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

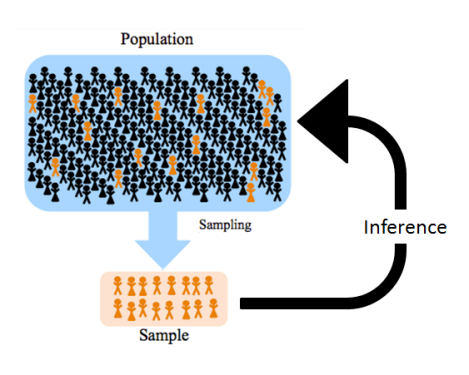
Recursos limitados nos impiden encuestar a toda una población.

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

Para tener una **muestra estadísticamente representativa** el supuesto de la **muestra aleatoria** es fundamental

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...

Y nos movemos dentro del mundo del **muestreo probabilísitico**

Muestreo probabilístico

Sí tenemos un listado de los elementos (marco muestral), y seleccionamos aleatoriamente solamente algunos elementos a estudiar, estamos en un **muestreo probabilístico**.

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
## 0.078 0.922
```

La selección aleatoria

¿Fue suerte?, veamos de nuevo...

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

```
##
##  ext  nac
## 0.08 0.92
```

De nuevo

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

```
##
##   ext   nac
## 0.076 0.924
```

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()
}

baseinicial <- base
```

La selección aleatoria

```
head(baseinicial,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

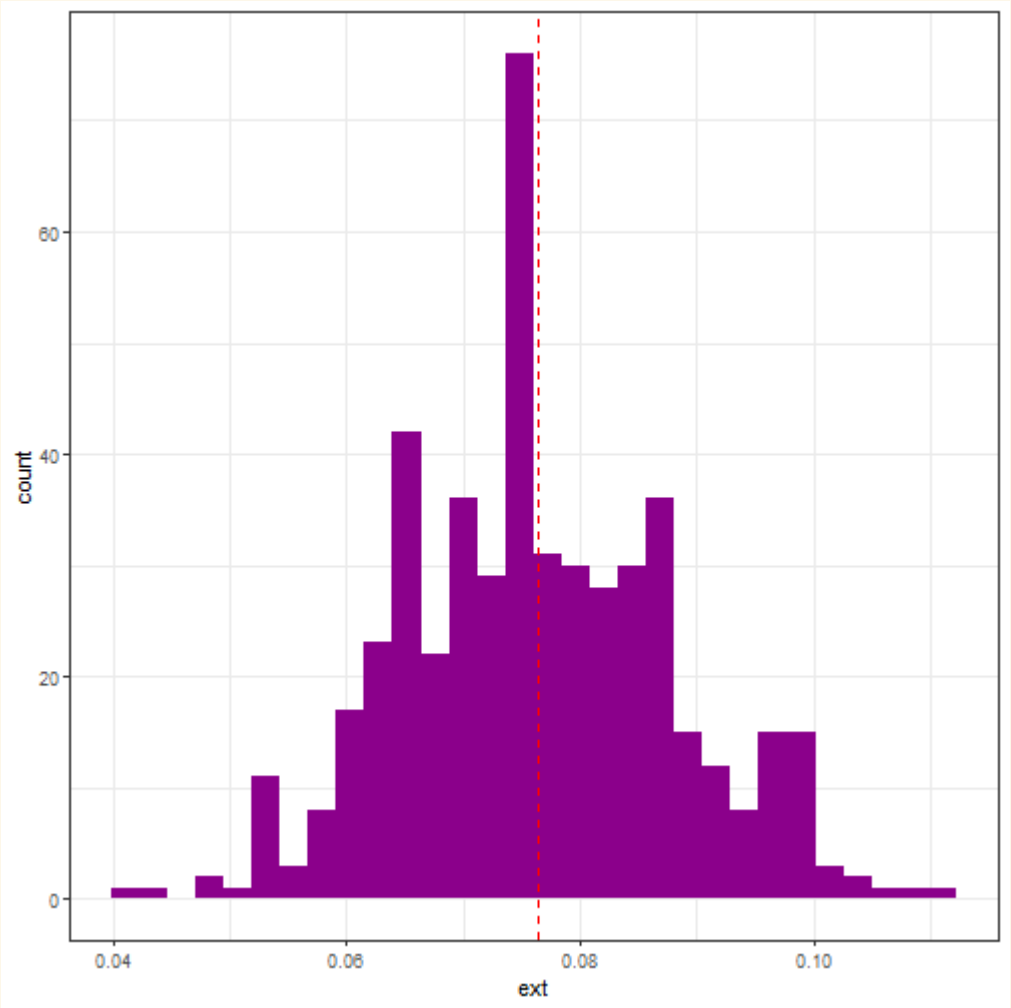
¿El promedio del % de extranjeros entre todas las muestras?

