Muestreo y selección de casos

¿Cómo seleccionar los casos del estudio?

Metodología I

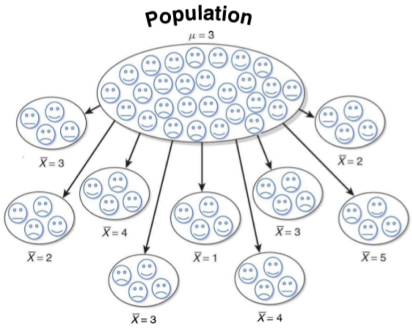
Facultad de Ciencias Sociales
Universidad de Chile
Rodrigo Medel Sierralta
Nicolás Ratto

24 de mayo de 2021

Tabla de Contenidos

¿Cómo conocer a una población?

- A menos que censemos a toda una población, no podemos conocer μ y σ^2.
- Para eso se suelen hacer muestras de la población. Para poder hablar de una población a partir de esa muestra se usa la **inferencia estadística**.



Preliminares: nomenclatura

- Para conocer los parámetros usamos estimadores. Un estimador es una regla para calcular el valor de un parametro poblacional en base a una serie de datos observados.

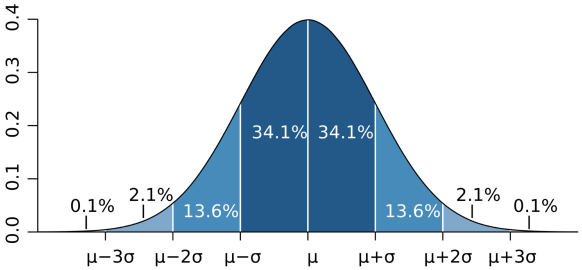
MEDIDAS	POBLACION (parámetro)	MUESTRA (estadístico)
Media aritmética	μ	$\overline{x}$
Varianza	σ^2	s^2
Desviación estándar	σ	s
Proporción	π	p
Tamaño	N	n

¿Cómo lograr representatividad?

- El supuesto central detrás de una muestra representativa es el de una muestra aleatoria (Vivanco, 2006).
  - La selección aleatoria es condición para obtener muestras representativas. Sin embargo, una muestra aleatoria no necesariamente es una muestra representativa
  - 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.
- En el diseño probabilístico de muestras todos los elementos tienen una probabilidad conocida y distinta de cero de ser elegidos, Podemos conocer el error asociado a la estimación

Probabilidad e intervalos

- La función teórica de la distribución normal nos permite establecer un intervalo de posibles valores para nuestra estimación.
- Dentro de las propiedades de la curva normal sabemos que el 68 % de los casos se encuentran a +/- 1 SD del promedio, y el 95 % a +/- 1.96 SD



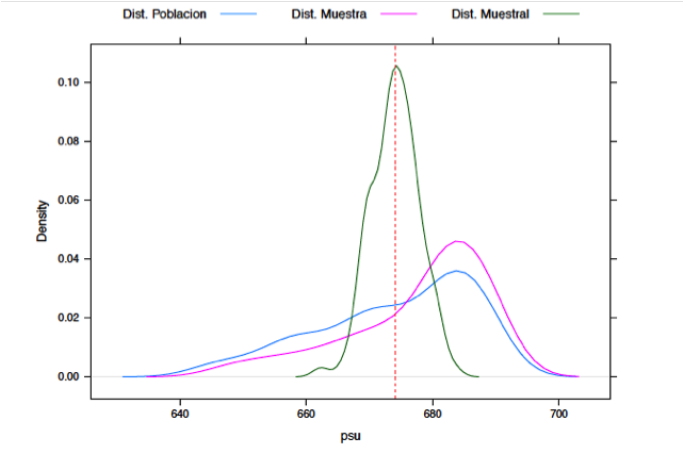
La inferencia estadística

- Sobre la base de un muestreo aleatorio, podemos usar a nuestro favor las propiedades de la curva normal.
  - **Consistencia:** A medida en que n de una muestra crece, el estimador se acerca cada vez más al verdadero valor del parámetro poblacional. Esta propiedad está asegurada por la **Ley de los Grandes Números**: A medida que el tamaño muestral crece, tiende a acercarse al parámetro poblacional.
  - **Teorema del Límite central**: A medida en que crece el tamaño de la muestra de donde se extrae el estimador, la distribución muestral de este estadístico tenderá a ser normal.

