Examen Final SOL114 — Probabilidad e Inferencia Estadística

Instrucciones:
Seleccione las preguntas que desea responder, de modo que la suma de sus ponderaciones alcance **exactamente 100%**. Debe elegir entre **7 y 10 preguntas**. Una vez alcanzado el 100%, respuestas adicionales no serán consideradas. Cada pregunta indica su ponderación entre paréntesis.

Probabilidad e Inferencia Estadística para Datos Sociales

---

1. (5%)
En una región del país, el 55% de los votantes prefiere a Jara y el 45% a Kast. Se sabe, además, que entre quienes tienen educación universitaria, el 60% vota por Jara. Si el 30% de los votantes tiene educación universitaria, ¿cuál es la probabilidad de seleccionar al azar una persona que tenga educación universitaria **y** vote por Jara?

Primero defino claramente los eventos para ordenar el problema:

- U: "tener educación universitaria".
- J: "votar por Jara".

La probabilidad que se pide es la probabilidad conjunta de que ambas cosas ocurran a la vez, es decir $P(U \cap J)$. A partir de las reglas de probabilidad sé que:

$$P(U \cap J) = P(J \mid U)\, P(U),$$

porque la probabilidad conjunta se puede expresar como "probabilidad condicional por probabilidad marginal".

Del enunciado tomo los valores:

- $P(J \mid U) = 0.60$ (60% de quienes tienen educación universitaria votan por Jara).
- $P(U) = 0.30$ (30% de los votantes tiene educación universitaria).

Entonces reemplazo en la fórmula:

$$P(U \cap J) = 0.60 \times 0.30 = 0.18.$$

Por lo tanto, la probabilidad de elegir al azar una persona que **tenga educación universitaria y vote por Jara** es 0.18, es decir, un 18% de todos los votantes se ubica simultáneamente en ambas categorías.

---

2. (5%)
Usando el contexto anterior, calcule la probabilidad de que un votante que apoya a Jara tenga educación universitaria. Interprete el resultado.

En este caso la probabilidad es condicional "al revés" del ejercicio anterior: ahora quiero $P(U \mid J)$, es decir, la probabilidad de que la persona tenga educación universitaria dado que **ya sabemos** que vota por Jara.

Uso la definición de probabilidad condicional:

$$P(U \mid J) = \frac{P(U \cap J)}{P(J)}.$$

Del ejercicio 1 ya calculé la probabilidad conjunta:

$$P(U \cap J) = 0.18.$$

Y del enunciado sé que:

$$P(J) = 0.55,$$

porque el 55% de los votantes prefiere a Jara. Entonces sustituyo:

$$P(U \mid J) = \frac{0.18}{0.55} \approx 0.3273.$$

Numéricamente, 0.3273 equivale aproximadamente a un 32.7%.

Interpretación: entre quienes apoyan a Jara (es decir, restringiendo la población solo a votantes de Jara), alrededor de un tercio tiene educación universitaria. Es importante notar que esto es distinto de decir "el 30% de la población tiene educación universitaria": aquí estoy reponderando sobre el subconjunto de personas que votan por Jara.

---

3. (5%)
Suponga que 5 votantes se seleccionan al azar y que cada uno vota por Kast con probabilidad 0.45, de manera independiente. ¿Cuál es la probabilidad de que exactamente 4 voten por Kast?

Primero identifico el modelo probabilístico apropiado. Cada votante tiene dos posibles resultados: "vota Kast" o "no vota Kast", con:

- $P(\text{vota Kast}) = 0.45$,
- $P(\text{no vota Kast}) = 1 - 0.45 = 0.55$,

y los votantes se consideran independientes. Por lo tanto, el número de personas que votan por Kast en una muestra de $n = 5$ sigue una distribución Binomial:

$$X \sim \text{Binomial}(n = 5, p = 0.45).$$

Quiero la probabilidad de que exactamente 4 voten por Kast, es decir:

$$P(X = 4).$$

La fórmula de la Binomial es:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}.$$

