COMP1433 Mid-term

1

For example, in the combat with Covid-19, data analytics (Big data) were used to determine whether a certain person have close contact with someone that was confirmed positive for Covid-19.

1.a

Motivation: this data analytics is very useful. Since the virus is invisible, one will never know if he ever exposed to virus unless certain symptoms manifest, which takes time. However, under the help of data analytics, we can instantly know whether ever have close contact with confirmed patient. If yes, we can go to hospital in advance, this may save our life.

1.b

Data: the data may mainly comprise the travel data, for example, flight information, train information.

Attributes: geographical location, travel time, seat location, etc.

1.c

Expected info: Whether someone stayed with patients same time and same location. The distance of seats between someone and patients on vehicle.

1.d

Possible technique: I think it may use Statistical Method.

Reason: The data is huge (As Spring Festival travel rush), so it is unpractical to examine every case individually. If someone looks up in system, I guess the system may undergo a hypothesis testing with a null hypo H0: “have contact with patient”, and then use several statistical variables (time, location..) to judge that hypothesis.

2

2.a

Let A, B denote the event that the integer is divisible by 2 and 3 respectively.
There are $\left\lfloor \frac{100}{2} \right\rfloor = 50$ integer divisible by 2
$\left\lfloor \frac{100}{3} \right\rfloor = 33$ integer divisible by 3,
and $\left\lfloor \frac{100}{2\times3} \right\rfloor = 16$ integers divisible by both 2 and 3.
So,

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = \frac{50}{100} + \frac{33}{100} - \frac{16}{100} = \frac{67}{100}$$

2.b

Let X denote the number of questions that are correct. To Pass the exam, he shall have $X \geq 6$.

for each question, the probability that he get correct answer is $\frac{1}{4}$
$P(X \geq 6) = p(X = 6) + p(X = 7) + p(X = 8) + p(X = 9) + p(X = 10)$
$X \sim B(10, \frac{1}{4})$, so $p(X = k) = \binom{10}{k}(\frac{1}{4})^k(\frac{3}{4})^{10-k}$ (1)
Apply aforementioned general formula (1), we can get $P(X \geq 6) \approx 0.0197$

2.c

This question can be solved by enumeration.
So the basic idea is to enumerate all cases and judge whether the case satisfy certain condition.

I used Python to generate all possible outcomes with following code.
There in all 17^3 possible cases. In the nested for loop, each iteration filters one case.

```
#python 3.7.4 on Windows
all = []  # All cases that x1 +x2 +x3 == 17
for i in range(1, 18):  # from 1 to 17
    for j in range(1, 18):
        for k in range(1, 18):
            if i+j+k == 17:
                all.append([i, j, k])# use python list to implement vector

ans = []  # Cases that x1 +x2 <= 8
for case in all:
    if case[0] + case[1] <= 8:
        ans.append(case)
prob = len(ans)/len(all)
```

Let A denotes $x_1 + x_2 + x_3 = 17$, B denotes $x_1 + x_2 \leq 8$
It turns out that the sample size of A is 120, and sample size of B is 28.
Given every case is evenly likely to happen, we have

$$P(x_1 + x_2 \leq 8 | x_1 + x_2 + x_3 = 17) = \frac{7}{30}$$

2.d

In each test, if the coin is fake, we call it is a failure
let X_i denotes failure at i-th iteration(test), then $p(X_1) = \frac{1}{2}$
if i-th step fails, then step i-1 are also fail.
At the i-th iteration, the probability of failure is $p(X_i) = \frac{i}{i+1}$
Then the general formula for $p(X_i)$ can be expressed recursively:

$$p(X_i) = \begin{cases} \frac{1}{2}, & i = 1 \\ p(X_{i-1})\frac{i}{1+i}, & i \geq 2 \end{cases}$$

So, failure at 9-th iteration is $p(X_9) = \frac{1}{2} * \frac{2}{3} * \ldots\ldots * \frac{9}{10} = \frac{1}{10}$

2.e

Let S, N denote the event that random selected news is sports news and non-sports news respectively. And A, B be the event than observe "ball" and observe "player" respectively.

Then $p(A|S) = 0.8, p(B|S) = 0.7, p(A|N) = 0.1, p(B|N) = 0.1$
$p(S) = 0.2, p(N) = 1 - p(S) = 0.8$

By the *law of total probability*,

$p(A) = p(A|S)p(S) + p(A|N)p(N) = 0.24$
$p(B) = p(B|S)p(S) + p(B|N)p(N) = 0.22$

$p(A \cap S) = p(A|S)p(S) = 0.16, p(B \cap S) = 0.14$
Given that A and B are independent, $p(AB|S) = p(A|S)p(B|S) = 0.56$
$p\big((A \cap B) \cap S\big) = p(AB|S)p(S) = 0.112$, Also we can get $p\big((A \cap B) \cap N\big) = 0.008$
So, $p(S|A \cap B) = \frac{p\big(S \cap (A \cap B)\big)}{p(A \cap B)} = \frac{p\big(S \cap (A \cap B)\big)}{p\big((A \cap B) \cap S\big) + p\big((A \cap B) \cap N\big)} = \frac{14}{15}$

3

3.a

Let X denotes temperature
Sample mean: $\bar{X} = \frac{\sum X}{N} = 37.9$
variance: $\sigma^2 = \frac{\sum (X_N - \mu)^2}{N} = 1.76$

standard deviation $\sigma = \sqrt{\sigma^2} = 1.327$
medium = 38.4
range = $39.2 - 36.2 = 3$

3.b

I will adopt *Normal Distribution* model to reflect people's average temperature in this city.

