Introducción a la Estadística y Ciencia de Datos

Práctica 4 - Bootstrap

---

1. Sean $X_1, \ldots, X_n \sim F$ variables aleatorias independientes. Sea $X \sim X_i$ y supongamos que queremos estimar a $\mu = \mathbb{E}_F(X)$ asumiendo sólo que μ existe. La media μ puede pensarse como una función de F: en efecto, podemos escribir $\mu = T(F) = E_F(X)$.

   En general, cualquier función de la distribución F recibe el nombre de *funcional estadístico*. Otros ejemplos son la varianza $\sigma^2 = T(F) = E_F(X^2) - (E_F(X))^2$ y la mediana $m = T(F) = F^{-1}(1/2)$.

   El **estimador plug–in** de $\theta = T(F)$ se define como

   $$\widehat{\theta}_n = T(F_n)\,,$$

   donde F_n es la distribución empírica, es decir $F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(X_i \leq x)$.

   a) Hallar el estimador plug–in de μ

   b) Hallar el estimador plug–in de σ^2

   c) Si $T(F) = \mathbb{E}_F(r(X))$ para alguna función r, hallar el estimador plug–in de $T(F)$.

   d) Si $A \subset \mathbb{R}$, podríamos estimar $\mathbb{P}_F(X \in A)$, ¿cuánto resulta $\widehat{\mathbb{P}}_F(X \in A) = \mathbb{P}_{F_n}(X \in A)$? (Sug.: aplicar c) con la función indicadora apropiada).

2. La cantidad de partículas que emite una fuente radiactiva en un minuto es una variable aleatoria con distribución de Poisson de parámetro λ. Se midió la cantidad de emisiones en intervalos de 1 minuto, obteniendo las siguientes 15 muestras

   $$500 - 488 - 426 - 510 - 450 - 368 - 508 - 514 - 426$$
   $$476 - 512 - 526 - 444 - 524 - 236$$

   a) A partir del Ejercicio 1 de la Práctica 2 estimar el valor del parámetro λ basado en la muestra utilizando un estimador de momentos

      1) basado en el momento de orden 1, que llamaremos T_1.
      2) basado en el momento de orden 2, que llamaremos T_2.

   b) Notar que el cálculo de la varianza del estimador basado en el momento de orden 2 es difícil y que en este caso el *método bootstrap* puede auxiliarnos. Dado que hemos asumido que los datos provienen de una muestra Poisson, podemos apelar al *método bootstrap paramétrico*.

      1) Explorar el comando `rpois` para obtener una muestra de tamaño 15 generada a partir del estimador $T_{1,obs}$, computar el estimador bootstrapeado y guardar el valor obtenido.
      2) Repetir $B = 1000$ veces el ítem anterior guardando los valores bootstrapeados de T_1 en un vector.
      3) Realizar un histograma basado en los valores bootstrapeados de T_1 del ítem anterior. ¿Qué observa?

4) A partir de los valores bootstrapeados de T_1 estimar la varianza de T_1. ¿Tiene alguna intuición de cuanto valdría esta varianza?

c) Repetir b) para T_2 y comparar los resultados obtenidos para los dos estimadores.

3. A partir del archivo *datos1.txt*, estimar el valor central con las siguientes medidas de posición: (a) Media (b) Mediana. Para estimar el desvío estándar de ambos estimadores, se propone el siguiente esquema de *bootstrap*:

   a) Generar una muestra del mismo tamaño muestral que datos1.txt formada por elementos tomados de *datos1.txt* elegidos al azar con reposición.

   b) Calcular la medida de posición en dicha muestra.

   c) Repetir el procedimiento $B = 1000$ veces o más y obtener 1000 medidas.

   d) Estimar el desvío estándar a partir de las medidas obtenidas en el ítem anterior. Comparar los resultados obtenidos.

   Intervalos Bootstrap

   Es muy común querer calcular un intervalo de confianza para un cierto parámetro θ basándonos en un estimador $\widehat{\theta}$. El problema es que muchas veces la distribución (exacta o asintótica) de $\widehat{\theta}$ no es conocida o contiene parámetros que desconocemos. Vamos a considerar los siguientes dos métodos:

   - **Método 1. (Método Bootstrap Normal)** Supongamos que $\widehat{\theta}$ tiene distribución aproximadamente normal con media θ. Luego, un intervalo de nivel aproximado $1-\alpha$ es
   $$[\widehat{\theta} - z_{\alpha/2}\, \widehat{\mathtt{se}}(\widehat{\theta}),\ \widehat{\theta} + z_{\alpha/2}\widehat{\mathtt{se}}(\widehat{\theta})],$$
   donde $\widehat{\mathtt{se}}(\widehat{\theta})$ (se son las siglas de *standard error*) es un estimador del desvío estándar de $\widehat{\theta}$. Por otra parte, el estimador $\widehat{\mathtt{se}}(\widehat{\theta})$ puede obtenerse con Bootstrap, es decir, considerando realizaciones $\widehat{\theta}^*_1, \ldots, \widehat{\theta}^*_B$ del estimador basadas en remuestreos de la distribución empírica y tomando $\widehat{\mathtt{se}}(\widehat{\theta}) = \mathtt{sd}(\widehat{\theta}^*_1, \ldots, \widehat{\theta}^*_B)$, el desvío estándar muestral.
   - **Método 2. (Método Bootstrap Percentil)** Se consideran realizaciones $\widehat{\theta}^*_1, \ldots, \widehat{\theta}^*_B$ del estimador $\widehat{\theta}$ obtenidos haciendo Bootstrap como en el método anterior. Luego, un posible intervalo de nivel aproximado $1-\alpha$ es
   $$[\widehat{\theta}^*_{(\alpha/2)}, \widehat{\theta}^*_{(1-\alpha/2)}],$$
   donde $\widehat{\theta}^*_{(\gamma)}$ es el $\gamma-$percentil de la muestra $\widehat{\theta}^*_1, \ldots, \widehat{\theta}^*_B$.

4. (Para hacer con el R) Tenemos un dado que no sabemos si es o no equilibrado. Para estudiarlo, lo tiramos $n = 100$ veces de manera independiente, obteniendo los siguientes valores, que se pueden cargar fácilmente al R con la instrucción `scan`.

```
2 2 4 6 1 3 1 3 2 4 4 4 4 4 6 3 3 4 1 2 1 6 3 2 3 4 1 1 5 4 1 4 6
4 1 2 1 5 4 3 3 1 3 1 6 5 1 3 2 3 6 2 4 2 6 6 5 2 4 4 1 4 3 1 2 1
6 1 1 3 1 6 6 1 2 6 1 1 4 5 4 1 5 2 2 1 6 6 1 2 1 3 1 3 3 4 3 3 3 5
```

¿Cómo podemos comprobar si el dado es equilibrado? Haremos dos pruebas sobre los resultados obtenidos.

a) Queremos estudiar la probabilidad de que un resultado del dado sea par. Llamemos θ a dicha probabilidad. ¿Cuánto debería valer θ si el dado fuera equilibrado?

1) Tome una "*muestra bootstrap*" seleccionando con reemplazo $n = 100$ (sí, la misma cantidad que el tamaño de la muestra original) de los valores observados. El comando `sample` puede ser útil. Estime la probabilidad buscada a partir de la muestra bootstrapeada. [1]
2) Repita el paso anterior $B = 5000$ veces, guardando las proporciones bootstrapeadas en un vector `titas.boot`, $\widehat{\theta}^*_1, \ldots, \widehat{\theta}^*_B$.
3) Realizar un histograma de las proporciones bootstrapeadas. Observar que este gráfico es una aproximación del gráfico de la función de densidad de $\widehat{\theta}$, $f_{\widehat{\theta}}$. ¿Tiene forma aproximadamente acampanada?
4) A partir de las proporciones bootstrapeadas calcular un estimador del standard error de $\widehat{\theta}$.
5) Armar dos intervalos de confianza para θ basados en las muestras bootstrap de nivel $1 - \alpha = 0{,}95$ según los dos métodos descriptos más arriba. A partir de lo obtenido, ¿contradice esto el supuesto de que el dado es equilibrado?

b) Estudiemos ahora la probabilidad de que un resultado del dado sea exactamente igual a 5. Llamemos θ a dicha probabilidad. ¿Cuánto debería valer θ si el dado fuera equilibrado? Repetir el análisis anterior de los ítems (i)–(v) y concluir.

5. (Para hacer con el R) Instalar el paquete `Lock5withR` correspondiente al libro de Lock [2]. En dicho paquete hay un conjunto de $n = 25$ datos que se denomina `MustangPrice`. Cargarlo al R con la instrucción `data("MustangPrice")`.

   Los datos corresponden a 25 automóviles Mustang usados que se ofrecen a la venta en un sitio de Internet en determinado momento. Se registraron las variables: *Age*, edad (en años), *Miles*, kilometraje (en millas) y *Price*, precio de venta (en miles de dólares) para cada automóvil de la muestra.

   Nos enfocaremos en la correlación entre las variables *Price* y *Miles*. Recordemos que si (X, Y) es un vector aleatorio con distribución conjunta F, la correlación entre X e Y es

$$\rho = \frac{\mathbb{C}ov(X, Y)}{\sqrt{\mathbb{V}(X)\mathbb{V}(Y)}}\,.$$

   Además, dada una muestra de vectores aleatorios $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ podemos estimar a ρ mediante la correlación muestral definida por

$$\widehat{\rho} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{S_X^2\ S_Y^2}}\,,$$

[1] Observe que esto equivale a seleccionar una muestra de los números 1 a 6 de modo que la probabilidad de elegir a cada uno sea la proporción observada en la muestra original en vez de ser 1/6 cada uno.

