University of Toronto Scarborough
Department of Computer and Mathematical Sciences
STAD29 (K. Butler), Final Exam
April 13, 2022

Aids allowed: my lecture overheads (slides); any notes that you have taken in this course; your marked assignments; my assignment solutions; non-programmable, non-communicating calculator.
This exam has 25 numbered pages of questions.
In addition, you have an additional booklet of Figures to refer to during the exam. Contact an invigilator if you do not have this.
The maximum marks available for each part of each question are shown next to the question part.
You may assume throughout this exam that the code shown in Figure 1 of the booklet of Figures has already been run.
The University of Toronto's Code of Behaviour on Academic Matters applies to all University of Toronto Scarborough students. The Code prohibits all forms of academic dishonesty including, but not limited to, cheating, plagiarism, and the use of unauthorized aids. Students violating the Code may be subject to penalties up to and including suspension or expulsion from the University.

Question 1 (17 marks)

The data shown in Figure 2 presents the results of an experiment conducted to study the influence of the operating temperature (100C, 125C and 150C) and three faceplate glass types (A, B and C) on the light output of an oscilloscope tube. The data are saved in a dataframe called `gtl`. There are 27 observations in total, three for each combination of glass and temperature.

(a) (2 marks) A plot is shown in Figure 3. The code used to produce the plot is shown above the plot. What do you conclude from the plot?

(b) (3 marks) What feature of the light output is *not* shown in Figure 3? Describe *in words* a different graph that would show this feature. (Hint: show your understanding by saying what role each variable plays in your graph.)

(c) (3 marks) An analysis of variance is shown in Figure 4. What do you conclude from it, in the context of the data?

(d) (3 marks) Some more analysis is shown in Figure 5. What is the (statistical) technical term for this kind of analysis? What do you conclude from it, in the context of the data?

(e) (2 marks) Some more analysis is shown in Figure 6. What do you conclude from it, in the context of the data?

(f) (4 marks) Explain briefly how your conclusions from each of the previous two parts are consistent (or are inconsistent) with the graph in Figure 3.

Question 2 (21 marks)

An Italian wine enthusiast analyzed the chemical composition of 178 wines from one region of Italy. These wines were all of the same type, but the grapes used to make the wines were grown from three different cultivars, labelled 1 through 3 in the data. Wines made from grapes of different cultivars might be expected to have a different chemical composition.

Thirteen different chemicals were measured. In the same order as in the data, these are: Alcohol, Malic acid, Ash, Alkalinity of ash, Magnesium, Total phenols, Flavonoids, Nonflavonoid phenols, Proanthocyanins, Color intensity, Hue, OD280/OD315 of diluted wines, Proline. I don't know any more about these variables than what you see here. Part of the dataset is shown in Figure 7.

(a) (2 marks) Figure 8 shows a MANOVA analysis. Why do you think I created the response variable as I did, rather than by using `cbind`? Explain briefly.

(b) (2 marks) What do you conclude from the MANOVA in Figure 8?

(c) (2 marks) A discriminant analysis is shown in Figure 9. Why is it a good idea to do this analysis?

(d) (2 marks) Why are there two linear discriminants in Figure 9? Are they both worth considering? Explain briefly.

(e) (2 marks) In Figure 9, which four of the measured variables make the greatest contribution to the first linear discriminant? Explain briefly.

(f) (2 marks) In Figure 9, which three of the measured variables make the greatest contribution to the second linear discriminant? Explain briefly.

(g) (2 marks) What kind of values on which variables would make a wine have a very negative score on LD2?

(h) (2 marks) A plot of the discriminant scores is shown in Figure 10. Based on this plot, do you think it is difficult or easy to distinguish the cultivars on the basis of their chemical composition? Explain briefly.

(i) (2 marks) In terms of the original variables, what seems to distinguish cultivar 2 from the others? Hint: you have actually already answered this.

(j) (3 marks) Some more output is shown in Figure 11. What is this output, and what do you learn from it? Explain briefly. (Hint: say what the top table is, what the results in it mean, and what the bottom table tells you.)

Question 3 (14 marks)

ACTIVE (Advanced Cognitive Training for Independent and Vital Elderly) was a multi-site randomized controlled trial conducted from 1999 to 2001. The primary aim of the trial was to test the effects of three distinct cognitive interventions on measures of cognitively demanding daily activities. The three cognitive interventions focused on memory, executive reasoning, and speed of processing. There was also a control intervention, designed to look like a real intervention, but with no actual cognitive effect. Each participant was randomly assigned to one of the interventions. Participants were assessed before training, immediately after training, and after one and two years, for a total of four time points. At each time point, the assessment was done using the Hopkins Verbal Learning Test, on which a lower score indicates better cognitive functioning.