```
mean(baseinicial$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

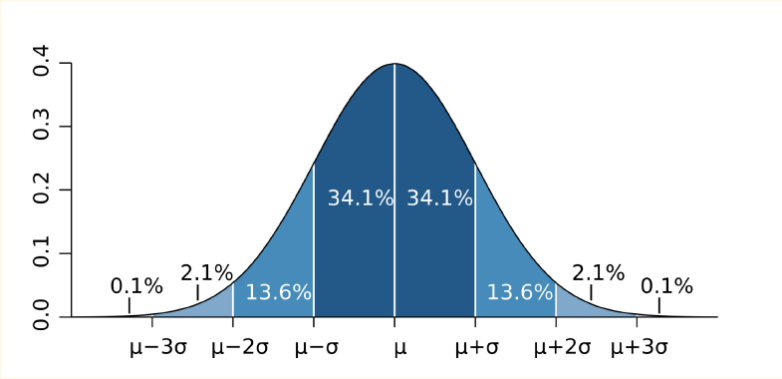


La selección aleatoria

¿Que forma tiene la distribución?

Distribución teórica normal. ¿Y que sabemos de esta?

El el 68% de los casos se encuentran a +/- 1 SD del promedio, y el 95% a +/- 1.96 SD (puntaje Z)



Estimar desde muestra

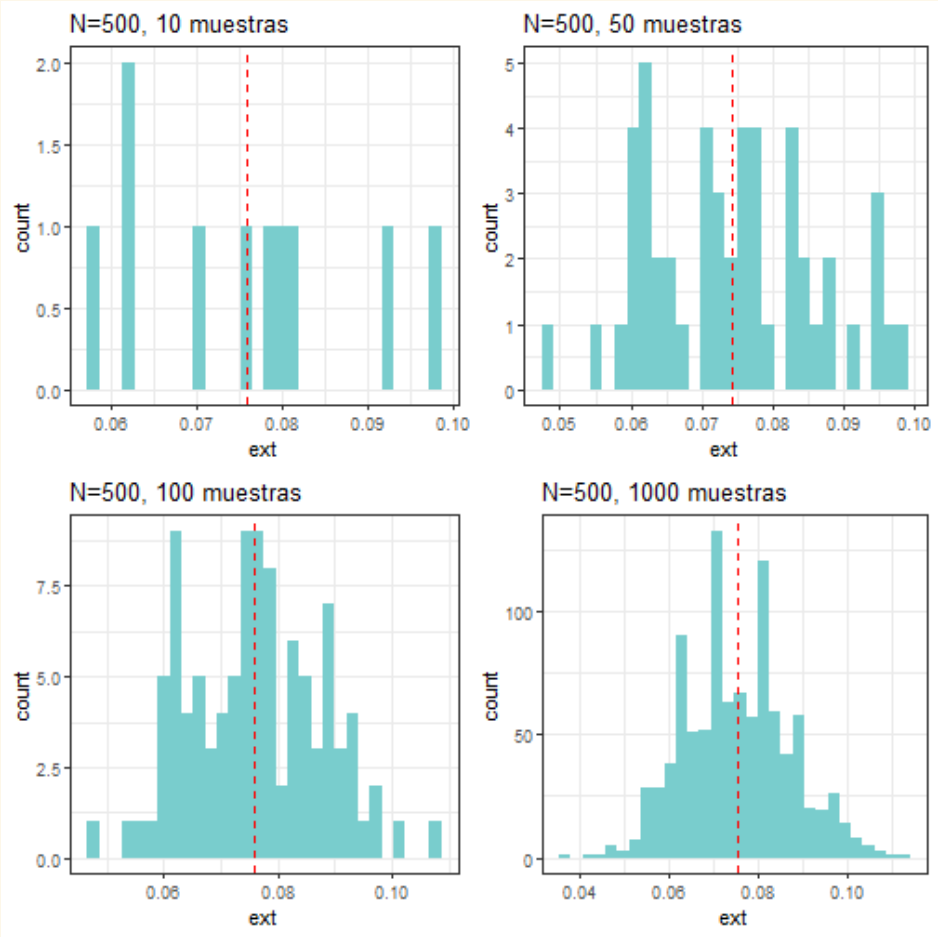
Además, mientras más muestras sacamos, más se distribuyen normal (**ley de los grandes números**)

