

Teorema del límite central

Carles Rovira Escofet

P08/75057/02306



Universitat Oberta
de Catalunya

www.uoc.edu

Índice

Sesión 1

La distribución de la media muestral	5
1. Distribución de la media muestral para variables normales	5
1.1. Caso de desviación típica poblacional conocida	5
1.2. Caso de desviación típica poblacional desconocida.	
La t de Student	8
2. Resumen	10
Ejercicios	11

Sesión 2

El teorema del límite central	13
1. Aproximación de la binomial a la normal	13
1.1. Estudio de la proporción	16
2. El teorema del límite central	17
2.1. Control de calidad	18
3. Resumen	19
Ejercicios	20

La distribución de la media muestral

En esta sesión estudiaremos el comportamiento de la media muestral de una variable. Por ejemplo, supongamos que queremos estudiar la media de la altura de los estudiantes de la UOC: entre ellos hemos seleccionado una muestra al azar, los hemos medido y hemos calculado la media de las alturas de los estudiantes de la muestra; ahora queremos ver cómo se comporta esta media muestral.

Veremos que si sabemos que la variable que se estudia es normal, entonces la media muestral también es normal, pero con desviación típica menor. Y también veremos que si la variable no es normal pero la muestra es lo bastante grande, la media también será aproximadamente normal.

1. Distribución de la media muestral para variables normales

Supongamos que tenemos una muestra x_1, \dots, x_n de una variable aleatoria normal. Recordemos que la media se define como:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

Esta media depende de la muestra. Normalmente tendremos sólo una muestra, pero podríamos tomar muchas diferentes, de manera que a cada una le correspondería una media diferente. Esto nos da pie a hablar de la distribución muestral de la media. Para indicar que se trata de una variable aleatoria, la denotaremos por \bar{X} .

Para estudiarla, deberemos distinguir dos casos: cuando la desviación típica de la variable que medimos es conocida y cuando es desconocida.

1.1. Caso de desviación típica poblacional conocida

Pensemos en el ejemplo de las alturas de los estudiantes de la UOC. Supongamos que en un estudio anterior se había demostrado que las alturas de los estudiantes de la UOC seguían una distribución normal de media 172 cm y desviación típica de 11 cm.

Intuitivamente vemos que la media de las observaciones de la muestra que tenemos debe de ser un valor cercano a 172. También parece razonable pensar que observaciones mayores que la media poblacional, 172, se compensarán con valores menores, y que cuanto mayor sea la muestra, más cercano será el valor de la media muestral a 172.

Observad que...

... para una colección de muestras, tendremos la correspondiente colección de medias muestrales $\bar{x}_1, \dots, \bar{x}_k$.

Desviación poblacional y desviación muestral

La desviación poblacional es la desviación real de la variable, que en este caso suponemos conocida. Cuando calculamos la desviación a partir de muestras, hablamos de *desviación muestral*.

Pensemos ahora que tenemos una muestra de cien estudiantes de la UOC. Hacemos diez grupos de diez estudiantes y hacemos la media aritmética para cada grupo. Obtenemos diez valores, correspondientes a las diez medias $\bar{x}_1, \dots, \bar{x}_{10}$. Parece razonable pensar que la media de estos nuevos datos sería también 172. Por otra parte, también parece razonable pensar que estos nuevos valores sean más cercanos a 172 que los datos originales, ya que en cada una de las medias se nos habrán compensado valores grandes con valores pequeños.

Si la variable que estudiamos sigue una distribución normal con media μ y desviación típica σ conocidas, entonces la media muestral es también normal con la misma media μ y desviación típica $\frac{\sigma}{\sqrt{n}}$, donde n es el tamaño de la muestra. Por tanto, tipificamos la variable \bar{X} y obtenemos que:

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

sigue una distribución normal estándar.

Demostración

La demostración de este resultado es consecuencia de una importante propiedad de las variables aleatorias normales. La propiedad es la siguiente: si X e Y son variables aleatorias independientes con leyes

$$N(\mu_1, \sigma_1^2) \text{ y } N(\mu_2, \sigma_2^2)$$

respectivamente, entonces $X + Y$ tiene una ley:

$$N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$$

En nuestro ejemplo la variable que recoge todas las posibles medias de cada grupo de diez estudiantes sigue una distribución normal de media 172 cm y desviación típica $11 / \sqrt{10} = 3,48$ cm. Observamos que, efectivamente, cuanto mayor es la muestra, menor resulta la desviación típica y, por tanto, hay menos dispersión.

Este cociente que nos da la desviación típica de la media muestral se conoce como *error estándar*.

Si σ es la desviación típica de la población y n el tamaño de la muestra, se define el **error estándar de la media muestral** como:

$$\frac{\sigma}{\sqrt{n}}$$

Observad que...

... el error estándar es cada vez menor cuanto mayor es el tamaño de la muestra.

Ejemplo de error estándar de una media muestral

Consideremos las alturas de los estudiantes de la UOC. Supongamos que sabemos que se trata de una variable aleatoria normal de media 172 cm y desviación típica 11 cm y que hemos tomado una muestra de trescientos estudiantes al azar. Entonces podemos contestar preguntas del tipo siguiente:

a) ¿Cuál es la probabilidad de que la media sea menor que 170 cm?

