

# Muestreo con probabilidades proporcionales

Agrupamos la base de datos por hogares y resumimos la información total por hogar

```
library(TeachingSampling)
library(dplyr)
data("BigCity")
Hogares <- BigCity %>% group_by(HHID) %>%
  summarise(Ingreso = sum(Income),
            Gasto = sum(Expenditure),
            EdadMedia = mean(Age),
            Personas = n())
head(Hogares)

## # A tibble: 6 x 5
##   HHID     Ingreso   Gasto   EdadMedia Personas
##   <chr>     <dbl>   <dbl>      <dbl>     <int>
## 1 idHH00001  2775.  2442.      27        5
## 2 idHH00002  1492.  1084.     19.4       5
## 3 idHH00003  4280.  2441.     38.2       4
## 4 idHH00004  2200.  1851.     29.8       4
## 5 idHH00005  3119.  3068.     32.4       5
## 6 idHH00006   675.  1098.     24.8       5
```

## Diseño de muestreo poisson

creamos las probabilidades proporcionales al número de personas por hogar

```
attach(Hogares)
N <- dim(Hogares)[1]
n <- 2000
pik <- n * Personas / sum(Personas)
which(pik > 1)

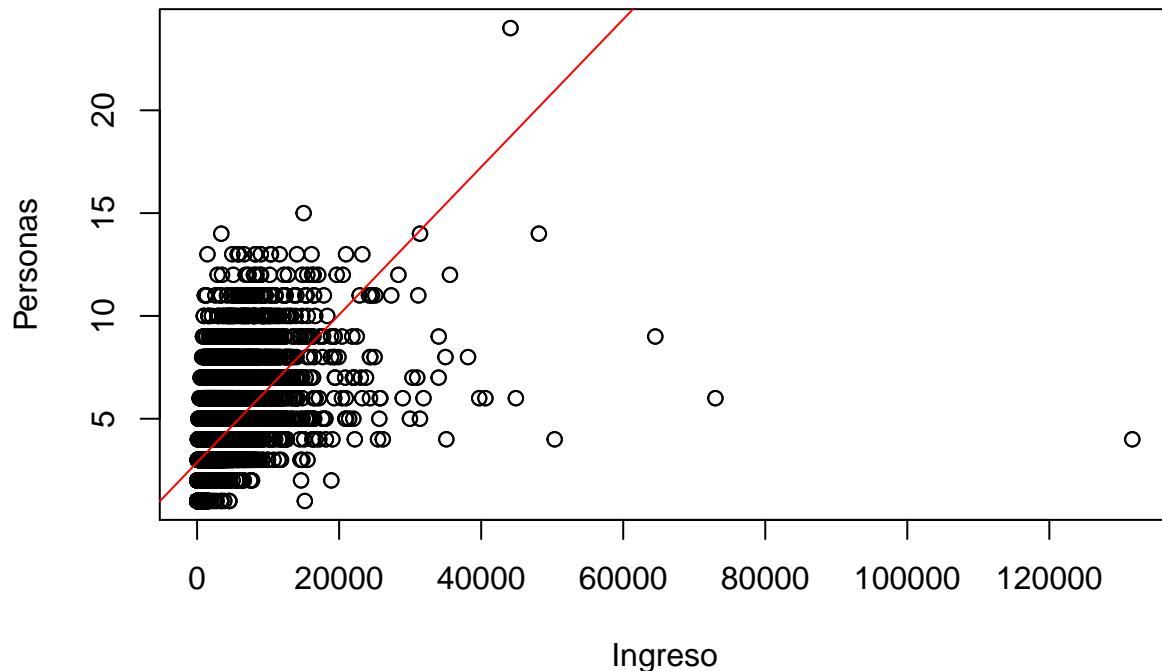
## integer(0)
sum(pik)

## [1] 2000

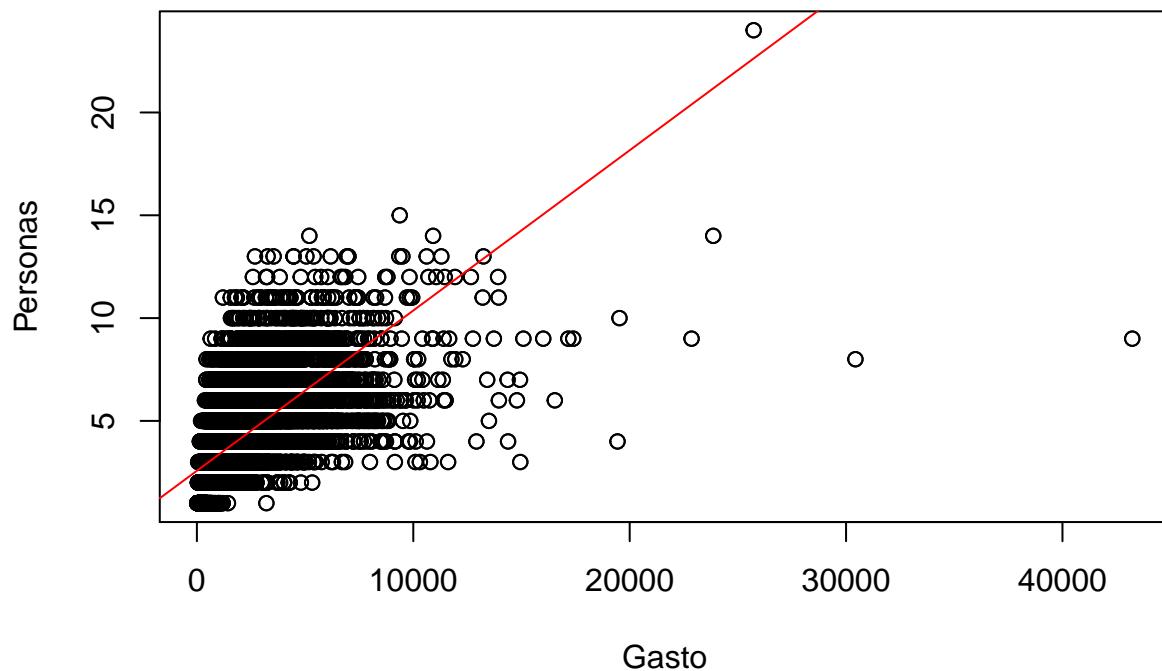
Observemos la matriz de correlación de las variables
matriz <- cbind(pik, Ingreso, Gasto, Personas)
cor(matriz)

##          pik    Ingreso    Gasto Personas
## pik 1.0000000 0.5637840 0.6436065 1.0000000
## Ingreso 0.5637840 1.0000000 0.7413683 0.5637840
## Gasto 0.6436065 0.7413683 1.0000000 0.6436065
## Personas 1.0000000 0.5637840 0.6436065 1.0000000
```

```
plot(Personas ~ Ingreso)
abline(lm(Personas ~ Ingreso), col=2)
```



```
plot(Personas ~ Gasto)
abline(lm(Personas ~ Gasto), col=2)
```



