Assignment 1 *

Ishaan Kapoor

January 23, 2022

1 Birthday Paradox

A:

k=65

B:

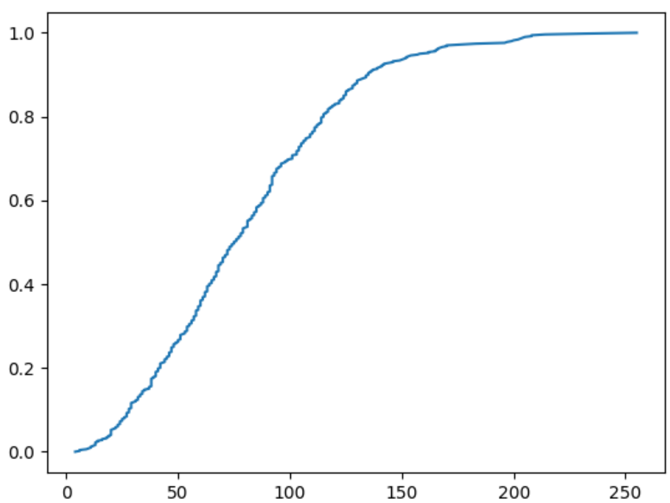

Figure 1: CDF of k for m=500

C:

$\mathbf{E}[k] = 70.782$

*CS 6140 Data Mining; Spring 2022 Instructor: Qingyao Ai, University of Utah

D:

The algorithm for implementing the same is as follows:

- initialise empty list countlist and timelist and set m, n
- loop through 0 to m
- select a random number from 1 to n and store it in num
- append num to numlist
- if num does not exist, continue current proces
- if num exists in the numlist append length of numlist to countlist and go back to step 2
- sort countlist
- calculate cdf of the number by leveraging sorted countlist
- use this to plot the graph

The time taken is:

$$timeelapsed(m = 500, n = 4000) = 0.150614023$$

Graph:

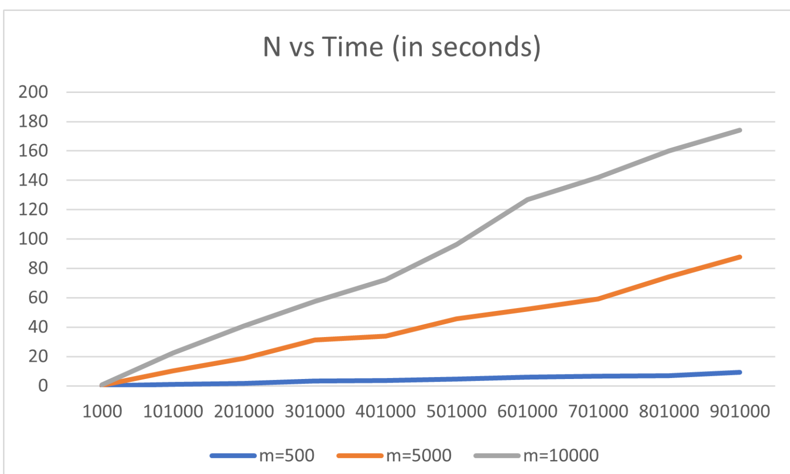

Figure 2: N vs Time for various values of M

2 Coupon Collectors

A:

$k = 1279$

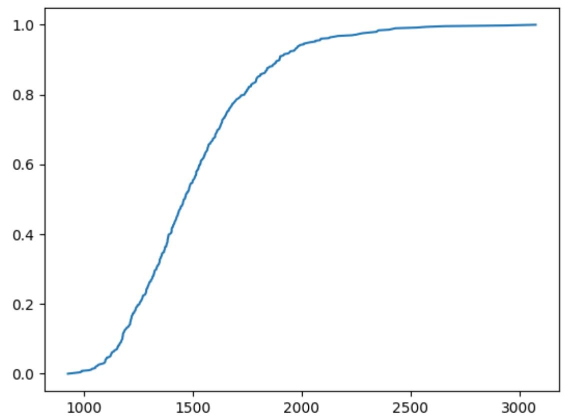

Figure 3: CDF of k for m=500

B:

C:

$\mathbf{E}[k] = 1429.19$

D:

The algorithm for implementing the same is as follows:

- initialise empty list countlist and timelist and set m, n and count=0
- loop through 0 to m
- select a random number from 1 to n and store it in num
- append num to numlist, increment count
- if count¿n convert numlist to set and see if length of set is equal to n
- if check returns true append len(numlist) to countlist and go to step 2
- sort countlist
- calculate cdf of the number by leveraging sorted countlist
- use this to plot the graph

The time taken is:

$$timeelapsed(m = 500, n = 250) = 0.015634536743164062$$

Graph:

