Practica Parcial 2

Estudiantes de Pregrado en Estadística

Ivan Santiago Rojas Martinez
Ronald Gabriel Palencia

Docente

Victor Ignacio Lopez Rios

Asignatura

Muestreo Estadistico

Sede Medellín
Octubre 23 de 2023

Índice

Trabajo Práctico Segundo Examen Parcial

Utilizando el texto **Elementary Survey Sampling** se aplica el M.A.E. para estimar el número promedio de palabras por párrafo. Para esto, se generan tres estratos utilizando los capitulos del texto de la siguiente manera:

- **Estrato 1:** Capítulos 1, 2 y 3.
- **Estrato 2:** Capítulos 4, 5 y 7.
- **Estrato 3:** Capítulos 7, 8, 9 y 10.

1 Construcción de marco muestral por estrato.

Con el fin de tratar de conservar la homogeneidad dentro de cada uno de los estratos, mantener una baja variabilidad en el número de palabras por párrafo y obtener una estimación más precisa del parámetro de interés, se utilizan los párrafos del libro que presenten características similares en su longitud. Por lo tanto, es claro que no se consideran como párrafos las fórmulas matemáticas, los gráficos y las tablas que aparezcan en el libro. De igual manera, cuando las fórmulas están dentro del párrafo no se tienen en cuenta como palabras.

Para el caso de las secciones de ejercicios en cada capítulo, se tiene en cuenta que existen algunos de estos que solo presentan una pregunta, por lo que no se pueden considerar directamente como párrafos. Sin embargo, existen otro tipo de ejercicios cuya formulación incluye una explicación de lo que se debe hacer, por lo tanto, si se pueden considerar como párrafos. Además, aquellos ejercicios que cuenten con literales cortos se consideran de manera general como parte del párrafo del ejercicio, mientras que aquellos que presenten un contexto de un problema, se consideran como párrafos independientes.

NOTA: Junto a este informe se anexa el libro con la enumeración realizada para cada párrafo.

2 Formato de EXCEL para el registro de información.

El formato creado contiene tres pestañas:

- **MAE piloto:** Contiene el conteo de palabras realizado en la muestra piloto.
- **MAE final:** Contiene el conteo de palabras realizado en la muestra final.
- **Conteo de párrafos:** Contiene el número de párrafos por estrato.

NOTA: Junto a este informe se anexa el formato descrito con el nombre **Registro de información.xlsx**.

3 Muestra piloto

Para obtener estimaciones de la varianza del número de palabras por párrafo en los tres estratos, se considera realizar una muestra piloto, donde en cada uno de los estratos se obtienen muestras aleatorias simples (M.A.S.) de 20 párrafos seleccionados aleatoriamente.

A continuación, se muestra el código utilizado para obtener las tres M.A.S. por cada estrato. Además, el número de párrafos por estrato es el siguiente:

- **Estrato 1:** 305 párrafos.
- **Estrato 2:** 560 párrafos.
- **Estrato 3:** 447 párrafos.

A continuación, se obtienen las M.A.S. para cada estrato:

```
# Proponer valor de semilla para obtener siempre los mismos valores aleatorios.
set.seed(123)

# MAS para estrato 1.
MAS1 = sample(1:305, 20)
sort(MAS1)
```

```
##  [1]   7  14  26  90  91 118 137 153 179 195 197 211 229 244 254 256 291 295 299
## [20] 304
```

```
# MAS para estrato 1.
MAS2 = sample(1:560, 20)
sort(MAS2)
```

```
##  [1]  23  41  72 135 141 143 153 166 217 224 277 290 294 309 373 431 463 490 544
## [20] 555
```

```
# MAS para estrato 1.
MAS3 = sample(1:447, 20)
sort(MAS3)
```

```
##  [1]   4  13  16  34  39  69  86  90  94 116 159 209 223 235 240 262 306 316 342
## [20] 374
```

