Universidad Nacional de Tres de Febrero

Maestría en Generación de Información Estadística
Teoría y Técnicas de Muestreo
TP Diseño de muestra y estimadores

Augusto E. Hoszowski

Ejercicio I

En la siguiente población de dieciocho alumnos se tienen definidas dos variables X e Y.

Alumno	X	Y
a	6	14.0
b	9	20.0
c	5	12.0
d	4	10.0
e	2	5.0
f	7	12.0
g	10	24.0
h	4	5.0
i	12	21.0
j	5	9.0
k	8	18.0
l	12	20.0
m	5	8.0
n	9	15.0
o	2	2.5
p	6	11.0
q	11	20.0
r	8	15.0

Se desea estimar mediante una Muestra Aleatoria Simple de 9 alumnos la media de X, la media de Y y la razón $R = \bar{Y}/\bar{X}$, mediante los estimadores usuales

$$\bar{y}$$

$$\bar{x}$$

$$r = \frac{\bar{y}}{\bar{x}}$$

- Cuántas muestras posibles hay?
- Hallar la varianza y CV de los estimadores $\bar{y}$ y $\bar{x}$
- Calcular la varianza aproximada y CV aproximado del estimador r
- Seleccionando 10,000 muestras aleatorias simples, estimar la varianza y CV de r. Comparar con el resultado obtenido en el punto anterior.

Ejercicio II

La tabla *tabla_muestras_posibles.xlsx* contiene 20 unidades, a las que se le midieron una variable Y. Será nuestro universo/marco de muestreo. Se desea estimar la media de Y mediante una MAS(10). Se compararán estos estimadores:

- Media muestral
- Media muestral truncada, eliminando 10% inferior y 10% superior (en nuestro caso resulta el menor valor de la muestra y el mayor valor de la muestra)
- Mediana

1. Listar con R todas las muestras posibles y calcular para cada una de ellas media, media truncada y mediana
2. Agregar a cada muestra la media, media truncada y mediana de los diez valores.
3. Verificar que la media muestral es un estimador insesgado de la media poblacional, lo que no se cumple para la mediana y la media truncada
4. Graficar mediante tres histogramas las tres series de estimaciones. Tienen una distribución aproximadamente normal? Incluir en los gráficos una línea vertical de referencia que indique la ubicación del parámetro a estimar.
5. Tabular CV y Error Medio Cuadrático de los tres estimadores
6. Qué estimador le parece preferible?

Sugererencia para la resolución del Ejercicio II
Para listar los subconjuntos posibles podemos hacer por ejemplo,

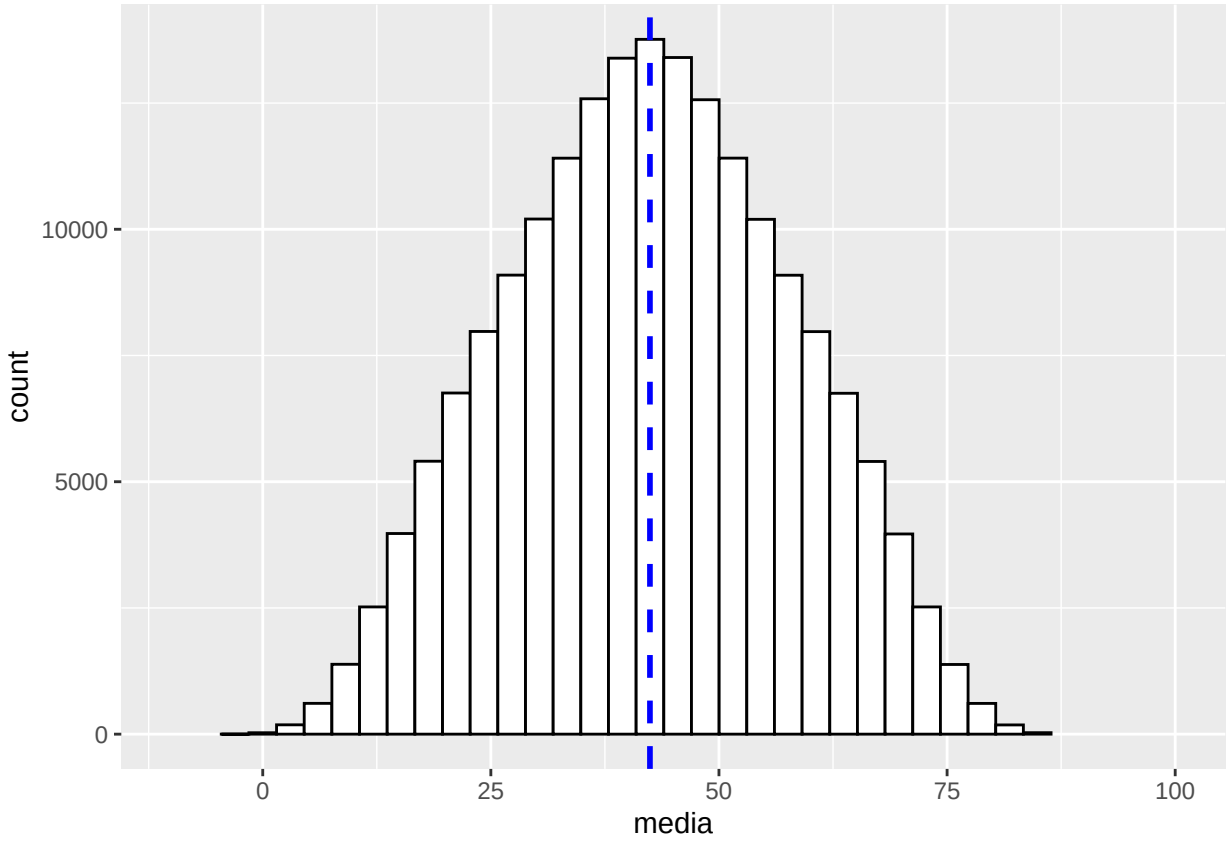
```
df_tabla <- read_excel("tabla_muestras_posibles.xlsx")
df_muestras <- data.frame(matrix(unlist(combn(df_tabla$Y,10, simplify = FALSE)),
                         ncol=10, byrow=TRUE))
```

