

# STAT847 Lab 5

Anh Pham - ewa589

October 5, 2023

## Contents

- 1) Load in the “agpop.csv” file to an object called ‘ag.data’

```
ag.data <- read.csv('D:agpop-1.csv')
head(ag.data)
```

```
##               county state acres92 acres87 acres82 farms92 farms87 farms82
## 1 ALEUTIAN ISLANDS AREA   AK  683533  726596  764514      26      27      28
## 2      ANCHORAGE AREA    AK   47146   59297  256709     217     245     223
## 3      FAIRBANKS AREA    AK  141338  154913  204568     168     175     170
## 4          JUNEAU AREA    AK    210    214    127      8      8      12
## 5 KENAI PENINSULA AREA    AK   50810   85712   98035     93     119     137
## 6      AUTAUGA COUNTY    AL  107259  116050  145044     322     388     453
##  largef92 largef87 largef82 smallf92 smallf87 smallf82 region
## 1      14      16      20        6        4        1      W
## 2       9      10      11       41       52       38      W
## 3      25      28      21       12       18       25      W
## 4       0       0       0        5        4        8      W
## 5       9      18      17       12       18       19      W
## 6      25      32      32        8       19       17      S
```

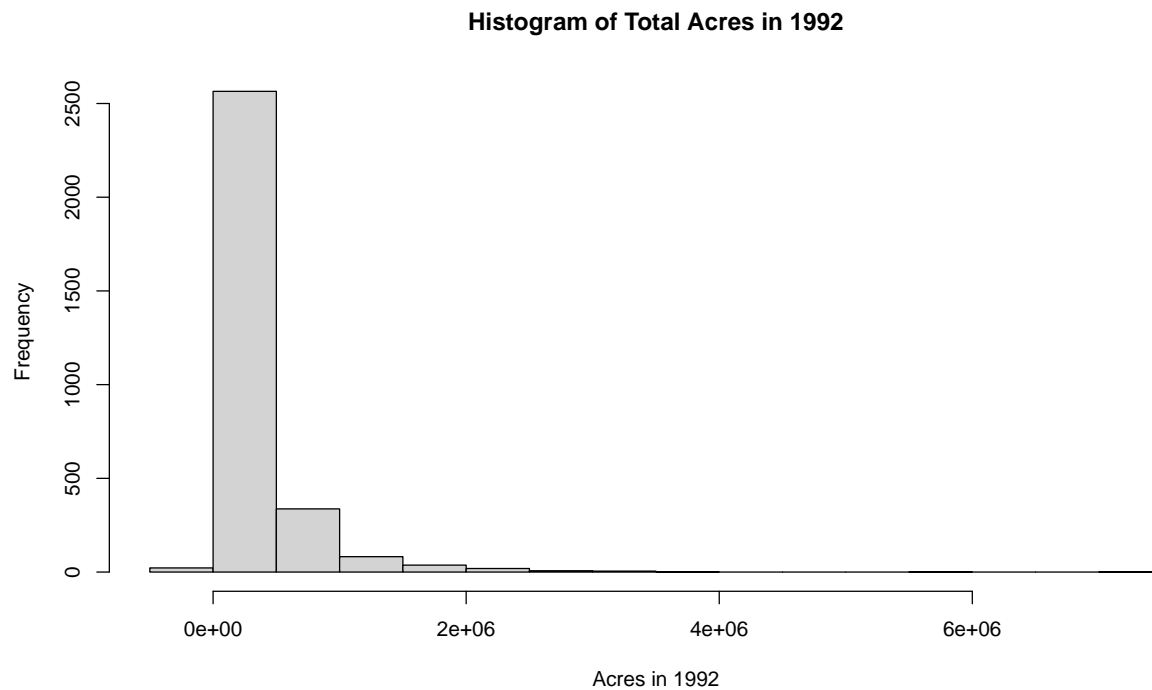
- 2) Use the `summary()` function to print out one summary of `acres92` and `acres87`

```
summary(ag.data[c('acres92', 'acres87')])
```

```
##      acres92      acres87
## Min.   :    -99  Min.   :    -99
## 1st Qu.:  80903  1st Qu.:  86236
## Median : 191648  Median : 199864
## Mean   : 306677  Mean   : 313016
## 3rd Qu.: 366886  3rd Qu.: 372224
## Max.   :7229585  Max.   :7687460
```

- 3) Print out a histogram of `acres92` using the base R `hist()` function. Change the title of the histogram to “Histogram of Total Acres in 1992”. Change the x-axis label to “Acres in 1992”

```
hist(ag.data$acres92, main = "Histogram of Total Acres in 1992", xlab = "Acres in 1992")
```



4) Use the `sample()` function to create sampling indexes that split the ‘ag.data’ into 80% for training and 20% for testing. Then use these sample indexes to split the ‘ag.data’ into training and testing datasets called “train.data” and “test.data”.

```
set.seed(1)
dt= sort(sample(nrow(ag.data), nrow(ag.data)*.8))
train.data <- ag.data[dt,]
test.data <- ag.data[-dt,]
dim(train.data)
```

```
## [1] 2462 15
```

```
dim(test.data)
```

```
## [1] 616 15
```