COMP1003/1433 Exercise Questions (Week 1 – Week 6)

1. (True or False) Data analytics is also known as data mining.

   False. Data analytics include both data mining and communication and concerns more with the entire methodology while data mining may focus on an individual analysis step. P57, Lecture 1.

2. (True or False) The sample mean approximates the population mean μ for any sample size n.

   False. The sample mean approximates the population mean for a very large sample size n. P17, Lecture 3.

3. (True or False) The angle of two vectors can be used to measure their distance.

   True. The cosine of the angle is the cosine similarity measure. P21, Lecture 4.

4. (True or False) For any function f(x), we will find its maximal or minimal solution via solving the equation of f'(x)=0, where f'(x) means the derivative of f(x).

   False. A variable resulting in the derivative of 0 might not an optimal solution. P23-24, Lecture 5.

5. (True or False) The differentiation process for a function f(x) allows the measurement of the instantaneous rate of change at (x, f(x)).

   True. P10, Lecture 5.

6. (True or False) In logistic regression, the sigmoid function only works for binary classification.

   True. As can be seen from the function graph, the outlier values squash towards two sides 0 or 1. P33, Lecture 5.

7. (True or False) If two discrete random variables X and Y are independent, then we can have E(XY)=E(X)E(Y).

   True. If X and Y are independent, we can have P(X=x, Y=y)=P(X=x)P(Y=y) for any given x and y. Then, with the definition of expected values, we can have E(XY)=E(X)E(Y).

8. (1 correct choice only) Suppose we are interested in predicting whether a news report concerns a ``vaccine'' topic or not (e.g., to work on COVID-19 related applications). In our prior knowledge, 30% of news reports are about ``vaccine'' while 70% are not. Besides, we know that the probability of observing the word `` Pfizer'' in a ``vaccine'' news report is 60% and that in a ``non-vaccine'' news report is 20%. Now, given a news report containing `` Pfizer '', the probability that the news report is about ``vaccine'' is ______.
   A. Larger than 50%
   B. Smaller than 50%
   C. Equal to 50%

   (A) V: the news report is about ``vaccine''; NV: the news report is not about ``vaccine''; P: the news report contains the word ``Pfizer''. Then P(V|P)=P(P|V)*P(V)/(P(P|V)*P(V)+P(P|NV)*P(NV))=0.6*0.3/(0.6*0.3+0.2*0.7)=0.5625>50%.

9. (1 correct choice only) There are five boxes, where one carries a paper cheque with $1M while the other four each carry plain paper. You will never know which box carries the cheque till one draws the paper out of the box and release the results. Your friend first selects a random box and announced that the paper drawn is plain paper. Now it is your turn to do the lucky draw.

The probability for you to draw the cheque becomes ____ compared to the moment before your friend's drawing result is announced.

A. Larger
B. Smaller
C. Unchanged

(A) The sample space becomes smaller because your friend helps you exclude a box with plain paper.

10. (1 correct choice only) Given two vectors a=(2,2, 2, 2) and b=(0, 4, 0, 3), the cosine similarity of a and b is: (C)

A. 0.3
B. 0.4
C. 0.7
D. 0.8

(C) 0+8+0+6/sqrt(16*25)=14/20=0.7

11. (1 correct choice only) For x in the range of [0, 1], the area above x-axis and under the curve $\mathrm{f(x)} = \mathrm{x}^3 + e^{2x}$ is in the range of _____.

A. [0,1]
B. [1,2]
C. [2,3]
D. [3,4]

(D) $\int_0^1 (x^3 + e^{2x})dx = \left(\frac{x^4}{4} + \frac{e^{2x}}{2}\right)|_0^1 = \frac{1}{4} + \frac{e^2}{2} - \frac{1}{2} = 3.445$ in the range of [3,4].

12. (1 correct choice only) There are 2 types of Happy Meal toys in the McDonald's. Each of the toy type will be given with equal chances to a customer who buys the Happy Meal. Suppose that there is only one toy type that has the castle and Little Mary wants to get the castle very much. Let X denotes the random variable indicating the number of Happy Meals Little Mary should buy till she gets the castle. Then, the expected value of X should be ______.

