Stat 184 HW 4

Brandon Montez

Q1

(a)

```
library(synthpop)
```

```
## Find out more at https://www.synthpop.org.uk/
```

```
data(SD2011)
```

(b)

```
library(diffpriv)
SD2011 <- SD2011[!is.na(SD2011$income),]

f <- function(X) mean(X)
M <- DPMechGaussian(target = f)

bs <- function(n, var=SD2011$income){var[sample.int(n=length(var),size=n, replace=TRUE)]}
M <- sensitivitySampler(M, oracle = bs, n = nrow(SD2011), m=10000)
```

```
## Sampling sensitivity with m=10000 gamma=0.0195261707735036 k=10000
```

```
sens_f <- M@sensitivity
```

(c)

```
sd_noise <- sens_f / 0.2

r <- function(x, M, mu=0.2){f(x) + rnorm(n=1, mean=0, sd = M@sensitivity / mu)}

replicate(5, r(SD2011$income, M))
```

```
## [1] 1413.124 1370.421 1420.574 1414.744 1429.144
```

(d)

```
SD2011 <- SD2011[SD2011$income > 0,]

f_log <- function(X) mean(log(X))
M_log <- DPMechGaussian(target = f_log)
bs_log <- function(n, var=log(SD2011$income)){var[sample.int(n=length(var),size=n, replace=TRUE)]}
M_log <- sensitivitySampler(M_log, oracle = bs_log, n = nrow(SD2011), m=10000)
```

```
## Sampling sensitivity with m=10000 gamma=0.0195261707735036 k=10000
```

```
sens_f_log <- M_log@sensitivity

sd_noise_log <- sens_f_log / 0.2
r_log <- function(x, M, mu=0.2){f_log(x) + rnorm(n=1, mean=0, sd = M@sensitivity / mu)}

replicate(5, exp(r_log(SD2011$income, M_log)))
```

```
## [1] 1355.969 1354.319 1353.808 1354.068 1352.805
```

The log income model is better as it centers the data and more closely resembles a normal distribution. This makes sense as our sensitive variable is income, which typically has a skewed distribution.

Q2

(a)

```
vars <- c("sex", "age", "placesize", "region", "edu", "socprof", "unempdur", "income", "marital")
SD2011_sub <- SD2011[,vars]

SD2011_syn <- syn(SD2011_sub, m=5)
```

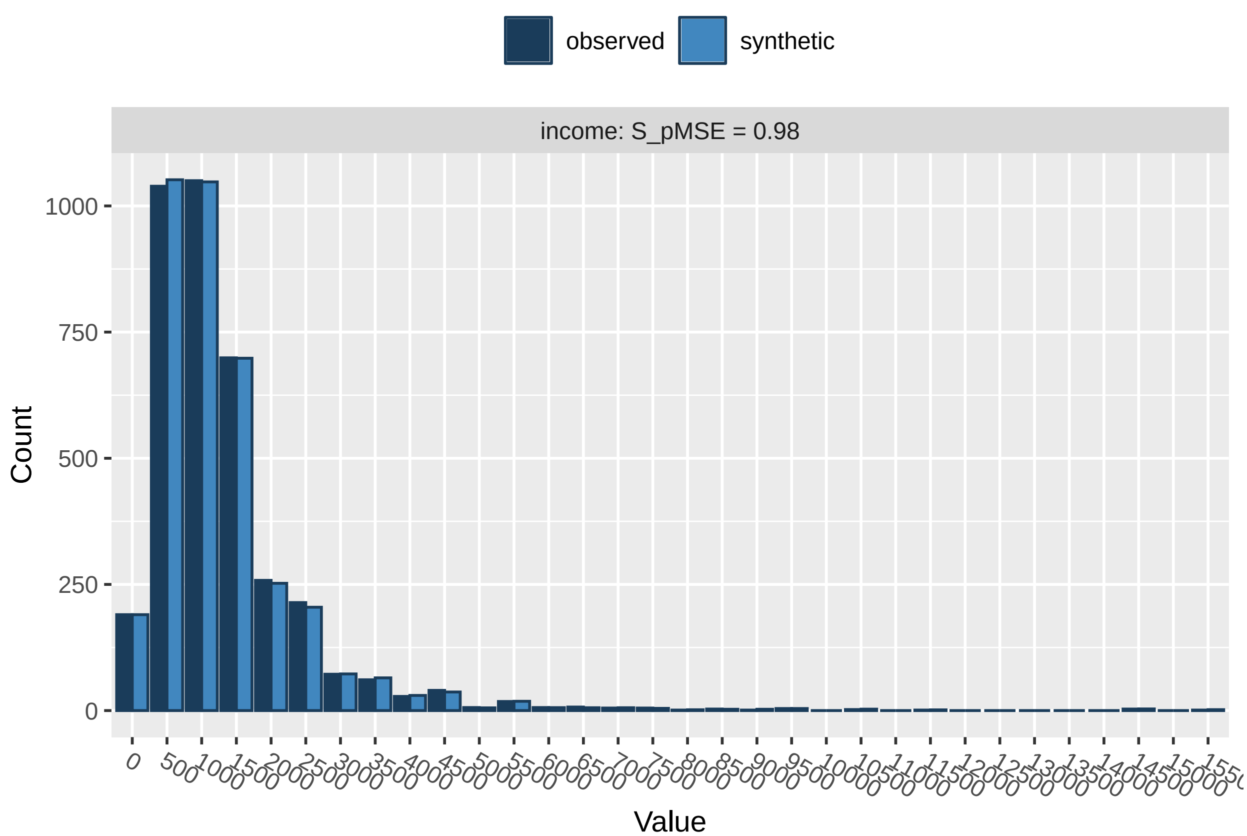
```
##
## Synthesis number 1
## --------------------
##  sex age placesize region edu socprof unempdur income marital
##
## Synthesis number 2
## --------------------
##  sex age placesize region edu socprof unempdur income marital
##
## Synthesis number 3
## --------------------
##  sex age placesize region edu socprof unempdur income marital
##
## Synthesis number 4
## --------------------
##  sex age placesize region edu socprof unempdur income marital
```

```
##
## Synthesis number 5
## --------------------
##  sex age placesize region edu socprof unempdur income marital
```

(b)

```
compare.synds(SD2011_syn, SD2011_sub, vars = "income", stat = "counts", breaks = 25, table = TRUE)
```

```
##
## Comparing counts observed with synthetic
##
## $income
##              0    500   1000  1500  2000  2500 3000 3500 4000 4500 5000 5500
## observed  190 1039.0 1050.0 699.0 258.0 214.0 72.0 61.0   28 40.0  6.0 18.0
## synthetic 190 1051.8 1047.6 698.2 252.2 204.8 72.6 64.8   30 36.8  5.4 18.4
##           6000 6500 7000 7500 8000 8500 9000 9500 10000 10500 11000 11500 12000
## observed   6.0  7.0  5.0  5.0  1.0  3.0  1.0    4     0   2.0     0   1.0     0
## synthetic  5.8  5.6  5.6  4.4  1.6  2.6  2.6    4     0   2.8     0   1.4     0
##           12500 13000 13500 14000 14500 15000 15500
## observed      0     0     0     0   3.0     0   1.0
## synthetic     0     0     0     0   3.2     0   1.8
```



```
##
## Selected utility measures:
##            pMSE  S_pMSE df
## income 6.6e-05 0.98196  4
```

The synthetic data appears to be extremely similar to the original data, given the very low S_pMSE value. This is supported by visual inspection of the histograms.