To approximate the parameter in Normal distribution,
I will randomly sample N testees and test their temperature, then we can get Sample mean $\bar{X}$ and sample variance $Var(\bar{X})$

Then I will use that to approximate populations' statistical parameter$\mu_{\bar{X}}, \sigma_{\bar{X}}$.
That is to say, use the unbiased estimation to approximate population mean.

$$\mu_{\bar{X}} = \bar{X}, \sigma_{\bar{X}} = \frac{Var[\bar{X}]}{N}$$

(BTW, by the *law of large number*, the larger N is, the more precise our approximation the population mean is.)

Put it together, my model is $N(\mu_{\bar{X}}, \frac{Var[\bar{X}]}{N})$

3.3

Let $\bar{X} = 37.3$ denote sample mean.
our null hypothesis $H_0: \mu = 37$
According to survey, we take $\sigma = 0.4$
the p-value is

$$P(|\bar{X} - \mu|) \geq 0.3 = P(\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}) \geq \frac{0.12}{(\frac{0.4}{\sqrt{16}})} = 2\Phi(-1.2) = 0.230 > 0.1 > 0.05$$

So we will reject H0 both at significance 0.05 and 0.10.

4

4.1

We can use *Adjacency matrix* to represent such data.
Each column represent a user entry, same as their indices.
If i know j, then row i, column j (i, j) will be set to 1, otherwise 0.
(Note that (i,i) will always be 1)
for example:

$$\begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \end{pmatrix}$$

means there are 3 users,
user 1 only know himself,
user 2 know himself and user 3,
user 3 know user 1, 2, 3.

4.2

suppose the data is stored as mentioned in 4.1.

To measure the similarity between two users, we shall calculate their distance.
(Generally Euclidean distance)
Steps:

1. extract two user entries (column) form the adjacency matrix (then the entry shall be a vector)

2. calculate their Euclidean distance$||x - y|| = \sqrt{(x_1 - y_1)^2 + (x_1 - y_1)^2 + \ldots + (x_N - y_N)^2}$, where N is the dimension of vector (the number of all users)
3. The similarity can then be demonstrated by that distance, the closer, the more similar.

For example, continue the example in 4.1,
user 1:(1,0,0) user3:(1,1,1)
distance:$\sqrt{2}$, which is larger than that between user2 and user3 ($\sqrt{1}$).
So user3 is more similar to user2 than to user1 in terms of the users they known.

5

5.1

The *probability distribution function* follows:

$$F_X(x) = \int_{-\infty}^{x} f_X(x)\, dx$$

1. if $(x < 0)$, $f_X(x) = \frac{d}{dx}0 = 0$
2. if $(0 \le x \le 1)$, $f_X(x) = \frac{d}{dx}x = 1$
3. if $(x > 1)$, $f_X(x) = \frac{d}{dx}1 = 0$

So, the *probability distribution function* is:

$$f_X(x) = \begin{cases} 1, & 0 \le x \le 1 \\ 0, & otherwise \end{cases}$$

5.2

$$E[X] = \int_{-\infty}^{x} x f_X(x)\, dx = \int_{0}^{1} x\ dx = \frac{x^2}{2}|_0^1 = \frac{1}{2}$$

6

As a programming language, R has its strengths and weaknesses.

Strengths:

1. R provide a great deal of built-in functions and data models, along with many user-distributed packages on R official website(i.e. cran.r-project.org). Whats more, R release integrates package manager, which makes it much more easier to use.
   All of these make R a strong, easy-to-use language for statistical analysis.

2. R provide many different format to store R objects, for example, .r is for r scripts, .RDS and .RData are used to store R objects. This design separates algorithm and data. More specifically, it is quite easy in R to reuse data, as R provides its own format(.RDS, .RData) and corresponding functions (`save( )`, `load( )`).
   These features make it easier to manage our project.

3. R is open-source and **free**, so it is friendly to learners. (compared with MATLAB).

Weakness:

1. R is designed for data processing, and it may not do some low-level work as other programming language. For example, R cannot used for embedded development, and it may also infeasible to build an OS or a compiler.

2. R, as an interpretive language, has relatively lower efficiency compared with compiled language. Things can get worser when the data set is huge.

3. The quality of R packages is guaranteed. As many developers are in fact statisticians (not computer scientists), the performance of their package can be really poor.