

Intervalos de confianza

Àngel J. Gil Estallo

P08/75057/02307

Índice

Sesión 1

Introducción a los intervalos de confianza.

El caso de la media aritmética	5
1. El concepto de intervalo de confianza	5
2. Intervalo de confianza para la media aritmética cuando la población es normal y conocemos la desviación típica	9
2.1. El efecto del tamaño de la muestra	10
2.2. Consideraciones sobre la normalidad	11
3. Intervalo de confianza para la media cuando la población es normal y desconocemos la desviación típica	11
3.1. El tamaño de la muestra	13
4. Comparación entre los casos estudiados	14
5. Resumen.....	15
Ejercicios	16

Sesión 2

Intervalo de confianza para la proporción	20
1. Procedimiento para construir un intervalo de confianza para la proporción	21
2. El efecto del tamaño de la muestra.....	22
3. Resumen.....	24
Ejercicios	25

Introducción a los intervalos de confianza. El caso de la media aritmética

La inferencia estadística proporciona métodos para obtener conclusiones a partir de un conjunto de datos. La mayoría de las veces no tendremos una certeza absoluta de las conclusiones a las que llegamos. La teoría de la probabilidad fundamentará las conclusiones obtenidas y permitirá establecer la precisión de los métodos utilizados. A continuación trabajaremos los llamados *intervalos de confianza*. Comenzaremos por el caso más sencillo, a partir del cual introduciremos las definiciones generales.

1. El concepto de intervalo de confianza

A partir de un caso concreto iremos introduciendo gradualmente las ideas y las técnicas que sustentan la construcción de los llamados *intervalos de confianza*. En concreto nos planteamos estudiar la media de las alturas de los estudiantes de la UOC. Veremos paso a paso cuál sería el procedimiento que seguiríamos:

1) Establecemos algunas *hipótesis previas*, que son las que determinan las distribuciones que hay que utilizar en la construcción del intervalo. Estas hipótesis previas permitirán, en definitiva, utilizar resultados de la teoría de la probabilidad.

En este ejemplo utilizaremos dos hipótesis:

a) La distribución de las alturas sigue una ley normal, que tendrá una media μ (que supondremos desconocida) y una desviación típica σ .

b) Gracias a estudios anteriores conocemos el valor de la desviación típica poblacional σ ; supongamos que tenemos $\sigma = 10$ cm.

2) A continuación efectuamos la recogida de los datos adecuados al problema. Normalmente seleccionaremos una muestra aleatoria simple de la población y obtendremos los datos requeridos a partir de los individuos de la muestra.

Supongamos que obtenemos una muestra aleatoria simple de 121 alumnos de la UOC, a quienes preguntamos su altura.

3) A partir de los datos obtenidos, calculamos resúmenes numéricos adecuados.

Puesto que estamos interesados en la media de las alturas, calculamos la media de estas 121 observaciones. Supongamos que obtenemos que esta media es $\bar{x} = 171$ cm.

La utilidad de las muestras

En la inmensa mayoría de los casos no podemos acceder a los datos de toda la población, bien porque es inviable, bien porque resulta demasiado caro.

En este caso la media poblacional μ es un **parámetro** de la población que queremos describir y la media muestral \bar{x} es un **estadístico** que nos permite aproximar el valor del parámetro.

4) Utilizamos la teoría de la probabilidad para obtener una relación entre el parámetro y el estadístico que permita construir cierto intervalo, llamado *intervalo de confianza*. Este intervalo debe tener las propiedades siguientes:

a) Debe estar centrado en el valor del estadístico (que es el valor que, efectivamente, conocemos).

b) Ya que podemos obtener muestras diferentes de la población, y cada muestra dará lugar a una media muestral diferente, debemos garantizar, para un porcentaje elevado de las posibles muestras, que el procedimiento produzca intervalos que contengan el auténtico valor del parámetro.

La confianza en el intervalo

Debemos medir de alguna manera la confianza que podemos tener en el intervalo.

Este porcentaje de muestras que dan lugar a intervalos que contienen el auténtico valor del parámetro es el llamado **nivel de confianza**.

Así pues, un **intervalo de confianza** para cierto parámetro con un nivel de confianza de $C\%$ es un intervalo calculado a partir de una muestra de manera que el procedimiento de cálculo garantiza que el $C\%$ de las muestras dé lugar a un intervalo que contenga el valor real del parámetro.

Veremos cómo se puede obtener un intervalo con estas características para el ejemplo de las alturas de los estudiantes de la UOC.

Ahora nos preguntamos qué podemos decir de la media aritmética de las alturas de todos los estudiantes de la UOC. Observemos que esta media es la media poblacional μ , ya que suponemos que la distribución de las alturas es normal con media precisamente μ . También sabemos que, salvo que preguntemos a todos y cada uno de los estudiantes de la UOC cuál es su altura, nunca conseguiremos saber cuál es el verdadero valor de la media poblacional μ .

La pregunta es, pues: ¿qué conclusiones podemos extraer sobre el valor de la media poblacional μ , a partir de los datos de que disponemos, y en concreto a partir de la media muestral $\bar{x} = 171$ cm? Es decir, nos preguntamos qué relación podemos establecer entre la verdadera media poblacional μ y la media muestral \bar{x} . Observad que no podemos decir que sean iguales, ni que una sea mayor que la otra. Además, la media muestral depende de la muestra, lo que significa que muestras diferentes proporcionan resultados diferentes.

Deberemos recurrir a nuestros conocimientos de probabilidad; en concreto, utilizaremos que para la media de las alturas de la UOC, la variable:

$$\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{\bar{X} - \mu}{\frac{10}{\sqrt{121}}}$$

siga una distribución normal estándar. Esta expresión permite establecer una relación indirecta entre μ y \bar{x} . Aprovecharemos este hecho para construir el intervalo de confianza siguiendo su definición.