Sustituyo $n = 5$, $k = 4$ y $p = 0.45$:

$$P(X = 4) = \binom{5}{4} (0.45)^4 (0.55)^{5-4}.$$

Calculo paso a paso:

1. El coeficiente binomial:

$$\binom{5}{4} = 5.$$

2. La potencia de la probabilidad de "éxito":

$$(0.45)^4.$$

3. La probabilidad de "fracaso" elevada a la cantidad correspondiente:

$$(0.55)^1 = 0.55.$$

Luego, la probabilidad total es:

$$P(X = 4) = 5 \times 0.45^4 \times 0.55 \approx 0.113.$$

Por lo tanto, la probabilidad de que exactamente 4 de los 5 votantes seleccionados voten por Kast es aproximadamente 0.113, es decir, un 11.3%.

---

4. (5%)
El ingreso mensual sigue $N(800, 200^2)$. ¿Cuál es la probabilidad de ingreso mayor a 1000?

Aquí se plantea una variable aleatoria continua con distribución Normal:

$$X \sim N(\mu = 800,\ \sigma^2 = 200^2).$$

Me piden $P(X > 1000)$. Para calcular esta probabilidad uso la estandarización a la distribución Normal estándar $Z \sim N(0, 1)$.

Primero calculo el puntaje z correspondiente a $x = 1000$:

$$z = \frac{x - \mu}{\sigma} = \frac{1000 - 800}{200} = \frac{200}{200} = 1.$$

Entonces, la probabilidad original se traduce a:

$$P(X > 1000) = P\left(Z > \frac{1000 - 800}{200}\right) = P(Z > 1).$$

Uso la tabla de la Normal estándar (o un software) para obtener $P(Z \leq 1) = \Phi(1) \approx 0.8413$. Por tanto:

$$P(Z > 1) = 1 - \Phi(1) \approx 1 - 0.8413 = 0.1587.$$

Por lo tanto, aproximadamente un 15.9% de las personas tiene un ingreso mensual superior a 1000 (en las mismas unidades del enunciado).

---

5. (10%)
Sea X el ingreso en 2010:
$\mathbb{E}[X] = 834$, $\text{Var}(X) = 270^2$
$Y = 200 + 1.5X$ (ingreso en 2030)

a) Calcular $\mathbb{E}[Y]$

Uso la propiedad de linealidad de la esperanza matemática. Si $Y = a + bX$, entonces:

$$\mathbb{E}[Y] = a + b\,\mathbb{E}[X].$$

En este caso $a = 200$, $b = 1.5$ y $\mathbb{E}[X] = 834$, entonces:

$$\mathbb{E}[Y] = 200 + 1.5 \cdot 834.$$

Calculo el producto:

$$1.5 \cdot 834 = 1.5 \cdot (800 + 34) = 1.5 \cdot 800 + 1.5 \cdot 34 = 1200 + 51 = 1251.$$

Luego:

$$\mathbb{E}[Y] = 200 + 1251 = 1451.$$

Es decir, el ingreso promedio esperado para 2030 es de 1451.

b) Calcular $\operatorname{Var}(Y)$

Para una transformación lineal $Y = a + bX$, la varianza cumple:

$$\operatorname{Var}(Y) = b^2 \operatorname{Var}(X),$$

porque sumar una constante no cambia la varianza y multiplicar por b la escala por b^2.