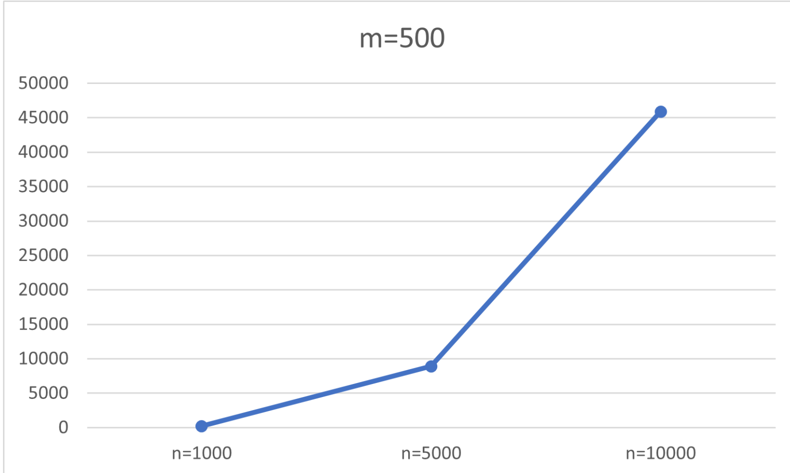

Figure 4: N vs Time for M=500

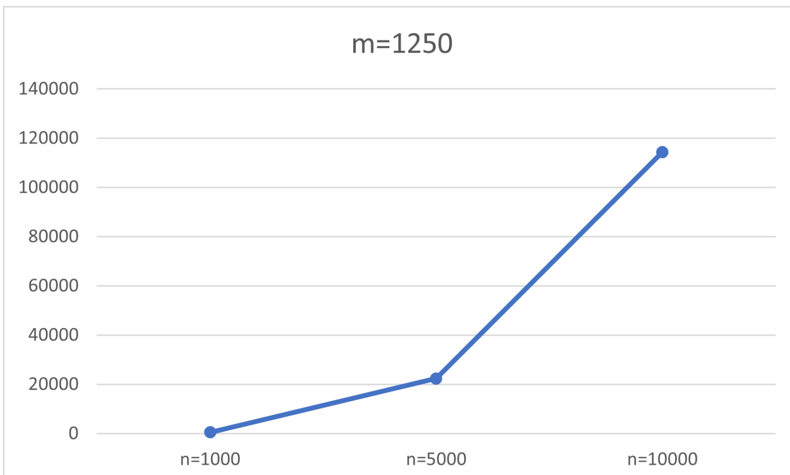

Figure 5: N vs Time for M=1250

3 Comparing Experiments to Analysis (30 points)

A

Probablity of atleast 1 collision after m trails
$Pr[atleast \quad 1 \quad collision] = 1 - Pr[no \quad collision]$
We know,
$Pr[no \quad collision] = 1 \times (1 - \frac{1}{n}) \times (1 - \frac{2}{n})...(1 - \frac{m}{n})$
We use the approximation that:
$1 - \frac{1}{x} \approx e^{-\frac{1}{x}}$
$\implies Pr[X] = 1 - (\sum_{i=1}^{m} e^{-\frac{1}{x}})$
$\implies Pr[X] = 1 - e^{\frac{\sum_{i=1}^{m} i}{n}}$
$\implies Pr[X] = 1 - e^{-\frac{m \times (m+1)}{2n}}$
$\because Pr[X] = 0.5$
$\implies 0.5 = 1 - e^{-\frac{m \times (m+1)}{2n}}$

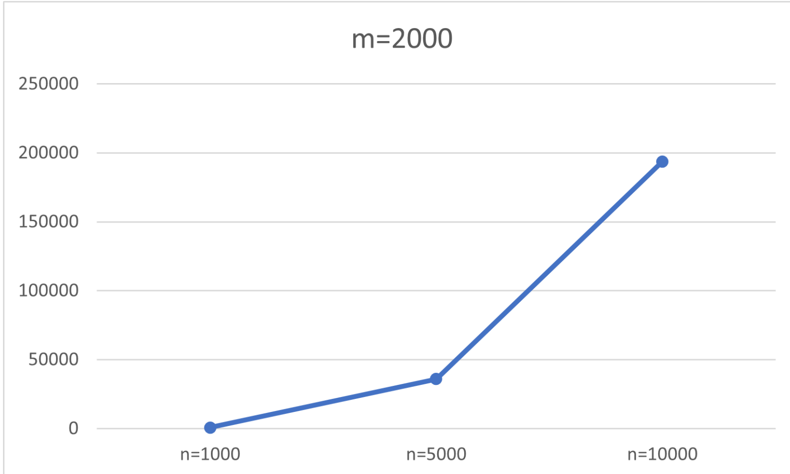

Figure 6: N vs Time for M=2000

$\therefore m = 73.968$
$m \approx 74$
According to 1C, the answer is 70.782. The answer received by empirical calculations is 74, which is very close to our observed answer

B

$E[X] = n \times \log n$
$here,$
$n = 250$
$\implies E[X] = 1380.36$
According to 2C, the answer is 1429. The answer is almost same as we received empirically.