Estimación basada en muestras probabilísticas

Tarea Examen 1. Tipo A

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

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

1. **Propiedades de $\widehat{t}_{y\text{WR}}$**

   Considere el estimador de Hansen y Hurwitz (1943)

   $$\widehat{t}_{y\text{WR}} = \frac{1}{m}\sum_{i=1}^{m}\frac{y_{k_i}}{p_{k_i}}. \tag{1}$$

   Usando la siguiente variable para cada elemento $k \in U$:

   a_k : número de veces que el elemento k aparece en la muestra ordenada $os = \{k_1, ..., k_m\}$.

   Demuestre que el estimador $\widehat{t}_{y\text{WR}}$ es insesgado para $t_y = \sum_{k=1}^{N} y_k$. Además que

   $$\text{V}(\widehat{t}_{y\text{WR}}) = \frac{1}{m}\sum_{k=1}^{N}\left(\frac{y_k}{p_k} - t_y\right)^2 p_k. \tag{2}$$

   Puede usar que $\text{E}(a_k) = mp_k$, $\text{V}(a_k) = mp_k(1-p_k)$ y $\text{Cov}(a_k, a_l) = -mp_kp_l,\ \forall\ k \neq l,\ k, l \in U$.

   Solución.

   Tenemos que esta variable puede tomar valores entre 0 y m, además, dado que los experimentos se realizan de forma independiente y para cada elemento k se tiene que la aparición de éste en el experimento i se puede describir como un experimento Bernoulli(p_k), entonces a_k tiene distribución Binomial(m, p_k).

   De igual manera tenemos que $\text{E}(a_k) = mp_k$, $\text{V}(a_k) = mp_k(1-p_k)$ y $\text{Cov}(a_k, a_l) = -mp_kp_l,\ \forall\ k \neq l,\ k, l \in U$.

   $$\Longrightarrow \widehat{t}_{y\text{WR}} = \frac{1}{m}\sum_{i=1}^{m}\frac{y_{k_i}}{p_{k_i}} = \sum_{k=1}^{N} a_k\frac{y_k}{mp_k}$$

$$\begin{aligned}
\implies \mathbb{E}\left[\widehat{t}_{y_{\text{WR}}}\right] &= \mathbb{E}\left[\sum_{k=1}^{N} a_k \frac{y_k}{mp_k}\right] \\
&= \sum_{k=1}^{N} \mathbb{E}\left[a_k \frac{y_k}{mp_k}\right] = \sum_{k=1}^{N} \mathbb{E}\left[a_k\right] \frac{y_k}{mp_k} \\
&= \sum_{k=1}^{N} mp_k \frac{y_k}{mp_k} = \sum_{k=1}^{N} y_k \\
&= t_y \quad \square
\end{aligned}$$

$$\begin{aligned}
\implies \mathbb{V}ar(\widehat{t}_{y_{\text{WR}}}) &= Var\left(\sum_{k=1}^{N} a_k \frac{y_k}{mp_k}\right) + Cov\left(\sum_{k=1}^{N} a_k \frac{y_k}{mp_k}, \sum_{l=1;l\neq k}^{N} a_l \frac{y_l}{mp_l}\right) \\
&= \sum_{k=1}^{N} Var\left(a_k \frac{y_k}{mp_k}\right) + \sum_{k=1}^{N} \sum_{l=1;l\neq k}^{N} Cov\left(a_k \frac{y_k}{mp_k}, a_l \frac{y_l}{mp_l}\right) \\
&= \sum_{k=1}^{N} \left(\frac{y_k}{mp_k}\right)^2 Var(a_k) + \sum_{k=1}^{N} \sum_{l=1;l\neq k}^{N} \frac{y_k}{mp_k} \frac{y_l}{mp_l} Cov(a_k, a_l) \\
&= \sum_{k=1}^{N} \frac{y_k^2}{m^2 p_k} mp_k(1-p_k) + \sum_{k=1}^{N} \sum_{l=1;l\neq k}^{N} \frac{y_k}{mp_k} \frac{y_l}{mp_l}(-mp_k p_l) \\
&= \sum_{k=1}^{N} \frac{y_k^2}{mp_k}(1-p_k) - \sum_{k=1}^{N} \sum_{l=1;l\neq k}^{N} \frac{y_k}{mp_k} y_l p_k \\
&= \sum_{k=1}^{N} \frac{y_k}{mp_k}\left(y_k(1-p_k) - \sum_{l=1;l\neq k}^{N} y_l p_k\right) = \frac{1}{m}\sum_{k=1}^{N}\left(\frac{y_k^2}{p_k}(1-p_k) - y_k \sum_{l=1;l\neq k}^{N} y_l\right) \\
&= \frac{1}{m}\sum_{k=1}^{N}\left(\frac{y_k^2}{p_k} - y_k^2 - y_k \sum_{l=1;l\neq k}^{N} y_l\right) = \frac{1}{m}\sum_{k=1}^{N}\left(\frac{y_k^2}{p_k} - y_k \sum_{l=1}^{N} y_l\right) = \frac{1}{m}\sum_{k=1}^{N}\left(\frac{y_k^2}{p_k} - y_k t_y\right) \\
&= \frac{1}{m}\left(\sum_{k=1}^{N} \frac{y_k^2}{p_k} - \sum_{k=1}^{N} y_k t_y\right) = \frac{1}{m}\left(\sum_{k=1}^{N} \frac{y_k^2}{p_k} - t_y t_y\right) = \frac{1}{m}\left(\sum_{k=1}^{N} \frac{y_k^2}{p_k} - t_y^2 + t_y^2 - t_y^2\right) \\
&= \frac{1}{m}\left(\sum_{k=1}^{N} \frac{y_k^2}{p_k} - 2t_y^2 + t_y^2\right) = \frac{1}{m}\left(\sum_{k=1}^{N} \frac{y_k^2}{p_k} - 2t_y \sum_{k=1}^{N} y_k + t_y^2 \sum_{k=1}^{N} p_k\right) \\
&= \frac{1}{m}\left(\sum_{k=1}^{N} \frac{y_k^2}{p_k} - \sum_{k=1}^{N} 2t_y y_k + \sum_{k=1}^{N} t_y^2 p_k\right) = \frac{1}{m}\sum_{k=1}^{N}\left(\frac{y_k^2}{p_k} - 2t_y y_k + t_y^2 p_k\right) \\
&= \frac{1}{m}\sum_{k=1}^{N}\left(\left(\frac{y_k}{p_k}\right)^2 - 2t_y \frac{y_k}{p_k} + t_y^2\right) p_k = \frac{1}{m}\sum_{k=1}^{N}\left(\frac{y_k}{p_k} - t_y\right)^2 p_k \quad \square
\end{aligned}$$