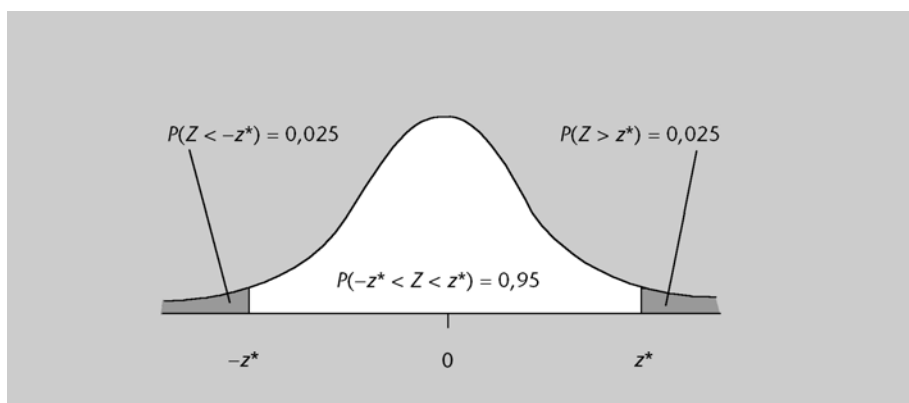
Supongamos que queremos un nivel de confianza del 95%. Comenzaremos por construir un intervalo de la forma $(-z^*, z^*)$ centrado en el valor 0, de manera que los valores que toma la variable aleatoria Z :

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \approx N(0,1)$$

pertenezcan a este intervalo con una probabilidad del 0,95; es decir, buscamos unos valores $-z^*$ y z^* para los cuales:

$$P\left(-z^* \leq \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq z^*\right) = 0,95$$

En estas condiciones, y tal como se ve en el gráfico siguiente:



el valor de z^* es aquel valor que $P(Z \geq z^*) = (1 - 0,95)/2 = 0,025$ (ya que Z sigue una ley normal estándar) y, por tanto, tenemos que $z^* = 1,96$ (este valor se puede obtener a partir de las tablas de la distribución normal estándar); es decir:

$$P\left(-1,96 \leq \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq 1,96\right) = 0,95$$

Ahora, operando esta expresión con el objetivo de aislar el valor de μ :

$$P\left(-\bar{X} - 1,96 \frac{\sigma}{\sqrt{n}} \leq -\mu \leq -\bar{X} + 1,96 \frac{\sigma}{\sqrt{n}}\right) = 0,95$$

y, finalmente:

$$P\left(\bar{X} - 1,96 \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 1,96 \frac{\sigma}{\sqrt{n}}\right) = 0,95.$$

Haced memoria

Recordad que la expresión $-a \leq b$ es equivalente a esta otra: $-b \leq a$.

Este resultado nos indica que la probabilidad de que la media poblacional μ pertenezca a un intervalo de la forma:

$$\left(\bar{X} - 1,96 \frac{\sigma}{\sqrt{n}}, \bar{X} + 1,96 \frac{\sigma}{\sqrt{n}}\right)$$

es de 0,95; o lo que es lo mismo: noventa y cinco de cada cien veces que escogemos una muestra aleatoria simple y calculamos el valor de la media muestral, el intervalo que obtendremos sustituyendo el valor de \bar{X} por la media correspondiente a la muestra de la que disponemos contendrá el verdadero valor de μ .

Finalmente, lo único que tenemos que hacer es sustituir \bar{X} , σ y n por los valores correspondientes, y recordar que trabajamos con una confianza del 95%. Así pues, haciendo la sustitución, obtenemos que el intervalo:

$$\left(171 - 1,96 \frac{10}{\sqrt{121}}, 171 + 1,96 \frac{10}{\sqrt{121}}\right) = (169,22, 172,78)$$

es un intervalo de confianza con un nivel de confianza del 95% para la media de las alturas de los alumnos de la UOC.

Antes de continuar, es importante remarcar algunos hechos:

- a) El intervalo está centrado en el valor de la media de la muestra.
- b) No sabemos si el intervalo contiene o no la media poblacional μ y no hay manera de saberlo, salvo que conozcamos el valor de μ (y si ya lo conocemos, no es preciso que nos esforcemos en buscar su intervalo de confianza).
- c) Para cada muestra obtenida de la población tenemos un valor de la media muestral y, por tanto, un intervalo de confianza que puede ser diferente del que hemos obtenido.
- d) La expresión “confianza del 95%” indica “confianza en el método” utilizado, de manera que el 95% de las veces que apliquemos el método a la misma población obtendremos intervalos que sí contienen la media poblacional μ .

No hay que confundir

Decir que μ pertenece al intervalo con una probabilidad del 0,95 es incorrecto, ya que la μ pertenece a este intervalo (con lo cual la probabilidad es uno) o no pertenece (con lo cual la probabilidad de que pertenezca es cero).

Resumimos a continuación las ideas y procedimientos que nos han llevado a la construcción de nuestro intervalo de confianza. Evidentemente, la construcción depende de las hipótesis de partida y, como veremos al final de la sesión, diferentes hipótesis dan lugar a diferentes tipos de intervalos.

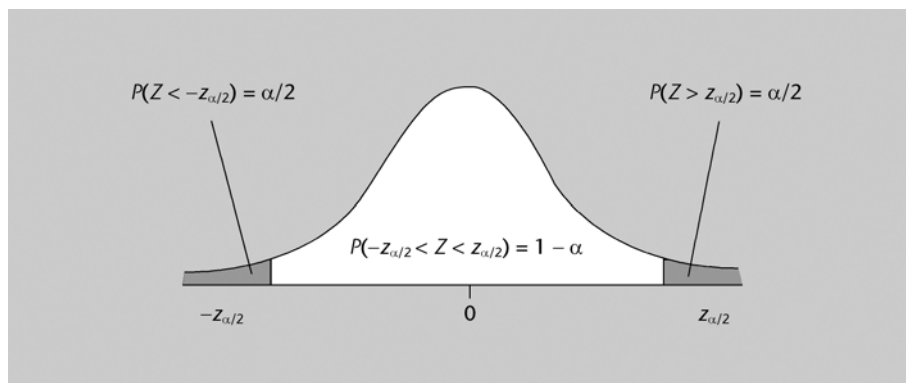
2. Intervalo de confianza para la media aritmética cuando la población es normal y conocemos la desviación típica

Supongamos que la variable que queremos estudiar sigue una ley normal de media μ (desconocida) y desviación típica σ conocida y que disponemos de una muestra aleatoria simple de tamaño n y del valor de la media de la muestra \bar{x} . Entonces:

a) Fijamos el nivel de confianza (en forma de porcentaje), que habitualmente escribiremos como $(1 - \alpha)$.

b) Calculamos el error estándar de la media $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$.

c) Calculamos el llamado *valor crítico*, que es aquel punto $z_{\alpha/2}$ tal que $P(Z \geq z_{\alpha/2}) = \alpha/2$, donde Z es una variable $N(0, 1)$. Gráficamente:



Notación

El nivel de confianza también se denota por $(1 - \alpha)100\%$; normalmente consideraremos $(1 - \alpha)$ igual a 90%, 95% o 99%.