Aquí $b = 1.5$ y $\operatorname{Var}(X) = 270^2$, entonces:

$$\operatorname{Var}(Y) = (1.5)^2 \cdot 270^2.$$

Primero calculo:

$$(1.5)^2 = 2.25,$$

$$270^2 = 72{,}900.$$

Ahora:

$$\operatorname{Var}(Y) = 2.25 \cdot 72{,}900 = 164{,}025.$$

c) Calcular $\operatorname{Cov}(X, Y)$

La covarianza entre X y una transformación lineal $Y = a + bX$ tiene la forma:

$$\operatorname{Cov}(X, a + bX) = b \operatorname{Cov}(X, X) = b \operatorname{Var}(X).$$

Por lo tanto, con $b = 1.5$:

$$\operatorname{Cov}(X, Y) = 1.5 \cdot \operatorname{Var}(X) = 1.5 \cdot 72{,}900.$$

Calculo:

$$1.5 \cdot 72{,}900 = 109{,}350.$$

Interpretación: el ingreso promedio esperado en 2030 es 1451, mayor que el promedio de 834 en 2010, lo que refleja un aumento sistemático. La varianza aumenta a 164,025, lo que indica que la dispersión de los ingresos también se amplifica con el factor 1.5. La covarianza positiva y grande, $\text{Cov}(X, Y) = 109{,}350$, muestra una fuerte asociación positiva: quienes tenían mayores ingresos en 2010 tienden, en promedio, a tener mayores ingresos también en 2030.

---

6. (15%)
La Ley de los Grandes Números establece:

$$\lim_{n\to\infty} P(|\bar{X}_n - \mu| < \varepsilon) = 1,$$

donde:

- $\bar{X}_n$ es el promedio muestral de una muestra de tamaño n,
- μ es la media poblacional verdadera,
- ε es un margen (tolerancia) positivo y pequeño.

Primero interpreto la expresión: para cualquier margen de error $\varepsilon > 0$, por pequeño que sea, la probabilidad de que la diferencia entre el promedio muestral $\bar{X}_n$ y la verdadera media μ sea menor a ε **tiende a 1** cuando el tamaño muestral n crece sin límite.

Esto significa, en palabras, que:

- Cuando el tamaño de la muestra es **muy grande**, el promedio observado en la muestra se aproxima mucho al promedio real de la población.
- La discrepancia entre $\bar{X}_n$ y μ se debe al azar del muestreo, pero a medida que incorporamos más observaciones, este azar "se compensa" y el promedio muestral se estabiliza alrededor de μ.

Aplicado a encuestas en ciencias sociales:

- Si tomamos una muestra pequeña, el promedio muestral puede ser muy distinto del promedio poblacional simplemente por azar.
- Si tomamos muestras grandes y repetimos el proceso, veremos que los promedios muestrales se concentran cada vez más cerca de μ.
- Por eso, la Ley de los Grandes Números justifica que el promedio muestral sea un **estimador consistente** de la media poblacional: al aumentar n, la estimación se vuelve más precisa y confiable.

En resumen, la Ley de los Grandes Números dice que "promediar mucho" reduce el impacto del azar de muestreo y hace que el promedio de la muestra represente mejor a la población.

---

7. (5%)
Distribución muestral del promedio ($\bar{x} = 834$, $s = 270$, $n = 3000$).

Primero identifico qué se me pide: la distribución muestral de la media. Asumo que la población tiene media $\mu \approx 834$ y desviación estándar $\sigma \approx 270$, y que el tamaño muestral es grande ($n = 3000$). En estas condiciones, por el Teorema Central del Límite, el promedio muestral $\bar{X}$ se distribuye aproximadamente Normal con:

- media igual a la media poblacional, y
- desviación estándar igual al error estándar.

El error estándar del promedio es:

$$SE(\bar{X}) = \frac{\sigma}{\sqrt{n}} \approx \frac{270}{\sqrt{3000}}.$$

Calculo $\sqrt{3000}$:

$$\sqrt{3000} \approx 54.77,$$

entonces:

$$SE(\bar{X}) \approx \frac{270}{54.77} \approx 4.93.$$

Por lo tanto, la distribución muestral del promedio se aproxima a:

$$\bar{X} \sim N(834,\ 4.93^2).$$

Interpretación: si repitiéramos muchas veces la encuesta, cada vez con una muestra aleatoria de tamaño 3000, los promedios obtenidos en cada muestra seguirían aproximadamente una distribución Normal con centro en 834 y variación alrededor de ese valor medida por un error estándar de aproximadamente 4.93.