2. **Estimadores insesgados bajo un diseño muestral probabilístico**

Considere el estimador

$$\widehat{\theta} = \frac{1}{Nn}\left[\sum_{k\in S} y_k + \frac{N-1}{n-1}\sum_{k\in S}\sum_{\substack{l\in S\\ l\neq k}} y_l\right]. \tag{3}$$

¿Para qué parámetro θ, que es una estadística descriptiva muy usada, es $\widehat{\theta}$ un estimador insesgado considerando un $m.a.s$.?

Solución.

Recordemos que para un $m.a.s$ las probabilidades de inclusión son:

$$\pi_k = \frac{n}{N} \text{ y } \pi_{kl} = \frac{n(n-1)}{N(N-1)}$$

Si $\widehat{\theta}$ es insesgado se satisface que $\mathbb{E}[\widehat{\theta}] = \theta$

$$\Rightarrow \mathbb{E}[\widehat{\theta}] = \mathbb{E}\left[\frac{1}{Nn}\left(\sum_{k\in S} y_k + \frac{N-1}{n-1}\sum_{k\in S}\sum_{\substack{l\in S\\ l\neq k}} y_l\right)\right]$$

$$= \frac{1}{Nn}\mathbb{E}\left[\sum_{k\in S} y_k + \frac{N-1}{n-1}\sum_{k\in S}\sum_{\substack{l\in S\\ l\neq k}} y_l\right]$$

Veamos que la doble suma puede ser expresada de una manera más sencilla, la suma interna en la que $l \in S$ y $l \neq k$ se puede expresar como una suma sobre todos los elementos en S y solo restando un y_k porque $l \neq k$, de esta manera:

$$\sum_{\substack{l\in S\\ l\neq k}} y_l = \sum_{l\in S} y_l - y_k$$

Así:

$$\sum_{k\in S}\sum_{\substack{l\in S\\ l\neq k}} y_l = \sum_{k\in S}\left(\sum_{l\in S} y_l - y_k\right)$$

$$= \sum_{k\in S}\sum_{l\in S} y_l - \sum_{k\in S} y_k$$

$$= n\sum_{l\in S} y_l - \sum_{k\in S} y_k$$

Entonces:

$$
\begin{aligned}
\mathbb{E}[\widehat{\theta}] &= \frac{1}{Nn}\mathbb{E}\left[\sum_{k\in S} y_k + \frac{N-1}{n-1}\left(n\sum_{l\in S} y_l - \sum_{k\in S} y_k\right)\right] \\
&= \frac{1}{Nn}\mathbb{E}\left[\sum_{k\in S} y_k + \frac{N-1}{n-1}\left(n\sum_{k\in S} y_k - \sum_{k\in S} y_k\right)\right] \\
&= \frac{1}{Nn}\mathbb{E}\left[\sum_{k\in S} y_k + \frac{N-1}{n-1}\left((n-1)\sum_{k\in S} y_k\right)\right] \\
&= \frac{1}{Nn}\mathbb{E}\left[\sum_{k\in S} y_k + (N-1)\sum_{k\in S} y_k\right] \\
&= \frac{1}{Nn}\mathbb{E}\left[(N-1+1)\sum_{k\in S} y_k\right] = \frac{1}{Nn}\mathbb{E}\left[N\sum_{k\in S} y_k\right] \\
&= \frac{N}{Nn}\mathbb{E}\left[\sum_{k\in S} y_k\right] \\
&= \frac{1}{n}\mathbb{E}\left[\sum_{k\in U} y_k I_k\right] \\
&= \frac{1}{n}\sum_{k\in U}\mathbb{E}[y_k I_k] \\
&= \frac{1}{n}\sum_{k\in U} y_k\mathbb{E}[I_k] \\
&= \frac{1}{n}\sum_{k\in U} y_k\pi_k \\
&= \frac{1}{n}\sum_{k\in U} y_k\frac{n}{N} \\
&= \frac{n}{Nn}\sum_{k\in U} y_k \\
&= \frac{1}{N}\sum_{k\in U} y_k \\
&= \bar{y}_U
\end{aligned}
$$

Por lo tanto para $\theta = \bar{y}_U$, $\widehat{\theta}$ un estimador insesgado considerando un $m.a.s.$

3. **Muestreo complementario**

Considere una población de tamaño N, $U = \{1, 2, ..., N\}$, y una muestra s_1 de tamaño fijo n_1 seleccionada usando un $m.a.s$. Se decide complementar la muestra s_1 selecionando una muestra s_2 de tamaño fijo n_2 del conjunto de elementos $U \setminus s_1$ usando un $m.a.s.$, $n_1 + n_2 \leq N$. Es decir, la muestra final s corresponde a $s = s_1 \cup s_2$.

¿Cuál es el conjunto de muestras posibles $\mathcal{S}_0$ asociado a s? Dé el diseño de muestreo asociado a s, es decir, indique:

$$p(s) = P(S = s) \quad \forall \quad s \in \mathcal{S}_0.$$

Solución.

