OPSO79-1-UCSH2021

Inferencia desde muestras complejas en R: la lógica del muestreo y la inferencia

26/11/2021

La lógica del muestreo e inferencia

Muestreo

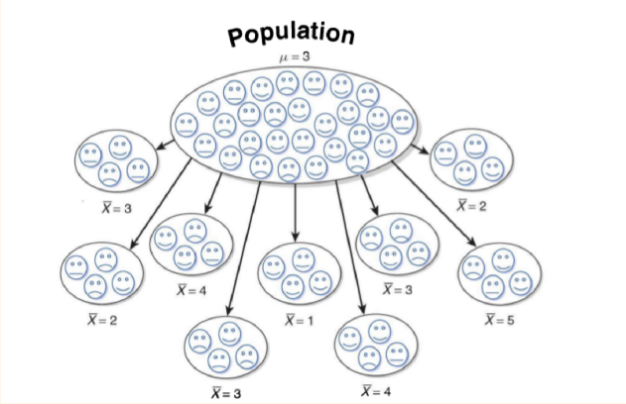
Recursos limitados nos impiden encuestar a toda una población.

Una salida es encuestar a una parte. ¿Que parte encuestar?, ¿A cuántas personas encuestar?

Si elegimos a la muestra de la forma correcta, podremos inferir hacia la población.

Esta inferencia la haremos con un error conocido.

Toda la lógica del muestreo es conocer y tratar de minimizar este error.



Un poco de nomenclatura

Para conocer los parámetros usamos estimadores.

Con el estimador calculamos el parametro poblacional en base a una serie de datos observados.

MEDIDAS	POBLACION (parámetro)	MUESTRA (estadístico)
Media aritmética	μ	$\overline{x}$
Varianza	σ^2	s^2
Desviación estándar	σ	s
Proporción	π	p
Tamaño	N	n

Muestra

El supuesto central detrás de una **muestra estadísticamente representativa** es el de una **muestra aleatoria**

La selección aleatoria no supone que el grado de imprecisión asociado a las estimaciones sea necesariamente pequeño. Sí permite conocer la magnitud de la imprecisión.

Además, todos los elementos tienen una probabilidad conocida y distinta de cero de ser elegidos

Con esto, podemos conocer el error asociado a la estimación

No todas las formas de seleccionar a una muestra son "probabilísticas". El muestreo por cuotas, por ejemplo, es "no probabilistico".

Muestreo probabilístico

Sí tenemos un listado de los elementos (marco muestral), y seleccionamos aleatoriamente solamente algunos elementos a estudiar, estamos en un **muestreo probabilistico**. Esta es la base de otros diseños más complejos.

La función teórica de la distribución normal nos permite establecer un intervalo de posibles valores para nuestra estimación.

Un ejemplo simple de selección aleatoria.

Creamos una base de datos de 19.678.363 casos, el número de personas en Chile según el INE, que contenga 1.492.522 extranjeros.

```
poblacion<-data.frame(id=c(seq(1:19678363)),
                      extranjers=c(rep("ext",1492522),
                                   rep("nac",19678363-1492522)))
```

Muestreo probabilístico

```
table(poblacion$extranjers)
```

```
##
##      ext      nac
##  1492522 18185841
```

```
prop.table(table(poblacion$extranjers))
```

```
##
##        ext        nac
## 0.07584584 0.92415416
```

La selección aleatoria

Saquemos una muestra de 500, ¿qué % de extranjeros aparecerá?

```
muestra<-sample_n(poblacion,500)
```

Y observamos su distribución:

```
prop.table(table(muestra$extranjers))
```

```
##
##   ext   nac
## 0.068 0.932
```

La selección aleatoria

¿Fue suerte?, veamos de nuevo...

```
muestra<-sample_n(poblacion,500)
prop.table(table(muestra$extranjers))
```

```
##
##   ext   nac
## 0.084 0.916
```

De nuevo

```
muestra<-sample_n(poblacion,500)
prop.table(table(muestra$extranjers))
```

```
##
##   ext   nac
## 0.084 0.916
```

La selección aleatoria

Cada vez que actualicemos esta presentación, la muestra seleccionada se nos va a modificar.

Esto generará que cualquier análisis que queramos hacer con los datos no será reproducible, ya que cada vez que saquemos una muestra esta se modificará.

Manteniendo el azar, R nos permite fijar la selección aleatoria. Esto se logra "fijando una semilla".

```
set.seed(17) ## número arbitrario
```

Con esto, la distribución de la muestra que saquemos siempre será 0,064 (6,4%).