Y mientras más grande es cada muestra de las muchas que saquemos, la distribución muestral de este del estadístico tenderá a ser normal (**Teorema del límite central**).

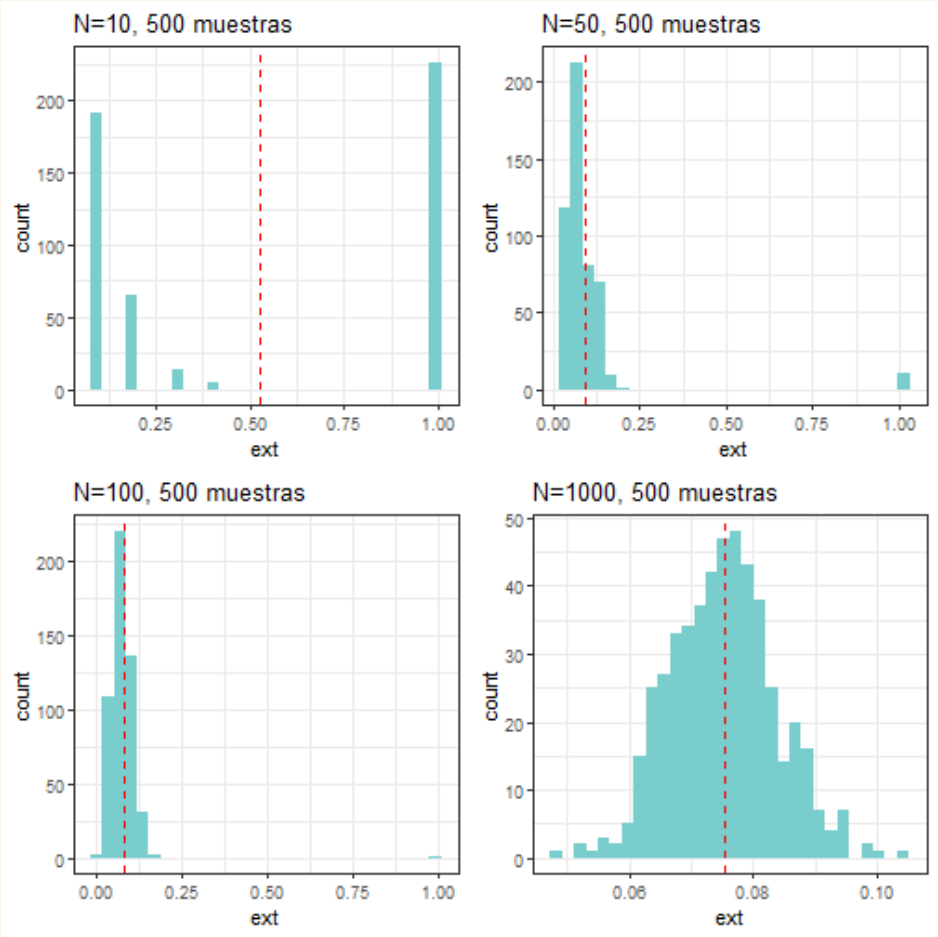
Las medias de muestras grandes y aleatorias son aproximadamente normales

Veamos ambas propiedades

Ley de los grandes números



Teorema del límite central



Estimar desde muestra

Conocemos la teoría de muestras y la demostramos sacando muchas muestras.

