Data Science

CSE558

Assignment - 2

Submit by 2359 of 29/10/23 Maximum marks: 40

- **While answering the questions, you are encouraged to discuss with your classmates. Mention their names in your submission. If you have used the internet to understand the solution you are writing, then mention the URL.**
- **Write answers and code of your own; do not copy from others. We will follow the standard plagiarism policy, which can be found link**
- **Clearly state any extra information (such as assumption or encoding) used to reach your answers from the given questions. Write all the steps that you followed.**
- **Prepare one zip file with all your answers to the theory questions, files of code, saved data and plots. Name it "roll_no-DSA1", e.g., "20001-DSA2". Submit the zip file through Google Classroom. A delay in submission would cost you 5 marks per delayed day. Based on a quick viva, you will be graded.**

Best wishes!

---

1. Consider you have an unbiased k-faced die, which you independently roll twice.
   (a) For $k = 6$, what is the probability that the face value of the second roll is a multiple of the face value from the first roll? **(6)**

   (b) Let, k be an integer divisible [1] by every integer from 1 to $k/2$. For a large value of k, what is the probability that the face value of the second roll is a multiple of the face value from the first roll? **(4)**

2. In probability theory, the birthday problem asks for the probability that, in a set of n randomly chosen people, at least two will share a birthday. The birthday paradox refers to the counter intuitive fact that only 23 people are needed for that probability to exceed 50%.
   (a) For a year with 365 days prove the birthday paradox. **(3)**

   (b) What is smallest integer value n such that with probability more than 75%, at least two will share a birthday. **(4)**

   (c) Your friend made a software that computes the value n for a given probability. Lets say there is a bug in the software. It takes the real birthday from the user and then it randomly generates another date. For a given probability, p software returns the smallest integer n such that at least one of the

[1] A number x is divisible by y, if there exists a positive integer c such that $c\dot{y} = x$. Assume k exists even if you cannot find or compute it.

two dates of two people is the same. For example, let dates of n people are $R1, F1, R2, F2, \ldots, Rn, Fn$, where Ri is the real date and Fi is the fake date of the person i. For a given probability p, the software returns an integer m (where $m \leq n$), such that there exists a pair (i, j) where $1 \leq i < j \leq n$ and the probability of $Ri = Rj$ or $Ri = Fj$ or $Fi = Rj$ or $Fi = Fj$ is at least p. **(8)**

3. Use z-test for proportion for the following questions.

   (a) You think at most 50% of email you receive are spam. Over a week you received 55 spam emails. At 5% level of significance, what is the minimum number of emails you need to see so you do not change your thoughts (i.e., do not reject the null hypothesis)? **(7)**

   (b) Let your friend claim she can correctly guess the suit of a randomly picked card with more than 1/3 of the time on an average. You test the correctness of the claim; you randomly picked x cards and out of which, she correctly guessed 28 of them. What is the maximum value of x such that with a 5% level of significance, you end up believing her claims (i.e., reject the null hypothesis)? **(8)**