Con las probabilidades obtenemos la muestra

```
sam <- S.PO(N, pik)
muestra <- Hogares[sam,]
n.s <- dim(muestra)[1]
n.s
```

```

## [1] 1945
attach(muestra)
head(muestra)

## # A tibble: 6 x 5
##   HHID      Ingreso  Gasto EdadMedia Personas
##   <chr>     <dbl>   <dbl>     <dbl>     <int>
## 1 idHH00026    402    223.      17         3
## 2 idHH00034    611.    659.     27.8       5
## 3 idHH00041   3295    1569.    22.5       4
## 4 idHH00064    677.    419.     54.4       5
## 5 idHH00075   1321.    858.     25.5       2
## 6 idHH00089   991.    317.     51.7       3

```

Con esta muestra calculamos las estimaciones correspondientes

```

pik.s <- pik[sam]
estima <- data.frame(Ingreso, Gasto, Personas)
E.PO(estima, pik.s)

##           N      Ingreso      Gasto Personas
## Estimation 39727.003457 8.554979e+07 5.584174e+07 1.461337e+05
## Standard Error 1041.212030 2.504686e+06 1.516966e+06 3.209788e+03
## CVE          2.620918 2.927753e+00 2.716546e+00 2.196474e+00
## DEFF          Inf 6.811473e-01 7.824254e-01 3.244628e+00