La selección aleatoria

```
muestra<-sample_n(poblacion,500)
prop.table(table(muestra$extranjers))
```

```
##
##   ext   nac
## 0.064 0.936
```

Efectivamente. Veamos de nuevo...

```
set.seed(17)
muestra<-sample_n(poblacion,500)
prop.table(table(muestra$extranjers))
```

```
##
##   ext   nac
## 0.064 0.936
```

La selección aleatoria

Las estimaciones que obtengamos de cada muestra de la población se ubicará en torno al parámetro poblacional.

Ahora veremos que tan cierta es esta afirmación con 500 muestras

```
set.seed(1917)  # semilla

base<-data.frame(muestra=c(1:500),  # data vacía
                 ext=rep(NA,500),
                 nac=rep(NA,500))

# loop
for(i in 1:nrow(base)){
  base[i,2:3]<-sample_n(poblacion,size=500) %>%
    select(extranjers) %>%
    table() %>% prop.table()
}
```

La selección aleatoria

```
head(base,10) %>% knitr::kable()
```

muestra	ext	nac
1	0.092	0.908
2	0.080	0.920
3	0.058	0.942
4	0.076	0.924
5	0.062	0.938
6	0.062	0.938
7	0.070	0.930
8	0.078	0.922
9	0.082	0.918
10	0.098	0.902

La selección aleatoria

¿El promedio del % de extranjeros entre todas las muestras?

```
mean(base$ext)   ## OMG! en la población era 0.07584584
```

```
## [1] 0.076344
```

Solo hay una diferencia de 0.049816% (menos de un 1%).

La selección aleatoria

Estimar desde muestra

Cada % de extranjeros estimado con 1 muestra tiene un error asociado.

Por ejemplo: *en la población se estiman entre un 6,5% y 8,5% de extranjeros*

Esta forma de medir el error utiliza un "intervalo de confianza" (+-1%)

La teoría de muestras nos permite conocer este error, en base a:

- tamaño de la muestra (n)
- nivel de confianza en puntaje Z ($z_{a/2}$)
- variabilidad de los datos ($\overline{p}(1-\overline{p})$)

$$[\overline{p} \pm z_{a/2}\sqrt{\frac{\overline{p}(1-\overline{p})}{n}}]$$

Estimar desde muestra

En simple, definamos objetos

```
p <- as.vector(prop.table(table(muestra$extranjers)))[1]
n <- nrow(muestra)
z <- 1.96 # para 95% de confianza
```

Calculamos intervalo

```
ci <- z * (sqrt(p*(1-p)/n))
ci
```

```
## [1] 0.02145354
```

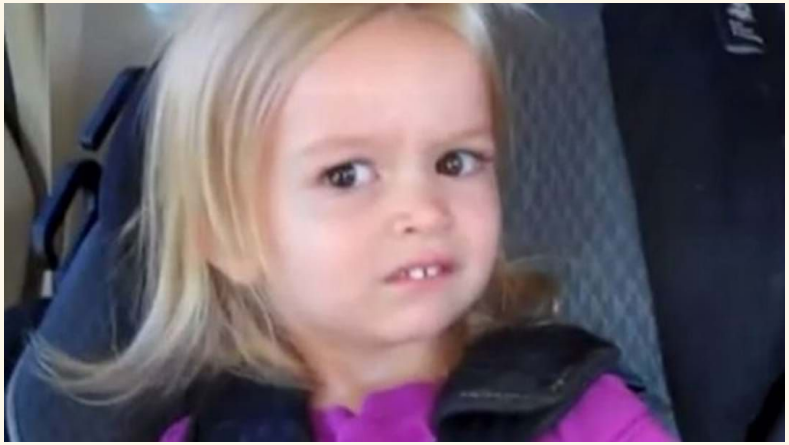
Con un 95% de confianza, y con una muestra probabilística de 500 observaciones, calculamos que el porcentaje de extranjeros en la población está entre 0.0425465 y 0.0854535 (0.064 ± 0.0214535).

La estimación parecer ser correcta. En la población la proporción es 0.075

Estimar desde muestra

¿Que significa el 95% de confianza?

Que si sacamos 20 muestras, las estimaciones desde una de estas no contendría dentro de sus intervalos el valor poblacional.



Veamoslo con 20 de las 500 muestras que sacamos.

