UNIVERSIDAD DE PUERTO RICO
RECINTO DE RIO PIEDRAS
FACULTAD DE ADMINISTRACION DE EMPRESAS
Instituto de Estadística y Sistemas Computadorizados de Información

MUESTREO con R

Preparado por:
José Carlos Vega Vilca, Ph.D.
jose.vega23@upr.edu

Contenido

INTRODUCCION AL MUESTREO

Censo.- es el estudio completo de los elementos de la población, con el fin de calcular sus parámetros

Muestreo.- es el estudio de una selección de elementos de una población, llamada muestra, con el fin de estimar los parámetros de la población, mediante la Inferencia Estadística.

VENTAJAS DEL METODO DE MUESTREO

Costo reducido.- Si los datos se obtienen únicamente de una pequeña fracción del total, los gastos son menores que los que se realizarían en un censo.

Mayor rapidez.- Los datos pueden ser recolectados y resumidos más rápidamente con una muestra que con un censo.

Mayor exactitud.- Si el volumen de trabajo es reducido se puede emplear personal capacitado al cual se le puede someter a entrenamiento intensivo

Cuidado de la población.- En estudios destructivos, conserva los elementos de la población; como por ejemplo, el estudio del tiempo de duración de baterías.

MUESTREO PROBABILISTICO

Todos los individuos tienen probabilidad conocida de ser elegidos.
Todas la posibles muestras de tamaño *n* tienen probabilidad conocida de ser elegidas.
Sólo estos métodos nos aseguran *representatividad* de la muestra.
Los tipos de muestreo probabilístico son:

1. Muestreo Aleatorio Simple
2. Muestreo Aleatorio Sistemático
3. Muestreo Aleatorio Estratificado
4. Muestreo Aleatorio por Conglomerados

MUESTREO NO PROBABILISTICO

Aplicado cuando el muestreo probabilístico resulta excesivamente costoso
Todos los individuos **no** tienen la misma probabilidad de ser elegidos.
No se tiene la certeza de que muestra extraída sea representativa
No se puede hacer generalizaciones.

SELECCIÓN ALEATORIA

Una muestra tiene *selección aleatoria* cuando el proceso de selección de unidades se hace por sorteo, ya que de esta manera todas las unidades tienen la misma probabilidad de ser seleccionadas.

MARCO DE MUESTREO

El marco muestral es una representación de todos los elementos de la población objetivo que consta de una lista de características que permitan identificar dicha población.

PARÁMETROS DE UNA POBLACIÓN

Total poblacional: T

$$T = \sum_{i=1}^{N} X_i$$

Media poblacional: μ

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N} = \frac{T}{N}$$

Proporción poblacional: P

$$P = \frac{Y}{N}$$

Donde: N = tamaño de la población, $Y = \sum_{i=1}^{N} X_i$, donde $X_i = \begin{cases} 1 & \textit{éxito} \\ 0 & \textit{fracaso} \end{cases}$

MUESTREO ALEATORIO SIMPLE

Si se tiene que seleccionar una muestra de *n* elementos de una población de tamaño *N*. El muestreo aleatorio simple es aquel en el que cada muestra posible de tamaño *n* tienen la misma probabilidad de ser seleccionada.

Estimación de la media poblacional: $\bar{x}$

Sean $x_1, x_2, \cdots, x_n$ los valores observados de una muestra de tamaño *n*, tomada de una población de tamaño *N*.

1) Estimación puntual de la media: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

2) Estimación de la varianza de la media muestral: $var(\bar{x}) = \frac{s^2}{n}\left(\frac{N-n}{N}\right)$

3) Estimación del error estándar de la media muestral: $se(\bar{x}) = \sqrt{\frac{s^2}{n}\left(\frac{N-n}{N}\right)}$

4) Estimación por intervalos de la media: $\bar{x} \pm z_0 \times se(\bar{x})$

Estimación del total de la poblacional: $\hat{T}$

Sean $x_1, x_2, \cdots, x_n$ los valores observados de una muestra de tamaño *n*, tomada de una población de tamaño *N*.

1) Estimación puntual del total: $\hat{T} = N\ \bar{x}$

2) Estimación por intervalos del total: $N\,\bar{x} \pm z_0 \times N\, se(\bar{x})$

Estimación de la proporción poblacional: $\hat{p}$

Sean $x_1, x_2, \cdots, x_n$ los valores observados ("1" y "0") de una muestra de tamaño *n*, tomada de una población de tamaño *N*.

1) Estimación puntual de la proporción: $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad , \quad \hat{q} = 1 - \hat{p}$

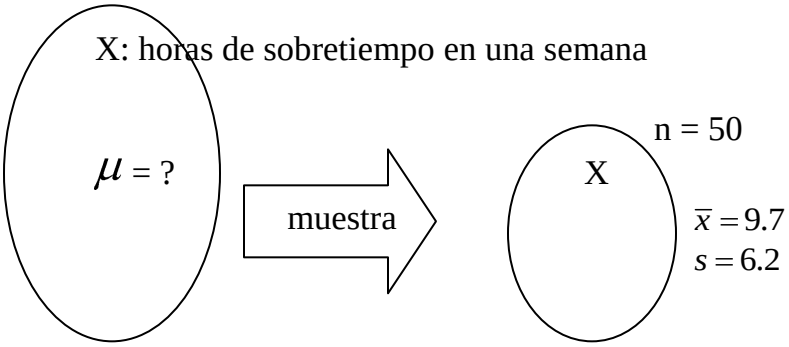
2) Estimación de varianza de la proporción muestral: $var(\hat{p}) = \frac{\hat{p}\,\hat{q}}{n-1}\left(\frac{N-n}{N}\right)$

3) Estimación del error estándar de la proporción muestral: $se(\hat{p}) = \sqrt{var(\hat{p})}$

4) Estimación por intervalos de la media: $\hat{p} \pm z_0 \times se(\hat{p})$

Ejemplo1
Una empresa tiene 189 contables. En una muestra aleatoria de 50 de ellos, el número medio de horas trabajadas en sobretiempo en una semana fue de 9.7 horas con una desviación estándar de 6.2 horas. Halle un intervalo del 95% de confianza para el número medio de horas trabajadas en sobretiempo en una semana.

Población: conjunto de contables de la empresa (N = 189)



Parámetro: μ = número medio de horas trabajadas en sobretiempo en una semana.

Estimación de la varianza de la media muestral:

$$var(\bar{x}) = \frac{s^2}{n}\left(\frac{N-n}{N}\right) = \frac{6.2^2}{50}\left(\frac{189-50}{189}\right) = 0.5654$$

Estimación del error estándar de la media muestral:

$$se(\bar{x}) = \sqrt{0.5654} = 0.7519$$

Intervalo de confianza: $IC(\mu) = [a, b]$

$$a = \bar{x} - z_0 \times se(\bar{x}) = 9.7 - 1.96 \times 0.7519 = 8.2263$$
$$b = \bar{x} + z_0 \times se(\bar{x}) = 9.7 + 1.96 \times 0.7519 = 11.1737$$

Intervalo buscado: $IC(\mu) = [8.2263, 11.1737]$

Interpretación: *El intervalo encontrado brinda un 95% de contener al verdadero valor del parámetro, tiempo medio trabajado en sobretiempo en una semana.*

Ejemplo2
En el ejemplo anterior, halle un intervalo del 95% de confianza para el número total de horas trabajadas en sobretiempo en una semana.

Parámetro: T_x = número total de horas trabajadas en sobretiempo en una semana.

Intervalo de confianza: $IC(T) = [c, d]$

$$c = N \times a = 189 \times 8.2263 = 1554.771$$
$$d = N \times b = 189 \times 11.1737 = 2111.829$$

Intervalo buscado: $IC(T) = [1554.771, 2111.829]$

Interpretación: *El intervalo encontrado brinda un 95% de contener al verdadero valor del parámetro, tiempo total trabajado en sobretiempo en una semana.*

Programa #1. Estimación de la media y el total poblacional en muestreo aleatorio simple

```
msa.m=function(N,n,media,desv)
{ f=n/N
  varmed=(desv^2/n)*(1-f)
  desmed=sqrt(varmed)
  a1=media-1.96*desmed
  b1=media+1.96*desmed
  a2=N*a1
  b2=N*b1

  cat("media: IC =  ",a1, "--",b1,"\n")
  cat("total: IC =  ",a2, "--",b2,"\n")
}
```

Aplicación del Programa #1.

```
> msa.m(189,50,9.7,6.2)
media: IC =   8.226198 -- 11.17380
total: IC =   1554.751 -- 2111.849
```

Ejemplo3

Una agencia bancaria que cuenta con un total de 4800 clientes, los que están clasificados como clientes tipo 1 ó tipo 0. Una muestra aleatoria de 120 clientes: 89 tipo "0" y 31 tipo "1", fue usada para hallar un intervalo de confianza del 95% para la proporción de clientes que fueron denominados "tipo 1".

