Diseños de muestreo unietápicos

Tarea Examen 2. Tipo A

Profesor. Gonzalo Pérez.
Ayudante. Juan Andrés Cervantes.
Ayudante. José Angel Román.

	Franco Zarraga Daniel
Integrantes:	Gomez Jimenez Aaron Mauricio
	Sandoval Mendoza Jorge

1. **Muestreo sistemático**

a) Considere el estimador HT del total y su varianza dada en la expresión (104) de las notas. Demuestre que una expresión equivalente de la varianza es:

$$\mathrm{V}(\widehat{t}_{y\pi}) = \frac{1}{2}\sum_{r=1}^{m}\sum_{r'=1}^{m}\left(t_{s_r} - t_{s_{r'}}\right)^2.$$

Solución.

Tenemos que:

$$t_{s_r} = \sum_{k\in S_r} y_k \qquad y \qquad \bar{t} = \sum_{r=1}^{m}\frac{t_{s_r}}{m}$$

Empezaremos a desarrollar la expresión (104)

$$\begin{aligned}\mathrm{V}(\widehat{t}_{y\pi}) &= m\sum_{r=1}^{m}(t_{s_r} - \bar{t})^2 \\ &= \frac{1}{2}(2m)\sum_{r=1}^{m}(t_{s_r} - \bar{t})^2\end{aligned}$$

Notemos que lo anterior puede ser separado en 2 sumas, en las que sin pérdida de generalidad, la primera correrá sobre un indice r y la segunda sobre un índice r'

$$= \frac{1}{2}\left[m\sum_{r=1}^{m}(t_{s_r} - \bar{t})^2 + m\sum_{r'=1}^{m}\left(t_{s_{r'}} - \bar{t}\right)^2\right] \ldots(1)$$

Se puede reescribir el primer sumando como sigue:

$$m\sum_{r=1}^{m}(t_{s_r} - \bar{t})^2 = \sum_{r'=1}^{m}\sum_{r=1}^{m}(t_{s_r} - \bar{t})^2$$

Pues al sumar sobre r' desde 1 hasta m, una expresión que no está en términos de r' y por lo tanto no depende de ésta, en realidad se está sumando m veces lo mismo, de manera análoga sucede lo mismo con la otra suma:

$$m\sum_{r'=1}^{m}\left(t_{s_{r'}}-\bar{t}\right)^2=\sum_{r=1}^{m}\sum_{r'=1}^{m}\left(t_{s_{r'}}-\bar{t}\right)^2$$

De esta manera la expresión (1) queda como:

$$\frac{1}{2}\left[\sum_{r'=1}^{m}\sum_{r=1}^{m}\left(t_{s_r}-\bar{t}\right)^2+\sum_{r=1}^{m}\sum_{r'=1}^{m}\left(t_{s_{r'}}-\bar{t}\right)^2\right]$$

$$=\frac{1}{2}\left[\sum_{r=1}^{m}\sum_{r'=1}^{m}\left(t_{s_r}-\bar{t}\right)^2+\sum_{r=1}^{m}\sum_{r'=1}^{m}\left(t_{s_{r'}}-\bar{t}\right)^2\right]\ldots(2)$$

Mostraremos que: $\sum_{r=1}^{m}\sum_{r'=1}^{m}\left(t_{s_r}-\bar{t}\right)\left(t_{s_{r'}}-\bar{t}\right)=0$ ya que este resultado permitirá resolver el problema.

$$\begin{aligned}
\sum_{r=1}^{m}\sum_{r'=1}^{m}\left(t_{s_r}-\bar{t}\right)\left(t_{s_{r'}}-\bar{t}\right)&=\sum_{r=1}^{m}\left(t_{s_r}-\bar{t}\right)\sum_{r'=1}^{m}\left(t_{s_{r'}}-\bar{t}\right)\\
&=\left[\sum_{r=1}^{m}t_{s_r}-\sum_{r=1}^{m}\bar{t}\right]\left[\sum_{r'=1}^{m}t_{s_{r'}}-\sum_{r'=1}^{m}\bar{t}\right]\\
&=\left[\sum_{r=1}^{m}t_{s_r}-m\bar{t}\right]\left[\sum_{r'=1}^{m}t_{s_{r'}}-m\bar{t}\right]\\
&=\left[\sum_{r=1}^{m}t_{s_r}-m\cdot\sum_{r=1}^{m}\frac{t_{sr}}{m}\right]\left[\sum_{r'=1}^{m}t_{s_{r'}}-m\sum_{r'=1}^{m}\frac{t_{s_{r'}}}{m}\right]\\
&=\left[\sum_{r=1}^{m}t_{s_r}-\sum_{r=1}^{m}t_{s_r}\right]\left[\sum_{r'=1}^{m}t_{s_{r'}}-\sum_{r'=1}^{m}t_{s_{r'}}\right]\\
&=(0)(0)=0
\end{aligned}$$

Con lo anterior, sumamos un 0 oportuno a la expresión (2)

$$=\frac{1}{2}\left[\sum_{r=1}^{m}\sum_{r'=1}^{m}\left(t_{s_r}-\bar{t}\right)^2-2\sum_{r=1}^{m}\sum_{r'=1}^{m}\left(t_{s_r}-\bar{t}\right)\left(t_{s_{r'}}-\bar{t}\right)+\sum_{r=1}^{m}\sum_{r'=1}^{m}\left(t_{s_{r'}}-\bar{t}\right)^2\right]$$

$$=\frac{1}{2}\sum_{r=1}^{m}\sum_{r'=1}^{m}\left[\left(t_{s_r}-\bar{t}\right)^2-2\left(t_{s_r}-\bar{t}\right)\left(t_{s_{r'}}-\bar{t}\right)+\left(t_{s_{r'}}-\bar{t}\right)^2\right]$$

