STAT 530 – Midterm Exam (Take-Home Portion) – Fall 2024

Note: For this midterm exam, **you are not allowed to receive help from *anyone except me* on the exams.** For example, you may not talk to other students about the exam problems, and you may not look at other students' exams. Violations of this policy may result in a 0 on this portion of the exam, an F for the course, and/or punishment by the USC Office of Academic Integrity.

I will answer queries asking for clarification about the exam questions. Since this is an exam (and not homework), I will probably decline to provide very much help in solving the problems, but I am happy to clarify questions if necessary.

In addition, the computer programming/coding and writing of the answers on this exam must be done entirely by you − not with the help of any other individual or any AI program such as ChatGPT. You are welcome to use the textbook, course website, and other STAT 530 materials as aids in doing the problems. If you use other background sources (it is unlikely that you will need to do so), then you must cite the sources you used.

1. You are working as a statistical assistant for the United Nations, helping with some statistical analysis to use in a magazine publication that will describe aspects of a sample of UN member countries. The UN has gathered data on 166 countries, measuring 7 variables on each, which represent various characteristics of the nations: pop = Population, in numbers of people; area (in square km); oil_prod = Crude oil production in barrels/day); gdp = Gross Domestic Product per capita, in dollars/person; educ = education spending in % of GDP; roadways = Roadways per unit area, in km/sq km; net_users = lower bound on the country's fraction of Internet users, in % of population. Although each country's true fraction of Internet users will exceed this lower bound, you can use this variable for the analysis as it is given.

Some questions that the Secretary-General of the United Nations would like answered include:

1. Are there particular countries(s) that are highly unusual in terms of the measured characteristics? If so, identify them.
2. Are there notable associations/relationships between some of the variables? (if so, describe them)
3. Is there a way to graphically represent the raw data for the 166 countries (or more sensibly, a selected subset of the countries) and draw conclusions about the data set from such a graph?
4. Can we find a few indices that describe the variation in the data set using a lesser dimension than the original set of variables? If so, what are those indices? Is there a convenient interpretation of any of the indices?
5. Can we graphically display the data in a low number of dimensions using such indices? What conclusions about the countries (individual countries or groups of countries) can you draw from such a graph?
6. Are there any countries that are similar or different from each other in any aspects that are surprising to you? What useful information, e.g., for a public relations campaign, could be gleaned from this data set as related to this?
7. What are any other potentially interesting aspects of the data set that may be gleaned for these data?

Note that you will likely not have space in your report to answer all these questions fully. You should consider providing the answers in your report that are most illuminating about the data set; it's all right if you cannot address all the questions above. Use your judgment and statistical insight to provide the most enlightening report that you can.

You will type a 2-page report detailing your analysis of the data and your conclusions. Keep in mind that the report should be written for two audiences: the publications staff member of the secretary-general, who knows about world politics but is not an expert in statistics; and your supervisor, the head statistician, who will be judging you and deciding on your possible promotion based on the statistical competence of the report. Your report should be understandable and meaningful to both audiences.

You may include graphs and BRIEF computer output that illustrate and/or support your findings. (The graphs and output do not have to count as part of the 2-page length.) Do NOT include computer code within the main body of your report. This will be incomprehensible to the publications staff member and would only annoy her. You may include such code in an appendix if you wish.

The data for this problem and code to read it into R are given at the link "Countries Data" on the course web page.

2. You are working as a consulting statistician for a company that has a contract with a medical researcher. She has gathered data on 60 adult female patients for a diabetes study. The variables measured include health and demographic variables for the females. The 7 variables she has are:

X1 = Number of times pregnant
X2 = Plasma glucose concentration (based on an oral glucose tolerance test)
X3 = Diastolic blood pressure (mm Hg)
X4 = Triceps skin fold thickness (mm)
X5 = Two-Hour serum insulin (mu U/ml)
X6 = Body mass index (weight in kg/(height in m)^2)
X7 = Age (years)

Some questions that the researcher would like answered include:

(1) Are there individual females who are highly unusual (in any way) based on the measured health variables, X2 through X6? If so, identify their numbers.
(2) Are there notable associations/relationships between some of the variables? (if so, describe them)
(3) Is there a way to graphically represent the raw data for the 60 patients and draw conclusions about the data set from such a graph?
(4) Are there a small number of underlying characteristics of patients that the observed variables might be connected to? If so, determine how many latent characteristics there seem to be in this set of variables. Also, try to interpret them the best you can, with the aid of statistical techniques.
(5) Can we graphically display the data in a low number of dimensions using such latent traits? What conclusions about the patients (individual patients or groups of patients) can you draw from such a graph?
(6) What are any other potentially interesting aspects of the data set?

Note that you will likely not have space in your report to answer all these questions fully. You should consider providing the answers in your report that are most illuminating about the data set; it's all right if you cannot address all the questions above. Use your judgment and statistical insight to provide the most enlightening report that you can.

You will type a 2-page report detailing your analysis of the data and your conclusions. Keep in mind that the report should be written for two audiences: the medical researcher, who has a sense for numbers but is not an expert in statistics; and your own supervisor at the statistical consulting company, who will be judging you and deciding on your possible promotion based on the statistical competency of the report. Your report should be understandable and meaningful to both audiences.

You may include graphs and BRIEF computer output that illustrate and/or support your findings. (The graphs and output do not have to count as part of the 2-page length.) Do NOT include computer code within the main body of your report. This will be incomprehensible to the researcher and would only annoy her. You may include such code in an appendix if you wish.

The data for this problem are given at the link "Diabetes Data 60" on the course web page. There is a link to a data file with patient ID numbers and code to read the data into R and to create a data frame without ID numbers.

Grading Scale:

Each problem will be worth 17 points, for a total of 34 points. For each problem, your report will be graded based on Writing, Analysis, and Context. For example:

Writing (out of 5 points): How organized, clearly written, comprehensible, and grammatically correct is the report? Would the client reading this report be confident that it was written by an educated, well-trained statistical scientist? IMPORTANT: It's critical that this writing, while reasonably formal, sound as if it's been written by a real person. The reader wants to report to reflect YOUR intellect and YOUR personality.

Analysis (out of 7 points): Were the graphs and data analyses appropriate for the problem? Were the analyses carried out correctly? Were your statistical conclusions about the data set sensible and clearly justified by numerical or graphical evidence?

Context (out of 5 points): Were the questions answered in terms of the variables of the data set? Although you are not an expert in the field as your client is, have you attempted to frame your conclusions and interpretations in a subject-matter context rather than treating the data as simply a meaningless set of numbers?