

9. Tamaño de la muestra

En estadística el tamaño de la muestra es el número de sujetos que componen la muestra extraída de una población, necesarios para que los datos obtenidos sean representativos de la población.

9.1 Objetivos de la determinación del tamaño adecuado de una muestra

1. Estimar un parámetro determinado con el nivel de confianza deseado.
2. Detectar una determinada diferencia, si realmente existe, entre los grupos de estudio con un mínimo de garantía.
3. Reducir costes o aumentar la rapidez del estudio.

Por ejemplo, en un estudio de investigación epidemiológico la determinación de un tamaño adecuado de la muestra tendría como objetivo su factibilidad. Así:

1. Si el número de sujetos es insuficiente habría que modificar los criterios de selección, solicitar la colaboración de otros centros o ampliar el período de reclutamiento. Los estudios con tamaños muestrales insuficientes, no son capaces de detectar diferencias entre grupos, llegando a la conclusión errónea de que no existe tal diferencia.
2. Si el número de sujetos es excesivo, el estudio se encarece desde el punto de vista económico y humano. Además es poco ético al someter a más individuos a una intervención que puede ser menos eficaz o incluso perjudicial.

El tamaño de una muestra es el número de individuos que contiene.

Una fórmula muy extendida que orienta sobre el cálculo del tamaño de la muestra para datos globales es la siguiente :

$$\frac{k^2 N p q}{e^2 (N - 1) + k^2 p q}$$

N: es el tamaño de la población o universo (número total de posibles encuestados).

k: es una constante que depende del nivel de confianza que asignemos. El nivel de confianza indica la probabilidad de que los resultados de nuestra investigación sean ciertos: un 95,5 % de confianza es lo mismo que decir que nos podemos equivocar con una probabilidad del 4,5%. Los valores de k se obtienen de la tabla de la distribución normal estándar N(0,1).

Los valores de k más utilizados y sus niveles de confianza son:

Valor de k	1,15	1,28	1,44	1,65	1,96	2,24	2,58
Nivel de confianza	75%	80%	85%	90%	95%	97,5%	99%

(Por tanto si pretendemos obtener un nivel de confianza del 95% necesitamos poner en la fórmula $k=1,96$)

e : es el error muestral deseado, en tanto por uno. El error muestral es la diferencia que puede haber entre el resultado que obtenemos preguntando a una muestra de la población y el que obtendríamos si preguntáramos al total de ella. Ejemplos:

Ejemplo 1: si los resultados de una encuesta dicen que 100 personas comprarían un producto y tenemos un error muestral del 5% comprarán entre 95 y 105 personas.

Ejemplo 2: si hacemos una encuesta de satisfacción a los empleados con un error muestral del 3% y el 60% de los encuestados se muestran satisfechos significa que entre el 57% y el 63% ($60\% \pm 3\%$) del total de los empleados de la empresa lo estarán.

Ejemplo 3: si los resultados de una encuesta electoral indicaran que un partido iba a obtener el 55% de los votos y el error estimado fuera del 3%, se estima que el porcentaje real de votos estará en el intervalo 52-58% ($55\% \pm 3\%$).

p : proporción de individuos que poseen en la población la característica de estudio. Este dato es generalmente desconocido y se suele suponer que $p=q=0.5$ que es la opción más segura.

q : proporción de individuos que no poseen esa característica, es decir, es $1-p$.

n : tamaño de la muestra (número de encuestas que vamos a hacer).

Altos niveles de confianza y bajo margen de error no significan que la encuesta sea de mayor confianza o esté más libre de error necesariamente; antes es preciso minimizar la principal fuente de error que tiene lugar en la recogida de datos. Para calcular el tamaño de la muestra suele utilizarse la siguiente fórmula:

Otra fórmula para calcular el tamaño de la muestra es:

$$n = \frac{N\sigma^2 Z^2}{(N-1)e^2 + \sigma^2 Z^2}$$

Donde: n = el tamaño de la muestra.

N = tamaño de la población.

σ = Desviación estándar de la población, que generalmente cuando no se tiene su valor, suele utilizarse un valor constante de 0,5.

Z = Valor obtenido mediante niveles de confianza. Es un valor constante que, si no se tiene su valor, se lo toma en relación al 95% de confianza equivale a 1,96 (como más usual) o en relación al 99% de confianza equivale 2,58, valor que queda a criterio del encuestador.

e = Límite aceptable de error muestral que, generalmente cuando no se tiene su valor, suele utilizarse un valor que varía entre el 1% (0,01) y 9% (0,09), valor que queda a criterio del encuestador.

