Tamaño de la muestra

Ya se estudió la estimación de la media y de la proporción mediante intervalos de confianza. Ahora toca detenerse en el tamaño de la muestra que conviene tomar para realizar tales estimaciones.

Muchas veces cuando alguien que no es experto en estadística tiene que tomar una muestra, la pregunta que se hace es “¿de qué tamaño la tomo?” Pero, por sentido común, sabe que quiere que el error en la inferencia sea pequeño. Es decir, desea que el proceso de la inferencia del estadístico al parámetro poblacional sea lo más preciso posible.

De esto se trata esta lectura.

Determinación del tamaño de la muestra para una estimación por intervalo de la media poblacional

Caso Delicity

Delicity es una cadena de casas de té, que opera en el país desde hace 10 años. Ha realizado un estudio de mercado en la ciudad de Córdoba para ver si es conveniente instalarse ahí. Tomó una muestra representativa de locales en alquiler y se encontró que, en promedio, los alquileres estaban en $8500 mensuales con una desviación estándar de $1500.

Ahora, Delicity quiere realizar un nuevo estudio para estimar la media poblacional de los alquileres, pero esta vez, especificó el margen de error que puede aceptar, que es de $500, y el nivel de confianza en la investigación, que es del 95 %. La empresa pide que se determine el tamaño de la muestra para estimar la media poblacional, utilizando la desviación estándar del estudio previo realizado.

Relación entre el tamaño de la muestra y la precisión de los resultados

Se decía en la introducción que queremos que el parámetro poblacional estimado sea lo más preciso posible. Por supuesto que va a existir un error por estimación y un error que es el que el investigador está dispuesto a aceptar al no estudiar la población completa.

Pero ¿cuál es el tamaño adecuado de la muestra? Si es demasiado pequeña se podría fracasar en el análisis y si es grande, la posibilidad de aproximación al parámetro poblacional será más precisa, pero, seguramente, se estarían desperdiciando recursos.

El error de muestreo se puede controlar con la selección adecuada del número de elementos para la muestra. A continuación, se darán pautas para definir el tamaño de la muestra,

- para estimar la media poblacional y
- para estimar la proporción poblacional.

¿Cómo obtener el tamaño de la muestra para obtener un margen de error deseado?

Ya se expuso, en lecturas anteriores, la relación del tamaño de la muestra con el intervalo de confianza para la media y la proporción. Se sabe cómo influye y qué relación tiene con el resto de los valores del intervalo de confianza.

Pero el enfoque que se propone en esta lectura es el de establecer el número mínimo de unidades que tiene que tener la muestra, si establecemos de antemano el error que el investigador está dispuesto a aceptar, así como el nivel de confianza que se requiere.

Determinación del tamaño de la muestra para la media poblacional

Ejemplo 1

Una empresa petrolera está haciendo una investigación sobre el precio promedio de venta al público de la nafta súper en todo el país, durante el primer trimestre del año pasado. Los datos históricos registran que la distribución de precios es, aproximadamente, normal y que la desviación estándar de la población (todas las empresas que operan en el país tanto a nivel nacional como provincial), durante ese lapso, es de $12. ¿Qué tamaño tiene que tener la muestra si la petrolera quiere estimar la media del precio al público de la nafta súper, vendida en ese trimestre, sabiendo que está dispuesta a aceptar un margen de error de ± $ 10, con un nivel de confianza del 95 %?

En otras palabras, la petrolera desea tomar una muestra para estimar la media poblacional.

Para esto, se calculará la media de la muestra $\bar{x}$ y se la utilizará como estimador puntual de la media poblacional. Pero la empresa quiere tener una certeza del 95% de que esa media (precio promedio) no exceda los ± $ 10 de la estimación puntual (es decir, que esté por arriba o por debajo de la estimación puntual.

Resolución del ejemplo 1

Se sabe que:

- el límite inferior de confianza está dado por: $\bar{x}-z.\sigma_{\bar{x}}$ y que
- el límite superior de confianza está dado por: $\bar{x}+z.\sigma_{\bar{x}}$.

Además, a $z.\sigma_{\bar{x}}$ se lo denomina error permitido (E) o margen de error, que es el error que el investigador está dispuesto a aceptar.

Por lo tanto, $E=z.\sigma_{\bar{x}}$.

