Muestreo Taller 1

Jhonatan Smith Garcia.

18/3/2021

Taller muestreo 1

MUESTREO - TAREA #01 MAS Profesor: RAUL ALBERTO PEREZ AGAMEZ

Muestreo Aleatorio Simple sin Reemplazo (MAS)

Ejercicio 19. Considere que la siguiente informacion corresponde a una muestra aleatoria piloto de una comunidad de 10000 personas.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
genero	f	m	f	m	f	m	f	f	f	m	f	f	m	m	m
EC	c	s	c	s	c	c	c	o	o	s	o	s	s	s	c
I	2.0	2.5	4.0	3.8	7.2	10.0	5.6	4.9	3.3	4.0	3.5	5.7	10.0	8.1	4.4

	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
genero	f	m	f	f	m	m	f	m	f	m	f	m	f	m	f
EC	o	s	o	c	s	c	s	o	c	s	o	s	o	c	s
I	6.6	7.3	8.0	9.0	3.9	7.1	4.9	2.3	3.9	11.1	7.3	6.5	5.8	4.0	3.0

Las preguntas que plantea el problema son:

a) Estimar ingreso promedio de la comunidad con un error absoluto no mayor a 0.2 salarios minimos por persona con una confiabilidad minima del 95%

b) Estimar el ingreso total de la comunidad con un error absoluto no mayor a 2000 salarios minimos en total con una confiabilidad minima del 95%

c) ¿Cual deberia ser el tamaño de la muestra necesaria para estimar μ, el numero promedio de salarios por personas, con un limite para el error de estimacion de magnitud 0.1 salario?

d) ¿Cual deberia ser ele tamaño de muestra necesaria para estimar a τ, el total de ingresos percibidos por la comunidad (o el numero total de salario percibidos por la comunidad), con un limite de error de estimacion de magnitud 1000 salario?

Solucion: Inciso C) Hasta ahora, la informacion que se dispone para responder el problema es la siguiente:

1) N= 10000 personas
2) $n = 30$ muestra piloto de la tabla
3) Se pide un error B no mayor al 0.1 salarios.

Para resolver el problema del tamaño de la muestra para cumplir los requisitos del inciso c, se deben calcular primero ciertos valores a computar para usar la siguiente dormula.

- $n = (N * S^2)/(N-1) * D + S^2$ con n el tamaño de la poblacion y $D = \frac{B^2}{Z_{\frac{\alpha}{2}}}$

Se conoce el valor de N pero se desconoce el de S^2 (La varizanza muestral).

Calculando la varianza muestral del ingreso de la muestra piloto con la funcion de R var(x) entonces:

```
var(data$I)
```

```
## [1] 5.997713
```

Por tanto, $S^2 \approxeq 5.98$.

Por otra parte $Z_{\frac{\alpha}{2}}$ con una confianza del 95% representa el quantil de una normal estandar N(0,1). Dicho valor es inmediato y es $\approxeq 1.96$ (calculado con la funcion qnorm en R)

El error B= 0.1 y es dado por el ejercicio.

Finalmente, D entonces será dada por:

$D = \frac{0.1^2}{1.96} \approxeq 0.005102041$

Asi, reemplzanado en la formula de n, tamaño de muestra necesario para cumplir con las condiciones pedidas tendremos que:

$n = \frac{10000*5.98}{(9999)*0.005102041+5.98} \approxeq 1050$ muestras a tomar (redondeando al entero superior)

Esto implica que, el tamaño de la muestra necesario para estimar el promedio de ingresos en la comunidad dada con un error de estimacion de 0.1 a un nivel de confianza del 95% es de 1050 personas.

El tamaño de la muestra es considerablemente grande puesto que se exige un nivel de confianza alto (95) y un error cometido muy bajo. Es esperable que, para disminuir dicho error, es necesario la toma de muchas mas muestras para entender mucho mejorel comportamiento de la poblacion.

Se debe tener en cuenta como se usó un estimador de la varianza poblacional, la muestra piloto no debe pertenecer a la muestra final.

Inciso D)

Nuevamente, para responder a la pregunta acerca del tamaño de la muestra dada, basta revisar la siguiente formula que es util para calcular el tamaño de n para estimar el total poblacional τ.

$n = \frac{1}{\frac{1}{N}+\frac{N-1}{N}*\frac{1}{n_0}}$ con $n_0 = \frac{N^2*\sigma^2*Z_{\frac{\alpha}{2}}}{B_\tau^2}$

En este caso, como se desconoce σ (la varianza poblacional) se usará nuevamente a S^2 como un estimador de la varianza. Nuevamente, la muestra piloto no ha de ser parte de la muestra final.

