Sociedad Ecuatoriana de Estadística

R Users Group - Ecuador®

"Análisis de encuestas por muestreo con R"

Introducción

Andrés Peña M.

agpena@colmex.mx

Mayo 2023

Tabla de contenidos

1. Introducción

Introducción

- "The universe cannot be read until we have learned the language and become familiar with the characters in which it is written. It is written in **mathematical** language" Galileo Galilei.

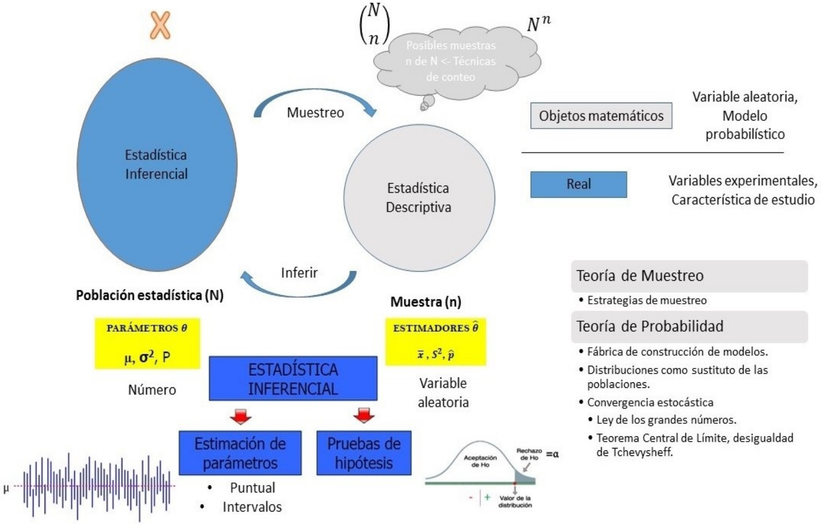
Introducción

- "The universe cannot be read until we have learned the language and become familiar with the characters in which it is written. It is written in **mathematical** language" Galileo Galilei.
- "**Statistics** is, or should be, about scientific investigation and how to do it better, but many statisticians believe it is a branch of mathematics. Now I agree that the physicist, the chemist, the engineer, and the statistician can never know too much mathematics, but their objectives should be better physics, better chemistry, better engineering, and in the case of statistics, **better scientific investigation**. Whether in any given study this implies more or less mathematics is incidental" George E. P. Box.

Introducción

- “The universe cannot be read until we have learned the language and become familiar with the characters in which it is written. It is written in **mathematical** language” Galileo Galilei.
- “**Statistics** is, or should be, about scientific investigation and how to do it better, but many statisticians believe it is a branch of mathematics. Now I agree that the physicist, the chemist, the engineer, and the statistician can never know too much mathematics, but their objectives should be better physics, better chemistry, better engineering, and in the case of statistics, **better scientific investigation**. Whether in any given study this implies more or less mathematics is incidental” George E. P. Box.
- El carácter de las Matemáticas es fundamentalmente **deductivo**. Por su parte, la Inferencia Estadística hace uso del conocimiento matemático pero tiene una naturaleza **inductiva**. El vínculo entre estas dos áreas lo provee la **Probabilidad**.

Visión general del método estadístico



Introducción

- En las encuestas por muestreo, el principal objetivo es estimar características de la poblacion usando los datos de una muestra.
- Mahalanobis (1965:45) resumió las ventajas de las encuestas por muestreo:
  - “... las encuestas por muestreo a grandes escalas, cuando se realizan de la manera apropiada con un diseno muestral satisfactorio, pueden proporcionar, **rápidamente** y a un **menor costo**, informacion con suficiente **precisión** para fines prácticos y con la posibilidad de **evaluar el margen de incertidumbre** con una base objetiva”.

Introducción

- ¿Qué es una muestra?

  Es una parte de una población de interés. Un subconjunto de esta.

- ¿Qué es la población de interés?

  Es un conjunto finito de objetos (elementos o unidades muestrales) identificables con ubicación en **tiempo y espacio**.

- Muestreo en la vida diaria

  Utilizamos muestreo, por ejemplo, al cocinar, al comprar, al comer.

- Objetivos del muestreo

  Las técnicas del muestreo se utilizan para conocer las características generales de la población de interés, al estudiar solo una parte de esta.