Pero el error esperado, en el ejemplo 1 es ± $ 10. Es decir, la petrolera está determinando que $z.\sigma_{\bar{x}}=\$10$

Si se busca en la tabla de la normal, tal como se vio en los intervalos de confianza para la media (en la lectura anterior), para una confianza del 95%, el valor debe ser z = 1,96. Entonces se sustituye:

$$1,96.\sigma_{\bar{x}}=\$10$$

Se despeja la desviación estándar de la distribución de medias muéstrales $\sigma_{\bar{x}}$:

$$\sigma_{\bar{x}} = \frac{\$\ 10}{1,96} = \$\ 5,1$$

que también llamamos error estándar de la media.

Además, se sabe que el error estándar de la media está inversamente relacionado con el tamaño de la muestra, pues por definición

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}\,.$$

Siendo σ la desviación estándar poblacional, que en este caso se tiene como dato. Entonces, se sustituye y sólo queda despejar *n*, que es lo que se quiere saber:

$$\$\,5{,}1 = \frac{\$12}{\sqrt{n}}$$

$$\sqrt{n} = \frac{12}{5{,}1}$$

$$\sqrt{n} = 2{,}353$$

$$n = (2{,}353)^2 = 5{,}54 \cong 6 \text{ (siempre se redondea hacia arriba).}$$

De esta manera, el tamaño de la muestra para la precisión especificada de antemano es, como mínimo, de 6 empresas. O, lo que es lo mismo, la empresa petrolera debe tomar una muestra de 6 empresas que comercializan en el país para obtener la precisión deseada en la estimación del precio promedio de la nafta súper en el primer trimestre del año pasado.

Hay que recordar que el tamaño de la muestra es igual a 6 empresas o que debe ser mayor o igual a 6 empresas. Lo importante es que el tamaño de la muestra obtenida cumpla con los requisitos que se propuso el investigador. Después se podrá hacer variar el número para ver cómo varía el error esperado y si el investigador está dispuesto a aceptarlo. A menor error permitido *E*, el tamaño de la muestra aumenta y viceversa.

Generalización de la expresión matemática del tamaño de la muestra para la media

Caso de σ conocida:

- El objetivo es determinar el tamaño de la muestra, sabiendo que existe un error que se acepta de antemano.

- El punto de partida es el mismo que se usa para resolver el ejemplo 1: la expresión que define un intervalo de confianza.

- Si se tiene el valor de σ, es posible encontrar el tamaño de la muestra n necesario para proporcionar cualquier margen de error deseado. A continuación, se presenta la deducción de la fórmula. En general:

$$\bar{x} \pm z_{\alpha/2} \cdot \sigma_{\bar{x}}$$

(El tema de escribir como subíndice de z, fue explicado en la lectura anterior, por lo que se aconseja remitirse a la lectura 3 y repasarlo para tenerlo presente).

La cantidad

$$z_{\alpha/2} \cdot \sigma_{\bar{x}}$$

se llama margen de error o error permitido *E*.

Entonces

$$E = z_{\frac{\alpha}{2}} . \sigma_{\bar{x}} .$$

Pero se sabe que el error estándar es

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}},$$

sustituyendo en la ecuación inmediatamente anterior:

$$E = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} .$$

Se despeja *n*

$$E . \sqrt{n} = z_{\alpha/2} . \sigma$$

$$\sqrt{n} = \frac{z_{\alpha/2} . \sigma}{E} .$$

Se elevan ambos miembros de la ecuación al cuadrado y se distribuye la potencia:

$$n = \frac{z_{\alpha/2}{}^{2}.\sigma^{2}}{E^{2}}.$$

- Aclaraciones:

1) Da lo mismo, por practicidad y según las características del problema, calcular primero

$$\frac{z_{\alpha/2}.\sigma}{E}$$

2) La ecuación de tamaño de la muestra se usa para determinar el tamaño de muestra adecuado. Pero la opinión del investigador es muy importante para determinar si el tamaño de la muestra final debe ser mayor.

3) n es inversamente proporcional al margen de error E.

Refuerzo de los conceptos, analizando brevemente la fórmula del tamaño de la muestra

Figura 1: Fórmula del tamaño de la muestra para una estimación por intervalo de la media poblacional

$$n = \frac{z_{\alpha/2}{}^{2} \cdot \sigma^{2}}{E^{2}}$$

$$n = \frac{z_{\alpha/2}{}^{2} \cdot \sigma^{2}}{E^{2}}$$

n:

tamaño de la muestra que proporciona el margen de error deseado al nivel de confianza elegido.

$$n = \frac{{z_{\alpha/2}}^2 \cdot \sigma^2}{E^2}$$

σ:

valor de la desviación estándar poblacional.

$$n = \frac{z_{\alpha/2}{}^{2} \cdot \sigma^{2}}{E^{2}}$$

z_(α/2)

valor de z, consecuencia directa del nivel de confianza que se va a usar para calcular la estimación por intervalo.

$$n = \frac{{z_{\alpha/2}}^2 . \sigma^2}{E^2}$$

E:

margen de error o error permitido que el investigador está dispuesto a aceptar.

