Universidad Panamericana
Maestría en Ciencia de Datos
Estadística

Enrique Ulises Báez Gómez Tagle Joel Vázquez Anaya
Luis Guillén Roberto Requejo

18 de abril de 2025

Tarea

Resuelve los siguientes ejercicios, incluye sus desarrollos en hojas aparte.

1. (2 puntos) Un desarrollador inmobiliario en la CdMx quiere comprar un terreno en la colonia Del Valle para construir unos edificios. Desea estimar el área A de dicho terreno. Cuando mide la longitud del terreno comete un error aleatorio de modo que la longitud observada X es una variable aleatoria con media μ y varianza σ^2. Preocupado por su posible error, decide hacer dos mediciones independientes X_1 y X_2. Para estimar el área A está en un dilema de cómo proceder por lo que propone dos estimadores:

$$\hat{A}_1 = \left(\frac{X_1 + X_2}{2}\right)^2 \qquad \text{y} \qquad \hat{A}_2 = \frac{X_1^2 + X_2^2}{2}$$

   a) Calcula el sesgo de cada estimador del área $A = \mu^2$.

   b) Determina cuál de los dos es mejor.

   Explicación: El estimador $\hat{A}_1$ presenta un sesgo de $\sigma^2/2$, menor que el sesgo de $\hat{A}_2$ igual a σ^2; por ello, $\hat{A}_1$ se considera el mejor estimador (menor sesgo y menor MSE).

   Solución paso a paso

   Sea X la medición individual con

$$E[X] = \mu, \quad \text{Var}(X) = \sigma^2, \quad X_1, X_2 \text{ i.i.d.}$$

   El área real es $A = \mu^2$.

   1. Sesgo de cada estimador

$$\hat{A}_1 = \left(\frac{X_1 + X_2}{2}\right)^2 = \frac{1}{4}\big(X_1^2 + 2X_1X_2 + X_2^2\big)$$

$$\begin{aligned} E[\hat{A}_1] &= \frac{1}{4}\big(E[X_1^2] + 2E[X_1]E[X_2] + E[X_2^2]\big) \\ &= \frac{1}{4}\big(2(\sigma^2 + \mu^2) + 2\mu^2\big) = \mu^2 + \frac{\sigma^2}{2}. \end{aligned}$$

$$\boxed{\text{Bias}(\hat{A}_1) = E[\hat{A}_1] - A = \frac{\sigma^2}{2}}$$

$$\hat{A}_2 = \frac{X_1^2 + X_2^2}{2}, \quad E[\hat{A}_2] = \frac{1}{2}\big(E[X_1^2] + E[X_2^2]\big) = \mu^2 + \sigma^2.$$

$$\boxed{\text{Bias}(\hat{A}_2) = \sigma^2}$$

2. Comparación de MSE

Para variables normales $N(\mu, \sigma^2)$, los sesgos y varianzas son:

Cantidad	$\hat{A}_1$	$\hat{A}_2$
Bias	$\sigma^2/2$	σ^2
Varianza	$\sigma^2\big(\mu^2 + \frac{\sigma^2}{2}\big)$	$2\,\sigma^2\big(\mu^2 + \frac{\sigma^2}{2}\big)$
MSE = Var + Bias^2	$\sigma^2\big(\mu^2 + \frac{\sigma^2}{2}\big) + \frac{\sigma^4}{4}$	$2\,\sigma^2\big(\mu^2 + \frac{\sigma^2}{2}\big) + \sigma^4$

Claramente el MSE de $\hat{A}_1$ es menor, por lo que $\hat{A}_1$ es el estimador preferido.
Esta conclusión también fue respaldada empíricamente mediante simulaciones computacionales desarrolladas, confirmando que $\hat{A}_1$ presenta un menor error cuadrático medio bajo diversas condiciones de entrada.

2. (2 puntos) Una persona saca del cajero automático \$3,300. El cajero le entrega dos billetes de \$50, seis de \$200 y cuatro de \$500; mismos que guarda en su cartera. Todos los billetes son indistinguibles.

   a) Si esta persona saca al azar dos de estos billetes de su cartera sin reemplazo, sea T la variable aleatoria que denota el monto total de dinero que tiene cada muestra de tamaño 2. Obtenga la distribución de muestreo de T.

   b) Calcule el valor esperado y la varianza de T.

Solución:
El siguiente análisis se llevó a cabo mediante procedimientos computacionales estadísticos, lo cual permitió obtener la distribución de muestreo de la variable aleatoria T, que representa el monto total al sacar dos billetes al azar sin reemplazo.

