Introducción a la Estadística y Ciencia de Datos

Práctica 3 - Intervalos de confianza

---

1. Para medir el período de un péndulo 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 son variables aleatorias con distribución $N(0, 1/4)$ y son independientes. Se realizan 10 mediciones obteniendo los siguientes datos:

   5,1 5,2 5,6 5,1 5,5 5,8 5,9 4,9 5,2 5,6

   a) Encontrar un intervalo de confianza de nivel 0,95 para μ y calcular la estimación por intervalos con los datos brindados.

   b) ¿Se podría haber anticipado la longitud de dicho intervalo antes de conocer la muestra?

   c) ¿Cuál debe ser el tamaño de la muestra si se desea que la longitud del intervalo fuese a lo sumo 0,10?

   d) Responder verdadero o falso y justificar: "si tomamos una nueva observación, la verdadera media 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). Se busca un 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 contruídos considerando $A = z_\beta$ y $B = z_{1-\delta}$, con $\beta + \delta = \alpha$.

3. La distribución del índice de colesterol en cierta población es una variable aleatoria con distribución $N(\mu, \sigma^2)$, 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. La longitud de dicho intervalo ¿depende de la muestra? Calcular la estimación por intervalos con los datos brindados.

   b) Encontrar un intervalo de confianza para σ de nivel 0,90 y calcular la estimación por intervalos con los datos brindados.

   c) Encontrar un intervalo de confianza para $e^{-\mu}$ de nivel 0,95 y calcular la estimación por intervalos con los datos brindados.

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 un intervalo de confianza para λ de nivel $1-\alpha$, basado en la muestra de tamaño n.

d) Se sabe que el tiempo de duración de cierto tipo de lámparas tiene distribución $\varepsilon(\lambda)$. Se han probado 20 lámparas y los tiempos de duración de los mismos (en días) fueron los siguientes

25	45	50	61	39	40	45	47	38	39
54	60	39	46	39	50	42	50	62	50

Calcular la estimación por intervalos para λ de nivel 0,99 con los datos brindados.

5. Sean $X_1, \ldots, X_n$ una muestra aleatoria con distribución $\mathcal{U}(0, \theta)$ y sea $T = \text{máx}(X_1, \ldots, X_n)$.

a) Mostrar que $W = T/\theta$ tiene una distribución que no depende de θ.

b) Usando *a*) hallar un intervalo de confianza para θ de nivel $1 - \alpha$ basado en la muestra aleatoria de tamaño n.

6. *a*) Dada una muestra aleatoria de tamaño n de una población con distribución $\mathcal{B}(1, p)$, construir un intervalo de confianza de nivel asintótico $1 - \alpha$ para p.

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

1) Calcular una estimación por intervalos para p de nivel asintótico 0,95 con los datos brindados.
2) ¿Qué tamaño de muestra debería tomarse si se desea una longitud menor a 0,1?

7. El 50 % de los bits emitidos por un canal de comunicación binario son 1. El receptor indica que hay un 1 cuando efectivamente se ha enviado un 1 con probabilidad p e indica que hay un 0 cuando efectivamente se ha enviado un 0 con probabilidad 0,6. ¿Cuántos bits deberán emitirse para que sea posible construir un intervalo de confianza de nivel asintótico 0,95 para p cuya longitud sea menor que 0,01? *Sugerencia*: considerar la variable aleatoria que es 1 cuando se recibió un bit con un 1 y 0 cuando se recibió un 0.

8. *a*) Sea $X_1, \ldots, X_n$ una muestra aleatoria de una población con 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 es una variable aleatoria con distribución de Poisson de media λ. Se ha registrado el número de llamadas durante 50 días, obteniéndo una cantidad total de 1761 llamadas. Calcular la estimación por intervalos para λ de nivel asintótico 0,90 con los datos brindados.

