1 - Warm Up

1. Why is it better to take a simple random sample, instead of just the first k rows?

It is better because the data may already be sorted.

2. Suppose we halt reservoir sampling at element m, with m < n, where n is the size of the entire stream. Can this be a sample of the entire data? Explain.

It can be a sample of the entire data if the data is already randomly ordered. However, most likely it wouldn't be, for example, a reservoir sample from 0-80 wouldn't accurately represent the numbers from 0-100.

3. I read on the internet that `shuf -n 100 data.txt` uses reservoir sampling. The following commands each produce 100 lines from `data.txt`. For each command, will it produce a simple random sample of the lines of the file `data.txt`? Why or why not?

```
head -n 100 data.txt | shuf         # 1
shuf -n 100 data.txt | head -n 100  # 2
shuf -n 200 data.txt | head -n 100  # 3
shuf -n 100 data.txt | head -n 100  | sort  # 4
```

#1 This will not produce a random sample because we are only shuffling the first 100 rows of data

#2 This will produce a random sample because if gets a random sample of 100 then gets the first 100 of those (all of them)

#3 This will get a random sample because it gets a sample of 200 items then takes the first 100 of these. Because they are already randomly sampled this is still random

#4 This will not be a random sample, because it takes a random set of 100 then takes the first 100 of those (all of them), then sorts them. Although they are sorted this is still a random sample, if the sorting happened first it wouldn't be a random sample.

2 - Implement Reservoir Sampling

Permalink

(10 pts)

Implement simple or optimal reservoir sampling by writing a program in Julia called `shuf.jl` that works like a simple version of `shuf`. It should accept one positional argument with the number of elements to sample, and default to 100.

Verify that it works for the following cases:

1. `seq 10 | julia shuf.jl` shuffles the integers from 1 to 10.
2. `seq 10 | shuf | julia shuf.jl` shuffles the integers from 1 to 10.
3. `seq 100 | julia shuf.jl 20` samples 20 random integers without replacement from 1 to 100.
4. `seq 1000 | julia shuf.jl` samples 100 random integers without replacement from 1 to 1000.
5. `seq -f "%.0f" 1e7 | julia shuf.jl` samples 100 random integers without replacement from 1 to 10 million.

3 - Hypothesis Testing

Permalink

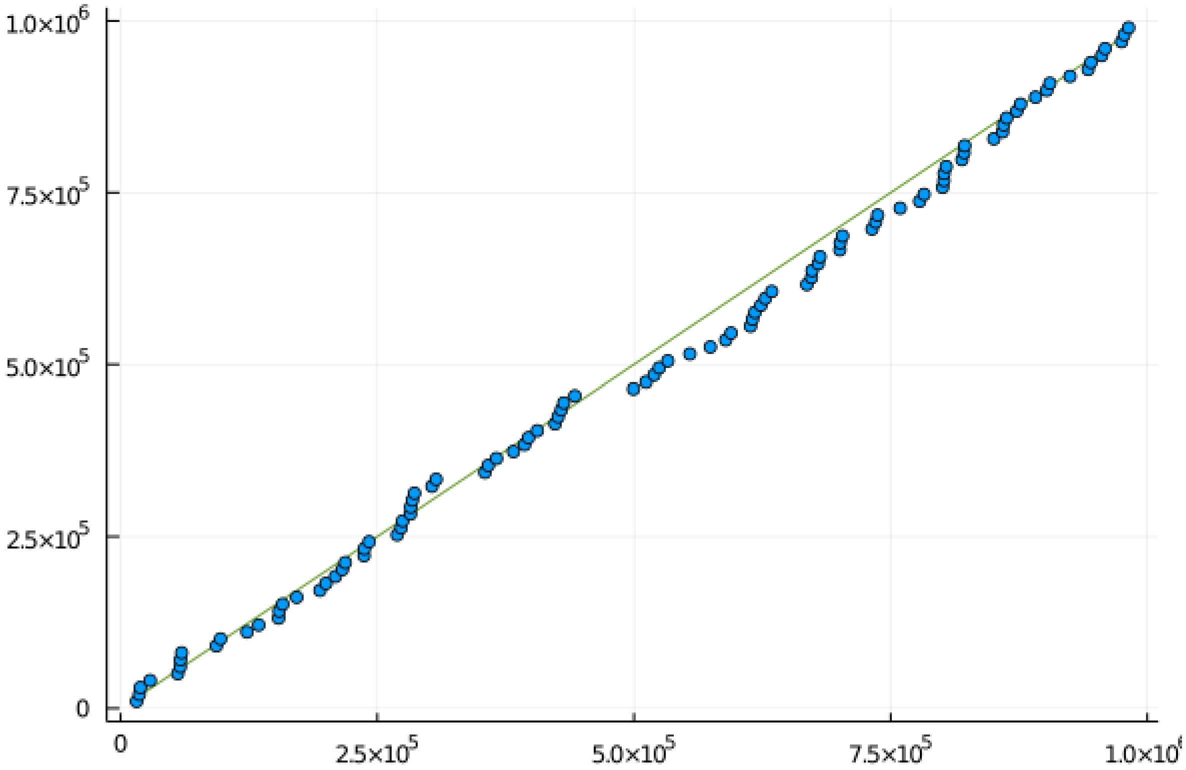
(7 pts)

_Note: I will explain this step further in subsequent classes.__

Use the Chi Square test or Kolmogorov Smirnov test together with `seq` to check if your implementation of reservoir sampling differs from the uniform distribution on the integers 1 to n. Describe how you designed the test, state the null hypothesis, show your calculations, and explain your conclusion.

I used a seq of 1,000,000 and passed it through my program to get a reservoir of 100. Then I compared this against a uniform distribution of 1 to 1,000,000

H0: The data will correlate with a uniform distribution



My plot shows that the data is correlated to the uniform distribution because the line seems to fit