Tarea N°2

Daniel Felipe Villa - 1005087556

2022-03-30

Caso de estudio n°3:

Habiendo hecho esto, procederemos con el intervalo de confianza (IC)

Contexto: Suponga que se desea realizar un estudio de muestreo en un municipio A del departamento de Antioquia para estimar la **proporción de votantes** registrados con intención de voto por un candidato X. Para ello:

- Se dispone del listado de todos los habitantes mayores de edad que conforman el municipio, $N = 10000$ Habitantes.
- Se opta por seleccionar una MAS sin remplazo de dicha población, $n = 500$.
- Se encuentra que en dicha muestra sólo 350 personas estaban inscritas para votar y de éstas 150 estaban a favor del candidato X.

¿Cómo estimar la proporción de votantes registrados con intención de voto por el candidato X?

Solución

extraemos los datos del enunciado anterior

N: 10.000 habitantes *(de la población total)*

n: 500 hab. tamaño de la muestra

N_k: Desconocido

a_k: 150 *(voto a favor del candidato X)*

n_k: 350

```
ak <- 150
nk <- 350
N <- 10000
n <- 500
```

$$\hat{p}_k = \frac{a_k}{n_k} = \frac{150}{350} = 0.4285714 \approx 0.43$$

```r
# "p" gorro:
pg <- round(ak/nk,2)
```

I.C.

Habiendo hecho esto, procederemos con el intervalo de confianza (IC)

Como N_k es desconocido utilizare:

$\widehat{Var}[\hat{p}_k] = \frac{N-n}{N} * \frac{\hat{p}_k(1-\hat{p}_k)}{n_k-1}$ *(1)*

```r
# Varianza estimada:
var_p <- (N-n)/N * pg*(1-pg)/(nk-1)
print("# Varianza estimada:")
```

```
## [1] "# Varianza estimada:"
```

```r
print(var_p)
```

```
## [1] 0.0006671777
```

un intervalo al 95% esta dado por:

$$\hat{p}_k \pm t_{1-\alpha/2\,,\,n_k-1}\sqrt{\hat{V}(\hat{p}_k)}$$

```
## [1] "# IC para P_k"
```

```
## [1] "#       Limite Inf ,   Limite Sup"
```

```
## [1] "( 0.379198383195855 , 0.480801616804145 )"
```

Con esto vemos un IC al 95% de la proporción de votantes que tienen intención de voto por el candidato X.

Caso de estudio n°4:

Contexto: Una multinacional de abrir nuevos puestos de trabajo en un barrio de Medellín.

para ello necesitas estimar de las personas que **NO trabajan**, el tiempo en (meses) que el jefe de hogar ha completado sin trabajar.

La empresa cuenta con lista de todos los hogares del barrio bajo estudio, conformado por 1000 hogares. se decide:

- seleccionar una muestra piloto 10 hogares y se entrevista al jefe de hogar

los datos que se obtuvieron fueron los siguientes:

Hogar	1	2	3	4	5	6	7	8	9	10
Trabaja (1: SI | 0: NO)	1	0	1	0	0	1	0	1	0	0
Tiempo (en meses)		3		12	5		7		8	2

¿Como estimar la **proporción** de hogares donde el jefe de hogar NO trabaja?

¿Como estimar el **tiempo promedio** en meses, de los jefes de hogar que NO han tenido trabajo?

Solución:

Primera pregunta:

Extraemos los datos del enunciado:

```
#Valores necesitados:
nk <- 6
N <- 1000
n <- 10

# proporción estimada
pg <- nk/n
pg
```

```
## [1] 0.6
```

```
#Tiempo en meses sin trabajar
tm <- c(3,12,5,7,8,2)
```

Para hallar IC, primero hallaremos el $\hat{V}(\hat{p})$

Nota: es la misma fórmula anterior (1)

```
# Varianza estimada:
var_p <- (N-n)/N * pg*(1-pg)/(n-1)
print("# Varianza estimada:")
```

```
## [1] "# Varianza estimada:"
```

```
print(var_p)
```

```
## [1] 0.0264
```