A. 1
B. 3/2
C. 2
D. 5/2

(C) From the question, we can have $\mathrm{E(X)} = \lim_{n\to+\infty} \sum_{i=1}^{n} \left(\frac{1}{2}\right)^n \cdot n$. Let $\mathrm{S}_n = \sum_{i=1}^{n} \left(\frac{1}{2}\right)^n \cdot n$ and hence $2\mathrm{S}_n = \sum_{i=1}^{n} \left(\frac{1}{2}\right)^{n-1} \cdot n$. So, we'll have $2\mathrm{S}_n - S_n = \mathrm{S_n} = \sum_{i=1}^{n} \left(\frac{1}{2}\right)^{n-1} = \frac{1-\left(\frac{1}{2}\right)^n}{1-\frac{1}{2}} = 2\left(1 - \left(\frac{1}{2}\right)^n\right)$. Therefore, $\mathrm{E(X)} = \lim_{n\to+\infty} S_n = 2$.

13. (1 correct choice only) Given the following short movie reviews, each labeled with a genre, either comedy or action (the genre name is in **[boldface]** and the word in the reviews are in *italic*):

- *fun, couple, love, love* **[comedy]**
- *fast, furious, shoot* **[action]**
- *couple, fly, fast, fun, fun* **[comedy]**
- *furious, shoot, shoot, fun* **[action]**
- *fly, fast, shoot, love* **[action]**

Given a new document D: *fast, couple, shoot, fly*, we should assign D to the class of ____ measured by a Naive Bayes classifier with add-1 smoothing. The likelihood of observing the words in D conditioned on that class is ______.

A. comedy, $1.714 \text{ x } 10^{-4}$
B. action, $2.858 \cdot 10^{-4}$
C. comedy, $2.858 \cdot 10^{-4}$
D. action, $1.714 \cdot\ 10^{-4}$

(B) The vocabulary $\mathrm{V} = \{\mathrm{fun, couple, love, fast, furious, shoot, fly}\}$. So its size $|\mathrm{V}| = 7$

Let C denotes comedy genre and A denotes action. So, the prior of the two class labels are:

$\mathrm{P(C)} = \frac{\mathrm{count(C)}}{\mathrm{count(C)+count(A)}} = \frac{2}{5}$ $\mathrm{P(A)} = \frac{\mathrm{count(A)}}{\mathrm{count(C)+count(A)}} = \frac{3}{5}$

For likelihoods of observing different words are calculated as following:

$$\mathrm{P(fast|C)} = \frac{\mathrm{count(fast, C)} + 1}{\mathrm{count(C)} + |\mathrm{V}|} = \frac{1+1}{9+7} = \frac{1}{8}$$

$$\mathrm{P(fast|A)} = \frac{\mathrm{count(fast, A)} + 1}{\mathrm{count(A)} + |\mathrm{V}|} = \frac{2+1}{11+7} = \frac{1}{6}$$

$$\mathrm{P(couple|C)} = \frac{\mathrm{count(couple, C)} + 1}{\mathrm{count(C)} + |\mathrm{V}|} = \frac{2+1}{9+7} = \frac{3}{16}$$

$$\mathrm{P(couple|A)} = \frac{\mathrm{count(couple, A)} + 1}{\mathrm{count(A)} + |\mathrm{V}|} = \frac{0+1}{11+7} = \frac{1}{18}$$

$$\mathrm{P(shoot|C)} = \frac{\mathrm{count(shoot, C)} + 1}{\mathrm{count(C)} + |\mathrm{V}|} = \frac{0+1}{9+7} = \frac{1}{16}$$

$$\mathrm{P(shoot|A)} = \frac{\mathrm{count(shoot, A)} + 1}{\mathrm{count(A)} + |\mathrm{V}|} = \frac{4+1}{11+7} = \frac{5}{18}$$

$$\mathrm{P(fly|C)} = \frac{\mathrm{count(fly, C)} + 1}{\mathrm{count(C)} + |\mathrm{V}|} = \frac{1+1}{9+7} = \frac{1}{8}$$

$$\mathrm{P(fly|A)} = \frac{\mathrm{count(fly, A)} + 1}{\mathrm{count(A)} + |\mathrm{V}|} = \frac{1+1}{11+7} = \frac{1}{9}$$

