SNU Fourth Industrial Revolution Academy

Basic Math for Big Data

Homework 3

Due: August 3, 10:00 AM

## Reminders

- T.A.: Chiwan Park (chiwanpark@snu.ac.kr)
- The points of this homework add up to 135.
- This has to be done individually like all the homeworks.
- Please answer clearly; illegible handwriting may get no points.
- Whenever you are making an assumption, please state it clearly.
- If you have a question about assignments, please upload your question in FIRA portal.

## Submissions

- You can submit your homework in the class or via email (only PDFs are accepted).
- Do not submit the homework in a photography form.

## Question 1 [13 points]

Table 1 shows the probability distribution for the number of years to earn a Doctor of Philosophy (Ph.D.) degree. Answer the following questions.

Table 1. Probability distribution

| $x$ | $P(X = x)$ |
|---|---|
| 3 | 0.05 |
| 4 | 0.40 |
| 5 | 0.30 |
| 6 | 0.15 |
| 7 | 0.10 |

(a) Fill in the missing value on Table 1. [3 points]

(b) What does it mean that the values 0, 1, and 2 are not included for $X$ on the probability distribution? [5 points]

There is no student which earns a Ph.D. degree in 2 years.

(c) How many years do we expect to earn a Ph.D. degree? [5 points]

$E(X) = 4.85$ years

## Question 2 [20 points]

Consider a biased coin that comes up heads with a probability $p > 0.5$. The probability of more than 25 heads in 45 tosses is approximately equal to $P(z > \frac{25-27}{3.29})$ where $z$ follows a standard normal distribution. Answer the following questions. Note that all your answers should be rounded off to three decimal points.

(a) What is the value of $p$? [10 points]

From the description above, we obtain $P(X > 25) \approx P\left(z > \frac{25-27}{3.29}\right)$ where $X$ follows a binomial distribution with 45 trials and the probability $p$. Recall that $P(Y > y) = P\left(z > \frac{y-\mu}{\sigma}\right)$ where $Y$ follows a normal distribution with mean $\mu$ and standard deviation $\sigma$. Therefore, the probability distribution of $X$ is approximately equal to a normal distribution with mean $\mu = 27$ and standard deviation $\sigma = 3.29$.

Setting the expected value in terms of the parameters of the binomial distribution equal to the mean of the normal distribution, we have $np = 45 \cdot p \approx 27$. Thus, $p \approx \frac{27}{45} = 0.6$.

(b) Find the probability of 65 heads in 100 tosses. [10 points]

$$P(65) = C(100, 65) \cdot 0.6^{65} \cdot 0.4^{35} \approx 0.0491$$

## Question 3 [25 points]

The probability density function of $X$, the decaying period (measured in seconds) of a certain type of radioactive molecule, is given by:

$$f(x) = \begin{cases} \frac{10}{x^2} & x \geq 10 \\ 0 & x < 10 \end{cases}$$

Answer the following questions.

(a) Find $P(X > 20)$. [10 points]

$$P(X > 20) = \int_{20}^{\infty} \frac{10}{x^2} dx = \left[-\frac{10}{x}\right]_{20}^{\infty} = \frac{1}{2}$$

(b) Assume that there are 6 molecules at the beginning. What is the probability that we observe at least 3 molecules after 15 seconds? (Note that the decaying of radioactive molecule occurs independently.) [15 points]

We first obtain $P(X \geq 15) = \frac{2}{3}$ for a molecule by similar calculation to (a). Since the decaying process occurs independently, we use the binomial distribution. Then, $P(Y \geq 3) = C(6,3) \cdot \left(\frac{2}{3}\right)^3 \cdot \left(\frac{1}{3}\right)^3 + C(6,4) \cdot \left(\frac{2}{3}\right)^4 \cdot \left(\frac{1}{3}\right)^2 + C(6,5) \cdot \left(\frac{2}{3}\right)^5 \cdot \left(\frac{1}{3}\right)^1 + C(6,6) \cdot \left(\frac{2}{3}\right)^6 \cdot \left(\frac{1}{3}\right)^0 = \frac{656}{729}$ where $Y$ is the number of molecules after 15 seconds.

