UNIVERSIDAD NACIONAL DE COLOMBIA
FACULTAD DE CIENCIAS, SEDE BOGOTÁ
DEPARTAMENTO DE ESTADÍSTICA
INFERENCIA ESTADISTICA
Lineamientos del trabajo final (Parte 2)

OJO: ¡LEER ANTES DE ENVIAR!

- **Fecha límite de entrega:** Domingo de la semana 15, 11:59pm (Penalidad de 0.1 sobre la nota final por cada hora o fracción de hora de retraso).
- **Formato:** Se debe enviar un único archivo del reporte en formato PDF; no Word, no Excel u otros. (Penalidad por incumplimiento: -0.2 sobre la nota final). También se deberá enviar una carpeta comprimida (.ZIP o .RAR) vía correo electrónico a mearrietap@unal.edu.co con los scripts debidamente documentados para responder a cada punto del trabajo (Si no se envía dicha carpeta, la calificación del trabajo será de 0.0).
- **Nombre del archivo:** "Trabajo final IE Grupo" *[Inserte el número de su grupo de trabajo tal y como aparece en la hoja de cálculo en el Drive].* (Penalidad por incumplimiento: -0.2 sobre la nota final).
- **Mecanismo de entrega:** El trabajo debe ser subido al classroom UNA SOLA VEZ por el líder del grupo u otro integrante si el líder tiene inconvenientes. Absténganse de enviar los trabajos a través del correo electrónico. (Penalidad por incumplimiento: -0.2 sobre la nota final). Las carpetas comprimidas deben ser enviadas por correo electrónico con el mismo nombre del archivo como asunto. Ambos elementos tienen la misma fecha límite de entrega.

- La idea de este trabajo es que ustedes hagan uso de los conceptos vistos en clase para aprender cómo usar la simulación para verificar propiedades de estimadores.
- Es recomendable mas no necesario que las implementaciones se hagan en R o Python. Si alguno desea utilizar uno diferente, hágamelo saber.
- Creen un script o archivo de código para dar solución a cada punto y documéntenlo bien, de modo que sea claro cómo y por qué estructuraron su código de dicha manera. Adicionalmente, en un documento de texto deberán registrar las respuestas a las preguntas hechas en cada situación.
- El número del grupo que les fue asignado en el documento drive se denominará κ a lo largo del informe. Revisen porque algunos puntos dependerán de ese número.
- Lo primero que deberán hacer es ver el video en la carpeta de drive denominada "proyecto", ya que, allí se explican algunos aspectos básicos. También aparecen archivos de código de los ejemplos desarrollados en el video que les pueden ser útiles.

A. Comparación de varios intervalos de confianza para una proporción en una muestra aleatoria Bernoulli.

Vimos en clase que para estimar por intervalo la proporción, p, en el modelo Bernoulli; usando una cantidad pivote asintótica hay tres posibilidades:

Posibilidad 1. A partir del teorema central del límite, se obtiene que:

$$p\left(-z_{1-\frac{\alpha}{2}} \leq \frac{\hat{p}_n - p}{\sqrt{p(1-p)}} \leq z_{1-\frac{\alpha}{2}}\right) \approx 1-\alpha,$$

siendo $\hat{p}_n = \overline{X}_n$, la media muestral. Aunque la variable aleatoria no es monótona en p, es posible obtener un intervalo de confianza despejando el parámetro de allí.

Posibilidad 2. A partir del teorema central del límite y reemplazando la varianza del modelo por un estimador consistente se obtuvo que el intervalo podía ser calculado como:

$$ICA_{100(1-\alpha)\%}(p) = \hat{p}_n \mp z_{1-\frac{\alpha}{2}}\sqrt{\hat{p}_n(1-\hat{p}_n)}.$$

Posibilidad 3. A partir del teorema central del límite y usando el método Delta con una transformación estabilizadora de varianza, se obtuvo que:

$$ICA_{100(1-\alpha)\%}(p) = \left[\sin^2\left\{\arcsin\left(\sqrt{\hat{p}_n}\right) - \frac{z_{1-\frac{\alpha}{2}}}{2\sqrt{n}}\right\}, \sin^2\left\{\arcsin\left(\sqrt{\hat{p}_n}\right) + \frac{z_{1-\frac{\alpha}{2}}}{2\sqrt{n}}\right\}\right].$$

Adicionalmente, se cuenta con la posibilidad de hacer muestreo por bootstrapping para obtener un intervalo de confianza para la proporción.

Posibilidad 4. A partir de la única muestra de datos que se pueda obtener, $x_1, x_2, \ldots, x_n$, calcule la estimación puntual $\hat{p}_{n,orig} = \overline{x}_n$. Ahora, genere $B = 1000$ muestras Bootstrap, cada una del tamaño de la muestra original y calcule las estimaciones Bootstrap de la proporción:

$$\left\{\hat{p}_{n,boot_i} = \overline{x}_{n,i}\right\}_{i=1}^{B}.$$

El *bootstrap basado en percentiles* simplemente definiría como límites del intervalo de confianza a:

- Límite inferior=percentil $\frac{\alpha}{2}$ de la secuencia de valores $\left\{\hat{p}_{n,boot_i} = \overline{x}_{n,i}\right\}_{i=1}^{B}$.
- Límite superior= percentil $1-\frac{\alpha}{2}$ de la secuencia de valores $\left\{\hat{p}_{n,boot_i} = \overline{x}_{n,i}\right\}_{i=1}^{B}$.

Exploren el uso de la función `quantile` para extraer percentiles empíricos de un arreglo de datos.

