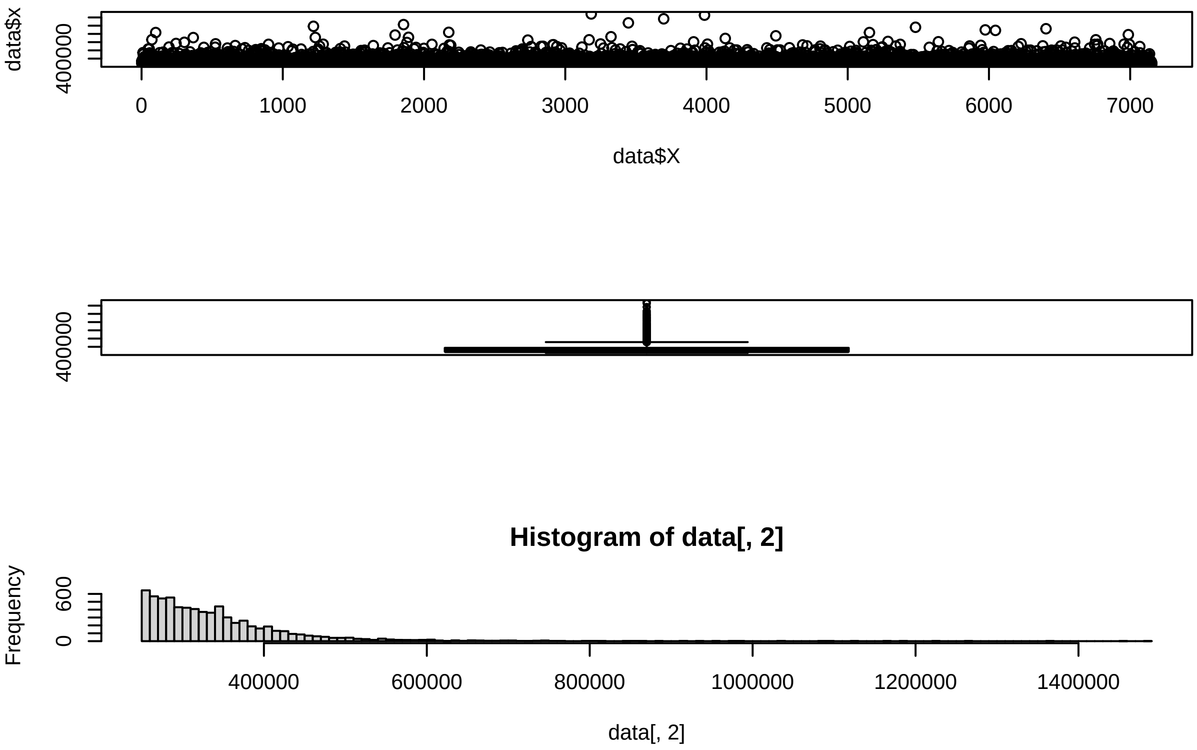
takehome

2024-11-02

Question 1

```
data <- read.csv("income20train-2.csv", header = T)
```

```
par(mfrow = c(3,1))
plot(data$X,data$x)
boxplot(data[,2])
hist(data[,2], breaks = 100)
```



(a)

```
summary(data[,2])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  250034  280433  320084  341938  373952 1484705
```

```
mean(data[,2])
```

```
## [1] 341938.1
```

Important to note that the incomes tend to be dense towards the minimum rather than being in the middle; these numbers skew heavily to the right. We can try to fit a Pareto distribution to this later to see how is models the data and run our test set against it. We have a median of 320084 and a mean of 341938.1 while he have a max of 1484705. Also note the amount of outliers in the boxplot and where these outliers are. Also note how these outliers range wider than that of the not unusual data. This shows support to the phenomenon of American income inequality (even among the 20% richest in the nation).

(b)

- consider an initial α_{b-1}
- sample β_{b-1} from $Mono(n\alpha_{b-1}+1, min(y_1, ..., y_n))$
- enter loop for B amount of times
    - sample α_b from $\Gamma(n+1, \sum_{i=1}^{n}[ln(y_i)] - nln(\beta_{b-1}))$
    - sample β_b from $Mono(n\alpha_b+1, min(y_1, ..., y_n))$
    - β_b is set to β_{b-1}
- these α's and β's are stored in a vector.

(c)