Distribuciones

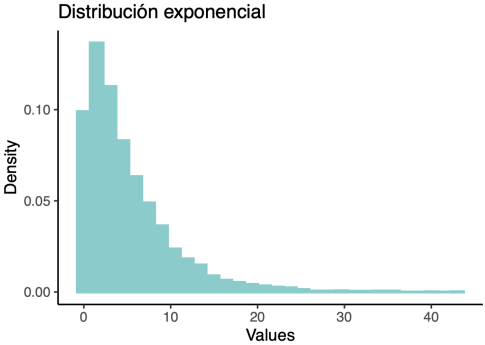
- Las estadísticas obtenidas por medio de los estimadores son aleatorias y varían de muestra en muestra. Por eso la importancia de distinguir entre distintos tipos de distribuciones.
  - La distribución poblacional.
  - La distribución en la muestra.
  - La distribución muestral de un estadístico.

Distribuciones



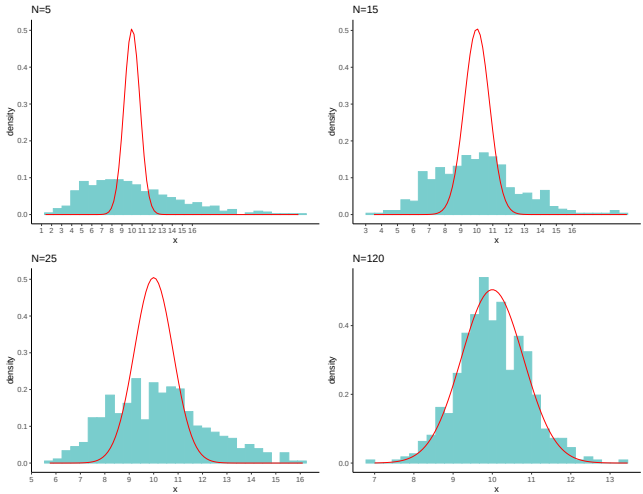
Histograma de la distribución exponencial

```
library(ggplot2)
library(dplyr)

expdist <- rexp(5000, 0.2)
ggplot(data.frame(x = expdist),
    aes(x = x)) + geom_histogram(aes(y = ..density..),
    colour = "darkslategray3",
    fill = "darkslategray3") +
    ggtitle("Distribución exponencial") +
    xlab("Values") + ylab("Density") +
    theme(plot.title = element_text(size = 9,
        face = "bold"), axis.text = element_text(size = 6),
        axis.title = element_text(size = 9)) +
    theme_classic(base_size = 6)
```



```r
expmatrix <- matrix(rexp(5 * 500, 0.1), 500)
expmeans <- apply(expmatrix, 1, mean)

plot1 <- ggplot(data.frame(x = expmeans), aes(x = x)) + geom_histogram(aes(y = ..density..), colour = "darkslategray3",
    fill = "darkslategray3") + scale_x_continuous(breaks = c(1:16)) + theme_classic() + stat_function(fun = dnorm,
    color = "red", args = list(mean = 10, sd = sqrt(0.625))) + ggtitle("N=5")


expmatrix <- matrix(rexp(15 * 500, 0.1), 500)
expmeans <- apply(expmatrix, 1, mean)

plot2 <- ggplot(data.frame(x = expmeans), aes(x = x)) + geom_histogram(aes(y = ..density..), colour = "darkslategray3",
    fill = "darkslategray3") + scale_x_continuous(breaks = c(1:16)) + theme_classic() + stat_function(fun = dnorm,
    color = "red", args = list(mean = 10, sd = sqrt(0.625))) + ggtitle("N=15")


expmatrix <- matrix(rexp(25 * 500, 0.1), 500)
expmeans <- apply(expmatrix, 1, mean)

plot3 <- ggplot(data.frame(x = expmeans), aes(x = x)) + geom_histogram(aes(y = ..density..), colour = "darkslategray3",
    fill = "darkslategray3") + scale_x_continuous(breaks = c(1:16)) + theme_classic() + stat_function(fun = dnorm,
    color = "red", args = list(mean = 10, sd = sqrt(0.625))) + ggtitle("N=25")



expmatrix <- matrix(rexp(120 * 500, 0.1), 500)
expmeans <- apply(expmatrix, 1, mean)

plot4 <- ggplot(data.frame(x = expmeans), aes(x = x)) + geom_histogram(aes(y = ..density..), colour = "darkslategray3",
    fill = "darkslategray3") + scale_x_continuous(breaks = c(1:16)) + theme_classic() + stat_function(fun = dnorm,
    color = "red", args = list(mean = 10, sd = sqrt(0.625))) + ggtitle("N=120")

library(gridExtra)
gridExtra::grid.arrange(plot1, plot2, plot3, plot4, ncol = 2)
```

