Practica Parcial 1

Estudiantes de Pregrado en Estadística

Ivan Santiago Rojas Martinez
Ronald Gabriel Palencia

Docente

Victor Ignacio Lopez Rios

Asignatura

Muestreo Estadistico

Sede Medellín
Septiembre 11 de 2023

Índice

Trabajo Práctico Primer Examen Parcial

Se busca calcular la proporción de libros en la sección 14, que ha sido seleccionada de forma aleatoria entre 62 secciones de la biblioteca, relacionados con economía y medio ambiente que fueron publicados después del año 2000. También queremos estimar el número total de libros en esta sección que cumplen con estos criterios. Para lograr esto, hemos decidido llevar a cabo un muestreo aleatorio simple, como se describe a continuación, con el objetivo de obtener estimaciones con un nivel de confianza del 90% y un error relativo máximo del 10%. ***(Enunciado del trabajo)***

1 Diseño del muestreo

Cada sección de la biblioteca está compuesta por 60 estantes, 30 a cada lado, y en promedio, cada estante alberga alrededor de 30 libros. Esto significa que aproximadamente hay unos 1800 libros en cada sección. Dado que la mayoría de las secciones son similares en estructura y tamaño, vamos a trabajar con $\boldsymbol{N = 1800}$ para la selección de la muestra inicial. Estos datos se tomaron gracias a la información suministrada por la persona encargada de la biblioteca y de organizar los estantes.

Además, **nuestro elemento** y **unidad de muestreo** será cada libro individual en la sección 14. La población de interés comprende todos los libros en dicha sección, y el **marco muestral** está representado por el estante correspondiente a la sección 14, que contiene todos los elementos de la población de interés. Es relevante señalar que al enumerar los libros se siguió un orden, comenzando por aquellos ubicados en la parte superior del estante y concluyendo en la zona inferior del mismo.

Tenemos como parámetros de interés la **proporción de libros** en la sección cuyo año de publicación fue después del año 2000 y el número total de libros de la misma sección que cumplen con dichas especificaciones.

2 Muestra piloto

Considerando la gran cantidad de libros disponibles y con el objetivo de obtener estimaciones precisas, optaremos por tomar una muestra aleatoria simple de 30 libros. Ya que este tamaño de muestra podría ser apropiado para calcular con precisión la varianza y determinar el tamaño definitivo de la muestra que necesitaremos para nuestras estimaciones.

Con ayuda de R se seleccionaron aleatoriamente 30 números entre 1 y 1800, se obtuvo como muestra piloto:

Id	titulo	Fecha	atributo
37	La Diosa Némesis	2003	1
129	Informe Del Estado De Los Recu	2002	1
270	Economía Del Medio Ambiente	1998	0
299	Reforma Agraria En América Lat	1972	0
330	Converting Land from Rural To	1755	0
382	Educación Ambiental Y Desarrol	1999	0
471	Guías Para La Evaluación De Im	1997	0
485	Gestión Ambiental En América L	2002	1
597	Manejo De Los Recursos Natural	1987	0
679	Futuro Ecológico De Un Contine	1995	0
729	Colombia Antioquia Y La Cuenca	1990	0
801	Economía Y Ambiente	2005	1
852	Restauración De Ecosistemas	2003	1
874	Introducción A La Valoración A	2003	1
878	Conservación Y Uso Sostenible	1995	0
930	Estrategias Ambientales De Los	2007	1
931	Our Natural Resources And Thei	1936	0
975	El Comercio Mundial Y El Medio	1992	0
1017	Indicadores De Impacto Ambient	2006	1
1129	Santander Y Su Desarrollo Econ	1930	0
1210	Historia Del Dinero	1973	0
1301	Los Trabajos Y Los Días De Rec	1977	0
1331	El Libro De Los Oficios De Ant	2002	1
1422	International Finance And Open	1994	0
1518	La Expropiación En El Derecho	1964	0
1533	Las Luchas Agrarias En Colombi	1972	0
1615	Teoría Monetaria Internacional	2001	1
1625	El Dinero La Banca Y La Activi	1977	0
1701	Mercado De Trabajo Y Deuda Soc	1991	0
1749	El Trabajo Infantil	2011	1

Donde **Id** es el número que identifica el libro dentro de la población, **titulo** presenta un fragmento del título del libro, **Fecha** corresponde al año de publicación del libro y **atributo** es una variable categórica que toma el valor de uno si la fecha de publicación es mayor a 2000 o cero en caso contrario.

$$\epsilon = 0.10 \qquad \alpha = 0.10 \qquad N = 1800 \qquad Z_{(1-\alpha/2)} = 1.644854$$

Además tenemos que una estimación para la proporción es:

$$\hat{\mathbf{p}} = \frac{1}{30}\sum_{i=1}^{30} atributo_i = \frac{11}{30} = 0.3666667$$

Con los datos anteriores, podemos calcular el tamaño de la muestra final como:

$$n = \frac{1}{\frac{1}{N} + \frac{N-1}{N}\left(\frac{1}{\mathbf{n_o}}\right)}; \qquad \mathbf{n_o} = \frac{Z^2_{(1-\alpha/2)}\ \hat{p}(1-\hat{p})}{B_p^2} = \frac{1.644854^2[0.366(1-0.366)]}{0.0366667^2} = 467.321354$$

Así, vemos que $\mathbf{n = 371.1646787\ \approx 372}$, Dado que los datos se toman el mismo día y en la misma sección se van a mantener las primeras 30 obervaciones como elementos de la muestra final. De esta forma se toman los otros 342 registros tomando de forma aleatoria los indices con la ayuda de la función sample() del R.