La distribución de la media muestral es normal de media 172 cm y desviación típica:

$$\frac{11}{\sqrt{300}} = 0,635$$

Tipificamos la variable para obtener una normal (0,1). Debemos calcular:

$$P(\bar{X} < 170) = P\left(\frac{\bar{X} - 172}{0,635} < \frac{-2}{0,635}\right) = P(Z < -3,149) = 0,0008$$

ya que Z es una variable aleatoria normal (0,1).

b) ¿Cuál es la probabilidad de que la distancia entre la media muestral (de esta muestra de trescientos estudiantes) y la media poblacional, 172 cm, sea menor que 1 cm?

Por un razonamiento parecido (si la distancia entre dos números a y b ha de ser menor que k , se debe cumplir: $|a - b| < k$):

$$P(|\bar{X} - \mu| < 1) = P(-1 < \bar{X} - \mu < 1) = P\left(-\frac{1}{0,635} < \frac{\bar{X} - \mu}{0,635} < \frac{1}{0,635}\right) = P(-1,57 < Z < 1,57)$$

donde Z es una variable aleatoria normal (0,1). Si buscamos en las tablas de la ley normal (0,1), vemos que esta probabilidad es igual a 0,8836.

Tenemos así una probabilidad del 0,8836 de obtener un valor para la media muestral que difiera en menos de 1 cm del valor real de la media cuando tomamos una muestra de trescientos individuos.

Observad que en ninguna parte hemos utilizado el hecho de que la media fuese exactamente 172 cm. Es decir, si sabemos que la variable “altura” sigue una normal con una desviación típica de 11 cm y tomamos una muestra de trescientos estudiantes, sabemos que la diferencia entre su media y la media poblacional μ (que quizá no conozcamos) será menor de 1 cm con una probabilidad del 0,8836.

c) Consideremos ahora el problema inverso. Supongamos que desconocemos la media μ de la altura de los estudiantes de la UOC y queremos estudiar una muestra de manera que la diferencia entre la media de la muestra y la de la población μ sea menor que 1 cm con una probabilidad del 0,95. ¿De qué medida tiene que ser nuestra muestra?

Sabemos que la variable estadística tipificada:

$$\frac{\bar{X} - \mu}{\frac{11}{\sqrt{n}}}$$

se distribuye como una normal (0,1). Por otra parte, si observamos las tablas, nos damos cuenta de que si Z es una normal (0,1):

$$P(-1,96 < Z < 1,96) = 0,95$$

Por tanto:

$$0,95 = P\left(-1,96 < \frac{\bar{X} - \mu}{\frac{11}{\sqrt{n}}} < 1,96\right) = P\left(-1,96 \frac{11}{\sqrt{n}} < \bar{X} - \mu < 1,96 \frac{11}{\sqrt{n}}\right)$$

Y si imponemos que la diferencia $\bar{X} - \mu$ debe ser menor que 1 cm, obtenemos:

$$1,96 \frac{11}{\sqrt{n}} < 1$$

Por tanto, $\sqrt{n} > 11 \cdot 1,96$, y así: $n > (11 \cdot 1,96)^2 = 464,8$. Entonces, si tomamos 465 individuos para llevar a cabo el estudio, sabemos que la diferencia entre la media muestral que obtendremos y la media real será menor de 1 cm, con una probabilidad del 0,95. Fijaos en que cuanto mayor sea el tamaño de la muestra, menor será la diferencia entre la media muestral y la poblacional.

Si se multiplican el numerador y el denominador por n , podemos escribir el resultado que hemos visto en este apartado de otra manera.

Si la variable que estudiamos sigue una distribución normal con media μ y desviación típica σ conocida, entonces:

$$\frac{\sum_{i=1}^n X_i - n\mu}{\sqrt{n}\sigma}$$

sigue una distribución normal estándar.

1.2. Caso de desviación típica poblacional desconocida. La t de Student

Fijémonos en que en los ejemplos estudiados anteriormente necesitábamos dos cosas:

- que la variable que se estudiaba fuese normal
- que el valor de la desviación típica de la variable fuese conocido

Estos dos hechos se conocen gracias a estudios previos. A menudo este estudio no se lleva a cabo, pero podemos suponer que la variable es normal. En este caso deberemos hacer una estimación de la desviación típica con la llamada **desviación típica muestral**:

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

de manera que en los cálculos del apartado anterior reemplazaremos la σ por la s . Entonces la distribución muestral de la media ya no es una distribución normal, como sucedía cuando en lugar de s conocíamos el auténtico valor σ de la desviación.

Varios estudios realizados por W.S. Gosset al final del siglo XIX demostraron que en este caso se obtiene una distribución diferente de la normal, aunque para tamaños lo bastante grandes se parecen bastante. Esta nueva distribución se conoce con el nombre de t de Student con $n-1$ grados de libertad. Esto significa que por cada medida de la muestra, n , en realidad tenemos una distribución diferente.

La **distribución t de Student con n grados de libertad**, que denotaremos por t_n , es muy parecida a la distribución normal $(0,1)$: es simétrica alrededor del cero, pero su desviación típica es un poco mayor que la de la normal $(0,1)$, es decir, los valores que toma esta variable están un poco más dispersos. No obstante, cuanto mayor es el número de grados de libertad, n , más se aproxima la distribución t_n de Student a la distribución normal $(0,1)$. Consideraremos que podemos aproximar la t_n por una normal estándar para $n > 100$.