## Question 4 [10 points]

Compute $E(X)$ if $X$ has a probability density function given by:

(a) $f(x) = \begin{cases} c(1-x^2) & -1 < x < 1 \\ 0 & \text{otherwise} \end{cases}$. [5 points]

$$E(X) = \int_{-1}^{1} c(1-x^2)x dx = c\int_{-1}^{1} x - x^3 dx = c\left[\frac{x}{2} - \frac{x^4}{4}\right]_{-1}^{1} = 0$$

(b) $f(x) = \begin{cases} \frac{5}{x^2} & x > 5 \\ 0 & x \leq 5 \end{cases}$. [5 points]

$$E(X) = \int_{5}^{\infty} \frac{5}{x} dx = [5 \ln x]_5^\infty = \infty$$

## Question 5 [16 points]

A scientist starts to wait for a signal from his/her co-worker at 7:30 AM. Assume that the signal will arrive at some time uniformly distributed between 7:30 AM and 8 AM. Answer the following questions.

(a) What is the probability that the scientist should wait longer than 10 minutes? [8 points]

Let $X$ be the waiting time of the scientist. Then, $X$ follows a uniform distribution with range 0 minute to 30 minutes. Thus, $P(X > 10) = 1 - P(X \leq 10) = 1 - \frac{10}{30} = \frac{2}{3}$.

(b) Assume the signal has not arrived at 7:45 AM. What is the probability that the signal will not arrive until 7:55 AM? [8 points]

If the signal has not arrived at 7:45 AM, the waiting time of the scientist is at least 15 minutes. Similarly, the waiting time of the scientist is at least 25 minutes if the signal will not arrive until 7:55 AM. Thus, our desired probability is $P(X \geq 25|X \geq 15) = \frac{P(X\geq 25 \cap X\geq 15)}{P(X\geq 15)} = \frac{P(X\geq 25)}{P(X\geq 15)}$. Using $P(X \geq 25) = \frac{5}{30} = \frac{1}{6}$ and $P(X \geq 15) = \frac{15}{30} = \frac{1}{2}$. The probability we want is $\frac{1}{3}$.

## Question 6 [16 points]

Figure 1 shows 5 data points $\{(-2,-2),(0,-1),(0,0),(0,1),(2,2)\}$ on a $xy$-plane. We use a simple linear regression model $y = ax + b$ to describe the relationship between $x$ and $y$. Answer the following questions.

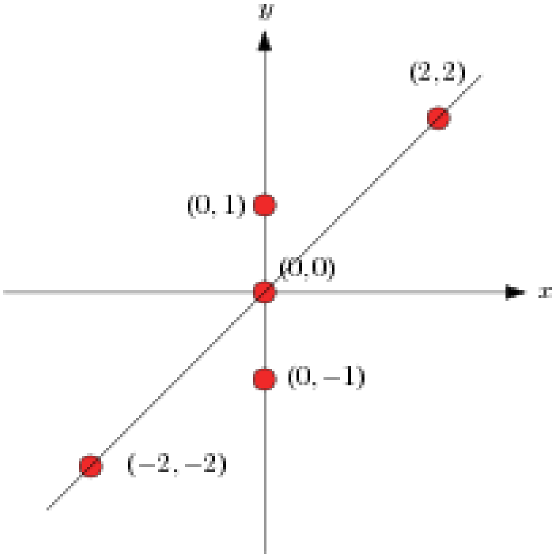

Figure 1. 5 data points on a $xy$-plane

(a) Derive the equations to find $a$ and $b$ using the least-squares method, and find the values of $a$ and $b$. [8 points]

From the equation of the least-squares method, $a = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2}$ and $b = \bar{y} - a\bar{x}$. Applying the given data to the equations, we obtain $a = \frac{(-2)^2+2^2}{(-2)^2+2^2} = 1$ and $b = 0 - 0 = 0$.

(b) Draw the best-fit line on Figure 1. [4 points]

(c) Calculate the squared error between the data points and the model. [4 points]

$$MSE = \frac{\sum\left(y_i - (ax_i + b)\right)^2}{n} = \frac{2}{5}$$

## Question 7 [35 points]

For this programming-based question, you will write one R script (subway.R) to analyze Seoul subway traffic data. The dataset and skeleton code are contained in a zip file that you can download the following link: https://datalab.snu.ac.kr/fira/subway.zip.

The zip file contains a comma-separated-value (CSV) file (subway.csv) containing the number of passengers entering and exiting subway stations during January 2015. Each line contains the number of passengers of a subway station for a day. There are five columns in the file:

- date: the date of the observation in YYYYMMDD format (e.g. 20150101).
- line: the line number of subway station (e.g. 2호선)
- station_name: the name of subway station (e.g. 강남)
- get_in: the number of passengers entering the station for the date
- get_off: the number of passengers exiting the station for the date

Fill in the R script to answer the following questions. After that, send an email with your script to the T.A.'s email address (chiwanpark@snu.ac.kr).

