2022 Methods Qualifying Exam

Jack Tubbs

August

Contents

Instructions

The BOX folder for your exam contains four (4) csv data files that you will need. I have not included any RStudio or SAS files for these data. Nor have I included an Overleaf file. Read the README file before you begin.

Your exam consists of five parts. The first three involve the analysis of simple data sets. The fourth involves the analysis of a slightly more difficult and somewhat open ended solution.

The final part of the exam does not involve any analysis as there is not any data for the problem. Rather, you will present your proposal for the analysis that answers the problem at hand[1]. For example, you might believe that the solution involves doing analysis, **A**, for which you will need to assume conditions, **B**. Tell me what you would need to do to verify that conditions **B** hold. What would you hope to determine from your analysis? Can you foresee potential issues that might arise since your study involves human subjects? Your solution will have strengths and weaknesses. Describe each and state why you think the advantages of your proposal out weight the disadvantages [2].

When answering the questions you can use either SAS or R. I do not need nor want to see the same analysis done using both SAS and R. Do Not exclusively use SAS or R. It is your choice as to how much of each that you use. Your choice will effect your grade! I do not want to see your failed attempts to answer the question at hand, nor do I need to see all the steps that you needed to arrive at your final answer (e.g., model selection procedures).

Answer the questions and present your solutions in a document similar to what I have produced using LaTeX in Overleaf (or your choice of a TeX/LaTeX format). A Word-type document is Not Acceptable! In some cases, the question has been intentionally left vague or open ended. If you encounter these, then tell me what you intend to do, why you did it, and what you found. Do Not redefine your problem so that it is either

[1]You have hear people say, "don't just sit there, do something". Here, I am saying "Don't do something, just sit there and think before you act. Tell me what you are thinking!

[2]This problem is similar to one that I gave in 2021, for which the collective performance of that cohort were uniformly pitiful! I expect better from you. Your discussion should reflect that you are a statistician and not an uninformed creative writer. The absence of data does not mean you don't have models based upon the data and the analysis that you plan to perform.

trivial or overly complex. Use the methods that were presented in STAT 5380/5381. There shouldn't be any Bayesian solutions. Save those for your PPP!. The level of difficulty for this exam is consistent with what one would expect of an MS in Statistics. You are welcomed to use any of the resources given in the methods sequence (or any other resource that does not consume Oxygen).

It is your responsibility to clearly communicate your solution to each of the problems. Make use of cites and labels when referring to output such as tables and graphs or figures. If I have to search or guess at your solution then it is wrong!

You have nearly 7 days in which to work on this exam! Budget your time and use it wisely. I suspect none of the questions can be answered and documented in one sitting. I attack problems of these type by thinking in terms of three disjoint but related tasks. 1. What is the problem, what is my initial approach and do I have the needed resources in place to do what I think will be needed. For example, suppose I think the problem involves ANOVA for which I intend to use SAS. What do I need? How will I try to do it?, etc. 2. The second task involves the analysis. Did my program do what I needed it to do? What are the results? Do I need more since some unexpected issues arose? 3. The third part consists of documenting what I found. Is my explanation clear? Is it relative to what I was attempting to find? etc.

My experience with activities such as this exam is that (10 - 15)% of your time will be spent on part 1. Yet, you can do this anywhere. As you are walking, riding a bike, or drinking coffee, think about your plan. You will be surprised at how valuable this type of activity can be. The analysis should take about 50% of your time. The writing will take the remainder of your time. It is amazing how long it takes to write up your results. I suspect none of you will have an acceptable first draft! So edit and rewrite....

Exam

Part 1 – Alex Rodriquez Home Runs

Alex Rodriguez (known to fans as A-Rod) was the youngest player ever to hit 500 home runs. The file contains the number of home runs hit by A-Rod during the 1994–2016 seasons.

1. Describe the data using simple descriptive methods (both numbers and graphs). Are there unusual values for the number of home runs? What characteristics of these data are you looking for? What happened in 2013-14? What influence do these data have upon your results? Explain. How about 2015?
2. Suppose that Y is the number of home runs in a given year. Is Y normally distributed? Justify, your answer.
3. Suppose Y is not normally distributed, how would you make Y more normally distributed? What would you do? Did it work? (show me)
4. Find a 95% CI for the statistic which is the 75^{th} percentile for the number of home runs hit in a given year. There may be several approaches to this problem, pick one and show me the results.

