MATH 4939 Final Exam April 20, 2020, 9 am to 12 noon

Instructions:

This exam is invigilated over Zoom.

If necessary, rename yourself on Zoom so your name matches your name on your student card.

Keep Zoom active and keep your video on while you write the exam. If you have questions, post them as a private question to the host. The instructor will respond directly to you as soon as possible although there might be a delay if there are other questions ahead of yours. Keep your audio on so you can hear any announcements made during the exam.

An invigilator may ask to see your student card at some time during the exam.

Each student writing this exam gets a similar but different set of questions.

The marks add up to 200. Questions are grouped into pages of related questions. The total marks for the questions in each page is shown at the top of the page.

This is an open-book, open-web exam in which you may use any resource EXCEPT any help originating after the start of the exam from any intelligent entity except yourself. You MAY NOT communicate with anyone except the Zoom host and co-hosts (the instructor and and invigilator) while you are taking the exam.

Whenever you use information from a web site or from the textbook, you must cut and paste the URL of the web site or cite the page of the textbook (e.g. Utts p.45) into the answer to your questions. There is no grade penalty for citing a source. Failing to cite a source is a breach of academic honesty. It is not necessary to cite information taken from your personal notes. Do include URLs to course documents or videos.

Try all questions. There are 3 types of questions:

1) Multiple choice questions with only one best answer. You get full marks for a correct answer, 0 for

an incorrect answer.
2) Open answer questions: either text or numerical answers, you simply type the answer in the answer box.
3) Questions with open answers that require supporting diagrams or rough work. Type your final answer in the box. Draw the diagram or do the rough work on one or more sheets of paper as you progress through the exam. You will photograph these sheets at the end of the exam and send them as a private post on Piazza to Instructors within 15 minutes after the end of the exam. Alternatively you can email them to georges@yorku.ca

Before starting the exam, make sure that you have at least 5 blank letter-size sheets of paper for scratch work and to draw diagrams as required during the exam.

Within 15 minutes after your finish the exam (by pressing the Done button at the end of the exam), you must photograph and upload your rough work and diagrams to Piazza as a private post to Instructors or, if this is not feasible, email the photographs to the instructor: georges@yorku.ca within 15 minutes of finishing the exam.

You are free to choose not to write this exam and, instead, you may choose to

- **write the exam in-class when in-class meetings resume at a future date,**
- **write a deferred online exam at a later date to be determined (no later than September)**

You may make this choice even if you have started writing this exam by finishing the exam and notifying the instructor through Piazza of your intention as soon as possible.

MATH 4939 Final Exam April 20, 2020, 9 am to 12 noon

Identification:

Family name as in York records

Given name(s) as in York records

York Student Number

York e-mail address (ending in yorku.ca)

Phone Number

Read each statement below and check the boxes to indicate that you have read and accept each statement before continuing with the exam:

- [ ] I understand that this is an open-book and open-web exam but I agree not to knowingly use information created and provided by anyone other than myself or an instructor after the start of this exam.
- [ ] Whenever I use information from a web site or from the textbook, I will cut and paste the URL of the web site or cite the page of the textbook (e.g. Utts p.45) into the answer to my questions. There is **no grade penalty for citing a source**. It is not necessary to cite information taken from my personal notes. It is necessary to include a URL from a course document or video.
- [ ] **Using a source without citing it is a breach of academic honesty.**
- [ ] After starting the exam I may not communicate about this exam with anyone except privately with the instructors. Communicating with anyone else about the exam will be considered a breach of the code of academic honesty and is subject to academic penalties.
- [ ] I have at least **5 blank sheets of paper** (approx. 8-1/2" x 11") that I will use for rough work and diagrams. I will photograph the sheets I have used at the end of the exam and **post them on Piazza as private posts to the "Instructors" within 15 minutes** of the end of the exam. Alternatively, I will email the photographs to georges@yorku.ca within 15 minutes of the end of the exam.
- [ ] Numerical answers should be given to at least **3 significant digits** or at least 2 digits beyond the decimal point whichever is more accurate.
- [ ] I will leave my video on and my microphone off throughout the exam. If I have a question during the exam, I will send a chat message to the host in Zoom. An instructor will contact me through the chat window as soon as possible.
- [ ] I have my York student card which I may be asked to show during the exam and which I will photograph and upload at the end of the exam together with my work.

After checking each paragraph above, press Next to proceed with the exam

[30] Write an essay on variable selection strategies in statistical analysis taking into account the factors discussed in our course. What major factors are relevant and how do the affect the choice of strategy?

[10] Suppose a test for Covid-19 use has a specificity of 95% and a sensitivity of 95%. Does this means that the test is incorrect 5% of the time.

[20] Therefore, if Carla takes the test and the result is ‘positive’ (i.e. the test indicates that Carla has Covid-19) the probability that she does not have it is only 5%. Is this statement true or false. Discuss thoroughly.

[20] Discuss this statement: "In a multiple regression, if you drop a predictor whose effect is not significant, the coefficients of the other predictors should not change very much, nor should the p-values associated with them."

[20] In a study of country X, it was found that the real wages (adjusted for inflation) of people with only a high school degree have gone down 3% and the real wages of people with a bachelors university degree have gone down 1% between the year 2000 and the year 2018.

However the real wages of the combination of these two groups together have gone up 3% in the same period of time.