En la práctica solo tenemos recursos para seleccionar 1 muestra.

Desde los datos de esta muestra debemos **estimar** el valor que nos interesa de la población.

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%)

Estimar desde muestra

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

- tamaño de la muestra (n)
- nivel de confianza en puntaje Z (z)
- variabilidad de los datos (p(1-p))

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

- tamaño de la muestra: 500 observaciones
- nivel de confianza en puntaje Z: 1,96
- variabilidad de los datos: 6,4% (o lo que de la muestra)

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).

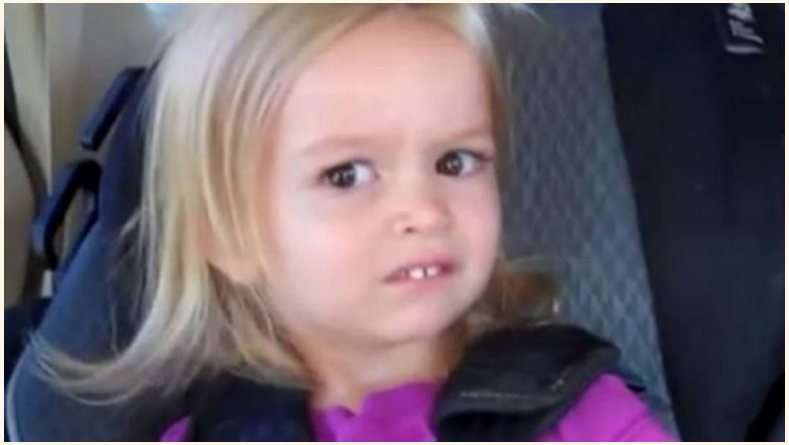
La estimación parecer ser correcta. En la población la proporción es 0.075

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.

O que si sacamos 100 muestras, las estimaciones desde cinco de estas no contendría dentro de sus intervalos el valor poblacional.



Veamoslo con 20 de las 500 muestras que sacamos.

