Key

BIOS 6250 Multivariate Analysis Mid-Term Exam

The following is an open-book, open-note mid-term exam. Collaboration is not allowed. Please show all your work, as no credit will be given for unsupported answers. Please try to fit your answers into the spaces provided. Papers will be collected promptly at 2:30 p.m.

Good luck!

(12pts) 1.(a) Consider the data on 98 patients undergoing radiation therapy contained in Table 1.7, p. 43 of our textbook. As a preliminary step in the data analysis, the marginal distributions of the 6 dependent variables were checked for univariate normality (UVN). For variable X_2 (Activity), the test statistic for the correlation coefficient test for normality was $r_Q = .945$. Interpolate a p-value for this result using Table 2 of the article by Looney and Gulledge (1985) that was distributed in class. Does the assumption of UVN appear to be reasonable for the Activity variable? Why or why not?

	lower-tail area		
n	.000	z	.005
95	.260		.979
98	x	.945	y
100	.255		.979

To find the 0 % point for n = 98:

$$\frac{x - .255}{.260 - .255} = \frac{98 - 100}{95 - 100} \Rightarrow \boxed{x = .257}$$

$\boxed{y = .979 \text{ by inspection}}$

To find the p-value:

$$\frac{z - .000}{.005 - .000} = \frac{.945 - .257}{.979 - .257} \Rightarrow z = .005\left(\frac{.688}{.722}\right) = .0048$$

(round to .005)

So, $p \approx .005 \Rightarrow$ Reject UVN for X_2. So the assumption of UVN is not reasonable for Activity.

(10 pts) (b) Based on an interim power analysis calculation for this study, it was discovered that n = 98 would not be sufficiently large for the purposes of the study. It was determined that a sample size of n = 150 would be required. So, an additional 52 patients were selected at random from the patient population. For this new sample of 150 patients, the test statistic for the correlation coefficient test for normality for the Activity variable was $r_Q = .964$. Would you accept or reject normality for X_2 in this case? Justify your answer.

Table 2 in Looney & Gulledge (1985) does not go beyond n = 100. However, in any given column (except the 1st), the critical values increase as n increases. So $r^*_{150}(.005) \geq r^*_{100}(.005) = .979$. Since $r_Q = .964 < r^*_{100}(.005) \leq r^*_{150}(.005)$, conclude that $p \leq .005$ and reject UVN for X_2.

2. Consider the data on the mineral content of the bones of n = 25 older women contained in Table 1.8, p. 44, of our textbook. The assessment of multivariate normality (MVN) for these data indicated no serious departures from the MVN assumption. Suppose that this sample of 25 women came from a nursing home in a large mid-western city. It is of interest to determine if the true mean vector for women residing in this nursing home is the same as the national mean vector: [.84 .82 1.86 1.74 .70 .70]. Consider the following SAS output for this problem:

Test $H_0: \underset{\sim}{\mu} = \underset{\sim}{\mu}_0$

Table 1.8, p. 44
DESCRIPTIVE STATISTICS

The MEANS Procedure

Variable	Mean	Variance	Std Dev
X1	0.8438000	0.0130016	0.1140245
X2	0.8183200	0.0114179	0.1068545
X3	1.7926800	0.0803572	0.2834735
X4	1.7348400	0.0694845	0.2635991
X5	0.7044000	0.0115684	0.1075566
X6	0.6938400	0.0105991	0.1029521

HOTELLINGS T-SQUARED TEST

Obs	T2CAL	FCAL	PVAL1
1	14.8872	1.96428	0.12192

CHI-SQUARE APPROXIMATION TO HOTELLINGS TEST

Obs	T2CAL	DF	PVAL2
1	14.8872	6	0.021153

(15 pts) (a) Using the information given above, perform the test of $H_0: \underset{\sim}{\mu} = \underset{\sim}{\mu}_0$ using Wilk's lambda with $\alpha = .05$. Give the test statistic, degrees of freedom (if any), and the p-value. Would you reject H_0? Why or why not?

From Result 5.1, p. 218 of our textbook:

$$\Lambda_{cal}^{2/n} = \left[1 + \frac{T^2_{cal}}{n-1}\right]^{-1} = \left[1 + \frac{14.8872}{24}\right]^{-1} = \frac{1}{1.6203} = .617$$

$$\Rightarrow \Lambda_{cal} = (.617)^{n/2} = (.617)^{25/2} = .0024 \text{ (round to .002)}.$$

There are no d.f. associated with Wilk's lambda. The exact p-value is the same as that for Hotelling's T^2, so $p = .122$. Therefore, I would not reject Ho since $p > .05$.

(6 pts) (b) For the data in Table 1.8, the Hotelling's T^2 and "robust" approaches yield different conclusions. Which is more appropriate? Why?

Since the MVN assumption appears to be reasonable for these data, the "normal-theory" approach (i.e., Hotelling's T^2) is the preferred method.

3. Consider Exercise 6.5, p. 333. Suppose that an examination of the raw data indicated that the MVN assumption was not appropriate for this study.

(15 pts) (a) Using the information provided in this exercise, find a 95% simultaneous confidence interval (CI) for the true difference between μ_2 and μ_3 using the χ^2 approximation to the distribution of Hotelling's T^2 statistic.