La expresión dentro de la doble suma es un trinomio cuadrado perfecto, lo factorizamos:

$$=\frac{1}{2}\sum_{r=1}^{m}\sum_{r'=1}^{m}\left[\left(t_{s_r}-\bar{t}\right)-\left(t_{s_{r'}}-\bar{t}\right)\right]^2$$

$$= \frac{1}{2}\sum_{r=1}^{m}\sum_{r'=1}^{m}\left[t_{s_r} - \bar{t} - t_{s_{r'}} + \bar{t}\right]^2$$

$$= \frac{1}{2}\sum_{r=1}^{m}\sum_{r'=1}^{m}\left[t_{s_r} - t_{s_{r'}}\right]^2$$

$$\therefore \quad \mathrm{V}(\widehat{t}_{y\pi}) = \frac{1}{2}\sum_{r=1}^{m}\sum_{r'=1}^{m}\left(t_{s_r} - t_{s_{r'}}\right)^2$$

b) Suponga que $N = 100$ y los valores de $y_1, ..., y_N$ son {1,2,3,...,9,10,1,2,3,...,9,10,...,1,2,3,...,9,10}, es decir, los valores del 1 al 10 se repiten 10 veces. Dé un ordenamiento tal que al usar un muestreo sistemático con $n = 20$, el estimador de HT del total tenga un ECM igual a cero.

Solución.

Primero tenemos que el valor real de la población es $\theta = 10\sum_{j=1}^{10} j = 10 \cdot \frac{10\cdot 11}{2} = 550$

Para la probabilidad de inclusión tenemos que $m = \frac{N}{n} = \frac{100}{20} = 5$

De este modo tendríamos sólo 5 muestras posibles y equiprobables.

$$S_0 = \{s_1, s_2, s_3, s_4, s_5\}$$

La probabilidad de cada muestra es $\frac{1}{m} = \frac{1}{5}$

Ahora para el error cuadrático medio tenemos:

$$ECM(\widehat{t}_{\pi y}) = \mathbb{E}\left[(\widehat{t}_{\pi y} - \theta)^2\right] = \sum_{c_i \in sop}(c_i - \theta)^2 \mathbb{P}\left(\widehat{t}_{\pi y} = c_i\right)$$

$$= \frac{1}{m}\sum_{c_i \in sop}(c_i - \theta)^2 = \frac{1}{m}\sum_{c_i \in sop}\left(m\sum_{k\in s} y_k - \theta\right)^2$$

$$= \frac{1}{5}\sum_{c_i \in sop}\left(5\sum_{k\in s} y_k - 550\right)^2$$

$$ECM(\widehat{t}_{\pi y}) = 0 \iff \frac{1}{5}\sum_{c_i \in sop}\left(5\sum_{k\in s} y_k - 550\right)^2 = 0$$

$$\iff \left(5\sum_{k\in s} y_k - 550\right)^2 = 0 \qquad \text{para toda muestra}$$

$$\iff \sum_{k\in s} y_k = 110$$

Sabemos que la suma de losprimeros 10 números naturales positivos es 55, por lo tanto una propuesta de ordenamiento sería tener una muestra con los primeros 10 números naturales dos veces, i.e.

$$s_j = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10\} \quad \text{con } j = \{1, 2, 3, 4, 5\}$$

Como nuestros "saltos"son de magnitud 5 (r=5), una propuesta de ordenamiento sería la siguiente:

$$\{1,1,1,1,1,2,2,2,2,2,...,10,10,10,10,10,1,1,1,1,1,2,2,2,2,2,...,10,10,10,10,10\}$$

Se repite 5 veces el 1, luego 5 veces el 2 y así hasta repetir 5 veces el 10, despúes repetimos el proceso del 1 al 10.Teniendo las siguientes 5 muestras iguales.

s_1	s_2	s_3	s_4	s_5
1	1	1	1	1
2	2	2	2	2
3	3	3	3	3
4	4	4	4	4
5	5	5	5	5
6	6	6	6	6
7	7	7	7	7
8	8	8	8	8
9	9	9	9	9
10	10	10	10	10
1	1	1	1	1
2	2	2	2	2
3	3	3	3	3
4	4	4	4	4
5	5	5	5	5
6	6	6	6	6
7	7	7	7	7
8	8	8	8	8
9	9	9	9	9
10	10	10	10	10

Cabe aclarar que para que el error cuadrático medio sea cero, el ordenamiento anterior no es el único posible, sin embrago las muestras si deben ser como estan en la tabla. Para obtener dichas muestras se podría poner primero los 10 unos y así con cada número hasta llegar a los 10 dieces.

2. **Selección con probabilidades diferentes y sin reemplazo**

Sean $p_1, ..., p_N$ son un conjunto de valores tal que $0 < p_k \leq 1 \; \forall \; k \in U$ y $\sum_{k=1}^{N} p_k = 1$. Considere el siguiente esquema de selección para obtener una muestra de tamaño $n = 2$.

a) Del conjunto de N unidades se seleciona de forma aleatoria una unidad, en esta primera selección la unidad k tiene probabilidad p_k de ser seleccionada.

b) Una vez seleccionada una unidad, digamos que fue la i, se selecciona otra unidad del conjunto de las $N-1$ unidades restantes (sin contar la i). En esta selección la unidad $k,\ k \neq i$, tiene probabilidad $p_k(1-p_i)^{-1}$ de ser seleccionada.

Indique cuáles son las probabilidades de primer y segundo orden si se usa un diseño de muestreo con este esquema de selección.

Solución.