8. (10%)
Explique qué es una distribución muestral y cómo se relaciona con el error estándar.

Una **distribución muestral** describe cómo se comporta un **estadístico** (por ejemplo, un promedio, una proporción o una diferencia de promedios) cuando repetimos el proceso de muestreo muchas veces bajo las mismas condiciones. La idea es la siguiente:

1. Fijamos una población y un tamaño muestral n.
2. Tomamos una primera muestra aleatoria y calculamos un estadístico, por ejemplo el promedio $\bar{X}_1$.
3. Repetimos el proceso muchas veces: tomamos otra muestra, calculamos $\bar{X}_2$; luego otra, obtenemos $\bar{X}_3$, y así sucesivamente.
4. Si miramos todos esos valores $\bar{X}_1, \bar{X}_2, ...$, veremos que no son idénticos: varían por efecto del azar del muestreo.

La distribución de todos esos posibles valores del estadístico (en este ejemplo, los distintos promedios muestrales) es la **distribución muestral** del estadístico.

El **error estándar (SE)** es un número que resume cuánta variabilidad hay en esa distribución muestral. Por ejemplo, en el caso del promedio:

$$SE(\bar{X}) = \frac{\sigma}{\sqrt{n}},$$

donde σ es la desviación estándar poblacional y n el tamaño de la muestra. Este SE es, en esencia, la **desviación estándar de la distribución muestral** del promedio.

Es crucial distinguir:

- La **desviación estándar muestral** s mide la variabilidad de **las personas** dentro de una misma muestra (qué tan dispersos están los ingresos, por ejemplo, entre los individuos).
- El **error estándar** mide la variabilidad de **un estadístico** (por ejemplo, del promedio) **entre muestras hipotéticas diferentes**.

En otras palabras, el error estándar no describe la dispersión de los datos individuales, sino la precisión del estimador: si el SE es pequeño, significa que, al repetir el muestreo, los valores del estadístico (por ejemplo, los promedios) se agruparían fuertemente alrededor del valor real poblacional; si es grande, esos valores fluctuarían más de una muestra a otra.

---

9. (10%)
Intervalo de confianza al 99% para la media poblacional.

Tengo:

- promedio muestral $\bar{x} = 834$,
- error estándar $SE \approx 4.93$,
- nivel de confianza 99%.

Para un intervalo de confianza al 99% con distribución Normal, el valor crítico es aproximadamente:

$$z_{0.005} = 2.576,$$

porque dejo 0.005 de probabilidad en cada cola.

La fórmula del intervalo de confianza es:

$$IC_{99\%} = \bar{x} \pm z_{0.005} \cdot SE.$$

Sustituyo los valores:

$$IC_{99\%} = 834 \pm 2.576 \cdot 4.93.$$

Primero calculo el margen de error:

$$2.576 \cdot 4.93 \approx 12.7.$$

Entonces los extremos del intervalo son:

$$\text{límite inferior} = 834 - 12.7 \approx 821.3,$$

$$\text{límite superior} = 834 + 12.7 \approx 846.7.$$

Así, el intervalo es:

$$IC_{99\%} \approx [821.3;\ 846.7].$$

Interpretación correcta: el procedimiento de construcción del intervalo está diseñado de modo que, si repitiéramos el muestreo muchas veces y cada vez construyéramos un intervalo usando la misma fórmula, aproximadamente el 99% de esos intervalos contaría con el verdadero promedio poblacional μ. Para **este** intervalo específico, no tiene sentido decir que hay un "99% de probabilidad" de que μ esté adentro, porque μ es un número fijo (no aleatorio). Lo que es aleatorio son las muestras y, por lo tanto, los intervalos resultantes. Lo que afirmamos es que estamos usando un método que, en el largo plazo, captura el valor verdadero en el 99% de los casos.

---

10. (10%)
Test de proporción: $\hat{p} = 0.52$, $p_0 = 0.50$, $n = 3000$, $\alpha = 0.01$.