Sea $s_1 = n_1$ y sea $s_2 = n_2$, primero obtenemos la muestra $s_1 = n_1$ que como sabemos es un $m.a.s$ por lo tanto tiene $p(s_1) = \frac{1}{\binom{N}{n_1}}$.

Ahora como queremos completar esa muestra con $s_2 = n_2$, es decir sabemos que $_2 = n_2$ esta condicionado ya que solo pueden salir los elementos que no salieron en $s_1 = n_1$, por lo tanto $s_1 \cap s_2 = \emptyset$, así podemos usar que $s = s_1 \cup s_2$.

Así por Teorema de Probabilidad Total tenemos que

$$P(S = s) = \sum_{s_1 \in S | s_1 = n_1} P(S_2 = s - s_1 \mid S_1 = s_1) P(S_1 = s_1)$$

Como podemos ver estamos condicionando a que ya salio s_1 y por lo tanto no puede salir un elemento de s_1, es decir queremos sacar muestras de tamaño n_2 del conjunto de interes menos s_1 que es la muestra que ya habíamos sacado, pero recordemos que es un $m.a.s$, por lo tanto tenemos que

$$P(S_2 = s - s_1 \mid S_1 = s_1) = \frac{1}{\binom{N - n_1}{n_2}}$$

Analogamente sabiamos que

$$P(S_1 = s_1) = \frac{1}{\binom{N}{n_1}}$$

Nos falta ver las combinaciones posibles de sacar la muestra s_1 del conjunto de interes total, es decir de $S = s_1 \cup s_2$, por lo tanto $n_s = n_1 + n_2$, así obtenemos que las combinaciones posibles son

$$\binom{n_1 + n_2}{n_1}$$

Sustituyendo

$$P(S = s) = \sum_{s_1 \in S | s_1 = n_1} P(S_2 = s - s_1 \mid S_1 = s_1) P(S_1 = s_1)$$

$$= \binom{n_1+n_2}{n_1} \frac{1}{\binom{N-n_1}{n_2}} \frac{1}{\binom{N}{n_1}}$$

$$= \frac{(n_1+n_2)!}{(n_1+n_2-n_1)!(n_1)!} \frac{(N-(n_1+n_2))!n_2!}{(N-n_1)!} \frac{(N-n_1)!n_1!}{N!}$$

Cancelando términos tenemos que

$$P(S=s) = \frac{(n_1+n_2)!(N-(n_1+n_2))!}{N!}$$

$$P(S=s) = \frac{(n_1+n_2)!(N-(n_1+n_2))!}{N!} = \frac{1}{\binom{N}{n_1+n_2}}$$

Por lo tanto S_0 lo podemos describir como

$$S_0 = \{\{S = s_1 \cup s_2\} = \{1, ..., n_1\} \cup \{m, ..., n_2\} = s\binom{N}{n_1+n_2}\}$$

4. **Estimación de un parámetro que es función de totales**

Suponga que el parámetro de interés es la media geométrica calculada para la variable y en una población $U = \{1, 2, ..., N\}$, es decir,

$$\theta = \left(\prod_{k \in U} y_k \right)^{\frac{1}{N}}.$$

1. Muestre que el parámetro θ se puede escribir como una función de totales.
Solución.

$$\theta = \left(\prod_{k \in U} y_k \right)^{\frac{1}{N}} \iff Ln(\theta) = Ln\left(\left(\prod_{k \in U} y_k \right)^{\frac{1}{N}} \right) \iff e^{Ln(\theta)} = exp\left\{ Ln\left(\left(\prod_{k \in U} y_k \right)^{\frac{1}{N}} \right) \right\}$$

$$\iff \theta = exp\left\{ \frac{1}{N} Ln\left(\prod_{k \in U} y_k \right) \right\}$$

$$= exp\left\{ \frac{1}{N} \sum_{k \in U} Ln(y_k) \right\} \quad \text{sea } z_k = Ln(y_k)$$

$$= exp\left\{ \frac{1}{N} \sum_{k \in U} z_k \right\} = exp\left\{ \frac{t_z}{N} \right\}$$

2. Suponga que N es desconocido y se estimará. Proporcione un estimador $\widehat{\theta}$ basado en estimadores HT que sea aproximadamente insesgado para θ.
Solución.
Como tenemos una comosición de funciones derivables podemos proponer el siguinete estimador:

$$\widehat{\theta} = exp\left\{ \frac{1}{\widehat{N}} \sum_{k \in S} \frac{z_k}{\pi_k} \right\} = exp\left\{ \frac{1}{\sum_{k \in S} \frac{1}{\pi_k}} \sum_{k \in S} \frac{z_k}{\pi_k} \right\} = exp\left\{ \frac{\hat{t}_{z\pi}}{\hat{t}_{x\pi}} \right\}$$

3. Dé la expresión de la varianza teórica o poblacional aproximada para el estimador dado en 2, es decir, $\mathrm{V}(\widehat{\theta})$.

Solución.