Como por hipotesis sabemos que la probabilidad de que el elemento k sea seleccionado es p_k, notemos que las p_k no son equiprobables, es decir la probabilidad no es uniforme para todas las k, ya que cada k tiene proba distinta de ser seleccionada. Sea k arbitraria en U y sea S_1 la muestra de tamaño $n = 1$ así tenemos que

$$P(k \in S_1) = P(I_k = 1) = \sum_{s \in S_1} I(k \in S_1)(p(s)) = I_k(p(s)) = p_k$$

Por lo tanto tenemos que

$$\pi_k = \sum_{s:k \in S_1} P(s) = \sum_{s:k \in S_1} p_k$$

Ahora para la probabilidad de que el elemento k sea seleccionado dado que ya seleccionamos la unidad i tenemos que dado que ya sacamos la unidad con proba p_i tomando el diseño de probabalidad clásica tendriamos que restar el valor de esa proba del total posible, es decir tendriamos que si sea S_2 el evento con muestra de tamaño $n = 2$ y la probabilidad de que sea seleccionada la unidad k dado que ya se seleccionó la unidad i

$$P(k \in S_2 \mid i \in S_1) = \frac{p_k}{1-p_i}$$

De esta manera vemos que la probabilidad de que i sea seleccionado y despues se seleccione a k es equivalente a que

$$P(i \in S_1, k \in S_2) = P(i \in S_1)P(k \in S_2 \mid i \in S_1) = (p_i)(\frac{p_k}{1-p_i})$$

Como buscamos la probabilidad de inclusión de segundo orden π_{ik} es decir la probabilidad de que en la muestra salga k e i, pero debemos observar que no es lo mismo que primero salga k y luego i, a que primero salga i y despues k, entonces para encontrar la probabilidad de inclusión sumamos estas probabilidades

$$P(k \in S_1, i \in S_2) = P(k \in S_1)P(i \in S_2 \mid k \in S_1) = (p_k)(\frac{p_i}{1-p_k})$$

Así, obtenemos que

$$\pi_{ki} = (p_i)(\frac{p_k}{1-p_i}) + (p_k)(\frac{p_i}{1-p_k})$$

3. **Muestreo Poisson**

Considerando un diseño de muestreo Poisson de tamaño esperado 10, se seleccionó una muestra de una población de tamaño $N = 284$ con el objetivo de estimar el total de la variable $REV84$. Las probabilidades de inclusión usadas en la selección de la muestra fueron proporcionales al tamaño considerando la variable $P75$ (el total de la variable $P75$ en la población es 8,182). La muestra seleccionada fue de tamaño $n_s = 12$ y la información recolectada de la variable $REV84$ en la muestra es la siguiente:

$REV84$	5246	59877	2208	2546	2903	6850	3773	4055	4014	38945	1162	4852
$P75$	54	671	28	27	29	62	42	48	33	446	12	46

Datos de las variables $REV84$ y $P75$ en las unidades seleccionadas

a) Dé un intervalo de confianza al 95 % para el total de la variable $REV84$ usando las fórmulas vistas en clase.

Solución.

Primero determinemos el valor del total de la variable $REV84$ usando un diseño de muestreo Poisson.

Tomamos a $x = P75$ como la variable tamaño de este diseño, ya que se encuentra relacionada con la variable $REV84$.

El valor esperado del tamaño de muestra deseado es $n = 10$. y el total de la variable tamaño es $t_x = 8182$.

Así las probabilidades de inclusión están dadas por:

$$\begin{aligned}\pi_k &= \text{mín}\left(\frac{n \cdot x_k}{t_x}, 1\right) \\ &= \text{mín}\left(\frac{10 \cdot x_k}{8182}, 1\right)\end{aligned}$$

Donde las x_k son los valores muestrales de la variable $P75$.

```
# valores muestrales de las variables estudiadas
rev_84 <-c(5246, 59877, 2208, 2546, 2903, 6850,
3773, 4055, 4014, 38945, 1162, 4852)

p75 <-c(54, 671, 28, 27, 29, 62, 42, 48, 33, 446, 12, 46)

# almacenmos todo en un data frame
muestra <- data.frame(rev_84, p75)
```

```
n <- 10
N <- 284
total_p75 <- 8182

# probabilidades de inclusion de primer orden
pi_k <- (n/total_p75)*p75
pi_k
```

```
##  [1] 0.06599853 0.82009289 0.03422146 0.03299927 0.03544366 0.07577609
##  [7] 0.05133219 0.05866536 0.04033244 0.54509900 0.01466634 0.05622097
```

Se observan los valores de las probabilidades de inclusión, a continuación se calcula el valor estimado del total de $REV84$, con:

$$\hat{t}_{y\pi} = \sum_{k \in S} \frac{y_k}{\pi_k}$$

```
# estimador del total

t<- sum(rev_84/pi_k)
t
```

```
## [1] 945598.2
```

$$\hat{t}_{y\pi} = 945598.2$$

Posteriormente calculamos el valor del total ajustado para comparar con el valor obtenido previamente:

$$\hat{t}_{alt} = N \frac{\hat{t}_{y\pi}}{\hat{N}_{\pi}}$$

Donde:

$$\hat{N}_{\pi} = \sum_{k \in S} \frac{1}{\pi_k}$$

```
# estimador del tamaÃ±o poblacional
N_pi <- sum(1/pi_k)
N_pi
```

```
## [1] 266.4325
```

```r
# estimador alternativo del total
t_a <- N*t/N_pi
t_a
```

```
## [1] 1007947
```

$$\hat{t}_{alt} = 1,007,947$$

Procedemos a calcular el estimador de la varianza de $\hat{t}_{\pi y}$

$$\hat{V}(\hat{t}_{\pi y}) = \sum_{k \in S} \frac{y_k^2}{\pi_k^2}(1 - \pi_k)$$

```r
# estimador de la varianza

s<-1-pi_k

var<-sum((rev_84/pi_k)^2*s)
var
```

```
## [1] 65324477213
```

El intervalo de confianza al 95 % está dado por:

$$\left[\hat{t}_{\pi y} - t_{(n-1,1-\alpha/2)}\sqrt{\hat{V}(\hat{t}_{\pi y})},\ \hat{t}_{\pi y} + t_{(n-1,1-\alpha/2)}\sqrt{\hat{V}(\hat{t}_{\pi y})}\right]$$

```r
ns=12
alpha<-0.05
# cuantil distribucion T
z<-qt(1-alpha/2,df=ns-1)
z
```

```
## [1] 2.200985
```

```r
# intervalo
c(t-z*sqrt(var),t+z*sqrt(var))
```

```
## [1]  383056 1508140
```

