408 TEXTOS DOCENTS

PROBLEMAS DE PROBABILIDADES Y ESTADÍSTICA

Vol. 2. Inferencia estadística

Carles M. Cuadras

Departamento de Estadística

UNIVERSITAT DE BARCELONA

PROBLEMAS DE PROBABILIDADES Y ESTADÍSTICA

Vol. 2. Inferencia estadística

Carles M. Cuadras

Departamento de Estadística

408 TEXTOS DOCENTS

UNIVERSITAT DE BARCELONA

Edicions

Índice

Prólogo

La primera versión de esta obra fue publicada por la Universidad de Barcelona en 1974, formando parte de la colección Publicaciones del Laboratorio de Cálculo. Pronto se convirtió en libro y después de un periplo de cinco lustros por las editoriales EUNIBAR, PPU y EUB S. L., y tras una docena de ediciones, la última (revisada y ampliada), apareció en 1999-2000. Agotadas estas ediciones, y ante la insistencia de varios colegas, la obra regresa a la Universidad de Barcelona, que se ha ofrecido a editarla y publicarla con nuevo formato, dentro de la colección de Textos Docentes.

Es este un libro de ejercicios y, por lo tanto, destinado a proponer y resolver problemas de probabilidades y estadística. Los problemas de naturaleza teórica se alternan con los aplicados. Se ha procurado presentarlos en orden de dificultad creciente. La extensión de la obra ha obligado a dividirla en dos volúmenes: el primero dedicado a Problemas de probabilidades y el segundo, a Problemas de inferencia estadística. La actual edición EUB de la Universidad de Barcelona mejora la edición de 1999-2000 en la presentación y revisión de algunos apartados.

A fin de facilitar la consulta del libro, se incluye en cada capítulo un extenso resumen de la teoría. Para ampliar la teoría, el lector deberá consultar libros adecuados y puede orientarse en la bibliografía que aparece al final de cada volumen. Los apartados de mayor dificultad teórica y los problemas de resolución más difícil se indican con un asterisco.

La Estadística se ha ido imponiendo como materia fundamental en la mayoría de los estudios universitarios. Esta obra, con un planteamiento teórico-práctico, sintetiza los principales temas de la Probabilidad y la Estadística. A pesar de los grandes avances tecnológicos, la base no ha cambiado, por lo que el contenido de este libro sigue teniendo plena actualidad.

Este segundo libro, dedicado a la inferencia estadística, contiene problemas teóricos y aplicados sobre la base de datos de procedencia muy variada.

Los dos volúmenes constituyen una obra que resulta adecuada para estudiantes y profesores de primer ciclo de carreras y grados universitarios (Estadística, Biología, Matemáticas, Física, Química, Geología, Ingeniería, Arquitectura, Informática, Economía, Veterinaria, Psicología, Pedagogía y Medicina).

Mi más sincero agradecimiento a todos aquellos que me han hecho comentarios, sugerido cambios y correcciones, lo que ha hecho posible la versión actual. En especial debo mencionar a J. Fortiana, P. Sánchez Algarra, G. Alonso, J. M. Oller, F. Carmona, J. Ocaña, C. Ruiz-Rivas, M. Ríos, R. Vélez, J. M. Font, S. Salvo y D. Cuadras.

Respecto a la presente edición, deseo agradecer a M. Aicart, J. M. Oller, M. C. Pallejá y Edicions de la Universitat de Barcelona su valiosa colaboración en la preparación del original.

Carles M. Cuadras

Barcelona, diciembre de 2016

1. MUESTREO

1.1. Muestreo aleatorio simple

Sea X una variable aleatoria (v.a.) definida sobre una población Ω. Dados n valores de X

$$x_1,\ x_2,\ \ldots,\ x_n$$

observados en condiciones independientes, diremos que constituyen una *muestra aleatoria simple* de X. Los valores de una muestra son aleatorios (a muestras distintas tendremos valores distintos), y desde el punto de vista probabilístico, deben considerarse como n v.a. $X_1,\ X_2,\ \ldots,\ X_n$ tales que:

1. Son v.a. independientes.
2. Cada X_i tiene la misma distribución que X.

Por lo tanto, si $f_X(x)$ es la función de densidad de X, la función de densidad de una muestra aleatoria simple $x_1, x_2, \ldots, x_n$ es

$$f(x_1,\ x_2,\ \ldots,\ x_n) = f_X(x_1) \cdot f_X(x_2) \cdots f_X(x_n).$$

El conjunto de posibles valores que podemos obtener mediante un muestreo aleatorio simple de tamaño n es un subconjunto del espacio euclídeo $\mathbb{R}^n$, llamado *espacio muestral*.