Part 2 – Blood Pressure by Age Data

Does blood pressure, on average, change with age? If so, what can you say? The data consists of two categorical variables: Blood pressure categorized as High, Normal, Low, and Age categorized as under 30, 30-49, and over 50.

1. Describe the data using descriptive methods.
2. State the hypothesis for testing the relationship between AGE and BLOOD PRESSURE? What conclusion can you reach? Assuming you used the CMH procedure what is the degrees of freedom that you used? Why?
3. Are these conclusions for populations or individuals? Explain.

Part 3 – Hand Washing Experiment

A student decided to investigate just how effective washing with soap is in eliminating bacteria. To do this she tested four different methods—washing with water only, washing with regular soap, washing with antibacterial soap (ABS), and spraying hands with antibacterial spray (AS) (containing 65% ethanol as an active ingredient). Her experiment consisted of one experimental factor, the washing Method, at four levels.

She suspected that the number of bacteria on her hands before washing might vary considerably from day to day. To help even out the effects of those changes, she generated random numbers to determine the order of the four treatments. Each morning, she washed her hands according to the treatment randomly chosen. Then she placed her right hand on a sterile media plate designed to encourage bacteria growth. She incubated each plate for 2 days at 36°C, after which she counted the bacteria colonies. She replicated this procedure 8 times for each of the four treatments.

1. Analyse the data and describe your findings. Were any of the washing methods preferred? If so, what are the differences? Make sure you describe your approach to the analysis and what assumptions that you made.
2. State any issues that you "see" with this investigator's approach. What modification would you suggest, if any? Keep in mind that she probably has more time than money for doing this experiment.
3. Since she "suspected the amount of bacteria on her hands might vary", what adjustments would you suggest to control for this confounding factor? Describe your new model.

Part 4 – Halloween Candies Study

The internet site Fivethirtyeight.com conducted an online survey of Halloween candies. Respondents were asked to rate which candy they prefer out of a random pairing of 2 candies selected from a full list of 85. About 269,000 candy comparisons were presented to respondents spanning 8,371 unique IP addresses. From these comparisons, an ultimate preference score was calculated as the percentage of times each candy was chosen as the better of the pair, recorded as the win percent in the data. Additional independent variables included sugar percentage, price, and the presence of various candy components. The data in this file contains the results for 85 candy choices.

1. Model the win percentage as a continuous normally distributed variable and determine which factors are used when creating a predictive model. There are multiple approaches, including multiple regression and CART/RF methods. Use methods from both areas.
2. Determine which independent characteristics of a HALLOWEEN candy best predict the event that a candy falls into the top 25% of the win percentage category.
3. (New method) Repeat # 1 above assuming that the win percentages are continuous decimal ($win_{dec} = win/100$) variables on the interval (0,1). Use the GLM for the Beta model[3]. Summarize your findings. How do these results compare with what you found in part # 1 above? There is no reason to repeat the machine learning parts you used in # 1, why? Explain.

Part 5 – Experimental Study

As the lead statistician for the following clinical trial, provide your initial analysis plan (IAP) for the study in order to get external approval. There is no analysis since there is no data at this time. Your IAP should have clearly stated goals and objectives, including specific models for which the needed assumptions have or will be satisfied. Write as a phD statistician and not as a sophomore creative writer.

1. A clinical study is performed to determine the effectiveness of two new experimental drugs when compared to the current standard of medical care. The study consists of randomly assigning subjects to one of three groups, **A, B** and control **C** for which the clinical response of interest is **Y**. Subjects in groups **A** and **B** receive two different experimental drugs, whereas subjects in group **C** receive the current standard FDA approved drug. Measurements are scheduled to be taken at study onset (time for subject randomization into the three groups) and every 6 weeks for the entire 24 week study. All the clinical subjects enter the trial at day 1 of the study. Additional subject-wise covariates, **X** are available for which some are time independent (e.g., gender) and some are measured at each clinical visit (e.g., blood pressure and weight). Since the measurement of **Y** is time consuming, the study was conducted at 3 medical centers located in Minnesota, Florida, and Arizona. The enrollment is such that the available subjects in each group is about the same for the individual medical centers.

[3]There will likely be many occasions in the future in which you are ask to do something for which you have never been exposed or taught. Get busy or get another job!