Ahora, se exporta la información obtenida con las muestras para poder estimar el número promedio de palabras por párrafo en cada estrato y su respectiva varianza muestral.

```r
# Se importa el formato de información para la muestra piloto:
DataMuestraPiloto = read_excel("Data/Registro de información.xlsx",
                               sheet = "MAE piloto")
```

```r
# Promedio y varianza muestral para el estrato 1.
MediaE1 = mean(DataMuestraPiloto[DataMuestraPiloto$Estrato == "Estrato 1",
                                 "Palabras"]$Palabras)
VarE1 = var(DataMuestraPiloto[DataMuestraPiloto$Estrato == "Estrato 1",
                              "Palabras"]$Palabras)
MediaE1
```

```
## [1] 92.15
```

```r
VarE1
```

```
## [1] 3712.976
```

```r
# Media y varianza muestral para el estrato 2.
MediaE2 = mean(DataMuestraPiloto[DataMuestraPiloto$Estrato == "Estrato 2",
                                 "Palabras"]$Palabras)
VarE2 = var(DataMuestraPiloto[DataMuestraPiloto$Estrato == "Estrato 2",
                              "Palabras"]$Palabras)
MediaE2
```

```
## [1] 73.85
```

```r
VarE2
```

```
## [1] 2140.029
```

```r
# Media y varianza muestral para el estrato 3.
MediaE3 = mean(DataMuestraPiloto[DataMuestraPiloto$Estrato == "Estrato 3",
                                 "Palabras"]$Palabras)
VarE3 = var(DataMuestraPiloto[DataMuestraPiloto$Estrato == "Estrato 3",
                              "Palabras"]$Palabras)
MediaE3
```

```
## [1] 67.25
```

```
VarE3
```

```
## [1] 1056.618
```

4 Cálculo del tamaño de muestra y tipo de asignación a estratos

Con el fin de elegir el tipo de asignación más apropiada a los estratos, se debe entender que dicha asignación depende de los siguientes factores:

1) El número de elementos en cada estrato.
2) La variabilidad de las observaciones dentro de cada estrato.
3) El costo de obtener una observación en cada estrato.

Relacionado al primer item, se encontró que el número de párrafos son diferentes para cada uno de los estratos (**Sección 2**). Por otra parte, con base en la muestra piloto, se pudo obtener una estimación de la varianza para cada uno de los estratos. Dichas varianzas estimadas son diferentes para cada uno de los estratos. Por último, en el contexto de este trabajo académico no se conoce el costo de contar el número de palabras por párrafo, por lo tanto, se asumen que estos costos son iguales para cada uno de los estratos.

La descripción anterior, permite concluir que el tipo de asignación más apropiado para el caso trabajado es la **Asignación óptima de Neyman** suponiendo costos iguales y tamaños de estratos y varianzas diferentes.

Con base en lo anterior, se procede a calcular el tamaño de muestra mínimo que se requiere para estimar el número promedio de palabras por párrafo del libro *Elementary Survey Sampling* con una confianza del 95% y un error relativo del 10%.

Para encontrar el tamaño de muestra n se utiliza la siguiente expresión:

$$n = \frac{\sum_{h=1}^{3} \frac{N_h^2 \sigma_h^2}{w_h}}{N^2 D + \sum_{h=1}^{3} N_h \sigma_h^2}$$

Donde $D = \frac{B^2}{Z_{\alpha/2}^2}$

En este caso, se tiene que el error relativo $\epsilon = 0.1$, es decir, $\epsilon = \frac{B}{\mu} \iff B = 0.1 * \mu$. Para este caso se debe estimar μ utilizando la información de la muestra piloto.

```
# Estimación de la media con la muestra piloto.
MediaPiloto = mean(DataMuestraPiloto$Palabras)
MediaPiloto
```

```
## [1] 77.75
```

Por tanto, $\hat{\mu}_{Piloto} = 77.75$, entonces $B = 0.1 * 77.75 = 7.775 \approx 8$. Este valor se puede nombrar como el límite del error absoluto para la estimación del número promedio de palabras por párrafo en el libro.

Con el valor de B encontrado, se calcula $D = \frac{8^2}{Z_{\alpha/2}}$, donde $\alpha = 1 - 0.95 = 0.05$. Por tanto, $D = \frac{8^2}{Z_{0.05/2}} = \left(\frac{8}{1.96}\right)^2$

Como se va a utilizar la **Asignación óptima de Neyman**, las proporciones para cada estrato se encuentran con la siguiente expresión:

$$w_h = \frac{N_h \sigma_h}{\sum_{h=1}^{3} N_h \sigma_h}$$

Ahora, de acuerdo a la asignación óptima se tiene que:

$$w_1 = \frac{305 * 60.9342}{(305 * 60.9342) + (560 * 46.26045) + (447 * 32.50567)} = 0.3148877$$

$$w_2 = \frac{560 * 46.26045}{(305 * 60.9342) + (560 * 46.26045) + (447 * 32.50567)} = 0.4389274$$

$$w_3 = \frac{447 * 32.50567}{(305 * 60.9342) + (560 * 46.26045) + (447 * 32.50567)} = 0.2461849$$