A mayor tamaño de la muestra mayor precisión



Tamaño de la muestra

- Fórmula estándar para calcular tamaño muestral en proporciones:

$$n = \frac{z^2 * p(1-p)}{m^2} \tag{1}$$

  - Descripción: n = tamaño de la muestra requerido;
  - z = nivel de confiabilidad de 95 % (valor estándar de 1,96);
  - p(1-p) = Varianza o dispersión de la muestra.
  - p = probabilidad de ocurrencia 0.5 (50 % de los casos de la curva normal);
  - m = margen de error admitido de 5 % (valor estándar de 0,05)

- Lo central es que a menor tamaño muestral, mayor error tolerado. Y viceversa.

Ej: tamaño de la muestra

- Una fórmula estándar para calcular tamaño muestral:

$$n = \frac{1{,}96^2 * 0{,}5(1 - 0{,}5)}{0{,}05^2} \tag{2}$$

```
((1.96*1.96) * (0.5*(1 - 0.5)))/(0.05*0.05)
```

```
## [1] 384.16
```

- Si quiero analizar varias variables simultáneamente, debo agrandar mi tamaño muestral. Por ejemplo, si quiero analizar como se comporta la popularidad de la presidenta en tres estratos socioeconómicos, debería idealmente contar con una muestra de 384*3 para poder seguir realizando estimaciones a un 95 % de confianza y con un error de 5.

Relación entre error y tamaño

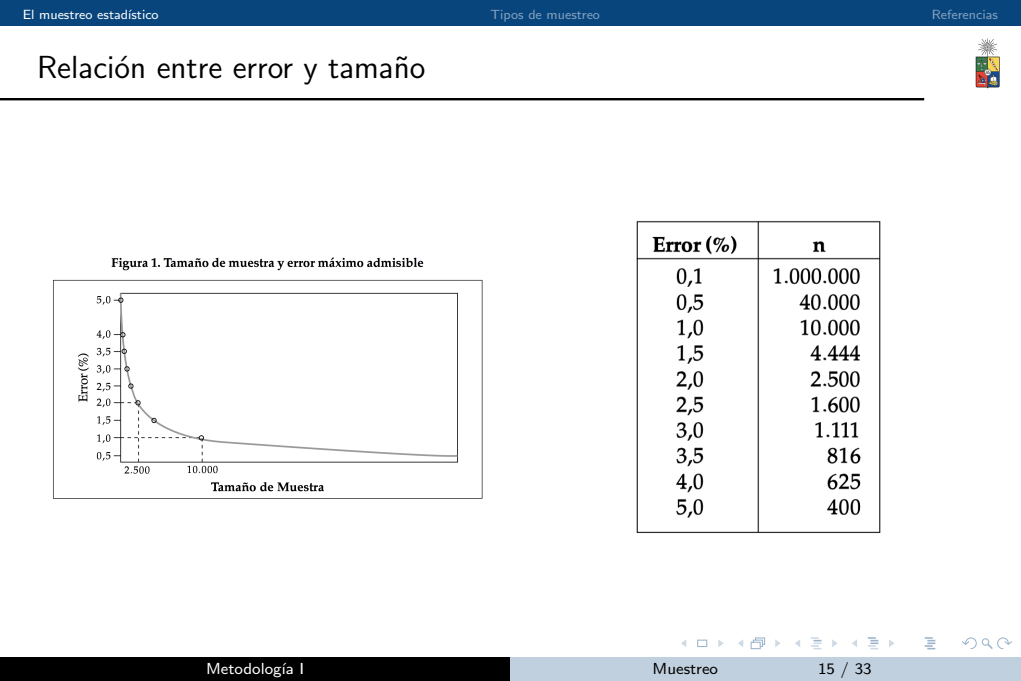


Figura 1. Tamaño de muestra y error máximo admisible

Error (%)	n
0,1	1.000.000
0,5	40.000
1,0	10.000
1,5	4.444
2,0	2.500
2,5	1.600
3,0	1.111
3,5	816
4,0	625
5,0	400