Las variables aleatorias normales son habituales

En muchos casos es habitual suponer que una variable aleatoria es normal. Algunos ejemplos son: el peso o la altura de las personas, el error que cometen los aparatos de medida, el peso de la fruta, las ventas semanales de una tienda, etc.

Observad que...

... en el caso de la desviación típica muestral se divide por $n-1$, no por n .

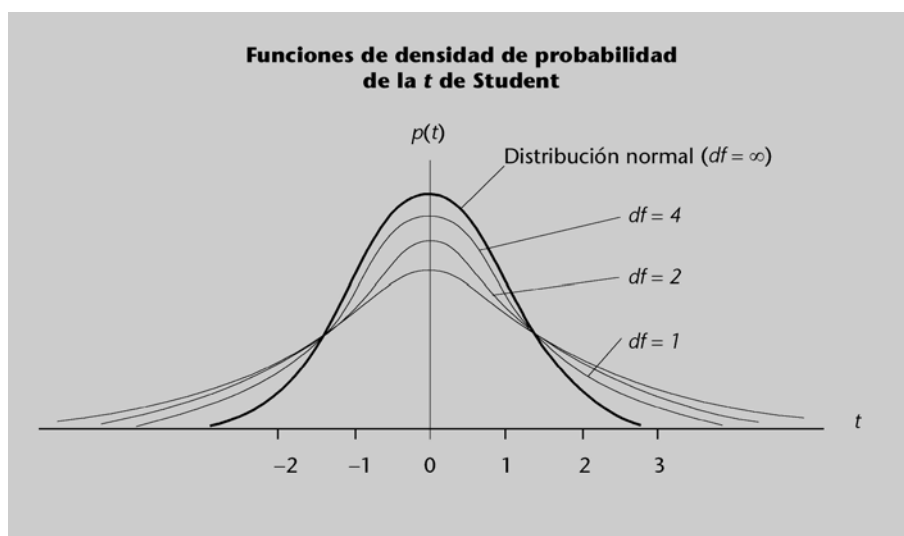
W.S. Gosset

W.S. Gosset trabajaba en la empresa cervecera Guinness y utilizaba el seudónimo de *Student* para firmar sus trabajos.

El valor real y la distribución t_n de Student

Observad que cuando conocemos el valor auténtico de σ , la variable \bar{X} sigue siempre una distribución normal, pero su varianza depende de n .

El gráfico siguiente representa las funciones de densidad de la t de Student para diferentes valores de n y con una línea más gruesa, la densidad de una distribución normal $(0,1)$.



Si σ es desconocida y n es el tamaño de la muestra, calcularemos el error estándar mediante el cociente:

El error estándar es menor cuanto mayor es el tamaño de la muestra.

$$\text{Error estándar} = \frac{s}{\sqrt{n}}$$

Este error estándar nos permite obtener un resultado nuevo importante.

Si la variable que estudiamos sigue una distribución normal con media μ y desviación típica desconocida, entonces:

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

sigue una distribución t_{n-1} , es decir, una t de Student con $n - 1$ grados de libertad.

Obviamente, la manera más fácil de calcular probabilidades relacionadas con una t de Student es con cualquier *software* estadístico o, incluso, una hoja de cálculo. De todos modos, como en el caso de la normal, comentaremos cómo podemos utilizar unas tablas estadísticas.

Las tablas que nos dan la distribución de la t de Student son parecidas a las de la distribución normal estándar. No obstante, y dado que para cada valor de los grados de libertad tenemos una distribución diferente, las tablas habituales sólo nos sirven para ocho probabilidades determinadas (para otros valores hay

que utilizar algún *software* apropiado). La forma de utilizar las tablas es la siguiente: buscamos en la primera columna el número de grados de libertad, nos situamos en aquella fila y determinamos qué puntos nos dejan la probabilidad acumulada que nos interesa.

Ejemplo de utilización de las tablas de la *t* de Student

Una empresa indica en un paquete de arroz que el peso medio del paquete es de 900 gramos. En una inspección hemos analizado el peso en gramos de 10 paquetes de arroz y hemos obtenido los datos siguientes:

890	901	893	893	896
895	894	895	904	899

a) ¿Cuál es la probabilidad de que la distancia entre la media poblacional y la media muestral sea mayor de 3 gramos?

Es razonable pensar que el peso en gramos de un paquete de arroz es una variable aleatoria normal con media del peso que indica el paquete, y con una desviación típica determinada. Es decir, de media los paquetes deberían tener 900 gramos, pero a causa de los errores de medida de los aparatos que los llenan, algunos contendrán un poco más de 900 gramos y otros, un poco menos. Supongamos, pues, que la variable de interés (el peso del paquete) es normal, pero no sabemos nada de su desviación típica. Con nuestros datos podemos estimar la desviación típica y obtenemos:

$$s = 4,19$$

Entonces podemos utilizar el hecho de que $(\bar{x} - \mu) / (s / \sqrt{n})$ es una observación de una *t* de Student con $n - 1$ grados de libertad (en nuestro ejemplo, puesto que tenemos diez datos, será una *t* de Student con nueve grados de libertad). Ahora podemos calcular:

$$\begin{aligned} P(|\bar{X} - \mu| > 3) &= 1 - P(-3 < \bar{X} - \mu < 3) = 1 - P\left(-\frac{3}{\frac{4,19}{\sqrt{10}}} < \frac{\bar{X} - \mu}{\frac{4,19}{\sqrt{10}}} < \frac{3}{\frac{4,19}{\sqrt{10}}}\right) = \\ &= 1 - P(-2,26 < t_9 < 2,26) \end{aligned}$$

donde ya sabemos que t_9 es una *t* de Student con nueve grados de libertad. Podemos calcular esta probabilidad en las tablas:

$$P(-2,26 < t_9 < 2,26) = 1 - 2P(t_9 \geq 2,26) = 1 - 2 \cdot 0,025 = 0,95$$

Entonces:

$$1 - P(-2,26 < t_9 < 2,26) = 1 - 0,95 = 0,05$$

Por tanto, a partir de estos datos, todo parece indicar que la empresa engaña a sus clientes. En efecto, si se toma una muestra de tamaño 10, la probabilidad de que la diferencia entre la media muestral y la real sea mayor de sólo 3 gramos es de un 5%. En cambio, la media de nuestra muestra es de 896 gramos, 4 gramos menos que la cantidad que indica el paquete.

En este caso los valores que nos han aparecido nos han permitido utilizar las tablas. En otras ocasiones necesitaremos utilizar el ordenador.

2. Resumen

En esta sesión hemos estudiado la distribución de la media de datos que provienen de una distribución normal, y hemos diferenciado dos casos: cuando la varianza poblacional es conocida y cuando la varianza es desconocida. Para estudiar este último caso, hemos tenido que introducir la distribución *t* de Student.

Ejercicios

1. El gasto mensual de la familia mexicana Robles sigue una distribución normal de media de 3.000 pesos y varianza 500. Supongamos que el gasto de cada mes es independiente del de los otros meses. Si el ingreso anual es de 37.000 pesos, ¿cuál es la probabilidad de que no gasten más de lo que ganan? ¿Cuánto deberían ganar para tener una seguridad del 99% de que no gastarán más de lo que han ganado?
2. Hemos hecho una encuesta entre los hombres de una población determinada y, a partir de los resultados, deducimos que el peso de los hombres de esta población sigue una distribución normal de media 72 kg. Para saber si los datos que hemos obtenido son fiables, pesamos a cuatro de los encuestados y obtenemos una media de 77,57 kg, con una desviación típica de 3,5 kg. ¿Tenemos suficientes motivos para pensar que los encuestados han mentado cuando nos han dicho su peso?

Solucionario

1. Llamamos X_A al gasto anual. Puesto que el gasto mensual X_M sigue una ley normal de media 3.000 y desviación típica $\sqrt{500}$ y:

$$12 \cdot 3.000 = 36.000 \text{ y } \sqrt{12 \cdot 500} = 77,4597$$

sabemos que $\frac{X_A - 36.000}{77,4597}$ sigue una distribución normal estándar.

Por tanto, la probabilidad de que la familia Robles gaste menos de 37.000 pesos es:

$$P(X_A < 37.000) = P\left(\frac{X_A - 36.000}{77,4597} < \frac{37.000 - 36.000}{77,4597}\right) = P(Z < 12,9099)$$

donde Z es una distribución normal estándar. Si observamos las tablas de la distribución normal estándar, observamos que la probabilidad de que sea menor que 3 ya es 1. Por tanto, la probabilidad es 1, es decir, podemos asegurar con casi un 100% de certeza que no gastarán más de lo que ganan.

Para responder a la segunda pregunta, debemos encontrar una cantidad G tal que:

$$P(X_A < G) = P\left(\frac{X_A - 36.000}{77,4597} < \frac{G - 36.000}{77,4597}\right) = 0,99$$

Si observamos las tablas de la normal, vemos que la cantidad:

$$\frac{G - 36.000}{77,4597}$$

debería ser igual a 2,33 y, por tanto, si resolvemos la ecuación siguiente:

$$\frac{G - 36.000}{77,4597} = 2,33$$

obtenemos que es preciso que $G = 36.180,4811$ para tener una seguridad del 99% de que esta familia no gastará más de lo que gana.

2. Observamos que la diferencia entre la media de nuestros datos y el valor poblacional es de 5,57. Calcularemos la probabilidad de que, si escogemos a cuatro de los encuestados al azar, la media del peso de estos individuos difiera en 5,57 kg o más de la media que conocemos de la población. Por tanto, debemos calcular:

$$P(|\bar{X} - \mu| \geq 5,57)$$

Si esta probabilidad fuese pequeña, nos indicaría que los encuestados seguramente han mentado sobre su peso. Con la ayuda de las tablas, calculamos la probabilidad del complementario:

$$\begin{aligned} P(|\bar{X} - \mu| < 5,57) &= P(-5,57 < \bar{X} - \mu < 5,57) = P\left(-\frac{5,57}{\frac{3,5}{\sqrt{4}}} < \frac{\bar{X} - \mu}{\frac{3,5}{\sqrt{4}}} < \frac{5,57}{\frac{3,5}{\sqrt{4}}}\right) = \\ &= P(-3,18 < t_3 < 3,18) = 1 - 2P(t_3 \geq 3,18) = 1 - 0,05 = 0,95 \end{aligned}$$

donde t_3 es una t de Student con tres grados de libertad. Debemos utilizar la t de Student porque sabemos que la variable de interés sigue una distribución normal, pero desconocemos su desviación típica (sólo tenemos la desviación típica de la muestra). Por tanto:

$$P(|\bar{X} - \mu| \geq 5,57) = 1 - P(|\bar{X} - \mu| < 5,57) = 0,05$$

Así pues, parece que nos han mentado, ya que la probabilidad de que la diferencia entre las medias de los pesos que nos han dicho y 72 es muy pequeña, del orden de 0,05.

