Apuntes - Muestreo e intervalos de confianza

Curso: 2 Bachillerato Ccss

La estadística inferencial permite conocer una población estudiando solo una muestra. Aprenderás a elegir muestras, a describir cómo varían la media y la proporción muestrales y a construir intervalos de confianza, como en las encuestas de opinión y los controles de calidad.

Actividad de introducción: la encuesta de la cafetería

La asociación de estudiantes de un instituto quiere saber si los alumnos están contentos con la cafetería. Preguntar a los 900 alumnos llevaría semanas, así que deciden preguntar a 90. Surgen enseguida varias dudas: ¿a quién se pregunta, a los que están en la cola a la hora del recreo o a alumnos elegidos por sorteo en las listas de clase? Si 54 de los 90 dicen que están contentos, ¿se puede afirmar que exactamente el 60 % del instituto lo está? ¿Y si otra muestra hubiera dado 57? Este tema responde a esas preguntas: cómo elegir bien una muestra, cuánto varían los resultados de una muestra a otra y cómo dar una estimación con un margen de error y un nivel de confianza, igual que hacen los institutos de estadística y las empresas de sondeos.

Población, muestra y técnicas de muestreo

Definición: población y muestra

La **población** es el conjunto de todos los individuos que se quiere estudiar; la **muestra** es un subconjunto de ella. La muestra debe ser **representativa**: sus características deben reflejar las de la población. El **tamaño muestral** n es el número de individuos de la muestra.

Definición: técnicas de muestreo

Aleatorio simple: todos los individuos tienen la misma probabilidad de ser elegidos. **Sistemático**: se elige uno al azar entre los k primeros y después uno de cada $k = N/n$. **Estratificado**: se divide la población en estratos homogéneos y se toma de cada uno una submuestra, proporcional a su tamaño. **Por conglomerados**: se eligen grupos enteros al azar.

Ejemplo 1

Un instituto tiene 1200 alumnos: 600 de ESO, 300 de Bachillerato y 300 de FP. Se quiere una muestra estratificada de 80 con afijación proporcional. ¿Cuántos de cada etapa?

La fracción de muestreo es $\frac{80}{1200} = \frac{1}{15}$. ESO: $600/15 = 40$; Bachillerato: $300/15 = 20$; FP: $300/15 = 20$. Total: $40 + 20 + 20 = 80$.

Ejemplo 2

En una lista de 500 clientes se quiere una muestra sistemática de 20. ¿Cada cuántos se elige?

$k = \frac{500}{20} = 25$. Se elige al azar un número entre 1 y 25 (por ejemplo, el 7) y después el 32, el 57, el 82, y así hasta 20 clientes.

Distribución de la media y de la proporción muestrales

Teorema: distribución de la media muestral

Si $X \sim N(\mu, \sigma)$ (o si $n \geq 30$), la media muestral sigue $\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$. La desviación típica $\frac{\sigma}{\sqrt{n}}$ se llama **error típico**: disminuye al aumentar n.

Teorema: distribución de la proporción muestral

Si p es la proporción en la población y n es grande ($np \geq 5$, $nq \geq 5$), la proporción muestral sigue $\hat{p} \sim N\left(p, \sqrt{\frac{pq}{n}}\right)$.

Ejemplo 3

Las alturas de una población siguen $N(170, 8)$. Para muestras de 64 personas, calcula $P(169 \leq \bar{X} \leq 172)$.

$\bar{X} \sim N\left(170, \frac{8}{\sqrt{64}}\right) = N(170, 1)$. $P = \Phi(2) - \Phi(-1) = 0{,}9772 - 0{,}1587 = 0{,}8185$.

Ejemplo 4

El 30 % de los usuarios de una app compra contenido. En una muestra de 100, halla $P(\hat{p} > 0{,}35)$.

$\hat{p} \sim N\left(0{,}3; \sqrt{\frac{0{,}3 \cdot 0{,}7}{100}}\right) = N(0{,}3;\ 0{,}0458)$. $z = \frac{0{,}35 - 0{,}3}{0{,}0458} = 1{,}09$.

$P = 1 - \Phi(1{,}09) = 1 - 0{,}8621 = 0{,}1379$.

