National Taiwan University
Institute of Industrial Engineering
Jakey BLUE

Data Analytics

111-2 Homework #01

Due at 23h59, March 5, 2023; PDF file uploaded to NTU-COOL

1. On a multiple-choice exam with four possible answers for each of the five questions, what is the probability that a student would get 16 or more (out of 20, i.e., get at least four questions correct) just by guessing randomly?

2. Suppose two teams, E and W, are playing the NBA finals (a series of 7 games), where the series is done when E or W wins four matches firstly. If each match is independently won by team E with probability p and by team W with probability $1-p$. Find the expected number of matches that are played, and evaluate this expected number when $p = 1/2$.

3. A fountain show starts every 80 minutes, you arrive the place at random and decide to wait for 20 minutes, what's the probability you will witness the show?

4. In the post office of A city, there are two clerks working with different efficiencies: clerk 1 has service time following an exponential distribution with mean $1/\mu_1$ while the service time of clerk 2 follows **a different** exponential distribution with mean $1/\mu_2$. One day, John enters the post office and he is served by clerk 1 at 8h00.
   a. Mary enters at 8h10, what is the probability she sees John is still being served by clerk 1?
   b. Since John is still in service, Mary goes to clerk 2 to be served. What is the probability that Mary finishes her service **before** John does?

   * Hint: Memoryless Property

5. John lives in A city and goes to work every morning by taking one train and then connecting to a local bus in B city. To avoid being late for work, he must arrive no later than 8h30. John always takes the train at 8h00. The trajectory between A and B takes **exactly 10 minutes**. According to the long-term observation, the train is of a delay probability distribution as the table below:

delay (min)	4	6	8	10	12
probability	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{16}$

   Another uncertainty to his office is that the bus is not stable either. Averagely, a bus arrives at the train station of B city to pick up passengers at 8h20 and then the trip is **exactly 10 minutes** to arrive John's office. According to another survey, the bus schedule is following the Normal distribution with the average departure time at 8h20 from the train station and a **standard deviation of 2 minutes**.
   Assume the train delay and the bus uncertainty are independent. What is the probability that John will be late for work?

   * Hint: Interaction of Two Random Variables

6. Assume in average 1 out of 100 people has cancer (1%) and the cancer detection rate with current X-ray scan is accurate up to 99%. One day, John is diagnosed as positive in the X-ray cancer scan.
   a. What is the probability that John really has the cancer?
   b. Since John knows Bayes' theorem very well, he decides to make an MRI check to know more details of his health. It is known that MRI check has 99.9% accuracy detecting cancer. Unfortunately, John is diagnosed again positive after the MRI check. What is the probability that John has the cancer now?

   * Hint: Bayes' Theorem

7. The amount of time that a customer spends waiting at an airport check-in counter is a random variable with mean 8.5 minutes and standard deviation 3.5 minutes. Suppose that a random sample of $n = 49$ customers is observed. Find the probability that the "average time waiting in line for these 49 customers" is
   a. less than 10 minutes;
   b. between 7 and 10 minutes;
   c. less than 7.5 minutes;
   * Hint: Central Limit Theorem

8. The proportion of people living in A city who are iPhone users is estimated to be $p = 0.4$. To test this hypothesis, a random survey of 600 people is conducted. After the statistical analysis, we decide that if the number of iPhone users is between 216 to 264, the hypothesis will be accepted; otherwise, we will conclude that $p \neq 0.4$. Please find the **type I error probability** for this analysis procedure, assuming that $p = 0.4$ for real.
   * Hint: z-test

9. Define X as the number of under-filled beer bottles from a filling operation in a carton of 24 bottles. 75 cartons are inspected and the following observation on X are recorded:

X	0	1	2	3
Frequency	39	23	12	1

   Based on these 75 observations, is a "Binomial distribution" an appropriate model? Perform a Goodness-of-Fit procedure with $\alpha = 0.05$.

10. Two courses: Probability & Statistics (Prob) and Operations Research (OR), are given in the same semester to the same group of students. Suppose we have 100 students with the following grades summary.

Prob \ OR	A	B	C
A	24	11	10
B	7	13	5
C	4	6	20

    Please setup a hypothesis testing with $\alpha = 0.01$ to conclude if the grades in Prob and OR are related.
    (Grades are classified into Three levels, where A: excellent; B: average; C: failed.)
    * Hint: χ^2 test

11. (15%) Simulate the averages of $[2, 3, 4, 5]$ dice for 1000 times. Draw the four histograms for the sample averages of $[2, 3, 4, 5]$ dice, respectively.
    *Reproduce the CLT results on p. 28 of DA01 slides.

12. To validate the Kruskal's count on p. 13 of DA01 slides, we play the game with one deck of cards, i.e., 52 cards, for 10000 times. Each time, the 52 cards are randomly shuffled. We then start from the first 10 cards, and the face cards (J/Q/K) are counted as 5 steps.
    a. (15%) What is the probability that all the first 10 cards reach the same end?
    b. (15%) Vary the simulation settings:
       the # of cards: $[52, 104]$;
       the # of steps for face cards $= [1, 3, 5, 7, 9]$.
       What are the $2 \times 5 = 10$ probabilities? Discuss your observation?