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

August, 2023

General directions

Complete both problems in this exam. Your report is to be typed, double spaced, no smaller than ten-point font with one-inch margins, and should be identified by your "CODE WORD" in the header on each page; *do not include your name or UNM ID number*. Each problem is to be no longer than four pages, and an additional four-page appendix is allowed for each problem but will be examined only at the discretion of the graders; the better constructed your appendix with cross-references from the text, the more likely it is to get examined. In your data analysis, raw uninterpreted computer output will be graded as the dross it is.

Write your answers completely, but concisely. Insert tables and figures to support your points. Tables and figures should be well-labelled and cross-referenced from text, such as, "in Table 1 …", or if in the appendix, "in Table A1 …" and each should have a caption that describes it and briefly tells the reader why it is of value. Figures should include appropriate symbols suitable for black-and-white reproduction (that is, avoid use of color if possible; consider symbols, line types, and distinct shades of gray to distinguish categories or values).

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.

As necessary:

1. Plot and describe the data (that is, plot all the individual observations, in addition to summaries of data you might present with the results, such as the mean and confidence intervals).
2. Clearly define population parameters and sample statistics.
3. Clearly specify hypotheses tested and explicitly state the associated model at least once (i.e., write the model equation).
4. Define and assess method assumptions.
5. Write a coherent evidence-based conclusion that a layperson can understand.

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.

Any points of clarification can be directed to Prof. Erik Erhardt, erike@stat.unm.edu.

Email solutions as a pdf file by **3 PM, Fri Aug 18, 2023** 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.

Problem 1, Pollution

Data from The U.S. Department of Labor Statistics gives the properties of 60 Standard Metropolitan Statistical Areas (SMSA, a standard Census Bureau designation of the region around a city) in the United States, collected from a variety of sources. The data include information on the social and economic conditions in these areas, on their climate, and some indices of air pollution potentials.

city : City name
JanTemp : Mean January temperature (degrees Farenheit)
JulyTemp : Mean July temperature (degrees Farenheit)
RelHum : Relative Humidity
Rain : Annual rainfall (inches)
Mortality : Age adjusted mortality
Education : Median education
PopDensity : Population density
%NonWhite : Percentage of non whites
%WC : Percentage of white collar workers
pop : Population
pop/house : Population per household
income : Median income
HCPot : HC pollution potential
NOxPot : Nitrous Oxide pollution potential
SO2Pot : Sulfur Dioxide pollution potential
NOx : Nitrous Oxide

Develop a regression model to determine whether the amount of pollution is associated with the population demographics of the city adjusted for environmental factors. That is, are people with less privilege subjected to greater pollution?

As the response, scale each of the 4 pollution variables (HCPot, NOxPot, S02Pot, NOx) so the maximum value observed in the dataset is 1, then take the average of the scaled pollution values so the total pollution is a value between 0 and 1. Do not use Mortality as a predictor of total pollution.

Download the data from
https://math.unm.edu/sites/default/files/files/qual-exams/stat/unm_exam_202308_stat_qual-takehome_dat1.csv.

Problem 2, Tobacco plant virus

A plant pathologist wanted to examine the effects of two (2) different strengths of tobacco virus (weak and strong) on the area of lesions (mm^2) on three (3) common types of tobacco leaf (Brightleaf, Shade, and Wild) at two (2) stages of growth (vegatative and flowering). She knew from pilot studies that leaves were inherently very variable in response to the virus. In an attempt to account for this leaf to leaf variability, both treatments were applied to each leaf. For each leaf type and stage of growth, eight (8) individual leaves were divided in half, with half of each leaf inoculated with weak strength virus and the other half inoculated with strong virus. Assess the effect of virus strength (difference between strong and weak) on lesion size conditional on leaf type and stage of growth.

Please address the following questions as part of your write-up:

1. What statistical design is being used, and why? Could a better design have been used, and why or why not?
2. What is/are the outcome(s)/response(s)?
3. What is/are the treatment(s)?
4. Is there blocking? If so, what is/are the block(s)?
5. What is/are the nuisance factor(s) to be "averaged out" in the design?
6. Plot the data (not only summaries of the data, but all of the values) in a way that helps you understand what the effects are.
7. Fit the simplest linear model that adequately explains the data in this experiment.
   - Write the model using the same parametrization that appears in the computer output; use of indicator variables may be necessary.
   - Explain the terms in the model, and list all the assumptions.
   - Fit the model parameters.
   - How many degrees-of-freedom are allocated to each source of variation?
   - State and assess model assumptions. (If assumptions are not met, try to address that. If you can not address unsatisfied model assumptions, state this and continue as though the model assumptions are met.)
8. State and conduct statistical tests for the parameters, and interpret the test results.
   - Examine the main effects, interaction terms, and do multiple comparisons of the interested treatments or combination of treatments.
9. Write a succinct, coherent, and complete summary of your analysis that addresses the original goal of the experiment and discuss anything else of interest.

Download the data from

https://math.unm.edu/sites/default/files/files/qual-exams/stat/unm_exam_202308_stat_qual-takehome_dat2.csv.