University of Toronto Scarborough
Department of Computer and Mathematical Sciences
STAC32 (K. Butler), Final Exam
December 12, 2015

Aids allowed:

- My lecture overheads
- Any notes that you have taken in this course
- Your marked assignments and labs, including the midterm exam
- The course R text
- The course SAS text
- Non-programmable, non-communicating calculator

Before you begin, complete the signature sheet, but sign it only when the invigilator collects it. The signature sheet shows that you were present at the exam.

This exam has 11 numbered pages of questions. Please check to see that you have all the pages.

In addition, you should have an additional booklet of output to refer to during the exam. Contact an invigilator if you do not have this.

Answer each question in the space provided (under the question). If you need more space, use the backs of the pages, but be sure to draw the marker's attention to where the rest of the answer may be found.

The maximum marks available for each part of each question are shown next to the question part. In addition, the total marks available for each page are shown at the bottom of the page, and also in the table on the next page.

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.

Last name: ____

First name: ____

Student number: ____

For marker's use only:

Page	Points	Score
1	14	
2	10	
3	8	
4	13	
5	9	
6	13	
7	13	
8	5	
9	11	
10	7	
11	17	
Total:	120	

1. A university employment office wants to compare the time taken by students of different majors to find their first full-time job after graduation. The data are shown in Figure 1. The data came from a spreadsheet and were saved into a file called `job.csv`, in CSV format, in the folder where R is running and also in your "home" folder on SAS Studio. Give code to accomplish the tasks below, in R or SAS as required. Each part can be done with three lines of code or fewer.

   (a) (2 marks) Read the data into a data frame, and display the first 6 (2017: or "few") lines of that data frame (in R).

   (b) (2 marks) In R, using the data frame that you read in in the previous part, draw side-by-side boxplots of number of days (until a student finds full-time employment) for each type of degree.

   (c) (2 marks) In R, obtain the mean number of days until a student finds full-time employment for each type of degree.

   (d) (3 marks) Read the data into a SAS data set and display that data set. (In SAS.) (2015 note) Do not worry about an "informat" for this question, just read the data as best you can without it.

   (e) (1 mark) (2015) Describe how SAS will read in the degree names without errors, but not as you see them in the data file.

   (f) (2 marks) Obtain the mean number of days required to find full-time employment for a student graduating with each degree type (in SAS).

   (g) (2 marks) Obtain side-by-side boxplots of number of days required to find full-time employment for each type of degree, in SAS.

2. A farmer decided to start growing blueberries. He bought 8 plants of each of 4 different varieties of highbush blueberries, with the aim of finding out which variety produced the most blueberries. He planted each plant, and measured the yield of blueberries from each one (that is, how many pounds of blueberries each plant produced).

   (a) (2 marks) An analysis of variance is shown in Figure 2. What *null* and *alternative* hypotheses are being tested here?

   (b) (2 marks) What do you conclude from Figure 2, in the context of the data? Explain briefly.

   (c) (2 marks) Is the analysis of Figure 3 worth doing for these data? If not, explain why not; if it is, explain briefly what you conclude from it.

   (d) (2 marks) Look at Figure 4. (2017: this is an old-fashioned boxplot, but the interpretation is the same as a `ggplot` one.) From what you see in this Figure, are you surprised by your conclusions from part (b), or not? Explain briefly.

   (e) (2 marks) Does Figure 4 suggest that there are any problems with the analysis in this question? Explain briefly why or why not.

3. I have an R data frame called `mydata` containing two variables `u` and `v`. I also have a SAS data set called `mydata` containing the same two variables. I want to draw a scatterplot of `v` against `u`, and on that plot, draw the regression line for predicting `v` from `u`.

   (a) (4 marks) Give code for producing this plot in R.

   (b) (4 marks) Give code for producing this plot in SAS.

4. Moissanite is a popular abrasive material because of its hardness. It has another important property, elasticity. A mixture containing moissanite was compressed at each of eleven different pressures, and the compressed volume was recorded each time. Some analysis and plots are shown in Figures 5 through 7 in the booklet of code and output.

   (a) (2 marks) In Figure 5, a linear regression model is fitted. What do you conclude from the value `2.83e-10` (2.83×10^{-10}) in that Figure? Explain briefly. (This number appears twice in Figure 5. You may describe your conclusion from either one.)

   (b) (2 marks) What is R-squared for this regression? What does that tell you? Explain briefly. (Don't look at the adjusted R-squared here.)

   (c) (3 marks) What do you learn from Figure 6? (2017: this is a plot of residuals vs. fitted values with a smooth trend.) Is it therefore worth trying the analysis shown in Figure 7? Explain briefly.

   (d) (2 marks) Is the second regression, in Figure 7, a statistically significant improvement over the first one? How can you tell? Explain briefly. (You don't need any more output than is already there.)

