STAT 530 – Midterm Exam (Take-Home Portion) – Fall 2025

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.

Data Set and Mini-report for the Take-home Portion

You are working as a statistician doing oversight for a food-safety and nutrition board, as part of a governmental investigation. The agency has selected 77 types of cereal produced by seven companies (A = American Home Food Products; G = General Mills; K = Kellogg's; N = Nabisco; P = Post; Q = Quaker Oats; R = Ralston Purina). Nine numerical nutritional characteristics have been measured for each cereal. These 9 variables are:
calories: calories per serving.
protein: grams of protein.
fat: grams of fat.
sodium: milligrams of sodium.
fiber: grams of dietary fiber.
carbo: grams of complex carbohydrates.
sugars: grams of sugars.
potass: milligrams of potassium.
vitamins: vitamins and minerals (this is discretely measured: 0, 25, or 100, indicating the typical percentage of FDA recommended)

In addition, the code on the webpage will produce a character vector of labels with the name of each cereal, another character vector naming the companies that produced the cereal, and a character vector with the type of the cereal ("hot" or "cold"). Finally, the code creates a variable called rating that is a subjective quality rating of the cereal (presumably by a Consumer Reports expert).

Some questions that the researcher would like answered include:

(1) Are there notable associations/relationships between some of the variables? (if so, describe them)
(2) Is there a way to graphically represent the raw data for some or all of the 77 cereals and draw conclusions about the data set from such a graph?
(3) Is there a way to graphically associate (some of) the numeric variables to (some of) the categorical variables in a way that illustrates patterns of interest across categories?
(4) Can you compare companies in terms of values for some or all of the numeric variables? Are there interesting differences between the companies' cereals that can be displayed using the variables (or combinations of variables) in the data set?
(5) Are there a small number of underlying characteristics of cereals that the observed numerical variables might be connected to? If so, determine how many latent characteristics there seem to be in this set of numerical variables. Also, try to interpret them the best you can, with the aid of statistical techniques.
(6) Can we graphically display the data in a low number of dimensions using such latent traits? What conclusions about the cereals (individual cereals or groups of cereals) can you draw from such a graph or graphs?
(7) 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 roughly 2-to-3-page report detailing your analysis of the data and your conclusions. Your report should be understandable and meaningful to the general public, to the leaders of the cereal industry, and to statisticians who will be reviewing your report. You may include graphs that illustrate and/or support your findings (such graphs don't count against the 3-page limit). Do NOT include computer code within the main body of your report. You may include such code in an appendix if you wish. The data for this problem are given at the link "Cereal Data for Midterm" on the course web page. There is a link to a data file with all variables. There is also a link with R code to read the data into R and to create a data frame of only the numeric nutritional variables, as well as to create individual vectors with several categorical variables and the numerical rating variable.

Grading Scale:

The take-home portion of the midterm will be worth a total of 24 points. Your report will be graded based on Writing, Analysis, and Context. For example:

Writing (out of 7 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 10 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 7 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?