Estadística (M) - Segundo cuatrimestre 2019

Práctica 3 - Intervalos de Confianza y Bootstrap

Se recomienda empezar por los ejercicios marcados sin †.

A lo largo de esta práctica puede usar los siguientes resultados sin probarlos.

- Si $X \sim N(\mu, \sigma^2)$, entonces $(X-\mu)/\sigma \sim N(0,1)$.
- Si $X \sim N(0,1)$, entonces $X^2 \sim \chi_1^2$. Además, si $Y \sim \chi_n^2$ y $Z \sim \chi_m^2$ son independientes, entonces $Y+Z \sim \chi_{n+m}^2$.
- Si $X \sim N(0,1)$ y $Y \sim \chi_n$ son independientes, entonces $\frac{X}{\sqrt{Y/n}} \sim t_n$.
- Si $X_1, \ldots, X_n$ son i.i.d. con distribución $N(\mu, \sigma^2)$, entonces $\overline{X}$ y $S_X^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X})^2$ son independientes. Además, $\frac{n-1}{\sigma^2} S_X^2 \sim \chi_{n-1}^2$.

A) *Intervalos de confianza de nivel exacto*

1. Se trata de medir el período de un péndulo y se tiene un cronómetro de precisión conocida (es decir, se conoce la varianza del error). Se supone que las observaciones son de la forma $Y_i = \mu + \varepsilon_i$, donde los ε_i tienen distribución $N(0, 1/4)$ y son independientes. Los datos obtenidos son:

   5.1 5.2 5.6 5.1 5.5 5.8 5.9 4.9 5.2 5.6

   (Observar que la media muestral es 5.39 y la varianza muestral es 0.11)

   (a) Encontrar un intervalo de confianza para μ de nivel 0.95 y probar que para esta muestra la estimación por intervalo (es decir, el intervalo de confianza observado) es [5.08, 5.70].

   (b) ¿Podría haber anticipado cuál era la longitud de dicho intervalo antes de conocer la muestra?

   (c) ¿Cuál debería haber sido el tamaño de la muestra si se hubiera deseado que la longitud del intervalo fuese a lo sumo 0.10?

   (d) Responder verdadero o falso y justificar: si tomamos una nueva observación, tendrá probabilidad 0.95 de pertenecer al intervalo [5.08, 5.70].

2. Sea $X_1, \ldots, X_n$ una muestra aleatoria de una población con distribución $N(\mu, \sigma_0^2)$ (σ_0^2 conocido). Supóngase que se pide un intervalo de confianza para μ de nivel $1-\alpha$. Mostrar que al elegir $A = z_{\frac{\alpha}{2}}$ y $B = -z_{\frac{\alpha}{2}}$ se obtiene el intervalo de longitud mínima entre los contruidos considerando $A = z_\beta$ y $B = z_{1-\delta}$, con $\beta + \delta = \alpha$.

3. La distribución del índice de colesterol en cierta población es $N(\mu, \sigma^2)$, donde μ y σ^2 son desconocidos. Se hacen análisis a 25 personas elegidas al azar entre esta población y se obtienen los siguientes valores:

   1.52 1.65 1.72 1.65 1.72 1.83 1.62 1.75 1.72 1.68 1.51 1.65 1.58
   1.65 1.61 1.70 1.60 1.73 1.61 1.52 1.81 1.72 1.50 1.82 1.65

   (a) Encontrar un intervalo de confianza para μ de nivel 0.95. ¿Es cierto que su longitud no depende de la muestra?

   † (b) ¿Cuántos datos adicionales deben obtenerse para poder construir un intervalo de confianza para μ de nivel 0.95, pero con una longitud no mayor que 0.05? ¿Qué forma tendría este intervalo?

(c) Encontrar un intervalo de confianza para σ de nivel 0.90.

(d) Encontrar un intervalo de confianza para $\exp(-\mu)$ de nivel 0.95.

4. (a) Probar que si X tiene distribución $\varepsilon(\lambda)$, entonces $Y = 2\lambda X$ tiene distribución χ^2_2.

   (b) Sean $X_1, \ldots, X_n$ una muestra aleatoria de una población con distribución $\varepsilon(\lambda)$. Mostrar que $T = 2\lambda \sum_{i=1}^n X_i$ tiene distribución χ^2_{2n}.

   (c) En base a *b*) hallar una familia de intervalos de confianza para λ de nivel $1-\alpha$.