```

## Diseño PPT (con reemplazo)

Veamos las correlaciones respectivas mencionadas en el modulo

```

attach(Hogares)
N <- nrow(Hogares)
m <- 2000
(N^2 / m) * cov(Personas, (Ingreso^2 / Personas))

## [1] 3.443098e+12
(N^2 / m) * cov(Personas, (Gasto^2 / Personas))

## [1] 1.100114e+12
cor(Personas, (Ingreso^2 / Personas))

## [1] 0.06754802
cor(Personas, (Gasto^2 / Personas))

## [1] 0.2574057

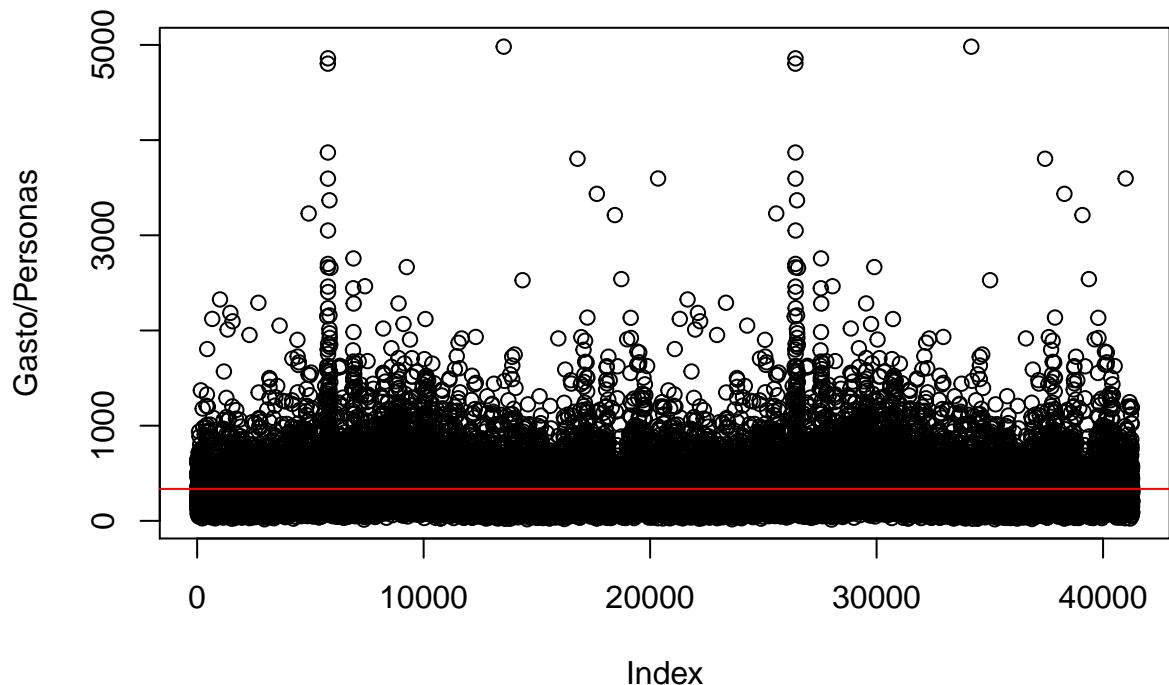
```