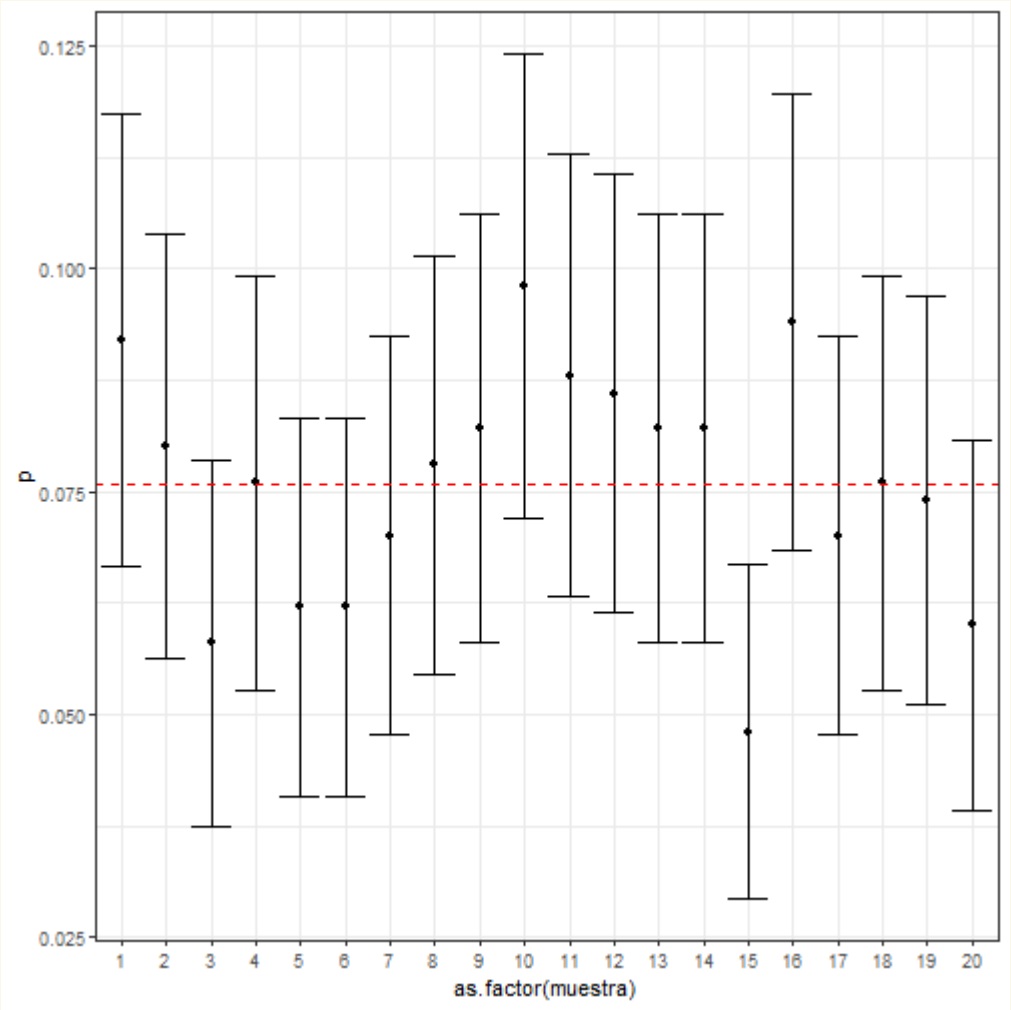
Estimar desde muestra

Calculamos los intervalos de confianza de cada muestra

```
submuestra <- base[1:20,c(1,2)] %>%
  rename(p=ext) %>%
  mutate(limite_inferior= p - (z * (sqrt(p*(1-p)/n))),
         limite_superior= p + (z * (sqrt(p*(1-p)/n))))
```

muestra	p	limite_inferior	limite_superior
1	0.092	0.0666658	0.1173342
2	0.080	0.0562201	0.1037799
3	0.058	0.0375115	0.0784885
4	0.076	0.0527719	0.0992281
5	0.062	0.0408618	0.0831382

Visualización estimaciones



Elementos a destacar

Vimos como estimar proporciones, para medias, medianas, varianzas y totales la formula debe ser ajustada.

Para calcular los IC no hemos consideramos el tamaño de la población.

¿No es relevante? Lo es, pero marginalmente

Lo que es más relevante es el tamaño de la muestra, nivel de confianza y variabilidad de los datos

Elementos que la formula ya vista resume y que podemos re estructurar para calcular el tamaño de una muestra:

$$n = (1 - \frac{n}{N}) * \frac{z^2_{a/2}(s^2)}{e^2}$$

Calcular muestras

Hay algunas páginas como SurveyMonkey o Calculator desde donde se puede calcular el tamaño de una muestra en base a los criterios de la formula.

Acá utilizaremos un paquete para eso

```
library(samplingbook)
```

Tipos de muestreo

Recursos web utilizados

Xaringan: Presentation Ninja, de Yihui Xie. Para generar esta presentación.

Bibliografía utilizada

Lohr, S. L. (2000). *Muestreo: Diseño y Análisis*. 519.52 L6. International Thomson Editores.

Vivanco, M. (2006). "Diseño de Muestras En Investigación Social". In: *Metodolog\'ias de La Investigación Social. Introducción a Los Oficios.* Santiago: LOM, pp. 141-168.