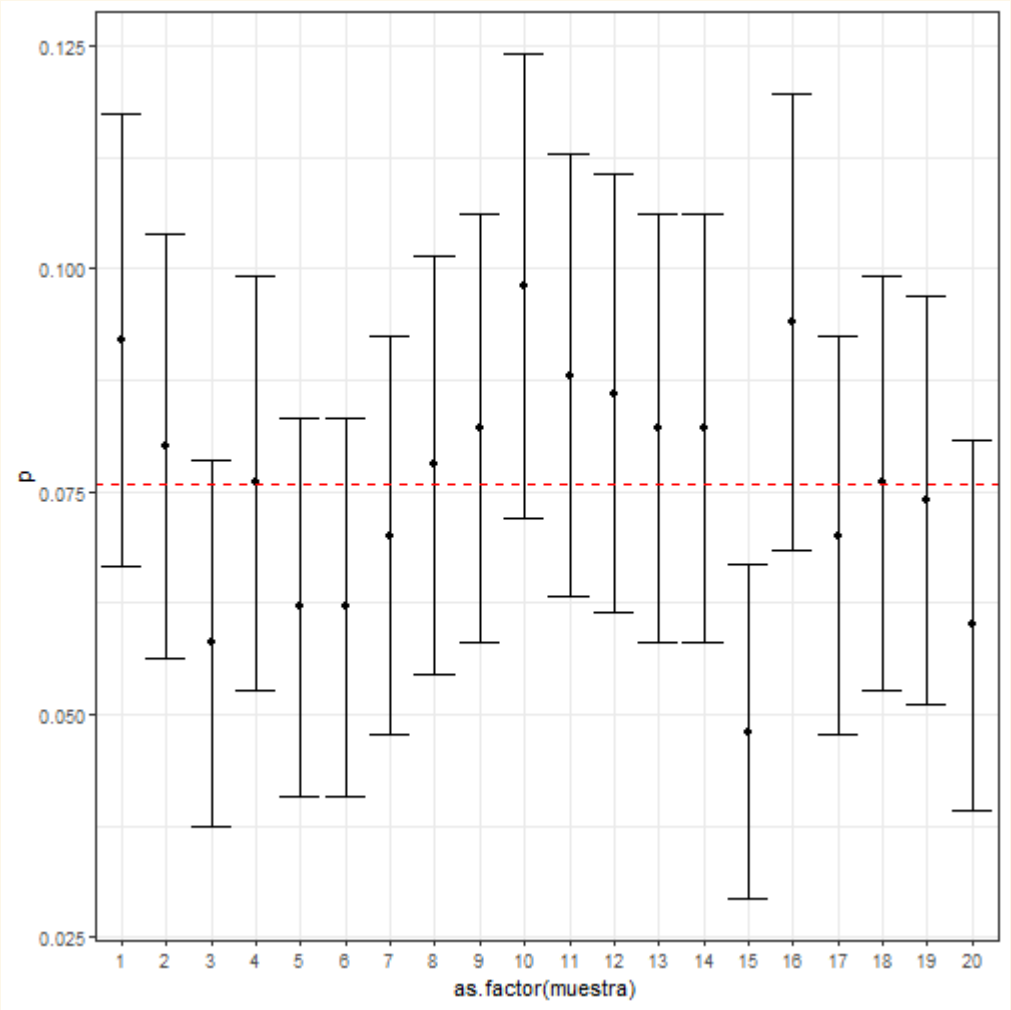
Estimar desde muestra

Calculamos los intervalos de confianza de cada muestra

```
submuestra <- baseinicial[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 estimar 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

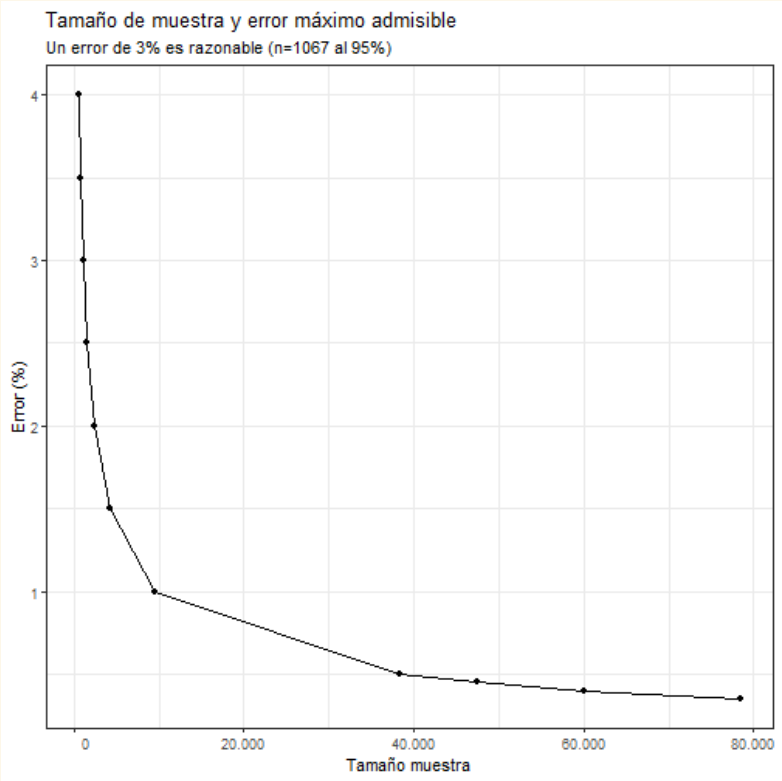
Elementos que la formula ya vista resume y que podemos re estructurar para calcular el tamaño de una muestra (agregando tamaño de población):

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

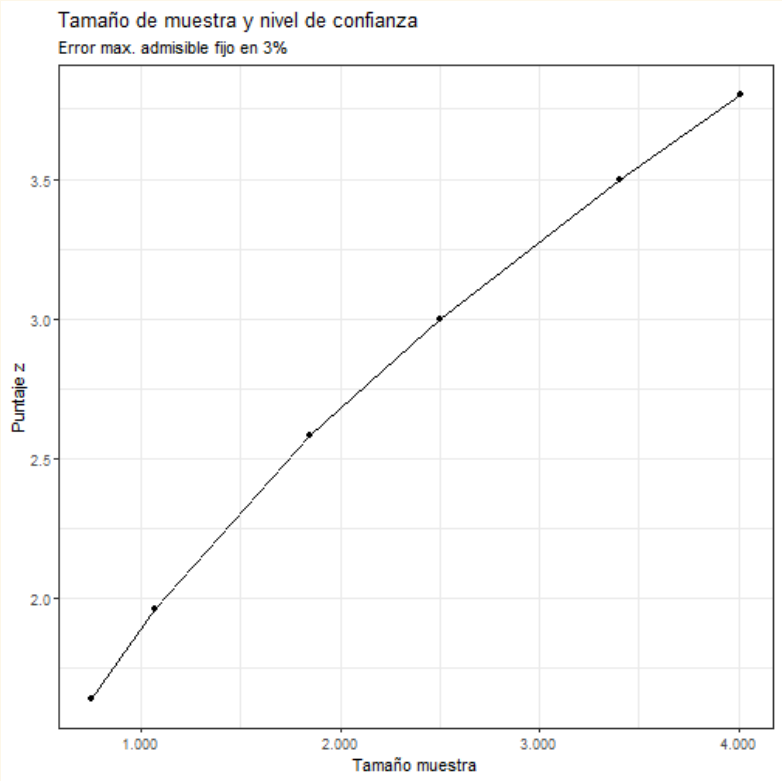
Determinando un nivel de confianza (z), un error esperado (e), teniendo antecedentes de la variabilidad de los datos (s2) y conociendo el N de la población... podemos calcular una muestra.