- **Definición de los billetes:** Se construyó el conjunto `bills` $= \{50, 50, 200, 200, 200, 200, 200, 200, 500, 500, 500, 500\}$.
- **Combinaciones sin reemplazo:** Se generaron todas las combinaciones posibles de dos elementos del conjunto sin reemplazo. El número total de combinaciones se determinó mediante $\binom{12}{2} = 66$.
- **Cálculo de montos:** A cada combinación (x_i, x_j) se le asignó el valor $T = x_i + x_j$.
- **Frecuencias y probabilidades:** Se construyó una tabla de frecuencias f_T y se calcularon las probabilidades relativas mediante $p_T = \frac{f_T}{66}$.

Los resultados obtenidos fueron:

Total ($)	Frecuencia	Probabilidad
100	1	0.0152
250	12	0.1818
400	15	0.2273
550	8	0.1212
700	24	0.3636
1000	6	0.0909

Con base en esta distribución, se calcularon las siguientes medidas:

$$\mathbb{E}[T] = \sum_i t_i \cdot p_i = 550 \mathbb{V}[T] = \sum_i t_i^2 \cdot p_i - (\mathbb{E}[T])^2 \approx 51136{,}36$$

donde t_i son los montos totales posibles y p_i sus respectivas probabilidades.

3. (2 puntos) Un nuevo medicamento está siendo desarrollado en un laboratorio. Se toma una muestra de 10 pacientes similares para probar dicho medicamento. En cada paciente se midió el tiempo de recuperación en días obteniéndose los siguientes valores:

$$10, 10, 7, 12, 5, 7, 2, 1, 7, 2$$

   a) Obtenga un intervalo de confianza del 95 % para el tiempo promedio real de recuperación.

   b) El laboratorio establece que no lanzará el medicamento al mercado si la desviación estándar de los tiempos de recuperación es mayor a 2 días. Determina si el laboratorio lanzará o no el medicamento al mercado mediante la construcción de un intervalo de confianza al 98 %.

c) ¿Cuál es el tamaño de muestra necesario para afirmar, con una probabilidad de 0.9 que el tiempo promedio estimado de recuperación no dista del tiempo real de recuperación en más de 3 días?

Solución:
Tiempos de recuperación (en días) 10, 10, 7, 12, 5, 7, 2, 1, 7, 2.

Estadísticos muestrales

$$n = 10, \quad \bar{X} = 6{,}3, \quad s = 3{,}773.$$

A. Intervalo de confianza del 95 % para μ

Para varianza desconocida utilizamos la distribución t con $\nu = n - 1 = 9$.

$$t_{0{,}975,9} = 2{,}262, \quad \text{M.E.} = t_{0{,}975,9}\,\frac{s}{\sqrt{n}} = 2{,}262\,\frac{3{,}773}{\sqrt{10}} = 2{,}699.$$

$$\boxed{\mu \in (\,\bar{X} - \text{M.E.},\ \bar{X} + \text{M.E.}\,) = (3{,}60,\ 9{,}00) \text{ días}}_{95\,\%}$$

B. ¿Se lanza el medicamento? IC al 98 % para σ

Para la desviación estándar usamos la distribución $\chi^2_{(\nu)}$ con $\nu = 9$ y $\alpha = 0{,}02$.

$$\chi^2_{0{,}01,9} = 21{,}666, \quad \chi^2_{0{,}99,9} = 2{,}088.$$

$$\sigma \in \left(\sqrt{\frac{(n-1)s^2}{\chi^2_{0{,}01,9}}},\ \sqrt{\frac{(n-1)s^2}{\chi^2_{0{,}99,9}}}\right) = (2{,}43,\ 7{,}83) \text{ días} \quad (98\,\%).$$

Como todo el intervalo está por encima de 2 días, la evidencia indica que $\sigma > 2$.

$$\boxed{\text{El laboratorio } \textit{no} \text{ debería lanzar el medicamento.}}$$

C. Tamaño muestral para un error ≤ 3 días con 90 % de confianza

Deseamos $\Pr\big(|\bar{X} - \mu| \leq 3\big) = 0{,}9$. Para un nivel de confianza del 90 % usamos $z_{0{,}95} = 1{,}645$ y aproximamos σ con s:

$$n \;\geq\; \left(\frac{z_{0{,}95}\,s}{d}\right)^2 = \left(\frac{1{,}645 \times 3{,}773}{3}\right)^2 \approx 4{,}28.$$

$$\boxed{n_{\text{mín}} = 5}$$

Por lo tanto, con al menos cinco pacientes se garantiza que la estimación del tiempo promedio de recuperación queda a no más de 3 días del valor real con una confianza del 90 %.

4. (2 puntos) Por la temporada navideña una tienda departamental recibe un lote de chocolates en forma de Santa Claus. El gerente de compras desea estimar la proporción de chocolates rotos en el lote, para ello toma una muestra aleatoria de 100 chocolates, de los cuales 10 están rotos.

   a) Construya un intervalo de confianza al 95 % para la verdadera proporción de chocolates rotos en el lote.

   b) La fábrica de chocolates acepta la devolución si el lote contiene más del 5 % de chocolates rotos. Plantea el problema como una prueba de hipótesis.

   c) ¿Cuál sería la recomendación al gerente de compras?