Por lo tanto el intervalo del 95 % de confianza pata el total es:

$$[383\ 056\ ,\ 1\ 508\ 140]$$

b) Obtenga el intervalo de confianza usando el paquete survey en R.
Solución.

```
#cargamos el paquete
library(survey)

#pesos muestrales
w<- 1/pi_k

# almacenmos todo en un data frame
muestra <- data.frame(rev_84, p75, pi_k, w)

#probabilidades de inclusion de segundo orden

pi_kl<-pi_k %*% t(pi_k)
diag(pi_kl) <- pi_k
```

Observemos que con el paquete *survey* el total estimado es igual al que encontramos

```
mPoiss <- svydesign(id=~1, weights = ~w, data=muestra,
                fpc = ~pi_k, pps=ppsmat(pi_kl), variance = "HT" )

# estimacion del total
total<-svytotal(~rev_84,mPoiss)
total
```

```
##         total     SE
## rev_84 945598 255587
```

Y el intervalo es:

```
confint(total, df=degf(mPoiss))
```

```
##         2.5 %  97.5 %
## rev_84 383056 1508140
```

$$[383056\ ,\ 1508140]$$

El mismo intervalo que obtuvimos en a)

4. **Comparación de diseños unietápicos.**

   Suponga que cada mes se levanta un censo en N=411 puntos de venta para obtener los precios del aguacate y calcular el precio promedio. Además de este cálculo, lo importante de este seguimiento es analizar la **diferencia de los precios promedio** asociada a los dos últimos meses.

   Dado que el movimiento de los precios ha sido muy variable, se plantea realizar este seguimiento con más oportunidad, de ser posible cada semana. Sin embargo, para hacer la comparación cada semana no existe mucho dinero y se plantea la opción de levantar una muestra del 10 % de los puntos de venta. Para esto se analizan dos características para definir la selección de la muestra.

   1. Considerar un m.a.s. del 10 % de los puntos de venta o un diseño Poisson con el mismo tamaño esperado de muestra, donde la variable tamaño se define por los precios de enero. Para el caso del m.a.s no se estimaría N, pero para el diseño Poisson, dado que se usan probabilidades diferentes, se considera estimar N.
   2. Considerar una sóla muestra que será visitada cada semana o bien seleccionar una muestra independiente cada semana.

   Para definir la mejor estrategia se acercan a usted y la información que le proporcionan son los precios del aguacate en enero y febrero recolectada en los 411 puntos de venta (DatosAguacatePobEneFeb.csv).

   Con esta información de apoyo, ayude a definir la mejor estrategia considerando como base la estimación de la **diferencia de los precios promedio** entre enero y febrero.

   Solución.

   Realizaremos los 4 estudios y compararemos las varainzas.

```
library(tidyverse)
library(psych)
library(latex2exp)
AguacatePob <- read.csv(file = 'DatosAguacatePobEneFeb.csv', sep=',',
                        fileEncoding="latin1")
set.seed(1971)
(N <- nrow(AguacatePob))

## [1] 411

(n=floor(N*0.1))

## [1] 41
```

Queremos la diferencia en el promedio de los precios, como tenemos toda la población podemos sacar el valor exacto o muestral.

```
(Diferencia=mean(AguacatePob[,2])-mean(AguacatePob[,1]))

## [1] 5.907105
```

- **Estudio Longitudinal**
  Al escoger la misma muestra en los dos meses podemos crear una nueva variabl restando los precios de Enero a los precios de Febrero

```
DiferenciaLongitudinal<-AguacatePob[,2]-AguacatePob[,1]
AguacatePob<-cbind(AguacatePob,DiferenciaLongitudinal)
```

  Nuetro parámetro de interes sera el siguiente:

$$\widehat{\theta} = \frac{1}{\widehat{N}} \sum_{k \in S} \frac{z_k}{\pi_k} \qquad \text{Con } z_k \text{ La diferencia longitudinal}$$

  - m.a.s
    Realizaremos el muestreo mediante sample, obteniendo como resultado los índices de las observaciones seleccionadas.
    En este caso tenemos equiprobabilidad y nuestro parámetro queda como:

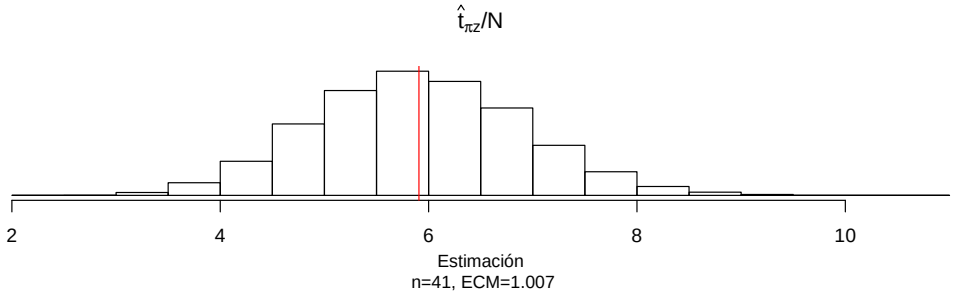
$$\widehat{\theta} = \frac{1}{n} \sum_{k \in S} z_k$$

```
(sindex=sample(1:N, n))
##  [1]  43  83 251  90 333  57 184 185 285  21 158  13 300  44 340 393  34 25
## [20] 246 118 101 287 154 343  38 303 320 163 330 156 177 345 365 113  59 14
## [39]  37 260 364
```

    Repetiremos la selección 10000 veces para analizar cómo se ve la distribución del estimador HT

```
B=10000
ParMuestra = replicate(B, sample(1:N, n))
Diferencia.HT= sapply(1:B,function(x, Indices, Pob)
      {mean(Pob[Indices[,x],"DiferenciaLongitudinal"])}
      , Indices=ParMuestra, Pob=AguacatePob)
h1=hist(Diferencia.HT, plot=FALSE, breaks=25)
plot(h1, freq=FALSE, main=TeX("$\\widehat{t}_{\\pi z}/N$"),
```

```
     xlab=paste("\n Estimación \n n=",n, ", ECM=",
    round(mean((Diferencia.HT-Diferencia)^2), 3) ,  sep=""),
    ylab="", cex.lab=.9, yaxt='n')
abline(v=Diferencia, col="red")
```




```
mean(Diferencia.HT)
```

```
## [1] 5.897464
```