Finally, we have:

$\mathrm{P(D|C)} \cdot \mathrm{P(C)} = \mathrm{P(fast|C)} \cdot \mathrm{P(couple|C)} \cdot \mathrm{P(shoot|C)} \cdot \mathrm{P(fly|C)P(C)} = \frac{1}{8} \cdot \frac{3}{16} \cdot \frac{1}{16} \cdot \frac{1}{8} \cdot \frac{2}{5} = 7.324 \cdot 10^{-5}$

And $\mathrm{P(D|A)} \cdot \mathrm{P(A)} = \mathrm{P(fast|A)} \cdot \mathrm{P(couple|A)} \cdot \mathrm{P(shoot|A)} \cdot \mathrm{P(fly|A)P(A)} = \frac{1}{6} \cdot \frac{1}{18} \cdot \frac{5}{18} \cdot \frac{1}{9} \cdot \frac{3}{5} = 1.714 \cdot 10^{-4}$

Therefore, we will classify the new document D into action genre ($1.714 \cdot 10^{-4} > 7.324 \cdot 10^{-5}$). The likelihood to observe words in D conditioned on action is $\frac{1}{6} \cdot \frac{1}{18} \cdot \frac{5}{18} \cdot \frac{1}{9} = 2.858 \cdot 10^{-4}$.

14. (1 correct choice only) In a new research paper published by University B, it takes 5 days on average for a COVID-19 patient to have > 30 CT value (tested negative). It is known that the time

for a COVID-19 patient to have > 30 CT value satisfies general normal with the standard deviation as 2.5 days. University P would be interested in knowing whether they can trust University B's results (the null hypothesis). So, they examined the sample of 64 COVID-19 patients and the time for their CT value to go back to a > 30 status is 5.5 days on average. Given the observations, if University P accepts University B's statement on the level of significance as x, then ______.

A. x<5.48%

B. x>5.48%

C. x<10.96%

D. x>10.96%

(C) Let $\bar{X}$ denotes the average days for the sampled 64 COVID-19 patients to obtain >30 CT value. The time for all COVID-19 patients to obtain >30 CT value satisfies general normal with the expected value of μ and standard deviation $\sigma = 2.5$ days. Let $\mathrm{Z} = \frac{\bar{X}-\mu}{\sigma/\sqrt{64}}$The p-value is

$$\mathrm{P}(|\bar{X} - \mu| \geq 0.5) = \mathrm{P}\left(|\mathrm{Z}| \geq \frac{0.5}{\frac{2.5}{\sqrt{64}}}\right) = 2\phi(-1.6) = 2 \cdot 5.48\% = 10.96\%.$$

15. (1 correct choice only) Given 3 clusters, the representative (centroid) of Cluster 1, 2, 3, and 4 are (1,3,3), (7,1,4), (0,0,0), and (5,8,1), respectively. For a new data point p=(3,5,2), according to cluster assignment strategy of k-means algorithm (based on Euclidian distance), which cluster should p belong to:

A. Cluster 1

B. Cluster 2

C. Cluster 3

D. Cluster 4

(A) Let c1, c2, c3, and c4 represent the centroids of Cluster 1, 2, 3, and 4. Then, we have the following: $\left||\mathrm{p} - \mathrm{c1}|\right| = \sqrt{2^2 + 2^2 + 1^2} = \sqrt{9}$, $\left||\mathrm{p} - \mathrm{c2}|\right| = \sqrt{4^2 + 4^2 + 2^2} = \sqrt{36}$, $\left||\mathrm{p} - \mathrm{c3}|\right| = \sqrt{3^2 + 5^2 + 2^2} = \sqrt{38}$, and $\left||\mathrm{p} - \mathrm{c4}|\right| = \sqrt{2^2 + 3^2 + 1^2} = \sqrt{14}$. So, p is closest to Cluster 1, we should assign it to this cluster.