5. Sean $X_1, \ldots, X_n$ una muestra aleatoria con distribución $U[0,\theta]$ y sea $T = \max(X_1, \ldots, X_n)$.

   (a) Mostrar que T/θ tiene distribución independiente de θ.

   (b) Usando *a*) hallar un intervalo de confianza de longitud mínima para θ de nivel $1-\alpha$.

† 6. Se desea comparar los rendimientos de dos variedades de trigo A y B. Se han cultivado 15 parcelas elegidas al azar con la variedad A y 20 con la variedad B, obteniéndose los siguientes rendimientos por hectárea:

Var. A:	250	252	245	258	240	247	251	249	250	243	247	260	238	241	239
Var. B:	330	335	327	329	320	332	337	328	334	326	331	332	328	329	337
	341	336	338	325	321										

Se supone que los rendimientos de la variedad A tienen distribución $N(\mu_1, \sigma^2)$ y los de la otra variedad $N(\mu_2, \sigma^2)$, independientes entre sí. Notar que ambas muestras tienen la misma varianza (desconocida).

(a) Hallar un intervalo de confianza para $\mu_1 - \mu_2$ de nivel 0.99.

(b) ¿Qué le sugeriría el hecho de que el 0 no pertenezca al intervalo hallado? ¿Qué pensaría en caso contrario?

Sugerencia: Probar que si $X_1, \ldots, X_n$ son i.i.d. con distribución $N(\mu_1, \sigma^2)$ y $Y_1, \ldots, Y_m$ son i.i.d. con distribución $N(\mu_2, \sigma^2)$ (también independientes de las X_i), entonces

$$\frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n} + \frac{1}{m}}} \sim N(0,1) \qquad \text{y} \qquad \frac{(n-1)S_1^2 + (m-1)S_2^2}{\sigma^2} \sim \chi^2_{n+m-2}, \tag{1}$$

donde $S_1^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X})^2$ y $S_2^2 = \frac{1}{m_1}\sum_{i=1}^m (Y_i - \overline{Y})^2$. Probar también que ambas cantidades en (**??**) son independientes y deducir que

$$\sqrt{\frac{nm}{n+m}}\,\frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{S_P} \sim t_{n+m-2},$$

donde $S_P^2 = ((n-1)S_1^2 + (m-1)S_2^2)/(n+m-2)$ es la varianza muestral "ponderada".

† 7. Se tienen dos variedades de trigo A y B. Se eligen al azar 15 parcelas, y cada una de ellas se divide en dos partes iguales. En una parte se cultiva la variedad A y en la otra la B. Se obtienen así 15 pares de datos:

Parcela:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Var. A:	41	37	36	39	44	42	38	37	35	32	39	30	40	41	37
Var. B:	39	35.3	33.5	36	42.5	38	36	34.8	33.2	29	29	36.6	28.4	38.5	39

Sea X_i el rendimiento de la variedad A en la parcela i e Y_i el rendimiento de la variedad B en la misma parcela. Se supone que (X_i, Y_i), $1 \le i \le 15$, es una muestra de una distribución normal bivariada con parámetros desconocidos. Hallar un intervalo de confianza para $\mu_X - \mu_Y$ de nivel 0.95.

B) _Intervalos de confianza de nivel asintótico_

1. (a) Dada una muestra aleatoria de una distribución $Bi\,(1,p)$, construir un intervalo de confianza de nivel asintótico $1-\alpha$.

   (b) Una droga cura cierta enfermedad con probabilidad p. En una prueba con 100 enfermos, se curaron 30.

   i. Hallar un intervalo de confianza para p de nivel asintótico 0.95.

   ii. ¿Qué tamaño de muestra debería tomarse si se desea una longitud menor que 0.1?

2. (a) Sea $X_1,\ldots,X_n$ una muestra aleatoria de una distribución $\mathcal{P}\,(\lambda)$. Hallar un intervalo de confianza para λ de nivel asintótico $1-\alpha$.

   (b) El número de llamadas diarias a una central telefónica sigue un proceso de Poisson con media λ. Se ha registrado el número de llamadas durante 20 días, obteniéndose los siguientes valores:

35	41	38	40	34	36	41	48	42	46
39	37	41	35	37	38	42	43	44	67

   Hallar un intervalo de confianza para λ de nivel asintótico 0.90.