Notación del valor crítico

En el ejemplo de las alturas de los estudiantes de la UOC se ha utilizado la notación z^* , que ahora es sustituida por $z_{\alpha/2}$ para poder precisar el valor crítico según el nivel de confianza.

Para los niveles de confianza habituales, los valores críticos correspondientes son los siguientes:

- $(1 - \alpha) = 90\% = 0,9$, $\alpha = 0,1$ y $z_{\alpha/2} = z_{0,05} = 1,645$
- $(1 - \alpha) = 95\% = 0,95$, $\alpha = 0,05$ y $z_{\alpha/2} = z_{0,025} = 1,96$
- $(1 - \alpha) = 99\% = 0,99$, $\alpha = 0,01$ y $z_{\alpha/2} = z_{0,005} = 2,575$

d) Calculemos el llamado **margen de error** (también llamado **precisión de la estimación**) como $z_{\alpha/2}$ para el error estándar, es decir:

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

e) El intervalo de confianza obtenido con la muestra de partida es el siguiente:

$$(\bar{x} \pm z_{\alpha/2} \sigma_{\bar{x}}) = \left(\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right)$$

Observación

Por tanto, el margen de error es la mitad de la longitud del intervalo de confianza.

Mientras que $(1 - \alpha)$ es el nivel de confianza, α es el llamado **nivel de significación** y se corresponde a la proporción de muestras a partir de las cuales el intervalo construido según el procedimiento explicado no contiene el auténtico valor del parámetro que se quiere aproximar.

2.1. El efecto del tamaño de la muestra

En muchas ocasiones, una vez fijado el nivel de confianza, nos marcaremos como objetivo dar el valor del parámetro μ con cierta precisión. La única manera de obtener la precisión deseada consiste en modificar de forma adecuada el tamaño de la muestra. Supongamos que deseamos una precisión o margen de error ME ; puesto que sabemos que:

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

aislando n obtenemos:

$$\text{Tamaño de la muestra} = n \geq (z_{\alpha/2})^2 \frac{\sigma^2}{ME^2}$$

Tamaño de la muestra

Es fácil ver que si queremos reducir el ancho del intervalo de confianza a la mitad, deberemos tomar una muestra cuatro veces mayor.

Tiempo de conexión al campus virtual

En un estudio llevado a cabo en la UOC se toma una muestra aleatoria de 150 estudiantes y se les pregunta cuánto tiempo estuvieron conectados al campus virtual durante el mes de abril de 2000. Se obtiene una media muestral de 120 minutos. Supongamos, además, que el tiempo de conexión al campus virtual durante el mes de abril de 2000 sigue una distribución normal con desviación típica de diez minutos. Podemos calcular un intervalo de confianza del 95% para el tiempo de conexión durante este mes, considerando la media muestral ($\bar{x} = 120$) y el error estándar de la media, que es:

$$\sigma_{\bar{x}} = \frac{10}{\sqrt{150}} = 0,816$$

Por tanto, el intervalo de confianza con un nivel de confianza del 95% es $(120 \pm 1,96 \cdot 0,816) = (120 \pm 1,59936) = (118,40, 121,60)$. Si queremos que la precisión de nuestro intervalo sea de cinco puntos porcentuales, deberemos conseguir un margen de error inferior a $5\% = 0,05$. Por tanto:

$$1,96 \sigma_{\bar{x}} = 1,96 \frac{\sigma}{\sqrt{n}} = 1,96 \frac{10}{\sqrt{n}} < 0,05$$

y aislando n obtenemos $n > 153.664$ y, por tanto, necesitaríamos una muestra inabarcable, ya que a día de hoy la UOC no tiene tantos estudiantes.

2.2. Consideraciones sobre la normalidad

El procedimiento que se ha presentado resulta válido para variables que siguen leyes normales de media μ , ya que en este caso la variable siguiente sigue una ley normal estándar:

$$\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Por otro lado, el teorema del límite central afirma que, dada cualquier variable aleatoria X con media μ , si el tamaño de las muestras consideradas es $n > 30$, entonces la variable siguiente también se comporta como una distribución normal estándar:

$$\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Consecuencia del teorema del límite central

El único caso en el que no sabremos calcular intervalos de confianza para la media (suponiendo conocida la desviación típica) será cuando las poblaciones sean no normales de medida menor que treinta.

Este hecho hace que el procedimiento de cálculo de un intervalo de confianza para la media sea válido, aunque la variable que hay que estudiar no sea normal, siempre a condición de que el tamaño de la muestra sea superior a treinta.

3. Intervalo de confianza para la media cuando la población es normal y desconocemos la desviación típica

En este caso procederemos tal como se hace cuando se estudia la distribución de la media muestral cuando la desviación típica es desconocida: estimaremos la desviación típica utilizando los valores muestrales y trabajaremos con la distribución de la media muestral \bar{X} , ya que, por un procedimiento parecido a la estandarización, podemos relacionarla con otra variable que sigue una distribución de Student. A continuación, repetiremos el caso de las alturas de los alumnos de la UOC para que podáis comparar ambos casos:

Media de las alturas de los estudiantes de la UOC si la desviación típica es desconocida

- 1) La hipótesis previa será la siguiente: la distribución de las alturas sigue una ley normal, que tendrá cierta media μ y cierta desviación típica σ , ambas desconocidas.
- 2) La recogida de datos consiste en seleccionar una muestra aleatoria simple en la población. Supongamos que obtenemos una muestra aleatoria simple de 121 individuos, a los que preguntamos su altura.
- 3) En el apartado de cálculos ahora necesitamos:
 - a) La media de estas 121 observaciones. Supongamos que obtenemos $\bar{x} = 171$ cm.
 - b) La desviación típica de estas observaciones:

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{121} (x_i - \bar{x})^2}$$

Supongamos que en nuestro ejemplo obtenemos $s = 11$ cm.