   (e) (2 marks) Are your answers to parts (b) and (d) consistent with each other, or inconsistent with each other? Explain briefly.

   (f) (2 marks) What do you hope that the residual plot from the model of Figure 7 looks like? Why is that a good thing to hope for? Explain briefly.

5. An environmental researcher is studying carbon monoxide concentrations in air. A carbon monoxide level (in milligrams per cubic metre) is supposed to have a mean of 10 or less; if the mean is more than 10, it is considered to be too high. Carbon monoxide measurements typically have a standard deviation of 1.2.

   The standard way of testing for environmental carbon monoxide pollution is to collect 18 air samples and measure the mean carbon monoxide concentration in all of them. A hypothesis test is carried out, with a null mean of 10, and if the null hypothesis is rejected (in favour of the appropriate one-sided alternative), the location where the samples were taken is declared to be polluted with carbon monoxide.

   The researcher is interested in the "sensitivity" of this procedure: that is, how frequently the location is declared to be polluted *when it actually is polluted*.

   (a) (2 marks) Explain briefly how the P-value of the test *does not* assess the researcher's interest.

   (b) (3 marks) Explain precisely what the output in Figure 8 tells us. You need to talk about carbon monoxide levels in your answer. (I am looking for what you learn *overall* from this output, so I do not need to know what each line of code is doing.)

   (c) (2 marks) Figure 9 shows a second analysis. What do you learn from this one?

   (d) (2 marks) Explain briefly how Figures 8 and 9 give results that are consistent with each other.

6. 2017: this is a randomization test, which we didn't do. Your can skip this. How does the brain respond to sounds? A researcher hypothesizes that the brain responds differently to "pure tones" (generated by a computer) than to recognizable sounds. 37 macaque monkeys were anesthetized, and pure tones and monkey calls were fed directly to their brains using electrodes. Response to the stimulus was measured by the firing rate (electrical spikes per second) of neurons in various areas of the brain. Note that each monkey received both kinds of sounds, in randomized order.

   Some of the data are shown in Figure 10. Some analysis is shown in Figures 11 through 14.

   (a) (2 marks) What, in words, is the number `obs` that I calculate in Figure 11?

   (b) (3 marks) In Figure 12, a function is defined. In that function, `abs(x)` takes the absolute value of `x`, that is, its negative sign (if it has one) is removed. The last line of code shows some examples of running that function. What, in words, is that function doing, and how is it doing it?

   (c) (3 marks) What is the code in Figure 13 doing? (Three things, one for each of the three lines.) Include in your answer for the last line a brief explanation of *why* we want to do this.

   (d) (2 marks) What does the histogram in Figure 13 tell you about the P-value of your test? Explain briefly.

   (e) (3 marks) Obtain a P-value from the information in Figure 14. What do you conclude from it, in the context of the original problem?

7. What does it cost to transport something by truck, and what does that depend on? In 1980, several states had rules controlling trucking within that state, rules that were later removed. Florida was one of the first states to institute a "deregulation" policy. We have data on 134 shipments made by a particular Florida carrier. The response variable of interest is the price charged per ton-mile (`pricptm` in the data set). The logarithm of this variable was previously found to have a more nearly linear relationship with the other variables. These are:

   - `mileage`: distance that the shipment was shipped.
   - `shipment`: weight of the shipment in tons.
   - `pctload`: how full the truck was, in percent of maximum loading.
   - `origin`: `MIA` (Miami) or `JAX` (Jacksonville).
   - `dereg`: `YES` (rules have been removed, deregulation applies) or `NO` (rules are still in effect, deregulation does not apply).

   (a) (3 marks) I obtained the original data as a SAS permanent data set called `trucking`, which I stored in my home folder on SAS Studio. I am going to do a regression predicting log-price per ton-mile from the other variables, which need to be numbers. In Figure 15, I create a new data set. Explain what each of the four lines of code in the `data` step (not including the `data truck2` line) are doing.

   (b) (2 marks) Figure 16 contains a regression. In that regression, the slope coefficient for `shipment` is 0.2245. How do you interpret that number, in the context of the data?

   (c) (2 marks) Is it more expensive or less expensive to ship an item when deregulation is in effect (as compared to when deregulation is not in effect), all else being equal? Explain briefly, using the output in Figure 16.

   (d) (1 mark) Figure 17 contains a second regression. How is this regression different from the one in Figure 16?

   (e) (3 marks) Which regression do you prefer, the one in Figure 16 or the one in Figure 17? Explain briefly, giving *two* reasons for your answer, one of which is based on P-values.

   (f) (2 marks) What additional analysis would you like to do before you present your preferred regression model to the president of this trucking company? How will this improve your presentation? Explain briefly.

8. In Question 2, we looked at an analysis of variance of blueberry yields. Unfortunately, the farmer provided us the data in the form shown in Figure 18.