16. (1 or multiple correct choice(s)) Suppose we know the probability of event A conditioned on C is p(A|C), the probability of event B conditioned on C is p(B|C), and the probability of C is p(C). Which of the following probabilities can be calculated for sure (there's no independence assumption among A, B, and C): (C)

A. p(A)

B. p(B)

C. p(AC)

D. p(ABC)

(C) P(AC)=P(A|C)P(C). Others cannot be calculated because there's no independence assumption.

17. (1 or multiple correct choice(s)) Given three vectors a, b and c and two scalars β and γ, find the correct statement(s) in the following: (A, B, C, D)

A. $-\beta\mathrm{a} - \gamma\mathrm{b} \ = -\gamma\mathrm{b} - \beta\mathrm{a}$

B. $\gamma\mathrm{a} \ + \beta(\mathrm{b} \ + \ \mathrm{c}) \ = \ (\gamma\mathrm{a} \ + \beta\mathrm{b}) \ + \beta\mathrm{c}$

C. $(\beta + \gamma)(\mathrm{a} \ + \ \mathrm{b}) = (\beta + \gamma)\mathrm{a} \ + (\beta + \gamma)\ \mathrm{b}$

D. $\beta\ \mathrm{a} \ + \gamma\ \mathrm{a} = (\beta \ + \gamma)\ \mathrm{a}$

(ABCD) Page 9 and 11, Lecture 4.

18. (1 or multiple correct choice(s)) Find the correct statement(s) in the following: (B, D)

A. [f(g(x))]' = f'(x)g'(x)

B. [f(g(x))]' = f'(g(x))g'(x)

C. [f(x)g(x)]’ = f’(x)g’(x)

D. [f(x)g(x)]’ = f’(x)g(x) + f(x)g’(x)

(BD) B is the chain rule while D is the product rule.

19. (1 or multiple correct choice(s)) For naive Bayesian classifier, which of the following statements are correct?

A. It is not sensitive to missing data, and the algorithm is relatively simple, which is often used in text classification

B. Naive Bayes is a discriminant model, which calculates the conditional probability by learning the known samples.

C. It has a solid mathematical foundation and stable classification efficiency.

D. It is relevant to the choice of a priori probability, so there is a certain error rate in classification.

(ACD) B is incorrect because Naïve Bayes is a generative model. Other statements are true derived from our discussions in Lecture 2.

20. (1 or multiple correct choice(s)) Which of the following is(are) the assumptions of a Naïve Bayes classifier?

A. Position of the words doesn’t matter.

B. The probability to observe words are independent conditioned on the class.

C. The probability of word occurrences in the documents are independent with each other.

D. A document can be represented by the count of words

(ABD) P34, Lecture 2.

21. (1 or multiple correct choice(s)) Which of the following statement about the definite integral $\int_{-\infty}^{\infty} e^{-\frac{1}{2}(2x-3)^2} dx$ is correct?

A. The result is in the range of [1,2].

B. The exact result is 1.5.

C. Chain rule can help solve the problem.

D. The properties of normal distribution may be helpful.

(ACD) Let $\mathrm{f(x)} = e^{-\frac{1}{2}(2x-3)^2}, \mathrm{u = 2x - 3}$, so $\mathrm{du = 2dx}$. We can then have $\int_{-\infty}^{\infty} f(x)dx = \int_{-\infty}^{+\infty} \frac{1}{2} \mathrm{e}^{-\frac{u^2}{2}} du = \frac{1}{2}\sqrt{2\pi} = 1.2533$

22. (1 or multiple correct choice(s)) Which of the following operations are FOR SURE doable in the linear algebra:

A. The Euclidean distance of two equal vectors.

B. The multiplication of two equal matrices.

C. The angle of two equal vectors.

D. The addition of two equal matrices.

(AD) Equal vectors have the same dimension, so A is true. Similarly, equal matrices have the same size, so D is true. B may not be doable if the row number does not equal to the column number. C may not be doable if the vector is a zero vector (which may correspond to the length of 0 in the denominator of cosine similarity).