IC

Por notación calcularemos el β (error de estimación) *(esto con el fin de que no sea tan repetitivo todo)*

$ = t_{/2 , n-1}$

```
# error de estimación:
b <- qt(0.025,n, lower.tail = F)*sqrt(var_p)
b

## [1] 0.3620297

#Calculando el IC
print("#       Limite Inf ,   Limite Sup")

## [1] "#       Limite Inf ,   Limite Sup"

paste("(", pg-b, ",", pg+b, ")")

## [1] "( 0.237970287912041 , 0.96202971208796 )"
```

Según el IC decimos con un nivel de confianza del 95% en el barrio muestreado de Medellín los jefes de hogar que no trabajan están dentro del 24% y el 96%; esto es preocupante ya que deja mucho que desear acerca de esta población, haciéndonos preguntar… ¿Cómo generan ingresos para el hogar?

Segunda pregunta

Estimar la media:

$$\hat{\mu} = \bar{y}_k = \frac{1}{n}\sum y_{ki}$$

```
#Media estimada
yk <- mean(tm)
yk

## [1] 6.166667
```

Ahora estimamos la varianza de los datos de interés:

$$S_k^2 = \frac{1}{n-1}\sum (y_{ki} - \bar{y})^2$$

```
# Varianza estimada
sk <- var(tm)
sk

## [1] 13.36667
```

Por último, calculo de la varianza de la media estimada:

$$\widehat{Var}(\bar{y}) = \frac{N-n}{N}\frac{S_k^2}{n_k}$$

Se utilizo una función creada para futuros trabajos tenerla presente y solamente llamarla cuando sea necesario:

```r
#Funcion de varianza estimada de la media:
varmu <- function(N,n,sk,nk){
  "Ya que no se conoce Nk..."
  (N-n)/N * (sk)/nk
}
```

```
## [1] "# Calculando la varianza estimada de la media con N_k
desconocida: "
```

```
## [1] 2.2055
```

IC

Un IC al 95% para $\bar{y}$ (promedio de meses sin trabajar) están dados por:

$$\bar{y} \pm t_{1-\alpha/2\ nk-1}\sqrt{\hat{V}(\bar{y})}$$

```r
LI <- yk - qt(0.025,nk-1,lower.tail = F)*sqrt(varyk)


LU <- yk + qt(0.025,nk-1,lower.tail = F)*sqrt(varyk)

print("# IC para y_k")
```

```
## [1] "# IC para y_k"
```

```r
print("#       Limite Inf ,   Limite Sup")
```

```
## [1] "#       Limite Inf ,   Limite Sup"
```

```r
paste("(", LI, ",", LU, ")")
```

```
## [1] "( 2.34911463133005 , 9.98421870200329 )"
```

Podemos afirmar con un $\alpha = 0.05$ que los tiempos promedios en los cuales no han tenido ningún tipo de trabajo los responsables del hogar entre 3 a 10 meses; unido con lo anterior me deja pensando como hacen para subsistir estas familias... ¿Dependerán completamente del ingreso solidario del gobierno?

Ejercicio n°20

Observemos el ejercicio:

20. Una muestra aleatoria simple sin reemplazo de 56 personas fue seleccionada de una población de 1000 trabajadores de una fábrica. Se midieron las variables Ingreso mensual (I) y Género (G) (H-Hombre, M-Mujer). La información obtenida es la siguiente:

(a) Estimar el ingreso promedio de los trabajadores. Calcule un I.C del 95% para dicho ingreso promedio.

(b) Estimar el ingreso total de todos los trabajadores de dicha empresa. Calcule un I.C del 95% para dicho ingreso total.

(c) Estimar la proporción y el número total de mujeres en la empresa.

(d) Calcule un I.C del 95% tanto para la proporción como para el número total de mujeres en la empresa.

(e) ¿Qué puedes decir acerca de la validéz de la aproximación normal en este caso? Justifique su respuesta.

(f) ¿Cómo estimaría el ingreso promedio y el total de las mujeres para toda la empresa si no se conociera el número total de ellas?