Programa #2. Estimación de la proporción poblacional mediante muestreo aleatorio simple.

```
msa.p=function(N,n,exitos)
{ f=n/N
  p=exitos/n ; q=1-p
  varp=(p*q/(n-1))*(1-f)
  desp=sqrt(varp)
  a=p-1.96*desp
  b=p+1.96*desp
  cat("proporcion: IC =  ",a, "--",b,"\n")
}
```

Aplicación del Programa #2.

```
> msa.p(4800,120,31)
proporción: IC =   0.1806765 -- 0.3359901
```

Ejemplo4

Un auditor, examinando un total de 840 facturas pendientes de cobro, de una empresa, tomó una muestra aleatoria de 120 facturas. Usando los datos del archivo "**muestreo1.xls**", mediante muestreo aleatorio simple.

a) Hallar un intervalo del 95% de confianza para estimar la cantidad total de cobros pendientes
b) Hallar un intervalo del 95% de confianza para estimar la proporción de facturas por cobrar con menos de 100 dólares

Selección de la muestra aleatoria de 120 facturas de un total de 840

```
  [1] 839 292 158 350 409  52 562 411 162 525 221  93 447 608 425 351 588 503
 [19] 359 202 122 571 443 838 295   6 398 143 178 774 538 452 229 787 110 149
 [37]  29 182   3  54 205 778 649 264 362 271 496 388 151 377 223 831 517 105
 [55] 702 830 832 531 544 396 506 239  23 415 512 600 468  47 160 491 201 812
 [73] 603 734  57 284 273 228 798 598 569 615   5 198 629 505 330 484 663 651
 [91] 688 259 405 451 722 645  49 331 736 686 490 101 145 813 667  75 423 680
[109] 422 287 207 144 412 470 597 431 188 303 550 806
```

Cantidad por cobrar en la factura seleccionada

```
  [1] 136.41 160.31 158.61 181.41 246.84 151.57 113.22 118.23 151.96 109.23
 [11] 113.71  61.80  75.84 152.89  93.07 159.22 139.15 168.94 122.68  88.05
 [21]  28.22 183.41 153.14 153.10 160.16 149.68 117.32 123.36 106.76  98.36
 [31]  70.33 188.78 156.88  72.12 171.73 149.75 104.62 103.75  89.10 133.97
```

```
 [41] 186.87 132.64 206.54  70.18 145.09 126.41 164.18 156.42 112.54 103.77
 [51] 109.30  82.04 172.80 120.91 130.67 112.04 122.79 132.39 111.01 212.56
 [61]  77.95 152.56 141.76 123.17 135.37 156.54 164.46 124.17 235.36 179.80
 [71] 148.95 150.84 177.42 182.05 111.60 202.87 197.98 183.64 145.23 112.16
 [81] 195.92 165.27  95.11 143.63  65.02 133.86 206.05 132.74 113.56 142.26
 [91] 175.68 152.40  98.18 188.58 153.09 104.04 132.97 109.89 142.03 110.18
[101] 170.91 127.99 181.44  71.54 149.91 145.45 165.68  96.51 113.26  54.50
[111] 189.80  89.19 126.58 109.86 123.85  51.71 201.91 209.89 140.74 114.47
```

Cálculos: $\bar{x} = 136.903$, $s = 40.50198$

Aplicación del Programa #1.

```
> msa.m(840,120,136.903,40.50198)
media: IC =   130.1938 -- 143.6122
total: IC =   109362.8 -- 120634.2
```

Las facturas por cobrar con menos de 100 dólares, son las siguientes 20 facturas de la muestra de 120:

```
[1]  12  13  15  20  21  30  31  34  39  44
[11] 52  61  83  85  93 104 108 110 112 116
```

Cálculos: #éxitos = 20

Aplicación del Programa #2

```
> msa.p(840,120,20)
proporcion: IC =   0.1046736 -- 0.2286597
```

MUESTREO SISTEMATICO de 1 en *k*

Si se tiene que seleccionar una muestra de *n* elementos de una población de tamaño *N*. El muestreo sistemático de 1 en *k*, donde $k = N/n$, se realiza de la siguiente manera:

1) El primer elemento es seleccionado aleatoriamente entre los primeros *k* elementos
2) Los próximos elementos son seleccionados cada *k*-elementos.

> *En un muestreo sistemático de 1 en k, el número de muestras posibles que se pueden obtener es igual a k.*

Ejemplo1

Desde una población de $N = 12$ hogares, se selecciona una muestra de 4 hogares para investigar acerca de la variable "número de personas que viven en el hogar"

hogares	1	2	3	4	5	6	7	8	9	10	11	12
#personas	4	3	5	6	3	4	3	4	7	5	2	1

Usando el muestreo sistemático de 1 en 3, las 3 muestras posibles que pueden ser seleccionadas son:

Muestra #1

hogar	1	4	7	10
#personas	4	6	3	5

Muestra #2

hogar	2	5	8	11
#personas	3	3	4	2

Muestra #3

hogar	3	6	9	12
#personas	5	4	7	1

Suponiendo que la muestra seleccionada fue la muestra #2.

Cálculos: $\bar{x} = 4.50$, $s = 1.290994$

Aplicación del Programa #1.

```
> msa.m(12,4,4.50,1.290994)
media: IC =   3.46699 -- 5.53301
total: IC =   41.60388 -- 66.39612
```

Ejemplo2

Un auditor, examinando un total de 840 facturas pendientes de cobro, de una empresa, tomó una muestra aleatoria de 120 facturas. Usando los datos del archivo "**muestreo1.xls**", mediante muestreo sistemático de 1 en 7

1) Hallar un intervalo del 95% de confianza para estimar la cantidad total de cobros pendientes
2) Hallar un intervalo del 95% de confianza para estimar la proporción de facturas por cobrar con menos de 100 dólares

Selección de la muestra aleatoria de 120 facturas de un total de 840

```
  [1]   3  10  17  24  31  38  45  52  59  66  73  80  87  94 101 108 115 122
 [19] 129 136 143 150 157 164 171 178 185 192 199 206 213 220 227 234 241 248
 [37] 255 262 269 276 283 290 297 304 311 318 325 332 339 346 353 360 367 374
 [55] 381 388 395 402 409 416 423 430 437 444 451 458 465 472 479 486 493 500
 [73] 507 514 521 528 535 542 549 556 563 570 577 584 591 598 605 612 619 626
 [91] 633 640 647 654 661 668 675 682 689 696 703 710 717 724 731 738 745 752
[109] 759 766 773 780 787 794 801 808 815 822 829 836
```

Cantidad por cobrar en la factura seleccionada

```
  [1]  89.10  92.41 136.10  72.26  94.57 171.37 119.14 151.57 125.82 131.24
 [11] 113.26 115.07 146.94 145.45 127.99  68.99 161.47  28.22 145.38 194.64
 [21] 123.36  73.22 213.26 195.10 182.54 106.76 132.95  97.61  68.30 178.63
 [31] 195.03 100.21  77.19 125.22 163.57 142.56  55.76 101.21  84.90  94.46
 [41]  57.80 144.17 175.62  94.80  95.28 115.49 161.26 198.14 101.48 111.66
 [51] 157.03 154.02  80.18 131.57 142.02 156.42 125.55 100.40 246.84 231.28
 [61] 165.68 125.40  94.05  56.34 188.58  82.17  66.06  87.92 151.16 135.11
 [71] 178.48  72.02  87.26 165.15 174.55 210.91  95.00 176.73 128.62 120.36
 [81]  88.16  98.87 177.92  96.30 157.03 112.16 211.89 145.35 113.36 222.20
 [91]  83.13 141.89 195.24 144.80 131.24 128.80 127.59 125.71 117.14  99.79
[101] 155.87 146.59 187.47 104.35  87.21 134.84 136.10 126.11  89.54 199.16
[111] 186.57 177.50  72.12  95.91  67.00 195.18 120.34 150.27 142.00 186.00
```