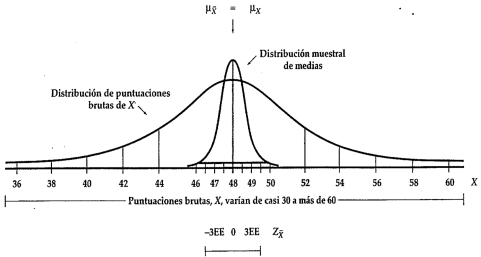
Error estándar

- Asumimos que nuestra estimación se basa en una muestra, y que podría variar si hubiera sido calculado en una muestra distinta
- Por lo tanto, existe una desviación muestral que se conoce como el **error estándar**.
- Con muestras mayores a un n de 120, la distribución muestral se aproximará a una distribución normal, con un promedio igual al promedio de la población y una desviación estándar de:

$$\sigma_{\overline{\chi}} = \frac{s}{\sqrt{n}} = SE(errorestándar)$$

Error estándar

- **Error estándar:** Desviación estándar de una distribución muestral, Mide la dispersión del error de muestreo que ocurre cuando se muestrea repetidamente una población.



- Gracias al teorema del límite central, mediante esta ecuación podemos obtener el error estándar del promedio muestral.

Probabilidad e intervalos

- En base a esto, por ejemplo si tenemos una muestra de 50 bebes, estatura= 64cm , y desviación estándar = 4:

$$\sigma_{\overline{X}} = \frac{s}{\sqrt{n}} = SE(errorestándar)$$

$$SE = \frac{4}{\sqrt{50}} = 0{,}56$$

- 64 +/- 0.56 nos da un intervalo de 63.43-64.56, donde se encuentra el 68 % de la población.
- Si sumamos +/- 1.96 SE (1.09) nos da un intervalo de 62,91 - 65.09 de estatura donde se encuentra el 95 % de los bebes.

¿Qué significa un 95 % de confianza?

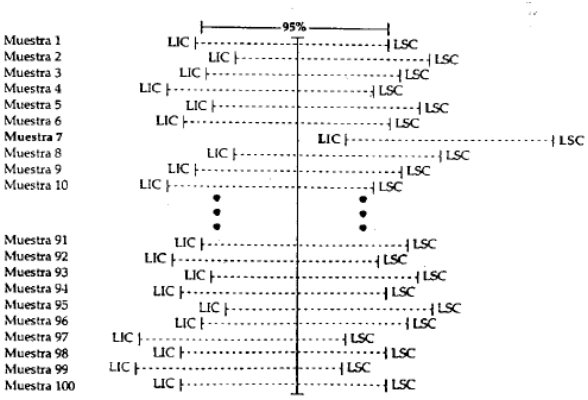
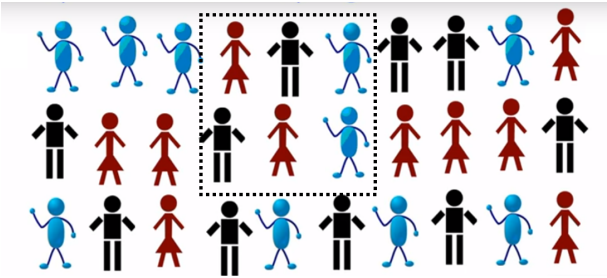


Figura: Ejemplo Intervalos

Tabla de Contenidos

Existen distintos tipos de muestreo

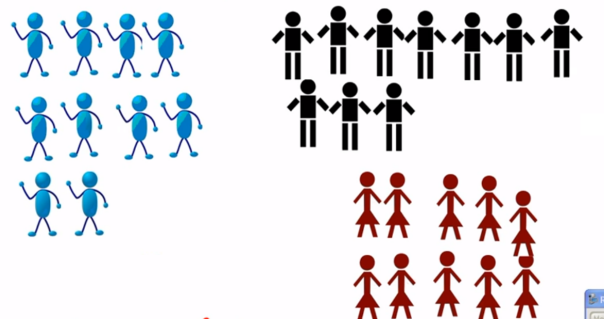
Muestreo aleatorio simple



Muestreo Aleatorio Simple

- Basado en la libre actuación del azar, sirve de referencia para los demás tipos de diseños.
- Es monetápica, autoponderada y equiprobable.
- Ventaja:
  - Sencillez de las fórmulas.
  - Precisión de la estimación
- Desventaja:
  - Necesitas el listado de todos los elementos.
  - Aumento de costo por la dispersión geográfica.

