DDA2001: Assignment 1

1. The assignment is due at **11:59 pm, Feb.20, 2022**.
2. Please submit your solution in PDF form. **Any other forms of solution will not be accepted and will be graded as 0.** Please leave enough time to make sure you have uploaded your solution as requirment before due.
3. If you submit the assignment late, you will get 0 for this assignment. **No excuses will be accepted for any late submission.**
4. Please make sure that your file could be downloaded successfully from BB after uploading your solution file.

---

1 Concept Questions [40 points]

1. **(4 points)** Data science is the process of a diverse set of data through?
   A. Organizing data
   B. Processing data
   C. Analysing data
   D. All of the above

   Solution: D
   Explanation: Data science is the process of deriving knowledge and insights from a huge and diverse set of data through organizing, processing and analysing the data. □

2. **(4 points)** What can we do with Data Science?
   A. To find unseen patterns.
   B. To derive meaningful information.
   C. To make business decisions.
   D. All of above.

   Solution: D □

3. **(4 points)** When rolling a fair dice, which of the following tools you will use to predict the outcome?
   A. Probability theory
   B. Statistics
   C. Simulation
   D. Optimization

   Solution: A
   Given the information of model(fair dice), predict numerical descriptions of how likely of each possible outcome to occur. This is Probability. □

4. **(4 points)** Sometimes we do not have access to the entire data set (population) and we have to infer our conclusions using sample data. Which of the following approaches addresses working with sample data to conclude about the population?
   A. Probability theory
   B. Statistics
   C. Simulation
   D. Optimization

   Solution: B
   Given the data(sample data), predict model to describe the population. This is Statistics. □

5. **(4 points)** This year, our company plans to manufacture two types of facial masks: A and B. To decide the amount of A and B, our company selected 1000 customers randomly and made a market survey. Given the collected data, which theory we may use to predict which type of facial mask is more popular?
   A. Probability
   B. Statistics
   C. Programming
   D. Optimization

   Solution: B
   Given the data(1000 market survey), predict model(customer's preference). This is Statistics. □

6. **(4 points)** The amounts of two types of facial masks and the profit of them are denoted as x_A, x_B and $R(x_A, x_B)$, respectively. To maximize $R(x_A, x_B)$, which theory we may use?
   A. Probability
   B. Statistics
   C. Sampling
   D. Optimization

   Solution: D
   Make decision(value for x_A, x_B) to maximize objective($R(x_A, x_B)$). This is Optimization. □

7. **(4 points)** In nucleic acid testing, if you want to arrange the transportation route of the test tube to the destination as soon as possible, which theory do you use?
   A. Probability theory.
   B. Statistics.
   C. Optimization.
   D. Simulation.

   Solution: C
   Make decision(transportation route) to maximize objective(time for transportation). This is Optimization. □

8. **(4 points)** Which of the following statement is false?
   A. Given the past data, we can use statistics to predict the data pattern.

B. Given the data pattern, we can use probability theory to predict new data.
C. We can use programming to collect the past data.
D. If the data pattern is complicated, we can always use a simplified model to predict new data.

Solution: D □

9. **(4 points)** What are statistics and probability theory? What is the relationship or difference between them?

Solution:
1. Statistics theory is a discipline that extracts correct information from data.
2. Probability theory is a formality to make sense of the world in terms of uncertainty.
Statistics is to predict model given data, while probability is to predict data given model.
Other reasonable answer is OK. □

10. **(4 points)** Describe the main steps when making a decision as a data scientist. Show what knowledge may be used in each step and point out the main purpose of the used knowledge in each step. Please give one practical example to show how to apply what you answered.

Solution:
Steps:

i. Collect the past data: Programming(automate decision)
ii. propose some models: Probability(quantify uncertainty)
iii. Choose the best model: Statistics(test credibility)
iv. Prediction for given input: Simulation/Sampling(calculate complicated objective)
v. Optimize input: Optimization(optimize objectives)

Example: ... □

2 Computation Questions [60 points]

1. **(3×2 points)** A cafeteria offers a three-course meal consisting of an entree, a starch, and a dessert. The possible choices are given in the following table:

Course	Choices
Entree	Chicken or Roast beef
Starch	Pasta or Rice or Potatoes
Dessert	Ice cream or Jello or Apple pie or Peach

Table 1: Choices for the three-course meal

A person is to choose one course from each category.