Cálculos: $\bar{x} = 131.3674$, $s = 43.71545$

Aplicación del Programa #1.

```
> msa.m(840,120,131.3674,43.71545)
media: IC =   124.1259 -- 138.6089
total: IC =   104265.8 -- 116431.5
```

MUESTREO ESTRATIFICADO

Si se tiene que seleccionar una muestra de *n* elementos de una población de tamaño *N*, la cual está dividida en *k* estratos, mutuamente excluyentes de tamaños $N_1, N_2, \ldots, N_k$, tal que:

$$N_1 + N_2 + \cdots + N_k = N$$

El muestreo estratificado consiste en seleccionar una muestra desde cada estrato de tamaños $n_1, n_2, \ldots, n_k$, tal que

$$n_1 + n_2 + \cdots + n_k = n$$

Estimación de la media poblacional: $\bar{x}_{str}$

Sean $\bar{x}_1, \bar{x}_2, \cdots, \bar{x}_k$ y $s_1^2, s_2^2, \cdots, s_k^2$ las medias y las varianzas muestrales desde cada estrato

1) Estimación puntual de la media: $\bar{x}_{str} = \frac{1}{N} \sum_{i=1}^{k} N_i \, \bar{x}_i$

2) Estimación de la varianza de la media muestral:

$$var(\bar{x}_{str}) = \frac{N_1^2 \, var(\bar{x}_1) + N_2^2 \, var(\bar{x}_2) + \cdots + N_k^2 \, var(\bar{x}_k)}{N^2}$$

Donde: $var(\bar{x}_i) = \frac{s_i^2}{n_i} \left(\frac{N_i - n_i}{N_i} \right) \qquad i = 1, 2, \cdots, k$

3) Estimación del error estándar de la media muestral: $se(\bar{x}_{str}) = \sqrt{var(\bar{x}_{st})}$

4) Estimación por intervalos de la media: $\bar{x}_{str} \pm z_0 \times se(\bar{x}_{str})$

Estimación del total de la poblacional: $\hat{T}_{str}$

Sean $\bar{x}_1, \bar{x}_2, \cdots, \bar{x}_k$ y $s_1^2, s_2^2, \cdots, s_k^2$ las medias y las varianzas muestrales desde cada estrato

1) Estimación puntual del total: $\hat{T}_{str} = N \, \bar{x}_{str}$

2) Estimación por intervalos del total: $N\,\bar{x}_{str} \pm z_0 \times N\,se(\bar{x}_{str})$

Estimación de la proporción poblacional: $\hat{p}_{str}$

Sean $\hat{p}_1, \hat{p}_2, \cdots, \hat{p}_k$ las proporciones muestrales desde cada estrato

1) Estimación puntual de la proporción: $\hat{p}_{str} = \frac{1}{N} \sum_{i=1}^{k} N_i\,\hat{p}_i$

2) Estimación de varianza de la proporción muestral:

$$var(\hat{p}_{str}) = \frac{N_1^2\,var(\hat{p}_1) + N_2^2\,var(\hat{p}_2) + \cdots + N_k^2\,var(\hat{p}_k)}{N^2}$$

Donde: $var(\hat{p}_i) = \frac{\hat{p}_i\,\hat{q}_i}{n_i - 1}\left(\frac{N_i - n_i}{N_i}\right) \qquad i = 1, 2, \cdots, k$

3) Estimación del error estándar de la proporción muestral: $se(\hat{p}_{str}) = \sqrt{var(\hat{p}_{str})}$

4) Estimación por intervalos de la media: $\hat{p}_{str} \pm z_0 \times se(\hat{p}_{str})$

Ejemplo1:

Una pequeña ciudad contiene un total de 1800 hogares. La ciudad está dividida en tres distritos que contienen 820, 540 y 440 hogares, respectivamente. Una muestra aleatoria estratificada de 310 hogares contiene 120, 100 y 90 hogares, respectivamente de estos tres distritos. Se pide a los miembros de la muestra que calculen su factura total de electricidad consumida en los meses de invierno. Las respectivas medias muestrales son \$290, \$352 y \$427, y las respectivas desviaciones estándar muestrales son \$47, \$61 y \$93.

Distritos	N_i	n_i	promedio	desviación estándar
1	820	120	290	47
2	540	100	352	61
3	440	90	427	93

Población: conjunto de hogares de una ciudad

Estratos: distritos de la ciudad

Variable: pago de electricidad consumida en los meses de invierno

Estimación del promedio del pago de electricidad consumida en los meses de invierno

1) Estimación puntual de la media:

$$\bar{x}_{str} = \frac{1}{N}\sum_{i=1}^{3} N_i\,\bar{x}_i = \frac{1}{1800}\ 820\times 290 + 540\times 352 + 440\times 427$$

$$\bar{x}_{str} = \frac{1}{1800}\times 615760 = 342.0888889$$

2) Estimación de la varianza de la media muestral, en cada estrato

$$var(\bar{x}_1) = \frac{s_1^2}{n_1}\left(\frac{N_1 - n_1}{N_1}\right) = \frac{47^2}{120}\left(\frac{820-120}{820}\right) = 15.71443089$$

$$var(\bar{x}_2) = \frac{s_2^2}{n_2}\left(\frac{N_2 - n_2}{N_2}\right) = \frac{61^2}{100}\left(\frac{540-100}{540}\right) = 30.31925926$$

$$var(\bar{x}_3) = \frac{s_3^2}{n_3}\left(\frac{N_3 - n_3}{N_3}\right) = \frac{93^2}{90}\left(\frac{440-90}{440}\right) = 76.44318182$$

3) Estimación de la varianza muestral de la media muestral estratificada

$$var(\bar{x}_{str}) = \frac{N_1^2\ var(\bar{x}_1) + N_2^2\ var(\bar{x}_2) + N_3^2\ var(\bar{x}_3)}{N^2}$$

$$var(\bar{x}_{str}) = \frac{820^2\times 15.71443089 + 540^2\times 30.31925926 + 440^2\times 76.44318182}{1800^2}$$

$$var(\bar{x}_{str}) = \frac{34206879.333}{1800^2} = 10.55767881$$

4) Estimación del error estándar de la media muestral

$$se(\bar{x}_{str}) = \sqrt{10.55767881} = 3.249258193$$

5) Estimación del intervalo de confianza para la media

$$a = \bar{x}_{str} - z_0 \times se(\bar{x}_{str}) = 342.0888889 - 1.96 \times 3.249258193 = 335.7203428$$

$$b = \bar{x}_{str} + z_0 \times se(\bar{x}_{str}) = 342.0888889 + 1.96 \times 3.249258193 = 348.4574349$$

Intervalo buscado: $IC(\mu) = [335.72, 348.46]$

Interpretación: *El intervalo encontrado brinda un 95% de contener al verdadero valor del parámetro, promedio del pago de electricidad consumida en los meses de invierno.*

Datos

```
N=c(820,540,440)
n=c(120,100,90)
media=c(290,352,427)
s=c(47,61,93)
```

Programa #3. Estimación de la media y el total poblacional en muestreo estratificado.

```
mstr.m=function(N,n,media,s)
{ Ntot=sum(N)
  f=n/N
  mestr=crossprod(N,media)/Ntot
  varm=(s^2/n)*(1-f)
  vstr=crossprod(N^2,varm)/Ntot^2
  setr=sqrt(vstr)
  a1=mestr-1.96*setr
  b1=mestr+1.96*setr
  a2=Ntot*a1
  b2=Ntot*b1
  cat("media: IC =  ",a1, "--",b1,"\n")
  cat("total: IC =  ",a2, "--",b2,"\n")
}
```

Aplicación del Programa #3.

```
> mstr.m(N,n,media,s)
media: IC =   335.7203 -- 348.4574
total: IC =   604296.6 -- 627223.4
```

Ejemplo2:
En el problema anterior hallar un intervalo del 95% de confianza para estimar el pago total de electricidad consumida en los meses de invierno.