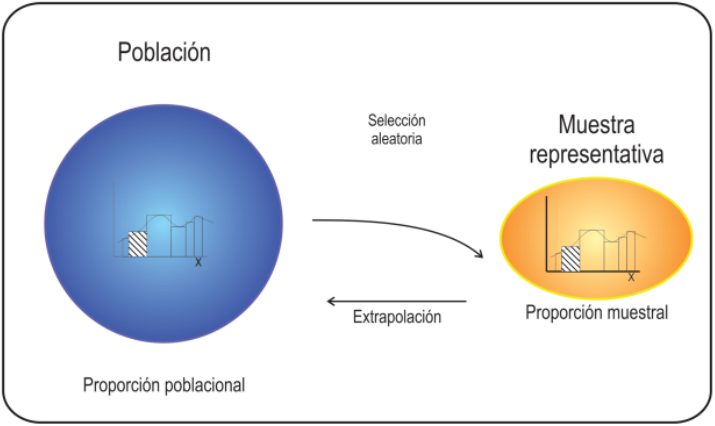
Introducción

- ¿Donde se usa?
  - Encuestas de opinión
  - Ratings de televisión
  - Industria. Control de calidad
  - Laboratorios. Estudios en sangre
  - Encuestas electorales
  - Encuestas del INEC. (Ingreso-Gasto, Empleo, Empresas, etc.)
  - Estudios de mercado
- ¿Por qué utilizar la información de solo una muestra?
  - Menos **Costo**
  - Mayor **Confiabilidad** en la informacion recabada
  - Si tenemos Pruebas destructivas, queremos destruir pocos elementos.
  - Mayor **Rapidez** en reunir la informacion

Introducción

- Objetivos del muestreo
  Seleccionar "buenas" muestras de un tamaño "apropiado", considerando la información que tenemos de la población que estamos estudiando y el presupuesto con que contamos
- ¿que es una "buena" muestra?
  Es una muestra representativa de la población, es decir, que las variables de interés en la muestra presenten una distribución semejante a las de la población.

Introducción



Introducción

- ¿Qué es un tamaño de muestra "apropiado"?
  Depende de:
  - La **variabilidad** que tiene en la población la característica que queremos estudiar.
  - La **precisión** con que queremos hacer la inferencia.
  - El **presupuesto** con que se cuenta.
  - El **tamaño** de la población.

2. Conceptos

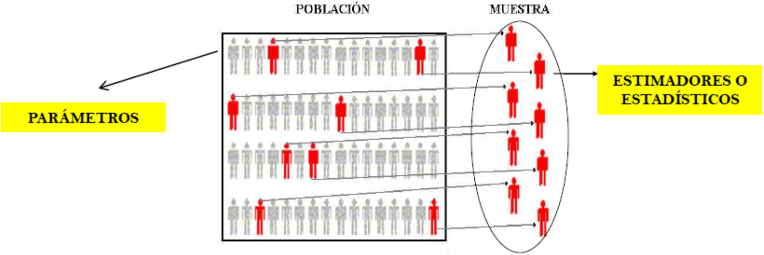
Conceptos

- **Población Objetivo.-** Conjunto de elementos identificables con ubicación en tiempo y espacio. La población se define al especificar que elementos son (a veces también cuáles no son) y que características deben tener.
- **Población muestreada.-** Es la población de donde se extrae la muestra.
- **Unidad de muestreo.-** Es la unidad donde realizamos la muestra, la que se selecciona.
- **Unidad de observación.-** Es el objeto (elemento) sobre el cual se realiza la medición.

  Muchas veces son iguales la unidad de muestreo y la unidad de observación.
- **Marco de muestreo.-** Es el medio físico que identifica a las unidades de muestreo de la población.

- **Muestra.**- Es un conjunto de unidades de la población seleccionadas del marco. Se puede seleccionar de forma probabilística o no probabilística.
- **Error de muestreo.**- Es el error de estimación que se controla con el diseño. Se debe a que tenemos una muestra solamente y no toda la poblacion.
- **Error de no muestreo.**- La no respuesta puede introducir sesgo a la estimación. Informacion falsa: preguntas mal redactadas, términos mal definidos, preguntas sensitivas. Sustitucion arbitraria de los elementos de la muestra. Se pueden controlar poniendo especial atencion a la construcción del cuestionario y a los detalles en el trabajo de campo a traves de una buena supervisión.
- **Estadístico(a).**- Es una función de la muestra que no tiene involucrados parametros desconocidos.
- **Estimador.**- Es un estadístico que se construye para estimar un parametro de la población (su valor varía de muestra a muestra).
- **Estimación.**- Es el valor que toma el estimador una vez observados los valores de la muestra.