A total of 1,575 older adults completed the trial. Some of the data is shown in Figure 12.

(a) (2 marks) What makes this a repeated measures analysis?

(b) (2 marks) An interaction plot is shown in Figure 14. Why might you expect not to see an significant interaction between treatment and time for these data? Explain briefly.

(c) (2 marks) The repeated measures ANOVA is shown in Figure 15. What feature of this data set do you think led to the interaction term in fact coming out significant?

(d) (3 marks) In Figure 15, is there a significant treatment effect? If there is, how would you describe the nature of that treatment effect? (Hint: see Figure 14.)

(e) (3 marks) In Figure 15, is there a significant time effect? If there is, how would you describe the nature of that time effect? (Hint: see Figure 14.)

(f) (2 marks) Why does it make practical sense that the effect of time would be as you found?

Question 4 (7 marks)

Figure 16 shows the air distances, in miles, between 10 US cities. We are going to see what happens when we perform a hierarchical cluster analysis on these distances.

(a) (4 marks) A complete-linkage cluster analysis is carried out, with the results shown in the dendrogram in Figure 17. Suppose that we want to group the cities into four clusters. Which cities would be in each cluster?

(b) (3 marks) Figure 18 shows the latitudes and longitudes of the US cities featured in this question. A city with a more negative longitude is further west, and a city with a more positive latitude is further north. Using this Figure, what do the cities in each of your clusters have in common?

Question 5 (12 marks)

A basketball fan collected stats on around a thousand past and present National Basketball Association (NBA) players. Some information about basketball in general and the stats listed below is given in Figure 19. We will consider the following stats:

- `fg_pct`: "field goal percent": the number of (2-point) shots made, divided by the number attempted
- `3fg_pct`: the number of 3-point shots made, divided by the number attempted
- `ft_pct`: the number of free throws made, divided by the number attempted
- `oreb`: the number of offensive rebounds per game
- `dreb`: the number of defensive rebounds per game
- `ast`: the number of assists per game
- `stl`: the number of steals per game
- `blk`: the number of shots blocked per game
- `tov`: the number of turnovers per game

The first three, labelled `pct`, are actually proportions, between zero and one; the others are made comparable for different players by dividing the total number made by the number of games that player played. (Some players have played many more games than others.)

Player stats in basketball generally fall into one of three categories: shooting effectiveness (the ones that end with `pct` in this question), rebounding, and the others (that benefit or hurt the team that the player belongs to).

Some of the data is shown in Figure 20, and a scree plot is shown in Figure 21.

(a) (2 marks) How many factors should be used for a factor analysis? Explain briefly.

(b) (4 marks) A factor analysis is shown in Figure 22. This uses two factors, which may or may not be the same as your preferred number of factors from the previous part. Which of the original variables are an important part of each of the two factors? Explain briefly.

(c) (6 marks) A plot of factor scores of all 1,000 players would be very hard to read, so we focus on just three players. A plot of the factor scores for the three players is shown in Figure 23. In addition, the original data for these three players are shown in Figure 24, and the percentile ranks on all the variables for those data are shown in Figure 25. For each of the three players, based on their positions on the plot, which of the original variables would you expect to be high or low, and is that indeed the case? Explain briefly.

Question 6 (11 marks)

The Scouts is an organization that provides outdoor activities and leadership opportunities for boys aged 11–14. It is believed that boys enrolled in Scouts are less likely to be involved in criminal activities. In a survey, 800 boys of Scouting age were classified by socioeconomic status, whether or not they were enrolled in Scouts, and whether or not they were classified as a juvenile delinquent. Juvenile delinquency is defined as "the act of participating in unlawful behavior as a minor or individual younger than the statutory age of majority". The data are shown in Figure 26.

(a) (2 marks) Two tables are shown in Figure 27. One of the numbers in the (bottom) table is 0.0877. What does that number tell you?

(b) (2 marks) In Figure 27, what is the purpose of the `margin = 1` in the code to make the second table? (This is connected with your answer to the previous part; if you answer this part elsewhere, you get credit for this part too.)

(c) (2 marks) What is your overall conclusion from Figure 27, in the context of the data?

(d) (3 marks) Some analysis is shown in Figure 28. What do you conclude from it, and how is that different from what you concluded earlier? Explain briefly.

(e) (2 marks) Some more tables are shown in Figure 29. Using these tables, how can you add to your conclusion from the previous part?

Use this page if you need more space to write your answers. Be sure to label any answers here with the question and part that they belong to.