STATISTICS MASTERS/Ph.D.-QUALIFYING EXAM: TAKE HOME

January, 2020

General Directions: The answer to each problem should be presented as a summary. It must be word processed, double spaced, and identified with your "Code Word" in the header on each page; *do not include your UNM ID or name.* The report for each problem should be no longer than 3 pages. You may create brief, **well-organized** appendixes for each problem. Remember, the appendix is not the report and *will only be examined* if the report draws interest to it and material is easy to find. In your data analysis, raw uninterpreted computer output will be graded as the dross it is. You should have a caption for every figure and table; one that describes it and briefly tells the reader why it is of value. Organize your sections to justify the validity of what you uncovered and the methods you used to uncover it. We want a summary of what you think is important, not a diary of how you spent your time. Remember that even the best data analysis is worthless if your reader does not understand it, so you are being graded on presentation as well as statistical content. You may use your course notes as well as any available books or web resources on general statistical methods for the exam. You may not consult any other person when working on this exam or discuss your exam with anyone else, regardless of whether or not the person is taking the exam nor are you allowed to use the internet to find analyses of these data. (No matter what you think, you will not find *these* problems on the web.) Questions can be directed to Ronald Christensen, fletcher@stat.unm.edu. Email solutions by **3 PM, Fri Aug 16, 2019** to Ana Parra Lombard, aparra@math.unm.edu, Department of Mathematics and Statistics, University of New Mexico. Please do not turn in a physical copy of your solutions.

1. https://www.math.unm.edu/~fletcher/pollution2020.txt contains the data. They are from various years in the early 1960s and relate air pollution to mortality rates for various standard metropolitan statistical areas in the United States. The dependent variable y is the total age-adjusted mortality rate per 100,000 as computed for different metropolitan areas. The predictor variables are explained in the text contained in the file. Find a good explanatory/predictive model for mortality. (This being a statistics program, good models are good statistical models. You can use a machine learning model – at your own risk – if you are capable of showing that it has good statistical properties.)

2. The ocean linear Lusitantic sank on May 1, 1914 after being struck by an ice-torpedo. Survival of passengers was related to their sex-age and their onboard status. Onboard status is either a crew member or a passenger and passengers were subdivided by the amount they paid for their tickets. First Class was the most expensive and Third Class (steerage) was the least expensive. (No discounts for children!) The numbers of survivors and the total numbers on board are given below.

		Saved	Aboard
Women	First Class	140	144
Women	Second Class	80	93
Women	Third Class	76	165
Women	Crew	20	23
Children	First Class	6	6
Children	Second Class	24	24
Children	Third Class	27	79
Children	Crew	7	8
Men	First Class	57	175
Men	Second Class	14	168
Men	Third Class	75	462
Men	Crew	192	885
Totals		1513	2224

Use your knowledge of statistical modeling to help explain these data. In addition to analyses and comments arising from your own curiosity, you must address the following items as part of your write-up. (We have tried to make these as generic as possible so do not try to infer things about the data from the questions being asked.)

(a) What is the your choice for a dependent variable?

(b) What tools do you have for analyzing such dependent variables?

(c) What are appropriate plots of the data and what do those plots suggest to you?

(d) Write out your best full statistical model (in notation, defining the notation you use) and state the model assumptions.

(e) Fit the model written in the previous part, summarize the fit, and, to the best of your ability, assess and address deviations from model assumptions. (Depending on your approach, this may need to be an iterative process in which case summarize each model fit and discuss the evidence guiding decisions to arrive at your final model. In other words, if model assumptions are not met, try to address those. If you can not address unsatisfied model assumptions, admit to this and continue as though the model assumptions are met.)

(f) State appropriate, interesting reduced models or parametric hypotheses and evaluate them statistically. Interpret the results. (This will probably involve testing but other statistical evaluations are possible.)

(g) Going back in time and based on these data, if you had limited financial resources and had to select one member of your family (mother, father, or young sibling) to travel on an ocean liner, what is the least expensive choice **statistically** consistent with maximizing their chance of survival? (Getting them a job on the crew is a possibility.)