[2] Lock, R. H., Lock, P. F., Lock Morgan, K., Lock, E. F., Lock, D. F. (2021) *Statistics: Unlocking the Power of Data*, 3rd Edition, Wiley ed.

que en R se obtiene mediante el comando `cor`.

a) Relizar un plot de *Price* vs. *Miles*. Estimar la correlación entre *Price* y *Miles* usando los datos originales.

b) Describir un mecanismo que permita computar un estimador bootstrapeado de la correlación entre *Price* y *Miles*.

c) Generar $B = 5000$ muestras bootstrap y computar para cada una de ellas el estimador bootstrapeado de la correlación. Con dichos valores realizar un histograma, ¿qué revela este gráfico?

d) Usar el ítem anterior para construir un intervalo bootstrap percentil de nivel 0.95 % para la correlación entre *Price* y *Miles*. Interpretar el intervalo obtenido.

6. (Para resolver en R) El archivo `eph21_3T.csv` tiene datos de la EPH, encuesta permanente de hogares, realizada en el tercer trimestre de 2021. [3] El archivo tiene información sobre 78217 personas encuestadas, esta encuesta incluye información de todos los habitantes de un hogar. Las tres columnas del archivo son:

   - `ïngresototal"`: ingreso total de las personas encuestadas
   - `"mujer"`: variable indicadora, vale 1 en el caso de que la persona encuestada sea mujer, 0 sino
   - `ïngresototal.conruido"`: es la variable ingreso total a la que se le sumó un pequeño ruido aleatorio para evitar el redondeo producido al reportar ingresos, y que la variable sea efectivamente continua. Esta es la variable que utilizaremos para el análisis.

   De la base de datos que figura en la página del INDEC excluimos (sin mayores discusiones) a los encuestados que reportaron ingresos nulos o que no reportaron su ingreso, esto da un total de 42348 personas. Queremos estudiar si el ingreso mediano poblacional de varones y mujeres es diferente. Como no conocemos la distribución de la variable ingreso, buscaremos hacerlo mediante bootstrap.

   a) Hallar una estimación de la mediana de los ingresos totales (con ruido) de las mujeres y de la de los hombres en la base de datos. Calcular la diferencia entre ellas.

   b) Tomar una muestra bootstrap de los ingresos totales (con ruido) de las mujeres, y tomar la mediana de dicha muestra. Elegir una muestra bootstrap de los ingresos totales (con ruido) de los hombres y calcular la mediana. Luego guardar la diferencia entre ambas medianas bootstrap.

   c) Repetir el análisis del ítem anterior una cantidad $B = 5000$ veces, guardando en un vector las diferencias de las medianas estimadas. A partir de ellas, graficar un histograma y hallar una estimación del error standard de la diferencia de medianas muestrales.

[3] Los datos fueron bajados de la página del INDEC, `https://www.indec.gob.ar/indec/web/Institucional-Indec-BasesDeDatos`

d) Hallar intervalos de confianza de nivel 0.95 para la diferencia de medianas de los salarios de mujeres y hombres basados en las muestras bootstrap con los dos métodos vistos en clase. En base a estos datos, ¿parecen diferentes los ingresos medianos de mujeres y hombres en la población?

7. Consideremos nuevamente los datos `MustangPrice` y en particular, la variable *Price*.

   a) Tomar una muestra bootstrap de *Price* y calcular la mediana muestral. Repetir este procedimento $B = 5000$ veces y guardar los resultados en un vector.

   b) Relizar un histograma con las medianas bootstrapeadas obtenidas en el ítem anterior. A partir de este gráfico explicar por qué no sería apropiado usar ninguna de las dos estrategias basadas en bootstrap para construir un intervalo de confianza para la mediana.

   c) ¿Se podrá usar bootstrap si en lugar de interesarnos la mediana nos interesa la media?

8. *a*) Se quiere construir intervalos de confianza para la mediana de una variable aleatoria X usando a la mediana muestral como estimador. Para eso, implementar dos funciones `boot.metodo.1` y `boot.metodo.2` que tengan como entrada el nivel $1 - \alpha$, B y datos $x_1, \ldots, x_n$ y devuelvan el intervalo de confianza estimado siguiendo la estrategia del método bootstrap normal y del método bootstrap percentil basados en B muestras bootstrap, respectivamente.

   b) Dado $n = 30$ generar una muestra de tamaño n de una variable con distribución $\mathcal{N}(0, 1)$ y obtener un intervalo de confianza de nivel 0,95 con el método 1 usando $B = 2000$.

   c) Repetir $Nrep = 1000$ veces el ítem anterior y completar la siguiente tabla.

   d) Repetir los dos ítems anteriores para cada $n = 50, 100, 1000$.

   e) Hacer lo mismo para el método 2 y comparar ambos métodos.

	$n = 30$	$n = 50$	$n = 100$	$n = 1000$
Longitud promedio				
Cubrimiento empírico				