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 del departamento de Antioquia para estimar la **proporción de votantes** registrados con intención de voto por un candidato . Para ello:

* Se dispone del listado de todos los habitantes mayores de edad que conforman el municipio, Habitantes.
* Se opta por seleccionar una MAS sin remplazo de dicha población, .
* Se encuentra que en dicha muestra sólo personas estaban inscritas para votar y de éstas estaban a favor del candidato .

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

## Solución

extraemos los datos del enunciado anterior

: 10.000 habitantes *(de la población total)*

: 500 hab. tamaño de la muestra

: Desconocido

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

: 350

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

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

## I.C.

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

Como es desconocido utilizare:

*(1)*

# Varianza estimada:  
var\_p <- (N-n)/N \* pg\*(1-pg)/(nk-1)  
print("# Varianza estimada:")

## [1] "# Varianza estimada:"

print(var\_p)

## [1] 0.0006671777

un intervalo al esta dado por:

## [1] "# IC para P\_k"

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

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

**Con esto vemos un IC al** 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 trabaja?

¿Como estimar el **tiempo promedio** en meses, de los jefes de hogar que 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

*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 (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 en el barrio muestreado de Medellín los jefes de hogar que no trabajan están dentro del y el ; 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:

#Media estimada  
yk <- mean(tm)  
yk

## [1] 6.166667

Ahora estimamos la varianza de los datos de interés:

# Varianza estimada  
sk <- var(tm)  
sk

## [1] 13.36667

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

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

#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 para (promedio de meses sin trabajar) están dados por:

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"

print("# Limite Inf , Limite Sup")

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

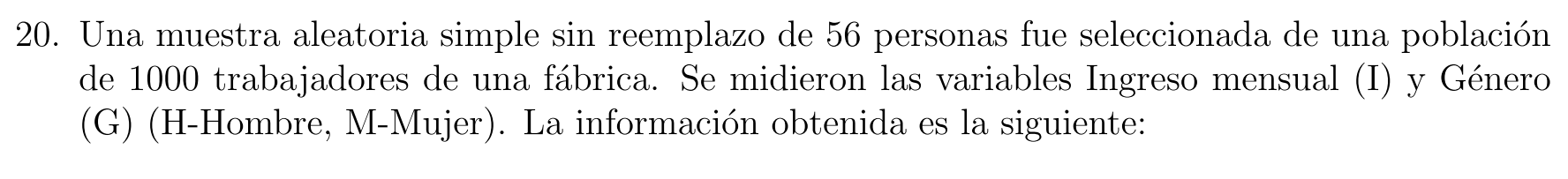
paste("(", LI, ",", LU, ")")

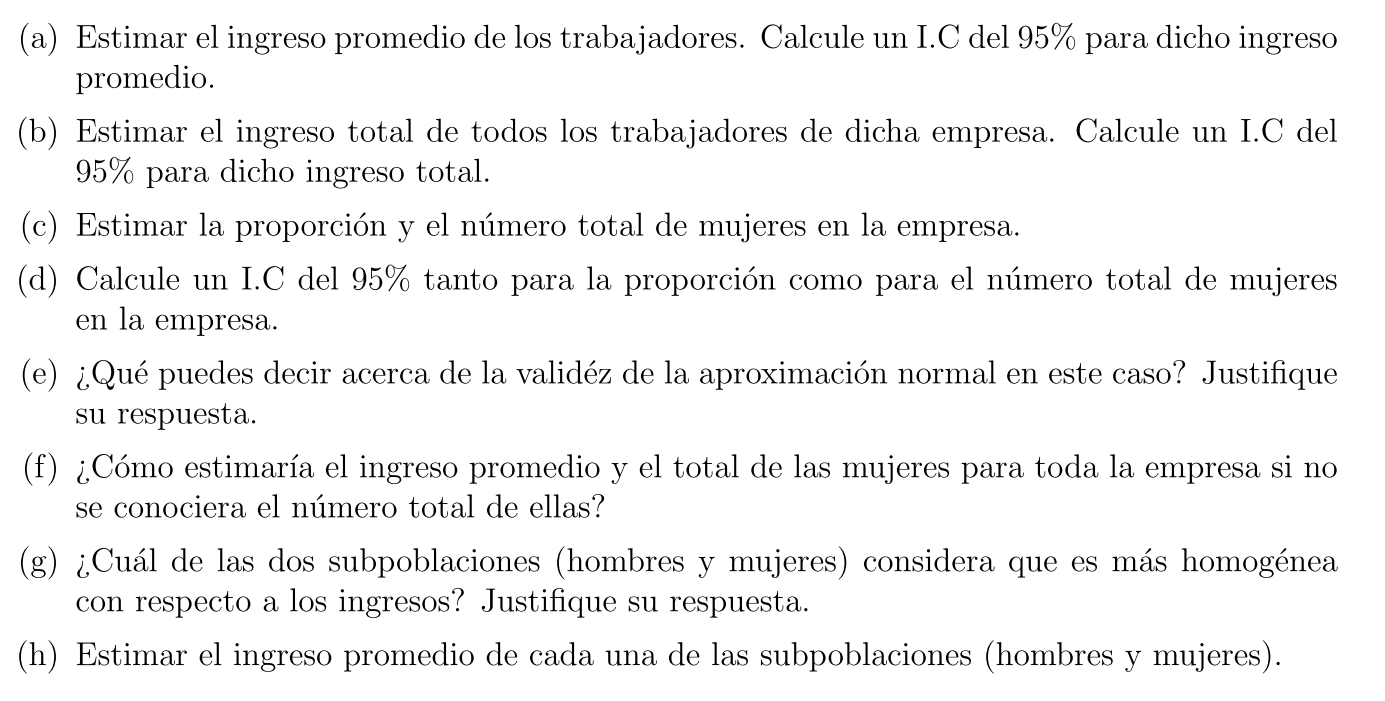
## [1] "( 2.34911463133005 , 9.98421870200329 )"