Is this even possible or must there be an error in these data? Explain clearly either why this is impossible or why it is possible and illustrate your answer with an appropriate diagram. Draw the diagram on a **sheet of paper to be uploaded at the end of the exam.**

This page is worth 20 marks

[5] Suppose the correlation between (sorry! grades again) first year GPA and second year GPA is 0.7. Assuming a close to linear relationship between GPAs in first year and in second year, what is the approximate average z-score in second year of students who had a z-score of 1.5 in their first-year GPA?

[5] What is the average first-year z-score of students who have a z-score of 1.5 in their second-year GPA?

[10] It seems that, on average, students who do well in first year do better than average in second year but not quite as well as they did in first year. And students who do better than average in second year did, on average, better than average in first year but not quite as well as they did in second year. Whichever way you go, from first year to second year, or from second year to first year, it looks like the grades are getting closer to the average.

Is this a contradiction? Is there an explanation for it? If you need a diagram to help explain it, go ahead a draw one and submit it by **uploading it at the end of the exam.**

[20] Consider a vector of strings containing names of people. Each string contains one name which can be in various formats: ‘Mary Ellen Brown’ (i.e. first name followed by middle name if any) and by last name), ‘Brown, Mary Ellen’ (last name, followed by first and middle names), ‘Paul Smith’ (if there is no middle name) or ‘Smith, Paul’.

Write a function in R that takes two arguments: a vector of such strings and a single character string. The function counts how often the second argument occurs as a last name in the vector that is the first argument.

[25] Prove that the following are equivalent:

a) A is a variance matrix

b) A is non-negative definite

c) all the eigenvalues in the spectral decomposition of A are non-negative

d) there exist a matrix B such that A = BB'

Note: you may use the spectral decomposition theorem: For any symmetric matrix A, there exists an orthogonal matrix G and a diagonal matrix D such that A = GDG'.

The Cowles data frame has 1421 rows and 4 columns. These data come from a study of the personality determinants of volunteering for psychological research.

Neuroticism (neuro) is classified in three levels: low, medium and high.

Extraversion (extra) is measured on a scale that ranges from 1 to 25.

The purpose of the study is to explore some personality predictors of the willingness to volunteer and how the prediction differs between men and women. This is some output from an R script:

The Cowles data frame has 1421 rows and 4 columns. These data come from a study of the personality determinants of volunteering for psychological research. Neuroticism (neuro) is classified in three levels: low, medium and high. Extraversion (extra) is measured on a scale that ranges from 1 to 25. The purpose of the study is to explore some personality predictors of the willingness to volunteer and how the prediction differs between men and women. This is some output from an R script:

```
> library(spida2)
> library(car)
> head(Cowles)
   neuro extra    sex volunteer
1 medium    13 female        no
2    low    14   male        no
3    low    16   male        no
4    low    20 female        no
5    low    19   male        no
6    low    15   male        no
> dim(Cowles)
[1] 1421    4
> Cowles %>% tab(~neuro+sex)
        sex
neuro    female male Total
  high       47   25    72
  low       234  280   514
  medium    499  336   835
  Total     780  641  1421
> Cowles %>% tab(~volunteer)
volunteer
   no   yes Total
  824   597  1421
> fit <- glm(volunteer ~ neuro*extra*sex, Cowles, family = binomial)
> print(fit)

Call:  glm(formula = volunteer ~ neuro * extra * sex, family = binomial,
    data = Cowles)

Coefficients:
              (Intercept)                   neurolow                neuromedium
                  -0.27178                   -0.61174                   -0.87311
                    extra                    sexmale              neurolow:extra
                  -0.00257                    5.45731                    0.06239
       neuromedium:extra          neurolow:sexmale        neuromedium:sexmale
                   0.07467                   -6.81494                   -5.24577
           extra:sexmale     neurolow:extra:sexmale  neuromedium:extra:sexmale
                  -0.43738                    0.52151                    0.39408
```

[15] Consider the first row and the fourth row of the following Anova table.
Specify both the null hypothesis and the alternative hypothesis being tested in each of these two rows. Use R's linear model formulas to express both the null and the alternative hypotheses for each of the two rows.

```
> Anova(fit)
Analysis of Deviance Table (Type II tests)

Response: volunteer
               LR Chisq Df Pr(>Chisq)
neuro            2.1124  2   0.347769
extra           21.1480  1  4.251e-06 ***
sex              5.6187  1   0.017770 *
neuro:extra      8.9289  2   0.011511 *
neuro:sex        1.9909  2   0.369556
extra:sex        0.1109  1   0.739082
neuro:extra:sex  9.4006  2   0.009092 **
```

[10] Contruct a linear hypothesis matrix to test whether neuroticism has the same relationship with the log odds of the propensity to volunteer in men as it has in women.

[10] Construct a linear hypothesis matrix to test whether there is no difference between the behaviour of female subjects with 'low' neuroticism and with 'medium' neuroticism.

This is the last page of the exam. If you have time left you can go back and check your answers. You can try the bonus question below if you wish. When you are finished press Done. Within 15 minutes of ending the exam, **upload images of your student card and of the work that needs to be uploaded as a private post(s) to the instructors in Piazza**, or if you run into problems doing that, send them in email messages to georges@yorku.ca.

Have a good, safe and healthy summer.