4) Fijamos un nivel de confianza; supongamos que interesa un nivel de $(1 - \alpha) = 95\%$. Para calcular el intervalo de confianza, en este caso debemos trabajar con la variable aleatoria siguiente:

$$\frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}}$$

Como sabemos, esta variable sigue una distribución t_{n-1} de Student con $n - 1$ grados de libertad. Ahora necesitamos un intervalo de la forma $(-t^*, t^*)$, centrado también en 0, ya que la distribución es simétrica en torno a 0, y de manera que:

$$P\left(-t^* \leq \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}} \leq t^*\right) = 0,95$$

En estas condiciones, y por un razonamiento análogo al seguido cuando la desviación típica era conocida, el valor de t^* es aquel que verifica:

$$P(t_{n-1} > t^*) = P(t_{120} > t^*) = (1 - 0,95)/2 = 0,025$$

Por tanto, consultando las tablas, por ejemplo, tenemos que $t^* = 1,98$; es decir:

$$P\left(-1,98 \leq \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}} \leq 1,98\right) = 0,95$$

Ahora, operando esta expresión con el objetivo de aislar el valor de μ , obtenemos:

$$P\left(\bar{X} - 1,98 \frac{s}{\sqrt{n}} \leq \mu \leq \bar{X} + 1,98 \frac{s}{\sqrt{n}}\right) = 0,95$$

Este resultado nos indica que el 95% de los intervalos de la forma siguiente:

$$\left(\bar{X} - 1,98 \frac{s}{\sqrt{n}}, \bar{X} + 1,98 \frac{s}{\sqrt{n}}\right)$$

contienen el verdadero valor de la media poblacional. Y ahora ya llegamos al final, lo único que tenemos que hacer es sustituir los valores de la media muestral, s y n , por los valores obtenidos en nuestro estudio, con la confianza del 95% de que el intervalo que obtengamos contendrá el verdadero valor de μ . Haciendo la sustitución, pues, obtenemos que el intervalo:

$$\left(171 - 1,98 \frac{11}{\sqrt{121}}, 171 + 1,98 \frac{11}{\sqrt{121}}\right) = (169,02, 172,98)$$

es un intervalo de confianza con un nivel del 95% para la media de las alturas de los estudiantes de la UOC. Es decir, en nuestro ejemplo la media de las alturas de los estudiantes de la UOC se encuentra entre 169,02 cm y 172,98 cm, con una confianza del 95%.

Resumimos a continuación el procedimiento para construir un intervalo de confianza para la media aritmética cuando la población es normal y desconocemos la desviación típica. Supongamos que la variable que queremos estudiar sigue una ley normal de media μ (desconocida) y desviación típica σ también desconocida y que disponemos de una muestra aleatoria simple de tamaño n y de valor de la media de la muestra \bar{x} . Entonces:

1) Fijamos el nivel de confianza, que habitualmente se escribe como $(1 - \alpha)\%$.

2) Calculamos la desviación típica muestral: $s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$

3) Calculamos el error estándar de la media $s_{\bar{x}} = \frac{s}{\sqrt{n}}$

4) Calculamos el llamado *valor crítico*, que es aquel punto $t_{\alpha/2, n-1}$ tal que:

$$P(t_{n-1} \geq t_{\alpha/2, n-1}) = \alpha/2$$

donde t_{n-1} es una variable Student con $n-1$ grados de libertad.

5) Calculamos el llamado *margen de error* (también llamado *precisión de la estimación*) como $t_{\alpha/2, n-1}$ para el error estándar, es decir, como:

$$t_{\alpha/2, n-1} \frac{s}{\sqrt{n}}$$

6) El intervalo de confianza obtenido con la muestra de partida es el siguiente:

$$(\bar{x} \pm t_{\alpha/2, n-1} s_{\bar{x}}) = \left(\bar{x} - t_{\alpha/2, n-1} \frac{s}{\sqrt{n}}, \bar{x} + t_{\alpha/2, n-1} \frac{s}{\sqrt{n}} \right)$$

3.1. El tamaño de la muestra

En este caso no tenemos un procedimiento directo para encontrar el tamaño de una muestra que produce un determinado margen de error, ya que el margen de error depende de la desviación típica muestral (s), que es un valor que calculamos precisamente una vez que la muestra ya ha sido seleccionada. En caso de que necesitemos aproximar el tamaño de la muestra, tomaremos una muestra de prueba que nos dé un valor aproximado de s . Después calcularemos el margen de error para este valor s y tomaremos una muestra de ancho suficiente para garantizar dicho margen.

Por otro lado, y tal como pasaba en el caso del intervalo de confianza de la media con la desviación típica conocida, el procedimiento de cálculo del intervalo de confianza cuando la desviación típica es desconocida también resulta válido, siempre que la muestra sea mayor que treinta, aunque la variable que hay que estudiar no siga una ley normal.

Tiempo de conexión al campus virtual con desconocimiento de la desviación típica poblacional

En un estudio llevado a cabo en la UOC se toma una muestra aleatoria de 150 estudiantes y se pregunta cuánto tiempo estuvieron conectados al campus virtual durante el mes de abril del año 2000. Se obtiene una media muestral de 120 minutos y una desviación típica muestral de 10 minutos. Dado que la muestra es grande ($n > 30$), podemos calcular un intervalo

de confianza del 95% para el tiempo de conexión durante este mes, considerando la media muestral ($\bar{x} = 120$) y el error estándar de la media, que es:

$$s_{\bar{x}} = \frac{10}{\sqrt{150}} = 0,816$$

Tenemos que calcular $t_{\alpha/2, n-1} = t_{0,025, 149}$, que es aquel valor para el que $P(t_{149} > t_{0,025, 149}) = 0,025$; utilizando algún programa de ordenador, obtenemos $t_{0,025, 149} = 1,976$. Por tanto, el intervalo de confianza con un nivel de confianza del 95% es $(120 \pm 1,976 \cdot 0,816) = (120 \pm 1,612416) = (118,39, 121,61)$.

En caso de que quisiéramos obtener un intervalo de confianza del 99%, deberíamos calcular $t_{0,005, 149} = 2,6092$ y el intervalo de confianza sería $(120 \pm 2,6092 \cdot 0,816) = (117,87, 122,13)$.