Primero obtenemos las variables de linealización

$$
\begin{aligned}
a_1 &= \left.\frac{\partial f(\hat{t}_{z\pi}, \hat{t}_{x\pi})}{\partial \hat{t}_{z\pi}}\right|_{(\hat{t}_{z\pi},\hat{t}_{x\pi})=(t_z,t_x)} \\
&= \left.\frac{\partial\, exp\left\{\frac{\hat{t}_{z\pi}}{\hat{t}_{x\pi}}\right\}}{\partial \hat{t}_{z\pi}}\right|_{(\hat{t}_{z\pi},\hat{t}_{x\pi})=(t_z,t_x)} \\
&= exp\left\{\frac{\hat{t}_{z\pi}}{\hat{t}_{x\pi}}\right\}\frac{1}{\hat{t}_{x\pi}}\Bigg|_{(\hat{t}_{z\pi},\hat{t}_{x\pi})=(t_z,t_x)} \\
&= \frac{1}{t_x} exp\left\{\frac{t_z}{t_x}\right\}
\end{aligned}
$$

$$
\begin{aligned}
a_2 &= \left.\frac{\partial f(\hat{t}_{z\pi}, \hat{t}_{x\pi})}{\partial \hat{t}_{x\pi}}\right|_{(\hat{t}_{z\pi},\hat{t}_{x\pi})=(t_z,t_x)} \\
&= \left.\frac{\partial\, exp\left\{\frac{\hat{t}_{z\pi}}{\hat{t}_{x\pi}}\right\}}{\partial \hat{t}_{x\pi}}\right|_{(\hat{t}_{z\pi},\hat{t}_{x\pi})=(t_z,t_x)} \\
&= -exp\left\{\frac{\hat{t}_{z\pi}}{\hat{t}_{x\pi}}\right\}\frac{\hat{t}_{z\pi}}{\hat{t}^2_{x\pi}}\Bigg|_{(\hat{t}_{z\pi},\hat{t}_{x\pi})=(t_z,t_x)} \\
&= -\frac{t_z}{t_x^2} exp\left\{\frac{t_z}{t_x}\right\}
\end{aligned}
$$

A partir de lo anterior, la variable de linealización v_k y su versión estimada $\widehat{v}_k$ se definen como

$$
\begin{aligned}
v_k &= a_1 z_k + a_2 x_k \\
&= \frac{z_k}{t_x} exp\left\{\frac{t_z}{t_x}\right\} - x_k \frac{t_z}{t_x^2} exp\left\{\frac{t_z}{t_x}\right\} \\
&= \frac{z_k}{t_x} exp\left\{\frac{t_z}{t_x}\right\} - \frac{t_z}{t_x^2} exp\left\{\frac{t_z}{t_x}\right\} \\
&= \frac{1}{t_x} exp\left\{\frac{t_z}{t_x}\right\}\left(z_k - \frac{t_z}{t_x}\right)
\end{aligned}
$$

$$
\begin{aligned}
\therefore \widehat{v}_k &= \frac{z_k}{\hat{t}_{x\pi}} exp\left\{\frac{\hat{t}_{z\pi}}{\hat{t}_{x\pi}}\right\} - \frac{\hat{t}_{z\pi}}{\hat{t}^2_{x\pi}} exp\left\{\frac{\hat{t}_{z\pi}}{\hat{t}_{x\pi}}\right\} \\
&= \frac{z_k}{\hat{t}_{x\pi}}\widehat{\theta} - \frac{\hat{t}_{z\pi}}{\hat{t}^2_{x\pi}}\widehat{\theta} \\
&= \frac{1}{\hat{t}_{x\pi}}\widehat{\theta}\left(z_k - \frac{\hat{t}_{z\pi}}{\hat{t}_{x\pi}}\right)
\end{aligned}
$$

Obteniendo así la varianza:

$$
\begin{aligned}
V(\widehat{\theta}) &= \sum_{k\in U}\sum_{l\in U} \Delta_{kl}\frac{v_k}{\pi_k}\frac{v_l}{\pi_l} \qquad \text{con } \Delta_{kl} = \pi_{kl} - \pi_k\pi_l \\
&= \sum_{k\in U}\sum_{l\in U} \Delta_{kl}\frac{1}{t_x} exp\left\{\frac{t_z}{t_x}\right\}\left(z_k - \frac{t_z}{t_x}\right)\frac{1}{t_x} exp\left\{\frac{t_z}{t_x}\right\}\left(z_l - \frac{t_z}{t_x}\right)\frac{1}{\pi_k\pi_l} \\
&= \frac{1}{t_x^2} exp\left\{2\frac{t_z}{t_x}\right\}\sum_{k\in U}\sum_{l\in U}\Delta_{kl}\left(z_k - \frac{t_z}{t_x}\right)\left(z_l - \frac{t_z}{t_x}\right)\frac{1}{\pi_k\pi_l}
\end{aligned}
$$

4. Proporcione un estimador de la varianza teórica usando
   - Resultado 2.9 de las notas, $\widehat{\mathrm{V}}(\widehat{\theta})$.
   - El estimador alternativo en ecuación (77) de las notas, $\widehat{\mathrm{V}}_a(\widehat{\theta})$.

   Solución.

$$\begin{aligned}
\widehat{\mathrm{V}}(\widehat{\theta}) &= \sum_{k\in U}\sum_{l\in U}\widehat{\Delta}_{kl}\frac{\widehat{v_k}}{\pi_k}\frac{\widehat{v_l}}{\pi_l} \qquad \text{con } \widehat{\Delta}_{kl} = \frac{\pi_{kl}-\pi_k\pi_l}{\pi_{kl}} \\
&= \sum_{k\in U}\sum_{l\in U}\widehat{\Delta}_{kl}\frac{1}{\hat{t}_{x\pi}}\widehat{\theta}\left(z_k - \frac{\hat{t}_{z\pi}}{\hat{t}_{x\pi}}\right)\frac{1}{\hat{t}_{x\pi}}\widehat{\theta}\left(z_l - \frac{\hat{t}_{z\pi}}{\hat{t}_{x\pi}}\right)\frac{1}{\pi_k\pi_l} \\
&= \frac{1}{\hat{t}_{x\pi}^2}\widehat{\theta}^2\sum_{k\in U}\sum_{l\in U}\widehat{\Delta}_{kl}\left(z_k - \frac{\hat{t}_{z\pi}}{\hat{t}_{x\pi}}\right)\left(z_l - \frac{\hat{t}_{z\pi}}{\hat{t}_{x\pi}}\right)\frac{1}{\pi_k\pi_l}
\end{aligned}$$