3. Distribuciones en el muestreo



- Un estadístico es una función de los valores muestrales, una variable aleatoria porque cambia de muestra a muestra.
- Los estadísticos pueden ser calculados con fines meramente descriptivos o para estimar parámetros poblacionales, en este último caso reciben el nombre de estimadores.
- Se designarán por $\widehat{\theta}$ pero para cada caso especial su identificación cambiará. Al valor que toma un estimador en una muestra se le denomina estimación.

Distribuciones muestrales

- La distribución de todas las estimaciones de un parámetro basadas en todas las muestras posibles que pueden ser generadas por el plan muestral particular se denomina *distribución muestral del estimador*.
- Dos estimaciones pueden coincidir, muestras con elementos "distintos" que sin embargo toman valores iguales, lo cual significa que el numero máximo de estimaciones distintas será igual al número total de muestras posibles que se pueda extraer.

La media de la distribución de un estimador $\widehat{\theta}$, se define como:

$$E(\widehat{\theta}) = \sum_{i=1}^{\upsilon} \widehat{\theta}_i \pi_i$$

donde:

$\upsilon =$ Número total de valores distintos tomados por el estimador,

$\widehat{\theta}_i =$ i-ésima estimación diferente del parámetro,

$\pi_i =$ Probabilidad de que el estimador tome el valor $\widehat{\theta}_i$.

Distribuciones muestrales

La varianza de la distribución de un estimador $\widehat{\theta}$, está dada por:

$$VAR(\widehat{\theta}) = \sum_{i=1}^{\upsilon} (\widehat{\theta}_i - E[\widehat{\theta}_i])^2 \pi_i$$

La desviación estándar de la distribución de un estimador $\widehat{\theta}$, se denomina frecuentemente error estándar de la estimacióny se define:

$$EE(\widehat{\theta}) = \sqrt{VAR(\widehat{\theta})}$$

El coeficiente de variación para un estimador $\widehat{\theta}$ está dado por:

$$CV(\widehat{\theta}) = \frac{EE(\widehat{\theta})}{E(\widehat{\theta})}$$

El $CV(\widehat{\theta})$ de un estimador mide la variabilidad muestral de la estimación relativa al parámetro a ser estimado.

Propiedades de los estimadores

Insesgabilidad

Un estimador es *insesgado*[a] si el valor promedio de las estimaciones obtenidas para todas las muestras posibles es igual al verdadero parámetro poblacional, es decir $B(\widehat{\theta}) = 0$ ó tambien $E(\widehat{\theta}) = \theta$.

[a] El sesgo de un estimador $\widehat{\theta}$ se define como $B(\widehat{\theta}) = E(\widehat{\theta}) - \theta$

Eficiencia relativa

$EFR(\widehat{\theta}_1, \widehat{\theta}_2) = \frac{VAR(\widehat{\theta}_1)}{VAR(\widehat{\theta}_2)}$, según $EFR(\widehat{\theta}_1, \widehat{\theta}_2)$[a] sea inferior, igual o superior a la unidad se dirá que $\widehat{\theta}_1$ es más, igual o menos eficiente que $\widehat{\theta}_2$.

[a] El *efecto de diseño* aproxima un "**n**" bajo un diseño específico si se desea la misma precisión del MAS.

Consistencia

$\underset{n\to\infty}{Lim}\, Pr(|\widehat{\theta} - \theta| < \varepsilon) = 1$, la magnitud de los errores de estimación probable se pueden reducir aumentando el tamaño de la muestra hasta eliminarlos completamente cuando este iguala el tamaño de la población.

Notación de los parámetros y estadísticos

Los **estadísticos** estiman **parámetros poblacionales**, es decir, que aunque no coincidan exactamente con el parámetro -si la muestra fue seleccionada correctamente- deberían asumir valores próximos a los mismos.

Tipo de variable	Medidas	Parámetros	Estadísticos
Cuantitativa	Media	μ	$\bar{X}$
	Varianza	σ^2	S^2
Cualitativa	Proporción	p	$\hat{p}$

Los estadísticos son **variables aleatorias**, ya que valor depende de la muestra seleccionada y podemos determinar su distribución con base en **todas las muestras posibles** de igual tamaño.