9. Una sustancia radiactiva emite partículas alfa de acuerdo con un proceso de Poisson de intensidad λ por segundo.

a) Se la observó durante 50 segundos y se registraron 4 emisiones. En base a esta información muestral construir un intervalo de confianza de nivel asintótico 0,95 para λ.

b) Se volvió a observar la misma sustancia radiactiva hasta que emitió la cuarta partícula alfa, lo que sucedió a los 50 segundos. En base a esta información muestral construir un intervalo de confianza de nivel 0,95 para λ.

10. *(Para hacer en R)* Generar una muestra aleatoria de una variable con distribución $\mathcal{B}(1,p)$ de tamanño n. Para esta muestra, calcular las dos estimaciones por intervalos para p de nivel asintótico 0,95, correspondientes a los dos métodos que se mencionan a continuación. Repetir todo este procedimiento k veces, obteniendo así k muestras independientes de tamaño n 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: la estimación por intervalos obtenida, 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 con ambos métodos.

    b) Calcular el cubrimiento empírico con ambos métodos.

11. 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 el rendimiento es una variable aleatoria con distribución $N(\mu_1, \sigma^2)$ para la variedad A y $N(\mu_2, \sigma^2)$ para la variedad B, independientes entre sí.

    a) Hallar un intervalo de confianza para $\mu_1 - \mu_2$ de nivel 0,99 y calcular la estimación por intervalos con los datos brindados.

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

    c) Hallar un intervalo de confianza para σ^2 de nivel 0,99 y calcular la estimación por intervalos con los datos brindados.

    d) Hallar un intervalo de confianza para σ de nivel 0,90 y calcular la estimación por intervalos con los datos brindados.

12. 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), $1 \leq i \leq 15$, es una muestra aleatoria de una población con distribución normal bivariada de parámetros desconocidos. Hallar un intervalo de confianza para $\mu_X - \mu_Y$ de nivel 0,95 y calcular la estimación por intervalos con los datos brindados.

13. *Cuidado, es más difícil de lo que parece.*

    a) *Opcional.* Sea $X_1, \ldots, X_{n_1}$ una muestra aleatoria de una población con distribución $N(\mu_1, \sigma_1^2)$ e $Y_1, \ldots, Y_{n_2}$ una muestra aleatoria de una población con 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 $n_1/n_2 \to c$ constante.

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

14. Una muestra aleatoria de 200 automovilistas de cierta localidad y que usan automóviles extranjeros indicó que 115 de ellos usan regularmente sus cinturones de seguridad, en tanto que otra muestra de 300 que usan automóviles de fabricación nacional indicó que 154 usan regularmente sus cinturones de seguridad. Hallar un intervalo de confianza de nivel asintótico 0.99 para la diferencia de las proporciones de quienes usan regularmente sus cinturones de seguridad entre los automóviles extranjeros y los de fabricación nacional, y luego realizar una estimación por intervalos con los datos brindados.

15. Sea $\mathbf{X} = (X_1, \ldots, X_n)$ una muestra aleatoria de una población con densidad $\beta(\theta, 1)$, o sea de la forma

    $$f_\theta(x) = \theta\, x^{\theta-1}\, \mathbb{I}_{(0,1)}(x), \qquad \theta > 0.$$

    Construir un intervalo de confianza de nivel asintótico $1 - \alpha$ para θ.

    Sugerencia: Revisar ejercicio 17 de la Práctica 2 y probar que $\dfrac{\overline{X}}{\left(2 - \overline{X}\right)\left(1 - \overline{X}\right)^2} \xrightarrow{P} \dfrac{\theta\,(\theta+1)^2}{\theta+2}$.