Para $\widehat{\mathrm{V}}_a(\widehat{\theta})$ tenemos quee $w_k = \frac{1}{w_k}$.Entonces:

$$\widehat{\theta} = exp\left\{\frac{1}{\sum_{k\in S}\frac{1}{\pi_k}}\sum_{k\in S}\frac{z_k}{\pi_k}\right\} = exp\left\{\frac{1}{\sum_{k\in S}w_k}\sum_{k\in S}z_kw_k\right\} = f(\hat{t}_{z\pi},\hat{t}_{x\pi})$$

$$\begin{aligned}
\hat{v}_{ak} = \frac{\partial f(\hat{t}_{z\pi})}{\partial w_k} &= \frac{\partial exp\left\{\frac{1}{\sum_{k\in S}w_k}\sum_{k\in S}z_kw_k\right\}}{\partial w_k} \\
&= exp\left\{\frac{1}{\sum_{k\in S}w_k}\sum_{k\in S}z_kw_k\right\}\frac{z_k\sum_{k\in S}w_k - \sum_{k\in S}z_kw_k}{\left(\sum_{k\in S}w_k\right)^2} \\
&= exp\left\{\frac{1}{\hat{t}_{x\pi}}\hat{t}_{z\pi}\right\}\frac{z_k\hat{t}_{x\pi} - \hat{t}_{z\pi}}{\left(\hat{t}_{x\pi}\right)^2} \\
&= \frac{1}{\hat{t}_{x\pi}}\widehat{\theta}\left(z_k - \frac{\hat{t}_{z\pi}}{\hat{t}_{x\pi}}\right)
\end{aligned}$$

$$\therefore \hat{v}_k = \hat{v}_{ak}$$
$$\therefore \widehat{\mathrm{V}}(\widehat{\theta}) = \widehat{\mathrm{V}}_a(\widehat{\theta})$$

5. **Tamaño de muestra, selección de muestra e intervalos de confianza.**

Continuación del ejercicio 4. Considere la media geométrica como el parámetro de interés. Suponga que se plantea usar un $m.a.s$ para seleccionar la muestra, así como el estimador $\widehat{\theta}$ propuesto en el ejercicio anterior.

Solución.

Como tenemos un m.a.s, N es conocida y además:

$$\pi_k = \frac{n}{N}$$

$$\begin{aligned}
\hat{\theta} &= exp\left\{\frac{\hat{t}_{z\pi}}{N}\right\} \\
&= exp\left\{\frac{1}{N}\sum_{k\in S}\frac{z_k}{\pi_k}\right\} \\
&= exp\left\{\frac{1}{N}\sum_{k\in S}\frac{z_k}{\frac{n}{N}}\right\} \\
&= exp\left\{\frac{N}{Nn}\sum_{k\in S} z_k\right\} \\
&= exp\left\{\frac{1}{n}\sum_{k\in S} z_k\right\} \\
&= exp\left\{\bar{z}_k\right\} \longrightarrow z_k = ln(y_k)
\end{aligned}$$

Para los estimadores de la Varianza, primero obtenemos las variables de linealización

$$\begin{aligned}
a_1 &= \left.\frac{\partial f(\hat{t}_{z\pi})}{\partial \hat{t}_{z\pi}}\right|_{\hat{t}_{z\pi}=t_z} = \left.\frac{\partial f(\hat{t}_{z\pi})}{\partial \hat{t}_{z\pi}}\right|_{\hat{t}_{z\pi}=t_z} \\
&= \left.\frac{\partial\, exp\left\{\frac{1}{N}\hat{t}_{z\pi}\right\}}{\partial \hat{t}_{z\pi}}\right|_{\hat{t}_{z\pi}=t_z} = exp\left\{\frac{1}{N}\hat{t}_{z\pi}\right\}\left.\frac{1}{N}\right|_{\hat{t}_{z\pi}=t_z} \\
&= \frac{1}{N}exp\left\{\frac{t_z}{N}\right\}
\end{aligned}$$

A partir de lo anterior, la variable de linealización v_k y su versión estimada $\widehat{v}_k$ se definen como

$$\begin{aligned}
v_k = a_1 z_k &= \frac{z_k}{N}exp\left\{\frac{t_z}{N}\right\} \\
\therefore \widehat{v}_k &= \frac{z_k}{N}exp\left\{\frac{\hat{t}_{z\pi}}{N}\right\} = \frac{z_k}{N}\widehat{\theta}
\end{aligned}$$

Obteniendo así la varianza:

$$V(\widehat{\theta}) = \sum_{k\in U}\sum_{l\in U} \Delta_{kl}\frac{v_k}{\pi_k}\frac{v_l}{\pi_l} \qquad \text{con } \Delta_{kl} = \pi_{kl} - \pi_k\pi_l$$

Que por el resultado **2.4**, se puede expresar como:

$$V(\widehat{\theta}) = -\frac{1}{2}\sum_{k\in U}\sum_{l\in U} \Delta_{kl}\left[\frac{v_k}{\pi_k} - \frac{v_l}{\pi_l}\right]^2$$

Y por (43), al desarrollar la expresión anterior tomando en cuenta un $m.a.s.$, se tiene:

$$V(\widehat{\theta}) = \frac{N^2}{n}\left(1 - \frac{n}{N}\right) S^2_{v_U}$$

Con:

$$S^2_{v_U} = \frac{1}{N-1}\sum_{k=1}^{N}(v_k - \bar{v}_U)^2$$

También podemos obteber la varianza con las v_k. Estos los usaremos en la pregunta 6.