(a) Let A be the event that ice cream is chosen. How many outcomes are in A?
(b) Let B be the event that chicken is chosen. How many outcomes are in B?
(c) List all outcomes in the event A and B (i.e., $A \cap B$)?

Solution:

(a) 6.

(b) 12.

(c) $AB = \{(\text{chicken}, \text{pasta}, \text{ice cream}), (\text{chicken}, \text{rice}, \text{ice cream}), (\text{chicken}, \text{potatoes}, \text{ice cream})\}$

□

2. **(4×2 points)** For two events A and B. State whether each of the following statements/situations is (i) necessarily true, (ii) necessarily false, or (iii) possibly true.

   (a) If A and B are disjoint, then A and B are independent.

   (b) If A and B are independent, then A and B are disjoint.

   (c) $\mathrm{P}(A) = 0.6$, $\mathrm{P}(B) = 0.6$, and A and B are independent.

   (d) $\mathrm{P}(A) = 0.6$, $\mathrm{P}(B) = 0.6$, and A and B are disjoint.

   Solution:

   (a) If A and B are disjoint, then when $P(A) = P(B) \neq 0$, we have $\mathrm{P}(AB) = 0 \neq \mathrm{P}(A)\mathrm{P}(B) > 0$, thus the statement is False, so choose (ii).
   You can also consider it's possible that A and B are independent when A and B are disjoint, which is when $P(A) = P(B) = 0$, so (iii) can also be judged right here.

   (b) If A and B are independent, then when $P(A) = P(B) \neq 0$, $\mathrm{P}(AB) = \mathrm{P}(A)\mathrm{P}(B) > 0$, thus the statement is False, so choose (ii).
   You can also consider it's possible that A and B are disjoint when A and B are independent, which is when $P(A) = P(B) = 0$, so (iii) can also be judged right here.

   (c) (iii)Possibly true.

   (d) If A and B are disjoint, i.e. $\mathrm{P}(AB) = 0$, then $\mathrm{P}(A \text{ or } B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(AB) = 1.2 > 1$. Thus the statement is (ii)necessarily false.

   □

3. **(6 points)** One-year-old catch is a Chinese tradition. On a baby's first birthday, parents will lay out a lot of items on the floor for the child to pick from and whatever he/she picks is an indicator of the child's future career. Now, the one-year-old catch party of the twin brothers, Tom and Bill, is held. Parents prepared two toy pianos, three toy cars, one dictionary, two pens and one wallet. Little Tom will catch one item firstly and then put it back to let little Bill catch. What is the probability that Tom and Bill both catch a toy piano or a wallet?

   Solution: Let T denote the event that a toy piano is caught, and W denote the event that a wallet is caught. Then

$$\mathrm{P}(T \text{ or } W) = \mathrm{P}(T) + \mathrm{P}(W) = \left(\frac{2}{9}\right)^2 + \left(\frac{1}{9}\right)^2 = \frac{5}{81}.$$

□

4. **(3×2 points)** A company has designed a new product. This company estimates that when it is introduced into the market it will be very successful with a probability 0.6, moderately successful with a probability 0.3, and not successful with a probability 0.1. The estimated yearly profit associated with the model being very successful is \$15 million and being moderately successful is \$5million; not successful would result in a loss of \$0.5 million. Let X be the yearly profit of the new model.

   (a) Determine the probability mass function (pmf) of X.

   (b) Determine the cumulative distribution function (cdf) of X.

   (c) Compute the mean $E[X]$ and the variance $Var[X]$.

   Solution:

   (a)
   $$P(x) = \begin{cases} 0.6 & x = 15 \text{ million} \\ 0.3 & x = 5 \text{ million} \\ 0.1 & x = -0.5 \text{ million} \end{cases}$$

   (b)
   $$F(x) = \begin{cases} 0 & x < -0.5 \text{ million} \\ 0.1 & -0.5 \text{ million} \le x < 5 \text{ million} \\ 0.4 & 5 \text{ million} \le x < 15 \text{ million} \\ 1 & x \ge 15 \text{ million} \end{cases}$$

   (c)
   $$E[X] = 10.45 \text{ million}$$
   $$E[X^2] = 142.525 \text{ million}^2$$
   $$Var[X] = E[X^2] - E[X]^2 = 33.3225 \text{ million}^2$$

   □

5. **(8 points)** Let X have the discrete uniform distribution on the integers $1, \ldots, n$. Compute the mean and variance of X. Hint: You may wish to use the formula $\sum_{k=1}^{n} k^2 = n(n+1) \cdot (2n+1)/6$.

