 25$. $25\times 0.9207 = 23.017 < 25.00$ so Hotelling's test doesn't reach significance. The P-value from `cumtrace()` is 0.10864 > .05.

Pillai's trace statistic is $(f_e + f_h)\sum_{i=1}^{s}\frac{\hat{\lambda}_i}{1+\hat{\lambda}_i}$ = (31+3)×(.82745/1.82745 + .069417/1.069417 + .023853/1.023853) = 34×.54100 = 18.394 < 25.00, so Pillai's, too, is non-significant. The P-value from `cumpillai()` is 0.22822.

2. $\boldsymbol{X}$ is a 50 by 4 data matrix whose rows are a random sample from a population with mean $\boldsymbol{\mu} = [\mu_1, \mu_2, \mu_3, \mu_4]'$ and variance matrix $\boldsymbol{\Sigma} = [\sigma_{ij}]$. The sample variance matrix is $\boldsymbol{S} = [s_{ij}]$ and sample mean is $\bar{\mathbf{x}} = [\bar{x}_1, \bar{x}_2, \bar{x}_3, \bar{x}_4]'$.

Give the dimensions of each of the following matrices and describe what they represent statistically (for example, "sample regression coefficients in the regression of the last column of $\boldsymbol{X}$ on the first 3 columns" or "variance of $\bar{x}_3$"). No justification is necessary.

(a) (10) $(1/50)\mathbf{1}_{50}'\boldsymbol{X}$

Solution

This is a particular case of $(1/n)\mathbf{1}_n'\boldsymbol{X} = (1/n)\sum\mathbf{x}_i' = \bar{\mathbf{x}}'$ = sample mean. Dimensions are 1 by 4.

(b) (10)

$$\frac{1}{50}\begin{bmatrix}1 & 1 & -1 & -1\end{bmatrix}\mathbf{S}\begin{bmatrix}1\\1\\-1\\-1\end{bmatrix}$$

Solution

This is a particular case of $(1/n)\mathbf{c}'\mathbf{Sc}$ where $\mathbf{c} = [1\ \ 1\ \ -1\ \ -1]'$ is a contrast vector and $n = 50$ is the sample size. Hence $\mathbf{c}'\mathbf{Sc}$ is the sample variance of $y = \mathbf{c}'\mathbf{x} = x_1 + x_2 - x_3 - x_4$ and $(1/n)\mathbf{c}'\mathbf{Sc}$ is the estimated variance of $\bar{\mathbf{y}} = \mathbf{c}'\bar{\mathbf{x}} = \bar{x}_1 + \bar{x}_2 - \bar{x}_3 - \bar{x}_4$. Dimensions are 1 by 1.

(c) (10)

$$\begin{bmatrix}1/\sqrt{\sigma_{11}} & 0 & 0 & 0\\ 0 & 1/\sqrt{\sigma_{22}} & 0 & 0\\ 0 & 0 & 1/\sqrt{\sigma_{33}} & 0\\ 0 & 0 & 0 & 1/\sqrt{\sigma_{44}}\end{bmatrix}\Sigma\begin{bmatrix}1/\sqrt{\sigma_{11}} & 0 & 0 & 0\\ 0 & 1/\sqrt{\sigma_{22}} & 0 & 0\\ 0 & 0 & 1/\sqrt{\sigma_{33}} & 0\\ 0 & 0 & 0 & 1/\sqrt{\sigma_{44}}\end{bmatrix}$$

This is the 4 by 4 population correlation matrix

$$\boldsymbol{R} = \begin{bmatrix} \frac{\sigma_{11}}{\sqrt{\sigma_{11}{}^2}} & \frac{\sigma_{12}}{\sqrt{\sigma_{11}\sigma_{22}}} & \frac{\sigma_{13}}{\sqrt{\sigma_{11}\sigma_{33}}} & \frac{\sigma_{14}}{\sqrt{\sigma_{11}\sigma_{44}}} \\ \frac{\sigma_{12}}{\sqrt{\sigma_{11}\sigma_{22}}} & \frac{\sigma_{22}}{\sqrt{\sigma_{22}{}^2}} & \frac{\sigma_{23}}{\sqrt{\sigma_{22}\sigma_{33}}} & \frac{\sigma_{24}}{\sqrt{\sigma_{22}\sigma_{44}}} \\ \frac{\sigma_{13}}{\sqrt{\sigma_{11}\sigma_{33}}} & \frac{\sigma_{23}}{\sqrt{\sigma_{22}\sigma_{33}}} & \frac{\sigma_{33}}{\sqrt{\sigma_{33}{}^2}} & \frac{\sigma_{34}}{\sqrt{\sigma_{33}\sigma_{44}}} \\ \frac{\sigma_{14}}{\sqrt{\sigma_{11}\sigma_{44}}} & \frac{\sigma_{24}}{\sqrt{\sigma_{22}\sigma_{44}}} & \frac{\sigma_{34}}{\sqrt{\sigma_{33}\sigma_{44}}} & \frac{\sigma_{44}}{\sqrt{\sigma_{44}{}^2}} \end{bmatrix} = \begin{bmatrix} 1 & \rho_{12} & \rho_{13} & \rho_{14} \\ \rho_{12} & 1 & \rho_{23} & \rho_{24} \\ \rho_{13} & \rho_{23} & 1 & \rho_{34} \\ \rho_{14} & \rho_{24} & \rho_{34} & 1 \end{bmatrix}$$

