Sociedad Ecuatoriana de Estadística

"Análisis de Encuestas por Muestreo con R"

Unidad 4: Distribuciones en el muestreo. Teorema del Límite Central

Andrés Peña M.

a.pena@rusersgroup.com

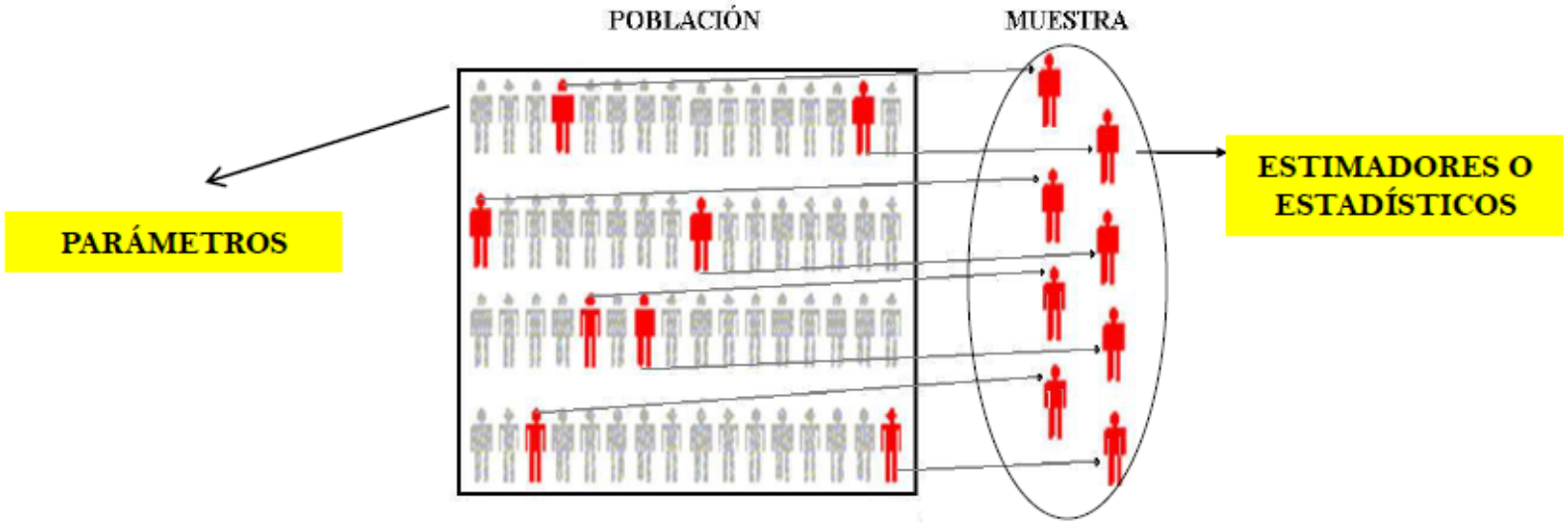
Mayo 2021

X Seminario Internacional de Estadística Aplicada

Tabla de contenidos

1. Distribuciones en el muestreo

Introducción



- 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

$\lim_{n \to \infty} Pr(|\widehat{\theta} - \theta| < \varepsilon) = 1$, 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.

Distribución conjunta de *n* observaciones muestrales

Si de la población se selecciona **una observación al azar**, se genera una variable aleatoria que tiene las misma distribución de la variable en la población. Si seleccionamos ***n*** **unidades con reemplazo,** se genera una muestra aleatoria simple.

Una **muestra aleatoria simple** es una sucesión X_1, X_2,.... X_n de *n* variables aleatorias independientes e igualmente distribuidas, es decir, que todas tienen la misma función de densidad o cuantía, que es la de la variable en la población.

$$f(x_1, x_2, \ldots\ldots, x_n) = \prod_{i=1}^{n} f(x_i) = [f(x_i)]^n$$

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 correctamente seleccionada, 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 su valor depende de la muestra seleccionada y podemos determinar su distribución en base a **todas las muestras posibles** de igual tamaño.

Distribución de la media muestral

Muestras posibles:

- Con reemplazo $VR_n^N = N^n$

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

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	$\frac{\sigma^2}{n}\frac{N-n}{N-1}$

Esperanza y varianza de la media muestral

Si X_1, X_2,.... 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.

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

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

Analice qué consecuencias tienen sobre la variabilidad de la distribución muestral de la media:

a) Un aumento de n

b) 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

```
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

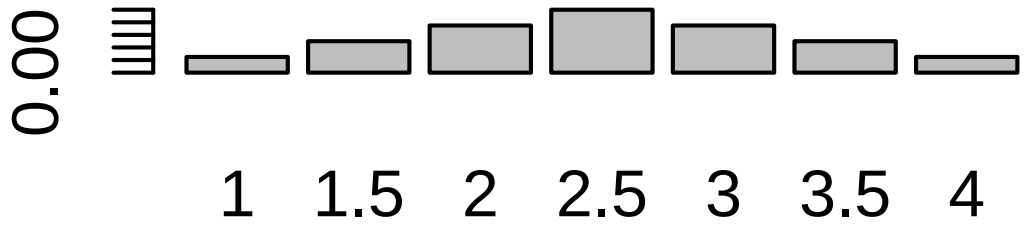
Distribución de la media muestral
(muestras de tamaño 2 con reemplazo)

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

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

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

y la varianza de la media muestral resulta igual a:

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

Compruebe en R

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

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

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

2. 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 $n.\sigma^2$

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

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

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

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

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

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 $nq \geq 5$

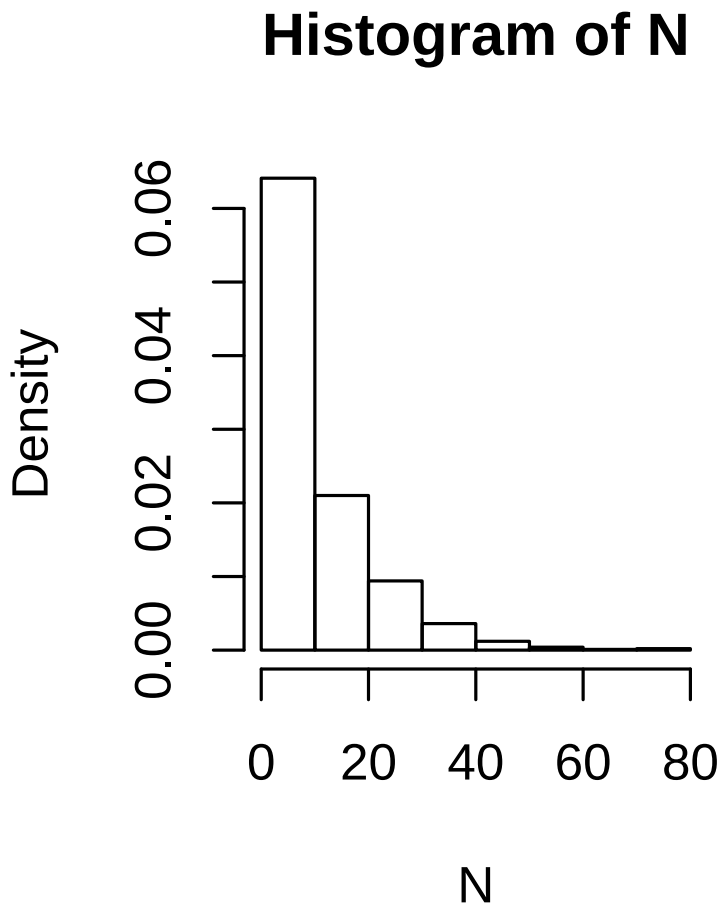
Teorema del Límite Central en *R*

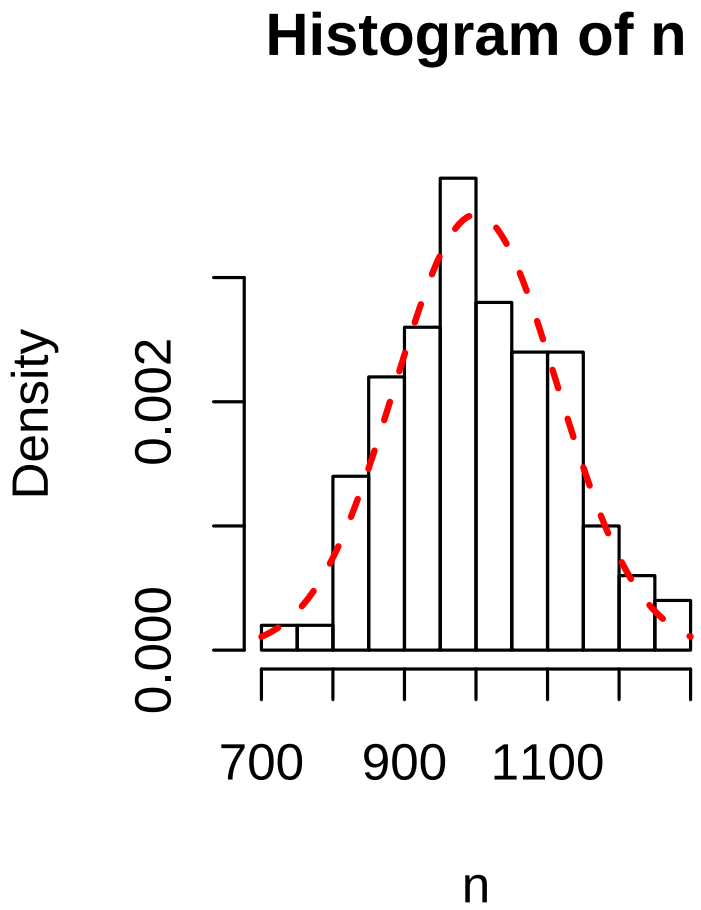
```
#Teorema del límite central
#N <- rbinom(1000, 100, 0.5)
N <- rexp(1000, 1/10)
#N <- runif(1000, 10, 50)

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*





Gracias!!!