UNIVERSIDAD NACIONAL DE COLOMBIA
FACULTAD DE CIENCIAS, SEDE BOGOTÁ
DEPARTAMENTO DE ESTADÍSTICA
INFERENCIA ESTADISTICA
Lineamientos del trabajo final (Parte 1)

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. Distribución de un estimador máximo verosímil cuando no hay condiciones de regularidad.

Considere una muestra aleatoria $X_1, X_2, \ldots, X_n$ de una distribución $U[0, \theta = \kappa]$, es decir, donde el verdadero valor del parámetro es igual al número del grupo. Se sabe que $\hat{\theta}_{MLE} = X^{(n)}$, sin embargo, como no se tienen condiciones de regularidad, este estimador no tiene una distribución asintótica normal (cuando es debidamente normalizado). Vamos a estudiar también la convergencia en distribución de la variable aleatoria $R_n = n\left(\theta - X^{(n)}\right)$.

a. Genere $m = 1000$ simulaciones de una muestra aleatoria de tamaño $n = 10$ de la distribución $U[0, \theta = \kappa]$, usando la función `runif` de R o su equivalente en otro lenguaje de programación. Para cada una de las m muestras, calcule y almacene la estimación máximo-verosímil del parámetro. Haga un histograma de las m estimaciones y superponga sobre este histograma la función de densidad teórica de $X^{(n)}$.
b. Repita el procedimiento descrito en **a.**, variando los tamaños de muestra por los valores $n = 50, 100, 200, 500, 1000$.
c. ¿Qué observa a medida que el tamaño de muestra aumenta? ¿En qué valor parecen concentrarse las realizaciones del estimador máximo-verosímil a medida que el tamaño de muestra aumenta? ¿Qué resultado de los vistos en clase explica ese fenómeno?
d. Genere nuevamente m muestras de tamaños $n = 10, 50, 100, 200, 500, 1000$; y calcule y almacene las m realizaciones de R_n para cada valor de n. Haga histogramas de los m valores obtenidos para cada n. ¿Observa que, a medida que n aumenta, dicho histograma se asemeja a alguna distribución conocida?
e. Pruebe formalmente su conjetura del apartado anterior, calculando el límite en distribución cuando $n \to \infty$:

$$R_n = n\left(\theta - X^{(n)}\right) \xrightarrow{d} ?$$

Nota: El estimador ML de este modelo es súper consistente y por eso su factor de escala no es $\sqrt{n}$ sino n.

B. Cálculo de una estimación máximo-verosímil cuando no hay una solución analítica

Considere una muestra aleatoria $X_1, X_2, \ldots, X_n$ de una distribución $Cauchy(\theta = \kappa, 1)$, es decir, donde el verdadero valor del parámetro de localización es igual al número del grupo. Se sabe que, en este caso, el estimador ML no tiene una expresión analítica, pero es posible igual estudiar las propiedades del estimador máximo-verosímil haciendo uso de la simulación y de métodos numéricos, como la función `optim` de R.

a. Genere $m = 1000$ simulaciones de una muestra aleatoria de tamaño $n = 10$ de la distribución $Cauchy(\theta = \kappa, 1)$, usando la función `rcauchy` de R o su

equivalente en otro lenguaje de programación. Para cada una de las m muestras, calcule y almacene la estimación máximo-verosímil del parámetro y la mediana (como una estimación por analogía). Haga un histograma de las m estimaciones ML y otro de las m estimaciones por analogía. Añada una línea vertical en cada histograma para indicar el verdadero valor del parámetro y otra línea (de diferente color) para indicar el promedio de las m estimaciones en cada caso. Reporte en una tabla el promedio y la varianza obtenidos de las m estimaciones para este tamaño de muestra. ¿Hay indicios de que ambos estimadores sean insesgados? Establezca qué estimador es mejor en términos de los valores de error cuadrático medio (estimado).

b. Repita el procedimiento descrito en **a.**, variando los tamaños de muestra por los valores $n = 50, 100, 200, 500, 1000$. ¿Se mantiene la misma conclusión acerca de qué estimador es mejor en términos de error cuadrático medio (estimado)?

c. ¿A qué convergen en distribución las siguientes secuencias?

$$V_n = \sqrt{n}\left(\hat{\theta}_{MLE} - \theta\right) \xrightarrow{d} ?$$

$$W_n = \sqrt{n}\left(Me - \theta\right) \xrightarrow{d} ?$$

Nota: Va a requerir calcular la información de Fisher, $I(\theta)$, así que muestre cómo la obtuvo.

d. Repitiendo el procedimiento de generar m muestras de tamaño $n = 1000$, calcule m realizaciones de V_{1000} y de W_{1000}, realice un histograma de cada una de ellas y superponga las distribuciones que encontró en el punto anterior. ¿Se ve que el resultado de la convergencia es correcto? ¿Cuál de los dos estimadores es asintóticamente más eficiente?

e. **Bonus [+0.2].** Incluya la Moda como estimador por analogía y compárelo con los otros dos estimadores en los apartados **a.** y **b.** La moda no puede ser estudiada como variable aleatoria de manera analítica ya que no tiene una dependencia funcional clara de la muestra aleatoria. Además, como las variables son continuas, no se puede calcular la moda como "el dato que más se repite", y es necesario usar la fórmula de la moda para datos agrupados. Explore la función `mlv` del paquete `modeest` en R y explore los métodos que allí se proponen (use el método '`naive`' como uno de ellos).