Distribución de la media muestral

Muestras posibles:

- Con reemplazo $VR_n^N = N^n$
- Sin reemplazo $C_n^N = \binom{N}{n} = \frac{N!}{n!(N-n)!}$

Muestreo	Estadístico	Esperanza	Varianza
Con reemplazo	$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$	μ	$\frac{\sigma^2}{n}$
Sin reemplazo	$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$	μ	$\frac{\sigma^2}{n}\frac{N-n}{N-1}$

Variaciones con repetición

```r
#Permutaciones y combinaciones
install.packages("gtools", dependencies = TRUE)
```

Usado en **muestreo con reemplazo**:

$$VR_n^N = N^n$$

```r
(x<-1:4)

## [1] 1 2 3 4

4^2 #Duplas posibles con repetición de una población de 4

## [1] 16
```

Variaciones con repetición

```
library(gtools)
permutations(n=4,r=2,v=x,repeats.allowed=T)
```

```
##       [,1] [,2]
##  [1,]    1    1
##  [2,]    1    2
##  [3,]    1    3
##  [4,]    1    4
##  [5,]    2    1
##  [6,]    2    2
##  [7,]    2    3
##  [8,]    2    4
##  [9,]    3    1
## [10,]    3    2
## [11,]    3    3
## [12,]    3    4
## [13,]    4    1
## [14,]    4    2
## [15,]    4    3
```

Combinaciones sin repetición

Usado en **muestreo sin reemplazo**:

$$C_n^N = \binom{N}{n} = \frac{N!}{n!(N-n)!}$$

```
choose(4,2) #Duplas posibles sin repetición

## [1] 6

combinations(4, 2, v=x)

##      [,1] [,2]
## [1,]    1    2
## [2,]    1    3
## [3,]    1    4
## [4,]    2    3
## [5,]    2    4
## [6,]    3    4
```

Distribución de la media muestral

Esperanza y varianza de la media muestral

Si $X_1, X_2, ...X_n$ representan observaciones de una muestra aleatoria extraida de **cualquier población** con media μ y varianza σ^2 entonces $\bar{x}$ es una variable aleatoria con media μ y varianza σ^2/n.

$$E(\bar{X}) = \mathrm{E}\left(\frac{\sum_{i=1}^{n} x_i}{n}\right) = \frac{1}{n}\sum_{i=1}^{n} E\left(x_i\right) = \frac{n\mu}{n} = \mu$$

$$V(\bar{X}) = \mathrm{V}\left(\frac{\sum_{i=1}^{n} x_i}{n}\right) = \frac{1}{n^2}\sum_{i=1}^{n} V\left(x_i\right) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

Que consecuencias sobre la variabilidad de la distribución muestral de $\bar{x}$ tendría:

- Un aumento de n
- Un aumento de σ^2

Ejemplo de distribución de $\bar{x}$ CR

Población teórica

Familia	¿Cuántos trabajan?
A	2
B	4
C	3
D	1
Media	$\mu = 2.5$
Varianza	$\sigma^2 = 1.25$

```
x<-1:4
n<-length(x)
(mu<-mean(x))

## [1] 2.5

(va<-sum((x-mu)^2)/n)

## [1] 1.2
```

Ejemplo de distribución de $\bar{x}$ CR

Muestras de tamaño 2 con reemplazo

Familias seleccionadas en la Muestra	Cuántos trabajan		Media
A,A	2	2	2
A,B	2	4	3
A,C	2	3	2.5
A,D	2	1	1.5
B,A	4	2	3
B,B	4	4	4
B,C	4	3	3.5
B,D	4	1	2.5
C,A	3	2	2.5
C,B	3	4	3.5
C,C	3	3	3
C,D	3	1	2
D,A	1	2	1.5
D,B	1	4	2.5
D,C	1	3	2
D,D	1	1	1.0

```r
library(gtools)
muestras<-permutations(n=4,r=2,v=x,repeats.allowed=T)
(xbar_n_i<-rowMeans(muestras))

##  [1] 1.0 1.5 2.0 2.5 1.5 2.0 2.5 3.0 2.0 2.5 3.0 3.5 2.5 3.0

(fx_i<-prop.table(table(xbar_n_i)))

## xbar_n_i
##     1   1.5     2   2.5     3   3.5     4
## 0.062 0.125 0.188 0.250 0.188 0.125 0.062
```

Ejemplo de distribución de $\bar{x}$ CR

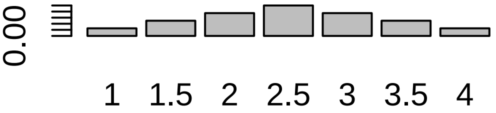
Distribución de la media muestral
(muestras de tamaño 2 con reemplazo)