Si queremos que la precisión de nuestro intervalo (al nivel de confianza del 95%) sea de cinco puntos porcentuales, deberemos conseguir un margen de error inferior al 5% = 0,05. Podemos utilizar la muestra que tenemos para dar una aproximación al valor de s y, por tanto, podemos aplicar la fórmula:

$$1,976 \sigma_{\bar{x}} = 1,976 \frac{s}{\sqrt{n}} = 1,976 \frac{10}{\sqrt{n}} < 0,05$$

aislando n obtenemos $n > 156.183,04$, lo que significa que volvemos a necesitar una muestra inabarcable.

4. Comparación entre los casos estudiados

Hemos visto cómo debemos construir, a partir de los resultados obtenidos de las observaciones de una muestra aleatoria simple, los llamados *intervalos de confianza*. Se parte de un parámetro poblacional desconocido y de algunas hipótesis sobre la distribución de la variable de interés. Fijado cierto nivel de confianza C%, el método de construcción de los intervalos garantiza que el C% de las muestras produzca un intervalo que contenga el auténtico valor del parámetro desconocido.

Notación

En muchos libros el nivel de confianza se denota por $(1 - \alpha)100\%$.

En caso de querer encontrar intervalos de confianza para la media aritmética y suponiendo que la variable que hay que considerar sigue una distribución normal, encontramos que los intervalos de confianza están centrados en la media muestral y, por tanto, presentan la forma siguiente:

$$(\bar{x} \pm \text{margen de error})$$

El margen de error se calcula multiplicando un factor asociado al nivel de confianza por el error estándar de la media. Para calcular correctamente el margen de error, hay que distinguir entre dos casos:

a) Si conocemos la desviación típica poblacional σ , el margen de error se calcula como:

$$Z_{\alpha/2} \sigma_{\bar{x}}$$

donde $z_{\alpha/2}$ es aquel valor en el que $P(Z \geq z_{\alpha/2}) = \alpha/2$, siendo Z una variable $N(0, 1)$ y $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$, el error estándar de la media.

b) Si desconocemos la desviación típica poblacional, el margen de error se calcula como:

$$t_{\alpha/2, n-1} s_{\bar{x}}$$

donde $t_{\alpha/2, n-1}$ es el valor tal que $P(t_{n-1} \geq t_{\alpha/2, n-1}) = \alpha/2$, donde t_{n-1} es una variable Student con $n - 1$ grados de libertad y $s_{\bar{x}} = \frac{s}{\sqrt{n}}$ es el error estándar de la media, calculado a partir de la desviación típica muestral s .

Finalmente, se ha considerado el efecto del tamaño de la muestra, tanto para obtener intervalos de confianza con los márgenes de error deseados como para observar que en muestras de medida mayor que treinta, y en virtud del teorema del límite central, se pueden calcular los intervalos de confianza, aunque la población no sea normal.

5. Resumen

En esta sesión se describen procedimientos para obtener intervalos de confianza para la media en los casos siguientes:

- Cuando la variable sigue una distribución normal de desviación típica conocida.
- Cuando la variable sigue una distribución normal de desviación típica desconocida.

También se muestra cómo se pueden extender estos procedimientos a variables que no siguen una distribución normal, a condición de que la medida de la muestra considerada sea lo bastante grande (en concreto, $n > 30$).

En todos los casos se considera el efecto que el tamaño de la muestra tiene sobre el intervalo de confianza y se muestra cómo hay que calcular la medida de la muestra que produce un margen de error estipulado *a priori*.

Para poder construir los intervalos de confianza, se han introducido los conceptos de parámetro, estadístico y los niveles de confianza y significación.

Ejercicios

1. El tiempo (en segundos) que tarda en arrancar la última versión del programa Macrohard Phrase sigue una distribución normal de desviación típica de 40. En 81 ordenadores se ha medido el tiempo que tarda en arrancar y se ha encontrado que la media de los tiempos de arranque medidos es de 158,3 segundos.

a) Dad un intervalo de confianza del 90% para la media de tiempo de arranque del programa.

b) Interpretad el intervalo de confianza.

c) El fabricante afirma que la media del tiempo de arranque del programa es de 140 segundos. ¿Es eso posible, según lo que hemos obtenido con el intervalo de confianza?

d) ¿Cuál debería ser la medida de la muestra para reducir la longitud del intervalo de confianza a la mitad?

2. El tiempo (en segundos) que tarda en arrancar la última versión del programa Macrohard Phrase sigue una distribución normal. En 81 ordenadores se ha medido el tiempo que tarda en arrancar y se ha encontrado que la media de los tiempos de arranque medidos es de 158,3 segundos y la desviación típica de la muestra es de 12 segundos.

a) Dad intervalos de confianza del 90% y del 95% para la media de los tiempos de arranque del programa.

b) Comparad los intervalos de confianza obtenidos en el apartado anterior.

c) El fabricante afirma que la media del tiempo de arranque del programa es de 140 segundos. ¿Es eso posible, según lo que hemos obtenido con el intervalo de confianza?

3. El fabricante de una determinada marca de yogures afirma que sus envases contienen de media 150 gramos de yogur. Hemos ido al supermercado, hemos comprado diez yogures, hemos pesado su contenido y hemos obtenido los datos siguientes (en gramos): 148, 149, 147, 146, 149, 146, 149, 148, 149, 149.

a) Construid un intervalo de confianza del 95% para el peso de los yogures, suponiendo que el peso sigue una distribución normal de desviación típica de 3 g.

b) De acuerdo con el resultado anterior, y ya que los pesos de los yogures que hemos comprado son todos menores que 150, ¿podemos afirmar que el fabricante no es lo bastante sincero en el peso de sus productos?

4. Repetid el ejercicio anterior, pero suponiendo ahora que el peso sigue una distribución normal de desviación típica desconocida.

Solucionario

1.

a) Puesto que la población es normal y conocemos la desviación típica poblacional, procederemos de la forma siguiente:

1) Fijamos el nivel de confianza $(1 - \alpha) = 0,9$.

2) Calculamos el error estándar de la media $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{40}{\sqrt{81}}$.