C. Comparación del estimador ML y de momentos en un modelo Doble Exponencial

Considere una muestra aleatoria $X_1, X_2, \ldots, X_n$ de una distribución $Doble\_Exp(\theta = \kappa, 1)$. Recuerde que, en este caso, $\hat{\theta}_{MLE} = Me$ y $\hat{\theta}_{MOM} = \bar{X}_n$.

a. Genere $m = 1000$ simulaciones de una muestra aleatoria de tamaño $n = 10$ de la distribución $Doble\_Exp(\theta = \kappa, 1)$, usando la función `rdexp` del paquete `nimble` de R o su equivalente en otro lenguaje de programación. Para cada una de las m muestras, calcule y almacene la estimación máximo-verosímil del

parámetro y la estimación de momentos. Haga un histograma de las m estimaciones ML y otro de las m estimaciones por momentos. Añada una línea vertical en cada histograma para indicar el verdadero valor del parámetro y otra línea (de diferente color) para indicar el promedio de las m estimaciones en cada caso. ¿Hay indicios de que ambos estimadores sean insesgados? Reporte en una tabla el promedio y la varianza obtenidos de las m estimaciones para este tamaño de muestra. ¿Hay indicios de que ambos estimadores sean insesgados? Establezca qué estimador es mejor en términos de los valores de error cuadrático medio (estimado).

b. Repita el procedimiento descrito en **a.**, variando los tamaños de muestra por los valores $n = 50, 100, 200, 500, 1000$. ¿Se mantiene la misma conclusión acerca de qué estimador es mejor en términos de error cuadrático medio (estimado)?
c. ¿A qué convergen en distribución las siguientes secuencias?

$$Z_n = \sqrt{n}\left(\bar{X}_n - \theta\right) \xrightarrow{d} ?$$

$$M_n = \sqrt{n}\left(Me - \theta\right) \xrightarrow{d} ?$$

d. Repitiendo el procedimiento de generar m muestras de tamaño $n = 1000$, calcule m realizaciones de Z_{1000} y de M_{1000}, realice un histograma de cada una de ellas y superponga las distribuciones que encontró en el punto anterior. ¿Se ve que el resultado de la convergencia es correcto? ¿Cuál de los dos estimadores es asintóticamente más eficiente?
e. **Bonus [+0.2].** Incluya la Moda como estimador por analogía y compárelo con los otros dos estimadores en los apartados **a.** y **b.** La moda no puede ser estudiada como variable aleatoria de manera analítica ya que no tiene una dependencia funcional clara de la muestra aleatoria. Además, como las variables son continuas, no se puede calcular la moda como "el dato que más se repite", y es necesario usar la fórmula de la moda para datos agrupados. Explore la función `mlv` del paquete `modeest` en R y explore los métodos que allí se proponen (use el método '`naive`' como uno de ellos).

D. Técnica de remuestreo (Bootstrap)

En ejercicios de simulación, cuando es necesario calcular algunas propiedades de una variable aleatoria, poder generar muestras de la verdadera distribución de los datos es de gran utilidad para verificar que los resultados teóricos se tienen. Sin embargo, en la vida real, rara vez se conoce la "verdadera" distribución de los datos y solo se cuenta con un conjunto de datos $x_1, x_2, \ldots, x_n$ recogidos.

Ahora bien, uno podría hacer un histograma de los datos, suponer un modelo para ellos a partir de la evidencia del histograma y luego, estimar los correspondientes parámetros. Sin embargo, es posible que la elección de ese modelo "adecuado" no sea tan sencilla y quisiéramos poder contar con un procedimiento que nos permitiera estudiar a una determinada estadística SOLO con la información que tienen los datos, sin necesidad de recurrir a hacer suposiciones sobre su distribución.

a. Genere **<u>una única</u>** muestra simulada de tamaño $n=10$ de la distribución $N\left(\mu=\kappa,\sigma^2=\kappa^2\right)$, usando la función `rnorm` de R o su equivalente en otro lenguaje de programación. Calcule para dicha muestra el promedio y la varianza muestrales. ¿Son cercanas las estimaciones muestrales a los verdaderos parámetros?

b. Haga un gráfico de la función de distribución empírica de los datos (explore la función `ecdf` de R), y superponga la función de distribución real de los datos con los parámetros reales. ¿Se parecen?

c. Ahora bien, vamos a olvidar que sabemos de qué distribución vienen los datos y solo se cuenta con esos $n=10$ datos tomados que llamaremos `data`. Calcule $B=1000$ muestras bootstrap, cada una de tamaño 10, con reemplazo de ese vector de `data`. Explore la función `sample` para tal fin.

d. Para cada una de las muestras de Bootstrap, calcule la media, $\left\{\overline{x}_i\right\}_{i=1}^{B}$; la expresión de la varianza, $\left\{\frac{(n-1)s_i^2}{\hat{\sigma}^2}\right\}_{i=1}^{B}$ ($\hat{\sigma}^2$ es la estimación de la varianza en **a.**); y el coeficiente de variación, $\left\{\frac{s_i}{\overline{x}_i}\right\}_{i=1}^{B}$; obteniendo $B=1000$ realizaciones de cada una. Realice un histograma de cada una y superponga a esos histogramas (en el caso de la media y de la expresión relacionada con la varianza) la función de densidad teórica de cada una.

e. ¿Qué conclusiones puede sacar de los resultados anteriores? ¿Cree que la similitud de las curvas de distribución del ítem **b.** está relacionada con sus hallazgos del ítem **d.**?

f. En relación con la muestra Bootstrap del coeficiente de variación, ¿qué tipo de distribución parece tener el estimador propuesto? ¿Parece ser un estimador insesgado? Si no, ¿de cuánto parece ser su sesgo?

g. Repita los pasos **a.-f.** pero con un tamaño de muestra $n=1000$. ¿Cómo cambian sus respuestas a los incisos anteriores? ¿Por qué? Comente.