Observad que podemos hacer todos estos cálculos con las tablas de la t de Student.

El teorema del límite central

La distribución de la media muestral de una población normal es una distribución normal con la misma media poblacional y con desviación típica el error estándar. Este hecho nos permite calcular probabilidades cuando tenemos una muestra de una variable con distribución normal y desviación típica conocida. Cuando no conocemos la desviación típica de la variable, también podemos hacer cálculos con la distribución t de Student.

En esta sesión veremos cómo debemos proceder cuando no sabemos si la variable de interés sigue una distribución normal o no, o cuando sabemos seguro que su distribución no es normal.

Cuando la muestra es lo bastante grande, la solución nos viene dada por uno de los resultados fundamentales de la estadística: el teorema del límite central. Lo introduciremos con un caso particular: el estudio de la binomial.

1. Aproximación de la binomial a la normal

Supongamos que jugamos diariamente a un número de una lotería que, entre otros premios, devuelve el importe jugado a todos los números que acaban en la misma cifra que el número ganador.

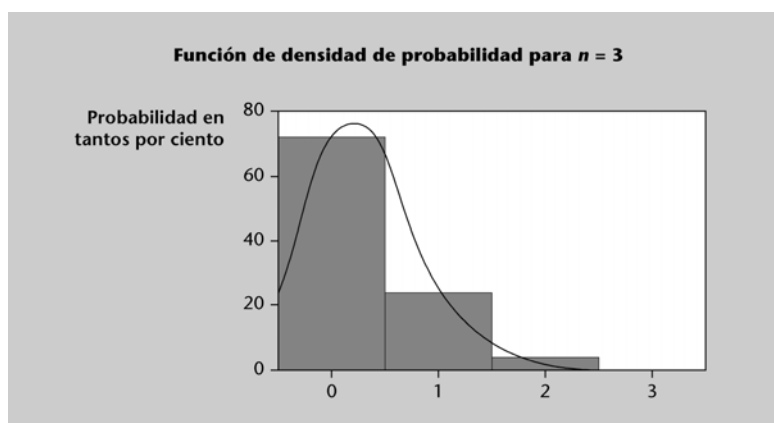
Consideremos la variable $X(n)$, que nos da el número de veces que nos han devuelto el importe jugado cuando se han realizado n sorteos. En este caso sabemos que la variable aleatoria $X(n)$ sigue una distribución binomial de parámetros n y $p = 0,1$. En efecto, se han hecho n sorteos (es decir, se ha repetido un mismo experimento n veces de manera independiente) y en cada sorteo la probabilidad de que nos devuelvan el dinero es $p = 1/10 = 0,1$ (probabilidad de éxito). Sin embargo, observemos qué sucede al aumentar el valor de n con la función de densidad de probabilidad de la variable $X(n)$. Si dibujamos esta función de densidad de probabilidad para $n = 3$, obtenemos el gráfico siguiente:

Binomial

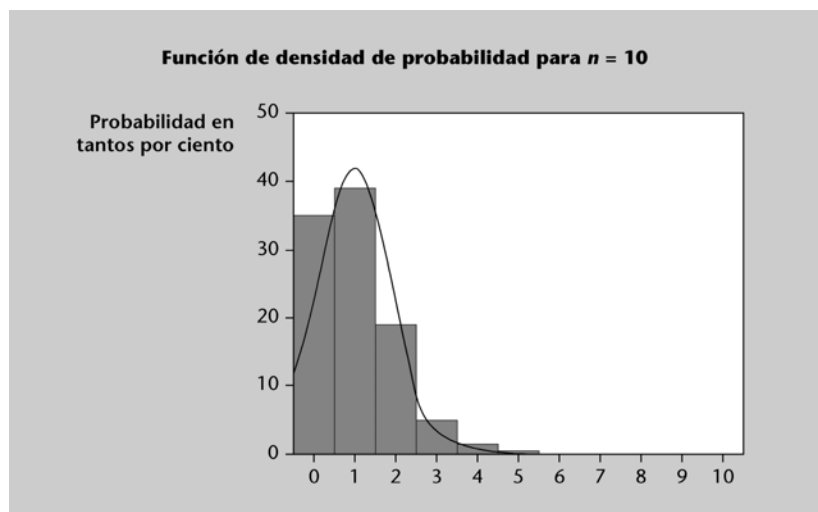
Si X sigue una distribución binomial de parámetros n y p , entonces:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