3) Calculamos el valor crítico; como $(1 - \alpha) = 0,9$, $\alpha = 0,1$ y $z_{\alpha/2} = z_{0,05} = 1,645$.

4) Calculamos el margen de error (también llamado *precisión de la estimación*) como $z_{\alpha/2}$ para el error estándar, es decir, como:

$$z_{\alpha/2} \frac{\sigma}{\sqrt{n}} = 1,645 \cdot \frac{40}{\sqrt{81}} = 7,31$$

5) El intervalo de confianza es:

$$(\bar{x} \pm z_{\alpha/2} \sigma_{\bar{x}}) = (158,3 - 7,31, 158,3 + 7,31) = (150,99, 165,61)$$

b) La interpretación es la siguiente: en el 90% de las muestras de 81 ordenadores el valor de la media muestral obtenida hace que el intervalo contenga el verdadero valor de la media del tiempo que tarda en arrancar el programa.

c) El intervalo de confianza obtenido no contiene el valor 140; además, el extremo izquierdo del intervalo está muy alejado del valor 140. Esto nos indica que es poco probable que la verdadera media sea de 140 segundos.

d) Si la longitud del intervalo de confianza ha de ser la mitad, entonces el error estándar también debe ser la mitad; por tanto, tenemos que:

$$z_{\alpha/2} \frac{\sigma}{\sqrt{n}} = 1,645 \cdot \frac{40}{\sqrt{n}} = \frac{7,31}{2} = 3,655$$

resolviendo la ecuación obtenemos:

$$n = 1,645^2 \frac{40^2}{3,655^2} \approx 324$$

Observamos que equivale a multiplicar por cuatro la medida de la muestra inicial.

2.

a) En este caso no conocemos la desviación típica poblacional, por lo que aplicaremos el procedimiento siguiente (calcularemos simultáneamente los dos intervalos pedidos):

1) Fijamos los niveles de confianza; en el primer caso, $1 - \alpha = 0,9$ y en el segundo, $1 - \alpha = 0,95$.

2) Calculamos la desviación típica muestral, que es $s = 12$, tal como nos dice el enunciado.

3) Calculamos el error estándar de la media $s_{\bar{x}} = \frac{s}{\sqrt{n}} = \frac{12}{\sqrt{81}} = 1,33$.

4) Calculamos los valores críticos:

- Caso $1 - \alpha = 0,9 \Rightarrow t_{\alpha/2, n-1} = t_{0,05;80} = 1,6641$
- Caso $1 - \alpha = 0,95 \Rightarrow t_{\alpha/2, n-1} = t_{0,025;80} = 1,9901$

5) Calculamos el margen de error como $t_{\alpha/2, n-1}$ para el error estándar, es decir:

- Caso $1 - \alpha = 0,9 \Rightarrow s_{\bar{x}} = t_{\alpha/2, n-1} \frac{s}{\sqrt{n}} = 1,664 \cdot 1,33 = 2,2132$
- Caso $1 - \alpha = 0,95 \Rightarrow s_{\bar{x}} = t_{\alpha/2, n-1} \frac{s}{\sqrt{n}} = 1,9901 \cdot 1,33 = 2,6468$

6) Finalmente, los intervalos de confianza pedidos son los siguientes:

- Caso $1 - \alpha = 0,9 \Rightarrow (158,3 - 2,2132, 158,3 + 2,2132) = (156,09, 160,51)$
- Caso $1 - \alpha = 0,95 \Rightarrow (158,3 - 2,6468, 158,3 + 2,6468) = (155,65, 160,95)$

b) Al aumentar el nivel de confianza, el intervalo se hace más largo. Esto se debe al hecho de que como tenemos que garantizar que más intervalos contengan la media poblacional, los intervalos, que siempre están centrados en la media muestral, deben ser más largos.

c) No parece posible, ya que los valores del intervalo de confianza están muy alejados del valor 140.

3.

a) Se trata de un intervalo de confianza para una variable normal cuya desviación típica poblacional conocemos; por tanto, tiene la forma siguiente:

$$\left(\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right) = \left(148 \pm 1,96 \cdot \frac{3}{\sqrt{10}} \right) = (146,14, 149,8)$$

b) El intervalo obtenido no contiene el valor 150; por tanto, la conclusión a la que llegamos es que, a pesar de que es posible que 150 sea la media poblacional (ya que está muy cerca de los límites del intervalo), podemos decir que no lo es con una confianza del 95%, ya que sabemos que el 95% de los intervalos contiene la media, y éste no la contiene.

4.

a) Se trata de un intervalo de confianza para una variable normal cuya desviación típica poblacional desconocemos; por tanto, tiene la forma siguiente:

$$\left(\bar{x} \pm t_{\alpha/2, n-1} \frac{s}{\sqrt{n}} \right) = \left(148 \pm 2,2621 \frac{1,247}{\sqrt{10}} \right) = (147,11, 148,89)$$

donde \bar{x} es la media de los valores de la muestra y s es la desviación típica de dichos valores.


b) Tampoco ahora el intervalo obtenido contiene el valor 150; por tanto, llegamos a la misma conclusión que en el ejercicio anterior, con la diferencia de que ahora el valor 150 se encuentra más lejos del intervalo.

Observemos que, al no asumir la desviación típica poblacional igual a 3, el margen de error se vuelve más pequeño, ya que la desviación típica muestral es mucho menor que 3.

Intervalo de confianza para la proporción

En esta sesión nos dedicaremos al estudio del intervalo de confianza para una proporción. Comenzaremos con un ejemplo. Supongamos que queremos estudiar la proporción de estudiantes de la UOC que han visitado el Valle de Nuria alguna vez (para ver si vale la pena hacer publicidad de dicho lugar, por ejemplo). Seguiremos los pasos que exponemos a continuación:

- 1) Supongamos que la proporción real de los de estudiantes de la UOC que han visitado alguna vez el Valle de Nuria es p .
- 2) Escogemos una muestra aleatoria de, pongamos por caso, $n = 136$ estudiantes.
- 3) En esta muestra el 75% de los estudiantes declara haber estado en Nuria.

Denotaremos por \hat{p} la proporción obtenida a partir de una muestra. 

Entonces, en nuestro caso $\hat{p} = 0,75$.