Primero planteo las hipótesis:

- $H_0 : p = 0.50$ (la proporción poblacional es 50%),
- $H_1 : p \neq 0.50$ (la proporción es distinta de 50%).

El estadístico de prueba para una proporción, usando aproximación Normal, es:

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}.$$

Sustituyo los valores:

- $\hat{p} = 0.52$,
- $p_0 = 0.50$,
- $n = 3000$.

Primero calculo el denominador (error estándar bajo H_0):

$$SE_0 = \sqrt{\frac{p_0(1-p_0)}{n}} = \sqrt{\frac{0.5 \cdot 0.5}{3000}} = \sqrt{\frac{0.25}{3000}}.$$

Entonces:

$$\frac{0.25}{3000} \approx 0.00008333,$$

y

$$SE_0 \approx \sqrt{0.00008333} \approx 0.00913.$$

Ahora calculo el estadístico z:

$$z = \frac{0.52 - 0.50}{0.00913} = \frac{0.02}{0.00913} \approx 2.19.$$

Para obtener el p-value en un test bilateral:

$$p\text{-value} = 2 \cdot P(Z > |z|) = 2 \cdot P(Z > 2.19).$$

De la tabla de la Normal estándar, $P(Z > 2.19) \approx 0.0143$, por lo tanto:

$$p\text{-value} \approx 2 \cdot 0.0143 = 0.0286.$$

Finalmente, comparo:

- p-value ≈ 0.0286,
- $\alpha = 0.01$.

Como $0.0286 > 0.01$, **no** rechazo H_0.

Conclusión sustantiva: con un nivel de significancia del 1%, la evidencia estadística no es suficientemente fuerte como para afirmar que la proporción real difiere del 50%. Aunque la estimación puntual es 52%, esa diferencia de 2 puntos porcentuales no alcanza el umbral de significancia exigente que fijamos ($\alpha = 0.01$).

---

11. (15%)
Test de diferencia de proporciones entre hombres y mujeres.

Supongo que el ejercicio entrega:

- $\widehat{p}_H = 0.55$ con $n_H = 1400$ (hombres),
- $\widehat{p}_M = 0.49$ con $n_M = 1600$ (mujeres),
- diferencia observada: $\widehat{p}_H - \widehat{p}_M = 0.06$.

Quiero probar si esta diferencia es estadísticamente distinta de cero al 1%.

Las hipótesis son:

- $H_0 : p_H - p_M = 0$,
- $H_1 : p_H - p_M \neq 0$.

El error estándar de la diferencia de proporciones se calcula (sin “pooled” en este caso) como:

$$SE(\widehat{p}_H - \widehat{p}_M) = \sqrt{\frac{\widehat{p}_H(1-\widehat{p}_H)}{n_H} + \frac{\widehat{p}_M(1-\widehat{p}_M)}{n_M}}.$$

Sustituyo valores:

$$SE = \sqrt{\frac{0.55 \cdot 0.45}{1400} + \frac{0.49 \cdot 0.51}{1600}}.$$

Calculo cada término:

$$\frac{0.55 \cdot 0.45}{1400} = \frac{0.2475}{1400} \approx 0.0001768,$$

$$\frac{0.49 \cdot 0.51}{1600} = \frac{0.2499}{1600} \approx 0.0001562.$$

Sumo:

$$0.0001768 + 0.0001562 \approx 0.000333.$$

Entonces:

$$SE \approx \sqrt{0.000333} \approx 0.01825.$$

El estadístico de prueba es:

$$z = \frac{(\hat{p}_H - \hat{p}_M) - 0}{SE} = \frac{0.06}{0.01825} \approx 3.29.$$

Para $\alpha = 0.01$ en un test bilateral, el valor crítico es aproximadamente $z_{0.005} = 2.576$. Comparo:

- $|z| = 3.29$,
- 2.576 es el umbral.

Como $3.29 > 2.576$, la estadística cae en la región de rechazo.