(g) ¿Cuál de las dos subpoblaciones (hombres y mujeres) considera que es más homogénea con respecto a los ingresos? Justifique su respuesta.

(h) Estimar el ingreso promedio de cada una de las subpoblaciones (hombres y mujeres).

Lo primero que haremos, será que ingresaremos la base de datos escrita en Excel.

```
library(readxl)
df20 <- read_excel("data_sets20.xlsx")
```

Nro.	1	2	3	4	5	6	7	8	9	10	11	12	13	14
I	800	960	670	688	1025	2346	780	1126	1780	2135	1446	2245	1760	2287
G	M	H	H	M	M	H	M	H	H	H	H	H	M	H
Nro.	15	16	17	18	19	20	21	22	23	24	25	26	27	28
I	686	997	1335	1567	1456	1234	2678	1456	1388	1785	1653	2121	880	984
G	H	M	H	M	M	H	H	H	M	H	M	H	M	H
Nro.	29	30	31	32	33	34	35	36	37	38	39	40	41	42
I	1256	946	2000	2037	3111	1042	1564	1222	1768	1984	2348	876	890	1452
G	H	M	M	M	H	M	H	M	H	H	H	M	H	H
Nro.	43	44	45	46	47	48	49	50	51	52	53	54	55	56
I	1678	1326	1843	880	760	1146	1680	2880	1890	1033	2668	3345	2156	1880
G	H	M	H	H	M	M	M	H	M	H	M	H	M	M

Literal (a):

Para estimar el promedio de los ingresos de los trabajadores utilizaremos:

un estimador puntual para μ

$$\hat{\mu} = \bar{y} = \frac{1}{n}\sum y_i$$

```
# La función mean() hace lo descrito anteriormente:
## Podemos suponer que el ingreso esta escalonado en dolares

mug <- mean(df20$ingreso)
mug

## [1] 1570.161
```

Podemos ver que el promedio del ingreso de los trabajadores muestreados es $1570.161 dólares, esto en la actualidad es un salario alto, pero no podemos decir quienes fueron los encuestados, es decir, si le preguntaron a los trabajadores y gerentes, socios (personas que naturalmente ostentan un salario mucho más alto que un trabajador regular de fabrica)

Para calcular el IC al 95% utilizaremos:

- La varianza muestral
- el límite en el error de la estimación
- un estadístico $Z_{\alpha/2}$, ya que $n > 30$

```
# varianza muestral:
s2 <- var(df20$ingreso)
s2

## [1] 422010.7

#tamaño de la muestra:
n <- length(df20$ingreso)
n

## [1] 56

#Tamaño de la población
N <- 1000

# Error
## Utilizamos una normal ya que el n > 30
B <- qnorm(0.025,lower.tail = F)*sqrt((s2*(N-n))/(N*n))
B

## [1] 165.311
```

Un IC para el ingreso promedio es: $\bar{y} \pm B$

```
LI <- mug-B
LU <- mug+B
print("#       Limite Inf ,   Limite Sup")

## [1] "#       Limite Inf ,   Limite Sup"

paste("(", LI, ",", LU, ")")

## [1] "( 1404.84969039249 , 1735.47173817894 )"
```

Podemos ver que nuestro ingreso promedio se encuentra dentro de los $1401 y los $1740 dólares es decir, estos trabajadores tienen muy buenos sueldos si su moneda de pago es el dólar.

