Week 2: 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) Do the Bayes rule reading assignment.
2) Read through this whole document.
3) Load the data and look at it.
4) Work through each of the questions, seeking resources as needed.

Readings and resources

- Reading about PMF and CDF that may be helpful for Part 2 of this assignment (you probably only need to read section 2 to 3.4). Or this is a very short description of PDF and CDF, which might be an easier entry.

Skills/Knowledge we're using

- Conditional probability and Bayes Rule
- Mean and standard deviation
- Binomial and normal distributions

Part 1: Conditional probabilities and Bayes' Rule

Where appropriate, please practice writing your answers in the format of conditional probabilities (you don't need to do this in your variable names in your code). For example, the probability of "this" given "that" can be written as P(this|that), where | can be typically found above the Enter key on your keyboard.

1. Write yourself some organized notes about true/false positives/negatives (from last assignment) and how they relate to conditional probabilities. Look up *sensitivity* and *specificity*, and express them as conditional probabilities.

For the next exercises, we will be revisiting your Parkinson's disease screening from last week.

2. If you randomly select someone from your study group who tested positive according to your screening test (screen ==1), what is the probability that they actually have Parkinson's disease?

3. Now, let's imagine this becomes a regular screening for adults 45 and over. This paper estimates "the overall prevalence of PD among those aged ≥45 years to be 572 per 100,000." What is the probability that a random person who gets a positive test result (screen=1) actually has Parkinson's disease? What are your thoughts on this number?

4. Now, let's assume that instead of being used as a regular screening, your test is only used in scenarios where people have expressed multiple symptoms. Let's assume the probability of a person who has expressed multiple symptoms having Parkinson's disease is 30%. Given that a random person has expressed multiple symptoms and gets a positive test result (screen=1), what is the probability that they actually have Parkinson's disease? What are your thoughts on this number?

5. Generate a plot to illustrate how the *probability a person having Parkinson's given that your screening result was positive* changes as function the prevalence of the disease in the population that you are testing.

6. At its current sensitivity and specificity, would you recommend that your screening test should be used for all people over 45 or only people who have displayed multiple symptoms of Parkinson's? Why? What factors did you consider? (Your answer does not need to be limited to the values computed here.)

7. Challenge: If you wanted each individual's probability of having Parkinson's if your screening test gave them a positive result to be 0.90, what would the sensitivity of your test need to be?

Part 2: Exploring distributions

In this part, we're going to look at the binomial distribution and start to explore the probability mass function (PMF), probability density function (PDF) and the cumulative distribution function (CDF). It may be helpful to do this reading about PMF and CDF (you probably only need to read section 2 to 3.4), or look up the meaning of the probability mass function (PMF) and the cumulative distribution function (CDF) for discrete variables (like the number of heads in 10 coin tosses).

These exercises suggest you use the disttool gui. If it's helpful for you to understand things by writing your own code, then check out the makedist function.

Binomial Distribution

Explore the PMF and CDF visually using the disttool function in Matlab (you can bring up the interface by typing disttool into your Matlab Command Window). Set the distribution type to binomial; leave the function type as PDF* (this is actually a PMF, see note at bottom*); set the Trials to 1; and leave the probability ("Prob") as 0.5. You may need to adjust the "Upper Bounds", which set the x-axis for the plot.
If you click on either of the blue +'s, you will see the density value displayed on the left. The density refers to the probability of getting exactly that many positive outcomes (e.g, heads) in a given number of trials (1 in our case).

8. As you change the value of "Prob", does the density change as you expect?

9. Set the "Prob" back to 0.5. Increase the number of trials a few times (go up to at least 50). Describe what happens to the plot and how this translates to the coin flip analogy.

10. Using the PMF (labeled PDF), what is the probability of getting exactly 19 heads in 37 flips of a fair coin? If you want, confirm this using math.

11. Now switch to the view of the CDF. You may want to return to 1 or 2 trials to make sure you have your head wrapped around what this represents. What is the probability of getting 19 or fewer heads in 37 flips of a fair coin? What if the coin is biased towards heads with a probability of 0.6?

Normal Distribution

Now, let's explore the "normal distribution" using the disttool interface. Notice that the plots become smooth instead of chunky (#peanutbutter). The variable mu (μ) refers to the mean, and sigma (σ) refers to the standard deviation.

12. Describe what happens in the PDF plot when you change the mean and what happens when you change the standard deviation.

13. Switch to the CDF plot. The cdf outputs the area under the PDF for values up to and including the input value. Describe what happens when you change the mean and what happens when you change the standard deviation of your distribution..

14. Consider a standard normal distribution (mu = 0, sigma = 1). What is the probability of randomly selecting a value of 1.96 or less?

15. What is the probability of having an x value of 4 or less when your normal distribution has a mean of 1 and a standard deviation of 3?

**Note: In the Matlab disttool() interface, the Function Type options are PDF or CDF. PDF stands for probability density function. The PMF is used for discrete variables and the PDF is used for continuous variables, so technical the label should say PMF for any of the discrete distributions, like the binomial distribution (but this would be annoying to program).*