Tarea 4

Christian Badillo Luis Nuñez Luz Maria Santana Sealtiel Pichardo

Tabla de contenidos

1 Ejericicio 1

Un investigador en bosques quiere estimar el promedio de altura de árboles en cierta región. La región se divide en parcelas de un cuarto de acre. Se selecciona una muestra aleatoria simple de 20 parcelas de las 386 que componen la región. Todos los árboles en las parcelas muestreadas se miden, los resultados se muestran en la siguiente tabla:

Tabla 1: Datos Ejercicio 1.

Parcela (Cluster)	Número de árboles (M_i)	Altura Promedio (en pies)
1	42	6.2
2	51	5.8
3	49	6.7
4	55	4.9
5	47	5.2
6	58	6.9
7	60	6.3
8	52	6.7
9	61	5.9
10	49	6.1
11	57	6.0
12	63	4.9
13	43	4.3
14	59	5.2
15	48	5.7
16	41	6.1
17	45	5.3
18	46	6.7
19	62	6.1
20	58	7.0

Estime la altura promedio de los árboles en la región y dé un intervalo del 95% de confianza (El total para el cluster i se encuentra multiplicando M_i por el promedio del cluster).

Dado que no conocemos el número de árboles totales M, usaremos un estimador de razón para estimar el promedio por elemento, el cuál se estima cómo:

$$\hat{\bar{Y}}_e = \frac{\bar{Y}}{\hat{M}} = \frac{\sum_{i=1}^{n} M_i \bar{y}_i}{\sum_{i=1}^{n} M_i}$$

La altura promedio de los árboles en la región es de 5.909 pies. La varianza de nuestro estimador se estima como:

$$\hat{\mathbb{V}}(\hat{\bar{Y}}_e) = \left(1 - \frac{n}{N}\right) \frac{1}{n\hat{\bar{M}}^2} \sum_{i=1}^{n} \frac{M_i^2(\bar{y}_i - \hat{\bar{Y}}_e)^2}{n-1}$$

Donde: $\hat{\bar{M}} = \sum_{i=1}^{n} \frac{M_i}{n} = 52.3$, $n = 20$, y $N = 386$. Reemplazando se obtiene que la varianza del estimador $\hat{\mathbb{V}}(\hat{\bar{Y}}_e)$ es 0.02599. El intervalo de confianza al 95% para el promedio de altura de los árboles es [5.5930035, 6.2249697].

2 Ejercicio 2

Una empresa está considerando la revisión de sus políticas de retiro y sedea estimar la proporción de empleados que están a favor de una nueva política. La empresa tiene 87 plantas localizadas por todo el país. Ya que los resultados deben obtenerse rápidamente y con poco costo, se decidió usar muestreo de conglomerados con cada planta como conglomerado. Se seleccionó una m.a.s. de 15 plantas y se preguntó a todos los empleados de cada planta en muestra su opinión. Los datos son:

Tabla 2: Datos ejercicio 2

Planta	Número de empleados	Número de empleados a favor
1	51	42
2	62	53
3	49	40
4	73	45
5	101	63
6	48	31
7	65	38
8	49	30
9	73	54
10	61	45
11	58	51
12	52	29
13	65	46
14	49	37
15	55	42

Estime la proporción de empleados a favor de la nueva política de retiro y dé un intervalo del 95% de confianza para esta proporción.

Tenemos un muestreo de conglomerados donde cada planta es un conglomerado y fue seleccionado por medio de m.a.s, por lo tanto la proporción estimada será:

$$\hat{P} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} M_i}$$

La proporción de empleados a favor de la nueva política es 0.7091, que representa el 70.91% de los empleados.

Con varianza estimada:

$$\hat{V}(\hat{P}) = \left(1 - \frac{n}{N}\right) \frac{1}{n\hat{\bar{M}}^2} \sum_{i=1}^{n} \frac{(y_i \ - \ \hat{P}M_i)^2}{n-1}$$

Donde recordemos $\hat{\bar{M}}^2 = \left(\sum_{i=1}^{n} \frac{M_i}{n}\right)^2 = 3688.5378$, es la estimación del tamaño promedio de los conglomerados, $n = 15$ el tamaño de muestra, y $N = 87$ el total de conglomerados. Reemplazando se obtiene que la varianza del estimador $\hat{\mathbb{V}}(\hat{P})$ es 0.00058.