¿Cómo se calcula n, si no se posee la desviación estándar poblacional σ?

Caso de σ desconocido

Para calcular el tamaño de la muestra, se necesita el valor de la desviación estándar de la población. Ésta debe especificarse antes de determinar el tamaño de la muestra, pues es requisito para poder utilizar la fórmula.

Seguidamente, se exponen algunos métodos para obtener el valor de σ.

1. Se puede obtener un valor de σ de estimaciones calculadas a partir de estudios anteriores.
2. Se puede tomar la desviación estándar muestral como estimador de la poblacional. Es aconsejable seleccionar una muestra preliminar y utilizarla luego como estimador.
3. Se puede calcular con una lógica que se basa en el rango de los datos de la población. Se sabe que la media más y menos 3 desviaciones estándar incluyen el 99,7% del área total bajo la curva normal

(regla empírica). Por lo tanto, el rango aproximadamente abarca 6 desviaciones estándar. Para una estimación un poco burda de la desviación estándar, podría ser el rango de la población dividido 6. No es una estimación precisa, pero puede funcionar si la distribución es normal.

Resolución del caso Delicity

Datos:

- $\hat{\sigma}$: estimador de σ = $1500
- E = $500.
- Para un 95% de confianza, z = 1,96.

$$n = \frac{{z_{\alpha/2}}^2 . \sigma^2}{E^2} = \left(\frac{1{,}96 \times 1500}{500}\right)^2 = 34{,}57$$

El tamaño de la muestra para estimar la media poblacional, de acuerdo a las especificaciones de la empresa, es de 37 locales.

¿Qué tan grande debe ser la selección de una muestra para tener un intervalo de confianza de 95 % con un margen de error de 10, suponiendo que la desviación estándar poblacional es 40?

- [ ] $n \geq 62$.
- [ ] La muestra tiene que ser de por lo menos 62 elementos.
- [ ] $n = 62$.
- [ ] $n \leq 61$.
- [ ] $n \leq 62$.

SUBMIT

Determinación el tamaño de la muestra para estimar una proporción de una población

Caso Seguros Kevin S.A.

La compañía de seguros Kevin S.A. desea hacer un estudio de mercado en la provincia de Chubut para saber qué proporción de la población está interesada en comprar un tipo de seguro para la vivienda. Con este fin, quiere realizar un muestreo significativo en dicha provincia. Como no cuenta con datos previos de muestras anteriores, ni con datos de la población, sólo establece el margen de error o error permitido que está dispuesto a aceptar, que es ± 10% clientes con un nivel de confianza del 99%. Los directivos de Kevin S.A. consultan sobre cuál debe ser el tamaño de la muestra más adecuado para realizar este estudio.

¿Cómo obtener el tamaño de la muestra para obtener un margen de error deseado?

Se estudió cómo calcular el tamaño de la muestra para estimar la media poblacional. Del mismo modo, se calcula el tamaño de la muestra para estimar la proporción poblacional, con un error permitido y un nivel de confianza dado. Es decir, el enfoque que se propone en esta lectura es establecer el número mínimo de unidades que tiene que tener la muestra para estimar la proporción poblacional, si se establece de antemano el error que el investigador está dispuesto a aceptar y el nivel de confianza que se requiere.

Una aclaración importante para cuando se quiere estimar una proporción poblacional, es que, por lo general, el margen de error deseado es de 0,1 o menor. Hay que tener en cuenta que se trata de una proporción y si se establece de 0,1, se permite un error del 10%.

Muchas encuestadoras admiten un error de 0,03 o 0,04, que es lo más común. Con estos márgenes de error, se asegura que la muestra calculada cumpla con las condiciones estudiadas para aproximar por la normal a la distribución binomial: $n.p \geq 5$ y $n.q \geq 5$, en este caso, se traduce en usar una distribución normal como una aproximación de la distribución muestral $\bar{x}$.

Determinación del tamaño de la muestra para la proporción poblacional

Hay varias situaciones que pueden presentarse al momento de calcular el tamaño de la muestra para estimar la proporción poblacional. Se explican, a continuación, los dos casos más comunes.

