Tarea 3

Christian Badillo, Luis Nuñez, Luz Maria Santana, & Sealtiel Pichardo

Tabla de contenidos

Problema 1

- Una compañía realiza un estudio entre sus 7500 empleados, se desea conocer la proporción de empleados que tienen automóvil. Se decide dividir a la población de empleados en 3 estratos. Estrato 1: Bajo ingreso (<10000 pesos al mes); estrato 2: Ingreso medio (10001 a 20000 pesos al mes) y estrato 3: Alto ingreso (>20000 pesos al mes). Los resultados de la encuesta se muestran a continuación:

Estrato	Nh	nh	ph
Estrato 1	3,500	500	0.13
Estrato 2	2,000	300	0.45
Estrato 3	2,000	200	0.50

Donde:

- ***Nh:*** es el número de empleados en el estrato h
- ***nh:*** es el tamaño de muestra del estrato h
- ***ph:*** es la proporción muestral de empleados con automóvil

1) Obtenga $\hat{P}$, la proporción de empleados en la compañía que tienen automóvil.

La estimación total de la proporción de empleados se calcula como la suma de las proporciones en cada estrato multiplicadas por el peso que ocupa dicho estrato:

$$\hat{P} = \sum_{h=1}^{L} W_h \hat{p_h} \quad con : \hat{p_h} = \sum_{i=1}^{n_h} \frac{y_{h_i}}{n_h} \quad y \quad W_h = \frac{N_h}{N}$$

```
N <- sum(datos$Nh)
Nh <- datos$Nh
nh <- datos$nh
ph <- datos$ph

#Calcular la proporción total

Wh <- c()
suma <- c()

for(i in 1:nrow(datos)){
  Wh[i] <- Nh[i] / N

  suma[i] <- Wh[i] * ph[i]

}

#Proporción total estimada de personas en la compañía que tienen automóvil

p_total <- sum(suma)
```

La estimación de la proporción de los empleados que tienen automóvil es de 0.314.

2) Construya el intervalo del 95% de confianza para P

Para ello usamos el estimador de la varianza del estimador

$$\hat{V}(\hat{P}) = \sum_{h=1}^{L} W_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{\hat{p_h}(1-\hat{p_h})}{n_h - 1}$$

```
z <- 1.96
var_est <- c()

for(i in 1:nrow(datos)){
  var_est[i] <- ((Wh[i])^2) * (1- (nh[i]/Nh[i]))* (ph[i] * (1-ph[i])/(nh[i]-1))
}


prec <- z * sqrt(sum(var_est))

ls <- p_total + prec
li <- p_total - prec
```

El intervalo de confianza obtenido es: $[0.2882, 0.3397]$, indicando que con un nivel del 95% de confianza, se estima que la proporción de empleados con automóvil en la compañía se encuentra entre el 28.82% y el 33.97%.

3) Considere que el criterio de estratificación fue adecuado? Explique el porqué de su respuesta.

Dado que la población de una compañía está compuesta por empleados con funciones y salarios heterogéneos, es adecuado estratificar los salarios cuando se estudia la posesión de bienes materiales, como un automóvil, ya que cada estrato presenta condiciones socioeconómicas similares y, por ende, patrones de consumo más homogéneos. En este caso, se confirma que el salario fue un buen criterio de estratificación, pues se observó una relación directa entre el nivel de ingresos y la probabilidad de poseer un automóvil: a mayor ingreso, mayor proporción de empleados con automóvil. Esta estratificación, al reducir la variabilidad dentro de cada grupo, mejora la precisión de las estimaciones, lo que no sería posible si no se hubiera aplicado este criterio.

Problema 2.