Tenemos un error cuadático medio de 1.007 y un valor promedio del estimador de HT de:

```
mean(Diferencia.HT)
```

```
## [1] 5.897464
```

Ahora para la varianza usarmos la siguiente función:

$$\widehat{Var}\left(\widehat{\theta}\right) = \frac{1}{N^2}\widehat{Var}(\widehat{t}_{\pi z}) = \frac{1}{N^2}\frac{N^2}{n}\left(1-\frac{n}{N}\right)S^2_{zS}$$

```
varMAS.HT= sapply(1:B,function(x, Indices, Pob)
  {(1/N^{2})*(N^{2}/n)*(1-n/N)*(1/(n-1))*
  sum((Pob[Indices[,x],"DiferenciaLongitudinal"]
  -mean(Pob[Indices[,x],"DiferenciaLongitudinal"]))^{2})},
  Indices=ParMuestra, Pob=AguacatePob)
```

Teniendo un valor promedio de:

```
mean(varMAS.HT)
```

```
## [1] 0.9937154
```

Por lo tanto hay una desviación estándar de:

```
sqrt(mean(varMAS.HT))
```

```
## [1] 0.9968527
```

- Poisson

  En este caso tenemos que estimar N por lo tanto nuestro parámetro de interés es el siguiente:

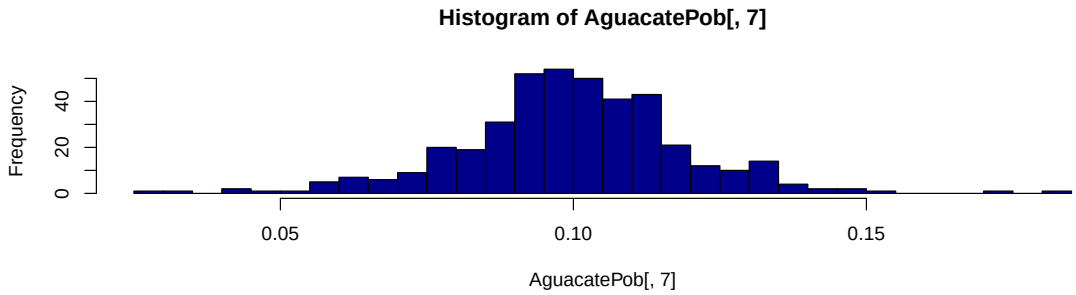
$$\widehat{\theta} = \frac{1}{\sum_{k \in S} \frac{1}{\pi_k}} \sum_{k \in S} \frac{z_k}{\pi_k}$$

  Primero calcularemos las probabilidades de inclusión con ayuda de los precios de Enero.

```
pik=pmin(n*AguacatePob[,1]/sum(AguacatePob[,1]), 1)
AguacatePob<-cbind(AguacatePob,pik)
```

  Obtenindo así el siguiente histograma de probabilidades:

```
hist(AguacatePob[,7], col = "darkblue",breaks=25)
```



  Ahora para hacer el muestreo Poisson creamos la siguiente función:

```
samplePoisson=function(pks){
    Nd=length(pks)
    index=1:Nd
    unif=runif(Nd)
    Unos=(unif<pks)*1
    indexs=index[Unos==1]
    return(indexs)
}
```

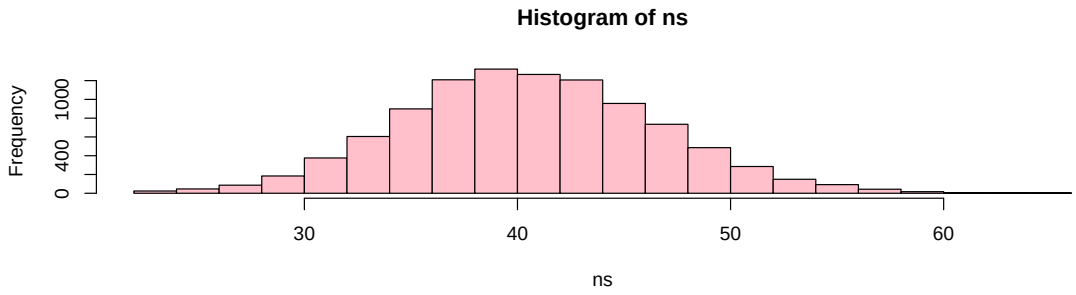
  Al aplicar samplePoisson obtenemos como resultado los índices de las observaciones seleccionadas.

```
(IndicesP=samplePoisson(AguacatePob[,7]))
```

```
##  [1]   4  26  57  60  65  66  70  72  79  88 109 120 153 154 161 171 198 21
## [20] 219 222 224 229 233 245 259 260 272 282 286 303 307 310 349 352 373 37
## [39] 384 385 395 404 410
```

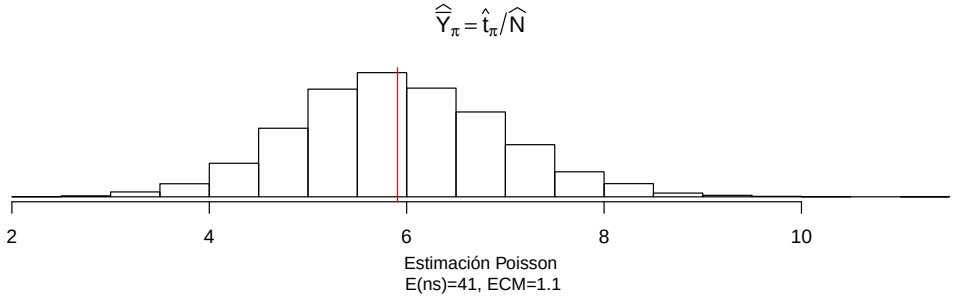
  Ahora al hacer las muestras no simepre son del mismo tamaño, por ejemplo, podemos ver los diferentes tamaños si se realizan 10,000 muestreos.

```
ParMuestraP = replicate(B, samplePoisson(AguacatePob[,7]))
ns= sapply(1:B,function(x, Indices){length(Indices[[x]])},
           Indices=ParMuestraP)
hist(ns,col="pink", breaks =25)
```



