University of Toronto Scarborough
Department of Computer and Mathematical Sciences
STAD29 (K. Butler), Final Exam
April 12, 2025

Aids allowed (on paper, no computers):

- 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 13 numbered pages of questions, including this cover page.

In addition, you have an additional booklet of Figures to refer to during the exam.

The maximum marks available for each question are shown next to the question number.

If you need more space, use the last page of the exam. Anything written on the back of the page will not be graded.

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.

Tom Izzo

Each year, 64 teams are selected to compete for the national championship in a US college sport. The teams are divided into four regions, and the sixteen teams within a region are seeded from 1 (believed to be best) to 16 (believed to be worst). The teams play a knockout tournament, with the winning team in each region advancing to the Final Four. (These winning four teams then play for the national championship). A total of 1664 teams were selected for the championship between 1985 and 2010. In the dataset shown in Figure 2, each row represents one of those teams; the year of participation of that team is shown, together with their seed number that year, and whether or not they advanced to the Final Four (1 is yes, 0 is no). The dataframe is called `FinalFourIzzo`.

Tom Izzo is a well-known college coach in this sport. Some experts think that his teams do better in the national championship than you would otherwise expect. In the dataframe, the column `Izzo` is 1 if the team was coached by Tom Izzo in that year, and 0 if coached by someone else. (Izzo only coached one team in any given year.)

(1) (2 points) A logistic regression is shown in Figure 3. Why is this logistic regression appropriate?

(2) (2 points) How do you know that the logistic regression in Figure 3 is predicting the probability that a team *does* advance to the Final Four, as opposed to the probability that the team does not do so?

(3) (2 points) In Figure 3, why does the negative sign of the Estimate for `Seed` make practical sense?

(4) (2 points) Some predictions are shown in Figure 4. Describe the effect of the seed number as shown in that Figure.

(5) (3 points) Assess the belief of the "experts" mentioned in the description of the data, using *both* Figure 3 and Figure 4. What do you conclude?

(6) (2 points) A graph of the predictions is shown in Figure 5. Why does it make sense that the envelope around the blue curve is bigger than the envelope around the red curve?

Using both sides of the brain

Eighty subjects were randomly assigned to one of two kinds of task (Visual and Verbal in column `Task` in our dataset), and they were instructed to report on the result in one of two randomly-chosen ways (also Visual and Verbal in column `Report`). The total Time needed to complete both the Task and the Report is recorded in the column `Time`, in seconds. A smaller time is better. The two types of task and the two types of report were designed to take the same amount of time to complete when done in isolation. Some randomly chosen rows of the data are shown in Figure 6.

According to psychological theory, visual and verbal activities are carried out by opposite sides of the brain. Thus, when the task and the report are of different kinds, the subject can use both sides of the brain to complete them at the same time, but when the task and the report are of the same kind, the subject's brain has to do them one after the other, taking a longer time. The data in Figure 6 were collected to investigate whether this psychological theory is correct.

(7) (2 points) In a two-way analysis of variance predicting `Time` from `Task`, `Report`, and their interaction, would you expect to see a significant interaction, according to the psychological theory? Explain briefly.

(8) (2 points) Some boxplots are shown in Figure 7. How do these plots suggest that it would be better to use the log of `Time` in the analysis rather than `Time` itself?

(9) (2 points) Some analysis is shown in Figure 8. What do you conclude from it, in the context of the data?

(10) (4 points) Some further analysis is shown in Figure 9 and Figure 10. What do you conclude, in the context of the data? (You might find Figure 7 helpful.)

(11) (2 points) Does this data support the psychological theory given in the description of the data? Discuss briefly.

Barley yields

Five varieties of barley (a grain) were grown in each of six locations in each of 1931 and 1932. The yield of barley (the total amount grown) was measured in each year. The scientists were mainly interested in whether the variety affected yield, allowing for any effect of location. The data, in dataframe `immer`, are shown in Figure 11. `Y1` contains the yields in 1931 and `Y2` contains the yields in 1932. The highest yield is best. The locations, in `Loc`, and the varieties, in `Var`, are indicated by the initial one or two letters of their names. For the varieties, their full names are: Manchuria (M), Svansota (S), Velvet (V), Trebi (T), and Peatland (P). I don't know what the full names of the locations are. There are 30 rows of data altogether.

(12) (2 points) Why will it *not* be possible to estimate an interaction effect between location and variety for these data?

(13) (2 points) Why is it appropriate to analyze these data using a MANOVA?

(14) (3 points) A MANOVA is shown in Figure 12. What do you conclude from this analysis?

(15) (2 points) Figure Figure 13 shows graphs of yields `Y1` and `Y2`, separately, against location and variety. There are too many varieties for colour to distinguish them clearly, so I used a different technique (which I explain in the next question, so you don't need to ask about it here). Why did I put location on the x-axis rather than variety?

(16) (2 points) The package `ggrepel` contains a function `geom_text_repel` that places text as close to a point as possible. The text that appears on the plot comes from the variable in `label`. Some of the points are close together on the plot, but you may assume that if a letter is further up the page, its corresponding point on Figure 13 is further up the page as well. Why do you think the significance (or not) of the MANOVA for variety in Figure 12 makes sense according to Figure 13? Explain briefly.

Cross-country ski grip

In cross-country skiing, the fastest skiers are the ones who can generate the most upper-body power. This might be influenced by how they grip the ski poles. There are three standard ways in which a cross-country skier might grip the poles, called Classic, Integrated, and Modern. These are shown in Figure 14, in column `grip.type`. 12 skiers (labelled in `id`) were randomly assigned to use one of the grip types. The researchers were concerned about a possible effect of practice, so they measured the upper body power (`UBP`) generated by each skier on three separate occasions, labelled 1, 2, and 3 in column `replicate`. The dataframe is called `grip`.

