Stat 184

Brandon Montez

```r
library(laeken)
library(sdcMicro)
```

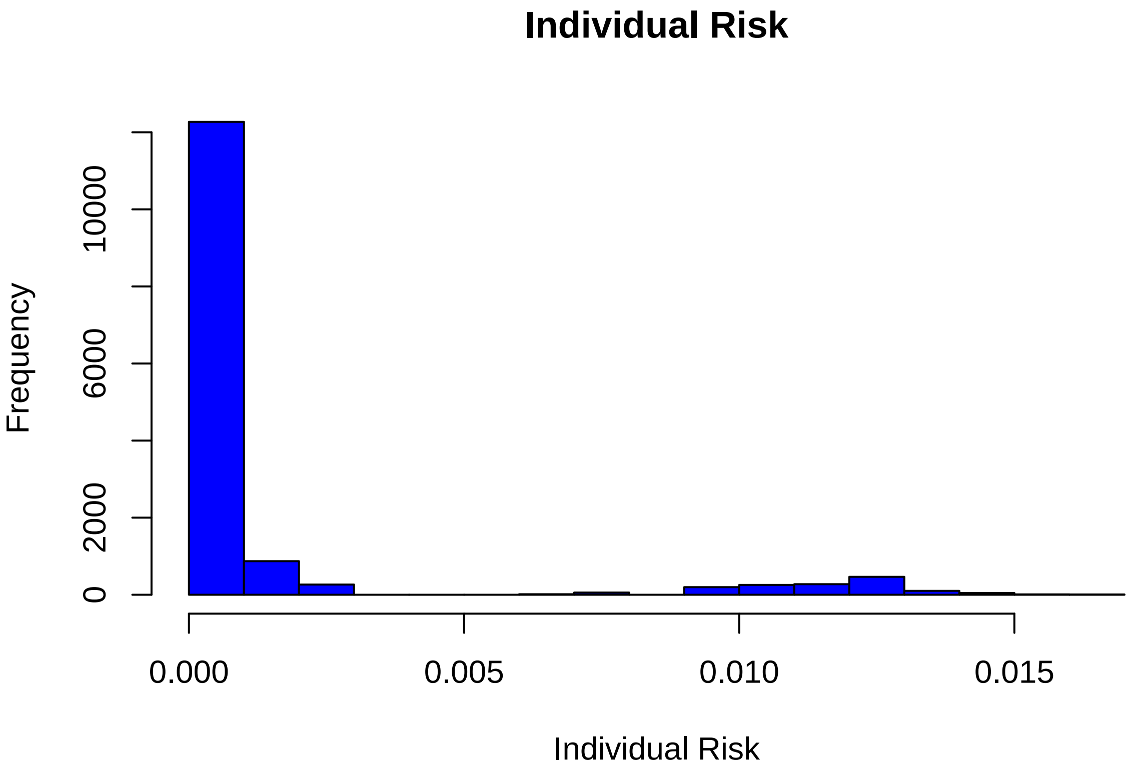
```r
require("laeken")
data("eusilc", package = "laeken")
```

(a)

```r
sdc <- createSdcObj(dat = eusilc,
  keyVars = c("age", "pb220a", "pl030", "rb090", "hsize"),
  weightVar = "rb050",
  hhId = "db030")
```

(b)