1.2. Concepto de estadístico. Características muestrales

Dada una muestra $(X_1,\ \ldots,\ X_n)$ se llama *estadístico* a toda variable que sea función de la muestra

$$U_n = g(X_1,\ \ldots,\ X_n).$$

Los estadísticos se presentan de forma natural en problemas de estimación de parámetros (y entonces reciben el nombre de estimadores), y en contraste de hipótesis. Son aspectos relacionados con la obtención de un estadístico:

1. Encontrar el estadístico adecuado al problema de inferencia estadística que pretendemos resolver.
2. Encontrar la distribución exacta de U_n para cualquier n.
3. Encontrar la distribución asintótica de U_n para $n \longrightarrow \infty$.
4. Determinar el tamaño de la muestra n para obtener una precisión suficiente en estimación de parámetros, o construir un test con potencia alta.

Los parámetros poblacionales $m = E(X)$, $\sigma^2 = \mathrm{var}(X)$, $m_k = E\left(X^k\right)$ están relacionados con ciertos estadísticos especialmente importantes llamados características muestrales:

$$\begin{array}{lrcl} \textit{Media muestral} & \overline{X}_n & = & \dfrac{1}{n}(X_1+\cdots+X_n) \\ \textit{Varianza muestral} & S_n^2 & = & \dfrac{1}{n}\sum_{i=1}^{n}\left(X_i-\overline{X}_n\right)^2 \\ \textit{Momento muestral} & A_k & = & \dfrac{1}{n}\sum_{i=1}^{n} X_i^k \end{array}$$

Se verifica:

$$E\left(\overline{X}_n\right)=m \qquad \operatorname{var}\left(\overline{X}_n\right)=\frac{\sigma^2}{n} \qquad E\left(S_n^2\right)=\frac{n-1}{n}\sigma^2 \quad E(A_k)=m_k$$

Las características muestrales son v.a. (que varían con el muestreo), algunos casos con distribución exacta conocida y otros con distribución asintótica, generalmente normal. $\overline{X}_n$, S_n^2 y A_k no deben confundirse con m, σ^2 y m_k, que son las características poblacionales de X y tienen valor constante. Las características muestrales calculadas sobre una muestra dada se indican por $\overline{x}_n$, s_n^2 y a_k.

Análogamente, se define un muestreo aleatorio simple de tamaño n del par (X, Y) con distribución bivariante, como n pares ordenados

$$(x_1,\ y_1),\ (x_2,\ y_2),\ \ldots,\ (x_n,\ y_n)$$

obtenidos en condiciones independientes. Son características muestrales importantes:

$$\begin{array}{lrcl} \textit{Covarianza muestral:} & S_{XY} & = & \dfrac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y}) \\ \textit{Correlación muestral:} & r & = & \dfrac{S_{XY}}{S_X\,S_Y} \end{array}$$

donde $S_X=\left(\dfrac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2\right)^{1/2}$, $S_Y=\left(\dfrac{1}{n}\sum_{i=1}^{n}(y_i-\overline{y})^2\right)^{1/2}$ son las desviaciones típicas muestrales. Tampoco debe confundirse r (que es una v.a.) con la correlación poblacional ρ.

1.3. Muestreo artificial

En ciertos problemas es útil simular la distribución de una v.a. X con función de distribución $F_X(x)$, obteniendo una muestra artificial de X. El proceso es el siguiente:

1. Se genera una muestra aleatoria simple

   $$z_1,\ z_2,\ \ldots,\ z_n$$

   de una v.a. uniforme en $(0,\ 1)$.

2. Si F_X es continua y creciente, entonces

   $$x_1=F_X^{-1}(z_1),\ x_2=F_X^{-1}(z_2),\ \ldots,\ x_n=F_X^{-1}(z_n)$$

   es una muestra aleatoria simple de X.