Repetiremos la selección 10,000 con los índices ya calculados veces para analizar cómo se ve la distribución del estimador HT

```
Diferencia.HT.P= sapply(1:B,function(x, Indices, Pob)
  {(1/sum(1/Pob[Indices[[x]],"pik"]))*sum(Pob[Indices[[x]],
    "DiferenciaLongitudinal"]/Pob[Indices[[x]],"pik"])},
  Indices=ParMuestraP, Pob=AguacatePob)

h2=hist(Diferencia.HT.P, plot=FALSE, breaks=25)
plot(h2, freq=FALSE,
main=TeX("$\\widehat{\\bar{Y}}_{\\pi}=\\widehat{t}_{\\pi}/\\widehat{N}$"),
xlab=paste("\n Estimación Poisson \n E(ns)=",n, ", ECM=",
  round(mean((Diferencia.HT.P-Diferencia)^2), 2) ,  sep=""),
      ylab="", cex.lab=.9, yaxt='n')
abline(v=Diferencia, col="red")
```



Tenemos un error cuadático medio de 1.1 y un valor promedio del estimador de HT de:

```
mean(Diferencia.HT.P)
```

```
## [1] 5.913393
```

Ahora para la varianza usarmos la siguiente función:

$$\widehat{Var}\left(\widehat{\theta}\right) = \frac{1}{\widehat{N}^2}\widehat{Var}(\widehat{t}_{\pi z}) = \frac{1}{\widehat{N}^2}\sum_{k\in S}\frac{z_k^2}{\pi_k^2}(1-\pi_k)$$

```
varP.HT= sapply(1:B,function(x, Indices, Pob)
         {(1/(sum(1/Pob[Indices[[x]],"pik"]))^{2})*
             sum((Pob[Indices[[x]],"DiferenciaLongitudinal"]/
                Pob[Indices[[x]],"pik"])^{2}*(1-Pob[Indices[[x]],"pik"]))},
        Indices=ParMuestraP, Pob=AguacatePob)
```

Teniendo un valor promedio de:

```
mean(varP.HT)
```

```
## [1] 1.970057
```

Por lo tanto hay una desviación estándar de:

```
sqrt(mean(varP.HT))
```

```
## [1] 1.403587
```

Adicionalmente podemos calcular la varianza mediante un estimador alternativo, de la siguiente forma:

$$Var\left(\widehat{\theta}_{alt}\right) = \frac{1}{N^2}Var(\widehat{t_{\pi z alt}}) = \frac{1}{N^2}\sum_{k\in U}\frac{(z_k-\bar{z_U})^2}{\pi_k}(1-\pi_k)$$

```
(varP.HTalt=(1/N^{2})*sum((AguacatePob[,6]-
          mean(AguacatePob[,6]))^{2}/
         AguacatePob[,7]*(1-AguacatePob[,7])))
```

```
## [1] 1.069653
```

Por lo tanto hay una desviación estándar de:

```
sqrt(mean(varP.HTalt))
```

```
## [1] 1.03424
```

- **Estudio Transversal**

  Nuetro parámetro de interes sera el siguiente:

  $\widehat{\theta} = \frac{1}{\widehat{N_y}}\sum_{k\in S_y}\frac{y_k}{\pi_k} - \frac{1}{\widehat{N_x}}\sum_{k\in S_x}\frac{x_k}{\pi_k}$ Con x_k & y_k Los precios en Enero y Febrero respectivamente

- m.a.s

  En este caso tenemos equiprobabilidad y N conocido, por lo tanto nuestro parámetro queda como:

$$\widehat{\theta} = \frac{1}{n}\sum_{k\in S_y} y_k - \frac{1}{n}\sum_{k\in S_y} y_k$$

  Realizaremos el muestreo mediante sample, obteniendo como resultado los índices de las observaciones seleccionadas. Haremos esto tanto para Enero como para Febrero

```
(sindexEne=sample(1:N, n))
```

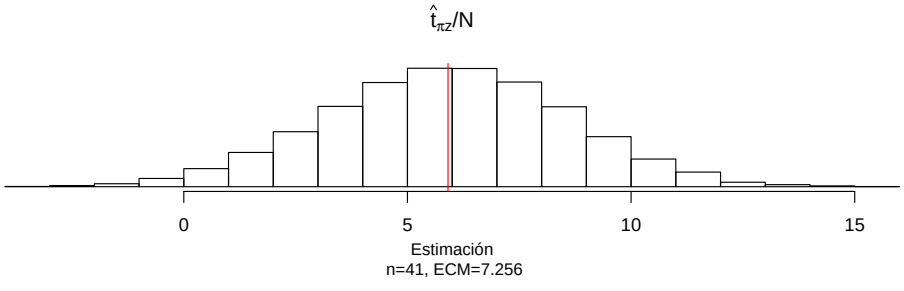
```
##  [1] 344 339 120 336 292 262  15 240 209 404  25 210 348 234  45 114 26
## [20] 255  39 161 194 247  17 249 195 406  47  57 154 142 196 244 111 36
## [39]  99 104  29
```

```
(sindexFeb=sample(1:N, n))
```

```
##  [1]  32 405 188  39 352 320  73  84 348 208 404  66 118 170  50 364 41
## [20] 291 175 229 132 399 397 366 342 287 219 407 250 167 327  19 117 24
## [39] 128 376 147
```