Tamaño muestra y error

Si definimos errores elevados la muestra es pequeña.

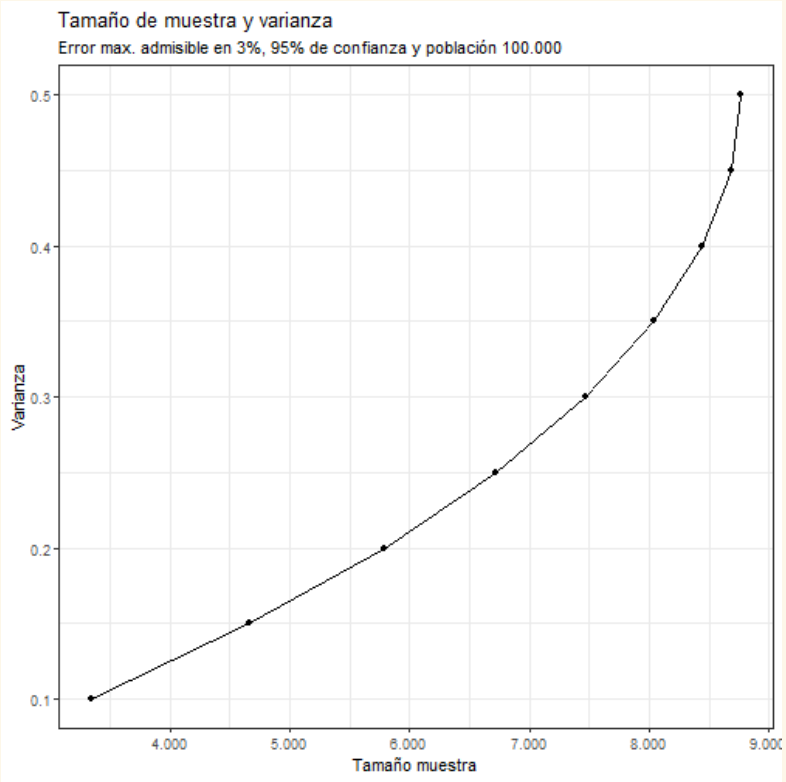


n muestra y nivel de confianza



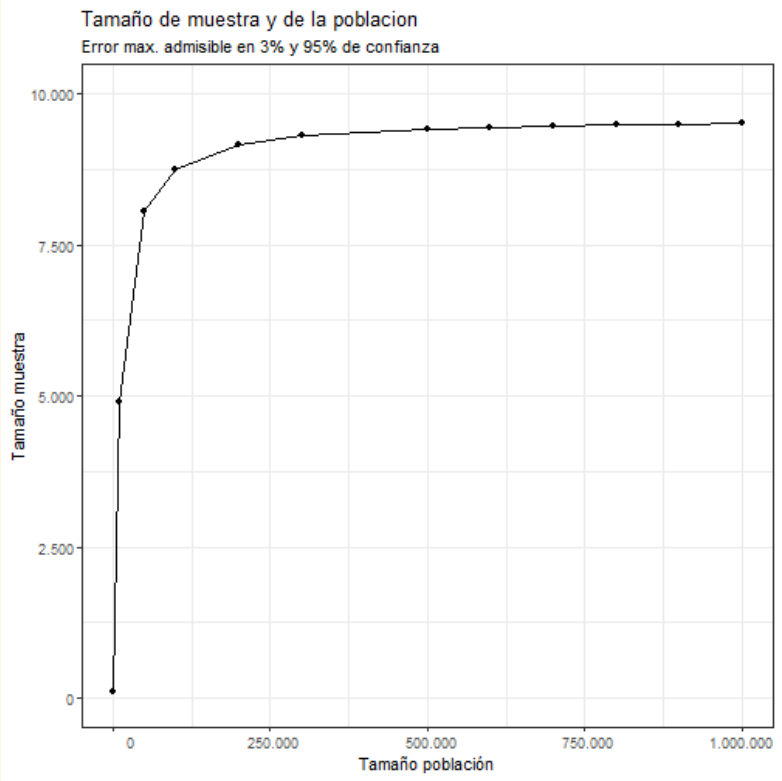
n muestra y variabilidad

Se considra la varianza de una pregunta de investigación previa. A mayor varianza, más casos se necesitarán. Si no hay información se supone varianza máxima (`0.5^2`)



n muestra y N población

En poblaciones superiores a 100.000 casos la influencia del tamaño de la población es ínfima



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)
```