Posibilidad 5. El método anterior es generalmente criticado por no generar un intervalo centrado en $\hat{p}_{n,orig} = \overline{x}_n$, la estimación puntual obtenida con la muestra original. Para ello, se usa el método de *Bootstrap empírico*. Calcule la secuencia de diferencias de la estimación puntual original con cada muestra bootstrap:

$$\left\{\hat{\delta}_{boot_i} = \hat{p}_{n,boot_i} - \hat{p}_{n,orig}\right\}_{i=1}^{B},$$

luego, calcule

- $\hat{\delta}_{\frac{\alpha}{2}}$ =percentil $\frac{\alpha}{2}$ de la secuencia de valores $\left\{\hat{\delta}_{boot_i}\right\}_{i=1}^{B}$,
- $\hat{\delta}_{1-\frac{\alpha}{2}}$ =percentil $1-\frac{\alpha}{2}$ de la secuencia de valores $\left\{\hat{\delta}_{boot_i}\right\}_{i=1}^{B}$,

para finalmente calcular el intervalo como

$$ICBoot_{100(1-\alpha)\%}(p) = \left[\hat{p}_{n,orig} - \hat{\delta}_{1-\frac{\alpha}{2}}, \hat{p}_{n,orig} - \hat{\delta}_{\frac{\alpha}{2}}\right]$$

Nota: Aunque existe el paquete `boot` que tiene la función `boot.ci`, en este ejercicio deben hacer sus propias implementaciones para generar los intervalos correspondientes.

Para el informe:

1. Muestren el procedimiento que siguieron para deducir la expresión del intervalo en la posibilidad 1.
2. Implementen el siguiente algoritmo:

```
a. Para cada valor de p entre 0.05,0.1,0.15,…,0.85,0.9,0.95:

   b. Para cada tamaño de muestra n entre 5, 10, 50, 100, 200, 500,
1000:

         c. Repetir este procedimiento m=1000 veces:

                  c1. Generen una muestra aleatoria de tamaño n de
                  una Bernoulli con parámetro p.

                  c2. Calculen para dicha muestra los intervalos del
                  95% de confianza obtenidos por cada una de las 5
                  posibilidades[1].

                  c3. Almacenen para cada método (posibilidad) y
                  para cada una de las m simulaciones, un 1 si dicho
                  intervalo contiene al verdadero parámetro (p), 0
                  en otro caso; y en caso de que contenga al
                  verdadero valor del parámetro, guarden la longitud
                  del intervalo (LS-LI).
```

[1] Para las posibilidades 4 y 5, es necesario que utilicen B=1000 muestras para cada una de las posibilidades. Lo idea, sería trabajar con funciones para cada posibilidad, que se llamen en el código principal.

b1. Una vez hecho esto para todas las m repeticiones, resuman para cada posible método de intervalos, la cobertura promedio y la longitud media del intervalo.

$$Cob.prom. = \frac{\text{Cantidad de intervalos que contuvieron al parámetro}}{1000}$$

$$Long.prom. = \frac{\text{Suma de longitud de los int. que contuvieron al parámetro}}{\text{Cantidad de intervalos que contuvieron al parámetro}}$$

Al final, ustedes deben contar con las medidas de cobertura promedio y longitud promedio para cada una de las posibilidades, y para cada combinación de valor del parámetro y del tamaño de muestra.

3. Para cada posibilidad (1-5) hagan dos gráficos. Uno que muestre en el eje horizontal los valores del parámetro p y en el eje vertical las coberturas promedio para cada tamaño de muestra (deben aparecer varias curvas, una para cada tamaño de muestra). El segundo gráfico debe mostrar en el eje vertical, la longitud promedio de los intervalos.

4. Concéntrense ahora en los tamaños de muestra n=5, 50, 200, 1000. Para cada tamaño de muestra, comparen en un mismo gráfico la cobertura promedio de las 5 posibilidades en función del parámetro p en el eje horizontal. Hagan otros cuatro gráficos comparando la longitud promedio de los intervalos de las 5 posibilidades.

5. ¿Qué conclusiones respecto a la efectividad de las diferentes posibilidades en función del tamaño de muestra y del valor del parámetro sacarían? Pueden hacer otros gráficos que ayuden a soportar sus conclusiones.

6. Para cada una de las siguientes dos situaciones, calculen un intervalo de confianza de la proporción de interés usando el método que consideran más adecuado, de acuerdo con la información suministrada.[2]

    a. En un pequeño estudio hecho, se verificó que, de 10 componentes de aire acondicionado testeados, 8 cumplieron con los estándares de producción. ¿Qué podría decirse de la proporción de componentes que cumplen con los estándares en la población con una confianza del 90%?
    b. Se realizó una encuesta virtual a 100 estudiantes de la UNAL-sede Bogotá, seleccionados al azar, con el fin de conocer cómo emplean su tiempo libre y cuáles son sus hobbies favoritos. Se les preguntó cuántas horas al día dedican a actividades ocio y qué tipo de actividades realizan. Con base en esta muestra se obtuvieron los siguientes resultados:

[2] Si el método seleccionado es un método de remuestreo, pueden reconstruir los datos originales creando un vector que tenga la cantidad de ceros y unos necesarios.

Tipo de actividad	Frecuencia absoluta	Frecuencia relativa
Jugar videojuegos	28	28%
Ver televisión	34	34%
Salir con amigos	15	15%
Leer un libro	10	10%
Dormir	7	7%
Otro	6	6%

¿Qué podría decirse de la proporción de estudiantes en la población que prefieren leer un libro? Evalúe este resultado con una confianza del 99%.