Media muestral $\bar{x}_i$	Probabilidad $f(\bar{x}_i)$
1	1/16
1,5	2/16
2	3/16
2,5	4/16
3	3/16
3,5	2/16
4	1/16
Total	1

$$E(\bar{X}) = \sum \bar{x}_i f(\bar{x}_i) = \frac{40}{16} = 2.5 = \mu$$

$$V(\bar{X}) = \sum (\bar{x}_i - 2.5)^2 f(\bar{x}_i) = \frac{10}{16} = \frac{1.25}{2} = \frac{\sigma^2}{n}$$

```
barplot(prop.table(table(xbar_n_i)))
```




```
xbar_i<-unique(xbar_n_i)
(esp_xbar<-sum(xbar_i*fx_i))
(var_xbar<-sum((xbar_i-esp_xbar)^2*fx_i))
```

Ejemplo de distribución de $\bar{x}$ SR

Muestras de tamaño 2 sin reemplazo

Viviendas seleccionadas en la Muestra	Cuántos trabajan		Media
A,B	2	4	3
A,C	2	3	2,5
A,D	2	1	1,5
B,C	4	3	3,5
B,D	4	1	2,5
C,D	3	1	2

Distribución de la media muestral (muestras de tamaño 2 sin reemplazo)

Media muestral $\bar{x}_i$	Probabilidad $f(\bar{x}_i)$
1,5	1/6
2	1/6
2,5	2/6
3	1/6
3,5	1/6

$$E(\bar{X}) = \sum \bar{x}_i f(\bar{x}_i) = \frac{15}{6} = 2.5 = \mu$$

$$V(\bar{X}) = \sum (\bar{x}_i - 2,5)^2 f(\bar{x}_i) = \frac{2.5}{6} = 0,42$$

Compruebe en R

y la varianza de la media muestral resulta igual a:

$$V(\bar{X}) = \frac{\sigma^2}{n}\frac{N-n}{N-1} = \frac{1,25}{2}\frac{2}{3} = 0,42$$

Para recordar...

En la práctica es imposible trabajar con la distribución empírica del estadístico, obtenida a partir de todas las muestras posibles de igual tamaño, por lo que es importante establecer un modelo teórico de probabilidad para los estadísticos muestrales.

Muestreo en poblaciones normales

Vimos que si $X_1, X_2, \ldots .X_n$ representan observaciones de una muestra aleatoria, extraída de **cualquier población** con media μ y varianza σ^2, entonces $\bar{x}$ es una variable aleatoria con media μ y varianza σ^2/n. Si X es normal, la distribución de $\bar{x}$ también lo es, para **cualquier tamaño de muestra**.

Sea una variable con distribución normal $X \sim N\left(\mu, \sigma^2\right)$ y $X_1, X_2, \ldots .X_n$ una muestra aleatoria de esa población, entonces $\bar{x}$ tiene distribución normal con media μ y varianza σ^2/n.

$$\text{Si} \quad X \sim N\left(\mu, \sigma^2\right) \quad \Rightarrow \quad \bar{x} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

4. Teorema del Límite Central

Teorema del Límite Central - Poblaciones no normales

A través de este teorema (TCL) se demuestra que, **cualquiera sea la población,** si el tamaño de la muestra es lo suficientemente grande, **la suma de variables** $Y = \sum_{i=1}^{n} x_i$ se distribuye aproximadamente normal con esperanza n.μ y varianza n.σ^2

Regla empírica: si n $\geq$ 30, se puede usar el TCL
Si se trabaja con la **media muestral**, cuya distribución también converge a la normal tenemos:

$$\lim_{n\to\infty} P\left(\frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} < z\right) = F(z)$$

$$Y = \sum_{i=1}^{n} X_i$$

$$\text{Si } n \to \infty \quad Y \sim N(n\mu, \sqrt{n\sigma^2})$$

$$\text{Si } n \to \infty \quad \bar{x} \sim N\left(\mu, \sqrt{\frac{\sigma^2}{n}}\right)$$

La importancia de este teorema radica en su generalidad, ya que puede aplicarse a la media proveniente de cualquier distribución.

Teorema del Límite Central - Poblaciones no normales

Recuerde que una variable binomial X es el número de éxitos en un experimento binomial que consiste en ensayos de éxito o fracaso independientes con probabilidad de éxito p para un determinado ensayo.

$$x_i = \begin{cases} 1 & si\ el\ i - ésimo\ ensayo\ produce\ un\ éxito \\ 0 & si\ el\ i - ésimo\ ensayo\ produce\ un\ fracaso \end{cases}$$