3. Si X tiene distribución discreta, se genera la muestra haciendo corresponder a z_i el valor x_i tal que $F_X(x_{i-1}) < z_i \leqslant F(x_i)$.

Un procedimiento para generar una muestra sobre $(0,\ 1)$ es iniciarla con un valor z_1 cualquiera, y utilizar la fórmula de recurrencia

$$z_i = \text{parte fraccionaria de } \left[(\pi + z_{i-1})^8\right].$$

El muestreo artificial se aplica para comprobar la calidad de un estimador, la potencia de un test, simular la distribución de un estadístico U_n cuya distribución exacta resulta complicada, simulación en modelos genéticos, etc.

1.4. Imagen empírica de una distribución continua

La función de distribución $F(x)$ describe totalmente la distribución de una v.a. X. Sin embargo, si nuestra información sobre X es una muestra aleatoria simple $x_1, \ldots, x_n$, podemos obtener una imagen empírica de la distribución ordenando la muestra

$$x_{(1)} < x_{(2)} < \cdots < x_{(n)}$$

y definiendo la *función de distribución empírica*

$$F_n(x) = \begin{cases} 0 & \text{si} \quad x < x_{(1)}, \\ \dfrac{k}{n} & \text{si} \quad x_{(k)} \leqslant x < x_{(k+1)}, \qquad k = 1, \ldots, n-1, \\ 1 & \text{si} \quad x \geqslant x_{(n)}. \end{cases}$$

Como $F_n(x)$ es la frecuencia relativa de la presencia del suceso $[X \leqslant x]$, se verifica, para cada $x \in \mathbb{R}$, como consecuencia del teorema de Bernoulli, la convergencia en probabilidad

$$F_n(x) \xrightarrow{P} F(x). \tag{1}$$

Por lo tanto, para n grande, $F_n(x)$ proporciona una imagen de la distribución de la variable X.

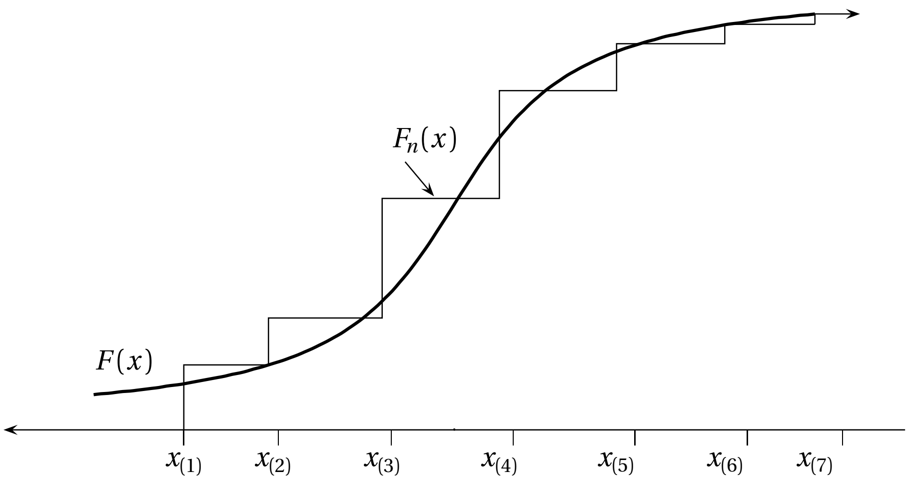

Figura 1.1.

Una manera de estudiar la diferencia entre $F_n(x)$ y $F(x)$ es introducir el estadístico

$$D_n = \sup_x |F_n(x) - F(x)|.$$

Se verifica, entonces, una propiedad todavía más fuerte que (1), conocida como teorema de Glivenko-Cantelli, también llamado teorema fundamental de la estadística.

Teorema. $F_n(x)$ converge casi seguramente a $F(x)$ uniformemente en x, es decir,

$$P\left(\lim_{n\to\infty} D_n = 0\right) = 1.$$

1.5. Muestreo sobre poblaciones finitas

Se entiende por población finita un conjunto finito de N elementos de naturaleza homogénea (N familias, N piezas fabricadas, N individuos de una especie). Si obtenemos una muestra de tamaño n, *con reemplazamiento*, cada elemento tiene una probabilidad $1/N$, las muestras son independientes y el muestreo es aleatorio simple. Si la muestra obtenida es *sin reemplazamiento*, las diferentes muestras no son independientes.