Ejemplo3:
En una ciudad que tiene tres distritos se quiere conocer la proporción de hogares con alguna persona profesional. Se toman muestras aleatorias de esos hogares en cada un de los tres distritos y se obtienen los resultados que muestra la tabla

Distritos	N_i	n_i	Profesionales (éxitos)	Proporción
1	1200	180	80	0.4444
2	1350	190	50	0.2632
3	1050	140	45	0.3214

Datos

```
N=c(1200,1350,1050)
n=c(180,190,140)
exitos=c(80,50,45)
```

Programa #4. Estimación de la proporción en muestreo estratificado

```
mstr.p=function(N,n,exitos)
{ Ntot=sum(N)
  f=n/N
  p=exitos/n; q=1-p
  pestr=crossprod(N,p)/Ntot
  varp=(p*q/(n-1))*(1-f)
  vstr=crossprod(N^2,varp)/Ntot^2
  setr=sqrt(vstr)
  a=pestr-1.96*setr
  b=pestr+1.96*setr
  cat("proporción: IC =  ",a, "--",b,"\n")
}
```

Aplicación del Programa #4.

```
> mstr.p(N,n,exitos)
proporción: IC =   0.3028843 -- 0.3782804
```

Ejemplo4:
Una empresa tiene tres divisiones y los auditores están intentando estimar la cantidad total en facturas pendientes de cobro de la empresa. Hay un total de 870 facturas y en cada división hay 250, 300 y 320 facturas respectivamente. Una muestra aleatoria estratificada de 195 facturas contiene 60, 65 y 70 facturas tomadas desde las tres divisiones respectivamente. Usar los datos del archivo "**muestra2.xls**"

MUESTREO POR CONGLOMERADOS

La población $U = \{1, 2, \cdots, N\}$ de N elementos, está dividida en conglomerados $C_1, C_2, \cdots, C_M$ los cuales forman las unidades primarias de muestreo, cada uno de estos conglomerados está constituido por elementos de la población, unidades finales.

N = número de elementos en la población
M = número de conglomerados en la población
m = número de conglomerados en la muestra

El muestreo por conglomerados puede ser realizado en una etapa o en dos etapas, de la siguiente manera:

Muestreo por conglomerado de una etapa
Consiste en seleccionar aleatoriamente un cierto número de conglomerados (m), y dentro de cada conglomerado se realiza un censo de las unidades finales.

Muestreo por conglomerado de dos etapas
Consiste en seleccionar aleatoriamente un cierto número de conglomerados (m), y dentro de cada conglomerado se realiza un muestreo de las unidades finales.

En el muestreo por conglomerados en una y dos etapas se pueden presentar cualquiera de los dos siguientes casos:

Caso 1: Conglomerados de igual tamaño
Cada conglomerado $C_1, C_2, \cdots, C_M$ de la población tiene igual número de unidades primarias. Sea u el número de unidades en cada conglomerado, entonces se cumple que $M = N/u$ y por lo tanto $N = M \times u$

Caso 2: Conglomerados de diferente tamaño
Cada conglomerado $C_1, C_2, \cdots, C_M$ de la población tiene diferente número de unidades primarias. Se u_i el número de unidades en el conglomerado C_i para $i = 1, 2, \cdots, M$

MUESTREO POR CONGLOMERADO DE UNA ETAPA

ESTIMACION DE LA MEDIA Y DEL TOTAL POBLACIONAL: conglomerados de igual tamaño

Estimación del total de la poblacional: $\hat{T}$
Sean $t_1, t_2, \cdots, t_m$ los totales en cada conglomerado de la muestra de m conglomerados de u unidades cada uno.

Cuadro No. 1

Estimación de …	**fórmula**
media del total por conglomerado	$\bar{t} = \frac{1}{m}\sum_{i=1}^{m} t_i$
varianza de la media del total por conglomerado	$var\ \bar{t}\ = \frac{s_t^2}{m}\left(\frac{M-m}{M}\right)$
error estándar de la media del total por conglomerado	$se\ \bar{t}\ = \sqrt{var\ \bar{t}\ }$
total de la población	$\hat{T} = M\,\bar{t}$
varianza de la estimación del total poblacional	$\text{var}\ \hat{T}\ = M^2 \times \text{var}\ \bar{t}$
error estándar de la estimación del total poblacional	$se\ \hat{T}\ = M \times se\ \bar{t}$

Intervalo de confianza: $IC(T) = \hat{T} \mp 1.96\ se\ \hat{T}$

Estimación de la media poblacional: $\bar{x}_{cluster}$

Cuadro No. 2

Estimación de …	**fórmula**
media muestral	$\bar{x}_{cluster} = \frac{\hat{T}}{N} = \frac{M\,\bar{t}}{N} = \frac{\bar{t}}{u}$
varianza de la media muestral	$var\ \bar{x}_{cluster}\ = \frac{1}{u^2}\,\text{var}\ \bar{t}$
error estándar de la media muestral	$se\ \bar{x}_{cluster}\ = \frac{1}{u}\ se\ \bar{t}$

Intervalo de confianza: $IC(\mu) = \bar{x}_{cluster} \mp 1.96\ se(\bar{x}_{cluster})$

Ejemplo1
Una ciudad está dividida en 30 distritos escolares con cuatro escuelas elementales en cada una. Mediante muestreo por conglomerados se seleccionaron al azar 3 distritos escolares. Construya un intervalo del 95% de confianza para el total de niños con daltonismo en la cuidad.

Tabla No. 1

Distrito Escolar seleccionado	Escuela del distrito escolar	Total de niños en la escuela	Número de niños daltónicos por escuela
1	1	130	2
	2	150	3
	3	160	3
	4	120	5
2	1	110	2
	2	120	4
	3	100	0
	4	120	1
3	1	89	4
	2	130	2
	3	100	0
	4	150	2

Conglomerado: distrito escolar con 4 escuelas

Unidad elemental: escuela

Variable 1: total de niños en la escuela

Variable 2: total de niños daltónicos

Los datos de la Tabla No. 1, están resumidos en las tres primeras columnas de la Tabla No. 2

Tabla No. 2

Distrito Escolar seleccionado	Total de niños en el distrito escolar (u_i)	Total de niños daltónicos en el distrito escolar (t_i)	$(t_i - \hat{p} \times u_i)^2$
1	560	13	5.7515650
2	450	7	2.3081807
3	469	8	0.7725923
	1479	**28**	

Estimación del total de niños con daltonismo

1) media del total por conglomerado: $\bar{t} = 28/3 = 9.3333$

2) varianza del total por conglomerado: $s_t^2 = 10.3333$

2) varianza estimada de la media del total por conglomerado:

$$var\ \bar{t} = \frac{s_t^2}{m}\left(\frac{M-m}{M}\right) = \frac{10.3333}{3}\left(\frac{30-3}{30}\right) = 3.1$$

3) error estándar de la media del total por conglomerado: $se\ \bar{t} = 1.760682$

4) estimación puntual del total poblacional: $\hat{T} = M\,\bar{t} = 30 \times 9.333 \approx 280$

5) estimación por intervalos del total poblacional:

$$a = M\,\bar{t} - 1.96\,M\,se\ \bar{t} = 30 \times 9.3333 - 1.96 \times 30 \times 1.760682 = 176.47$$

$$b = M\,\bar{t} + 1.96\,M\,se\ \bar{t} = 30 \times 9.3333 + 1.96 \times 30 \times 1.760682 = 383.53$$

Estimación de la media de niños con daltonismo por unidad (escuela)

1) media de niños con daltonismo por unidad: $\bar{x}_{clus} = \bar{t}\,/\,4 = 9.3333/4 = 2.333325$

2) error estándar: $se(\bar{x}_{clus}) = se(\bar{t})/4 = 1.761/4 = 0.44025$

3) estimación por intervalos de la media poblacional

$$a = \bar{x}_{clus} - 1.96\,se\ \bar{x}_{clus} = 2.333325 - 1.96 \times 0.44025 = 1.471$$

$$b = \bar{x}_{clus} + 1.96\,se\ \bar{x}_{clus} = 2.333325 + 1.96 \times 0.44025 = 3.196$$

Programa #5.

```
cluster81=function(clus,dat,M)
{ #M : número de cluster en la población
  #u : número de unidades en el cluster
  m=max(clus)
  u=NROW(clus)/m
```

```
  datos=data.frame(clus,dat)
  t=rep(0,m)
   for(i in 1:m)
      {a=subset(datos,clus==i,select=dat)
       t[i]=sum(a)}

mediat=mean(t)
vart=var(t)
f=m/M
var.mt=(1-f)*vart/m
se.mt=sqrt(var.mt)

#Estimación del total poblacional
  T=M*mediat
  se.T=M*se.mt
a1=T-1.96*se.T
b1=T+1.96*se.T

#Estimación de la media poblacional
  media=mediat/u
  se.media=se.mt/u
a2=media-1.96*se.media
b2=media+1.96*se.media

cat("total: IC",a1,"--",b1,"\n")
cat("media: IC",a2,"--",b2,"\n")
}
```

Aplicación del Programa #5.

```
> cluster81(clusdal,daltonico,30)
total: IC 176.4719 -- 383.5281
media: IC 1.470599 -- 3.196067
```

Interpretación: *El intervalo encontrado* $IC(T) = (176.5, 383.5)$, *brinda un 95% de contener al verdadero valor del parámetro, número total de niños con daltonismo en la ciudad.*

El intervalo encontrado $IC(\mu) = (1.47, 3.20)$, *brinda un 95% de contener al verdadero valor del parámetro, promedio de niños con daltonismo por escuela.*

Ejemplo2

Se quiere estimar el GPA promedio de los estudiantes que viven en un hotel colegial. En vez de obtener una lista de todos los estudiantes del hotel y conducir un muestreo aleatorio simple, se observa que el hotel tiene 100 habitaciones con 4 estudiantes

alojados por habitación. Se elige aleatoriamente 5 de estas habitaciones y se pregunta por el GPA a cada estudiante de cada habitación.

