CS544 Module 2 Assignment

General Rules for Homework Assignments

- You must work on your assignments individually. You are not allowed to copy the answers from the others.
- Each assignment has a strict deadline. If there is a delay, you must be in touch with the instructor. Late submissions without reasons will result in a grade deduction.
- When the term *lastName* is referenced in an assignment, please replace it with your last name.

Part1) Probability - 60 points

For questions a) and b), show the solutions with the calculations without using R. Then, verify with R code.

a) A disease affects 5 out of 100 people on average. The sensitivity of a clinical test to detect the disease is 92%, which means 92% of people who have the disease get positive test results. Its false positive rate is 6%, which means 6% of people who do not have the disease get positive results in the tests. What is the chance that a randomly selected person with a positive result does not have the disease? What is the chance that a randomly selected person with a negative result actually has the disease?

- **P(D) = 5/100 = 0.05 (prevalence of the disease)**
- **P(N) = 1 - P(N) = 1 - 0.05 = 0.95 (probability of not having the disease)**
- **P(S) = 0.92 (sensitivity - probability of testing positive given that the person has the disease)**
- **P(F) = 0.06 (false positive rate - probability of testing positive given that the person does not have the disease)**

From these, we can also derive:

- **P(W) = 1 - 0.92 = 0.08 (false negative rate - probability of testing negative given that the person has the disease)**
- **P(C) = 1 - 0.06 = 0.94 (specificity - probability of testing negative given that the person does not have the disease)**

The chance of a randomly selected person with a positive result does not have the disease:

P(T+) and P(T-) are positive and negative test results…

- **$P(T+| D) = P(S) = 0.92$**
- **$P(T+| N) = P(F) = 0.06$**
- **$P(T-| D) = P(W) = 0.08$**
- **$P(T-| N) = P(C) = 0.94$**

We are looking for P(N|T+)

First, we need to calculate P(T+), the overall probability of a positive test result:

P(T+) = P (T+| D) ∗ P(D) + P(T+| N) ∗ P(N)

P(T+) = (P(S) ∗ P(D)) + (P(F) ∗ P(N))

P(T+) = (0.92 ∗ 0.05) + (0.06 ∗ 0.95)

P(T+) = 0.046 + 0.057 = 0.103

Now apply Bayes' theorem to find P(N|T+):

$$\mathbf{P(N|\ T+) = \frac{P(F)*P(N)}{P(T+)}}$$

$$\mathbf{P(N|\ T+) = \frac{0.06*0.095}{0.103}}$$

$$\mathbf{P(N|\ T+) = \frac{.057}{.103} \approx .05534}$$

The chance that a redomly selected person with a positive result does not have the disease is about 55%.

What is the change that a randomly selected person with a negative result actually has the disease?

We are now looking for P(D| T-)

First, we need to calculate P(T-), which is the overall probability of a negative test which will just be subtracted to out findings in the previous step…

P(T−) = 1 − P(T+)

P(T−) = 1−0.103 = 0.897

Now apply Bayes' theorem to find P(D|T−):

$$\mathbf{P(D|\ T-) = \frac{P(W)*P(D)}{P(T-)}}$$

$$\mathbf{P(D|\ T-) = \frac{0.08*0.05}{0.897}}$$

$$\mathbf{P(D|\ T-) = \frac{.004}{.897} \approx .004459}$$

The chance that a randomly selected person with a negative result actually has the disease is about 0.45%.

b) Suppose that in a particular state, among the registered voters, 42% are Democrats, 48% are Republicans, and the rest are independents. A ballot question is whether to provide universal healthcare to citizens. Suppose that 85% of Democrats, 50% of Republicans, and 60% of Independents favor universal healthcare. If a person chosen at random **does not** favor universal healthcare, what is the probability that the person is ***i) a Democrat? ii) a Republican iii) an Independent.***

D = Democrat R = Republican and I = Independepent
Probability P(D) = 0.42
P(R) = 0.48

P(I) = 0.10

Let *Y* represent Yes for favoring universal health care and *N* for not favoring universal healthcare

Probability of favoring healthcare:

P(Y| D) = 0.85 (85% of Democrats favor)
P(Y| R) = 0.50 (50% of Republicans favor)
P(Y| I) = 0.60 (60% of Independents favor)

From these we can derive percentages for registered voters that do *not* favor heathcare:

P(N| D) = 1 - P(Y| D) = 1 - 0.85 = 0.15 (15% of Democrats do *not* favor)
P(N| R) = 1 - P(Y| R) = 1 - 0.50 = 0.50 (50% of Republicans do *not* favor)
P(N| I) = 1 - P(Y|I) = 1 - 0.60 = 0.40 (60% of Independents do *not* favor)

First we find the overall probability of a person *not* favoring universal healthcare, P(N):