23. (1 or multiple correct choice(s)) Given the following data observations: 6, 3, 2, 4, 9, 1, 7, 6, which of the following is correct?

A. The sample mean of these numbers is 4.75.

B. The sample median of these numbers is 5.

C. The sample range of these numbers is 8.

D. The sample standard deviation of these numbers is in the range of [7,8].

(ABC) Following the formula in page 11, Lecture 3, we can verify that ABC all correct. The sample variance is 7.357 while the sample standard deviation is 2.712 (not in the range of [7,8]).

24. (True or False) The research of big data focuses on the challenging problems of data analytics in large volume.

False. In addition to data volume, it also concerns data in velocity, variety, and veracity. P50, Lecture 1.

25. (True or False) Naive Bayes classifier is one of the most effective methods in text classification, which usually exhibits high accuracy.

False. Naive Bayes ignores the effects of word orders and assumes that features (e.g., words) are independent of each other. These assumptions are often not tenable in practical applications. P34, Lecture 2.

26. (True or False) In a hypothesis test, we reject a null hypothesis (H0) at the 5% level of significance, then we will for sure reject H0 at the 10% level of significance.

True. We reject H0 at 5%, meaning that the p-value of H0 should be smaller than 5%. Then the p-value of H0 is smaller than 10% and we will reject it at the 10% level of significance. P33-34 Lecture 3.

27. (True or False) In linear algebra, a vector is a list of numbers without orders.

False. A vector is an ordered list of numbers. P6 Lecture 4.

28. (True or False) The gradient descent algorithm always converges to the global minimum of the loss function.

    False. It may converge to a local minimum (the valley) if the function has multiple valleys (non-convex). P23-24, Lecture 5.

29. (True or False) In most supervised machine learning, the training process is to maximize the decision function, which predicts the labels (y) for any data input (x).

False. The training process is to minimize the loss function, which measures the distance between the predicted labels (y) and their ground truth annotations (y*). P35, Lecture 5.

30. (True or False) In the application of a Naïve Bayes classifier, when it meets the words absent in the training data or a priorly given vocabulary, it is safe to let the classifier simply ignore these words.

True. We'll ignore them because they are unknown words not used for training and knowing which class exhibits more unknown words is generally not a useful thing to know. P43, Lecture 3.

31. (True or False) The clustering results of K-means are very sensitive to how we initialize the clusters.

True. There is no guarantee to minimize the clustering objective of K-means. It simply goes down in each step and the initialization (how we start) is crucial to the clusters we may obtain at the end (how we end). P28, Lecture 4.

32. (1 correct choice only) Bag A contains 4 white and 6 red balls, and bag B contains 6 white and 8 red balls. We randomly select a bag with equal chances and draw a ball from it, which is found to be red. What is the probability that it was drawn from the bag A.
    A. 5/12
    B. 3/7
    C. 20/41
    D. 21/41

    (D) Let E=the drawn ball is red， F=the drawn ball is from bag A.

$$P(F|E)=\frac{P(E|F)P(F)}{P(E)}=\frac{P(E|F)P(F)}{P(E|F)P(F)+P(E|\bar{F})P(\bar{F})}=\frac{\frac{6}{10}}{\frac{6}{10}+\frac{8}{14}}=\frac{21}{41}$$

33. (1 correct choice only) Suppose we know the probability of event A conditioned on event B is 0.5 (p(A|B)=0.5), and the probability that event B happens is 0.8. The probability of A and B happening together is:
    A. 0.3
    B. 0.4
    C. 0.5
    D. 0.8

    (B) P(A,B)=P(A|B)*P(B)=0.8*0.5=0.4

34. (1 correct choice only) Given two vectors a=(1.2 3.3 5.1 2.2) and b=(0.2 1.3 1.1 0.2), the Euclidean distance of a and b is:

    A. 3

    B. 4

    C. 5

    D. 6

    (C) sqrt(1^2+2^2+4^2+2^2)=sqrt(1+4+16+4)=5

35. (1 correct choice only) Dr. Ling submitted two papers A and B to a conference with an acceptance rate of 25%. On the date of acceptance notification, she received two emails about the results of A and B, respectively. She read the first email and was excited to know that A was accepted to appear at the conference. Conditioned on what she observed so far, what is the probability Ling got both A and B accepted.
    A. 1/16
    B. 1/4
    C. 1/2
    D. 1

(B) Assume that event A means paper A accepted while event B means paper B accepts. Then P(AB|A)=0.25*0.25/0.25=0.25

36. (1 correct choice only) Given a function f(x)=K (for any x), where K is a constant. The derivative for f(x) is:
    A. K

B. 1
C. 0
D. x

(C) P12, Lecture 5.