Tabla No. 3

	GPA de los estudiantes					
Habitación	**Est.1**	**Est.2**	**Est.3**	**Est.4**	promedio	total
1	3.08	2.60	3.44	3.04	3.04	12.16
2	2.36	3.04	3.28	2.68	2.84	11.36
3	2.00	2.56	2.52	1.88	2.24	8.96
4	3.00	2.88	3.44	3.64	3.24	12.96
5	2.68	1.92	3.28	3.20	2.77	11.08
						56.52

Conglomerado: habitación con 4 estudiantes alojados

Unidad elemental: un estudiante

Variable: GPA de los estudiantes

M = 100, habitaciones (conglomerados)

u = 4, número de estudiantes por habitación

$N = M \times u = 400$, número total de estudiantes alojados en el hotel

Estimación del GPA promedio de los estudiantes del hotel estudiantil

1) media del total por conglomerado: $\bar{t} = 56.52/5 = 11.304$

2) varianza del total por conglomerado: $s_t^2 = 2.25568$

3) varianza estimada de la media del total por conglomerado:

$$var\ \bar{t} = \frac{s_t^2}{m}\left(\frac{M-m}{M}\right) = \frac{2.25568}{5}\left(\frac{100-5}{100}\right) = 0.428579$$

4) error estándar de la media del total por conglomerado: $se\ \bar{t} = 0.65466$

5) estimación puntual del promedio poblacional: $\bar{x}_{cluster} = \frac{\hat{T}}{N} = \frac{M\,\bar{t}}{N} = \frac{\bar{t}}{u} = \frac{11.304}{4} = 2.826$

6) estimación del error estándar de la media muestral

$$se\ \bar{x}_{cluster} = \frac{1}{u}\ se\ \bar{t} = \frac{1}{4} \times 0.65466 = 0.163665$$

7) estimación por intervalos de la media poblacional:

$$a = \bar{x}_{cluster} - 1.96\ se\ \bar{x}_{cluster} = 2.826 - 1.96 \times 0.163665 = 2.505217$$

$$b = \bar{x}_{cluster} + 1.96\ se\ \bar{x}_{cluster} = 2.826 + 1.96 \times 0.163665 = 3.146783$$

Aplicación del Programa #5

```
> cluster81(clusgpa,gpa,100)
total: IC 1002.087 -- 1258.713
media: IC 2.505217 -- 3.146783
```

Interpretación: *El intervalo encontrado* $IC(\mu) = (2.51, 3.15)$ *brinda un 95% de contener al verdadero valor del parámetro, GPA promedio de los estudiantes alojados en el hotel colegial*.

Ejemplo3:
El administrador de circulación de un nuevo periódico desea estimar el número promedio de periódicos comprados por los hogares de una comunidad. Los costos de viaje de hogar a hogar son sustanciales. Por lo tanto, los 4000 hogares en la comunidad son listados 400 conglomerados geográficos de 10 casas cada uno; una muestra aleatoria de 4 conglomerados es seleccionada.

Tabla No.7

	Número de periódicos comprados por hogar										
Cluster	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	Total
1	1	2	1	3	3	2	1	4	1	1	19
2	1	3	2	2	3	1	4	1	1	2	20
3	2	1	1	1	1	3	2	1	3	1	16
4	1	1	3	2	1	5	1	2	3	1	20

ESTIMACION DE LA MEDIA, TOTAL Y PROPORCION POBLACIONAL: conglomerados de diferente tamaño

Una idea sobre la distribución de los datos se da en la siguiente tabla

Tabla No. 4

Conglo-merado	Datos	Número de unidades en el conglomerado	total en el conglomerado	promedio en el conglomerado
1	$x_{11}, x_{12}, \cdots, x_{1,u_1}$	u_1	t_1	$\bar{x}_1$
2	$x_{21}, x_{22}, \cdots, x_{2,u_2}$	u_2	t_2	$\bar{x}_2$
3	$x_{31}, x_{32}, \cdots, x_{3,u_3}$	u_3	t_3	$\bar{x}_3$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
m	$x_{m1}, x_{m2}, \cdots, x_{m,u_m}$	u_m	t_m	$\bar{x}_m$

Las variables, número de unidades en el conglomerado (u_i) y el total en el conglomerado (t_i) están usualmente correlacionados positivamente.

Estimación de la media poblacional: $\bar{x}_{clu}$

Cuadro No. 3

Estimación de …	**fórmula**
media muestral	$\bar{x}_{clu} = \dfrac{\sum_{i=1}^{m} t_i}{\sum_{i=1}^{m} u_i}$
promedio de unidades por conglomerado	$\bar{u} = \dfrac{1}{m}\sum_{i=1}^{m} u_i$
varianza de la media muestral	$var\ \bar{x}_{clu} = \left(1 - \dfrac{m}{M}\right)\dfrac{1}{m\,\bar{u}^2}\dfrac{\sum_{i=1}^{m} t_i - u_i\,\bar{x}_{clu}^{\ 2}}{m-1}$
error estándar de la media muestral	$se\ \bar{x}_{clu} = \sqrt{\text{var}\ \bar{x}_{clu}}$

Ejemplo4
Considere una población de 187 salones de clase de un curso de algebra, de una ciudad. Un investigador coge una muestra aleatoria de 12 de estas clases y da un *test* para evaluar el conocimiento sobre el tema "funciones", se desea estimar la media del puntaje en dicho *test*. Las 12 clases seleccionadas fueron: 23, 37, 38, 39, 41, 44, 46, 51, 58, 62, 106, 108. Los datos son resumidos en el siguiente cuadro

Tabla No.5

Clases	número de estu-diantes por clase (u_i)	Puntaje total por clase (t_i)	Promedio obte-nido por clase $(\overline{x}_i)$	$t_i - u_i\,\overline{x}_{clu}{}^2$
1	20	1230	61.500	456.7298
2	26	1670	64.231	1867.7428
3	24	1402	58.417	9929.2225
4	34	1972	58.000	24127.7518
5	26	1508	58.000	14109.3082
6	28	1816	64.857	4106.2808
7	19	1048	55.158	19825.3937
8	32	2308	72.125	93517.3218
9	17	989	58.176	5574.9446
10	21	1398	66.571	7066.1174
11	26	1621	62.346	33.4386
12	26	1746	67.154	14212.7867
Total	**299**	**18708**	**---**	**194827.0387**

Conglomerado: salón de clase de un curso de álgebra con u_i estudiantes

Unidad elemental: un estudiante

Variable: puntaje en el examen de álgebra

M = 187, salones de clase (conglomerados)

m = 12, tamaño de muestra (salones de clase)

Estimación el promedio del puntaje del *test* de álgebra

1) media muestral del puntaje del *test*: $\bar{x}_{clu} = 18708/299 = 62.5686$

2) promedio de estudiantes por salón: $\bar{u} = 299/12 = 24.9167$

3) varianza del promedio del puntaje del *test* de álgebra

$$var\ \bar{x}_{clu} = \left(1 - \frac{m}{M}\right)\frac{1}{m\,\bar{u}^2}\frac{\sum_{i=1}^{m} t_i - u_i\,\bar{x}_{clu}\ ^2}{m-1}$$

$$var\ \bar{x}_{clu} = \left(1 - \frac{12}{187}\right)\times\frac{1}{12\times 24.9167^2}\times\frac{194827.0387}{12-1} = 2.224804503$$