P(N) = P(N| D) P(D) + P(N|R) P(R) + P(N|I) P(I)
P(N) = (0.15)(0.42) + (0.50)(0.48) + (0.40)(0.10)
P(N) = 0.063 + 0.24 + 0.04 = 0.343

Probability that the voter is a Democrat and do *not* favor universal healthcare (P(D|N)):
P(D|N) = [P(N|D) * P(D)] / P(N)
P(D|N) = (0.15 * 0.42) / 0.343
P(D|N) = 0.063 / 0.343 ≈ 0.1837

Probablility that the person is a *Democrat* is 18.3%

Probability that the voter is a *Republican* and do *not* favor universal healthcare (P(R|N)):
P(R|N) = [P(N|R) * P(R)] / P(N)
P(R|N) = (0.50 * 0.48) / 0.343
P(R|N) = 0.24 / 0.343) ≈ 0.6997

Probablility that the person is a *Republican* is 69.97%

Probability that the voter is a *Independent* and do *not* favor universal healthcare (P(I|N)):
P(I|N) = [P(N|I) * P(I)] / P(N)
P(I|N) = (0.40 * 0.10) / 0.343
P(I|N) = 0.04 / 0.343 ≈ 0.1166

Probablility that the person is a *Independent* is 11.6%

c) In a drama class, there are 12 females and 10 males. A group of 15 is to be chosen at random for a play. i) What is the probability that all males were chosen? ii) What is the probability that at least 5 males were chosen? Iii) What is the probability that at most 5 males were chosen?

i) **Total number of students = 12 females + 10 males = 22 students. A group of 15 is to be chosen.**

The total number of ways to choose 15 students from 22 is given by the combination formula:

C(n,k)= $\frac{n!}{k!(n-k)!}$

C(22,15) = $\frac{22!}{15!(22-15)!} = \frac{22!}{15!(7)!}$ **= 170,544**

If the first 10 students are male the rest of the 5 students must be females chosen from the 12

C(n,k)= $\frac{n!}{k!(n-k)!}$

C(12,5) = $\frac{12!}{5!(12-5)!} = \frac{12!}{5!(7)!}$ **= 792.** $\frac{792}{170,544} \approx$ **0.00464**

The probability that all males were chosen is 0.4%.

ii) **In order to find if at least 5 males were chosen we use all of the combos from 5-10 males**

So

- **5 males and 10 females: C(10,5) × C(12,10) =252 × 66 = 16,632**
- **6 males and 9 females: C(10,6) × C(12,9) = 210 × 220 = 46,200**
- **7 males and 8 females: C(10,7) × C(12,8) = 120 × 495 = 59,400**
- **8 males and 7 females: C(10,8) × C(12,7) = 45 × 792 = 35,640**
- **9 males and 6 females: C(10,9) × C(12,6) = 10 × 924 = 9,240**
- **10 males and 5 females: C(10,10) × C(12,5) = 1×792 = 792**

Total number of ways for at least 5 males = 16,632 + 46,200 + 59,400 + 35,640 + 9,240 + 792 = 167,904. $\frac{167,904}{170,544} \approx$ **0.98452 or 98% that at least 5 males were chose**

The probability that at most 5 males were chosen we use all of the combos of 5,4,and 3 as the minimum as there are only 12 females.

- **3 males and 12 females: C(10,3) × C(12,12) = 120 × 1 = 120**
- **4 males and 11 females: C(10,4) × C(12,11) = 210 × 12 = 2,520**
- **5 males and 10 females: C(10,5) × C(12,10) = 252 × 66 = 16,632**

Total number of ways for at most 5 males = 120 + 2,520 + 16,632 = 19,272. $\frac{19,272}{170,544} =$ **0.11299 or 11% that at most 5 males were chosen.**

Part2) R - 40 points

Using function data() to load R data set *airquality*.

Provide R code and output for all of the following.

a) Use the diff function to calculate the temperature differences between consecutive days. Insert the value 0 at the beginning of these differences. Add this result as the DIFFS column of the data frame.

b) Calculate the number of days that are warmer than the previous day?

c) Show the mean and median temperature for each month (May to Sept.). Do not hard code the month. Print out your result properly. For example, “The average and median temperature for May are xxx and xxx; …”.

d) Show the coldest and hottest days each month (May to Sept.). Do not hard code the month. Print out your result properly. For example, “The coldest and hottest days in May are xxx and xxx; …”.

Submission:

Upload your result file to the Assignments section of Blackboard.

Provide all R code in a single file, CS544_lastName.R. Clearly mark each subpart of each question and add appropriate comments.

If you need to submit more than one file, create a folder, CS544_lastName, and place all files in this folder. Archive the folder (CS544_A1_lastName.zip).