Se presentan las primeras 10 observaciones de la base de datos:

Id	Fecha	atributo
1	1996	0
3	2006	1
9	2000	0
13	1959	0
20	1994	0
24	2003	1
37	2003	1
40	1996	0
47	2013	1
48	1997	0

Ahora, tenemos que la estimación para la proporción con la muestra final es:

$$\hat{\mathbf{p}} = \frac{1}{372}\sum_{i=1}^{372} atributo_i = \frac{90}{372} = 0.2419$$

3 Estimación de la proporción de libros cuyo año de publicación fue después de 2000.

Para realizar dichas estimaciones recordemos que tenemos $\hat{\mathbf{p}} = 0.2419$ y que los intervalos de confianza para la proporción tienen la forma:

$$\hat{p} \quad \pm \quad t_{(1-\alpha/2,\ n-1)}\sqrt{\hat{V}(\hat{p})} \quad ; \text{con} \quad \hat{V}(\hat{p}) = \frac{\hat{p}(1-\hat{p})}{n-1}\left(\frac{N-n}{N}\right)$$

De esta forma y retomando los datos anteriores tenemos que un intervalo al 90% para dicha proporción estará dado por:

$$0.2419 \quad \pm \quad t_{(1-\alpha/2,\ n-1)}\sqrt{\frac{0.2419(1-0.2419)}{371}\left(\frac{1800-372}{1800}\right)} \quad = \quad 0.2419 \quad \pm \quad 0.0326555$$

Así, con una confianza del 90% la proporción de libros en la sección 14 cuyo año de publicación fue después de 2000 estará entre 0.2092799 y 0.274591

4 Estimación del número total de libros de la sección 14 cuyo año de publicación fue después de 2000.

Teniendo en cuenta que ya tenemos las estimaciones y los intervalos para la proporción podemos calcular dichos valores para el total multiplicandolos por N = 1800, de esta forma una estimación para el total de libros en la sección 14 cuyo año de publicación fue después de 2000 es:

$$\hat{\mathbf{A}} = 0.2419355(1800) = 435.483871 \ \approx \ 436$$

Del mismo modo, un intervalo de confianza al 90% para el total es:

$$N(0.2092799, 0.274591) = (376.703905, 494.2638369) \ \approx \ (376, 495)$$

Por lo tanto, con una confianza del 90% el total de libros en la sección 14 cuyo año de publicación fue después de 2000 estará entre 376 y 495

5 Observaciones

Durante el transcurso de la práctica, nuestra comprensión sobre el proceso de muestreo fue totalmente diferente. No teníamos idea de la cantidad de control y atención a los detalles que se requieren al llevar a cabo la toma de los datos. Además, descubrimos que elaborar una estrategia adecuada para el muestreo es bastante complejo.

Uno de los desafíos más notables se presenta al tomar la muestra, especialmente cuando la población a muestrear es considerablemente grande. Mantener el control para asegurarse de no dejar ningún elemento fuera o simplemente perder la cuenta o el orden en el que se seleccionan los elementos requiere de mucho cuidado y concentración, ya que la calidad de los resultados depende de ello.

Lo que pudimos observar es que es muy útil buscar la orientación de personas con experiencia o que están en contacto continuo con la población de interés. Estas personas pueden guiar al investigador en la realización de procesos que pueden resultar desconocidos para él o proporcionar información sobre la clasificación y distribución de los elementos en dicha población. Gracias a estos aspectos, fue mucho más fácil llevar a cabo el muestreo y realizar las estimaciones necesarias.

6 Anexo Código

```
set.seed(0)
libros <- sample(1:1800,30) %>% sort(decreasing = F)
```

```
muestra_piloto <- readxl::read_xlsx("Data/Muestra_piloto.xlsx")
muestra_piloto %<>%  mutate(titulo = str_sub(titulo,1,30)
                            , atributo = ifelse(Fecha>2000,1,0))
kable(muestra_piloto)
```

```
eps <- 0.10
alfa <- 0.10
N <- 1800
p_est <- sum(muestra_piloto[,4])/30
```

```
n_o <- ((1.644854^2)*(p_est*(1-p_est)))/((eps*p_est)^2)
n <- 1/((1/N)+(((N-1)/N)*(1/n_o)))
```

```
set.seed(578)
indices <- seq(1:1800)[-libros]
libros_index <- sample(indices,342) %>% sort(decreasing = F)
```

```
muestra_final <- readxl::read_xlsx("Data/Muestra_final.xlsx")  %>%
  mutate(atributo = ifelse(Fecha>2000,1,0))
```

```
kable(head(muestra_final, n=10))
```

```
p_est <- sum(muestra_final[,3])/372
```

```
B <- qt(1-alfa/2, 371)*sqrt(((p_est*(1-p_est))/371)*((1800-372)/1800))
lim_inf <- p_est - B
lim_sup <- p_est + B
```

```
A <- p_est * 1800
lim_inf_A <- lim_inf * 1800
lim_sup_A <- lim_sup * 1800
```