Week 1: Data analysis application

Commentary on this assignment

This document is available for comments and questions. It is truly a working document. Everyone in the class can comment on it, point out errors, ask clarifying questions, and add helpful resources. I will try to update things, fix mistakes, and clarify throughout the week.

You are welcome to approach this assignment however works best for you. I will offer one possible path:

1) Read through this document.
2) Load the data and look at it.
3) Work through each of the questions, seeking resources as needed.

Readings and resources

- An explanation of z-scores
  - Khan academy explanation
  - https://en.wikipedia.org/wiki/Standard_score
  - https://www.simplypsychology.org/z-score.html
  - Adding variables to a table in Matlab
- Removing errors and outliers
  - Some info on outliers (don't worry about the part after z-scores unless you're really interested)
  - Removing rows from a table in Matlab
- False positives, negatives
  - Video on false positives and negatives
  - Wikipedia on false positive rate
- The Matlab On Ramp (how to code in Matlab)
- A Matlab notebook with sample code (I found some errors in this and will posted an UPDATED: 1/24 at 4:35 PM, it's worth getting a new version.)
- You can suggest other things here by highlighting this statement (or other topics above) and making a comment. Be a pal, share good stuff with your coursemates (and me!)

Skills/knowledge we're using

- Load and manipulate real data in using software tools
- Compute and understand summary statistics (mean, median, mode, standard deviation, variance, interquartile range) for data.

- Identify outliers or potential errors in data, understand their effects on summary statistics, and justify decisions.
- Create meaningful visualizations of distributions of data (e.g., histograms).
- Compare individual data points from two different distributions using z-scores.
- Compute true positive, false positive, true negative, and false negative counts and rates.
- Calculate probabilities for a known group.

The situation

You and your team have developed a new screening for Parkinson's disease. Hooray, you're amazing! The attached data (parkinsons.txt, also on Canvas) are from a study to evaluate your screening. Since you are trying to evaluate the effectiveness of your screening, you recruit a group of people with known Parkinson's disease diagnoses (let's assume that these "known" diagnoses are accurate and attainable). While the actual prevalence of Parkinson's disease is the world is much lower, in your selected group, 50% of participants have Parkinson's disease and 50% do not.

Variable	**Format**	**Header in Text File**
Age	Integer age value	age
Postural stability test type	PSI or ES	PStype
Postural stability score (PSS)	number	PSS
Actual true diagnosis	1=Parkinson's, 0 = no Parkinson's	actual
Screening diagnosis (predicted based on your test)	1=Parkinson's, 0 = no Parkinson's	screen

Postural instability is one symptom of Parkinson's disease. Inconveniently, there are two different measures of postural stability, and people in your study had one of these two tests, depending on which collaborator did the testing (luckily you know which is which). When tested in the general population, the ES has a mean score of 70 with a standard deviation of 8, and the PSI test has a mean score of 61 with a standard deviation of 11.

It may be helpful to use data = readtable('Parkinsons.txt'); to load the data.

1. Summarize the age data with basic summary statistics and a plot. What does the difference between the mean and median suggest about the distribution?
2. Since our data has two different measures of postural stability, we'll need a way to appropriately compare these values. In a new variable (let's call it Common Stability Score or CSS), represent the stability scores of each participant on the all on the same scale. This scale can be ES, PSI, or a shared scale such as z-score.
3. Plot and comment on the CSS distribution.
4. Identify any extreme outliers or errors in the data. Decide how to deal with these and justify your decision. Use this revised data for the rest of your analysis.
5. Determine the counts of the true positive, true negative, false positive, and false negatives for your screening test.
6. If you randomly selected someone from your study group, what is the probability that your tests will flag them as having Parkinson's?
7. If you randomly select someone from your study group who you know does not have Parkinson's, what is the probability that your test will incorrectly diagnose them with the disease?
8. If you randomly select someone from your study group, what the probability that your test will diagnose them correctly?
9. Pose at least 3 additional questions that you could try to answer with these data. For example, how do the age distributions compare across diagnoses? A wide variety of questions are welcome, this is to help you think about what could be answered and what could not, given a particular data set.
10. Attempt to answer 1 of your 3 additional questions.

References

Paper on postural stability tests:
https://www.rehab.research.va.gov/jour/04/41/5/pdf/Chaudhry.pdf
Prevalence of Parkinson's disease in the US population:
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6039505/