Conclusión: se rechaza H_0 al 1% de significancia. Hay evidencia estadísticamente significativa de que la proporción difiere entre hombres y mujeres. Sustantivamente, la proporción de hombres (55%) es mayor que la de mujeres (49%), y esa diferencia de 6 puntos porcentuales no puede atribuirse solo al azar de muestreo con este nivel de confianza.

---

12. (10%)
Tamaño muestral necesario para estimar una proporción con error máximo $E = 0.001$ al 99% de confianza, usando $p = 0.5$ como caso conservador.

La fórmula general para el tamaño muestral en estimación de una proporción es:

$$n = \frac{z_{\alpha/2}^2 \, p(1-p)}{E^2},$$

donde:

- $z_{\alpha/2}$ es el valor crítico (para 99%, $\alpha = 0.01$ y $\alpha/2 = 0.005$),
- p es una suposición inicial sobre la proporción,
- E es el error máximo tolerado (margen de error).

Para 99% de confianza:

$$z_{0.005} = 2.576.$$

El problema indica usar $p = 0.5$, que es el valor que maximiza $p(1-p)$ y, por tanto, da el tamaño muestral más grande (más conservador).

Sustituyo:

$$n = \frac{(2.576)^2 \cdot 0.5(1-0.5)}{(0.001)^2} = \frac{(2.576)^2 \cdot 0.25}{(0.001)^2}.$$

Calculo paso a paso:

$$(2.576)^2 \approx 6.64,$$

entonces:

$$n = \frac{6.64 \cdot 0.25}{0.000001} = \frac{1.66}{0.000001} \approx 1{,}660{,}000.$$

Por lo tanto, se requerirían aproximadamente 1,660,000 casos.

Comentario: este resultado muestra que exigir un margen de error extremadamente pequeño ($E = 0.001$) con un nivel de confianza muy alto (99%) implica muestras gigantescas en encuestas reales.

---

13. (5%)
Cálculo de la correlación a partir de la covarianza y las varianzas.

Supongamos que se entregan:

- $\text{Cov}(X, Y) = 312.4$,
- $\text{Var}(X) = 72{,}900$,
- $\text{Var}(Y) = 6.76$.

Primero obtengo las desviaciones estándar de cada variable:

$$\sigma_X = \sqrt{\mathrm{Var}(X)} = \sqrt{72{,}900},$$

$$\sigma_Y = \sqrt{\mathrm{Var}(Y)} = \sqrt{6.76}.$$

Numéricamente:

$$\sigma_X \approx 270,$$

$$\sigma_Y \approx 2.6.$$

El coeficiente de correlación se define como:

$$\rho_{XY} = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y}.$$

Sustituyo:

$$\rho = \frac{312.4}{\sqrt{72{,}900 \cdot 6.76}}.$$

Como ya tengo aproximaciones:

$$\sqrt{72{,}900 \cdot 6.76} \approx 270 \cdot 2.6 = 702,$$

entonces:

$$\rho \approx \frac{312.4}{702} \approx 0.445.$$

Interpretación: la correlación es aproximadamente 0.445, lo que indica una asociación positiva de magnitud moderada entre las dos variables. En el contexto de ingreso y escolaridad, esto significa que quienes tienen más años de escolaridad tienden, en promedio, a tener mayores ingresos, aunque la relación está lejos de ser perfecta (no todos los casos siguen exactamente esa tendencia).

14. (10%)
Modelo cuadrático y correlación.

Supongo que el modelo ajustado es:

$$\text{Ingreso} = 50 + 4.1\,(\text{Escolaridad} - 8)^2.$$

Primero observo que se trata de una relación cuadrática centrada en 8 años de escolaridad. El término $(\text{Escolaridad} - 8)^2$ es siempre no negativo y toma su valor mínimo cuando $\text{Escolaridad} = 8$. Eso significa que el ingreso mínimo según el modelo se alcanza en 8 años de escolaridad y luego crece tanto para menos como para más de 8.