Intervalos de confianza

Definición: intervalo de confianza

Un **intervalo de confianza** al nivel $1-\alpha$ es un intervalo calculado con la muestra que contiene al parámetro desconocido con probabilidad $1-\alpha$. Los niveles habituales y su valor crítico $z_{\alpha/2}$ son: 90 %, $1{,}645$; 95 %, $1{,}96$; 99 %, $2{,}575$.

Regla: intervalo para la media

Con σ conocida: $\left(\bar{x} - z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}},\ \bar{x} + z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}}\right)$. El **error máximo** es $E = z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}}$.

Regla: intervalo para la proporción

$\left(\hat{p} - z_{\alpha/2}\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}},\ \hat{p} + z_{\alpha/2}\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}\right)$, con $E = z_{\alpha/2}\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}$.

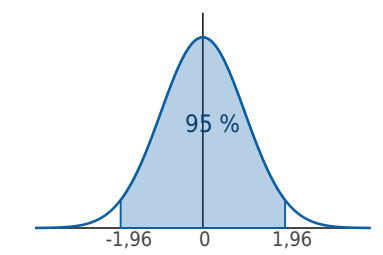
Método: construir un intervalo de confianza

1. Identifica el parámetro (media o proporción) y los datos: $\bar{x}$ o $\hat{p}$, σ, n y el nivel de confianza.
2. Busca $z_{\alpha/2}$ (90 %: 1,645; 95 %: 1,96; 99 %: 2,575).
3. Calcula el error máximo E.
4. Escribe el intervalo $(\bar{x} - E,\ \bar{x} + E)$ o $(\hat{p} - E,\ \hat{p} + E)$.
5. Interpreta el resultado con una frase.

Ejemplo 5

El tiempo diario de uso del móvil de los jóvenes tiene $\sigma = 1{,}5$ horas. En una muestra de 100 jóvenes, la media fue 4 horas. Halla el intervalo de confianza al 95 % para la media poblacional.

$E = 1{,}96 \cdot \dfrac{1{,}5}{\sqrt{100}} = 1{,}96 \cdot 0{,}15 = 0{,}294$. Intervalo: $(4 - 0{,}294;\ 4 + 0{,}294) = (3{,}706;\ 4{,}294)$ horas. Con una confianza del 95 %, el tiempo medio está entre 3,71 y 4,29 horas.



Ejemplo 6

En una encuesta a 400 personas, 220 están a favor de una medida. Halla el intervalo de confianza al 95 % para la proporción.

$\hat{p} = \frac{220}{400} = 0{,}55$. $E = 1{,}96\sqrt{\frac{0{,}55 \cdot 0{,}45}{400}} = 1{,}96 \cdot 0{,}0249 = 0{,}0487$. Intervalo: $(0{,}5013;\ 0{,}5987)$, es decir, entre el 50,1 % y el 59,9 %.

¡Atención!

Un intervalo al 95 % *no* significa que el 95 % de los individuos esté dentro de él, ni que el parámetro «tenga un 95 % de probabilidad» de estar en este intervalo concreto: significa que el 95 % de los intervalos construidos así contienen al parámetro.

Error máximo y tamaño de la muestra

Propiedad: efecto de n, σ y el nivel

El error E disminuye al aumentar n (con $4n$, el error se divide por 2), crece con σ y crece al aumentar el nivel de confianza. Para un error máximo E fijado: $n \geq \left(\frac{z_{\alpha/2}\,\sigma}{E}\right)^2$ para la media y $n \geq \frac{z_{\alpha/2}^2\,\hat{p}(1-\hat{p})}{E^2}$ para la proporción (si $\hat{p}$ es desconocida, se toma $\hat{p} = 0{,}5$). El resultado se redondea siempre hacia arriba.

Ejemplo 7

¿Qué tamaño de muestra hace falta para estimar una media con $\sigma = 10$, nivel del 95 % y error máximo 2?

$n \geq \left(\frac{1{,}96 \cdot 10}{2}\right)^2 = 9{,}8^2 = 96{,}04$. Se necesitan al menos 97 individuos.