4) error estándar del promedio del puntaje del *test* de álgebra: $se\ \bar{x}_{clu} = 1.491577857$

5) estimación por intervalos de la media poblacional:

$$a = \bar{x}_{clu} - 1.96\ se\ \bar{x}_{clu} = 62.5686 - 1.96\times 1.4916 = 59.6451$$
$$b = \bar{x}_{clu} - 1.96\ se\ \bar{x}_{clu} = 62.5686 - 1.96\times 1.4916 = 65.4921$$

Programa #6.

```
cluster83=function(mi,total,M)
{ #mi: número de unidades por cluster
  #M : número de cluster en la población
  #m : número de cluster en la muestra
  #mm: promedio de unidades por cluster

  m=NROW(ui)
  mm=mean(ui)
  media=sum(total)/sum(ui)
  aa=mi*media
  vart=crossprod(total-aa)/(m-1)
  f=m/M
  varmedia=(1-f)*(m*mm^2)^-1*vart
  se.media=sqrt(varmedia)

#Estimación de la media poblacional
  a1=media-1.96*se.media
  b1=media+1.96*se.media

cat("media: IC",a1,"--",b1,"\n")
}
```

Aplicación del Programa #6.

```
> cluster83(ui,total,187)
media =  62.56856 se.media =  1.491578
media: IC 59.64507 -- 65.49205
```

Interpretación: *El intervalo encontrado* $IC(\mu) = (59.65,\ 65.49)$, *brinda un 95% de contener al verdadero valor del parámetro, puntaje promedio del test de álgebra.*

Estimación del Total poblacional: $\hat{T}$

1) Si se conoce el número total de unidades en la población: N

Cuadro No. 4

Estimación de …	**fórmula**
total poblacional	$\hat{T} = N \times \bar{x}_{clus}$
error estándar	$se(\hat{T}) = N \times se(\bar{x}_{clus})$

$$IC(T) = \hat{T} \pm 1.96 \times se(\hat{T})$$

2) Si no se conoce el número total de unidades en la población (N). Siempre se conoce el número de conglomerados en la población: M

Cuadro No. 5

Estimación de …	**fórmula**
media del total en conglomerado	$\bar{t}_c = \dfrac{\sum_{i=1}^{m} t_i}{m}$
total poblacional	$\hat{T} = M \times \bar{t}_c$
varianza del total estimado	$var\ \hat{T} = M^2 \left(1 - \dfrac{m}{M}\right) \dfrac{1}{m} \dfrac{\sum_{i=1}^{m} t_i - \bar{t}_c^{\ 2}}{m-1}$
error estándar del total estimado	$se\ \hat{T} = \sqrt{\text{var}\ \hat{T}}$

$$IC(T) = \hat{T} \pm 1.96 \times se(\hat{T})$$

Estimación de la proporción de niños con daltonismo: Desde los datos de la Tabla No.2; el número de estudiantes por distrito escolar es tomado como el tamaño del conglomerado

Cuadro No. 6

Estimación de …	**fórmula**
proporción muestral	$\hat{p} = \frac{\sum_{i=1}^{m} t_i}{\sum_{i=1}^{m} u_i}$
promedio de unidades por conglomerado	$\bar{u} = \frac{1}{m}\sum_{i=1}^{m} u_i$
varianza de la media muestral	$var\ \hat{p} = \left(1-\frac{m}{M}\right)\frac{1}{m\,\bar{u}^2}\frac{\sum_{i=1}^{m} t_i - u_i\,\hat{p}^{\,2}}{m-1}$
error estándar de la media muestral	$se\ \hat{p} = \sqrt{\text{var}\ \hat{p}}$

$$IC(P) = \hat{p} \pm 1.96 \times se(\hat{p})$$

1) proporción estimada de niños con daltonismo: $\hat{p} = 28/1479 = 0.0189$

2) varianza estimada de la proporción:

$$var\ \hat{p} = \left(1-\frac{m}{M}\right)\times\frac{1}{m\,\bar{u}^2}\times\frac{\sum_{i=1}^{m}(t_i - u_i\,\hat{p})^2}{m-1} = \left(1-\frac{3}{30}\right)\times\frac{4.4162}{3\times 493^2} = 5.45096\text{E-}06$$

$$\bar{u} = \frac{1479}{3} = 493$$

3) error estándar estimado de la proporción: $se\ \hat{p} = 0.002334729$

4) estimación por intervalos de la proporción poblacional:

$$a = \hat{p} - 1.96\ se\ \hat{p} = 0.0189 - 1.96 \times 0.002334729 = 0.014355641$$

$$b = \hat{p} + 1.96\ se\ \hat{p} = 0.0189 + 1.96 \times 0.002334729 = 0.023507780$$

Aplicación del programa #6.

```
> cluster83(ui,tot,30)
media =  0.01893171 se.media =  0.002334729
media: IC 0.01435564 -- 0.02350778
```

Interpretación: *El intervalo encontrado* $IC(P) = (0.0144,\ 0.0235)$, *brinda un 95% de contener al verdadero valor del parámetro, proporción de niños con daltonismo en la ciudad.*

Ejemplo5:
Un sociólogo quiere estimar el ingreso promedio por familia en una cierta ciudad pequeña en la que no hay disponible una lista de residentes. **En este caso un muestreo por conglomerados es lo más adecuado**. La ciudad está formada por bloques rectangulares, excepto por dos áreas industriales y tres parques que contienen pocas casas. El sociólogo decide que cada bloque de la ciudad será considerado un conglomerado, las dos áreas industriales serán consideradas como un conglomerado y finalmente los tres parques serán considerados como un conglomerado.
Los conglomerados son numerados en un mapa de la ciudad del 1 al 415; se selecciona una muestra de 25 conglomerados, reportándose los siguientes datos:

Tabla No. 6

Cluster	Número de residentes	Ingreso total por cluster	Cluster	Número de residentes	Ingreso total por cluster
1	8	96000	14	10	49000
2	12	121000	15	9	53000
3	4	42000	16	3	50000
4	5	65000	17	6	32000
5	6	52000	18	5	22000
6	6	40000	19	5	45000
7	7	75000	20	4	37000
8	5	65000	21	6	51000
9	8	45000	22	8	30000
10	3	50000	23	7	39000
11	2	85000	24	3	47000
12	6	43000	25	8	41000
13	5	54000			

MUESTREO POR CONGLOMERADO CON PROBABILIDAD PROPORCIONAL AL TAMAÑO (*pps*, por siglas en inglés)

Es aplicado cuando los tamaños de los conglomerados son extremadamente diferenciados. Sea $\bar{x}_i$ el promedio en el conglomerado *i*, para $i = 1, 2, \cdots, m$

Cuadro No. 7

Estimación de …	**fórmula**
media poblacional	$\hat{\mu}_{pps} = \frac{1}{m}\sum_{i=1}^{m} \bar{x}_i$
varianza de la media estimada	$var\ \hat{\mu}_{pps} = \frac{1}{m} \times \frac{\sum_{i=1}^{m} \bar{x}_i - \hat{\mu}_{pps}^{\ 2}}{m-1}$
total poblacional	$\hat{T}_{pps} = N \times \hat{\mu}_{pps}$
varianza del total estimado	$var(\hat{T}) = N^2 \times var(\hat{\mu}_{pps})$

Ejemplo 6
Se desea muestrear registros de permisos por enfermedad de una empresa grande para estimar el número promedio de días de permisos por enfermedad por empleado sobre el pasado trimestre. La empresa tiene 8 divisiones, con 1200, 450, 2100, 860, 2840, 1910, 390 y 3200 empleados en cada división, respectivamente. Porque el número de días de permiso por enfermedad dentro de cada división puede ser altamente correlacionado con el número de empleados, se decide muestrear *m* = 3 divisiones, con **probabilidades proporcional al número de empleados**

Tabla No. 7

División	No. de empleados	Acumulado	Rango acumulado
1	1200	1200	0001 – 1200
2	450	1650	1201 – 1650
3	2100	3750	1651 – 3750
4	860	4610	3751 – 4610
5	2840	7450	4611 – 7450
6	1910	9360	7451 – 9360
7	390	9750	9361 – 9750
8	3200	12950	9751 – 12950

12950

Las dos primeras columnas de la Tabla No. 7 son los datos de empresa; para seleccionar una muestra de $m = 3$ divisiones, con probabilidad proporcional al tamaño de la división, se siguen los siguientes pasos:

1) Se construyen las dos últimas columnas de la Tabla No. 7.