Solución:

Sea $n = 100$ el tamaño de la muestra y $x = 10$ los chocolates rotos. Por tanto, la proporción muestral es $\hat{p} = \dfrac{10}{100} = 0{,}10$.

A. Intervalo de confianza del 95 % para p

$$\hat{p} \pm z_{0{,}975}\sqrt{\frac{\hat{p}\,(1-\hat{p})}{n}}, \qquad z_{0{,}975} = 1{,}96.$$

$$\text{M.E.} = 1{,}96\sqrt{\frac{0{,}10 \cdot 0{,}90}{100}} = 0{,}0588.$$

$$\boxed{p \in (0{,}041,\ 0{,}159)}_{95\,\%}$$

B. Prueba de hipótesis

$$H_0 : p \leq 0{,}05 \qquad\qquad H_1 : p > 0{,}05$$

Estadístico:

$$z = \frac{\hat{p} - 0{,}05}{\sqrt{\dfrac{0{,}05 \cdot 0{,}95}{100}}} = 2{,}294.$$

$$\text{valor-}p = P(Z > 2{,}294) = 0{,}0109.$$

Como p-value $< 0{,}05$, se **rechaza** H_0.

C. Recomendación

Se recomienda devolver el lote, pues la proporción de chocolates rotos supera el 5 %.

5. (1 punto) La cantidad de agua consumida por un adulto sano sigue una distribución Normal con media de 1.4 litros. Una campaña de salud promueve el consumo de cuando menos 2 litros diarios. Después de la campaña, una muestra de 10 adultos muestra el siguiente consumo en litros:

$$1{,}5,\ 1{,}6,\ 1{,}5,\ 1{,}4,\ 1{,}9,\ 1{,}4,\ 1{,}3,\ 1{,}9,\ 1{,}8,\ 1{,}7$$

a) Con el nivel de significancia de 0.01, ¿se puede concluir que se ha incrementado el consumo de agua?

b) Calcule e interprete el valor-p.

Solución:
Consumo muestral (en litros): 1,5, 1,6, 1,5, 1,4, 1,9, 1,4, 1,3, 1,9, 1,8, 1,7.

Estadísticos muestrales

$$n = 10, \quad \bar{X} = 1{,}60, \quad s = 0{,}216.$$

A. Prueba unilateral al 1 %

Hipótesis

$$H_0 : \mu = 1{,}4 \qquad H_1 : \mu > 1{,}4.$$

Estadístico

$$t_{\text{obs}} = \frac{\bar{X} - 1{,}4}{s/\sqrt{n}} = 2{,}93.$$

Valor crítico para $\alpha = 0{,}01$ (gl $= 9$):

$$t_{0{,}99{,}9} = 2{,}82.$$

Como $t_{\text{obs}} > t_{0{,}99{,}9}$, se **rechaza** H_0.

B. Valor-p

$$p = P\big(T_9 > 2{,}93\big) \approx 0{,}0084.$$

La probabilidad de observar una media tan alta o mayor si $\mu = 1{,}4$ L es solo 0.84 %, por lo que existe evidencia al 1 % de significancia de que el consumo medio de agua ha aumentado.

Parte B: Opción múltiple (1 punto)

Defina si las siguientes aseveraciones son verdaderas (V) o falsas (F).

1. A mayor tamaño de muestra menor longitud del intervalo de confianza. V

   El error estándar es inversamente proporcional al tamaño de muestra; al disminuir, reduce el margen de error y el ancho del intervalo.

2. Al disminuir el nivel de confianza, la longitud del intervalo disminuye. V

   Un nivel de confianza menor implica un valor crítico más pequeño z o t, reduciendo el margen de error.

3. Si el coordinador de la Maestría dice que tiene una confianza del 95 % de que la media de las calificaciones de todos los estudiantes de Estadística está entre 7 y 10, ¿qué es lo que realmente está diciendo?

   Si se repite muchas veces el muestreo y la construcción del intervalo, aproximadamente el 95 % de estos intervalos contendrá la verdadera media poblacional. El intervalo calculado ya está fijo; podría o no incluir la media.

4. En una prueba de hipótesis, la estadística de prueba sigue siempre una distribución Normal. F

   No siempre es normal, puede ser t, χ^2, F, binomial, etc.

5. En una prueba de hipótesis de dos colas, la zona de no rechazo es equivalente al intervalo de confianza para el parámetro de interés. V

   En una prueba bilateral, los valores del parámetro que quedan en la zona de no rechazo son exactamente los que conforman el intervalo de confianza al mismo nivel $1 - \alpha$.