$$\sum_{h=1}^{3} \frac{N_h^2 \sigma_h^2}{w_h} = \frac{305^2(3712.976)}{0.3148877} + \frac{560^2(2140.029)}{0.4389274} + \frac{447^2(1056.618)}{0.2461849} = 3483456042$$

$$\sum_{h=1}^{3} N_h \sigma_h^2 = 305(3712.976) + 560(2140.029) + 477(1056.618) = 2834881$$

$$N^2 D = 1312^2 \left(\frac{8}{1.96}\right)^2 = 28677118$$

Para hallar n se unen los resultados anteriores:

$$n = \frac{3483456042}{28677118 + 2834881} = 110.5438 \approx 111$$

Por lo tanto, el tamaño de muestra necesario para estimar el número promedio de palabras por párrafo del libro con una confianza del 95% y un error relativo del 10% es $n = 111$.

5 Estimación con base en la M.A.E. obtenida.

La asignación de $n = 111$ para cada estrato es:

- $n_1 = nw_1 = 111 * (0.3148877) = 34.95253 \approx 35$
- $n_2 = nw_2 = 111 * (0.4389274) = 48.72094 \approx 49$
- $n_3 = nw_3 = 111 * (0.2461849) = 27.32652 \approx 27$

Ahora, se realizan las muestras aleatorias simples por cada estrato para poder estimar el parámetro de interés.

```
# Proponer valor de semilla para obtener siempre los mismos valores aleatorios.
set.seed(345)

# MAS para estrato 1.
MAS1_G = sample(1:305, 35)
sort(MAS1_G)
```

```
##  [1]   2   5  18  19  27  32  40  69  73  75  87  90  93  97 105 106 112 135 144
## [20] 145 147 161 175 188 211 215 218 235 259 263 272 284 288 291 293
```

```
# MAS para estrato 1.
MAS2_G = sample(1:560, 49)
sort(MAS2_G)
```

```
##  [1]   7  33  34  44  46  62  88  93  95 101 124 126 128 129 148 165 166 180 187
## [20] 195 210 212 218 226 227 232 239 247 275 285 289 295 327 333 404 410 438 445
## [39] 447 470 483 486 487 493 496 498 511 526 530
```

```
# MAS para estrato 1.
MAS3_G = sample(1:447, 27)
sort(MAS3_G)
```

```
##  [1]   3  17  29  52  66  98 118 127 140 164 178 180 222 254 264 273 291 314 332
## [20] 346 347 354 390 404 407 408 430
```

```
# Se importa el formato de información para la muestra final:
DataMuestraFinal = read_excel("Data/Registro de información.xlsx",
                              sheet = "MAE final")
DataMuestraFinal = DataMuestraFinal[,c("Estrato","Pagina","Parrafo","Palabras")]
```

De la M.A.E. generada, se obtuvieron los siguientes resultados:

```r
# Media y varianza muestral para el estrato 1.
MediaE1_G = mean(DataMuestraFinal[DataMuestraFinal$Estrato == "Estrato 1",
                                  "Palabras"]$Palabras)
VarE1_G = var(DataMuestraFinal[DataMuestraFinal$Estrato == "Estrato 1",
                               "Palabras"]$Palabras)
MediaE1_G
```

```
## [1] 106.4571
```

```r
VarE1_G
```

```
## [1] 3476.255
```

```r
# Media y varianza muestral para el estrato 2.
MediaE2_G = mean(DataMuestraFinal[DataMuestraFinal$Estrato == "Estrato 2",
                                  "Palabras"]$Palabras)
VarE2_G = var(DataMuestraFinal[DataMuestraFinal$Estrato == "Estrato 2",
                               "Palabras"]$Palabras)
MediaE2_G
```

```
## [1] 73.36735
```

```r
VarE2_G
```

```
## [1] 1803.862
```

```r
# Media y varianza muestral para el estrato 3.
MediaE3_G = mean(DataMuestraFinal[DataMuestraFinal$Estrato == "Estrato 3",
                                  "Palabras"]$Palabras)
VarE3_G = var(DataMuestraFinal[DataMuestraFinal$Estrato == "Estrato 3",
                               "Palabras"]$Palabras)
MediaE3_G
```

```
## [1] 68.18519
```

```r
VarE3_G
```

```
## [1] 1189.464
```

La información anterior se resume en la siguiente tabla:

Estrato	Párrafos totales	Párrafos muestreados	Media muestral	Varianza muestral
Estrato 1	305	35	106.4571	3476.255
Estrato 2	560	49	73.36735	1803.862
Estrato 3	447	27	68.18519	1189.464

El número promedio de palabras por párrafo en el libro se puede estimar como sigue:

$$\hat{\mu}_{est} = \frac{1}{N}\sum_{h=1}^{3} N_h \bar{y}_h = \frac{1}{1312}\left(305 * 106.4571 + 560 * 73.36735 + 447 * 68.18519\right) = 79.29414 \approx 79$$

Se estima que el número promedio de palabras por párrafo en el libro es igual a 79.