Exhibit 3 (for problem 3)

```
Cmd> data <- read("","t04_03")
T04_03    30    4 format
) Data from Table 4.3 p. 187 in
) Applied Mulivariate Statistical Analysis, 5th Edition
) by Richard A. Johnson and Dean W. Wichern, Prentice Hall, 2002
) These data were edited from file T4-3.DAT on disk from book
) Omitted was the last column (d^2) as this can be computed directly
) using distcomp(T04_03)
) Four measurements of stiffness
) Col. 1: x1 (from shock wave down board)
) Col. 2: x2 (from vibrating board)
) Col. 3: x3 (from static test)
) Col. 4: x4 (from static test)
Read from file "TP1:Stat5401:Data:JWData5.txt"

Cmd> n <- nrows(data) # sample size

Cmd> stats <- tabs(data,covar:T,mean:T);stats
component: mean
(1)      1906.1      1749.5      1509.1        1725
component: covar
(1,1)  1.0562e+05       94614       87290       94231
(2,1)       94614  1.0151e+05       76137       81064
(3,1)       87290       76137       91917       90352
(4,1)       94231       81064       90352  1.0423e+05

Cmd> xbar <- stats$mean; s <- stats$covar

Cmd> cor(data) #correlation matrix
(1,1)           1     0.91376     0.88593     0.89812
(2,1)     0.91376           1     0.78821      0.7881
(3,1)     0.88593     0.78821           1      0.9231
(4,1)     0.89812      0.7881      0.9231           1

Cmd> eigen(s)
component: values
(1)  3.6322e+05       26814      7688.5      5550.9
component: vectors
(1,1)     0.52638    -0.19881    -0.23971     0.79116
(2,1)     0.48659    -0.72687     0.13627    -0.46511
(3,1)     0.47569     0.44462     0.75856    0.025065
(4,1)     0.50977     0.48421    -0.59039    -0.39637

Cmd> c1 # previously entered 3 by 4 matrix
(1,1)           1           1          -1          -1
(2,1)           1          -1           0           0
(3,1)           0           0           1          -1

Cmd> c1xbar <- c1 %*% xbar; c1xbar'
(1,1)      421.53      156.57     -215.83

Cmd> c1sc1 <- c1 %*% s %*% c1'; c1sc1
(1,1)       95759      -20213       442.6
(2,1)      -20213       17899     -2013.8
(3,1)       442.6     -2013.8       15440
```

(**Exhibit 3** continued on following page)

Exhibit 3 (continued)

```
Cmd> c1xbar' %*% solve(c1sc1/n) %*% c1xbar
(1,1)      254.72

Cmd> sqrt(diag(c1sc1/n))
(1)      56.498      24.426      22.686

Cmd> c2 # previously entered matrix
(1,1)           1          -1           0           0
(2,1)           1           0          -1           0
(3,1)           1           0           0          -1
(4,1)           0           1          -1           0
(5,1)           0           1           0          -1
(6,1)           0           0           1          -1

Cmd> c2xbar <- vector(c2 %*% xbar)

Cmd> c2sc2 <- c2 %*% s %*% c2'

Cmd> vhat2 <- c2sc2/n

Cmd> print(format:"10.4f",c2xbar:c2xbar',format:"10.3f",c2sc2,vhat2)
c2xbar:
(1,1)   156.5667   396.9667   181.1333   240.4000    24.5667  -215.8333
c2sc2:
(1,1)  17899.357   -149.843  -2163.595 -18049.200 -20062.953  -2013.753
(2,1)   -149.843  22953.964  14448.246  23103.807  14598.089  -8505.718
(3,1)  -2163.595  14448.246  21382.809  16611.841  23546.405   6934.563
(4,1) -18049.200  23103.807  16611.841  41153.007  34661.041  -6491.966
(5,1) -20062.953  14598.089  23546.405  34661.041  43609.357   8948.316
(6,1)  -2013.753  -8505.718   6934.563  -6491.966   8948.316  15440.282
vhat2:
(1,1)    596.645     -4.995    -72.120   -601.640   -668.765    -67.125
(2,1)     -4.995    765.132    481.608    770.127    486.603   -283.524
(3,1)    -72.120    481.608    712.760    553.728    784.880    231.152
(4,1)   -601.640    770.127    553.728   1371.767   1155.368   -216.399
(5,1)   -668.765    486.603    784.880   1155.368   1453.645    298.277
(6,1)    -67.125   -283.524    231.152   -216.399    298.277    514.676
```