Ejemplo 2: se cuenta con una estimación de la proporción poblacional

Un candidato a presidente sabe, por estudios previos realizados sobre una muestra, que lo votaría un 65% de los encuestados. Determina el tamaño de la muestra para que en un nuevo estudio, que se está por realizar sobre la estimación de la proporción poblacional que lo votaría, tenga un error no mayor a ± 0,04, con un 90% de certeza.

Resolución del ejemplo 2

Se parte de la fórmula del error permitido o margen de error en la estimación de intervalos de confianza para la proporción: $\hat{p}$ y $\hat{q}$ son los estimadores de p y q

$$E = z_{\alpha/2} \cdot \sqrt{\frac{\hat{p} \cdot \hat{q}}{n}}$$

Se sustituye según los datos dados:

$$0{,}04 = 1{,}64 \cdot \sqrt{\frac{0{,}65 \times 0{,}35}{n}}$$

$$\frac{0{,}04}{1{,}64} = \sqrt{\frac{0{,}2275}{n}}$$

Se elevan ambos miembros al cuadrado:

$$(0,024)^2 = \frac{0,2275}{n}$$

Ahora hay que despejar n:

$$(0,024)^2 = \frac{0,2275}{n}$$

Para estimar la proporción poblacional con un 90 % de certeza y un ± 0,04 de margen de error, debe tomarse una muestra de 395 personas.

Ejemplo 3: no se cuenta con ninguna información de la proporción poblacional

Un candidato desea saber de qué tamaño debe tener una muestra para estimar la proporción poblacional que lo votaría. Se solicita a los investigadores que el margen de error no sea mayor del ± 0,04, con un 90% de certeza.

Resolución del ejemplo 3

Se inicia la resolución igual que en el ejemplo 2:

$$E = z_{\alpha/2} . \sqrt{\frac{p.q}{n}}$$

$$0{,}04 = 1{,}64 . \sqrt{\frac{p.q}{n}}$$

$$\frac{0{,}04}{1{,}64} = \sqrt{\frac{p.q}{n}}$$

$$(0{,}024)^2 = \frac{p.q}{n}$$

$$n = \frac{p.q}{0{,}024^2}$$

Era esperado este problema, ya que se quieres estimar el valor poblacional de *p* mediante un tamaño de muestra que se está buscando.

Faltan datos para calcular *n*, se necesita una estimación de los parámetros *p* y *q* de la población. Si se tiene una buena idea de la proporción real de votantes del candidato, se puede considerar la mejor estimación de *p* para calcular *n*.

Pero si no se tiene idea del valor de *p*, entonces la mejor estrategia es darle un valor de una forma conservadora: que *n* sea lo suficientemente grande para dar, al menos, la precisión que se necesita sin importar el verdadero valor de *p*.

La manera de obtener la *n* más grande es generando el numerador más grande posible de la expresión a la que se arriba, esto sucede cuando elegimos $p = 0{,}5$ y $q = 0{,}5$.

La diferencia con el resto de las opciones de *p* y *q* cercanas a 0,5, no son significativas, pero esta es la más segura, si se desconoce la proporción poblacional. Entonces, se toma $p = q = 0{,}5$:

$$n = \frac{0{,}5 \times 0{,}5}{0{,}024^2} = 434$$

Para estimar la proporción poblacional con un 90 % de certeza y un ± 0,04 de margen de error, sin tener datos previos de ésta, debe tomarse una muestra de 434 personas. Es decir, para tener una seguridad del 90 % de que se estima la proporción verdadera dentro de 0,04, se debe escoger una muestra aleatoria simple de 434 votantes para entrevistar.

Se observa que la diferencia con el ejemplo 2 no es grande, esto se mantiene si le damos a *p* un valor entre 0,3 y 0,7.

En conclusión ¿cómo se estima $\hat{p}$, si no se lo tiene?

En la práctica, el valor estimado de $\hat{p}$ se determina mediante algunos de los siguientes métodos:

1) Utilizar la proporción poblacional de una muestra previa de las mismas unidades o de unidades similares.

2) Utilizar un estudio piloto y elegir una muestra preliminar. La proporción muestral de esta muestra se usa como valor estimado de *p*.

3) Emplear como valor estimado de *p* = 0,5, que es la forma más conservadora, como se acaba de indicar.