Literal (b):

Para hallar el total poblacional estimado, solamente tenemos que multiplicar nuestro estimador insesgado para μ con nuestro N (población donde se extrajo la muestra)

$$\hat{\tau} = \hat{\mu} * N$$

```
# total poblacional tau:
taug <- mug*N
taug

## [1] 1570161
```

Para hallar el IC para el total del ingreso de los trabajadores de la fábrica basta con multiplicar los extremos del intervalo de confianza anterior por $N = 1000$

```
print("#       Limite Inf ,   Limite Sup")

## [1] "#       Limite Inf ,   Limite Sup"

paste("(", LI*N, ",", LU*N, ")")

## [1] "( 1404849.69039249 , 1735471.73817894 )"
```

Concluimos que el total de ingreso de los trabajadores esta entre los 1401131 y los 1739189 dólares

Literal (c):

1. Para estimar la proporción de las mujeres en la empresa primero tenemos que hacer un conteo de cuantas mujeres hay en la muestra

```
# Convertiremos en factor la columna genero
df20$genero %<>% as.factor()

# Conteo de las mujeres de la muestra:
## El dato que nos interesa es el primero ya que el otro representa número de columnas
df20 %>% filter(genero == "M") %>% dim(.)

## [1] 25  3

fem <- 25
```

Ahora procederemos a calcular $\hat{p} = \frac{1}{n}\sum y_i$

```
# Proporción estimada:
pg <- fem/n
pg

## [1] 0.4464286
```

la proporción estimada de mujeres en la fábrica representa aproximadamente $44.64\% \approx 45\%$

2. para el caso del total de mujeres en la empresa, solo basta con tomar el resultado anterior y multiplicarlo por el $N = 1000$

```
# Total poblacional:
N*pg
```

```
## [1] 446.4286
```

el total estimado de mujeres en la población de la fábrica es $446.42 \approx 447$

Literal (d):

Un IC para la proporción sería:

```
# Hallar el límite en el error de estimación
B <- qnorm(0.025,lower.tail = F)*sqrt(((pg*(1-pg))*(N-n))/((n-1)*N))
B
```

```
## [1] 0.1276485
```

```
LI <- pg-B
LU <- pg+B

print("# IC para p")
```

```
## [1] "# IC para p"
```

```
print("#       Limite Inf ,   Limite Sup")
```

```
## [1] "#       Limite Inf ,   Limite Sup"
```

```
paste("(", LI, ",", LU, ")")
```

```
## [1] "( 0.318780114957863 , 0.57407702789928 )"
```

podemos notar que la proporción de mujeres abarca desde los 32% hasta más de la mitad con un 58% esto nos da un indicio que hay una cantidad significativa de mujeres.

Para el caso del total de las mujeres, hacemos como en los casos anteriores, simplemente multiplicamos nuestro IC por nuestro N

```
print("# IC para tau")
```

```
## [1] "# IC para tau"
```

```
print("#       Limite Inf ,   Limite Sup")
```

```
## [1] "#       Limite Inf ,   Limite Sup"
```

```
paste("(", LI*N, ",", LU*N, ")")
```

```
## [1] "( 318.780114957863 , 574.07702789928 )"
```

Este caso no se diferencia en el anterior al ver una cantidad significativa en este intervalo, superando la mitad del valor de N

Literal (e):

La aproximación es buena, es decir, gracias a que tenemos un $n > 30$ podremos decir que nos aproximamos a una tendencia normal, esto se debe al **TCL** que nos garantiza normalidad cuando las muestras respectivas son lo suficientemente grandes (significativas) como para por medio de modificaciones aproxime a una distribución normal.

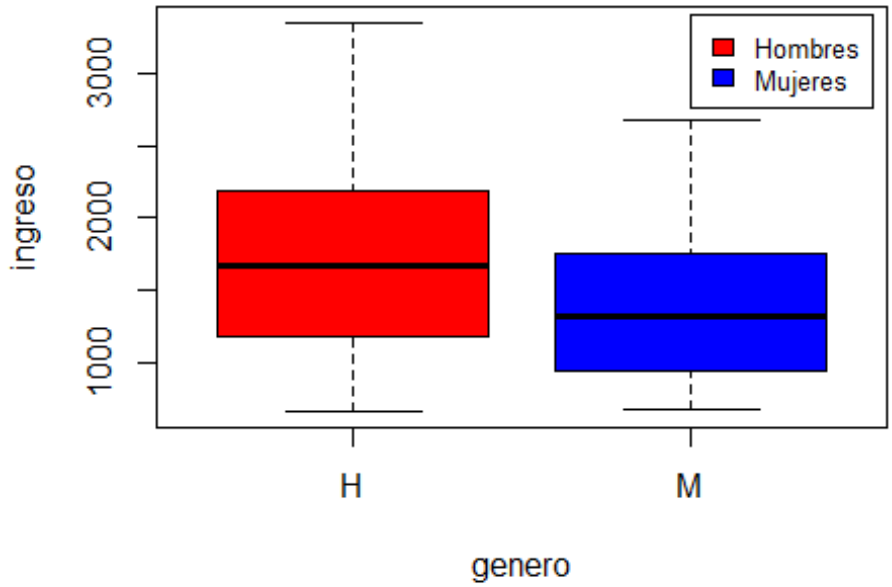
Literal (f):