Por lo tanto el intervalo del 95% de confianza para la proporción de empleados a favor de la nueva política de retiro es $\hat{P} \pm z_{1-\alpha/2}\sqrt{V(\hat{P})} = [\ 0.6619, 0.7563\ ]$.

3 Ejercicio 3

Para una encuesta de los gastos de las familias en cierta ciudad, se utilizó un muestreo bietápico (para propósitos de este ejercicio se supone que en cada casa habita una sola familia). La ciudad tiene 118 manzanas, se seleccionaron por m.a.s. 5 de ellas. De cada manzana seleccionada se seleccionaron por m.a.s. 4 casas. Se preguntó a cada casa sobre aspectos financieros de la familia. Los datos son:

Tabla 3: Datos ejercicio 3

Id de Manzana (i)	Número de Casas en la Manzana (M_i)	Gasto promedio en alimentos durante el mes anterior de las familias seleccionadas ($\hat{y}_i$)
25	82	102
54	138	132
76	101	106
107	173	110
110	84	146

Estime el gasto promedio en alimentos durante el mes anterior de las familas en la ciudad. Obtenga un intervalo del 95% de confianza para este gasto promedio.

En este caso, tenemos un muestreo bietápico de tipo m.a.s.-m.a.s. del cual no se conoce el valor de M el número total de casas en la población. Por ello la estimación de la media, usando razón, es:

$$\hat{\bar{Y}}_e = \frac{\bar{Y}}{\hat{M}} = \frac{\sum_{i=1}^{n} M_i \bar{y}_i}{\sum_{i=1}^{n} M_i}$$

Reemplazando, $\hat{\bar{Y}}_e = 118.6505$. El gasto promedio estimado de alimentos del mes anterior en las familias de la ciudad es de $118.65

Para obtener el intervalo de confianza, el estimador de la varianza se calcula como:

$$\hat{\mathbb{V}}(\hat{\bar{Y}}_e) = \left(1 - \frac{n}{N}\right) \frac{1}{n\hat{\bar{M}}^2} \sum_{i=1}^{n} \frac{M_i^2(\bar{y}_i - \hat{\bar{Y}}_e)^2}{n-1} + \frac{1}{nN\hat{\bar{M}}^2} \sum_{i=1}^{n} M_i^2 \left(1 - \frac{m_i}{M_i}\right) \frac{\bar{S}_{wi}^2}{m_i}$$

Dado que no tenemos el valor de cada elemento (el gasto en alimentos del mes anterior de cada familia seleccionada) y tampoco se dan los datos acerca de la varianza dentro de cada unidad secundaria de muestreo, entonces para el cálculo de la varianza se despreciará la parte de la variación dentro de las casas de una misma manzana con la justificación de que la mayor parte de la variana la abarca el cálculo de la variación entre unidades secundarias de muestreo. Quedando la varianza como:

$$\hat{\mathbb{V}}(\hat{\bar{Y}}_e) = \left(1 - \frac{n}{N}\right) \frac{1}{n\hat{\bar{M}}^2} \sum_{i=1}^{n} \frac{M_i^2(\bar{y}_i - \hat{\bar{Y}}_e)^2}{n-1}$$

Donde: $\hat{\bar{M}} = \sum_{i=1}^{n} \frac{M_i}{n} =$, $n = 5$, $N = 118$, $m_i = 4, i = 1, \ldots, 5$.

Tabla 4: Cálculos

Id de Manzana (i)	Número de Casas en la Manzana (M_i)	Gasto promedio en alimentos durante el mes anterior de las familias seleccionadas ($\hat{y}_i$)	$(\hat{y}_i - \hat{\bar{Y}}_e)^2$	$M_i^2(\hat{y}_i - \hat{\bar{Y}}_e)^2$
25	82	102	277.2398	1864160
54	138	132	178.2086	3393805
76	101	106	160.0356	1632523
107	173	110	74.8315	2239631
110	84	146	747.9941	5277846

Reemplazando se tiene que, $\hat{\mathbb{V}}(\hat{Y}_e) = 51.6242168$. Y por tanto, su intervalo de confianza al 95% es: $[101.5177, 129.6823]$. Teniendo en cuenta que hay una subestimación del valor real debido a la despreciación de la varianza dentro de cada unidad secundaria de muestreo.