Si tenemos variabilidad como proporción *dummy* (P)

```
sample.size.prop(e=0.03, P = 0.75, N = 19000000, level = 0.95)
```

```
##
## sample.size.prop object: Sample size for proportion estimate
## With finite population correction: N=1.9e+07, precision e=0.03 and expected
##
## Sample size needed: 801
```

Si tenemos variabilidad de variable numérica (SD) (y no tenemos el N)

```
sample.size.mean(e=0.03, S = 0.5, N = Inf, level = 0.95)
```

Tipos de muestreo

Presentación general del estratificado y por conglomerados

Muestreo

Probabilístico

Muestreo aleatorio simple

Muestreo estratificado

Muestreo por conglomerados

No probabilístico

Muestreo por cuotas (no lo veremos)

Muestreo aleatorio simple (MAS)

Se deja actuar libremente al azar (con o **sin reemplazo**, unidades de la muestra distintas)

Sirve de referencia y base para los demás tipos de diseños

Es monetápica, autoponderada y equiprobable.

Ventaja:

1. Sencillez de las fórmulas.
2. Precisión de la estimación

Desventaja:

1. Necesitas el listado de todos los elementos (marco muestral)
2. Aumenta el error al estimar desagregaciones

Con diseños más complejos podemos mantener la misma precisión, pero necesitando menos elementos en la muestra e incluso no necesitar un marco muestral

Marcos muestrales

La población esta claramente definida y acotada: alumnas/os del OPSOUCSH.