   (a) (2 marks) Explain briefly how the data in Figure 18 are not "tidy" as we defined it in class.

   (b) (3 marks) Give code, using one or more functions from the `tidyverse`, that will produce a tidy data frame `blueberry`. As input, use the data frame that was read in from the data file (Figure 18). This data frame is called `blueberry`.

   You can assume that `library(tidyverse)` has already been run.

9. 2017: this question is about detecting multivariate outliers, which we didn't do this year.

   Some data were collected on life expectancies in 38 countries. (The life expectancy is the number of years a baby born in the year of the survey can be expected to live.) Specifically, the variables were:

   - `lifeexp` overall life expectancy
   - `logpertv` logarithm of number of people per TV set in the country
   - `logperdr` logarithm of number of people per family doctor in the country
   - `lifeexpf` life expectancy for females
   - `lifeexpm` life expectancy for males

   The logarithms were taken because the original variables were very right-skewed. The data, with these five variables plus the `country` as text, are stored in a data frame called `life`.

   Figure 19 shows summaries of the variables in the data set, and correlations between all pairs of variables in the data set. I removed the first column in each case; it contains the names of the countries, so there is no point in including it in the numerical summaries.

   Figure 20 shows some R code to calculate leverages.

   (a) (3 marks) Why would I want to calculate leverages for this data set? What would leverages tell me that boxplots (or five-number summaries) would not? Explain briefly.

   (b) (2 marks) What is the reason for calculating the variable called `cutoff` in Figure 20? What is the reason for the numbers 5 and 38 being used in the calculation? Explain briefly.

   (c) (2 marks) Explain in words what the last line of code in Figure 20, with the two "pipes" `%>%`, does.

   (d) (2 marks) What specifally is it about Ethiopia that makes it appear at the end of Figure 20?

   (e) (2 marks) What is it about Sudan's `logperdr` and `logpertv` variable values that makes it appear at the end of Figure 20? I am looking for a specific thing for the second mark.

10. (7 marks) A study was conducted to evaluate the performance of a diesel engine run on three different types of fuel. The response variable was called the Mass Burning Rate, and it was thought to depend on both the Fuel type and on the Brake Power. The data are shown in Figure 21.

    The data have been read into a data frame called `synfuels`.

    Give R code to create a plot of these data according to the following specifications: there should be a scatterplot of Mass Burn Rate against Brake Power, with the points being different colours and different plotting characters according to the Fuel. There should be a legend enabling the reader to tell which points on the plot correspond to which Fuel type. In addition, there should be a lowess curve for predicting Mass Burn Rate from Brake Power, (2017) for each fuel, (2015) for all the fuels together, shown in brown. Finally, the plot should have the title "Mass Burn Rate for different Fuels", in larger than normal red text.

11. Soldering is a process in which two or more metal items are joined together by melting and flowing a filler metal (solder) into the joint, the filler metal having a lower melting point than the metal being joined. Antimony is sometimes added to tin-lead solder to replace the more expensive tin and to reduce the cost of soldering. The question then is whether adding antimony reduces the strength of the solder joints. An experiment was conducted to assess the strength of solder joints under various combinations of conditions: the amount of antimony added (0, 3, 5 and 10 per cent), and the cooling method: AB: air-blown, FC: furnace-cooled, OQ: oil-quenched, WQ: water-quenched.

    Under each combination of experimental conditions, three solder joints were tested. The measured strengths of these solder joints are labelled `s1, s2, s3` in the data set. The entire data set is shown in Figure 22 in the booklet of code and output.

    Your aim is to give code to extract and/or summarize parts of the data set as described below. You may use either R's `dplyr` tools or SAS to do this, whichever you like, but once you have made your choice you must use that same choice throughout the question. If you choose R's `dplyr` tools you may *not* use `aggregate`. If you choose SAS you need to show how to obtain a new data set (if necessary) satisfying the conditions. (You don't need to give code to list out any new data sets that you create.) Assume that the data have been read into an R data frame or a SAS data set (as appropriate) called `tinlead` with variable names as the column names shown in Figure 22.

    (a) (1 mark) To solve this question, I am going to use **R** / **SAS** (circle one).

    (b) (2 marks) Two of the variables are antimony percents and the strength measurements from the first solder joint. Show only these two variables for all the observations.

    (c) (2 marks) Show all of the variables *except* for cooling method. Do this *without* naming all the other variables. (Show all of the observations.)

    (d) (2 marks) Show all of the variables for the joints that were air-blown.

    (e) (3 marks) Show the variable `s2` for the joints that were water-quenched and had antimony greater than 4 (percent).

    (f) (3 marks) Find the mean value of `s3` for each method. (In SAS, assume that you are doing all the parts of this question in sequence.)

    (g) (4 marks) Find the mean value of `s3` for each method, but only for those observations where the antimony percentage is greater than 1.