Para agregar a cada muestra la media muestral y el correspondiente histograma podemos hacer

```
media <- apply(df_muestras[,1:10],1,mean)
df_muestras$Media <- media


p <- ggplot(df_muestras, aes(x=media)) +
  geom_histogram(bins=30, color="black", fill="white") +
  coord_cartesian(xlim = c(-10, 100))

p <- p+ geom_vline(aes(xintercept=mean(Media)),
              color="blue", linetype="dashed", linewidth=1)
p
```



Ejercicio III

Supongamos que la tabla de radios censales 2010 es nuestro universo bajo estudio. Deseamos estimar, encuestando en su totalidad una Muestra Aleatoria Simple de n=240 radios censales:

- Total de población
- Total de hogares que habitan en viviendas tipo Casa
- Total de hogares que habitan en viviendas rancho/ casilla
- Proporción de hogares que habitan en viviendas rancho/ casilla

Los tres primeros parámetros los estimaremos con el estimador usual del total en un MAS ($N \cdot \bar{y}$), que sabemos es el estimador de Horvitz-Thompson. El cuarto parámetro es una razón, lo estimaremos como es usual en el MAS con la razón muestral.

1. Hallar los cuatro parámetros (o sea los cuatro valores poblacionales)
2. Hallar los CV de los cuatro estimadores (el del estimador (d) será una aproximación)
3. Qué estimador tiene el CV más alto?
4. Cómo son los CV de los estimadores (c) y (d)?
5. Qué tamaño de muestra se necesitaría para que el estimador del total de población sea (aproximadamente..) 2%?.
6. Qué tamaño de muestra se necesitaría para que el estimador del total de hogares que habitan en rancho/casilla sea (aproximadamente..) 2%?.

7. Seleccionar una MAS de tamaño n=240 y con *survey* estimar total de población, total de hogares que habitan en rancho/casilla y proporción de hogares que habitan en rancho/casilla, y los respectivos CV e intervalos de confianza con un nivel de confianza de 90%.
    - Los intervalos contienen al parámetro en cuestión?
8. Comentar los resultados hallados

Ejercicio IV

Supongamos que la tabla de radios censales 2010 es nuestro universo bajo estudio. Deseamos estimar, encuestando en su totalidad una Muestra Sistemática de n=240 radios censales:

- Total de población
- Total de hogares que habitan en viviendas tipo Casa
- Total de hogares que habitan en viviendas rancho/ casilla

Se desea comparar dos estrategias (las dos utilizando como estimador la media muestral):

1. Muestreo sistemático, ordenando la tabla por Provincia-Total de viviendas del radio
2. Muestreo sistemático, ordenando la tabla por un número pseudo aleatorio

- Hallar los tres parámetros (o sea los tres valores poblacionales)
- Las estrategias 1 y 2 son insesgadas?
- Hallar CV, deff, sesgo relativo y EMC de cada estrategia, *seleccionando todas las muestas posibles*
- Presentar en una tabla los resultados
- Comentar los resultados

Ejercicio V (continuación del ejercicio IV)

Deseamos ahora probar otra estrategia para estimar los parámetros del ejercicio IV: selección de la muestra mediante Madow, con la cantidad de viviendas del radio como variable auxiliar

1. Seleccionar mediante **sampling** una muestra de n=240 radios mediante Madow, con total de viviendas del radio como variable auxiliar. Ordenando la tabla según código de radio (jurisdicción+departamento+fracción+radio)
2. Con **survey** estimar los parámetros, CV y deff correspondiente
3. Repetir los pasos 1 y 2 diez veces
4. En una tabla resumir los resultados del ejercicio anterior y lo hallado en este ejercicio

Ejercicio VI

En un ballotage, un candidato X encarga una estadístico una muestra aleatoria simple de electores para saber si gana o pierde la elección. Supongamos que la gente no miente y que no cambia el voto luego de la encuesta. Por motivos de costo se encuestan a 400 personas. De ellas, 212 afirman que votarán por X.

1. Qué le informa el estadístico al candidato?. La información que da la encuesta es útil?
2. Cuál es el CV del estimador?

Ejercicio VII

En un ballotage, un candidato X encarga una estadístico una muestra aleatoria simple de electores para saber si gana o pierde la elección. El candidato afirma que supone que obtendrá un porcentaje de votos cercano al 50%. Que es consciente que con eso no se puede determinar un tamaño de muestra para garantizar si gana o pierde la elección, pero que está conforme si el intervalo de confianza al 95% de la estimación tiene una

amplitud total de 1%. Qué tamaño de muestra sería necesario? (como suponemos que se trata de una gran ciudad, podemos obviar el factor de corrección por población finita)

Ejercicio VIII

Se selecciona una muestra aleatoria simple de 24 hogares de una localidad pequeña para indagar cierta característica rara. En la muestra ningún hogar la presenta. Puede dar un intervalo de confianza al 90% para la proporción de hogares con esa característica? (puede utilizarse el paquete de R *binom*)