2) Se elige una muestra aleatoria de 3 elementos entre los números 0001 – 12950. Los números seleccionados son: 2011, 7972 y 10281

3) El número 2011, se ubica en la división 3; el número 7972, se ubica en la división 6 y el número 10281, se ubica en la división 8. Las divisiones seleccionadas son: 3, 6 y 8.

Suponiendo que el total de días de permiso por enfermedad, en los conglomerados seleccionados son 4320, 4160 y 5790

Programa #7.

```
cluster89=function(ui,total)
{ m=NROW(ui)
  prom=total/ui
  m.pps=mean(prom)
  varm=var(prom)/m
  se.mpps=sqrt(varm)
a1=m.pps-1.96*se.mpps
b1=m.pps+1.96*se.mpps

  cat("media.pps = ",m.pps,"se.mpps = ",se.mpps,"\n")
  cat("media.pps: IC =  ",a1, "--",b1,"\n")
}
```

Aplicación del Programa #7.

```
> cluster89(ui,total,8)
media.pps =  2.014843 se.mpps =  0.1084973
media.pps: IC =   1.802188 -- 2.227498
```

Interpretación: *El intervalo encontrado* $IC(\mu) = (1.802,\ 2.227)$, *brinda un 95% de contener al verdadero valor del parámetro, promedio de días de permiso por enfermedad por empleado, en la empresa*.

MUESTREO POR CONGLOMERADO DE DOS ETAPAS

ESTIMACION INSESGADA DE LA MEDIA Y DEL TOTAL POBLACIONAL: conglomerados de diferente tamaño

Ejemplo 7
El gerente de una empresa de confección de ropa tiene 90 talleres que están ubicados en diferentes lugares de una ciudad, él quiere estimar el promedio del número de horas que las máquinas de coser estuvieron inactivas, esperando ser reparada. Se decidió usar el muestreo por conglomerados de dos etapas, considerando cada taller como un conglomerado de máquinas, en la primera etapa se selecciona $m = 10$ talleres y en la segunda etapa se seleccionan el 10% de las máquinas. El número total de máquinas, propiedad de la empresa es 4500.

Tabla No. 8

Taller	Total de máquinas	Máquinas en la muestra	Datos: tiempos de inactividad de las máquinas
1	50	10	5, 7, 9, 0, 11, 2, 8, 4, 3, 5
2	65	13	4, 3, 7, 2, 11, 0, 1, 9, 4, 3, 2, 1, 5
3	45	9	5, 6, 4, 11, 12, 0, 1, 8, 4
4	48	10	6, 4, 0, 1, 0, 9, 8, 4, 6, 10
5	52	10	11, 4, 3, 1, 0, 2, 8, 6, 5, 3
6	58	12	12, 11, 3, 4, 2, 0, 0, 1, 4, 3, 2, 4
7	42	8	3, 7, 6, 7, 8, 4, 3, 2
8	66	13	3, 6, 4, 3, 2, 2, 8, 4, 0, 4, 5, 6, 3
9	40	8	6, 4, 7, 3, 9, 1, 4, 5
10	56	11	6, 7, 5, 10, 11, 2, 1, 4, 0, 5, 4

Programa #8

```
cluster93=function(ui,clus,dat,N,M)
{ #N : número de unidades en la población
  #M : número de cluster en la población
  m=max(clus)
  ni=table(clus)

  datos=data.frame(clus,dat)
  ym=rep(0,m) ; s2=rep(0,m)
   for(i in 1:m)
      {a=subset(datos,clus==i,select=dat)
       ym[i]=mean(a) ; s2[i]=var(a)}

  mu=N/M
  media=crossprod(ui,ym)/(mu*m)
  aa=ui*ym
```

```
  bb=mu*media
  sb2=crossprod(aa-bb)/(m-1)

  cc=sum(ui*(ui-ni)*s2/ni)

  var.media=((M-m)/M)*(m*mu^2)^-1*sb2+(m*M*mu^2)^-1*cc
  se.media=sqrt(var.media)

#Estimación de la media poblacional
  a1=media-1.96*se.media
  b1=media+1.96*se.media

#Estimación del Total poblacional
  total=N*media
  se.total=N*se.media
  a2=total-1.96*se.total
  b2=total+1.96*se.total

  cat("media = ",media,"se.media = ",se.media,"\n")
  cat("total = ",total,"se.total = ",se.total,"\n")
  cat("\n")
  cat("media: IC =  ",a1, "--",b1,"\n")
  cat("total: IC =  ",a2, "--",b2,"\n")
}
```

Aplicación del Programa #8.

```
> cluster93(ui,clus,dat,4500,90)
media =  4.800359 se.media =  0.1925865
total =  21601.62 se.total =  866.639

media: IC =   4.422890 -- 5.177828
total: IC =   19903.00 -- 23300.23
```

Interpretación: *El intervalo encontrado* $IC(\mu) = (4.423,\ 5.178)$, *brinda un 95% de contener al verdadero valor del parámetro, promedio de horas que la máquina de coser estaba inactiva*.

El intervalo encontrado $IC(T) = (19903,\ 23300)$, *brinda un 95% de contener al verdadero valor del parámetro, total de horas que las máquinas de coser estaban inactivas.*

ESTIMACION DE RAZON DE LA MEDIA POBLACIONAL: conglomerados de diferente tamaño

Ejemplo 8
Hallar un intervalo de confianza para el estimador de razón de la media poblacional, usando los datos del Ejemplo 7.

Programa #9.

```
cluster94=function(ui,clus,dat,M)
{ #M : número de cluster en la población
  m=max(clus)
  ni=table(clus)

  datos=data.frame(clus,dat)
  ym=rep(0,m) ; s2=rep(0,m)
   for(i in 1:m)
      {a=subset(datos,clus==i,select=dat)
       ym[i]=mean(a) ; s2[i]=var(a)}

  qq=sum(ui)
  mu=mean(ui)
  mediar=crossprod(ui,ym)/qq
  aa=ui*ym
  bb=ui*mediar
  sr2=crossprod(aa-bb)/(m-1)

  cc=sum(ui*(ui-ni)*s2/ni)

  var.mediar=((M-m)/M)*(m*mu^2)^-1*sr2+(m*M*mu^2)^-1*cc
  se.mediar=sqrt(var.mediar)

#Estimación de razón de la media poblacional
  a1=mediar-1.96*se.mediar
  b1=mediar+1.96*se.mediar

  cat("mediar = ",mediar,"se.mediar = ",se.mediar,"\n")
  cat("mediar: IC =  ",a1, "--",b1,"\n")

}
```

Aplicación del Programa #9.

```
> cluster94(ui,clus,dat,90)
mediar =  4.598045 se.mediar =  0.2218872
mediar: IC =   4.163146 -- 5.032944
```

Interpretación: *El intervalo encontrado* $IC(\mu) = (4.163,\ 5.033)$, *brinda un 95% de contener al verdadero valor del parámetro, promedio de horas que la máquina de coser estaba inactiva*.

ESTIMACION DE LA PROPORCION POBLACIONAL: conglomerados de diferente tamaño

Ejemplo 9
Con la información del Ejemplo 7, se ha encontrado que la proporción muestral, por taller, de máquinas que requieren mayor reparación es significativa. Hallar un intervalo de confianza para la proporción poblacional

Tabla No. 9

Taller	Total de máquinas	Máquinas en la muestra	Proporción de máquinas que requieren mayor reparación
1	50	10	0.40
2	65	13	0.38
3	45	9	0.22
4	48	10	0.30
5	52	10	0.50
6	58	12	0.25
7	42	8	0.38
8	66	13	0.31
9	40	8	0.25
10	56	11	0.36

Programa #10.

```
cluster95=function(ui,ni,pi,M)
{ #ui: tamaño del cluster i
  #ni: tamaño de muestra dentro del cluster i
  #M : número de cluster en la población
  m=NROW(ui)
  qi=1-pi

  qq=sum(ui)
  mu=mean(ui)
  prop=crossprod(ui,pi)/qq
  aa=ui*pi
  bb=ui*prop
  sr2=crossprod(aa-bb)/(m-1)

  cc=sum(ui*(ui-ni)*pi*qi/(ni-1))

  var.prop=((M-m)/M)*(m*mu^2)^-1*sr2+(m*M*mu^2)^-1*cc
  se.prop=sqrt(var.prop)

#Estimación de la proporción poblacional
  a1=prop-1.96*se.prop
```

```
  b1=prop+1.96*se.prop

  cat("proporción = ",prop,"se.prop = ",se.prop,"\n")
  cat("proporción: IC =  ",a1, "--",b1,"\n")
}
```

Aplicación del Programa #10.

```
> cluster95(tam.clus,m.clus,pi,90)
proporción =  0.337318 se.prop =  0.02843061
proporción: IC =   0.281594 -- 0.393042
```

Interpretación: *El intervalo encontrado* $IC(P) = (0.282,\ 0.393)$, *brinda un 95% de contener al verdadero valor del parámetro, proporción de máquinas que requieren mayor reparación.*

ESTIMACION DE LA MEDIA Y TOTAL POBLACIONAL: conglomerados de igual tamaño

Cada conglomerado contiene *u* unidades; en este caso es común tomar muestras de igual tamaño (*n*) desde cada conglomerado seleccionado.