† 3. (a) Supóngase que $X_1,\ldots,X_{n_1}$ es una m.a. de una distribución $N\,(\mu_1,\sigma_1^2)$ y que $Y_1,\ldots,Y_{n_2}$ es una m.a. de una distribución $N\,(\mu_2,\sigma_2^2)$, independiente de la anterior. Mostrar que

$$\frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{\sqrt{\dfrac{s_1^2}{n_1}+\dfrac{s_2^2}{n_2}}} \xrightarrow{\mathcal{D}} N\,(0,1)$$

   cuando $n_1\to\infty$ y $n_2\to\infty$ de modo que $n_1/n_2\to\lambda$ constante.

   (b) Hallar un intervalo de confianza de nivel asintótico $(1-\alpha)$ para $\mu_1-\mu_2$.

† 4. (Para hacer con el R) Realizar el siguiente estudio de simulación. Generar una muestra aleatoria de variables $X_1,\ldots,X_n\sim Bi(1,p)$. Para esta muestra, construir los dos intervalos de confianza para p de nivel asintótico 0.95 vistos en clase. Repetir todo este procedimiento k veces, obteniendo así k muestras independientes y k intervalos para cada uno de los dos métodos.

   - Método 1: de nivel asintótico cuando se sustituye el valor de p en la varianza por $\bar{X}$.
   - Método 2: de nivel asintótico cuando no se sustituye a p y se calcula los extremos del intervalo como raíces de una cuadrática.

   Tomar $k=2000$ y los siguientes valores de n y p.

   - $n=20;50;100$.
   - $p=0.10;0.50$.

   Para cada muestra, guardar los siguientes resultados: el intervalo de confianza obtenido, la longitud de dicho intervalo, un 1 si el IC hallado contiene al verdadero valor de p y un 0 en caso contrario.

   Para cada combinación de n y p:

   (a) Estimar la longitud esperada para ambos métodos.

   (b) Estimar la probabilidad de cobertura para ambos métodos.

   (c) Sacar conclusiones.

C) <u>*Intervalos Bootstrap*</u>

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.** Supongamos que $\widehat{\theta}$ tiene distribución aproximadamente normal con media θ. Luego, un intervalo de nivel aproximado $1-\alpha$ es

$$[\widehat{\theta} - z_{\alpha/2}\sqrt{v^2}, \widehat{\theta} + z_{\alpha/2}\sqrt{v^2}],$$

  donde v^2 es un estimador de la varianza de $\widehat{\theta}$. Por otra parte, el estimador v^2 puede obtenerse con Bootstrap, es decir, considerando realizaciones $\widehat{\theta}^*_1, \ldots, \widehat{\theta}^*_{n_{boot}}$ del estimador basadas en remuestreos de la distribución empírica y tomando $v^2 = \textsc{Var}(\widehat{\theta}^*_1, \ldots, \widehat{\theta}^*_{n_{boot}})$.

- **Método 2.** Se consideran realizaciones $\widehat{\theta}^*_1, \ldots, \widehat{\theta}^*_{n_{boot}}$ 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 $\widehat{\theta}^*_1, \ldots, \widehat{\theta}^*_{n_{boot}}$.

1. (Para hacer con el R) **(Intervalos de confianza bootstrap para la mediana)**

   (a) Se quiere hacer intervalos de confianza para la mediana m de una distribución X usando a la mediana muestral como estimador. Para eso, implementar dos funciones `boot.metodo.1` y `boot.metodo.2` que tengan como input el nivel $1-\alpha$ y datos $x_1, \ldots, x_n$ y devuelvan el intervalo de confianza siguiendo la estrategia del método correspondiente.

   (b) Para cada $n = 30, 50, 100, 1000$, generar 1000 muestras de una distribución $N(0,1)$ y para cada una de ellas obtener los intervalos de nivel 0.95 con el método 1. Luego, para cada n completar la siguiente tabla Hacer lo mismo para el método 2 y comparar ambos métodos.

	n=30	n=50	n=100	n=1000
Longitud promedio				
Proporción de cubrimiento				

   (c) Repetir el item anterior para muestras de distribución Cauchy (también llamada t_1). Esta distribución es simétrica respecto del 0 y no tiene esperanza finita.

† 2. (Para hacer con el R) Modificando levemente el código del ejercicio anterior, obtener intervalos de confianza de nivel 0.95 para los percentiles 0.25, 0.5 y 0.75 del precio del ticket del Titanic (usar los datos que están subidos a la página de la materia).