Ahora, se calcula la varianza de la media estimada de la siguiente manera:

$$Var[\hat{\mu}_{est}] = \frac{1}{N^2} * \sum_{h=1}^{3} N_h^2 Var[\hat{\mu}_h]$$

Donde:

$$Var[\hat{\mu}_h] = \left(1 - \frac{n_h}{N_h}\right)\frac{\sigma_h^2}{n_h}$$

Entonces,

$$\hat{Var}[\hat{\mu}_1] = \left(1 - \frac{35}{305}\right)\frac{3476.255}{35} = 87.92401$$

$$\hat{Var}[\hat{\mu}_2] = \left(1 - \frac{49}{560}\right)\frac{1803.862}{49} = 33.59233$$

$$\hat{Var}[\hat{\mu}_3] = \left(1 - \frac{27}{447}\right)\frac{1189.464}{27} = 41.39671$$

Con la información calculada, se obtiene la estimación de la $Var[\hat{\mu}_{est}]$:

$$\hat{Var}[\hat{\mu}_{est}] = \frac{1}{1312^2}\left[(305^2)(87.92401) + (560^2)(33.59233) + (447^2)(41.39671)\right] = 15.67677$$

Ahora, se obtiene el respectivo intervalo de confianza. Para esto, se debe obtener el **Error Estándar** con base en el tamaño de la muestra. Como el tamaño de muestra obtenido es mayor a 30 ($n = 111$), entonces se utiliza el **Límite del error estándar** $B = Z_{\alpha/2}\sqrt{\hat{Var}[\hat{\mu}_{est}]}$

Por lo tanto, para obtener un intervalo de confianza para el valor medio de palabras por párrafo en el libro con una confianza del 95%, se tiene lo siguiente:

$$B = Z_{0.025}\sqrt{15.67677} = 1.96 * 3.95939 = 7.760404$$

$$\hat{\mu}_{est} \pm B \Longleftrightarrow 79 \pm 7.760404$$

Al realizar los cálculos y redondear al entero más cercano, se obtiene el siguiente intervalo $(71, 87)$, es decir, el número promedio de palabras por párrafo para el libro *Elementary Survey Sampling* se encuentra entre 71 y 87 palabras, esto con una confianza del 95%.

6 Inferencias con base en la M.A.E. obtenida

Con los datos obtenidos en la M.A.E. se responde la siguiente pregunta:

¿Existe diferencia significativa entre el número promedio de palabras por párrafo entre los estratos 1 y 2?

Para realizar dicha verificación, se utiliza la siguiente prueba de hipótesis con una significancia del 5%:

$$H_0 : \mu_1 - \mu_2 = 0$$
$$H_1 : \mu_1 - \mu_2 \neq 0$$

Como los tamaños de muestra para ambos estratos son mayores o iguales a 30 párrafos y no se conocen las varianzas poblacionales sino las muestrales, entonces se utiliza el siguiente estadístico de prueba aproximado:

$$Z_c = \frac{(\bar{y}_1 - \bar{y}_2)}{\sqrt{\frac{S_1^2}{n_1}(1 - \frac{n_1}{N_1}) + \frac{S_2^2}{n_2}(1 - \frac{n_2}{N_2})}} \sim n(0, 1)$$

Al reemplazar los respectivos valores en la expresión anterior, se obtiene:

$$Z_c = \frac{(106.4571 - 73.36735)}{\sqrt{\frac{3476.255}{35}(1 - \frac{35}{305}) + \frac{1803.862}{49}(1 - \frac{49}{560})}} = 3.001761$$

Ahora, se calcula el **valor P** para la prueba de la siguiente manera:

$$Val_P = 2P(Z > |Z_c|) = 2P(Z > |3.001761|)$$

```r
# Cálculo de valor P.
2*pnorm(3.001761, lower.tail = FALSE)
```

```
## [1] 0.002684228
```

Por lo tanto, $Val_P = 0.002684228$

Al comparar el Valor P obtenido con el nivel de significancia de la prueba del 5%, se puede concluir que se rechaza la hipótesis nula y, por tanto, el número promedio de palabras por párrafo en el Estrato 1 es significativamente diferente con respecto al Estrato 2 con una siginificancia del 5%.