Para el ingreso promedio de las mujeres tendríamos que considerar una subpoblación, que en este caso sería de las mujeres. Se suman todos los valores de los ingresos de esa población y se divide entre la cantidad de elementos de la subpoblación. Para el total dado que no conocemos el valor de la subpoblación para determinar el factor de corrección se trabaja con la población y la muestra (N/n) multiplicado por el promedio así tendríamos un estimador para el total

Literal (g) y (h):

Para responder a esta pregunta tendremos que observar cuál de los datos contiene mayor varianza respecto a la media estimada $\hat{V}(\hat{\mu})$, pero primero observemos un gráfico de boxplot para analizar a priori este enunciado:

```
boxplot(ingreso ~ genero, data = df20, col =c("red", "blue"))
legend("topright", inset = .02,legend=c("Hombres", "Mujeres"),
       col=c("red", "blue"), fill=c("red", "blue"), cex=0.8)
```



Podemos notar por medio de la gráfica que los hombres tienen ingresos más altos y variables, por otro lado, las mujeres tienen ingresos más centrados y con menos variabilidad

```
# Estimar el ingreso promedio de las subpoblaciones:

# Filtrar las bases de datos por hombre y mujer
dfH <- df20 %>% filter(genero == "H")
dfH

## # A tibble: 31 x 3
##      nro ingreso genero
##    <dbl>   <dbl> <fct>
##  1     2     960 H
##  2     3     670 H
##  3     6    2346 H
##  4     8    1126 H
##  5     9    1780 H
##  6    10    2135 H
##  7    11    1446 H
##  8    12    2245 H
##  9    14    2287 H
## 10    15     686 H
## # ... with 21 more rows

dfM <- df20 %>% filter(genero == "M")
dfM
```

```
## # A tibble: 25 x 3
##      nro ingreso genero
##    <dbl>   <dbl> <fct>
##  1     1     800 M
##  2     4     688 M
##  3     5    1025 M
##  4     7     780 M
##  5    13    1760 M
##  6    16     997 M
##  7    18    1567 M
##  8    19    1456 M
##  9    23    1388 M
## 10    25    1653 M
## # ... with 15 more rows
```

```
# mirar promedios por genero:

## Promedio del ingreso Mujer:
mean(dfM$ingreso)
```

```
## [1] 1384.92
```

```
## Promedio del ingreso Hombre:
mean(dfH$ingreso)
```

```
## [1] 1719.548
```

los hombres tienen un promedio más alto en cuestión de ingresos, pero lo que nos importan es cual es menos variable, para determinar cuál es más homogéneo

para esto utilizaremos: $({y}) = * () $

```
# Datos a utilizar:
skH <- var(dfH$ingreso)
nkH <- length(dfH$ingreso)
skM <- var(dfM$ingreso)
nkM <- length(dfM$ingreso)


# Varianza de la media muestral:

## Hombres
varmuH <- varmu(N,n,skH,nkH)
varmuH
```

```
## [1] 15228.74
```

```
## Mujeres:
varmuM <- varmu(N,n,skM,nkM)
varmuM
```

```
## [1] 10475.3
```

Como lo vimos en el Boxplot, los datos de los hombres presentan una variabilidad mucho más alta que el de las mujeres, por ende, nuestra población más homogénea es la de las mujeres.