4 Ejercicio 4

Con la información del ejercicio 3, estime la proporción de familias que cuentan con internet dentro de su casa. Calcule un intervalo del 95% de confianza para esta proporción.

Tabla 5: Datos.

Id de Manzana (i)	Número de Casas en la Manzana (M_i)	Proporción familias con internet (p_i)
25	82	0.38
54	138	0.18
76	101	0.26
107	173	0.25
110	84	0.30

Se observa que tenemos un caso de muestreo bietapico siendo el muestreo aleatorio simple el método de selección en ambas etapas. Por tanto nuestra proporción estimada es:

$$\hat{P} = \frac{\sum_{i=1}^{n} M_i \hat{p}}{\sum_{i=1}^{n} M_i}$$

Reemplazando, $\hat{P} = 0.2607$. Es decir, el 26.07% de las familias de la ciudad cuentan con el servicio de conexión a Internet.

Tabla 6: Varianzas

Id de Manzana (i)	Número de Casas en la Manzana (M_i)	Proporción familias con internet (p_i)	$\frac{M_i^2(p_i-\hat{P})^2}{n-1}$	$M_i^2\left(1-\frac{m_i}{M_i}\right)\left(\frac{p_i(1-p_i)}{m_i-1}\right)$
25	82	0.38	23.9072	410.9569
54	138	0.18	31.0397	1188.1513
76	101	0.26	0.0014	629.4786
107	173	0.25	0.8637	1878.5410
110	84	0.30	2.7184	431.7746

La varianza de nuestro estimador puede ser calculada por la siguiente expresión:

$$\hat{\mathbb{V}}(\hat{P}) = \left(1-\frac{n}{N}\right)\frac{1}{n\hat{\bar{M}}^2}\frac{\sum_{i=1}^{n}(\hat{p}_i-\hat{P})^2}{n-1} + \frac{1}{nN\hat{\bar{M}}^2}\sum_{i=1}^{n} M_i^2\left(1-\frac{m_i}{M_i}\right)\left(\frac{\hat{p}_i(1-\hat{p}_i)}{m_i-1}\right)$$

Donde: $\hat{\bar{M}} = \sum_{i=1}^{n}\frac{M_i}{n} = 115.6$, $n = 5$, $N = 118$, $m_i = 4, i = 1, \ldots, 5$.

Reemplazando se tiene que, $\hat{\mathbb{V}}(\hat{P}) = 0.0014146$. Y por tanto, su intervalo de confianza al 95% es: $[0.187, 0.3345]$.

5 Ejercicio 5

La tabla siguiente muestra gastos de consumo personal en el país para una selección de bienes y servicios (en miles de millones de pesos). Seleccione una muestra de 3 categorías con probabilidades proporcionales a los gastos de 2023:

1. Utilizando el algoritmo con reemplazo.
2. Utilizando el algoritmo ppt sistemático.
3. Con las categorías en muestra de incisos 1 y 2 estime el gasto total en 2024 (suponiendo que no lo conoce). Obtenga un intervalo del 95% de confianza par este gasto total.

Tabla 7: Datos Ejercicio 5.

Categoría	2023	2024
Vehículos automotores	116.8	26.4
Muebles y enseres domésticos	107.3	107.5
Alimentos	432.3	456.4
Ropa	132.6	136.9
Gasolina y aceite	108.8	105.2
Combustible y carbón	23.8	23.0
Vivienda	347.3	384.2
Funcionamiento de la vivienda	147.7	165.9
Transporte	75.3	78.7

Para realizar el muestro de probabilidad proporcional al tamaño, se usará la librería `pps` de `R`.

```
library(pps)
# La probabilidad de extracción es de:
prob_ex <- data[2] / sum(data[2])
prob_ex
```

```
       2023
1 0.07828943
2 0.07192171
3 0.28976473
4 0.08887995
5 0.07292714
6 0.01595281
7 0.23279040
8 0.09900127
9 0.05047255
```

```
# La "probabilidad" de inclusión es:
n <- 3
```

```r
prob_in <- n*prob_ex
prob_in
```

```
       2023
1 0.23486829
2 0.21576513
3 0.86929419
4 0.26663986
5 0.21878142
6 0.04785844
7 0.69837120
8 0.29700382
9 0.15141766
```