$$\begin{aligned}
V(\widehat{\theta}) &= \sum_{k\in U}\sum_{l\in U} \Delta_{kl}\frac{v_k}{\pi_k}\frac{v_l}{\pi_l} \qquad \text{con } \Delta_{kl} = \pi_{kl} - \pi_k\pi_l \\
&= \sum_{k\in U}\sum_{l\in U} \Delta_{kl}\frac{z_k}{N}exp\left\{\frac{t_z}{N}\right\}\frac{z_l}{N}exp\left\{\frac{t_z}{N}\right\}\frac{1}{\pi_k\pi_l} \\
&= \frac{1}{N^2}\sum_{k\in U}\sum_{l\in U} \Delta_{kl}z_k exp\left\{\frac{t_z}{N}\right\} z_l exp\left\{\frac{t_z}{N}\right\}\frac{1}{\pi_k\pi_l} \\
&= \frac{1}{N^2}\sum_{k\in U}\sum_{l\in U} \Delta_{kl} exp\left\{2\frac{t_z}{N}\right\}\frac{z_k z_l}{\pi_k\pi_l} \\
&= \frac{1}{N^2} exp\left\{2\frac{t_z}{N}\right\}\sum_{k\in U}\sum_{l\in U} \Delta_{kl}\frac{z_k z_l}{\pi_k\pi_l}
\end{aligned}$$

$$\begin{aligned}
\widehat{\mathrm{V}}(\widehat{\theta}) &= \sum_{k\in U}\sum_{l\in U} \widehat{\Delta}_{kl}\frac{\widehat{v_k}}{\pi_k}\frac{\widehat{v_l}}{\pi_l} \qquad \text{con } \widehat{\Delta}_{kl} = \frac{\pi_{kl} - \pi_k\pi_l}{\pi_{kl}} \\
&= \sum_{k\in U}\sum_{l\in U} \widehat{\Delta}_{kl}\frac{z_k}{N}\widehat{\theta}\frac{z_l}{N}\widehat{\theta}\frac{1}{\pi_k\pi_l} \\
&= \frac{1}{N^2}\sum_{k\in U}\sum_{l\in U} \widehat{\Delta}_{kl}z_k\widehat{\theta}z_l\widehat{\theta}\frac{1}{\pi_k\pi_l} \\
&= \frac{1}{N^2}\sum_{k\in U}\sum_{l\in U} \widehat{\Delta}_{kl}\widehat{\theta}^2\frac{z_k z_l}{\pi_k\pi_l} \\
&= \frac{1}{N^2}\widehat{\theta}^2\sum_{k\in U}\sum_{l\in U} \widehat{\Delta}_{kl}\frac{z_k z_l}{\pi_k\pi_l}
\end{aligned}$$

Para $\widehat{\mathrm{V}}_a(\widehat{\theta})$ tenemos quee $w_k = \frac{1}{w_k}$.
Entonces $\widehat{\theta} = exp\left\{\frac{1}{N}\sum_{k=1}^{N}\frac{z_k}{\pi_k}\right\} = exp\left\{\frac{1}{N}\sum_{k=1}^{N} z_k w_k\right\} = f(\hat{t}_{z\pi})$

$$\begin{aligned}\hat{v}_{ak} = \frac{\partial f(\hat{t}_{z\pi})}{\partial w_k} &= \frac{\partial exp\left\{\frac{1}{N}\sum_{k=1}^{N} z_k w_k\right\}}{\partial w_k} \\ &= exp\left\{\frac{1}{N}\sum_{k=1}^{N} z_k w_k\right\}\frac{t_z}{N} \\ &= \frac{t_z}{N} exp\left\{\frac{1}{N}\sum_{k=1}^{N}\frac{z_k}{\pi_k}\right\} = \frac{t_z}{N}\widehat{\theta}\end{aligned}$$

$$\begin{aligned}\widehat{\mathrm{V}}_a(\widehat{\theta}) &= \sum_{k\in U}\sum_{l\in U}\widehat{\Delta}_{kl}\frac{\widehat{v_{ak}}}{\pi_k}\frac{\widehat{v_{al}}}{\pi_l} \qquad \text{con } \widehat{\Delta}_{kl} = \frac{\pi_{kl}-\pi_k\pi_l}{\pi_{kl}} \\ &= \sum_{k\in U}\sum_{l\in U}\widehat{\Delta}_{kl}\frac{t_z}{N}\widehat{\theta}\frac{t_z}{N}\widehat{\theta}\widehat{\theta}\frac{1}{\pi_k\pi_l} \\ &= \frac{1}{N^2}\sum_{k\in U}\sum_{l\in U}\widehat{\Delta}_{kl} z_k\widehat{\theta} z_l\frac{1}{\pi_k\pi_l} \\ &= \frac{t_z^2}{N^2}\sum_{k\in U}\sum_{l\in U}\widehat{\Delta}_{kl}\widehat{\theta}^2\frac{1}{\pi_k\pi_l} \\ &= \frac{t_z^2}{N^2}\widehat{\theta}^2\sum_{k\in U}\sum_{l\in U}\widehat{\Delta}_{kl}\frac{1}{\pi_k\pi_l}\end{aligned}$$

1. Dé la expresión del tamaño de la muestra que se debería usar considerando una aproximación Normal, una confianza de 95 % y un error absoluto d.
   Solución.

   Con los resultados anteriores y usando **(94)**

   $$n \geq \frac{\left(\frac{kNS_{v_U}}{d}\right)^2}{1+\frac{1}{N}\left(\frac{kNS_{v_U}}{d}\right)^2}$$

   Como se busca una confianza de 95 %, entonces $k \approx 1.96$, entonces:

   $$n \geq \frac{\left(\frac{1.96NS_{v_U}}{d}\right)^2}{1+\frac{1}{N}\left(\frac{1.96NS_{v_U}}{d}\right)^2}$$

   y $S^2_{v_U}$ como se definió anteriormente.