37. (1 correct choice only) Given the function $f(x) = x^3 \cdot e^{(x^4+2)}$ and we want to calculate its indefinite integral with the chain rule. Which of the following is a good alternative to construct the composite function $f(x) = f(g(x))$?
   A. $g(x) = x^4 + 2$
   B. $g(x) = x^3$
   C. $g(x) = e^{x^4+2}$
   D. $g(x) = e^x$
   (A) If $g(x) = x^4 + 2$, then we can have $dg(x) = 4x^3dx$. Then $f(x) = \frac{1}{4}e^{g(x)}dg(x)$, which allows easy integration with the exponential rule.

38. (1 correct choice only) If x and y are both word count vectors derived from two sentences, which of the following describes the most precise range of the angle between them?
   A. $[0,\frac{\pi}{2}]$
   B. $[0,\frac{\pi}{4}]$
   C. $[0,\pi]$
   D. $[0,2\pi]$

(A) Because both x and y are vectors where all entries are non-negative, their cosine similarity will be in the range of [0,1]. So the angle between them should be in the range of [0,π/2].

39. (1 or multiple correct choice(s)) Naïve Bayes is a(n) ____ classifier.
   A. discrete
   B. generative
   C. linear
   D. non-linear

   (BC) It is a generative classifier because it builds the model for each class (measured with the posterior P(c|d) P30, Lecture 2). It is a linear classifier because the model just maximizes the sum of weights (P38, Lecture 2).

40. (1 or multiple correct choice(s)) Which of the following is a factor allowing data analytics to become popular in the last decade.
   A. Better models.
   B. More power machines.
   C. The availability of large-scale data.

   (ABC) P47 Lecture 1.

41. (1 or multiple correct choice(s)) Given a discrete random variable X, find the correct statement(s):
   A. E(aX) = aE(X)
   B. E(aX+b) = aE(X)+b
   C. Var(X) = E((X-E(X))^2)
   D. Var(X) = E(X^2) – E(X)^2

   (ABCD) Page 5 and 7, Lecture 3.

42. (1 or multiple correct choice(s)) Which of the following are supervised machine learning algorithms?

A. Linear Regression

B. Logistic Regression

C. Naïve Bayes

D. K-means

(ABC) The algorithms in ABC are all supervised learning methods, which aim to learn the map between data (x) and labels (y) (P30, Lecture 2). K-means is unsupervised learning, where only the data is given without labels (P27, Lecture 4).

43. (1 or multiple correct choice(s)) Which of the following statements must be wrong for any given events A and B?

A. P(B|A)<P(AB)

B. P(B)=P(B|A)

C. P(AB)=P(A)P(B)

D. P(A|A)=0

(AD) For A, P(B|A)=P(AB)/P(B)>P(AB). B and C might be correct if the two events are independent. For D, P(A|A)=P(A)/P(A)=1.

44. (1 or multiple correct choice(s)) Given two vectors x, y and a scalar a, find the correct statement(s) in the following:

A. ||ax|| = |a| ||x||

B. ||x+ y|| =||x|| + ||y||

C. ||x|| = 0 only If x = 0

D. It is possible for ||x||<0.

(AC) P18, Lecture 4.

45. (1 or multiple correct choice(s)) Given a function $f(x, y, z) = -\frac{1}{\sqrt{x^2+y^2+z^2+xyz}}$, which of the following is an entry in its gradient.