Dado que ninguna tiene una "probabilidad" de inclusión mayor o igual que 1, no se tendrán categorías autorepresentadas.

```r
# Se añaden estos datos a los originales.
data["prob.ext"] <- prob_ex
data["prob.inc"] <- prob_in
data
```

```
                     Categoría  2023  2024   prob.ext   prob.inc
1         Vehículos automotores 116.8  26.4 0.07828943 0.23486829
2  Muebles y enseres domésticos 107.3 107.5 0.07192171 0.21576513
3                    Alimentos 432.3 456.4 0.28976473 0.86929419
4                         Ropa 132.6 136.9 0.08887995 0.26663986
5            Gasolina y aceite 108.8 105.2 0.07292714 0.21878142
6          Combustible y carbón  23.8  23.0 0.01595281 0.04785844
7                      Vivienda 347.3 384.2 0.23279040 0.69837120
8 Funcionamiento de la vivienda 147.7 165.9 0.09900127 0.29700382
9                    Transporte  75.3  78.7 0.05047255 0.15141766
```

5.1 PPT no sistemático.

Para el muestro de probabilidad proporcional al tamaño no sistemático se usa la función `ppswr()` de `R`.

```r
 # Semilla para garantizar reproducibilidad.
set.seed(14082001)
 # indices de los registros.
idx_ns <- ppswr(data$`2023`, n)
idx_ns
```

```
[1] 3 5 1
```

Entonces nuestra muestra es:

Tabla 8: Muestra PPT No Sistemático.

	Categoría	2023	2024
3	Alimentos	432.3	456.4
5	Gasolina y aceite	108.8	105.2
1	Vehículos automotores	116.8	26.4

Para estimar el total del gasto en 2024, se usará la libreria `survey` de `R`.

```
library(survey, warn.conflicts = F)
dppt <- svydesign(id=~1,probs=~prob.inc,data=muestra)
summary(dppt)
```

```
Independent Sampling design (with replacement)
svydesign(id = ~1, probs = ~prob.inc, data = muestra)
Probabilities:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.2188  0.2268  0.2349  0.4410  0.5521  0.8693
Data variables:
[1] "Categoría" "2023"      "2024"      "prob.ext"  "prob.inc"
```

Estimamos el total:

```
tot2024 <- svytotal(~`2024`, dppt, deff=T)
CI <- confint(tot2024)
```

Tabla 9: Total Estimado con PPT No Sistemático.

	Total Estimado	SE	DEff
2024	1118.272	392.4006	0.2628894

Con un intervalo de confianza de [349.1811541, 1887.3634184].

5.2 PPT Sistemático.

Para el muestro de probabilidad proporcional al tamaño sistemático se usa la función `ppss()` de `R`.

```
 # Semilla para garantizar reproducibilidad.
set.seed(14082001)
 # indices de los registros.
idx_s <- ppss(data$`2023`, n)
idx_s
```

```
[1] 2 3 7
```

Nuestra muestra es:

Tabla 10: Muestra PPT Sistemático.

	Categoría	2023	2024
2	Muebles y enseres domésticos	107.3	107.5
3	Alimentos	432.3	456.4
7	Vivienda	347.3	384.2

Para estimar el total del gasto en 2024, se usará la libreria `survey` de `R`.

```
dppts <- svydesign(id=~1,probs=~prob.inc,data=muestra.s)
summary(dppt)
```

```
Independent Sampling design (with replacement)
svydesign(id = ~1, probs = ~prob.inc, data = muestra)
Probabilities:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.2188  0.2268  0.2349  0.4410  0.5521  0.8693
Data variables:
[1] "Categoría" "2023"      "2024"      "prob.ext"  "prob.inc"
```

Estimamos el total:

```
tot2024.s <- svytotal(~`2024`, dppts, deff=T)
CI.s <- confint(tot2024.s)
```

Tabla 11: Total Estimado con PPT Sistemático.

	Total Estimado	SE	DEff
2024	1573.388	44.96351	0.0059363

Con un intervalo de confianza de [1485.2609418, 1661.514668].

El gasto total real es: 1484.2 que para nuestro caso es incluido en el intervalo de confianza del PPT no sistemático pero no incluido en el PPT sistemático.