Nuevamente $N = 10000$, $S^2 \approxeq 5.98$, $B_\tau^2 = 1000$, $Z_{\frac{\alpha}{2}} \approxeq 1.96$, entonces reemplazando en n_0

```
n0 = (10000^2*1.96^2*5.98^2)/1000^2

nt =1/((1/10000)+((9999/1000)*1/13737.72))
```

$n_0 = \frac{10000^2*1.96^2*5.98^2}{1000^2} \approxeq 13737.7$

Luego, teniendo a n_0 podemos calcular a n. Reemplazando en la formula anterior se tiene que:

$n = \frac{1}{\frac{1}{10000}+\frac{9999}{1000}*\frac{1}{13737.7}} \approxeq 1208$

Con esto, se tiene que el tamaño de la muestra necesaraio para estimar el total de los ingresos de la comunidad es de 1208 muestras, asegurando una confianza del 95% y un error de 1000 salarios.

NOTA: Como la muestra piloto se usó para calcular la varianza (una estiamcion de la misma) entonces la muestra piloto no debe pertenecer a la muestra final.

Inciso A: Un estimador de μ (media poblacional) para estimar el promedio de ingresos de los salarios, es la media muestral.

Sea X la variable aleatoria que mide el ingreso de los encuestados en la muestra piloto, entonces $\bar{X}$ es un estimador insesgado de μ

```
xbar<- mean(data$I)
```

$\bar{X} = \sum_i^n \frac{x_i}{n}$ y para este caso, n=30 y x_i es cada uno de los salarios en la muestra piloto. Asi, $\bar{x} \approxeq 5.66$

Ahora, un intervalo de confianza para la media muestral está dado por:

$(\bar{x} - t_{1-\frac{\alpha}{2},n-1} * \sqrt{\frac{S^2}{n}\frac{N-n}{n}}, \bar{x} + t_{1-\frac{\alpha}{2},n-1} * \sqrt{\frac{S^2}{n}\frac{N-n}{n}})$

Este intervalo de confianza se puede usar debido a que, n=30, asi que una aproximacion de dicho intervalo a traves de una t-student del (1-α)*100% y n-1 grados de libertad es efectivo por el tamaño de la muestra piloto. De esta forma:

```
t.test(data$I)$conf.int ##Un IC calculado con la funcion t.test.
```

```
## [1] 4.742187 6.571147
## attr(,"conf.level")
## [1] 0.95
```

Un IC calculado con R será (4.742187 ;6.571147).

Ahora, calculando de manera manual dicho intervalo, reemplazando a la media, el valor de la varianza muestra, la poblacion (N) y la muestra (n):

Limite inferior: 5.66-2.04523 * $\sqrt{\frac{5.98^2 * 9999}{10000 * 30}}$ Limite inferior $\approxeq$ 3.428

Limite superior:5.66+2.04523 * $\sqrt{\frac{5.98^2 * 9999}{10000 * 30}}$ Limite superior $\approxeq$ 7.9

Finalmente, un intervalo de confianza para μ al 95% será:

(3.428 ; 7.9) lo que significa que, con una confianza del 95% el valor de μ se encuentra en ese intervalo "un 95% de las veces".

Ahora, si miramos la amplitud de los dos intervalos calculados; la amplitud del intervalo de confianza dado por la funcion t.test es un intervalo calculado con una t-student es menor que la amplitud del intervalo calculado manualmente. Sin embargo, estos valores estan "relativamente" cercanos entre si, brindando una muy buena idea de donde podria encontrarse el valor real del promedio de salarios.

El nivel de confianza y la amplitud del intervalo varían conjuntamente, de forma que un intervalo más amplio tendrá más probabilidad de acierto (mayor nivel de confianza), mientras que para un intervalo más pequeño, que ofrece una estimación más precisa, aumenta su probabilidad de error.

Inciso B: Se pide calcular el total poblacional, para esto, se usará el estimador insesgado $\widehat{\tau}$ del total poblacional τ.

Ahora para calcular una estimacion del total ingreso de la poblacion, se define:

$\widehat{\tau} = \bar{x} * N$ donde N es el total poblacional (10 mil) y $\bar{x}$ es la media muestral.

En este caso, $\widehat{\tau}$= 10000*5.66=56600.

Ahora un IC de este ultimo estimador será dado por por el IC anterior.

$\bar{x}$*(3.428 ; 7.9)=(19.40249,44.714) será un IC con una confianza del 95%

Analogamente para el otro intervalo calculado con la funcion de R, se tinene que:

$\bar{x}$*(4.742187 ;6.571147)=(26.84078,37.19269)

En ambos casos, la interpretacion es la misma per con valores diferentes. Con un nivel de confianza del 95%, el valor real del total de ingresos de la comunidad está dentro de los intervalos dados.