Observación

En la PAU de Ciencias Sociales se pide casi siempre un intervalo de confianza para la media o la proporción, el error máximo, el tamaño de muestra necesario o la interpretación del nivel de confianza, junto a la distribución de la media o la proporción muestrales.

De la muestra a la población: ideas clave

Observación: por qué funciona el muestreo

Una muestra bien elegida se parece a la población, pero nunca es idéntica. La media de una muestra es una variable aleatoria: si se repite el muestreo, sale un valor distinto cada vez. Lo sorprendente es que esa variabilidad se puede medir. El teorema central del límite garantiza que, si la muestra tiene al menos 30 individuos, la media muestral sigue aproximadamente una distribución normal aunque la población no lo sea. Por eso las fórmulas de este tema se aplican también a variables como el gasto o el tiempo de espera, que no son exactamente normales. Del mismo modo, la proporción muestral es aproximadamente normal cuando $np \geq 5$ y $nq \geq 5$.

Ejemplo 8

En una población de jubilados el gasto mensual en farmacia tiene media 40 € y desviación típica 15 €, pero su distribución no es normal. Se toma una muestra de 100 jubilados. ¿Qué probabilidad hay de que el gasto medio supere los 43 €?

Como $n = 100 \geq 30$, por el teorema central del límite $\bar{X} \approx N\left(40;\ \frac{15}{10}\right) = N(40;\ 1{,}5)$.

Tipificamos: $z = \frac{43 - 40}{1{,}5} = 2$ y $P(Z > 2) = 1 - 0{,}9772 = 0{,}0228$. Solo un 2,28 % de las muestras daría una media tan alta.

Método: elegir la fórmula correcta

1. Lee si el enunciado habla de una media (euros, minutos, kilos) o de una proporción (porcentaje de personas que cumplen algo).
2. Si es una media, necesitas $\bar{x}$, σ y n; si es una proporción, necesitas $\hat{p} = \frac{\text{casos favorables}}{n}$.
3. Busca el valor crítico según el nivel de confianza: 1,645, 1,96 o 2,575.
4. Calcula el error máximo E y escribe el intervalo $(\text{centro} - E,\ \text{centro} + E)$.
5. Termina con una frase de interpretación en el contexto del problema.

Ejemplo 9

Se pregunta a 250 clientes de un supermercado si usarían la caja de autopago y 150 dicen que sí. Halla el intervalo de confianza al 90 % para la proporción de clientes que la usarían.

Es una proporción: $\hat{p} = \frac{150}{250} = 0{,}6$ y $\hat{q} = 0{,}4$. El error es

$E = 1{,}645\sqrt{\frac{0{,}6 \cdot 0{,}4}{250}} = 1{,}645 \cdot 0{,}031 \approx 0{,}051$. El intervalo es $(0{,}549;\ 0{,}651)$: con una confianza del 90 %, entre el 54,9 % y el 65,1 % de los clientes usaría el autopago.

¡Atención!

Errores muy frecuentes en los exámenes: usar σ en lugar de $\frac{\sigma}{\sqrt{n}}$ al tipificar una media muestral; redondear hacia abajo el tamaño de la muestra; confundir el nivel de confianza con la probabilidad de que la media esté en un intervalo ya calculado (una vez calculado, la media está o no está dentro); y olvidar que el intervalo para una proporción se expresa entre 0 y 1, no en porcentaje dentro de la fórmula.

¿Sabías que?

En 1936 la revista estadounidense Literary Digest envió diez millones de cuestionarios para predecir las elecciones presidenciales y recibió más de dos millones de respuestas, pero falló estrepitosamente: había elegido a los encuestados a partir de listines telefónicos y registros de coches, que en plena crisis solo tenían las familias acomodadas. Ese mismo año George Gallup acertó con una muestra de unas 50 000 personas bien elegidas. La lección sigue vigente: una muestra pequeña pero representativa vale más que una enorme y sesgada. En España, el Centro de Investigaciones Sociológicas y el Instituto Nacional de Estadística aplican cada mes estas técnicas.

El consejo de Otto

Un intervalo siempre se escribe como $(\text{centro} - E,\ \text{centro} + E)$. Si te dan el intervalo, el centro es la media de los extremos y el error es la mitad de su longitud. Y recuerda: confianza mayor significa intervalo más ancho.