Muestreo estratificado



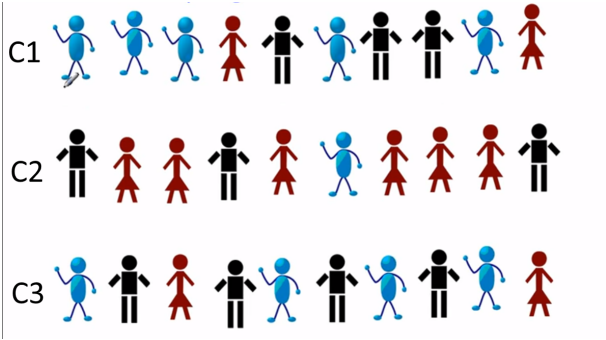
Muestreo Estratificado

- **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.
- Algunas cuestiones previas
  1. Agrupar en grupos homogéneos, aumenta la precisión.
  2. Puede ser con afijación igual o con afijación proporcional.
  3. Cada estrato es una agrupación independiente de las demás.
  4. Se establece como referencia un rango de 3 a 10 estratos.
  5. Dentro de cada estrato se aplica un MAS.

Muestreo por conglomerado

- El muestreo por conglomerado se caracteriza porque las unidades de muestreo no son los elementos de la población.
- Se consideran como unidades de muestreo conjuntos de elementos que constituyen un conglomerado.
- En cierto sentido es un muestreo jerarquizado en el cual hay conglomerados que incluyen a otros conglomerados.
- **Ejemplo de conglomerados:** Las comunas de una ciudad, distritos censales, departamentos de grandes tiendas, compañías de bomberos, áreas geográficas o superficies de un territorio.

Muestreo por conglomerado



Muestreo estratificado vs por conglomerados

Estratificado

- Los subgrupos se seleccionan de acuerdo con algún criterio relacionado con las variables estudiadas.
- Homogeneidad dentro de los subgrupos
- Heterogeneidad entre subgrupos

Conglomerados

- Los subgrupos se seleccionan de acuerdo con algún criterio de facilidad o disponibilidad en la recopilación de datos.
- Heterogeneidad dentro de subgrupos
- Homogeneidad entre subgrupos
- Selección aleatoria de subgrupos

Muestreo no probabilístico

- 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. En muestreo no probabilística es imposible conocer la magnitud del error.
- El más común es el muestreo por cuotas

Symbol	Age Group	No.
	11-21 Years	11
	22-31 Years	16
	32-41 Years	15
	42-51 Years	18
Total	11-51 Years	60

Muestreo por cuotas

- Se sustenta conceptualmente en que si se replican los porcentajes de la población, tendré una muestra probabilística.
- El trabajo del muestrista es establecer las cuotas. El del entrevistador llenarlas.
- Ventajas:
  1. No requiere del marco muestral (listado de elementos)
  2. Resuelve el problema de las no-respuestas.
- Desventajas:
  1. El muestrista no genera un dispositivo aleatorio perfecto.
  2. El entrevistador va a acceder, por lo general, a personas de fácil acceso.

Softwares para muestreo

- Excel: Comando “Data Analysis”
- R: se pueden incluso simular muestras falsas
- Stata y SPSS

Referencias

Vivanco, Manuel (2006). "Diseño de muestras en investigación social". En: *Manuel Canales Cerón (Coordinador-Editor), Metodologías de Investigación Social*, págs. 141-167.

Muestreo y selección de casos

¿Cómo seleccionar los casos del estudio?

Metodología I

Facultad de Ciencias Sociales
Universidad de Chile
Rodrigo Medel Sierralta
Nicolás Ratto

24 de mayo de 2021