Algunas evidencias gráficas

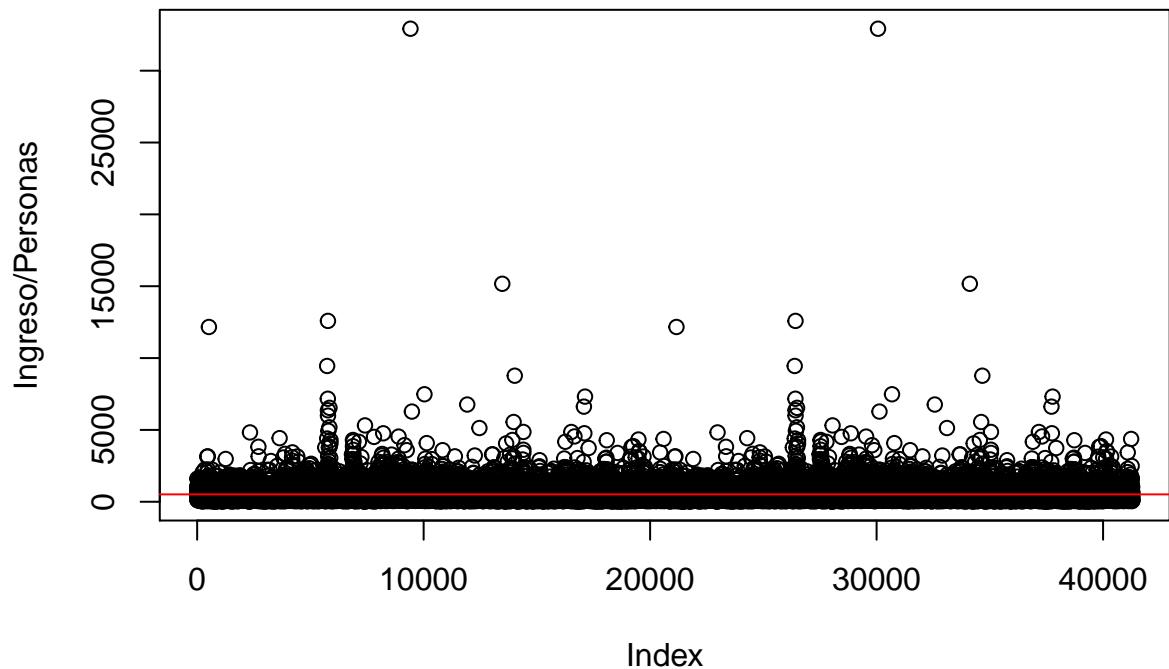
```

plot(Gasto/Personas)
abline(h = mean(Gasto/Personas), col = 2)

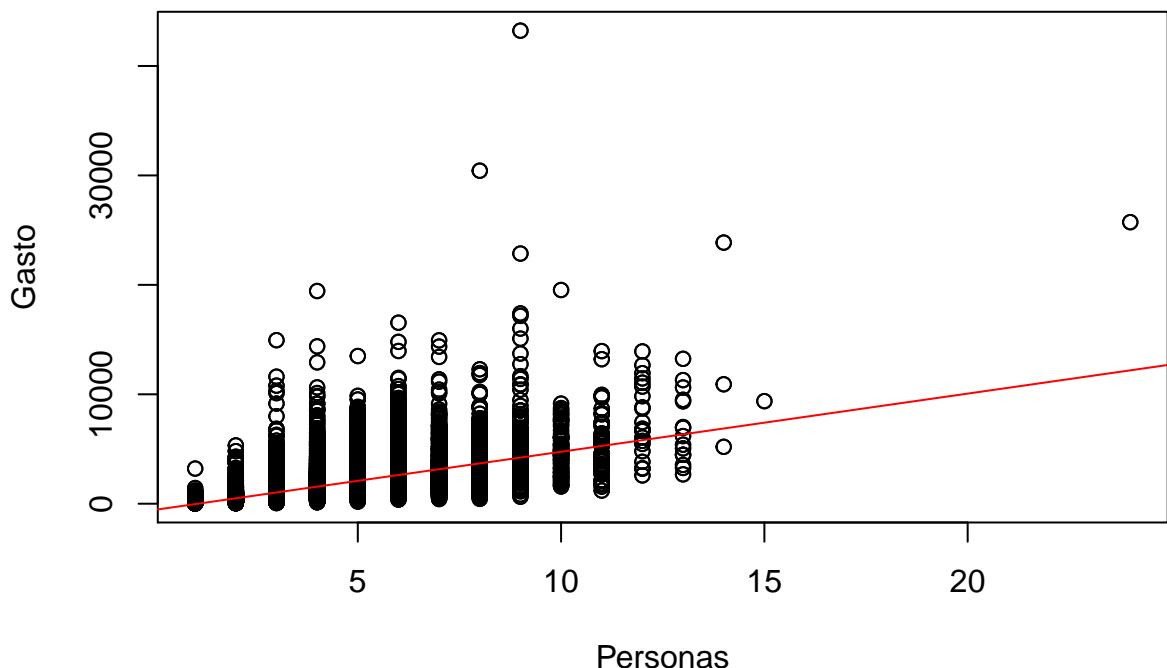
```