16. (Cotas de confianza o intervalos de confianza unilaterales)
    Definición: Sean $X_1, \ldots, X_n$ variables aleatorias iid con distribución F_θ con $\theta \in \Theta \subset \mathbb{R}$ y $\alpha \in (0,1)$. Sea $L(X_1, \ldots, X_n)$ un estadístico tal que

    $$P_\theta\left(L(X_1, \ldots, X_n) \leq \theta\right) = 1 - \alpha, \qquad \text{para todo } \theta \in \Theta.$$

    Entonces, decimos que la región $[L(X_1, \ldots, X_n), +\infty)$ es un intervalo de confianza unilateral a izquierda para θ de nivel $1 - \alpha$ o una cota inferior de confianza. Análogamente, si

    $$P_\theta\left(U(X_1, \ldots, X_n) \geq \theta\right) = 1 - \alpha, \qquad \text{para todo } \theta \in \Theta\,,$$

    decimos que la región $(-\infty, U(X_1, \ldots, X_n)]$ es un intervalo de confianza unilateral a derecha para θ de nivel $1 - \alpha$ o una cota superior de confianza.

a) Asumiendo que $X_1, \ldots, X_n$ son v.a.i.i.d. con distribución normal con varianza conocida, escribir los dos intervalos de confianza unilaterales de nivel 0.95 para la media.

b) Según la reglamentación, el agua potable tiene como máximo una concentración media de arsénico de 0.01mg/l. Se quiere saber si la concentración media de arsénico en el agua de una localidad la hace potable. Para ello se mide la concentración de arsénico en $n = 10$ muestras de agua de dicha localidad extraídas de forma independiente. Asumiendo que la distribución de las determinaciones de la cantidad de arsénico pueden suponerse normales con varianza conocida σ_0 hallar una región de confianza para la concentración media de arsénico en el agua de dicha localidad de nivel 0.95. ¿Tiene sentido construir un intervalo de confianza unilateral a izquierda, a derecha o uno bilateral para dar respuesta a este problema?

17. La longitud en metros de cada rollo de tela en un lote es una variable aleatoria con distribución uniforme sobre el intervalo $[15, 15 + \theta]$. Se examinaron 4 rollos y la máxima longitud observada resultó ser 25 metros. En base a la información muestral construir una cota superior de confianza de nivel 0.99 para θ.

18. (Bonferroni)

a) Sean $A_1, \ldots, A_k$ eventos de un espacio muestral. Probar la Desigualdad de Bonferroni:

$$P\left(\bigcap_{i=1}^{k} A_i\right) \geq 1 - \sum_{i=1}^{k} P(A_i^c)$$

Definición (Región de confianza de nivel simultáneo): Se puede extender la noción de intervalo de confianza de nivel $1-\alpha$ para un parámetro unidimensional θ a la noción de *región de confianza de nivel simultáneo* $1-\alpha$ *para un vector* $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k) \in \mathbb{R}^k$. Diremos que la región aleatoria $R(\mathbf{X}) \subset \mathbb{R}^k$ es una región de confianza de nivel $1-\alpha$ si

$$\mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{\theta} \in R(\mathbf{X})) = 1 - \alpha \text{ para todo } \boldsymbol{\theta} \in \mathbb{R}^k\,.$$

b) Una consecuencia directa de esto es que a partir de un procedimiento para obtener intervalos de confianza para un parámetro se pueden obtener regiones de confianza para un vector utilizando un método muy general conocido como *Método de Bonferroni*.

Sea $\mathbf{X} = (X_1, \ldots, X_n)$ una muestra aleatoria de una distribución $F_{(\theta_1,\ldots,\theta_k)}$. Sean $I_1, \ldots, I_k$ intervalos de confianza de nivel $1 - \frac{\alpha}{k}$ para los parámetros $\theta_1, \ldots, \theta_k$, respectivamente. Probar que $I_1 \times \cdots \times I_k$ es una región de confianza de nivel simultáneo mayor o igual a $1-\alpha$ para $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$.

Observar que el intervalo realizado para cada parámetro θ_j tiene un nivel $(1 - \frac{\alpha}{k})$ que resulta mayor al nivel simultáneo buscado, $1-\alpha$.