3. This problem deals with the analysis of data in Table 4.3 in Johnson and Wichern. The stiffness of 30 boards was measured in four different ways, two dynamic testing methods (sending a shock wave down the board and vibrating a board), and using two static methods. Exhibit 3 contains MacAnova output related to analysis of these data.

There is interest in comparing static with dynamic testing methods, comparing the dynamic methods and comparing the static methods.

(a) (15) Describe in words what the 3 elements of variable `c1xbar` on p. 3 of the exhibits are and why they might be relevent things to compute. Is there statistical evidence that expectations of the elements of `c1xbar` are non-zero?

Solution

The three rows of `c1` define contrasts. Row 1 defines the contrast $y_1 + y_2 - y_3 - y_4$, a comparison of the dynamic tests with the static tests and `c1xbar[1] =` $\bar{y}_1 + \bar{y}_2 - \bar{y}_3 - \bar{y}_4$. Row 2 defines the contrast $y_1 - y_2$, a comparison of the two dynamic tests and `c1xbar[2] =` $\bar{y}_1 - \bar{y}_2$.

Row 3 defines the contrast $y_3 - y_4$, a comparison of the two static tests and `c1xbar[3] =` $\bar{y}_3 - \bar{y}_4$. These contrast thus match the interests summarized in paragraph two of the question.

Since $V[\mathbf{Cy}] = \mathbf{C\Sigma}_y\mathbf{C}$, and $\hat{V}[\mathbf{C}\bar{\mathbf{y}}] = \frac{1}{n}\mathbf{CSC}'$, the MacAnova output can be recognized as computing $T^2 = 254.72$ for testing H_0: $\mathbf{C\mu} = 0$, that is $\mu_1 + \mu_2 - \mu_3 - \mu_4 = 0$, $\mu_1 - \mu_2 = 0$ and $\mu_3 - \mu_4$.
Since $f_e = n - 1 = 29$ and the length of $\mathbf{C\mu} = p' = p - 1 = 3$, you compare
$(f_e - p' + 1)\ T^2/(f_e p') = 27\times 254.72/(29\times 3) = 79.051$ with $F_{3,27}(.05) = 2.96$, so there is strong evidence that $\mathbf{C\mu} \neq \mathbf{0}$. The null hypothesis is equivalent to $\mu_1 = \mu_2 = \mu_3 = \mu_4$.

Another way to do it would be to Bonferronize the three *t*-statistics whose numerators are the elements of `c1xbar` and whose denominators are the estimated standard errors, `sqrt(diag(c1sc1/n))`. These are $t_1 = 421.53/\sqrt{(95759/30)} = 7.461$, $t_2 = 156.57/\sqrt{(17899/30)} = 6.41$ and $t_3 = -215.83/\sqrt{(15440/30)} = -9.5137$. The critical value is $t_{29}(.025/3) = 2.541$. Since all three *t* statistics exceed this, you can reject H_0. Moreover you can conclude there is a difference between the dynamic and static tests, between the two dynamic tests, and betseen the two static tests.

(b) (15) Find a 95% confidence interval for the difference between the mean measurement using the shock-wave measurement (variable 1) and the mean measurement using the first static test (variable 3). (Hint: Look at row 2 of matrix `c2` on p. 3 of exhibits.) Do it using a method that would be appropriate for simultaneous confidence intervals of all comparisons of two measurement methods. Use the shortest limits that would be appropriate.

Solution
The most appropriate method is Bonferronized t-based intervals. There are 6 comparisons defined in `c2` Berronizing factor is 6.

Each comparison is of the form $\bar{y}_j - \bar{y}_k \pm t\sqrt{\hat{V}[\bar{y}_j - \bar{y}_k]} = \bar{y}_j - \bar{y}_k \pm t\sqrt{\mathbf{c}_i'(\mathbf{S}/n)\mathbf{c}_i}$, where $\mathbf{c}_i'$ is the row of `c2` defining the comparison, in this case row 2. From the output $\mathbf{c}_2'(\mathbf{S}/n)\mathbf{c}_2 = 765.132$ and from tables of Bonferronized t with d.f. = 29, $t = 2.83$. So the interval is

$$396.9667 \pm 2.83\times\sqrt{765.132} = (318.69, 475.25)$$

A critical value based on the F-distribution, whether or not Bonferronized, in not appropriate.