  Repetiremos la selección 10000 veces para analizar cómo se ve la distribución del estimador HT

```
ParMuestraEne = replicate(B, sample(1:N, n))
ParMuestraFeb = replicate(B, sample(1:N, n))
Diferencia.HT.T= sapply(1:B,function(x, IndicesEne,IndicesFeb, Pob)
{mean(Pob[IndicesFeb[,x],"AguacateFeb2022"])-mean(Pob[IndicesEne[,x],
"AguacateEne2022"])},
  IndicesEne=ParMuestraEne,IndicesFeb=ParMuestraFeb, Pob=AguacatePob)
h3=hist(Diferencia.HT.T, plot=FALSE, breaks = 25)
plot(h3, freq=FALSE, main=TeX("$\\widehat{t}_{\\pi z}/N$"),
     xlab=paste("\n Estimación \n n=",n, ", ECM=",
        round(mean((Diferencia.HT.T-Diferencia)^2), 3) ,  sep=""),
     ylab="", cex.lab=.9, yaxt='n')
abline(v=Diferencia, col="red")
```



Tenemos un error cuadático medio de 7.256 y un valor promedio del estimador de HT de:

```
mean(Diferencia.HT.T)
```

```
## [1] 5.873142
```

Ahora para la varianza usarmos la siguiente función:

$$\begin{aligned}\widehat{Var}\left(\widehat{\theta}\right) &= \widehat{Var}\left(\frac{1}{N}(\widehat{t}_{\pi y}-\widehat{t}_{\pi x})\right)\\ &= \frac{1}{N^2}\left(\widehat{Var}(\widehat{t}_{\pi y})+\widehat{Var}(\widehat{t}_{\pi x})\right)\\ &= \frac{1}{N^2}\frac{N^2}{n}\left(1-\frac{n}{N}\right)S^2_{yS_y}+\frac{1}{N^2}\frac{N^2}{n}\left(1-\frac{n}{N}\right)S^2_{xS_x}\end{aligned}$$

```
varMAS.HT.T= sapply(1:B,function(x, IndicesEne, IndicesFeb, Pob)
{(1/N^{2})*(N^{2}/n)*(1-n/N)*(1/(n-1))*sum((Pob[IndicesFeb[,x],
"AguacateFeb2022"] -mean(Pob[IndicesFeb[,x],"AguacateFeb2022"]))^{2})+
    (1/N^{2})*(N^{2}/n)*(1-n/N)*(1/(n-1))*sum((Pob[IndicesEne[,x],
"AguacateEne2022"]-mean(Pob[IndicesEne[,x],"AguacateEne2022"]))^{2})},
IndicesEne=ParMuestraEne, IndicesFeb=ParMuestraFeb, Pob=AguacatePob)
```

Teniendo un valor promedio de:

```
mean(varMAS.HT.T)
```

```
## [1] 7.289067
```

Por lo tanto hay una desviación estándar de:

```
sqrt(mean(varMAS.HT.T))
```

```
## [1] 2.699827
```

- Poisson

  En este caso tenemos que estimar N por lo tanto nuestro parámetro de interés es el siguiente:

$$\widehat{\theta}=\frac{1}{\sum_{k\in S_y}\frac{1}{\pi_k}}\sum_{k\in S_y}\frac{y_k}{\pi_k}-\frac{1}{\sum_{k\in S_x}\frac{1}{\pi_k}}\sum_{k\in S_x}\frac{x_k}{\pi_k}$$

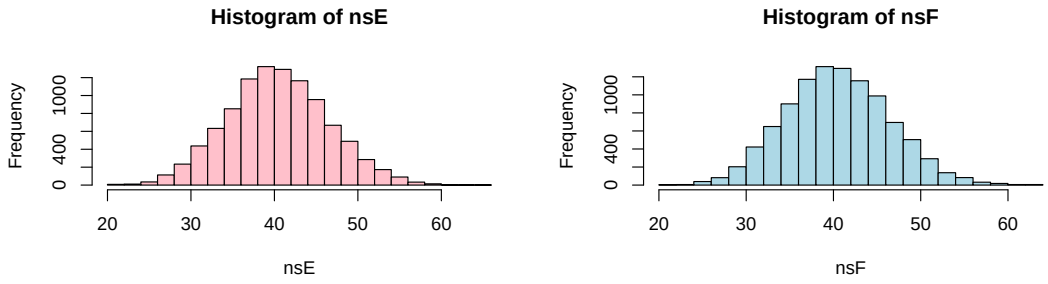
  Usaremos las probabilidades de inclusión de primer orden obtenidas en el estudio Transversal.De igual manera para hacer el muestreo Poisson usamos la función ya creada.

  Tendresmos que aplicar samplePoisson para Enero y Febreo y obtener así el resultado de los índices de las observaciones seleccionadas.

```
(IndicesPEne=samplePoisson(AguacatePob[,7]))
##  [1]   7  14  29  34  46  56  95 116 137 146 147 152 153 155 156 157 16
## [20] 196 198 209 218 223 229 230 243 244 247 256 257 282 284 285 288 29
## [39] 332 334 339 350 355 372 380 387 395 398
(IndicesPFeb=samplePoisson(AguacatePob[,7]))
##  [1]   1   3  33  45  46  54  65  72  82  96 100 106 107 138 159 163 16
## [20] 189 194 196 198 220 228 233 240 244 249 260 262 268 273 274 275 27
## [39] 303 305 312 321 342 343 345 348 356 375 389
```

Ahora al hacer las muestras no simepre son del mismo tamaño, por ejemplo, podemos ver los diferentes tamaños si se realizan 10,000 muestreos.

