R-assignment 3

The birthday problem

1. Suppose k people are in a classroom. We want to find the probability that at least two of the k people share the same birthday for $k = 2, 3, ..., 50$.

Since there are k people in the room and each has 365 possibilites for a birthday, there are 365^k possible outcomes. The number of possible outcomes for which no 2 have the same birthday is

$$(365)(364)...(365 - k + 1).$$

Thus, the probability, under the assumption that each outcome is equally likely, that no 2 people in the room have the same birthday is,

$$\frac{(365)(364)...(365 - k + 1)}{365^k}.$$

The probability that at least two of the k people share a birthday is

$$1 - \frac{(365)(364)...(365 - k + 1)}{365^k}.$$

Note that we can write

$$p = \frac{(365)(364)...(365 - k + 1)}{365^k} = \prod_{i=1}^{k} \left(1 - \frac{i-1}{365}\right).$$

Thus, the probability that at least two of the k people share a birthday is

$$pc = 1 - \prod_{i=1}^{k} \left(1 - \frac{i-1}{365}\right).$$

Do the following in R:

$> \quad k = 50.$

$> \quad pc = numeric(k)$ (to initialize the vector pc)

$> \quad for$ (i in 2:k) {
$p = prod(1 - (0 : (i - 1))/365)$
$pc[i] = 1 - p\}$

This will compute the probability for $k = 2, 3, ..., 50$.
Now do

$> \quad x = c(1 : k)$

(A) Put x and pc in a data frame and print out the data frame to turn in.
(B) Plot pc and print out the plot to turn in.
(C) What is the probability that at least two people share a birthday if there are 50 people in the classroom?
(D) How many people must be in the room for the probability that at least two of the k

people have the same birthday to be greater than 95%?

2. Suppose now that we are considering $10,000$ classrooms each with $k = 50$ students. We will simulate the problem given in part 1 for $k = 50$. Problem 1 gave an "exact" answer. We will now get an approximate answer.

Do the following in R:

> $k = 50.$

> $n = 10000.$

> $y = numeric(n)$

> for (i in 1:n) {
$s = sample(1 : 365, k, replace = T)$
$y[i] = k - length(unique(s))\}$

The sample command, is randomly chosing $k = 50$ numbers between 1 and 365 (365 days in a year).
The command length(unique(s)) counts the number of entries of different values in the vector. Thus, $y[i]$ counts the number of entries with equal value in the vector for each i. In other words, $y[i]$ counts the number of birthday matches for each classroom for classrooms, $i = 1, 2, ..., 10,000$.

Now do:

> $mean(y == 0)$. This will compute the proportions of classrooms with no birthday matches.
> $1 - mean(y == 0)$. This will compute the proportions of classrooms with at least two birthday matches.
The command $y == 0$ finds the entries of the vector y for which $y = 0$.

(A) What is the proportions of classrooms with at least two birthday matches?
(B) How does this compare with what you find in problem number 1.?

Do

> $hist(y, prob = T)$
This gives a relative frequency histogram for the number of birthday matches. The histogram shows the distribution of the number of birthday matches.
(C) Print out the histogram to turn in.

3. In problem 1 and 2, we assumed that there were an equally chance to be born in any day of the year. However, the reality is that more people are born in July and August. Suppose there are approximately three times as many births on days in July and August compared to

the other days of the year. We will now take that into consideration when we simulate the birthday problem for $10,000$ classrooms each with 50 students. There are a total of 61 days in July and August and 304 for rest of the year.
Do the following in R:

> $k = 50.$

> $n = 10000.$

> $y = numeric(n)$

> $weight = c(rep(3, 61), rep(1, 304))$

(This vector repeats 3, 61 times and repeats 304 one time. Thus, it gives 3 times as much weight to 61 of the days compared to the other 304 days). Do

> for (i in 1:n) {
$s = sample(1 : 365, k, replace = T, prob = weight)$
$y[i] = k - length(unique(s))\}$

> $mean(y == 0)$

> $1 - mean(y == 0)$

> $hist(y, prob = T)$

(A) What is the proportions of classrooms with at least two birthday matches?
(B) Print out the histogram to turn in.