En la reserva de fauna de Tanzania, se tiene un lote de 100 elefantes, el director de la reserva desea estimar el total del peso del lote debido a que la sequía los obliga a trasladar en un bote por río a los elefantes hacia zonas en donde exista alimento. El año pasado se obtuvieron los siguientes resultados respecto al peso (en toneladas) de los elefantes:

	Número	Peso Medio	Varianza
Macho	60	6	4.00
Hembra	40	4	2.25

1) Calcule la varianza poblacional para la variable "peso del elefante".

```
#Promedio ponderado
media_pond <- sum(df$Número * df$Peso_Medio)/N

##Distinguimos
N <- sum(df$Número)
Nh <- df$Número
Ymed_h <- df$`Peso Medio`
S2_h <- df$Varianza

media_pond <- sum(Nh * Ymed_h) / N
#La varianza poblacional se calcula como la suma de la
# varianza dentro cada estrato más la varianza entre estratos,
# cada una multiplicada por su respectivo peso.
#La varianza poblacional se calcula como la suma de la varianza dentro cada estra
# la varianza entre estratos, cada una multiplicada por su respectivo peso.
Wh <- Nh/N
varxestr <- Wh * S2_h
var_entre_estr <- Wh *((Ymed_h-media_pond)^2)

var_pob <- sum(var_entre_estr)+sum(varxestr)
```

Se obtuvo una varianza poblacional de 4.26 y una desviación estándar de 2.064 lo que indica que en promedio los pesos se desvían esa magnitud de la media poblacional. Esto ocurre por las diferencias de peso entre machos y hembras.

2) Si asumimos que el peso de los elefantes no ha cambiado significativamente de un año a otro, obtenga el estimador de la varianza del total de peso de lote, considere un m.a.s. de tamaño 10.

```
#estimador de la varianza del total del lote,
# bajo m.a.s. sin estratos con n= 10:
# Necesitamos la varianza total poblacional (var_pob).
n <- 10

var_mas <- (N^2)*(1- n/N)*(var_pob/n)
```

La varinza del estimador del total suponiendo un muestreo aleatorio simple es $\mathbb{V}\left(\hat{Y}\right) \approx 3834$. Este valor tan elevado es debido a un tamaño de muestra pequeño (10) que apenas corresponde al 10% de la población incial (100). Al ser una muestra tan reducida, la incertidumbre en la estimación aumenta. Además, se genera una mayor variabilidad en el peso totalal no tomar en cuenta la diferencia de los pesos entre machos y hembras, y la diferencia en las varianzas del peso.

3) Considere un muestreo estratificado (distribución proporcional) de tamaño 10 y obtenga el estimador de la varianza del total del peso del lote.

```r
#Dada una muestra n = 10, calculamos el nh de manera proporcional a cada estrato
nh <- c()

var_estrat <- c()

for(i in 1:nrow(df)){
  #Cuánto de n=10 le toca a cada estrato:
  nh[i] <- Wh[i] * n

# Calculo de la varianza
# Se calcula como la suma del tamaño del estrato al cuadrado por el factor de corrección
# por finitud de cada estrato y la varianza de cada estrato entre el tamaño de muestra de cada estrato

  var_estrat[i] <- (df$Número[i]^2)*(1- nh[i]/df$Número[i])*(df$Varianza[i]/nh[i])
}

var_estrato <- sum(var_estrat)
```

El resultado obtenido es de 2970, lo que implica la variabilidad esperada en la estimación del peso total de la población de elefantes al utilizar un diseño de muestreo estratificado con el tamaño de muestra proporcional.

4) Dé sus conclusiones.

Al comparar el estimador del peso total bajo muestreo aleatorio simple (3834) y el estimador con estratos (2970) se observa que la varianza en muestreo estratificado es menor, lo que indica que el muestreo estratificado es más preciso dado que toma en cuenta las diferencias entre machos y hembras, lo que explica mejor a la población.

Problema 3

En el estado de Sonora los productores de trigo conformaron estratos de acuerdo a su tamaño y al número promedio de hectáreas de trigo sembradas por granja en cada estrato, con los siguientes resultados:

	Número de Granjas	Promedio de hectarias sembradas	Desviación Estándar
0-30	482	6.2	3.5
31-60	590	11.7	4.8
61-90	532	18.5	6.7
91-120	479	26.3	8.1
>120	166	49.3	9.3

Para una muestra de 150 granjas, distribuya la muestra en los 5 estratos utilizando:

1) La distribución de Neyman.
2) La distribución proporcional.
3) Compare las precisiones de estos métodos con la del m.a.s.

Distribución de Neyman.