A. $-\frac{1}{2}(yz + 2x)(x^2 + y^2 + z^2 + xyz)^{-\frac{3}{2}}$

B. $\frac{1}{2}(xy + 2z)(x^2 + y^2 + z^2 + xyz)^{-\frac{3}{2}}$

C. $-(x^2 + y^2 + z^2 + xyz)^{-\frac{1}{2}}$

D. $-\frac{1}{2}(xz + 2y)(x^2 + y^2 + z^2 + xyz)^{-\frac{1}{2}}$

(B) The three entries of the gradient are:

$$\frac{\partial f(x,y,z)}{x} = \frac{1}{2}(yz + 2x)(x^2 + y^2 + z^2 + xyz)^{-\frac{3}{2}}$$

$$\frac{\partial f(x,y,z)}{y} = \frac{1}{2}(xz + 2y)(x^2 + y^2 + z^2 + xyz)^{-\frac{3}{2}}$$

$$\frac{\partial f(x,y,z)}{z} = \frac{1}{2}(xy + 2z)(x^2 + y^2 + z^2 + xyz)^{-\frac{3}{2}}$$

46. (1 or multiple correct choice(s)) Prof. K was concerned that over 10% of people in HK caught COVID-19 (the null hypothesis H0). So, he invited 400 people in HK to do a COVID-19 test, where the results from 38 of them were positive. Suppose that the accuracy of this COVID-19 test is

100% and it is known that the infection rate of COVID-19 satisfies the normal distribution with the standard deviation of 0.1. Then, we will ______.

A. reject H0 at the significance level of 10%

B. reject H0 at the significance level of 5%

C. accept H0 at the significance level of 10%

D. accept H0 at the significance level of 5%

(AB) Suppose the infection rate at the sample test $\overline{X} = \frac{38}{400} = 0.095$ and the infection rate at the population satisfies $N(\mu, \sigma^2)$, where $\sigma = 0.1$. The p-value is $P(\overline{X} \leq 0.095) = P\left(\frac{\overline{X}-\mu}{\sigma} \leq \frac{0.095-0.1}{\frac{0.1}{\sqrt{400}}}\right) = \phi(-1) = 0.1587 > 0.1 > 0.05.$

47. (True or False) The matrix in R programming can be understood as a two-dimensional array. Each element must have the same data type and be created using the command *matrix*.

True. P34-36, Lecture 6.

48. (True or False) In R programming, the symbol NaN can be used to represent missing values of the data for some imperfect dataset.

False. The symbol NA is used to represent missing values. P45, Lecture 6.

49. (True or False) The R code `x=seq(-4,4,0.01); plot(x, pnorm(x, 0, 1), col = "red");’ draws the density function diagram of normal distribution.

False. The R function `pnorm(q)’ for the definition of cumulative probability function instead of the density function. P56, Lecture 6.

50. (True or False) Given the following R code,

```r
 patientID<-c(1,2,3,4);

 age<-c(25,34,28,52);

 diabetes<-c("Type1","Type2","Type1","Type1");

 status<-c("Poor","Improved","Excellent","Poor");

 patientdata<-data.frame(patientID,age,diabetes,status);
```

The command of `patientdata[1:2][2,2]` queries the age of the patient with the ID 2.

True. The system will return the second row (corresponding to patient ID 2) and second column (corresponding to the age attribute) of the `patientdata' dataframe. P37-38, Lecture 6.

51. (1 correct choice only) If the running result of the following R code is 65535, the n value at line `` x.n(x=2,n=?)'' should be _____.

```r
x.n <- function(x,n){
  h <- 0
  for(i in 0:n){
    h <- h+x^i
  }
  return(h)
}
x.n(x=2, n=?)
```

A. 13

B. 14

C. 15

D.16

(C) (x.n=1+2+4+..+2^{n}=2^{n+1}-1=65535\\n=log$_2$65536 -1=15)

52. (1 correct choice only) Which result does the following code describe? (A)

```r
r.n <- function(r,n){
  a <- prod(2:r)/(prod(2:(r-n)*prod(2:n))
  return(a)
}
```

A. n choose r
B. n permute r
C. r choose n
D. r permute n

(C) The code is to calculate $\frac{r!}{(r-n)!\,n!}$. So it is r choose n.