(a) What is the average number of passengers of Sinchon (신촌) station with line number 2? [5 points]

97537.4

(b) Find top 10 subway stations in terms of the average number of passengers. [10 points] (Hint: you may use sort function in R)

강남, 신림, 구로디지털단지, 삼성(무역센터), 서울대입구(관악구청), 영등포, 신촌, 잠실(송파구청), 역삼, 을지로입구

(c) Find top 3 subway lines in terms of the average number of passengers. [10 points]

2호선, 7호선, 4호선

(d) Draw scatter plots for the number of passengers of Nakseongdae (낙성대) station and Incheon Int'l Airport (인천국제공항) station during January 2015. You may see two different patterns in the plots. Give a reason why the patterns are different from each other. [10 points]

Figure 2 and Figure 3 show the number of passengers of Incheon Int'l Airport station and Nakseongdae station, respectively. Note that the day of the week of beginning is Thursday. The number of passengers of Nakseongdae station during weekdays is larger than the number of passengers during weekends and holidays, while the number of passengers of Incheon Int'l Airport station during weekdays is lower than the number of passengers during weekends. One of possible explanation is that most of the passengers of Nakseongdae station are students and faculty, while most of the passengers of Incheon Int'l Airport station are tourists.

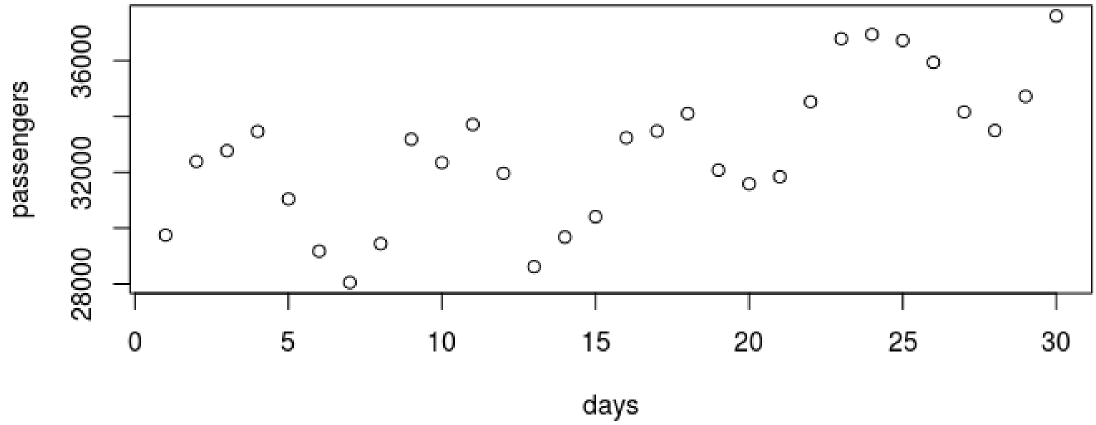

Figure 2. The number of passengers of Incheon Int'l Airport station

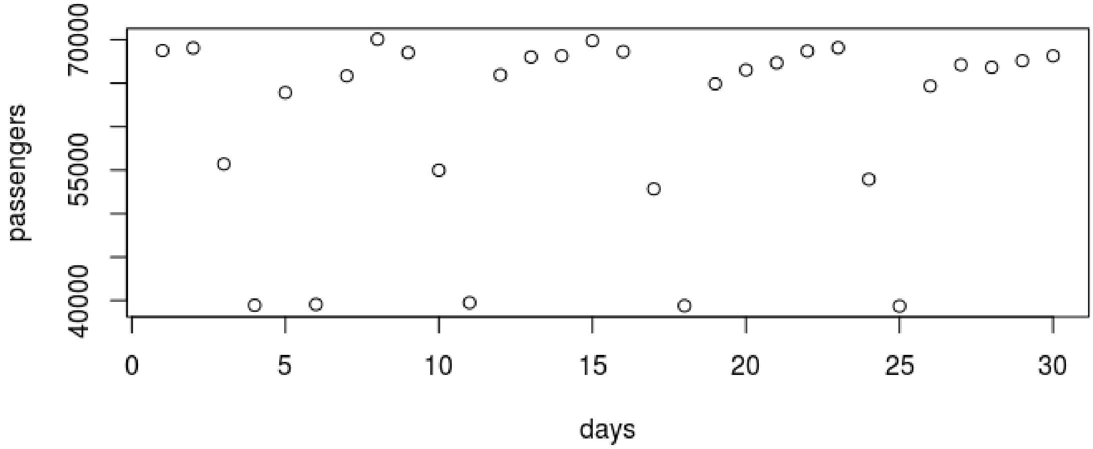

Figure 3. The number of passengers of Nakseongdae station