Bajo la distribución de Neyman, el tamaño de muestra para el extracto h es:

$$n_h = n\frac{N_h S_h}{\sum_{h=1}^{L} N_h S_h}$$

```
n <- 150
Nh <- df["Número de Granjas"]
Sh <- df["Desviación Estándar"]
nh <- n * (Nh*Sh) / sum(Nh*Sh)
nh <- round(nh)
```

	Tamaño de muestra n_h
0-30	19
31-60	31
61-90	40
91-120	43
>120	17

Distribución Proporcional.

Bajo la distribución de Proporcional, el tamaño de muestra para el extracto h es:

$$n_h = n\frac{N_h}{N} = nW_h$$

```
n <- 150
N <- sum(df["Número de Granjas"])
Nh <- df["Número de Granjas"]
nh <- n * Nh / N
nh <- round(nh)
nh[3,] <- nh[3,] + 1
```

	Tamaño de muestra n_h
0-30	32
31-60	39
61-90	36
91-120	32
>120	11

Comparación de Precisión.

Varianza del estimador de la media en la distribución de Neyman.

La varianza para el estimador de la media con tamaño de muestra obtenido por la distribución de Neyman es:

$$\mathbb{V}\left(\hat{\bar{Y}}\right) = \frac{\left(\sum_{h=1}^{L} W_h S_h\right)^2}{n} - \frac{\sum_{h=1}^{L} W_h S_h^2}{N}$$

```
N <- sum(df["Número de Granjas"])
Nh <- df["Número de Granjas"]

Wh <- Nh / N
Sh2 <- df["Desviación Estándar"] ** 2

var_ymean <- (sum(Wh * df["Desviación Estándar"]) ** 2) /n - sum(Wh * Sh2) / N
```

La precisión para la distribución proporcional $\mathbb{V}\left(\hat{\bar{Y}}\right) \approx 0.2228$.

Varianza del estimador de la media en la distribución proporcional.

Para estimar la varianza del estimador de la media bajo la distribución proporcional, se utiliza:

$$\mathbb{V}\left(\hat{\bar{Y}}\right) = \left(1 - \frac{n}{N}\right)\left(\frac{\sum_{h=1}^{L} W_h S_h^2}{n}\right)$$

```
N <- sum(df["Número de Granjas"])
Nh <- df["Número de Granjas"]

Wh <- Nh / N
Sh2 <- df["Desviación Estándar"] ** 2

var_ymean2 <- (1 - n/N) * sum(Wh * Sh2) / n
```

La precisión para la distribución proporcional $\mathbb{V}\left(\hat{\bar{Y}}\right) \approx 0.2467$.

Varianza del estimador de la media bajo muestro aleatorio simple.

Para estimar la varianza del estimador de la media bajo muestreo aleatorio simple, seusa la siguiente expresión:

$$\mathbb{V}\left(\hat{\bar{Y}}\right) = \left(1 - \frac{n}{N}\right)\frac{S^2}{n}$$

Dado que tenemos estratos, para calcular S^2 se debe usar:

$$S^2 = \sum_{h=1}^{L} W_h S_h^2 + \sum_{h=1}^{L} W_h \left(\bar{Y_h} - \bar{Y}\right)^2$$

```
N <- sum(df["Número de Granjas"])
Nh <- df["Número de Granjas"]

Wh <- Nh / N
Sh2 <- df["Desviación Estándar"] ** 2
Ybar <- sum(df$`Promedio de hectarias sembradas` * df$`Número de Granjas`) / sum(df$`Número de Granjas`)

S2 <- sum(Wh * Sh2) + sum(Wh * (df$`Promedio de hectarias sembradas` - Ybar) ** 2)

var_ymean3 <- (1 - n/N) * S2 / n
```

La precisión para muestreo aleatorio simple $\mathbb{V}\left(\hat{\bar{Y}}\right) \approx 1.0387$.

Comparación

Podemos observar que se cumple la siguiente desigualdad:

$$V_{opt}\left(\hat{\bar{Y}}\right) \leq V_{prop}\left(\hat{\bar{Y}}\right) \leq V_{m.a.s.}\left(\hat{\bar{Y}}\right)$$

En este caso:

$$0.2228 \leq 0.2467 \leq 1.0387.$$