2. Suponga que se quiere estimar la media geométrica de los precios del aguacate en un conjunto de 411 puntos de venta para la segunda quincena de marzo de 2022. Suponga

que para encontrar el tamaño de muestra adecuado se ha realizado una muestra piloto en 20 puntos de venta en la primer quincena de marzo de 2022. La información de los precios en esa pequeña muestra se encuentra en el archivo. *AguacateMuestraPiloto.csv*. Con esta información, calcule el tamaño de muestra mínimo necesario para poder estimar la media geométrica con un error absoluto $d = 3$. Es decir, con una confianza de 95 % se desea que la estimación de la media geométrica de los precios del aguacate en la segunda quincena de marzo de 2022 se realice con un error de ± 3 pesos.

Solución.

Importamos los datos y_k, realizamos una transformación:

$$z_k = ln(y_k)$$

```
aguacate <-read.csv("AguacateMuestraPiloto.csv",encoding = "Latin1")
y<-aguacate[,1]
z<-log(y)
```

Posteriormente definimos los valores que serán usados para este problema en especifico: $n = 20$, $N = 411$ $d = 3$ y k igual al cuantil al 95 % de una distribución normal.

```
aguacate <-read.csv("AguacateMuestraPiloto.csv",encoding = "Latin1")
y<-aguacate[,1]
z<-log(y)
N<-411
n<-length(y)
d<-3
k<-qnorm(0.975) #k=1.959964
```

Después calcular la estimación de n, pero como no contamos con los valores poblacionales S_{v_U} , realizaremos una estimación sobre v_k, usando:

$$\widehat{v}_k = \frac{z_k}{N}\widehat{\theta}$$

Donde: $\widehat{\theta} = \exp \bar{z}$

```
theta<-exp(mean(z))
v<-(z/N)*theta

Sv<-var(v)
Sv
```

```
## [1] 0.0007161418

n<-(Sv*(k*N/d)^2)/(1+(1/N)*(Sv*(k*N/d)^2))
#Tomando el entero mayor o igual a n
n<-ceiling(n)
n

## [1] 46
```

Así se obtiene que la cota del tamaño muestral es $n = 46$

3. La base *DatosAguacatePob.csv* contiene la información de los 411 puntos de venta. Por ahora suponga que no tiene la columna *Aguacate_q2Mar2022*, es decir, sólo se tiene la información de los puntos de venta y su localización. Seleccione una muestra de tamaño n usando un *m.a.s.*, donde n es el valor que se encontró en el inciso anterior.
Solución.

```
# 5.3
# Cargamos los datos poblacionales
aguacate_pob <- read.csv(file = 'DatosAguacatePob.csv', sep=','
, fileEncoding="latin1")

set.seed(2021) # para preservar los resultados
# n=46
indices=sample(1:N, n)
#generación de la muestra
muestra=aguacate_pob[indices,]
```

4. Considerando la muestra seleccionada y la información asociada a los elementos en la muestra en la columna *Aguacate_q2Mar2022*, proporcione la estimación de la media geométrica de los precios del aguacate en la segunda quincena de marzo de 2022. También proporcione un intervalo de confianza al 95 %
Solución.

```
# 5.4
#la estimación de la media geometrica sobre la muestra
y<-muestra[,1]
z<-log(y)
#probabilidades de inclusion (m.a.s)
```

```
pi_k<-n/N
pi_kl<-n*(n+1)/(N*(N+1))
#El estimador de la media geometrica es:
mediaG<-exp(1/N*sum(z/pi_k))
mediaG

## [1] 64.80261
```

La media geométrica estimada es:

$$\hat{\theta} = 64.803$$

Calculamos la varianza, que queda expresada como:

$$V(\widehat{\theta}) = \frac{N^2}{n}\left(1 - \frac{n}{N}\right) S^2_{v_U}$$

Usando de nuevo los estimadores $\hat{v}_k = \frac{z_k}{N}\hat{\theta}$

```
# calculamos las vk
v_k<-(z/N)*mediaG

#la varianza del estimador
V<-(N^2/n)*(1-n/N)*var(v_k)
```

Y su invervalo de confianza es:

$$\left[\hat{\theta} - k \cdot \sqrt{V(\hat{\theta})}, \hat{\theta} + k \cdot \sqrt{V(\hat{\theta})}\right]$$

```
IC<-(c(mediaG-k*sqrt(V),mediaG+k*sqrt(V)))
IC

## [1] 61.59930 68.00592
```

$$= (61.59930, \quad 68.00592)$$

6. **Evaluación del desempeño del estimador y de los intervalos de confianza.**

Continuación de los ejercicios 4 y 5.

Considere la base *DatosAguacatePob.csv* como la población ficticia. Aquí se tiene la información para los 411 elementos en la población, así que se puede analizar el desempeño de los estimadores usados. Para esto repita la selección de la muestra como en ejercicio 5, $B = 10000$ veces y presente lo siguiente.

1. Histograma de las $B = 10000$ estimaciones de la media geométrica, incluyendo el valor poblacional, así como la aproximación a la distribución normal asociada (usando $\mathrm{V}(\widehat{\theta})$).
Solución.
Cargamos nuestra base y creamos la variable z_k que se llama Log

```
library(tidyverse)
library(psych)
library(latex2exp)
Aguacate <- read.csv(file = 'DatosAguacatePob.csv', sep=',',
                     fileEncoding="latin1")
Log<-log(Aguacate[,1])
AguacatePob <- cbind(Aguacate,Log)
```

Obtenemos el valor muestral de la media geométrica

```
(MediaGeometrica=exp(mean(Log)))

## [1] 64.16803
```

Definimos semilla y nuestra muestra será de 20 con una población total de 411

```
set.seed(2021)
n=46
N=length(Log)
```

Usamos la función sample en R para seleccionar con un m.a.s., y así tener los 20 índices selecionados.