This can be treated as a 1-group repeated measures design. An approx. 95% CI $(\mu_1 - \mu_2)$ using the "robust" method is given by $(\bar{X}_2 - \bar{X}_3) \pm \sqrt{\chi^2_{q-1}(\alpha)} \sqrt{\dfrac{S_{22} + S_{33} - 2S_{23}}{n}}$

$$= (57.3 - 50.4) \pm \sqrt{\chi^2_2(.05)} \sqrt{\frac{80.2 + 97.4 - 2(55.6)}{40}}$$

$$= 6.9 \pm \sqrt{5.99}\,(1.288) = 6.9 \pm 3.2 = \boxed{(3.7,\ 10.1)}$$

(5 pts) (b) Why is the method of multiple comparisons you used in Part (a) above the most appropriate one to use for these data?

Because the assumption of MVN appears to be violated and $40 = n > p + 30 = 3 + 30 = 33$

(12 pts) 4. Consider the Egyptian skull data presented in Table 6.13, p. 344. (See also Exercise 6.24, pp. 344-5). The results of the MANOVA analyses for these data were presented in class. (Note that the SAS output is mistakenly labelled "`CALCIUM DATA - TABLES 6.5 & 6.6`" on the printouts I distributed in class.) As indicated on p. 12 of the SAS handout for this example, the value of the Wilk's lambda test statistic is 0.830. Verify that the value of the F test statistic corresponding to this value of Λ is 2.05 and that the degrees of freedom for the corresponding F test are $df_1 = 8$ and $df_2 = 168$.

Note that $g = 3$, $p = 4$, $n = 90$

From Table 6.3, p. 300

$$F_{cal} = \left(\frac{n-p-2}{p}\right)\left(\frac{1-\sqrt{\Lambda_{cal}}}{\sqrt{\Lambda_{cal}}}\right) = \left(\frac{90-4-2}{4}\right)\left(\frac{1-\sqrt{.830}}{\sqrt{.830}}\right)$$

$$= 21(.0976) = 2.05 \checkmark$$

$df_1 = 2p = 2(4) = 8 \quad df_2 = 2(n-p-2) = 2(90-4-2) = 2(84) = 168 \checkmark$

(10 pts) 5. Briefly answer ANY ONE of the following questions (a) – (c) using information presented in the indicated article that was distributed in class. You may answer the other questions for extra credit. Please indicate your choice for the required question by circling (a), (b), or (c).

(a) In repeated measures analysis, SAS routinely provides the results of both the multivariate and "adjusted univariate" approaches when testing any hypotheses involving the repeated factor (typically, "TIME"). What recommendations do Looney and Stanley (1989) make regarding the use of these methods when testing H_{OTG}, i.e., the hypothesis of no trial x group interaction, at significance level α?

Use the appropriate multivariate test (Hotelling's T^2, etc.) at level $\alpha/2$ and use either (1) the $\tilde{\epsilon}$-adjusted F test if $g=2$ or if it is known that $\epsilon \gtrsim .75$ OR (2) the $\hat{\epsilon}$-adjusted F test if $g \geq 2$ and nothing is known about the true value of ϵ, also at significance level $\alpha/2$ (p. 221).

(b) In Example 2 presented in Looney (1995), the Box-Cox methodology was used to find a transformation for 5 of the 6 dependent variables in Royston's hematology data set. Was this transformation successful? In other words, did the transformed variables appear to satisfy the MVN assumption? Justify your answer.

No, the marginal distributions of the transformed variables all appear to be UVN (Table 3); however, Srivastava's b_{1p} and b_{2p} tests and Srivastava & Hui's M_2 test all indicate a violation of MVN. The marginal results for the PC's should be checked to see where the apparent violation of MVN occurred (p. 68).

(c) What justification do Looney and Gulledge (1985) provide for recommending the use of Blom's plotting position in constructing normal probability plots?

Based on their simulation results, the correlation coefficient test based on Blom's position has power that is comparable to that of Hazen's for shorter-tailed alternatives & is generally the most powerful against longer-tailed & skewed alternatives. It is also preferable to Hazen's position when the plot is to be used for estimation purposes (p. 77).

(15 pts) 6. In univariate analysis, if the population standard deviation σ is known, the appropriate statistic for testing $H_0: \mu = \mu_0$ is $z_{cal} = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$. To perform a 2-tailed test, the result is then compared with $z_{\alpha/2}$. (This is an exact test.) What is the multivariate analogue of the above procedure for performing a 2-tailed test of $H_0: \underset{\sim}{\mu} = \underset{\sim}{\mu}_0$ when the population covariance matrix Σ is known? Justify your answer.

First, square the univariate test statistic:

$$z^2_{cal} = \left(\frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}} \right)^2 = \frac{n(\bar{X} - \mu_0)^2}{\sigma^2}$$

This is known to have a χ^2_1 distribution, so rewrite as

$$\chi^2_{cal} = n(\bar{X} - \mu_0)(\sigma^2)^{-1}(\bar{X} - \mu_0).$$

Now, make each mean a vector & each variance a covariance matrix. Also make the 1st vector involving means a row vector and the last vector involving means a column vector so that the resulting product will be a scalar:

$$\chi^2_{cal} = n(\bar{\underset{\sim}{X}} - \underset{\sim}{\mu}_0)' \Sigma^{-1} (\bar{\underset{\sim}{X}} - \underset{\sim}{\mu}_0) \quad (1)$$

In the general multivariate case with p variables, this latter quantity is known to have a χ^2_p distribution. [Result 4.7 (a), p. 163].

Therefore, to test $H_0: \underset{\sim}{\mu} = \underset{\sim}{\mu}_0$, calculate the test statistic in (1) & then use a χ^2_p distribution to calculate the exact p-value.