Si el muestreo se realiza sobre una variable cuantitativa X, son características poblacionales importantes la media y la varianza

$$m = \frac{1}{N}\sum_{i=1}^{N} x_i \qquad \sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - m)^2.$$

En el muestreo sin reemplazamiento, la media muestral $\overline{x}'$ verifica

$$E\left(\overline{x}'\right) = m \qquad \operatorname{var}\left(\overline{x}'\right) = \frac{N-n}{N-1}\cdot\frac{\sigma^2}{n}.$$

Si lo que se desea estudiar es la presencia de una característica cualitativa de probabilidad p, de modo que N_p es el número de elementos que poseen esta característica, si $\overline{x}$ es la frecuencia relativa de esta característica en el muestreo con reemplazamiento tendremos:

$$E(\overline{x}) = p \qquad \operatorname{var}(\overline{x}) = \frac{pq}{n},$$

siendo $q = 1-p$. Si $\overline{x}'$ es la frecuencia relativa en el muestreo sin reemplazamiento, entonces:

$$E\left(\overline{x}'\right) = p \qquad \operatorname{var}\left(\overline{x}'\right) = \frac{N-n}{N-1}\cdot\frac{pq}{n}.$$

Problemas resueltos

1.1. Sea X una v.a. con distribución de Poisson de parámetro λ. Dada una muestra aleatoria simple de tamaño n

$$X_1, \ldots, X_n$$

resolver las siguientes cuestiones:

a) Hallar la función de densidad conjunta de la muestra.

b) Decir cuáles de las siguientes funciones son estadísticos:

$$\begin{aligned} U &= X_1 \cdot X_2 \cdots X_n - n\lambda \\ V &= X_1 + X_2 \\ W &= X_1 + \cdots + X_n \\ Z &= \max\{X_1, \ldots, X_n\} \end{aligned}$$

y hallar la distribución de W y Z.

c) ¿Cuál es la característica muestral correspondiente a λ?

Solución

a) $f_X(x) = e^{-\lambda} \dfrac{\lambda^x}{x!}, \quad x = 0, 1, 2, \ldots$ luego la densidad de una muestra $x_1, \ldots, x_n$ es

$$f(x_1, \ldots, x_n) = e^{-n\lambda} \frac{\lambda^{\Sigma x_i}}{x_1! \cdots x_n!}.$$

b) U no es un estadístico porque depende de λ y su valor dada una muestra $x_1, \ldots, x_n$, no puede ser calculado si λ es desconocido.

V es estadístico. Sin embargo, es un estadístico que no utiliza toda la información de la muestra.

W y Z son también estadísticos. Como la suma de v.a. Poisson independientes es también una Poisson, la función de densidad de W es

$$f_W(k) = e^{-n\lambda} \frac{(n\lambda)^k}{k!} \qquad k = 0, 1, 2, \ldots$$

Por otra parte, como $X_1, \ldots, X_n$ son independientes,

$$\begin{aligned} P(Z \leqslant k) &= P(\max\{X_1, \ldots, X_n\} \leqslant k) \\ &= P([X_1 \leqslant k] \cap \cdots \cap [X_n \leqslant k]) \\ &= P(X_1 \leqslant k) \cdots P(X_n \leqslant k) = [F_X(k)]^n, \end{aligned}$$

siendo $F_X(k) = \sum_{i \leqslant k} e^{-\lambda} \dfrac{\lambda^i}{i!}$, luego la densidad de Z es

$$f_Z(k) = [F_X(k)]^n - [F_X(k-1)]^n.$$

c) Como

$$\lambda = E(X) = \mathrm{var}(X)$$

tanto $\overline{X}_n$ como S_n^2 son características muestrales relacionadas con λ. Veremos más adelante que $\overline{X}_n$ es más eficiente en problemas de inferencia sobre λ.