En cambio, en relación al **marco muestral**, el listado de los elementos de la población, no tenemos la certeza de su cobertura, actualización, etc.

Siempre es necesario evaluar la calidad del marco muestral, dado que será la fuente de la información.

Además, es relevante que se explicite como acceder a las unidades seleccionadas.

Muestreo estratificado

Estrato: Grupo homogéneo de elementos compuesto por una variable auxiliar

Los estratos son diferentes entre si, la varianza dentro de los estratos es pequeña, pero entre los estratos es grande.

- sector económico, región y rural/urbano

Agrupar en grupos homogéneos, aumenta la precisión de la estimación.

Con esto, se puede asegurar precisión para cada estrato.

La selección de elementos se realiza después de establecidos los estratos y asignado cada elemento al estrato de pertenencia.

Es clave contar con una variable de estratificación, que asigne un estrato a cada unidad.

Muestreo estratificado

Lo más básico es elegir para cada estrato un número proporcional de elementos (**Afijación proporcional**).

Si el estrato es más grande, debemos seleccionar más elementos de ese estrato

En la práctica esto nunca se hace.

Se ocupa la **afijación óptima**.

El tamaño de la muestra seleccionada para cada estrato varía proporcionalmente y por la varianza de cada estrato

Si el estrato es muy heterogéneo necesitamos más casos

Como resultado, nuestras distintas observaciones tendrán diferentes "pesos" por los estratos a los que pertenecen.

Estrato pequeño y homogeneo tendrá pocos casos en la muestra, los que pesarán mucho (factor de expansión)

Muestreo por conglomerados

La lógica aplicar dos veces el muestreo:

- para seleccionar conglomerados a estudiar
- para seleccionar a los elementos dentro de los conglomerados seleccionados

"Los conglomerados son parecidos entre sí, por tanto se seleccionan sólo algunos para conformar la muestra. No se requiere de la lista de todos los elementos de la población, sólo el listado de elementos de los conglomerados elegidos." (Vivanco, 2006: 144).

Si no se dispone del listado de elementos a encuestar se recomienda conglomerados.

Es más barato producir el marco muestral de los conglomerados seleccionados que de todos.

Reduce los costos y recolección. Facilita supervisión.

Desventaja – incrementa el error de estimación (**efecto diseño**).

Muestreo por conglomerados

En ciertas situaciones, las unidades muestrales están “agrupadas” en forma natural.

Por ejemplo:

Viviendas de una ciudad – agrupadas en Manzanas.

Alumnos de la Universidad – agrupados en Facultades (o cursos).

Luteranos - agrupados en iglesias

Usuarios del metro – agrupados por horario-estación

CASEN y ENE

Muestreos no probabilísticos

Se caracterizan por que los elementos **no** tienen una probabilidad conocida de selección

Esto anula las herramientas elaboradas para inferir de la muestra a la población.

Es imposible conocer la magnitud del error.

El más común es el muestreo por cuotas

Se sustenta conceptualmente en que si se replican los porcentajes de la población, tendré una muestra representativa (**nunca estadísticamente**)

El trabajo del muestrista es establecer las cuotas. El del entrevistador llenarlas.

No requiere del marco muestral (listado de elementos)

Sesgo de participación (no todos los elementos tienen probabilidad de ser seccionados y tampoco se conoce esta probabilidad)

Conclusiones

Para inferir estadísticamente necesitamos un muestreo probabilístico

- Unidades elegidas aleatoriamente
- Todas las unidades de la población tienen probabilidad conocida y distinta de cero de ser seleccionadas

Desafíos del terreno: calidad de respuestas, unidades no elegibles, no contacto y no respuesta

Inferir desde muestras seleccionadas con MAS puede ser relativamente simple (similar a como lo vimos acá)

Pero hacerlo desde muestreos más complejos como conglomerados y estratos, y cuando las unidades tienen diferentes probabilidades de selección, el tema se complica.

Ahí es cuando los paquetes `survey` y `srvyr` en R serán fundamentales (próxima clase los veremos)

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.