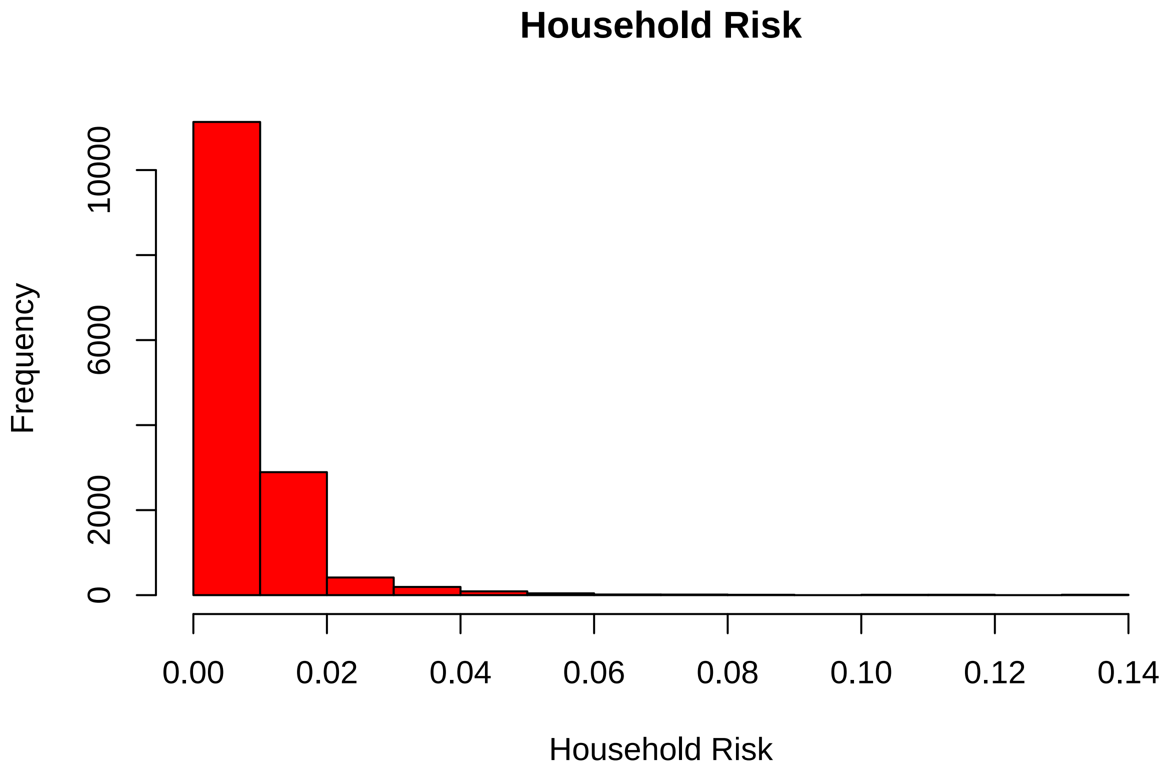
```r
risk <- get.sdcMicroObj(sdc, type="risk")$individual
hist(risk[, "risk"], main="Individual Risk", xlab="Individual Risk", col="blue")
```



It appears that most people have an individual risk that is small, but there is a cluster around 0.01 risk which is worth investigating. As there may be a high risk of re-identification.

c

```
hh_risk <- risk[, "hier_risk"]
hist(hh_risk, main="Household Risk", xlab="Household Risk", col="red")
```



If we consider 0.05 the threshold for too high of risk then the household risk evidently does exceed 0.05 for some observations. The household risk is generally below 20% but but there are some edge cases possibly suggesting non-uniformity in the distribution of key variables.

d

```
print(sdc, "risk")
```

```
## Risk measures:
##
## Number of observations with higher risk than the main part of the data: 0
## Expected number of re-identifications: 20.94 (0.14 %)
##
## Information on hierarchical risk:
## Expected number of re-identifications: 78.59 (0.53 %)
## ----------------------------------------------------------------------
```

The household risk is higher in general. We can see this from the histograms or from the information above. Given the hierarchical information, the expected number of re-identification is much higher.

e

```r
global_risk <- sum(risk[, "risk"])
global_risk
```

```
## [1] 20.93697
```

The estimated global risk using the simple summation of individual risk method says that we have around 20.94 expected individual re-identifications.

```r
set.seed(123)
subset_indices <- sample(nrow(eusilc), size = 0.1 * nrow(eusilc), replace = FALSE)
eusilc_subset <- eusilc[subset_indices, ]

sdc_subset <- createSdcObj(dat = eusilc_subset,
  keyVars = c("age", "pb220a", "pl030", "rb090", "hsize"),
  weightVar = "rb050",
  hhId = "db030")

risk_subset <- get.sdcMicroObj(sdc_subset, type="risk")$individual
global_risk_subset <- sum(risk_subset[, "risk"])

global_risk
```

```
## [1] 20.93697
```

```r
global_risk_subset
```

```
## [1] 8.066208
```

It does appear that smaller and smaller subsets reduce the global risk. That is to be expected, as the global risk is a proportion of the number of individuals in the dataset. If we only have 1 individual, then our expected number of identifications could be at most 1.

```r
print(sdc_subset, "risk")
```

```
## Risk measures:
##
## Number of observations with higher risk than the main part of the data: 0
## Expected number of re-identifications: 8.07 (0.54 %)
##
## Information on hierarchical risk:
## Expected number of re-identifications: 9.98 (0.67 %)
## ----------------------------------------------------------------------
```

This confirms our observation that the reduced subset has a lower global risk. The subset has redcued risk in both individual and household risk.