(17) (1 point) How, specifically, do you know that a repeated measures analysis will be necessary here?

(18) (3 points) An interaction plot is shown in Figure 15, and a spaghetti plot is shown in Figure 16. Using both Figures, do you think there will be a significant interaction between grip type and time, or not? Explain briefly.

(19) (2 points) Some code is shown in Figure 17. It was necessary to run this code before running the analysis in Figure 18. Why, specifically?

(20) (4 points) The analysis in Figure 18 created `grip.1`, and output from `summary(grip.1)` is shown in Figure 19. The people who collected the data wanted to know whether there was any difference in upper body power between the types of grip, whether there was a practice effect, and whether the upper body power depended on the combination of grip type and replicate. What does Figure 19 tell you about the answers to these questions? Describe your process clearly.

(21) (2 points) From Figure 15 and Figure 16, what seems to be the reason for any significant results you found in the previous question? Explain briefly.

NBA players 2015

Basketball players play in one of three positions: Centre, Guard, or Forward. The NBA basketball league compiles statistics on its players. Do players in different positions tend to have different statistics? In basketball, players can score in one of three ways: a field goal (2 points), a 3-point field goal from outside of a line on the court (3 points), and a free throw (1 point), awarded after a foul (illegal play). After a field goal attempt is missed, the player who catches the ball is awarded a rebound. A player who intercepts a pass by a member of the opposing team is credited with a steal.

In the data shown in Figure 20, each player's name and their position is shown. Next are their success rate (number made divided by attempts) at field goals, three-point field goals, and free throws, and finally the number of rebounds, steals, and fouls committed per game. (Questions continue on the next page)

(22) (2 points) Some analysis is shown in Figure 21. What do you conclude from this analysis?

(23) (2 points) Some further analysis is shown in Figure 22. Why is this a helpful analysis, given what we know so far?

(24) (2 points) In Figure 22, what are the two most important of the original variables in LD1? Explain briefly.

(25) (2 points) Based on your answer to the previous question, what would make a player have a very *negative* score on LD1? Explain briefly.

(26) (3 points) Figure 23 shows the results of some further analysis. What do you learn from this table specifically? (You have some choices here; for full credit, find **two** distinct insightful choices.)

(27) (2 points) A plot is shown in Figure 24. Using your conclusions from the previous question, describe whether this plot leads you to the same conclusions as in that question. Explain briefly. (If you are having trouble distinguishing the three colours, an invigilator can tell you the colour of a point you indicate. I have tried to make the graph colourblind-friendly.)

(28) (2 points) In Figure 24, what is the most important *one* thing you can say about values on the original variables for the players at the bottom of the plot? Explain briefly.

(29) (2 points) What are *two* pieces of evidence you have seen so far that the second linear discriminant is in fact not worth considering?

(30) (2 points) Look at Figure 25. For the player Kawhi Leonard, was his position correctly predicted? Was the prediction a close decision or a clear decision? Explain briefly in each case.

Hearing test

One hundred males of age 39 with no history of hearing disorders did a hearing test. Each individual is exposed to signals of varying frequencies with an increasing loudness until the individual indicates that they can hear the signal. Each individual was exposed to signals of 500, 1000, 2000, and 4000 hertz (frequency) in each ear. The loudness values were recorded in columns with names starting with L or R (left or right ear) followed by the frequency. The loudness was measured on a log scale, so some of the values are negative. A larger loudness value indicates that the individual had more difficulty hearing the signal. Each row of the data shown in Figure 26, in dataframe `hearing`, is for a different individual, identified in `Sl_No`. The dataframe `hearing0` contains all the columns of `hearing` except for `Sl_No`.

Interest was in what distinguished the individuals' ability to hear amongst the eight variables recorded.

(31) (2 points) A screeplot is shown in Figure 28. Someone suggests to you that four principal components is a good number. Do you agree or disagree? Explain briefly.

(32) (1 point) A principal components analysis is carried out as shown in Figure 27. The first component is often a measure of "overall size". How do you know that this is the case here?

(33) (2 points) Looking at Figure 27, which two of the original variables are the most important in component 2? What, if anything, do they have in common? Explain briefly.

(34) (4 points) A graph of scores on the first two principal components is shown in Figure 29. Find individual 66 on this graph. Using Figure 26 and Figure 30, explain briefly why it is not surprising that this individual appears on Figure 29 where they do. (Or "is surprising", if that's what you think.) You only need to consider the one most relevant component in this question.

Chest pain

Each of 10186 New Zealand adults was asked their age and whether they experienced any pain or discomfort in their chest over the last six months. If yes, they indicated whether it was on their left and/or right side of their chest. The data are shown in Figure 31. For each combination of age group and whether or not pain was experienced on the left side or the right side, the frequency `n` is shown. The researchers were interested in whether age had any influence on the presence of any sort of chest pain.

In the data, `(30,40]` means "strictly greater than 30 and less than or equal to 40", so that a person aged exactly 40 would be in this age group and not the `(40,50]` age group.

(35) (3 points) Some log-linear modelling is shown in Figure 32. Describe my process, and say why I stopped where I did.

(36) (2 points) Why do the researchers care more about `age:left_side` than `left_side:right_side`?

(37) (2 points) A graph is shown in Figure 33. In the code above the graph, why did I use `x` and `fill` as shown, rather than having the variables the other way around?

(38) (2 points) Interpret the graph in Figure 33, in the context of the data.

(39) (2 points) Interpret the graph in Figure 34, in the context of the data.

If you need any more space, use this page, labelling each answer with the question number it belongs to.