Fundamentals of Data Analysis: Assignment (Final)

Deadline: 2/12/2004 (Thursday)

Please post to the mailbox next to the IS management office (2nd floor of IS building)

1. **All students should solve this problem.**

   Run a numerical experiment on PCA (principal component analysis), according to the following steps.

   a. Generate 100 pseudo-random numbers, obeying a normal distribution whose mean and standard deviation are 0 and 2, respectively. (Let us denote these numbers a_i)
   b. Generate 100 pseudo-random numbers, obeying a normal distribution whose mean and standard deviation are 0 and 5, respectively. (Let us denote these numbers b_i)
   c. Operate an arbitrary 2D rotation matrix to two-dimensional vectors, made by paring two numbers generated in steps a and b, and make 100 two-dimensional data $\boldsymbol{x}_i$.
   d. Calculate correlation matrix of data $\boldsymbol{x}_i$.
   e. Calculate eigen values and eigen vectors of the correlation matrix and find the first and second principal components. Discuss the direction of the principal vectors.

2. **Answer question A if the first digit of your student ID is an odd number. Otherwise, answer question B.**

   A. A car manufacturer examines the effect of gasoline types on the mileage, using 16 sample cars for two types of gasoline. The result is as below. Test the hypothesis that the type of gasoline affects the mileage.

      Regular: average 30 km, unbiased estimate of variance 9 km^2

      High-octane: average 33 km, unbiased estimate of variance 10 km^2

   B. A psychologist wants to know whether the experience of an experiment affects the performance of the next experiment. He uses six subjects and measures their performance of the first and second session of the experiment. The following table summarizes the result. Test the hypothesis that the experience affects the performance.

	1	2	3	4	5	6
First Session	8	5	3	4	5	9
Second Session	10	8	8	7	9	9

3. **Answer question A if the second digit of your student ID is an odd number. Otherwise, answer question B.**

   A. Assume that the true mean and variance of height of the Japanese are 165cm, and 60cm^2, respectively. You want to know whether the variance of the height of the American is

different from that of the Japanese. You measure the height of 12 American people, and find that their mean is 175cm and the unbiased estimate of the variance is 120cm^2. Test the hypothesis that their variances are different.

B. You want to know whether the variance of weight of male new-born baby is smaller than that of female baby. You choose six pairs of male and female babies so that the socio-economic situations of their parents are similar in each pair, and measure their weights. The results are as follows. Test the hypothesis that their variances are different.

	1	2	3	4	5	6
Male Infant	3.2	3.3	3.6	3.9	3.5	3.6
Female Infant	3.1	3.1	3.4	3.8	3.3	3.2

4. Answer question A if your birth date is an odd number. Otherwise, answer question B.

A. You want to know what method is the best for education of French language. You divide 20 students into four groups, and teach the students using the following four methods for the groups: 1) a lecture-base method, 2) a textbook-base method, 3) a method using audio tape listening, and 4) a method using films on French life and culture. The following table shows the result of the examination at the end of the semester. Make a variance analysis table and test the effect of the education method. Run a post-hoc multiple comparison if the effect is significant.

Group 1	75	70	90	80	75
Group 2	68	73	70	60	65
Group 3	80	65	70	68	72
Group 4	87	90	85	75	80

B. In order to know the effect of the amount of TV watching on the ability to understand the documents of high-school student, you examine the amount of TV watching and socio-economic situation of the students. You divide each of the amount and socio-economic situation into three levels, and sample six students for each combination of two factors. The following table summarizes the score of document understanding for each group. Make a variance analysis table and test the main effects and their interaction.

		Amount of TV watching		
		Small	Medium	Large
Socio-economic level	Low	17, 17, 16, 16, 16, 15	16, 16, 15, 14, 13, 13	12, 12, 12, 9, 8, 8
	Medium	18, 18, 17, 17, 16, 16	16, 15, 15, 15, 15, 14	11, 11, 10, 10, 9, 9
	High	18, 18, 18, 17, 15, 14	18, 16, 15, 15, 14, 12	12, 12, 10, 10, 9, 8

5. Write your comments and requests on this lecture (if any). END.