Ejemplo 10
Durante la temporada alta, el número de visitantes a un parque estatal fue registrado en la Tabla No. 10. Mediante un muestreo por conglomerado de dos etapas se seleccionó las semanas 2, 6 y 8; dentro de la semana 2 se seleccionó los días 2, 3 y 5; dentro de la semana 6 se seleccionó los días 1,3 y 6; dentro de la semana 8 se seleccionó los días 3, 4 y 6 .Hallar un intervalo de confianza para el número total de visitantes durante la temporada, el número total de visitantes por semana y el número total de visitantes por día.

Tabla No. 10

	Número de visitantes al parque					
Semana	Día 1	Día 2	Día 3	Día 4	Día 5	Día 6
1	200	150	130	140	150	190
2	120	105	111	103	111	130
3	310	200	180	130	125	208
4	200	107	101	98	103	137
5	170	160	130	121	107	114
6	250	237	209	212	231	180
7	380	378	325	330	306	331
8	495	400	315	302	350	395

9	206	200	108	95	107	190
10	308	300	293	206	200	300

Programa #11.

```
cluster96=function(clus,visitas,M,u)
{ #M : número de conglomerados en la población
  #u : número de unidades en el conglomerado

  m=max(clus)
  n=table(clus)[1]

  media=mean(visitas)
  ss=lm(visitas~clus)
  b=anova(ss)
  MSB=b$"Mean Sq"[1]
  MSW=b$"Mean Sq"[2]
  f1=m/M ; f2=n/u
  var.media=(1-f1)*MSB/(m*n)+(1-f2)*(1/M)*MSW/n
  se.media=sqrt(var.media)

#Total de visitantes por estación
 t1=media*M*u
 se.t1=se.media*M*u
 a1=t1-1.96*se.t1
 b1=t1+1.96*se.t1

#Total de visitantes por semana
 t2=media*u
 se.t2=se.media*u
 a2=t2-1.96*se.t2
 b2=t2+1.96*se.t2

# Total de visitantes por dia
 t3=media
 se.t3=se.media
 a3=t3-1.96*se.t3
 b3=t3+1.96*se.t3

cat("Total1: media = ",t1, "Std.Err = ",se.t1,"\n")
cat("Total2: media = ",t2, "Std.Err = ",se.t2,"\n")
cat("Total3: media = ",t3, "Std.Err = ",se.t3,"\n")
cat("\n")
cat("Total1_IC: ",a1,"--",b1,"\n")
cat("Total2_IC: ",a2,"--",b2,"\n")
cat("Total3_IC: ",a3,"--",b3,"\n")
}
```

Aplicación del Programa #11.

```
> cluster96(clus,visitas,10,6)
```

```
Total1: media =  13420 Std.Err =  2677.534
Total2: media =  1342 Std.Err =  267.7534
Total3: media =  223.6667 Std.Err =  44.62557

Total1_IC:  8172.033 -- 18667.97
Total2_IC:  817.2033 -- 1866.797
Total3_IC:  136.2005 -- 311.1328
```

Interpretación: *El intervalo encontrado* $IC(T_1) = (8172, 18668)$, *brinda un 95% de contener al verdadero valor del parámetro, total de visitantes durante la temporada*.

El intervalo encontrado $IC(T_2) = (817, 1867)$, *brinda un 95% de contener al verdadero valor del parámetro, promedio de visitantes por semana durante la temporada*.

El intervalo encontrado $IC(T_3) = (136, 311)$, *brinda un 95% de contener al verdadero valor del parámetro, promedio de visitantes diarios durante la temporada*.

MUESTREO POR CONGLOMERADO CON PROBABILIDAD PROPORCIONAL AL TAMAÑO (*pps*, por siglas en inglés)

Ejemplo 8

Se desea tomar una muestra de estudiantes de un curso introductorio de estadística, agrupados en 15 clases con la finalidad de estimar el promedio de horas de estudio dedicado el fin de semana previo al examen por estudiante. Se decidió tomar una muestra de $m = 5$ clases con reemplazo, con probabilidad proporcional al tamaño de la clase; en cada clase se elegirá una muestra de $n = 5$ estudiantes

Tabla No. 11

clase	No. de estudiantes	Acumulado	Rango acumulado
1	44	44	01 – 44
2	33	77	45 – 77
3	26	103	78 – 103
4	22	125	104 – 125
5	76	201	126 – 201
6	63	264	202 – 264

7	20	284	265 – 284
8	44	328	285 – 328
9	54	382	329 – 382
10	34	416	383 – 416
11	46	462	417 – 462
12	24	486	463 – 486
13	46	532	487 – 532
14	100	632	533 – 632
15	15	647	633 – 647
	647		

Las dos primeras columnas de la Tabla No. 11 son los datos de las clases de estadística; para seleccionar una muestra de $m = 5$ clases, con probabilidad proporcional al tamaño de la división, se siguen los siguientes pasos:

1) Se construyen las dos últimas columnas de la Tabla No. 11.

2) Se elige una muestra aleatoria de 5 elementos entre los números 01 – 647. Los números seleccionados son: 471, 612, 595, 189, 37

3) Según los números seleccionados y el rango acumulado de la tabla No. 11, las clases seleccionadas deben ser: 12, 14, 14, 5 y 1.

Suponiendo que el total de horas dedicadas a estudiar, en cada clase seleccionada es: 12.0, 8.0, 10.0, 14.0 y 18.5

Programa #12.

```
cluster97=function(ni,total,N)
{ m=NROW(total)
  prom=total/ni
  m.pps=mean(prom)
  varm=var(prom)/m
  se.mpps=sqrt(varm)
a1=m.pps-1.96*se.mpps
b1=m.pps+1.96*se.mpps

to=N*m.pps
se.to=N*se.mpps
a2=to-1.96*se.to
b2=to+1.96*se.to
```

```
  cat("media.pps = ",m.pps,"se.mpps = ",se.mpps,"\n")
  cat("total.pps = ",to,"se.total = ",se.to,"\n")
  cat("\n")
  cat("media.pps: IC =  ",a1, "--",b1,"\n")
  cat("total.pps: IC =  ",a2, "--",b2,"\n")
}
```

Aplicación del Programa #12.

```
> cluster97(5,total,647)
media.pps =  2.5 se.mpps =  0.3605551
total.pps =  1617.5 se.total =  233.2792

media.pps: IC =   1.793312 -- 3.206688
total.pps: IC =   1160.273 -- 2074.727
```

Interpretación: *El intervalo encontrado* $IC(\mu) = (1.79,\ 3.21)$, *brinda un 95% de contener al verdadero valor del parámetro, promedio de horas dedicadas a estudiar.*

El intervalo encontrado $IC(T) = (1160.27,\ 2074.73)$, *brinda un 95% de contener al verdadero valor del parámetro, total de horas dedicadas a estudiar.*

REFERENCIAS

Cochran, W. G., (1977), "Sampling Techniques", Thirds Edition, Wiley, Ney York.

Levy P. S., Lemeshow S., (1999), "Sampling of Populations, Methods and Applications", Thirds Edition, John Wiley & Sons, Inc.

Lohr, S. L., (2010), "Sampling: Design and Analysis", Second Edition, Thomson, Brooks/Cole Cengage Learning.

Newbold, P., Carlson, W., Thorne, B., (2008), "Estadística para Administración y Economía", Sexta Edición, Pearson Educación, S. A. Madrid, España.

Scheaffer, R., Mendenhall III, W., Lyman Ott, R., (2006) "Elementary Survey Sampling", Sixth Edition, Thomson, Brooks/Cole.