1.2. Una v.a. tiene una distribución $N(m,\ \sigma^2)$. Hallar la función de densidad conjunta de una muestra aleatoria simple de tamaño n y calcular $\overline{x}_n$, s_n^2 y a_3 para la muestra

$$1{,}74 \quad 0{,}45 \quad 2{,}52 \quad 1{,}19 \quad 1{,}24 \quad 2{,}68 \quad 3{,}51 \quad 1{,}83 \quad 1{,}00 \quad 0{,}87.$$

Solución

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-(x-m)^2/2\sigma^2}$$

$$f(x_1,\ \ldots,\ x_n) = \frac{1}{\left(\sigma\sqrt{2\pi}\right)^n}\, e^{-\sum\limits_{i=1}^{n}\frac{(x_i-m)^2}{2\sigma^2}}$$

$$n = 10 \qquad \sum_{i=1}^{n} x_i = 17{,}03 \qquad \sum_{i=1}^{n} x_i^2 = 37{,}14 \qquad \sum_{i=1}^{n} x_i^3 = 95{,}23$$

$$\overline{x}_n = 1{,}703 \qquad s_n^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{x}_n^2 = 0{,}814 \qquad a_3 = 9{,}523.$$

1.3. Hallar la distribución exacta del estadístico media muestral $\overline{X}_n$ en los casos de que la distribución de la variable X sea:

a) Bernoulli de parámetro p.

b) Exponencial de parámetro α.

c) Cauchy.

d) Normal $N\left(m,\ \sigma^2\right)$.

¿Cuál es la distribución asintótica de $\overline{X}_n$ si la distribución X es desconocida?

Solución

a) X toma el valor 1 con probabilidad p y el valor 0 con probabilidad $1-p$. Una muestra de X estará formada por una sucesión de ceros y unos. Luego, $\overline{X}_n$ será la frecuencia relativa de unos en una muestra de tamaño n y su función de densidad será:

$$f_{\overline{X}_n}\left(\frac{k}{n}\right) = \binom{n}{k} p^k (1-p)^{n-k} \qquad \frac{k}{n} = 0,\ \frac{1}{n},\ \frac{2}{n},\ \ldots,\ \frac{n-1}{n},\ 1.$$

b) Por las propiedades de la distribución exponencial $S = X_1 + \cdots + X_n$ sigue la distribución gamma $G(\alpha,\ n)$, con función de densidad

$$f_S(x) = \frac{\alpha^n}{\Gamma(n)}\, e^{-\alpha x}\, x^{n-1} \qquad \text{si} \quad x > 0\,,$$

luego, la función de densidad de $\overline{X}_n = S/n$ es

$$f_{\overline{X}_n}(x) = n\, f_S(nx) = n\, \frac{\alpha^n}{\Gamma(n)}\, e^{-\alpha n x}(nx)^{n-1} \qquad \text{si} \quad x > 0.$$

c) Si X sigue la distribución de Cauchy, entonces $\overline{X}_n$ sigue también la distribución de Cauchy (véase problema 10.6., volumen 1).

d) $\overline{X}_n$ es $N\left(m,\ \dfrac{\sigma^2}{n}\right)$.

Finalmente, si X posee esperanza m y varianza σ^2, por el teorema central del límite, la distribución asintótica de $\overline{X}_n$ (distribución para n grande) es $N(m,\ \sigma^2/n)$.

1.4. Los siguientes números: 0,8663, 0,9948, 0,0367, 0,1918, 0,2859, 0,1493, 0,2187, 0,6653, 0,8778 y 0,7912 han sido extraídos al azar e independientemente del intervalo $(0,\ 1)$. Generar una muestra aleatoria simple del mismo tamaño de una v.a. X con distribución exponencial de parámetro $\alpha = 2$.

Solución. Las funciones de densidad y de distribución son:

$$\begin{aligned} f(x) &= 2\,e^{-2x} \qquad \text{si} \quad x > 0,\\ F(x) &= 1 - e^{-2x} \qquad \text{si} \quad x > 0. \end{aligned}$$

Si Z es uniforme en $(0,\ 1)$, se verifica que $X = F^{-1}(Z)$ es exponencial de parámetro $\alpha = 2$. Entonces

$$Z = 1 - e^{-2X} \quad \Rightarrow \quad e^{-2X} = 1 - Z \quad \Rightarrow \quad X = -\ln(1 - Z)/2.$$

Aplicando esta transformación inversa obtenemos la muestra:

1,0062,	2,6295,	0,0187,	0,1064,	0,1683
0,0808,	0,1234,	0,5472,	1,051,	0,7832.