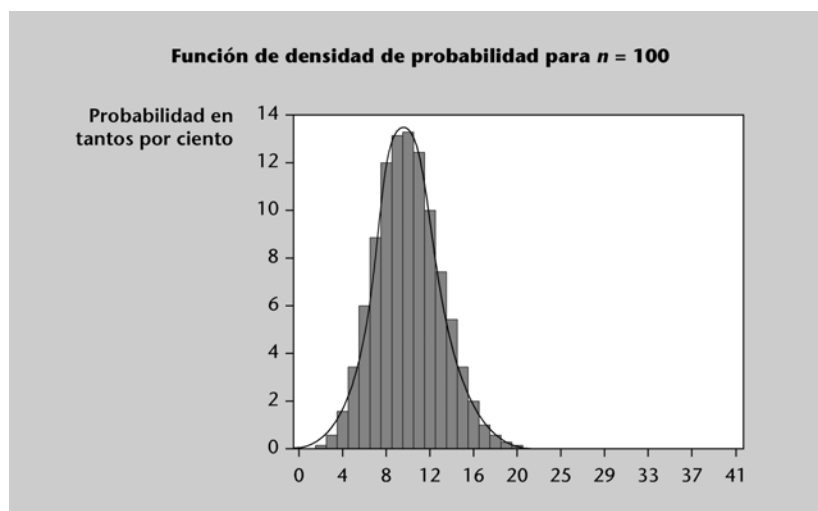
para los $k \in \{0, \dots, n\}$



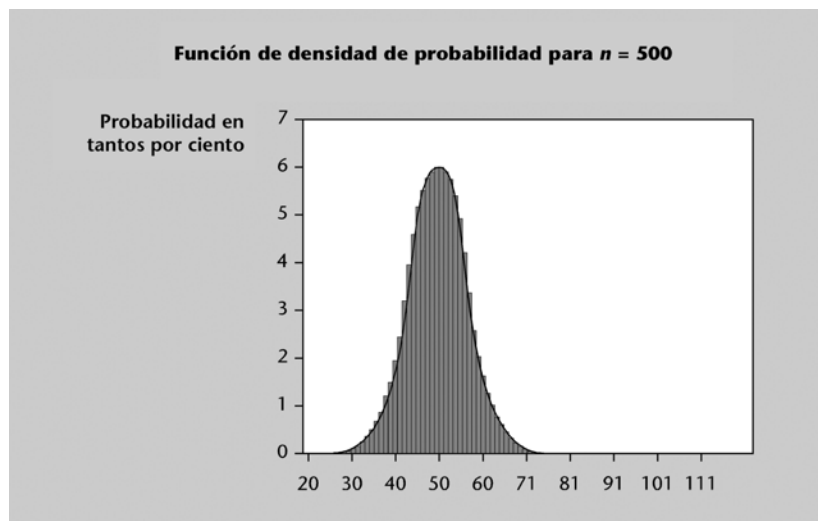
Si ahora consideramos $n = 10$, los posibles valores van del 0 al 10, y el gráfico de la función de densidad de probabilidad es:




Si tomamos $n = 100$, el gráfico es:



Y si por ejemplo tomamos $n = 500$, el gráfico de la función de probabilidad es:



Vemos, pues, que el perfil de este gráfico cada vez se parece más al de la función de densidad de probabilidad de una variable aleatoria normal. La conclusión que extraemos de este experimento es que si n es lo bastante grande, la variable aleatoria $X(n)$ es aproximadamente normal. Determinaremos ahora la media y la desviación de esta variable aleatoria, que serán las correspondientes a la misma $X(n)$: 

- La esperanza de esta variable es:

$$n \cdot p = 0,1 \cdot n$$

- y la varianza:

$$np(1 - p) = n(0,1) \cdot (0,9) = 0,09n$$

Éstos serán los parámetros de la variable aleatoria normal que aproxima la distribución de $X(n)$. Así pues, si n es lo bastante grande, $X(n)$ se comporta como una $N(0,1n; 0,09n)$.

Sea X una variable aleatoria con distribución binomial de parámetros n y p . Si n es grande, entonces la distribución de X es aproximadamente normal con esperanza $\mu = np$ y varianza $\sigma^2 = np(1 - p)$. En la práctica se suele utilizar esta aproximación cuando np y $n(1 - p)$ son mayores que 5, o bien cuando $n > 30$.

Este resultado nos permite simplificar bastante los cálculos en algunas situaciones.

Ejemplo de la lotería

¿Cuál es la probabilidad aproximada de que en un año nos hayan devuelto el dinero al menos cincuenta veces? De hecho, debemos calcular la probabilidad $P(X(365) \geq 50)$. Si quisiéramos obtener el valor exacto de esta probabilidad, por el hecho de que $X(365)$ es una binomial de parámetros 365 y $p = 0,1$, deberíamos hacer el cálculo siguiente:

$$P(X(365) \geq 50) = 1 - P(X(365) < 50) =$$

$$= 1 - P(X(365) = 0) - P(X(365) = 1) - P(X(365) = 2) - \dots - P(X(365) = 49)$$

donde cada una de estas probabilidades se encontraría mediante la fórmula de la binomial que ya conocemos, en nuestro caso:

$$P(X(365) = k) = \binom{365}{k} (0,1)^k (0,9)^{365-k}$$

En cambio, si renunciamos a pedir que la probabilidad sea exacta y nos conformamos con una muy buena aproximación, podemos utilizar el hecho de que la distribución de $X(365)$ se puede aproximar por una normal de parámetros $\mu = 365 \cdot 0,1 = 36,5$ y $\sigma^2 = 365 \cdot 0,09 = 32,85$. Así:

$$P(X(365) \geq 50) = P\left(\frac{X(365) - 36,5}{\sqrt{32,85}} \geq \frac{50 - 36,5}{\sqrt{32,85}}\right)$$

y si llamamos Z a una variable aleatoria normal $(0,1)$, esta probabilidad será aproximadamente:

$$P\left(Z \geq \frac{50 - 36,5}{\sqrt{32,85}}\right) = P(Z \geq 2,36) = 0,0091$$

Por tanto, la probabilidad aproximada de que nos devuelvan el dinero cincuenta veces o más a lo largo del año es únicamente del 0,0091.