¿Qué conclusiones podemos sacar ahora sobre la relación entre la proporción real, con respecto a toda la población p , y la proporción \hat{p} obtenida a partir de una muestra?

Como en el caso de la media, deberemos recurrir a la teoría de la probabilidad para conocer la distribución de la proporción, según la cual, si la medida de la muestra es lo bastante grande, la variable siguiente sigue una distribución normal estándar:

$$\frac{\hat{P} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

Notación

p es el parámetro poblacional y \hat{p} es el estadístico que utilizamos para estimarlo. Y denotamos por \hat{P} todos los posibles valores de las proporciones en todas las muestras del mismo tamaño.

Ahora, si fijamos un nivel de confianza del 95%, por ejemplo, y puesto que estamos utilizando una distribución normal estándar, tenemos que:

$$P\left(-1,96 \leq \frac{\hat{P} - p}{\sqrt{\frac{p(1-p)}{n}}} \leq 1,96\right) = 0,95$$

Operando esta expresión con el objetivo de aislar el valor de p , tenemos que:

$$P\left(\hat{P} - 1,96 \sqrt{\frac{p(1-p)}{n}} \leq p \leq \hat{P} + 1,96 \sqrt{\frac{p(1-p)}{n}}\right)$$

A diferencia de lo que pasaba en los ejemplos anteriores, ahora nos encontramos con un problema añadido, y es que el error estándar:

$$\sqrt{\frac{p(1-p)}{n}}$$

también depende del valor de p , que es precisamente el valor que queremos estimar y que, por tanto, es desconocido.

En este caso lo que haremos será aproximar el valor del error estándar por el valor que obtendríamos a partir del valor de la proporción en la muestra que tenemos, es decir, aproximaremos el valor del error estándar por:

$$\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Por tanto, en el caso del intervalo de confianza de la proporción, si el nivel de confianza es del 95%, trabajaremos con intervalos de la forma:

$$\left(\hat{p} - 1,96 \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + 1,96 \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right)$$

Ahora, lo único que tenemos que hacer es sustituir los valores de \hat{p} y n por los valores obtenidos en nuestro estudio, con lo que construimos el intervalo siguiente:

$$\left(\hat{p} - 1,96 \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + 1,96 \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right) =$$

$$= (0,75 - 1,96 \cdot 0,03713, 0,75 + 1,96 \cdot 0,03713) = (0,6772, 0,8227)$$

que es un intervalo de confianza con un nivel del 95% para la proporción de los estudiantes de la UOC que han visitado alguna vez el Valle de Nuria.

Por tanto, la proporción de estudiantes de la UOC que, en nuestro ejemplo, ha visitado el Valle de Nuria está entre el 67,72% y el 82,27%, con una confianza del 95%.

Como en el caso de la media aritmética, describiremos detalladamente el procedimiento de construcción de los intervalos de confianza.

1. Procedimiento para construir un intervalo de confianza para la proporción

Supongamos que disponemos de una muestra aleatoria simple de tamaño n y del valor de la proporción calculado a partir de la muestra \hat{p} y que el tamaño de la muestra es lo bastante grande, en concreto nosotros exigiremos siempre que el tamaño sea superior a cien. A continuación:

1) Fijamos el nivel de confianza, que habitualmente escribiremos como $(1 - \alpha)\%$.

Nivel de confianza

La confianza aproximada será del 95%, que nos dice que noventa y cinco de cada cien veces que escogemos una muestra aleatoria simple y hacemos la sustitución, el intervalo que obtendremos contendrá el auténtico valor de p .

Muestra siempre superior a 100

El tamaño de la muestra debe ser superior a cien para garantizar que se pueda aplicar el teorema del límite central y para obtener simultáneamente márgenes de error aceptables.

2) Calculamos la aproximación al error estándar siguiente:

$$s_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Error estándar de la muestra

Escribimos $s_{\hat{p}}$ en lugar de $\sigma_{\hat{p}}$, ya que se calcula a partir del valor de la proporción en una muestra.

3) Calculamos el llamado **valor crítico**, que es aquel punto $z_{\alpha/2}$ por el que $P(Z \geq z_{\alpha/2}) = \alpha/2$, donde Z es una variable $N(0, 1)$.

4) Calculamos el llamado *margen de error* (también llamado *precisión de la estimación*) como $z_{\alpha/2} s_{\hat{p}}$ para nuestra aproximación al error estándar, es decir, como:

$$z_{\alpha/2} s_{\hat{p}} = z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

5) El intervalo de confianza obtenido con la muestra de partida es:

$$(\hat{p} \pm z_{\alpha/2} s_{\hat{p}}) = \left(\hat{p} - z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right)$$

2. El efecto del tamaño de la muestra

Muchas veces, una vez fijado el nivel de confianza, nos marcaremos como objetivo dar el valor del parámetro p con una cierta precisión. La única forma de obtener la precisión deseada será modificando de forma adecuada el tamaño de la muestra. Supongamos que deseamos una precisión o margen de error ME ; si igualamos el margen de error con el error estándar, tenemos que:

$$ME = z_{\alpha/2} \sqrt{\frac{p(1-p)}{n}}$$

aislando n obtenemos:

$$n \geq (z_{\alpha/2})^2 \frac{p(1-p)}{ME^2}$$

Ahora volvemos a encontrarnos con el problema de antes, ya que, antes de hacer el estudio, no sabemos cuál es el valor de p .

En este caso se suele escoger un valor de \hat{p} que nos parezca apropiado (cuanto más cerca del verdadero y desconocido valor de p , mejor) de acuerdo con anteriores estudios o con algún hecho destacado.

