INFO370 Problem Set 3: Data description, probability

Your name:

Deadline: Wed, Oct 30th 2pm

Instructions

This problem set is about three topics: data description from statistical point of view, probability, and conditional probability (Bayes theorem). In terms of readings, it covers the first three chapters of Diez *et al.* (2019).

The first question is something you have to do with computer but the following two include operating with formulas, so you may want to do those on paper. If this is the case, please scan your solutions and include into the final markdown document (a cellphone photo works, you should be familiar how to include images in markdown). Even better, learn how to do it in latex!

1. Be sure to include well-documented (i.e. commented) code chucks, figures, tables, and clearly written text explanations as necessary. All figures should be clearly labeled and appropriately referenced within the text. Be sure that each visualization (graph or table) adds value to your written explanation; avoid redundancy – you do no need four different visualizations of the same pattern.

2. Don't output irrelevant, or too much of relevant, information. A few figures is helpful. A few thousand figures is only noise.

3. All materials and resources that you use (with the exception of lecture slides) must be appropriately referenced within your assignment. In particular, note that Stack Overflow is licensed as Creative Commons (CC-BY-SA). This means you have to attribute any code you pick from SO (a link to the question/answer webpage will normally do).

4. Remember partial credit will be awarded for each question for which a serious attempt at finding an answer has been shown. Students are *strongly* encouraged to attempt each question and to document their reasoning process even if they cannot find the correct answer. If you would like to include code to show this process, but it does not run withouth errors you can just comment it out and explain what do you attempt to do.

   If you go the knitr/rmarkdown way, you can also set the chunk option `eval` to `FALSE`:

   ~~~
   ```{eval=FALSE}
   a + b # these object don't exist
   # if you run this on its own it with give an error
   ```
   ~~~

5. When you have completed the assignment and have *checked* that your code both runs and builds correctly, convert the notebook/markdown to HTML (or pdf), name it ‘ps3-YourLastName-YourFirstName.html‘, and submit the html file on Canvas. Please submit also your original files (original notebook, or code files, or rmarkdown or whatever else do you have).

   html/pdf is much easier and quicker to check, but in case of questions, we also want to see your original code.

6. Please keep data file in the same folder as your code, and read these w/o any path like `"data.csv"` (or `"./data.csv"`). This makes the checking your code much easier!

7. Working together is fun and useful but you have to submit your own work. Discussing the solutions and problems with your classmates is all right but do not copy-paste their solution! First understand it, and thereafter create your own solution. Please list all your collaborators on the solution.

1 Describe data, compute probabilities

1.1 Where and when are people walking?

Your first task is to explore public data from the viewpoint of answering the question: *where and when are people walking?* Note: we do not expect you to actually answer this question, just to do a data exploration while keeping this question in mind.

1.1.1 Data

The data for this task comes from City of Seattle, https://data.seattle.gov/. There are various datasets, in this assignment you will use *Public Life Data 2018*.

1. Download the relevant datasets and other information. Explain what did you download and why did you chose these files.
2. Describe the dataset(s) you plan to use for this task. What was the purpose of collecting tese data? What do the dataset(s) contain? How was the data collected? What is the sampling scheme (consult OI 1.3.5)? Do you understand the data usage conditions?
3. Describe the potential issues: do you see any problems when trying to answer the question? Do you see issues with privacy, ethics?
4. Which variables do you see as relevant for this study? These are the ones you will select and work on below.

   Note: please answer this question solely based on documentation, not by analyzing the content of the data. If the documentation is vague or unclear, err toward including more information.

1.1.2 Descriptive statistics

Next, let's look at the actual data. Read the data in computer, extract the variables you considered relevant, and describe all those in a meaningful way.

Note: it is all right to discover later that some of the variables you selected are not really useful, and the way around–you overlooked some of the necessary data. It happens all the time.

I recommend to answer the following questions in a table (or multiple table) form.

1. Load the dataset(s) and select the relevant variables. Inspect the results and ensure these look reasonable.

   Note: you may have to go back and forth a number of times as you may find some of the variables you initially selected as virtually useless for your task.

2. For all variables report the number of valid observations and missings.

3. Report basic statistics for the most important (in the sense of answering the research question) variables. You answer should contain some of the following: (a) number of unique values; (b) central tendency; (c) variation; (d) range. Pick suitable indicators depending on the variable type. Explain your choice!

4. Comment your results. Does your data correspond to the documentation? Do you find any issues from the perspective of the analysis you are asked to do?

1.1.3 Analysis

We do not ask you to do the actual analysis, just discuss and explore:

1. Explain, how you might conduct this analysis. Which variables you might use? Which methods? Which problems you may run into?

2. Create a brief exploratory analysis (say, 1 table and 1 figure) along the lines you envisioned above.

1.2 Overbooking flights

You are hired by *Air Nowhere* to recommend the optimal overbooking rate.

The airline uses a 100-seat plane and tickets cost $100 each. So a fully booked plane generates $10,000 revenue. The sales team has found that the probability that passengers who have paid their fare actually show up is 98%, and individual show-ups can be considered independent. The additional costs, associated with finding an alternative solutions for passengers who are refused boarding are $500 per person.

1. Which distribution would you use to describe the actual number of show-ups for the flight?

2. Assume the airline never overbooks. What is it's expected revenue?

3. Now assume the airline sells 101 tickets for 100 seats. What is the probability that all 101 passengers will show up?

4. What are the expected profits (= revenue − expected additional costs) in this case? Would you recommend overbooking over booking the just right amount?

5. Now assume the airline sells 102 tickets. What is the probability that all 102 passengers show up?

6. What is the probability that 101 passengers – still one too many – will show up?

7. Would it be advisable to sell 102 tickets?

8. What is the optimal number of seats to sell for the airline? How big are the expected profits?

Hint: some of the expressions may be hard to write analytically. Feel free to use computer for the calculations, just show the code and explain what are you doing.

1.3 Does the student know the answer?

In the exam, there is a multiple-choice question with four (mutually exclusive) options. In average, 80% of the students know the answer, but in 10% of time they still answer wrong because of exam stress.

1. If a student get's the answer right, what is the probability that she actually knows the material?

References

Diez, D. M., Barr, C. D. and Çetinkaya Rundel, M. (2019) *Openintro Statistics*, Openintro, 4th edn.