Generalización de la expresión matemática del tamaño de la muestra para la proporción

De la fórmula del margen de error, se despeja *n* (no hay que olvidar que $\sigma_{\bar{p}}$ es la desviación estándar de la distribución de proporciones muestrales):

$$E = z_{(\alpha/2)} . \sigma_{\bar{p}}$$

$$E = z_{\alpha/2} . \sqrt{\frac{\hat{p} . \hat{q}}{n}}$$

Se elevan ambos miembros al cuadrado:

$$E^2 = {z_{\alpha/2}}^2 . \frac{\hat{p} . \hat{q}}{n}$$

$$n = {z_{\alpha/2}}^2 . \frac{\hat{p} . \hat{q}}{E^2}$$

- z: valor que corresponde a un determinado nivel de confianza.
- $\hat{p}$: estimador de la proporción poblacional de éxitos.
- q:̂ estimador de la proporción poblacional de fracasos.
- *E*: margen de error o error permitido.

Resolución Caso Seguros Kevin S.A.

Datos:

- E = ± 0,1
- 99 % nivel de confianza.

- Z = 2,58 desviaciones estándar.

Se debe determinar el tamaño de muestra más adecuado, que cumpla con los criterios anteriores. Se parte de la fórmula del error permitido despejamos *n*:

$$n = \frac{{z_{\alpha/2}}^2 . \hat{p} . \hat{q}}{E^2}$$

Como no se poseen los datos previos de p y q s estiman. Los valores de p y q que dan el tamaño máximo de la muestra, de manera tal que se cumplan los criterios de error y confianza, son $\hat{p}=\hat{q}=0,5$

Procede a sustitución en la fórmula de *n*:

$$n = \frac{2,58^2 \times 0,5 \times 0,5}{0,1^2} = 166,41$$

Para tener una seguridad del 99% de que se estimó la proporción verdadera, dentro de un margen de error del 10%, se debe escoger una muestra aleatoria simple de 167 potenciales clientes.

Este comentario que hacen los autores del texto básico, en el que relacionan el tamaño de la muestra con la desviación estándar resulta muy práctico e interesante para terminar la presente exposición:

Desde una perspectiva de sentido común, si la desviación estándar de una población es muy pequeña, los valores se agrupan muy cerca de la media y casi cualquier tamaño de muestra los captará y producirá información precisa. Por otro lado, si la desviación estándar de la población es muy grande y los valores están bastante dispersos, será necesaria una muestra muy grande para incluirlos y obtener información correcta. ¿Cómo puede tenerse una idea de la desviación estándar de la población antes de iniciar el muestreo? Las compañías que planean realizar estudios de mercado casi siempre hacen una investigación preliminar de la población para estimar la desviación estándar. Si el producto se parece a otro que ha estado en el mercado, a menudo es posible apoyarse en los datos anteriores acerca de la población sin más estimaciones. (Levin y Rubín, 2004, p. 307).

Para estimar la proporción poblacional con un margen de error determinado y una certeza dada, el error permitido se hace cada vez más grande a medida que aumenta el tamaño de la muestra.

- Es verdadero porque al permitir un error más grande la muestra se hace grande también.
- Es falso porque al permitir un error más grande la muestra se hace cada vez más pequeña.

SUBMIT

LECCIÓN 3 de 4

Revisión del módulo

Hasta acá aprendimos

Muestreo

En esta lectura has considerado la necesidad del muestreo, sus ventajas y desventajas. Así también a definir, qué es una muestra representativa y los tipos de errores que se cometen en el proceso de muestro. Has aprendido a clasificar las muestras en probabilísticas y no probabilísticas. A sub clasificar y definir los tipos de muestreos que más se usan para poder elegir el que más se adapte a cada situación.

Distribución de muestreo

En esta lectura se abordó el concepto de distribuciones de muestreo: para la media y para la proporción. Se diferenció entre poblaciones finitas e infinitas. Aprendiste a calcular el error estándar y el error estimado en distintos casos concretos. Se abordó el teorema del Límite Central, en el que es importante que comprendas las conclusiones a las que se arribaron, pues son la base para inferir los parámetros de la población a partir de estadísticos muestrales correspondientes.

Estimación

En esta lectura se introdujo el concepto de estimación, en el cuál se cristaliza la aplicación del Teorema del Límite Central. Diferenciaste una estimación puntual de otra mediante intervalos. Calculaste intervalos de confianza para la media y para la proporción de poblaciones finitas e infinitas. Todo a través de una muestra pequeña o grande, poniendo atención en cada una de las fórmulas que se utiliza en cada caso.

Tamaño de la muestra

En esta lectura aprendiste a calcular el tamaño aproximado que debe tener una muestra para que sea representativa en una investigación. También diferenciaste entre obtener un tamaño de la muestra para estimar la media y para la proporción poblacionales. Identificaste los estadísticos necesarios que intervienen y de qué forma se relacionan entre sí.

Referencias

Levin, R. y Rubin, D. (2004). *Estadística para Administración y Economía* (7ª edición). México: Pearson.