Tutorial 1 Worksheet AY 22/23 Sem 1

DSA2101

Birthday paradox

The birthday paradox is commonly introduced in probability classes. It revolves around the question:

> In a set of n randomly chosen people, what is the probability that at least two will share a birthday?

Assuming that *all birthdates are equally likely*, it is possible to compute that we only require 23 people for this probability to be more than half. In this tutorial, we shall relax these assumptions and obtain some answers using simulation.

In all that follows, assume that everyone was born in a non-leap year, and that birthdates are coded 1 to 365.

Classical set up

1. Write an R function `generate_unif_bdates` that
   - takes in a single argument, `n`,
   - generates `n` birthdates using `sample`, and finally
   - checks if there is any "clash". The function should return 1 if there *was* a clash, and 0 otherwise.

```
generate_unif_bdates(20)
```

```
## [1] 0
```

2. Suppose we were to randomly select n students from NUS, and check to see if there were any birthday clashes. Let $X_{i,n}$ be a Bernoulli random variable that takes on the value 1 if there is a clash and 0 otherwise. Then, if all birthdays are equally likely, and are independent of one another, we can compute that

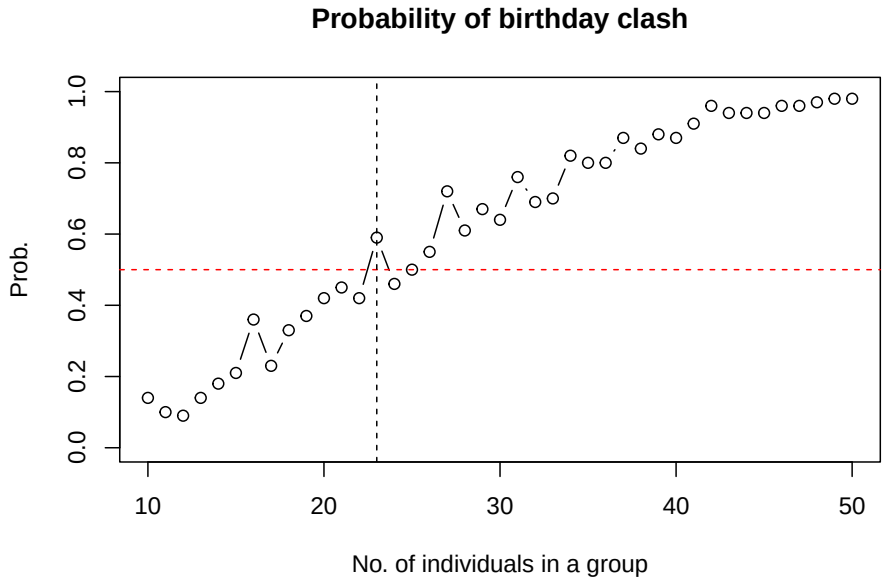
$$P(X_{i,23} = 1) \approx 0.50$$

   Note that i is simply an indication that this random variable corresponds to this particular group of n individuals.

   To estimate this probability using simulation, we would generate multiple observations of $X_{i,23}$ and compute the proportion of times the value 1 (or TRUE) occurs.

   `generate_unif_bdates(23)` generates a single observation of $X_{i,23}$. Repeat this 100 times to see if you obtain a number close to 0.50.

3. Now run 100 simulations each for `n` from 10 to 50. Create a similar graph to this one:



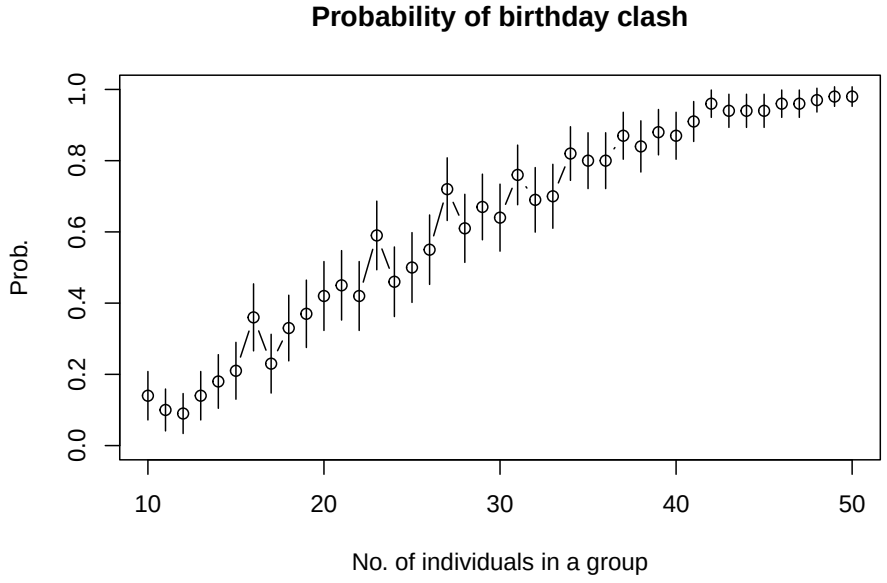
4. As you observed, we do not obtain the same graph every time we re-run the code. We can compute Confidence Intervals to quantify this uncertainty. To compute 95% confidence intervals for the probabilities above, we can use this formula:

$$\hat{p} \pm z_{0.975}\sqrt{\frac{\hat{p}(1-\hat{p})}{100}}$$

where:

- $z_{0.975}$ is the .975-quantile of the N(0,1) distribution. (`qnorm()` function in R)
- $\hat{p} = \frac{1}{100}\sum_{i=1}^{100} X_{i,n}$

Apply the formula above to update your plot:



6. Use the same approach above to estimate the number of people we need in a group in order for the probability of
   - 3 birthdays on one day to be more than 0.75?
   - birthdays within 7 days to be more than 0.5?

References

1. Wikipedia entry on Birthday paradox