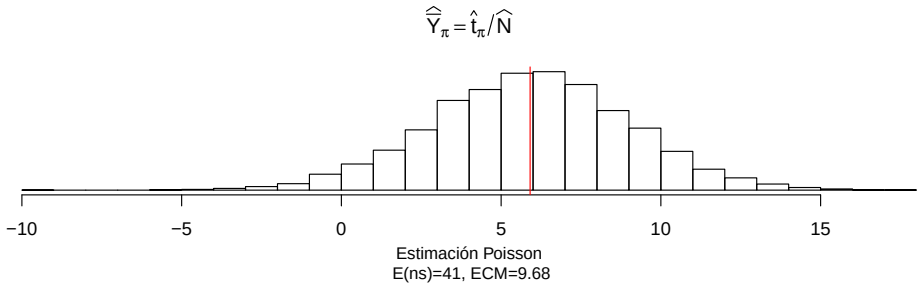
```
par(mfrow=c(1,2))
ParMuestraPEne = replicate(B, samplePoisson(AguacatePob[,7]))
nsE= sapply(1:B,function(x, Indices){length(Indices[[x]])},
   Indices=ParMuestraPEne)
hist(nsE,col="pink",breaks=25)

ParMuestraPFeb = replicate(B, samplePoisson(AguacatePob[,7]))
nsF= sapply(1:B,function(x, Indices){length(Indices[[x]])},
    Indices=ParMuestraPFeb)
hist(nsF,col="lightblue",breaks=25)
```



Repetiremos la selección 10,000 con los índices ya calculados veces para analizar cómo se ve la distribución del estimador HT

```
Diferencia.HT.P.T= sapply(1:B,function(x, IndicesEne, IndicesFeb, Pob)
{(1/sum(1/Pob[IndicesFeb[[x]],"pik"]))*sum(Pob[IndicesFeb[[x]],
"AguacateFeb2022"]/Pob[IndicesFeb[[x]],"pik"])-
    (1/sum(1/Pob[IndicesEne[[x]],"pik"]))*sum(Pob[IndicesEne[[x]],
 "AguacateEne2022"]/Pob[IndicesEne[[x]],"pik"])},
IndicesEne=ParMuestraPEne, IndicesFeb=ParMuestraPFeb, Pob=AguacatePob)
```

```
h4=hist(Diferencia.HT.P.T, plot=FALSE, breaks=25)
plot(h4, freq=FALSE,
    main=TeX("$\\widehat{\\bar{Y}}_{\\pi}=\\widehat{t}_{\\pi}/\\widehat{N}$
     xlab=paste("\n Estimación Poisson \n E(ns)=",n, ", ECM=",
        round(mean((Diferencia.HT.P.T-Diferencia)^2), 2) ,  sep=""),
     ylab="", cex.lab=.9, yaxt='n')
abline(v=Diferencia, col="red")
```



Tenemos un error cuadático medio de 9.68 y un valor promedio del estimador de HT de:

```
mean(Diferencia.HT.P.T)
## [1] 5.893936
```

Ahora para la varianza usarmos la siguiente función:

$$\begin{aligned}\widehat{Var}\left(\widehat{\theta}\right) &= \widehat{Var}\left(\frac{1}{\widehat{N}_y}\widehat{t}_{\pi y} - \frac{1}{\widehat{N}_x}\widehat{t}_{\pi x}\right)\\ &= \left(\frac{1}{\widehat{N}_y^2}\widehat{Var}(\widehat{t}_{\pi y}) + \frac{1}{\widehat{N}_x^2}\widehat{Var}(\widehat{t}_{\pi x})\right)\\ &= \frac{1}{\widehat{N}_y^2}\sum_{k\in S_y}\frac{y_k^2}{\pi_k^2}(1-\pi_k) + \frac{1}{\widehat{N}_x^2}\sum_{k\in S_x}\frac{x_k^2}{\pi_k^2}(1-\pi_k)\end{aligned}$$

```
varP.HT.T= sapply(1:B,function(x, IndicesEne, IndicesFeb, Pob)
{(1/(sum(1/Pob[IndicesFeb[[x]],"pik"]))^{2})*
    sum((Pob[IndicesFeb[[x]],"AguacateFeb2022"]/
 Pob[IndicesFeb[[x]],"pik"])^{2}*(1-Pob[IndicesFeb[[x]],"pik"]))+
    (1/(sum(1/Pob[IndicesEne[[x]],"pik"]))^{2})*
    sum((Pob[IndicesEne[[x]],"AguacateEne2022"]/
 Pob[IndicesEne[[x]],"pik"])^{2}*(1-Pob[IndicesEne[[x]],"pik"]))},
IndicesEne=ParMuestraPEne, IndicesFeb=ParMuestraPFeb, Pob=AguacatePob)
```

Teniendo un valor promedio de:

```
mean(varP.HT.T)
```

```
## [1] 212.0657
```

Por lo tanto hay una desviación estándar de:

```
sqrt(mean(varP.HT.T))
```

```
## [1] 14.56247
```

Adicionalmente podemos calcular la varianza mediante un estimador alternativo, de la siguiente forma:

$$
\begin{aligned}
Var\left(\widehat{\theta}_{alt}\right) &= \widehat{Var}\left(\frac{1}{\widehat{N}_y}\widehat{t}_{\pi yalt} - \frac{1}{\widehat{N}_x}\widehat{t}_{\pi xalt}\right) \\
&= \left(\widehat{Var}(\frac{1}{\widehat{N}_y}\widehat{t}_{\pi yalt}) + \widehat{Var}(\frac{1}{\widehat{N}_x}\widehat{t}_{\pi xalt})\right) \\
&= \frac{1}{N^2}\sum_{k\in U}\frac{(y_k - \bar{y_U})^2}{\pi_k}(1-\pi_k) + \frac{1}{N^2}\sum_{k\in U}\frac{(x_k - \bar{x_U})^2}{\pi_k}(1-\pi_k)
\end{aligned}
$$

```
(varP.HTalt.T=(1/N^{2})*sum((AguacatePob[,2]-
                       mean(AguacatePob[,2]))^{2}/
                      AguacatePob[,7]*(1-AguacatePob[,7]))+
       (1/N^{2})*sum((AguacatePob[,1]-
           mean(AguacatePob[,1]))^{2}/
          AguacatePob[,7]*(1-AguacatePob[,7])))
```

```
## [1] 9.497978
```

Por lo tanto hay una desviación estándar de:

```
sqrt(mean(varP.HTalt.T))
```

```
## [1] 3.081879
```

Para concluir podemos ver que de los cuatro métodos simulados, el que tiene menor varianza y error cuadrático medio es el el Muestreo Aleatorio Simple Longitudinal, por lo tanto este es el mejor esquema de selección.