En resumen, determinaremos el tamaño de la muestra así:

$$\text{Tamaño de la muestra} = n = (z_{\alpha/2})^2 \frac{\hat{p}(1-\hat{p})}{ME^2}$$

En caso de que no dispongamos de ninguna información sobre el posible valor de \hat{p} , deberemos asegurarnos de que la medida de la muestra es lo bastante grande para lograr nuestro objetivo. Dado que $z_{\alpha/2}$ y el margen de error están fijados, utilizaremos el hecho de que el producto $\hat{p}(1-\hat{p})$ es siempre menor o igual que $1/4$ y que precisamente $\hat{p}(1-\hat{p}) = 1/4$, en el caso de que $\hat{p} = 0,5$. Así pues, suponiendo que $\hat{p} = 0,5$, obtenemos el tamaño muestral máximo para cada margen de error. Éste es el procedimiento que siguen muchos sondeos y estudios electorales, que en definitiva se basa en el hecho de que:

Esto se debe a que $x(1-x)$ alcanza su valor máximo cuando $x = 0,5$.

$$\text{Tamaño de la muestra} = n = (z_{\alpha/2})^2 \frac{\hat{p}(1-\hat{p})}{ME^2} \leq \frac{1}{4} \frac{(z_{\alpha/2})^2}{ME^2}$$

Por tanto una muestra de tamaño $n = \frac{1}{4} \left(\frac{z_{\alpha/2}}{ME} \right)^2$ garantiza siempre el margen de error deseado.

Tabla de tamaños

Utilizando la fórmula anterior, obtenemos la siguiente tabla de tamaños muestrales según el margen de error (para un nivel de confianza del 95%):

Margen de error	Tamaño máximo
5%	384
2%	2.401
1%	9.640
0,5%	38.416
0,02%	≈24.010.000

Proporción de catalanes que irán a votar

De cien catalanes escogidos al azar, 65 nos dicen que irán a votar en las próximas elecciones. Con estos datos podemos determinar un intervalo de confianza con un nivel de confianza del 95% para la proporción de catalanes que irán a votar de esta manera:

$$\left(\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right) = \left(\frac{65}{100} \pm 1,96 \sqrt{\frac{\frac{65}{100} \left(1 - \frac{65}{100} \right)}{100}} \right) = (0,5565, 0,7434)$$

Por tanto, podemos esperar que el porcentaje de votantes estará entre el 55% y el 74%, con un nivel de confianza del 95%.

Realmente obtenemos un intervalo de confianza muy amplio, pero es que sólo hemos preguntado a cien individuos.

Tamaño de la muestra

Si sabemos que en las anteriores elecciones fue a votar un 70% de los catalanes, podemos utilizar esta información para determinar cuál debe ser el tamaño de una muestra que permita

construir un intervalo de confianza como el del ejemplo anterior, pero con margen de error menor del 5%. En concreto deberemos aplicar la fórmula siguiente para obtener el tamaño de la muestra:

$$n = (z_{\alpha/2})^2 \frac{\hat{p}(1-\hat{p})}{ME^2} = 1,96^2 \frac{0,7 \cdot 0,3}{0,05^2} = \frac{0,806736}{0,05^2} = 322,69$$

Es decir, necesitaríamos datos sobre la intención de participar de 323 catalanes.

3. Resumen

En esta sesión hemos construido el intervalo de confianza para la proporción y hemos estudiado la relación entre el margen de error y el tamaño de la muestra.

Ejercicios

1.

Hemos cogido una muestra aleatoria de 400 estudiantes de la UOC y hemos observado que 180 de ellos son fumadores.

a) Dad un intervalo de confianza del 95% para la proporción de estudiantes de la UOC que fuman. ¿Cuál es la longitud del intervalo de confianza?

b) ¿Cuántos estudiantes deberíamos seleccionar para reducir la longitud del intervalo de confianza del 95% a la mitad? ¿Y si queremos que el margen de error sea inferior al 1%?

c) El Departamento de Sanidad de la Generalitat de Cataluña afirma que el 40% de la población catalana es fumadora. ¿Qué consecuencias podemos extraer de los estudiantes de la UOC, comparando este dato con el intervalo de confianza?

2.

En la edición de Cataluña del día 28-4-2001, el diario *El Periódico* publicaba la siguiente ficha técnica de un sondeo llevado a cabo con motivo de las elecciones autonómicas en el País Vasco del día 13-5-2001:

“Error de la muestra: el error en la muestra asociado a un nivel de confianza del 95,5% ($\sigma = 2$) y $p = q = 50\%$ es del $\pm 2,57$ ”. Tamaño de la muestra: 1.514.

Comprobad que el error (margen de error) está calculado correctamente.

Solucionario

1.

a) El intervalo de confianza tiene la forma siguiente:

$$\left(\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right) = \left(\frac{180}{400} \pm 1,96 \sqrt{\frac{\frac{180}{400} \left(1 - \frac{180}{400} \right)}{400}} \right) = (0,4012, 0,4988)$$

y su longitud es (0,4012, 0,4988).

b) Si queremos reducir la longitud del intervalo a la mitad, deberemos multiplicar la medida de la muestra por cuatro. Esto se ve fácilmente sustituyendo en la fórmula del tamaño muestral ME por $ME/2$, con lo que tenemos que:

$$\text{Tamaño de la muestra} = n = (z_{\alpha/2})^2 \frac{\hat{p}(1-\hat{p})}{(ME/2)^2} = 4(z_{\alpha/2})^2 \frac{\hat{p}(1-\hat{p})}{ME^2}$$

Y, por tanto, el nuevo tamaño debe ser cuatro veces el antiguo.

Si queremos que el margen de error sea inferior a 0,01, se deberá cumplir:

$$n \geq (z_{\alpha/2})^2 \frac{\hat{p}(1-\hat{p})}{(0,01)^2} = 1,96^2 \frac{\frac{180}{400} \left(1 - \frac{180}{400}\right)}{1,01^2} = 9.507,96$$

c) Dado que el 95% de los intervalos de confianza contienen la verdadera proporción poblacional y puesto que este intervalo no contiene el 40%, debemos pensar que la proporción de fumadores en la UOC supera la proporción de fumadores en la población catalana.

2.

En este caso tenemos $p = q = 0,5$, es decir, se calcula el tamaño máximo de la muestra para el nivel de confianza dado. El nivel de confianza es 95,5%, habitualmente utilizado en los sondeos, ya que en este caso el valor $z_{\alpha/2} = 2$ (o “2 sigma”, como dice la ficha técnica, ya que representa dos desviaciones típicas de la media). Aplicando la fórmula del margen de error obtenemos que:

$$ME = z_{\alpha/2} \sqrt{\frac{p(1-p)}{n}} = 2 \sqrt{\frac{1/4}{1.514}} = 0,025700$$

por tanto, el margen de error es 2,57% (en los sondeos se utiliza la expresión $\pm 2,57$, que indica una posible variación hacia arriba o hacia abajo del 2,57%).