NOTA: Ambas aproximaciones del IC son validas. El segundo fue el que se saca de la funcion de R y el primero fue el calculado a mano. La pregunta para el docente es, ¿en que radica la diferencia? ¿por qué toman valores ligeramente diferentes?

Con respecto a este punto. Si bien los incisos c y d fueron inmediatos de resolver, los otros dos no. Todavia representan algo de confusion y tengo dudas respecto a los planteamientos que les di. Si bien, intenté aplicar todo lo que se vio en clase, de ser posible, me gustaria que me explicase en su debido espacio el como resolver especificamente este ejercicio. Gracias de antemano.

Caso de estudio #3: Suponga que se desea realizar un estudio de muestreo en un municipio A del departamento de Antioquia para estimar la proporcion de votantes registrados con intencion de voto por un candidato X. Para ello:

a) Se dispone del listado de los habitantes mayores d edad que conforman el municipio, N = 10000 habitantes
b) se opta por seleccionar una muestra aleatoria simple sin reemplazo de dicha poblacion, m = 500
c) Se encuentra que en dicha muestra solo 350 estaban inscritas para votar y de esas 150 estaban a favor del candidato X ¿Como estimar la porporcion de registrados con intencion de voto por el candidato X?

Solucion: Para responder a la pregunta debemos identificar **elemento, poblacion, unidad de muestreo y marco.**

Elemento: Cada uno de los habitantes del municipio que puede votar. Poblacion: Todos los habitantes registrados en el municipio con condicion de votante (pertenecientes a N) Unidad de muestreo: Cada uno de los habitantes que tiene facultad de votar. Marco: El listado de todos los habitantes que pueden votar.

Teniendo esto claro, se debe tener en cuenta un analisis de subpoblaciones ya que la muestra tiene un total de 500 personas seleccionadas, de las cuales **350 estaban inscritas para votar** es decir, de la poblacion total N, existe un porcentaje que se encuentra inscrito en el municipio A para votar allí y otro porcentaje que no lo está.

Por consiguiente, tenemos una M.A.S de tamaño 500 ($n = 500$) pero de esta muestra, tenemos que tener en cuenta la poblacion $n_1 = 350$ que representa la cantidad de personas registradas que SI puede votar dentro de el municipio A. Esta es la muestra de interes.

Ahora, gracias al listado se dispone de los N habitantes enumerados. Esto asegura una M.A.S correctamente recolectada.

y toma valor 1 si el habitante tiene intencion de voto y 0 otro caso.

Finalmente, se nos pide calcular la proporcion de votantes por el candidato X del municipio A. Para ello, tenemos que:

$\widehat{p}$ será el calculo de dicha proporcion tal que:

$\widehat{p}$= $\frac{150}{350}$=0.4285714, aproximando seria igual a 0.43

Este es el calculo de la proporcion de los habitantes que **estan inscritos** que tienen intencion de voto por el candidato X.

Ahora, segun esto, la proporcion es de cerca del 43%. Ahora, calculemos un IC al 95% para dicha proporcion.

Como $\widehat{p}$ es 0.3 y la muestra es de $n = 350$ podemos utilizar la aproximacion a la normal para calcular dicho intervalo. Dicho de otra manera, un IC para la proporcion de habitantes a votar por el candidato X será dada por:

($\widehat{p}$-B;$\widehat{p}$+B) donde B se conoce como el limite del error donde $B = Z_{1-\frac{\alpha}{2}}\sqrt{\frac{\widehat{p}*(1-\widehat{p}*(N-n))}{(n-1)*N}}$

Ahora, reemplazando obtenemos que B= 1.96 * $\sqrt{0.43*0.57*9650/(349*10000)} \approxeq 0.05102$

Ahora, un IC al 95% para la proporcion pedida estará dado por:

(0.43-0.05102;0.43+0.05102)= (0.37898,0.48102)

Ahora, esto se interpreta como, en un 95% de las veces, la poroprcion de votantes que estará a favor del candidato X es un valor perteneciente al intervalo anterior.

Caso de estudio 4: Una multinacional desea abrir nuevos puestos de trabajo en un barrio de Medellin.

Para ello necesita estimar de las personas que NO trabajan, el tiempo (en meses) que el jefe de hogar ha completado sin trabajar. La empresa cuenta con un listado de los hogares del barrio bajo estudio, conformado por 1000 hogares. Se decide:

a) Se selecciona una muestra piloto de 10 hogares y se entrevista al jefe del hogar
b) los datos que se obtivieron fueron los siguientes:

```
data2 <- data.frame( Trabaja=c(1,0,1,0,0,1,0,1,0,0), Tiempo=c(0,3,0,12,5,0,7,0,8,2))
attach(data2)
```

	1	2	3	4	5	6	7	8	9	10
Trabaja	1	0	1	0	0	1	0	1	0	0
Tiempo	0	3	0	12	5	0	7	0	8	2

¿Como estimar la proporcion de hogares donde el jefe de hogar No trabaja?