Observad que hemos calculado $P(X(365) \geq 50)$, pero que esta cantidad es la misma que $P(X(365) \geq 49,5)$, ya que la variable sólo toma valores naturales. Fijaos en que si la aproximamos por la normal, obtendremos:

$$\begin{aligned} P(X(365) \geq 49,5) &= P\left(\frac{X(365) - 36,5}{\sqrt{32,85}} \geq \frac{49,5 - 36,5}{\sqrt{32,85}}\right) \\ &= P\left(Z \geq \frac{49,5 - 36,5}{\sqrt{32,85}}\right) = P(Z \geq 2,26) = 0,0119 \end{aligned}$$

que es una cantidad ligeramente diferente de la obtenida antes. Se dice que este valor se ha obtenido haciendo una corrección de continuidad, ya que aproximamos una variable discreta por una continua. Podemos considerar buenos los dos resultados.

1.1. Estudio de la proporción

Hemos visto que cuando n es grande, podemos aproximar una binomial (n,p) por una normal de parámetros $\mu = np$ y $\sigma^2 = np(1-p)$. Por otro lado, sabemos que podemos considerar la variable aleatoria binomial como la suma de n variables aleatorias con distribución de Bernoulli de parámetro p . Si dividimos esta suma por n , obtenemos claramente la proporción de éxitos.

Una **proporción** corresponde a hacer la media de n variables aleatorias de Bernoulli de parámetro p , donde n es el tamaño de la muestra y p , la probabilidad de éxito de cada acontecimiento individual.

Ejemplo de cálculo de una proporción

Si queremos calcular la proporción de catalanes que se ha conectado hoy a Internet, podemos considerar que a cada catalán le corresponde una variable Bernoulli que vale 1 si se conecta o 0 si no lo hace. Para calcular la proporción, debemos dividir el número de catalanes que se han conectado por el número total de catalanes.

Puesto que hemos visto que la suma de n distribuciones de Bernoulli de parámetro p , que es una binomial (n,p) , es aproximadamente una distribución normal con media np y varianza $np(1-p)$, está claro que la proporción (que es la suma de las n distribuciones de Bernoulli dividida por n), tendrá esperanza p y desviación típica $\sqrt{p(1-p)/n}$.

Por tanto, cuando el tamaño de la muestra, n , es grande, la distribución de la proporción es aproximadamente una distribución normal de esperanza p y desviación típica $\sqrt{p(1-p)/n}$. En este caso $\sqrt{p(1-p)/n}$ corresponde al error estándar.

Ejemplo de la lotería

En el ejemplo de la lotería podemos pensar que $X(n)$, el número de veces que nos han devuelto el dinero en n sorteos, es una suma de n variables, cada una de las cuales vale 1 si aquel día concreto nos han devuelto el dinero, y 0 en caso contrario. La suma de las n variables nos da el número de veces que nos han devuelto el dinero en los n sorteos, y si dividimos por n obtenemos la proporción de sorteos en los que esto sucede.

Utilidad de las proporciones

La estadística cada vez se utiliza más y las encuestas aparecen todos los días en los diarios. Nos interesa saber qué proporción de electores votarán a un determinado partido, qué proporción de ciudadanos rechaza un determinado plan o una determinada ley que está preparando el gobierno, qué proporción de consumidores estarán interesados en un nuevo producto que queremos lanzar al mercado, etc.

Ejemplo de distribución de la proporción

Preguntamos a una muestra de habitantes de una población su opinión sobre la posible construcción de un pantano. La probabilidad de que un individuo concreto de la población esté de acuerdo con la construcción del pantano es p , y n es el número de habitantes entrevistados. El 30% de los encuestados está a favor de la construcción del pantano, es decir, podemos establecer que $p = 0,3$. Si hemos preguntado a cuatrocientos habitantes, entonces encontramos que la distribución de la proporción de habitantes que están a favor de la construcción del pantano, que denotaremos por p , es:

$$N\left(0,3; \frac{0,3(1-0,3)}{400}\right) = N(0,3; 0,0005)$$

Para calcular la probabilidad de que la proporción de habitantes a favor sea mayor del 40%, deberíamos hacer:

$$P(\hat{p} > 0,4) = P\left(\frac{\hat{p} - 0,3}{\sqrt{0,0005}} > \frac{0,4 - 0,3}{\sqrt{0,0005}}\right) = P(Z > 4,47) = 0$$

donde Z indica una distribución normal estándar.

