Tutorial Problem 4: Simulation and Bootstrap

1. Use a simulation to estimate the standard error on the *median* for data sampled from the standard normal distribution (μ=0, σ=1) for four different sample sizes: n=10, n=50, n=250, and n=1250. Summarize the result as a table, and plot it graphically (n on the x-axis, standard error on the y-axis).

2. Compare the *numerical* result above to the *theoretical* result given in Lecture 3 by adding the curve corresponding to the expected standard error on the median (i.e., the equation for SE_{med} as a function of n, assuming σ=1: see slide 24) to the plot above.

For the remaining problems, you will be using the measurements in the data set weirddata.csv.

3. Plot these data as a histogram and, separately, as an ECDF.

4. Use three different normality tests to confirm that this data is not drawn from a normal distribution.

5. Suppose we are interested in a quantity which we'll call *QG*, which is equal to the geometric mean of the first quartile and third quartile. (The geometric mean of two numbers X and Y is equal to sqrt(X*Y).) Calculate QG for this sample using a custom R function.

6. Perform a bootstrap to estimate a 95% confidence interval on the value of QG for the *population*.

Tutorial Problem 5: Covariance of Radiation Damage

A study is conducted in which 45 objects, treated in one of three different ways (designated A, B, or C), are subjected to a high-radiation environment. After being removed from this environment, the amount of radiation received and the amount of damage to each object is recorded. The data are stored in `radiation.csv`.

1. Calculate the mean amount of damage for each treatment group (A, B, and C) and the standard error on each. Looking only at this information (*not* considering the radiation dosages), does there appear to be a relationship between the treatment group and damage?

2. Plot exposure (explanatory variable) versus damage (response variable) and colour-code (and/or symbol-code) by treatment. Include a legend.

3. Perform a linear model to examine the dependence of damage on radiation and treatment, including interactions between radiation and treatment. (Treat the relation as a simple linear one, without curvature.) Then answer the following questions:

(a) How many parameters are there in the model?
(b) Of these parameters, for how many can the null hypothesis of zero be ruled out?
(c) What does each parameter mean, in the context of your model?
(Explain using 1-2 short sentences, without relying on technical terminology.)

4. Does your model contain significant interactions? Provide a *single* p-value describing the collective significance of model interactions. Describe what the presence (or absence) of interactions implies about the nature between the response variable and explanatory variables in a single, nontechnical sentence.

5. Is your conclusion from part (4) consistent with what you found in part (3b)? Explain.

6. Generate another plot of exposure versus damage (you can re-use the code from part 2), but this time overplot the full ANCOVA linear model on the plot as three different lines, one for each treatment. The lines should be colour-coded to match the treatment each describes.

7. Which treatment group, if any, would you advise if you wish to minimize damage for an object expected to be subjected to a high-radiation (exposure > 10) environment?

Problem 6: "Fun" with p-Hacking: A Cautionary Exercise

The file nutrition.csv contains data from an online survey conducted by fivethirtyeight.com on medical conditions, lifestyle choices, and various other self-reported information, along with data from a detailed food intake questionnaire that asks about their eating/drinking habits regarding a wide variety of different foods and beverages. Columns can be grouped as follows:

Column 1: ID number of the respondent.
Columns 2-27: Characteristics of the respondent, including e.g. medical conditions, handedness, smoking habits, if they own a cat, if they are atheist, if they have a favourable or unfavourable relationship with their internet provider, if their belly button is an "innie" or an "outie", whether the film "Crash" deserved to win Best Picture at the 2006 Oscars, and many others.
Columns 28-283: Frequency and quantity of consumption of various foods/beverages (FREQ = frequency weekly; QUAN = standardized amount consumed per sitting).
Columns 284-381: Information on brand preferences and cooking/eating habits (for example, whether they add milk to coffee or not).
Columns 382-414: Frequency of taking certain nutritional supplements and how long they have been doing so.
Columns 415-419: Information on vitamin intake calculated from the supplement responses.
Columns 420-427: Information on types of supplements chosen.
Columns 428-429: The number of meals eaten per day and the number of snacks per day.
Columns 430-451: Behavioural information about how frequently and for how long they engage in certain activities.
Columns 452-458: Ethnicity data.
Columns 459-1093: Information on the average daily intake of certain vitamins/minerals and other substances, calculated from the other columns of the survey.

1. Perform a two-sample test to compare the average frequency at which breakfast sandwiches are consumed (using BREAKFASTSANDWICHFREQ, the first food intake column) between participants who have had cancer and participants who have not had cancer (cancer, the first characteristics column). Is the difference in means statistically significant?

2. Calculate the correlation coefficient (Pearson r) between cancer occurrence (converted to a binary 0 or 1) and the frequency of consuming breakfast sandwiches, and its standard error. Then calculate the associated p-value. How does it compare with the mean comparison p-value from part 1?

3. Generalize the correlation-test method above to the entire data set by correlating every "characteristic" (columns 2-26) with every food quantity/frequency (columns 28-283). Collect the output as two different 25x256 element matrices: one containing the correlation coefficients (r-values), and one containing the associated p-values. The columns of the rows should be the user characteristics (25 of them), and the rows should be the food intake variables (256 of them); the values of the matrix cells will then be the r-values or p-values. You will need to use a nested loop. (To avoid warnings, you may want to check for any columns for which the respondents all gave the same answer and skip the calculations for these.)

4. How many of your correlations above are "highly significant" ($p<0.005$)? Make a four-column table of all such two-variable correlations containing the characteristic variable, the food variable, the correlation coefficient, and the p-value.

5. Can these "correlations" be trusted, given the methodology that was used to calculate the p-values? If not, explain why not.

6. Can it be concluded that any of these correlations are *not* present (i.e. that in reality there is no association between the two variables)? If so, list them. If not, explain why not.

7. Use Bonferroni's correction to identify the effective (conservative) adjusted α threshold on which to test to achieve a "true" significance of $\alpha=0.05$, given the number of comparisons actually performed. Do any of the correlations identified above meet this criterion?

8. There were only 54 different individuals who submitted complete, useable responses to this survey. Would increasing the sample size by a factor of 100 (to 5400 individuals) help to reduce the number of spurious correlations identified by the "method" in question 3-4? If so, by how much?

9. Would increasing the sample size increase the power of a survey like this to identify a correlation that meets the stricter threshold in #6, should a genuine correlation actually exist?

10. Suppose that, after such an expanded study, a significant correlation between a medical condition and a specific food was identified. Does this necessarily imply that the food causes the medical condition, that the condition causes an individual to consume the food more / more often, or neither?

11. Write a misleading/wrong tabloid headline based around one of the "correlations" identified in question 4. Start your headline with the words: "LJMU Data Scientists..."

12. After completing #11, go take a cold shower. (Optional, but recommended)