A partir de la variable x podemos definir la proporción como:

$$p = \frac{X}{n} \text{ donde } X = x_1 + x_2 + \cdots + x_n$$

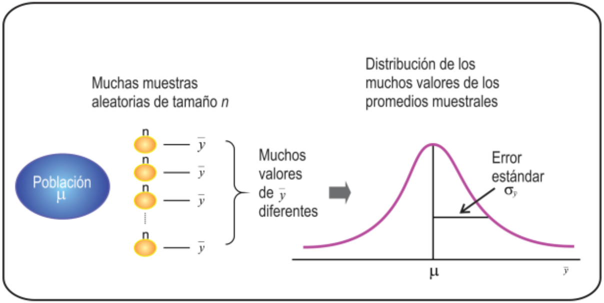
Dado que la variable binomial se define como la suma de variables bipuntuales, de acuerdo al TCL:

$$\lim_{n \to \infty} P\left(\frac{\hat{p} - p}{\sqrt{\frac{p.q}{n}}} < z\right) = F(z)$$

Los resultados empíricos muestran que se obtienen buenas aproximaciones de probabilidad utilizando el modelo normal, siempre que $np \geq 5$ y $nq \geq 5$

Teorema del Límite Central - Poblaciones no normales

Para que se alcance una distribucion parecida a la normal en el conjunto de posibles valores del promedio muestral, se requiere que *n* sea grande.

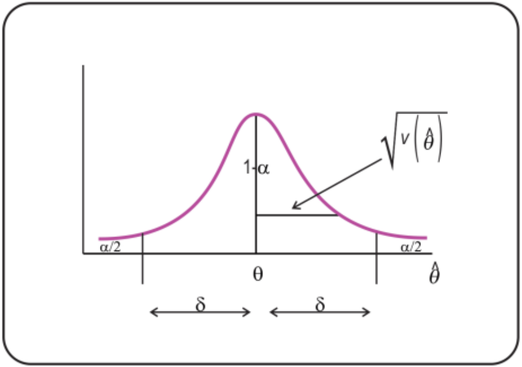


Sin embargo, la rapidez de acercamiento a la normal (velocidad de convergencia) tambien depende de la forma de la distribucion de la variable de interés en la población.

Teorema del Límite Central - Poblaciones no normales

En general, en la poblacion se tendrá un parámetro θ, que al tomar muchas muestras posibles con un diseno de muestra específico y una forma de estimador dada, produce muchos valores de $\hat{\theta}$.

Por el Teorema Central del Límite:



Teorema del Límite Central - Poblaciones no normales

$$E(\hat{\theta}) = \theta$$

$$V(\hat{\theta}) = E[\hat{\theta} - E(\hat{\theta})]^2 = E[\hat{\theta} - \theta]^2$$

$$P[\theta - \delta \leq \hat{\theta} \leq \theta + \delta] = 1 - \alpha$$

equivalente a:

$$P[|\hat{\theta} - \theta| \leq \delta] = 1 - \alpha$$

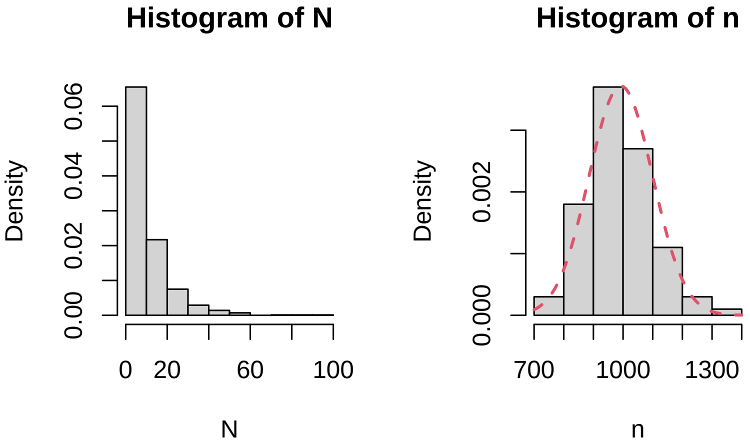
En palabras, la probabilidad de una discrepancia de a lo más δ entre θ y $\hat{\theta}$ es $1 - \alpha$.

A δ se le conoce como **precisión** del muestreo o **error de estimación**, y a $1 - \alpha$ como **confianza**.