   Solution: We have $E[X] = \dfrac{n+1}{2}$ and $E[X^2] = \sum_{k=1}^{n} \dfrac{k^2}{n} = \dfrac{(n+1)(2n+1)}{6}$. So,
   $$\text{Var}(X) = \frac{(n+1)(2n+1)}{6} - \frac{(n+1)^2}{4} = \frac{n^2-1}{12}.$$

   □

6. **(4×3 points)** A city publishes two newspapers A and B. Among the residents of the city, 55% subscribe to A, 65% subscribe to B, and 25% subscribe to both A and B. Please find the probability of the following events for one resident:

   (a) Only subscribe to A

   (b) Only subscribe to one type of newspaper

   (c) At least subscribe one type of newspaper

(d) Subscribe no newspaper at all

Solution:

Here we denote A,B as subscribing to Newspaper A, Newspaper B, then $P(A) = 0.55, P(B) = 0.65$, $P(AB) = 0.25$.

(a)
$$P(\text{ Only subscribe to } A) = P(AB^c) = P(A) - P(AB)$$
$$= 0.55 - 0.25 = 0.30$$

(b) As P(Only subscribe to one type of newspaper) $= P(AB^c) + P(A^cB)$, where

$$P(A^cB) = P(B) - P(AB) = 0.65 - 0.25 = 0.40$$

therefore, P(Only subscribe to one type of newspaper) $= 0.30 + 0.40 = 0.70$

(c)
$$P(\text{Subscribe to at least one type of newspaper})$$
$$= P(A \cup B) = P(A) + P(B) - P(AB) = 0.55 + 0.65 - 0.25 = 0.95$$

(d) P(Subscribe to no newspaper at all) $= P(A^cB^c) = 1 - P(A \cup B) = 0.05$

□

7. **(8 points)** A person wants to invest 10,000 RMB in a stock, and the current price of the stock is 2 RMB per share. Assume that the stock may be 1 RMB/share and 4 RMB/share after one year, with equal probability. The advice given to him by a financial advisor is: if you expect to have the largest stock market value in a year, buy it now; if you expect to have the largest number of stocks in a year, buy it a year later. Is the advice of a financial advisor correct? why?

Solution:

If he buys the stock now with cost 2 yuan/share, then he can buy 5000 share. Denote the market value of stocks owned in a year as X. The distribution column of X is:

X	5000×1	5000×4
P	$\frac{1}{2}$	$\frac{1}{2}$

Then we can get $E(X) = 2,500 + 10,000 = 12,500$, which means the expected stock market value in a year is 12,500 RMB if he buys it now. And since the number of stocks won't change, so there are still 5000 shares in a year.

If he buys the stock in a year, denote the number of stocks as Y. The distribution column of Y is:

Y	$10,000/1$	$10,000/4$
P	$\frac{1}{2}$	$\frac{1}{2}$

Then we can get $E(Y) = 5,000 + 1,250 = 6,250$, which means the expected number of stocks in a year is 6,250 if he buys it a year later, which is larger than 5,000 if he buys it now. And the stock market value in a year will be exactly 10,000 if he buys it a year later, which is smaller than 12,500 if he buys it now.

Therefore, the advise is right.

8. **(6 points)** Tom flips a fair coin 5 times, and Jerry flips the same coin 4 times. Please find the probability that the number of heads Tom flips is larger than the number of heads Jerry flips. (Hint: for a fair coin, the probability that "the number of heads Tom flips is larger than the number of heads Jerry flips" equals the probability that "the number of tails Tom flips is larger than the number of tails Jerry flips".)

Solution: We denote

$X_1 =$ the number of heads Tom flips, and $X_0 =$ the number of tails Tom flips $= n+1-X_1$
$Y_1 =$ the number of heads Jerry flips, and $Y_0 =$ the number of tails Jerry flips $= n - Y_1$
And we note events:

$$E = \{X_1 > Y_1\}, \quad F = \{X_0 > Y_0\}$$

Since this is a fair coin, we have that $P(E) = P(F)$. Besides,

$$\begin{aligned} F &= \{X_0 > Y_0\} = \{n + 1 - X_1 > n - Y_1\} \\ &= \{X_1 - 1 < Y_1\} = \{X_1 \leqslant Y_1\} = \bar{E}, \end{aligned}$$

Therefore, from $P(E) = P(F) = P(\bar{E})$, we know that $p(E) = 0.5$.

□