La fórmula anterior se obtiene de la fórmula para calcular la estimación del intervalo de confianza para la media:

$$\bar{X} - Z \frac{\sigma}{\sqrt{n}} \sqrt{\frac{(N-n)}{(N-1)}} \leq \mu \leq \bar{X} + Z \frac{\sigma}{\sqrt{n}} \sqrt{\frac{(N-n)}{(N-1)}}$$

En donde el error es:

$$e = Z \frac{\sigma}{\sqrt{n}} \sqrt{\frac{(N-n)}{(N-1)}}$$

Elevando al cuadrado el error se tiene: $[(e)]^2 = (Z \frac{\sigma}{\sqrt{n}} \sqrt{\frac{(N-n)}{(N-1)}})^2$ $e^2 = Z^2 \frac{\sigma^2}{n} \frac{(N-n)}{(N-1)}$

Multiplicando fracciones: $e^2 = \frac{[Z^2 \sigma]^2 (N-n)}{n(N-1)}$

Eliminando denominadores: $e^2 n(N-1) = [Z^2 \sigma]^2 (N-n)$

Eliminando paréntesis: $e^2 nN - e^2 n = [Z^2 \sigma]^2 N - [Z^2 \sigma]^2 n$

Transponiendo n a la izquierda: $e^2 nN - e^2 n + [Z^2 \sigma]^2 n = [Z^2 \sigma]^2 N$

Factor común de n :

$$n(e^2 N - e^2 + Z^2 \sigma^2) = [Z^2 \sigma]^2 N$$

Despejando n:

$$n = \frac{(Z^2 \sigma^2 N)}{(e^2 N - e^2 + Z^2 \sigma^2)}$$

Ordenando se obtiene la fórmula para calcular el tamaño de la muestra:

$$n = \frac{(N \sigma^2 Z^2)}{((N-1) e^2 + \sigma^2 Z^2)}$$

Ejemplo ilustrativo: Calcular el tamaño de la muestra de una población de 500 elementos con un nivel de confianza del 99%

Solución: Se tiene $N=500$, para el 99% de confianza $Z = 2,58$, y como no se tiene los demás valores se tomará $\sigma=0,5$, y $e = 0,05$.

Reemplazando valores en la fórmula se obtiene:

$$n = \frac{(N \sigma^2 Z^2)}{((N-1) e^2 + \sigma^2 Z^2)}$$

$$n = \frac{(500 \cdot [0,5]^2 \cdot [2,58]^2)}{((500-1) \cdot [0,05]^2 + [0,5]^2 \cdot [2,58]^2)} = \frac{832,05}{2,9116} = 285,77 = 286$$

9.2 Estimación de parámetros

La estimación de parámetros consiste en el cálculo aproximado del valor de un parámetro en la población, utilizando la inferencia estadística, a partir de los valores observados en la muestra estudiada. Para el cálculo del tamaño de la muestra en una estimación de parámetros son necesarios los conceptos de Intervalo de confianza, variabilidad del parámetro, error, nivel de confianza, valor crítico y valor α (véase estimación por intervalos).

9.3 Estimación de una proporción

Los datos que tenemos que incluir en la fórmula para calcular el número de sujetos necesarios de la muestra (N) son:

1. $Z_{\alpha/2}$: valor de Z correspondiente al riesgo α fijado. El riesgo α fijado suele ser 0,05 y $Z_{\alpha/2}$ de 1,96.
2. P : Valor de la proporción que se supone existe en la población.
3. i : Precisión con que se desea estimar el parámetro ($2i$ es la amplitud del intervalo de confianza).

9.4 Estimación de una media

Los datos que tenemos que incluir en la fórmula para calcular el número de sujetos necesarios en la muestra (**N**) son:

1. $Z_{\alpha/2}$: valor de Z correspondiente al riesgo α fijado. El riesgo α fijado suele ser 0,05 y $Z_{\alpha/2}$ de 1,96.
2. s^2 : Varianza de la distribución de la variable cuantitativa que se supone que existe en la población.
3. i : Precisión con que se desea estimar el parámetro ($2i$ es la amplitud del intervalo de confianza).

9.5 Contraste de hipótesis

Para conocer el tamaño de la muestra en un estudio de investigación en el que queremos conocer las diferencias existentes entre dos hipótesis, debemos conocer previamente:

- **error tipo I y tipo II**: Hay que establecer el riesgo de cometer un error de tipo I que se está dispuesto a aceptar. Normalmente de forma arbitraria se acepta un riesgo del 5%. Además hay que establecer el riesgo que se acepta de cometer un error tipo II, que suele ser entre el 5 y el 20%.
- **Si la hipótesis es unilateral o bilateral**: El planteamiento de una hipótesis bilateral o "de dos colas" requiere mayor tamaño muestral.
- Definir la **Magnitud de la diferencia** efecto o asociación que se desea detectar: A mayores diferencias preestablecidas en el planteamiento de la hipótesis, menor tamaño muestral, y a menor diferencia, mayor espacio muestral.
- Conocer la variabilidad del criterio de evaluación en la población.

9.6 Comparación de dos proporciones

Para calcular el número de sujetos necesarios en cada una de las muestras (n), debemos prefijar:

- **1,96** = Valor Z correspondiente al riesgo deseado
- **1,96** = Valor Z correspondiente al riesgo deseado, si es de dos colas.
- **0,13** = Valor de la proporción en el grupo de referencia, placebo, control o tratamiento habitual.

- **0,44** = Valor de la proporción en el grupo del nuevo tratamiento, intervención o técnica.
- **0,29** = Media de las dos proporciones p_1 y p_2 .

Coeficiente de correlación[editar]

La asociación entre dos variables cuantitativas necesita normalmente la utilización del coeficiente de correlación r de Pearson.