Teorema del Límite Central en *R*

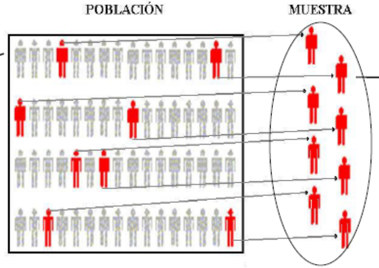
```
#Teorema del límite central
#N <- rbinom(1000, 100, 0.5)
N <- rexp(1000, 1/10)
#N <- runif(1000, 10, 50)

n <- numeric(100)
for (i in 1:100) {
  n[i] <- sum(sample(N, 100, replace = TRUE))
}

par(mfrow=c(1,2))
hist(N, probability = T)
hist(n, probability = T)
curve(dnorm(x, mean(n), sd(n)), col = 2, lty = 2,
      lwd = 2, add=T)
par(mfrow=c(1,1))
```

Teorema del Límite Central en *R*



5. Estimación de parámetros

Inferencia estadística



PARÁMETROS θ

μ, σ^2, P

ESTIMADORES $\hat{\theta}$

$\bar{x}$, s^2, $\hat{p}$

ESTADÍSTICA INFERENCIAL

Estimación de parámetros

Pruebas de hipótesis

Estimación Puntual

El valor de la media muestral de cualquier muestra se puede considerar como una **estimación puntual** ("puntual", porque se trata de un solo número, que corresponde a un solo punto de la recta numérica) de la media poblacional μ.

Población
Estudiantes de la Facultad de Cs.

Muestra
100 estudiantes seleccionados al azar

Estimador puntual
La media del gasto en libros: $25

Una estimación puntual no proporciona por sí misma información acerca de la precisión y confiabilidad de la estimación. Por la variabilidad, es poco probable que $\bar{x} = \mu$.
La estimación puntual no indica cuán cerca podría estar la media muestral con respecto a la media poblacional.

Estimación Puntual

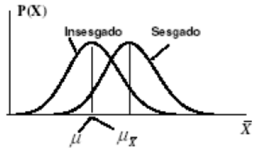
Propiedades de los buenos estimadores:

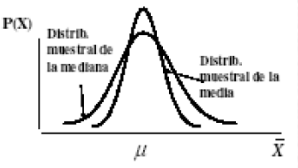
Insesgabilidad $E(\hat{\theta}) = \theta$

Consistencia $\lim_{n \to \infty} \left(\Pr\left|\hat{\theta} - \theta\right| < \varepsilon \right) = 1$

Eficiencia $V(\hat{\theta}_1) < V(\hat{\theta}_2)$

Suficiencia cuando utiliza toda la información que surge de la muestra





Estimación por intervalos

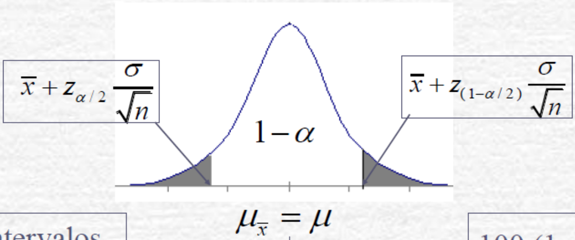
Una alternativa respecto a informar un solo valor sensible del parámetro es calcular un intervalo de valores (intervalo de confianza IC)

Estimación por intervalos

Un intervalo de confianza siempre se calcula al seleccionar primero un **nivel de confianza** que es una medida del grado de confiabilidad del intervalo.

Estimación por intervalos

Distribución muestral de la media



Los intervalos van de $\bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ a $\bar{x} + z_{(1-\alpha/2)}\frac{\sigma}{\sqrt{n}}$

100 (1-α) % de los intervalos construidos contienen a μ y 100(α) % no.

6. Tipos de muestreo

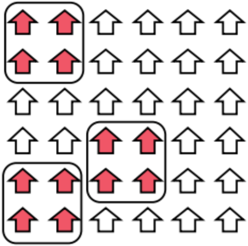
Tipos de muestreo

En el muestreo probalilístico cualquier elemento de la población objetivo tiene una cierta probabilidad de selección, las conclusiones son válidas para todo el universo de estudio y no solo para aquellos casos que fueron seleccionados.

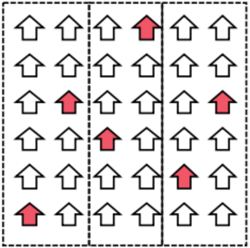
- Tipo de muestreo
  - Probabilístico
    - Aleatorio simple
    - Sistemático
    - Por conglomerados
    - Estratificado
  - Determinístico
    - Convencional o accidental
    - Por cuotas
    - Por juicio