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

##  [1] 391 166 231 396  70 192 251 102 110 325 103 361  23 332 146 123 188 125 3
## [20] 302 360  26 164 101 342 159 240 354  68 373 393 133  73 191 294 274 171 3
## [39] 242 407 381 198 378  79 162 150
```

Primero calcularemos la varianza del estimador y la estimación de la varainza del estimador.

Para esto tenemos que calcular las probabilidades de inclusión

```
pi_k=n/N
pi_kl=(n*(n-1))/(N*(N-1))
```

Total poblacional

```
VarPob=0
t_z=sum(Log)
t_z
```

```
## [1] 1710.379
```

Población por el estimador de HT

```
DatosS=AguacatePob[sindex,]
B=10000
ParMuestra = replicate(B, sample(1:N, n))
t_piz=sapply(1:B,function(x, Indices, Pob){((1/pi_k)*sum(Pob[Indices[,x],"Log"]))
             Indices=ParMuestra, Pob=AguacatePob)
(mean(t_piz))
```

```
## [1] 1710.539
```

Función para calcular la Varianza del estimador y la estimación de la varianza del estimador. Toma como valor tz que será el total poblacional o el estimador de HT

```
Varianza=function(t_z){
  for(k in 1:N) {
    for(l in 1:N){
      if(k==l){
        VarPob=VarPob + (pi_k - pi_k**2)*(Log[k]**2/pi_k**2)
      }else{
        VarPob=VarPob + (pi_kl-pi_k**2)*(Log[k]/pi_k)*(Log[l]/pi_k)
      }
    }
  }
  VarPob=(exp(2*t_z/N))/(N^{2})*VarPob
}
```

Aplicamos la función para calcular la Varianza del estimador

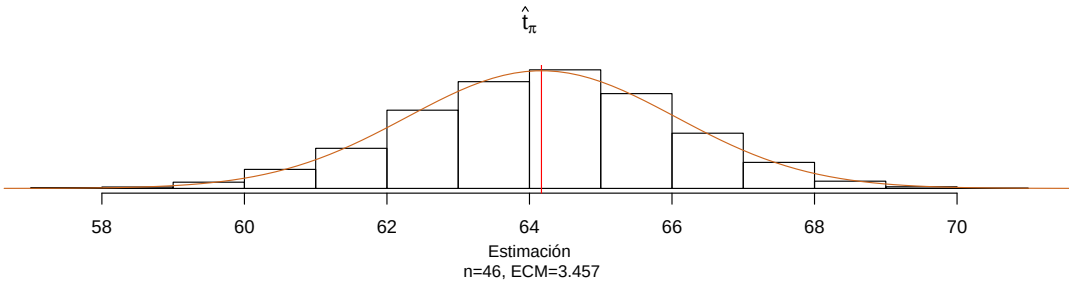
```
VarPob=Varianza(t_z)
VarPob

## [1] 3.552779
```

Aplicamos la función para calcular el estimador de la Varianza del estimador

```
VarPobEst=mean(Varianza(t_piz))
VarPobEst

## [1] 3.64851
```

Con estos índices se puede obtener la estimación. Para ésta repetiremos la selección 10000 veces para analizar cómo se ve la distribución del estimador HT

```
DatosS=AguacatePob[sindex,]
B=10000
ParMuestra = replicate(B, sample(1:N, n))

MediaGeometrica.HT= sapply(1:B,function(x, Indices, Pob)
    {exp(mean(Pob[Indices[,x],"Log"])) }, Indices=ParMuestra,
    Pob=AguacatePob)
h1=hist(MediaGeometrica.HT, plot=FALSE)
plot(h1, freq=FALSE, main=TeX("$\\widehat{t}_{\\pi}$"),
    xlab=paste("\n Estimación \n n=",n, ", ECM=",
    round(mean((MediaGeometrica.HT-MediaGeometrica)^2), 3) ,  sep=""),
    ylab="", cex.lab=.9, yaxt='n')
abline(v=exp(mean(AguacatePob[,"Log"])), col="red")

#Aproximación normal
x <- seq(-4,4,length=100)*sqrt(VarPob) + MediaGeometrica
hx <- dnorm(x,MediaGeometrica,sqrt(VarPob))
lines(x, hx,col = "chocolate3")
```



Tenemos una aproximación normal que se ajusta bien a los resultados del muestreo, con una media muy parecida el valor poblacional. Con un erros cuadrático medio igual a 3.457

2. Calcule los $B = 10000$ intervalos de confianza y proporcione el porcentaje de intervalos de confianza que contienen al valor poblacional.
   Solución.
   Calculamos los intervalos de confianza con el estimador de la varianza del estimador de HT.

```
IntervaloConfianza=c(MediaGeometrica.HT-qnorm(0.975)*sqrt(VarPobEst),
                     MediaGeometrica.HT+qnorm(0.975)*sqrt(VarPobEst))
```

   Ahora vemos cuantos de los intervalos tienen al valor poblacional.

```
mean(MediaGeometrica>=MediaGeometrica.HT-qnorm(0.975)*sqrt(VarPobEst)
     & MediaGeometrica<=MediaGeometrica.HT+qnorm(0.975)*sqrt(VarPobEst))
```

```
## [1] 0.9559
```

   El 95.59 % de los intervalos de confianza contienen al valor poblacional. Siendo esto consistente con el nivel de confianza.

3. Calcule el porcentaje de veces que $|\widehat{\theta} - \theta| \leq d$.
   Solución.
   Usamos d =3 porque es la que se usa en el ejercicio 5.

```
d=3
mean(abs(MediaGeometrica.HT-MediaGeometrica)<=d)
```

```
## [1] 0.8911
```

   Tenemos que nuestro estimador varia en ±3 pesos el 89.11 % de las veces, por lo tanto, la estimación de la media geométrica de los precios del aguacate en la segunda quincena de marzo de 2022 se realiza con un error de ±3 pesos pesos el 89.11 % de las veces.
   En este cso nos quemdos cortos pue slo ideal sería que se llegue a un 95 %.