1.5. La distribución de una v.a. X es uniforme en el intervalo $(0,\ 1)$. En muestras aleatorias simples de tamaño n hallar la distribución asintótica del estadístico «media geométrica»

$$G_n = (X_1 \cdot X_2 \cdot \dots \cdot X_n)^{1/n}.$$

Solución

$$Y_n = \ln G_n = \frac{1}{n}(\ln X_1 + \ln X_2 + \dots + \ln X_n)$$

$$\begin{aligned} E(\ln X_i) &= \int_0^1 \ln x\ d\,x = -1\\ E\left(\ln^2 X_i\right) &= \int_0^1 \ln^2 x\ d\,x = 2\\ \operatorname{var}(\ln X_i) &= 2 - (-1)^2 = 1. \end{aligned}$$

Luego cada v.a. $\ln X_i$ tiene un valor medio $m = -1$ y varianza $\sigma^2 = 1$; por lo tanto, el valor medio de Y_n es también -1 y la varianza es $1/n$. Por el teorema central del límite, la distribución de Y_n es asintóticamente normal $N(-1,\ 1/n)$, es decir, para n grande

$$F_{Y_n}\left(\tilde{y}\right) \simeq \int_{-\infty}^{\tilde{y}} \frac{\sqrt{n}}{\sqrt{2\pi}}\, e^{-n(y+1)^2/2}\, d\,y.$$

Además,

$$F_{G_n}(x) = P(G_n \leqslant x) = P(Y_n \leqslant \ln x) = F_{Y_n}(\ln x)$$

$$f_{G_n}(x) = f_{Y_n}(\ln x) \cdot \frac{1}{x} \qquad 0 < x < 1.$$

Luego, para n grande, la función de densidad asintótica de G_n es

$$f(x) = \frac{\sqrt{n}}{\sqrt{2\pi}\, x}\, e^{-n(\ln x+1)^2/2} \qquad 0 < x < 1.$$

1.6. El porcentaje de individuos daltónicos varones de una población es P. Se desea estimar el porcentaje de la población a partir del porcentaje observado sobre una muestra de tamaño n. Calcular el tamaño de la muestra a fin de que el error cometido sea inferior al 1 % con probabilidad 0,90 en los casos:

a) Se sabe que P es inferior al 16 %.

b) No se sabe nada acerca de P.

Solución. Podemos suponer $n \geqslant 30$. Entonces, la frecuencia f del número de daltónicos sobre una muestra de n individuos es aproximadamente normal $N(np,\ npq)$, siendo $p = P/100$. Por lo tanto, el porcentaje de daltónicos $(100\, f/n)$ es aproximadamente $N(P,\ 100^2\, pq/n)$. Luego, con probabilidad 0,90 tendremos:

$$P - 1{,}64 \times 100\sqrt{pq/n} < \hat{P} < P + 1{,}64 \times 100\sqrt{pq/n},$$

siendo $\hat{P} = 100\, f/n$.

a) Si se sabe que $p < 0{,}16$, el máximo de pq se alcanza para $p = 0{,}16$, $q = 1 - 0{,}16$. Tomemos pues $p = 0{,}16$, $q = 0{,}84$. Como el error debe ser inferior al 1 %, hacemos:

$$\begin{aligned} 1{,}64 \times 100\sqrt{0{,}16 \times 0{,}84/n} = 1 \quad \Rightarrow \quad n &= (1{,}64 \times 100)^2 \times 0{,}16 \times 0{,}84 \\ &= 3.614{,}82 \end{aligned}$$

Primera respuesta: $n = 3.615$ individuos.

b) Si no se tiene información sobre P, estudiemos en qué casos $pq = p(1-p)$ es máximo. Derivando e igualando a cero:

$$\frac{d}{dp}\left[p(1-p)\right] = 1 - 2p = 0 \quad \Rightarrow \quad p = 1/2$$

$$\begin{aligned} 1{,}64 \times 100\sqrt{0{,}5 \times 0{,}5/n} = 1 \quad \Rightarrow \quad n &= (1{,}64 \times 100)^2 \times 0{,}5 \times 0{,}5 \\ &= 6.724 \end{aligned}$$

Segunda respuesta: $n = 6.724$ individuos.