Solucion

Podemos identificar que la poblacion N=1000 el total de hogares en el barrio.

La muestra n=10 es una muestra pequeña.

De manera rapida de calcular el estimador de la porporcion pedida es $\widehat{p}$ de la siguiente manera

```
p1 = sum(data2$Trabaja)/length(data2$Trabaja)
p2 =1-p1
```

$\widehat{p}$=0.6 es el estimador de la proporcion de cabezas de hogar que NO trabajan. Sin embargo al ser una muestra tan pequeña la estimacion de esta proporcion no ha de ser muy buena.

Como la muestra es pequeña, utilicemos la aproximacion a una t-student para calcular un IC al 95%, teniendo en cuenta que dicha estimacion no es la recomendable a usar y se espera que dicho intervalo se comporte de manera "extraña"

Calculando el error $B = t_{1-\frac{\alpha}{2},n-1}\sqrt{\widehat{p} * (1 - \widehat{p}(N-n)/N(n-1))}$

De esta manera obtenemos que $B \approxeq 0.37$

De esta manera, un IC calculado aproximando por T-student será

(0.6-0.37;06+0.37)=(0.23,0.97)

Claramente es un intervalo que practicamente abarca todo el rango de posibilidades. Es muy poco util para hacer predicciones acertivas.

NOTA: Dando otro enfoque a travez de una funcion de R del paquete epitools.

```
library(epitools)
```

```
## Warning: package 'epitools' was built under R version 4.0.3
```

```
binom.exact(6,10)
```

```
##   x  n proportion     lower     upper conf.level
## 1 6 10        0.6 0.2623781 0.8784477       0.95
```

Esta funcion da un IC para cuando la muestra es pequeña y la variable de interes se "comporta" como una Bernoullie. Aun asi, es poco efectivo. La muestra sigue siendo demasiado pequeña para hacer estimaciones mas adecuadas.

NOTA PARA EL DOCENTE: La aproximacion binomial no me fue muy claro el como operarla. A nivel teorico es claro el argumento detras de esto, sin embargo no me queda muy claro el como se operan las ecuaciones para hacer los calculos. En un comentario al final de recibir el trabajo dando una breve explicacion seria de mucha ayuda. Gracias!!

Ahora, para estimar el tiempo en meses que lleva el jefe del hogar sin trabajar, entendemos que, la poblacion son las familias, con dos respectivas sub poblaciones, los que trabajan y los que no.

Se entiende entonces que si n es el tamaño de la muestra total (que son 10) ahora sea n_1 los que no trabajan y sacamos una media muestral del tiempo que llevan sin trabajar, tendriamos que:

$\sum_{i=1}^{n} \frac{3+12+5+7+8+2}{6} \approxeq 6.17$

Esta media muestral nos dice, en promedio, de los cabeza de familia, cuanto tiempo llevan sin trabajar. Segun esto, seria aproximadamente 6.17 meses. Redondeando de manera que la cifra cobre mas sentido, unos 6 meses.

Ahora, si queremos estimar un IC al 95% para esta media, podriamos hacerlo pero, se ha de tener en cuenta que por ser una muestra muy pequeña, tendremos tambien el mismo inconveniente, el IC no será muy determinista.

Ahora, teniendo en cuenta esto, el IC al 95% estará dado por:

$\bar{y_1} +_- t_{1-\frac{\alpha}{2},n-1}\sqrt{\widehat{V}(\bar{y_1})}$

Ahora, si $\bar{y_1}$ lo definimos como la media muestral de los datos de la subpoblacion objetivo (tiempo en meses del jefe de familia sin empleo) podriamos reemplazar los valores, calculando con una T-student con 5 grafos de libertad al cuantil 0.975. Sin embargo, para una muestra de tamaño 6 este cuantil no será deterministico. De hecho, no se realizará el calculo debido a que no será un valor representativo.

La conclusion mas coherente con los datos actuales, es aumentar el tamaño de la muestra.

NOTA: Sinceramente ante este ejercicio, la muestra tan pequeña evita sacar una conclusion mas allá de ampliar la muestra. En este caso profe, ¿el ejercicio requeria otro analisi? Debido a que no hay mucha "maniobrabilidad" con tan pocos datos. Mas allá de ver que se puede calcular (proporciones, medias muestrales, IC). Nuevamente, gracias y agradezco la paciencia. Nunca he usado antes Latex y tampoco era particularmente buen con el Markdown. Gracias!