Este tipo de relación genera una forma de "U" simétrica en torno a 8. Esto es importante para la correlación porque:

- La **correlación de Pearson** mide la **asociación lineal** entre dos variables.
- Si la relación es fuertemente curvilínea y simétrica alrededor de un punto, los "aumentos" y "disminuciones" pueden cancelarse en promedio al calcular el producto de desviaciones.

En otras palabras, aunque haya una relación sistemática (el ingreso varía claramente con la escolaridad), la forma cuadrática hace que el ajuste lineal no capture bien la tendencia global. El coeficiente de correlación podría incluso ser cercano a cero porque la parte "ascendente" y la parte "descendente" de la curva se compensan al resumir todo en una sola línea recta.

Por eso, en este ejemplo puede ocurrir que:

- Exista una relación clara y fuerte entre escolaridad e ingreso (descripta por el modelo cuadrático),
- Pero la correlación de Pearson entre ambas variables sea baja o incluso cercana a cero, ya que la correlación únicamente mide el componente lineal de esa relación.

15. (10%)
Explique con precisión qué es el p-value.

El p-value se define formalmente como:

$$p\text{-value} = P(T \geq t_{\text{obs}} \,\big|\, H_0 \text{ verdadera}),$$

en el caso de un test unilateral con cola superior (en un bilateral se consideran ambas colas). En palabras:

- T es el estadístico de prueba (por ejemplo, un z o un t),
- t_{obs} es el valor observado de ese estadístico en nuestros datos,
- H_0 es la hipótesis nula.

El p-value es la probabilidad, **suponiendo que la hipótesis nula es verdadera**, de obtener un valor del estadístico de prueba tan extremo como el observado (o más extremo) solo por azar de muestreo.

Es importante enfatizar que el p-value:

- Es una **probabilidad condicional sobre los datos**: se calcula bajo el supuesto de que H_0 es cierta.
- **No** es la probabilidad de que la hipótesis nula sea verdadera o falsa.
- **No** es la probabilidad de que los resultados se deban "solo al azar" en un sentido general; es una probabilidad calculada dentro de un modelo específico donde H_0 se asume cierta.

En la práctica, si el p-value es muy pequeño (menor que α), interpretamos que sería muy raro observar datos tan extremos si H_0 fuera verdadera, lo cual nos lleva a **rechazar** H_0 a ese nivel de significancia.

---

16. (10%)
Explique por qué, si rechazamos H_0 cuando $p < \alpha$, la probabilidad de cometer un error tipo I es igual a α.

Recordemos que un **error tipo I** consiste en **rechazar** H_0 **siendo** H_0 **verdadera**. El nivel de significancia α se define justamente como:

$$\alpha = P(\text{rechazar } H_0 \mid H_0 \text{ verdadera}).$$

En un test clásico, construimos una **regla de decisión** basada en el estadístico de prueba T y una región de rechazo. Por ejemplo:

- En un test bilateral, rechazamos H_0 si $|T| > t_{\alpha/2}$.
- En la versión basada en p-value, rechazamos si p-value $< \alpha$.

Ambas reglas son equivalentes: el p-value es precisamente la probabilidad (bajo H_0) de caer en la región de valores tan extremos o más que el observado. Cuando fijamos un umbral α y definimos la región de rechazo de modo que bajo H_0:

$$P(\text{estadística cae en la región de rechazo}) = \alpha,$$

estamos garantizando que, si H_0 es verdadera y repetimos el experimento muchas veces, en aproximadamente un $\alpha \times 100\%$ de los casos terminaremos rechazando H_0 **por error** (es decir, cometiendo un error tipo I).

Por lo tanto, si usamos la regla "rechazar cuando $p < \alpha$", estamos construyendo una prueba donde, por diseño, la probabilidad de cometer un error tipo I en el largo plazo es igual a α. Esa es justamente la interpretación de α como nivel de significancia: el control de la tasa de falsos positivos cuando H_0 es verdadera.