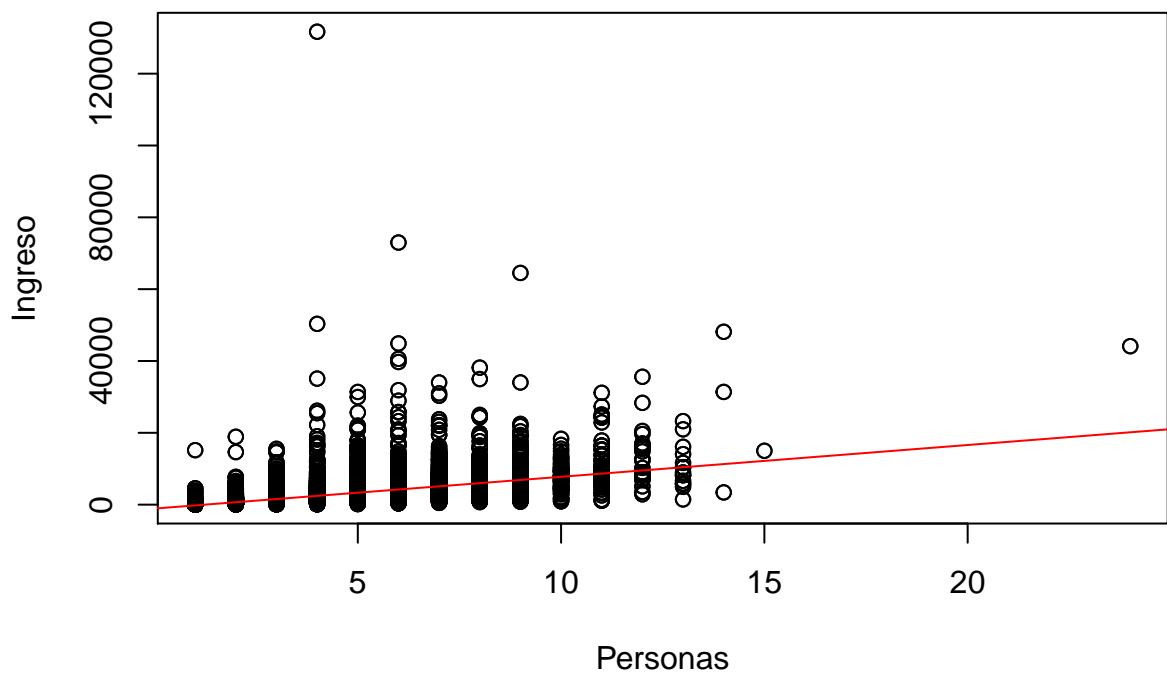
```
plot(Ingreso/Personas)
abline(h = mean(Ingreso/Personas), col = 2)
```



```
plot(Gasto ~ Personas)
abline(lm(Gasto ~ Personas), col=2)
```



```
plot(Ingreso ~ Personas)
abline(lm(Ingreso ~ Personas), col=2)
```



```
M.I <- lm(Gasto ~ Personas)
summary(M.I)
```

```
##
## Call:
## lm(formula = Gasto ~ Personas)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -300000 -100000 -50000  100000  300000
```

```

## -4076   -557   -135    245  39010
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -566.95     12.70  -44.64  <2e-16 ***
## Personas      531.47      3.11  170.87  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1170 on 41288 degrees of freedom
## Multiple R-squared:  0.4142, Adjusted R-squared:  0.4142
## F-statistic: 2.92e+04 on 1 and 41288 DF, p-value: < 2.2e-16
M.E <- lm(Ingreso ~ Personas)
summary(M.E)

```

```

##
## Call:
## lm(formula = Ingreso ~ Personas)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -8955 -1007  -239     451 129232
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1091.65     26.05  -41.91  <2e-16 ***
## Personas      884.88      6.38  138.70  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2400 on 41288 degrees of freedom
## Multiple R-squared:  0.3179, Adjusted R-squared:  0.3178
## F-statistic: 1.924e+04 on 1 and 41288 DF, p-value: < 2.2e-16

```

A continuación definimos las probabilidades para obtener la muestra respectiva

```

pk <- Personas / sum(Personas)
sam <- S.PPS(m, Personas)
muestra <- Hogares[sam,]
attach(muestra)
head(muestra)

```

```

## # A tibble: 6 x 5
##   HHID      Ingreso Gasto EdadMedia Personas
##   <chr>     <dbl>  <dbl>     <dbl>    <int>
## 1 idHH20553  1700   641.     12.4      5
## 2 idHH33660   624    554      14.8      4
## 3 idHH41196  1200   556      35.5      2
## 4 idHH12314  1806.  1997.    16.4      5
## 5 idHH08703  1794.  1555.    25        2
## 6 idHH14071  2657.  2440.    22.8      5

```

Con esta muestra obtenemos las estimaciones respectivas

```

pk.s <- pk[sam]
estima <- data.frame(Ingreso, Gasto, Personas)

```

```

E.PPS(estima, pk.s)

##           N      Ingreso      Gasto Personas
## Estimation 41502.486983 8.762016e+07 5.530696e+07 150266
## Standard Error 552.533535 1.683790e+06 9.189094e+05 0
## CVE          1.331326 1.921692e+00 1.661472e+00 0
## DEFF          Inf 2.880118e-01 3.352254e-01 0

length(pk.s)

## [1] 2000

```

## Diseño de muestreo piPT sin reemplazo

Obtenemos una muestra sin reemplazo con probabilidades proporcionales y estimamos las variables respectivas

```

attach(Hogares)
N <- dim(Hogares)[1]
n <- 2000
res <- S.piPS(n, Personas)
sam <- res[,1]
muestra <- Hogares[sam,]
attach(muestra)

pik.s <- res[, 2]
estima <- data.frame(Ingresa, Gasto, Personas)
E.piPS(estima, pik.s)

##           N      Ingreso      Gasto Personas
## Estimation 41208.647441 8.802506e+07 5.772221e+07 1.502660e+05
## Standard Error 545.478659 1.706839e+06 9.469803e+05 2.709153e-13
## CVE          1.323699 1.939037e+00 1.640582e+00 1.802905e-16
## DEFF          Inf 2.172386e-01 2.297960e-01 1.816374e-32

```