Tipos de muestreo

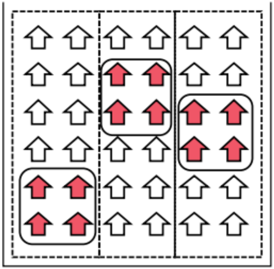
Muestreo por conglomerados



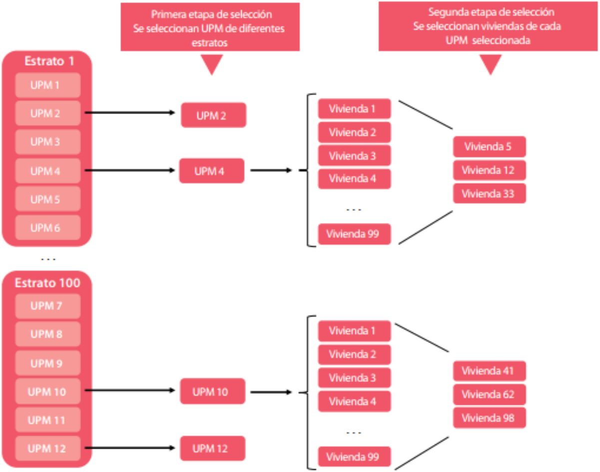
Muestreo estratificado



Muestreo estratificado y por conglomerados



Muestreo bietápico en las encuestas de hogares



El hecho de que en cada etapa se elijan ciertos casos y se descarten otros afecta a las estimaciones ya que se generan probabilidades de selección desigual. Para corregir ésta y otras situaciones se requiere del uso de ponderadores.

7. Ponderación de la muestra

Ponderación de la muestra

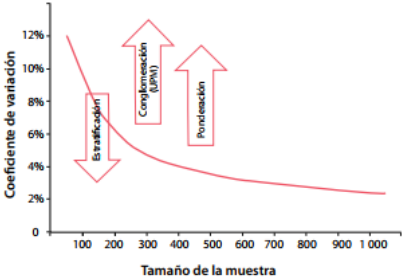
- El único caso donde las viviendas tienen una probabilidad de selección igual es en el muestreo aleatorio simple, donde no existe estratificación ni conglomeración, pero en encuestas de gran escala es difícil que se aplique (Heeringa et al., 2010, Naciones Unidas, 2009).
- Los **factores de expansión** se definen como el inverso de la probabilidad de selección (Naciones Unidas, 2009:61).
- En una encuesta probabilística cada persona u hogar seleccionado y entrevistado representa a un número determinado de personas u hogares. Esa representación se refleja en el factor de ponderación o de expansión (INEC, 2021).

Ponderación de la muestra

Factor de expansión = Factor teórico x Ajuste por cobertura x Calibración

Al igual que sucede con la conglomeración y la estratificación, cuando se aplica una ponderación a la muestra se afecta la precisión (Martínez, 2017).

Efecto de la conglomeración, estratificación y ponderación en el efecto de diseño

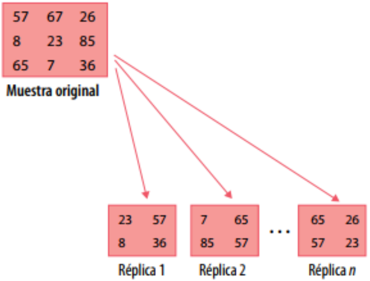


8. Cálculo de la varianza

Cálculo de la varianza

Métodos para la estimación de la varianza

- Cálculo de la varianza
  - Replicación de la varianza
    - Replicación repetida y equilibrada (*BRR*)
    - *Jackknife*
      - JK1
      - JK2
      - JKn
    - *Bootstrap*
  - Linealización por series de Taylor



Réplicas para una muestra con tres estratos y dos UPM

Estrato	UPM	Réplicas							
		1	2	3	4	5	6	7	8
Estrato 1	UPM_1	√			√			√	√
	UPM_2		√	√		√	√		
Estrato 2	UPM_1	√				√	√		√
	UPM_2		√	√	√			√	
Estrato 3	UPM_1	√		√			√	√	
	UPM_2		√		√	√			√
√	Sí está en la réplica.								

Coeficiente de variación

- Existen algunas medidas de dispersión que son útiles para evaluar la calidad de un dato que se genera a partir de una encuesta compleja. Dentro de éstas se encuentran los errores estándar (SE), el intervalo de confianza (IC) y el coeficiente de variación (CV).
- El CV refleja la magnitud relativa que el error estándar con respecto al estimador de referencia, entre más pequeño sea este valor mejor es la precisión.

$$CV(\widehat{\theta}) = \frac{EE(\widehat{\theta})}{E(\widehat{\theta})}; \quad EE(\widehat{\theta}) = \sqrt{VAR(\widehat{\theta})}$$

- Si bien no existe un consenso unánime sobre qué valores son los más adecuados, algunos INE's consideran que un dato es de buena calidad si el coeficiente de variación está por debajo de 15%.

Gracias!!!