Podemos afirmar con un 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

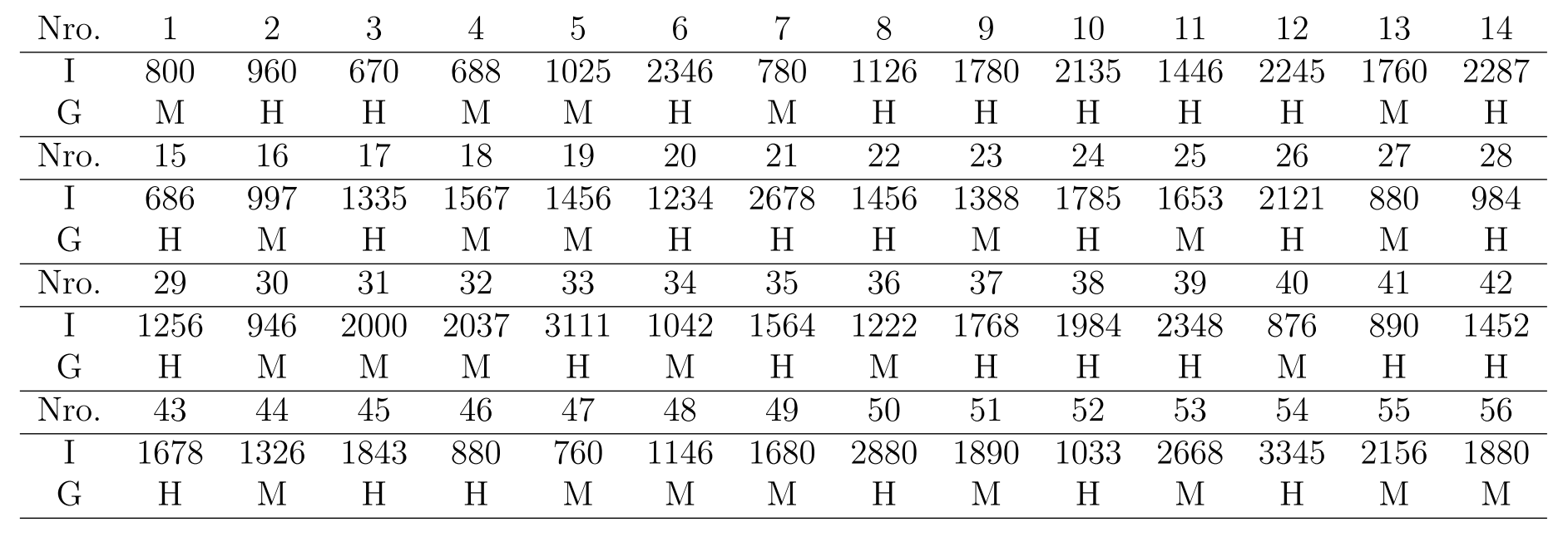
Observemos el ejercicio:





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

library(readxl)  
df20 <- read\_excel("data\_sets20.xlsx")



## Literal (a):

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

un estimador puntual para

# 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 utilizaremos:

* La varianza muestral
* el límite en el error de la estimación
* un estadístico , ya que

# 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:

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 y los 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 con nuestro (población donde se extrajo la muestra)

# 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

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

# Proporción estimada:  
pg <- fem/n  
pg

## [1] 0.4464286

la proporción estimada de mujeres en la fábrica representa aproximadamente

1. para el caso del total de mujeres en la empresa, solo basta con tomar el resultado anterior y multiplicarlo por el

# Total poblacional:  
N\*pg

## [1] 446.4286

el total estimado de mujeres en la población de la fábrica es

## 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

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

## Literal (e):

La aproximación es buena, es decir, gracias a que tenemos un 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.

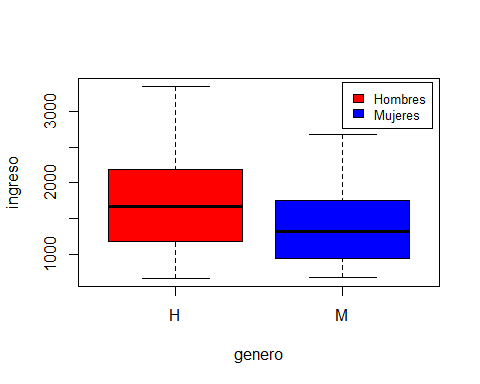
## Literal (f):

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 , 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.