2. El teorema del límite central

Sabemos que la distribución de la media muestral de una variable normal o bien tiene distribución normal o bien se corresponde con una t de Student. También hemos visto que si las variables originales siguen una distribución de Bernoulli, entonces su media es una proporción y , en este caso, cuando n es lo bastante grande, su distribución muestral también es una normal.

El último resultado es cierto sea cual sea la distribución de los datos originales. Es decir, no es preciso que partamos ni de distribuciones normales ni de distribuciones de Bernoulli, ya que para muestras de tamaños lo bastante grandes, la distribución de la media muestral es normal sea cual sea la distribución original. Este resultado fundamental de la estadística tiene un nombre propio: el *teorema del límite central*.

El **teorema del límite central** dice que si una muestra es lo bastante grande ($n > 30$), sea cual sea la distribución de la variable de interés, la distribución de la media muestral será aproximadamente una normal. Además, la media será la misma que la de la variable de interés, y la desviación típica de la media muestral será aproximadamente el error estándar.

¿Qué significa n bastante grande?

Consideraremos que n es lo bastante grande cuando, como mínimo, $n > 30$.

Una consecuencia de este teorema es la siguiente:

Dada cualquier variable aleatoria con esperanza μ y para n lo bastante grande, la distribución de la variable $(\bar{X} - \mu)/(\text{error estándar})$ es una normal estándar.

Cálculo del error estándar

Recordemos que si la variable tiene una desviación típica conocida σ , el error estándar se puede calcular como σ/\sqrt{n} . Cuando σ es desconocida, calculamos el error estándar como s/\sqrt{n} .

Ejemplo de aplicación del teorema del límite central

Una empresa de mensajería que opera en la ciudad tarda una media de 35 minutos en llevar un paquete, con una desviación típica de 8 minutos. Supongamos que durante el día de hoy han repartido doscientos paquetes.

- a) ¿Cuál es la probabilidad de que la media de los tiempos de entrega de hoy esté entre 30 y 35 minutos?
- b) ¿Cuál es la probabilidad de que, en total, para los doscientos paquetes hayan estado más de 115 horas?

Consideremos la variable $X = \text{"Tiempo de entrega del paquete"}$. Sabemos que su media es 35 minutos y su desviación típica, 8. Pero fijaos en que no sabemos si esta variable sigue una distribución normal. Durante el día de hoy se han entregado $n = 200$ paquetes. Es decir, tenemos una muestra x_1, x_2, \dots, x_n de nuestra variable.

Por el teorema del límite central sabemos que la media muestral se comporta como una normal de esperanza 35 y desviación típica:

$$\frac{8}{\sqrt{200}} = 0,566$$

Si utilizamos esta aproximación, ya podemos contestar a la pregunta a. Debemos calcular:

$$P(30 \leq \bar{X} \leq 35) = P\left(\frac{30-35}{0,566} \leq \frac{\bar{X}-35}{0,566} \leq \frac{35-35}{0,566}\right)$$

que es aproximadamente igual a la probabilidad siguiente:

$$P\left(\frac{30-35}{0,566} \leq Z \leq \frac{35-35}{0,566}\right) = P(-8,83 \leq Z \leq 0) \approx 0,5$$

donde Z es una normal (0,1). Es decir, tenemos una probabilidad aproximada del 0,5 de que la media del tiempo de entrega de hoy haya estado entre 30 y 35 minutos.

Por lo que respecta a la segunda pregunta, de entrada debemos pasar las horas a minutos, ya que ésta es la unidad con la que nos viene dada la variable. Observad que 115 horas por 60 minutos nos dan 6.900 minutos. Se nos pide que calculemos la probabilidad siguiente:

$$P\left(\bar{X} > \frac{6.900}{200}\right) = P(\bar{X} > 34,5)$$

y como que sabemos que la media se distribuye aproximadamente como una normal de media 35 y desviación típica 0,566 (supondremos siempre que la distribución de la media es normal, ya sea porque la variable de interés es normal o porque la muestra es lo bastante grande), esta probabilidad se puede aproximar por la probabilidad de una distribución normal estándar Z :

$$P\left(Z > \frac{34,5-35}{0,566}\right) = P(Z > -0,88) = 1 - P(Z < -0,88) = 1 - 0,1894 = 0,8106$$

2.1. Control de calidad

Uno de los casos más habituales en los que podemos aplicar el teorema del límite central es a la hora de hacer un proceso de control de calidad.

Entenderemos por **control de calidad** el seguimiento de cierta variable aleatoria en un proceso de producción a partir de la media de muestras sucesivas.

Estableceremos un intervalo, de manera que las medias que caigan fuera de este intervalo nos indicarán que existe alguna anomalía en el proceso de pro-

ducción en aquel instante. Los límites de este intervalo se denominan **límites de control**.

Si μ es la esperanza de la variable de interés, σ la desviación típica y consideramos una muestra de esta variable de tamaño n , los límites de control vendrán dados por $\mu + 3 \frac{\sigma}{\sqrt{n}}$ y $\mu - 3 \frac{\sigma}{\sqrt{n}}$. Es decir, calculamos tres veces el error estándar a lado y lado de la media. Por tanto, la longitud del intervalo es dos veces el triple del error estándar.

¿Por qué tomamos este intervalo? Si aplicamos el teorema del límite central sobre la variable de interés, sabemos que la media de n datos se distribuye como una normal con media μ y desviación típica $\frac{\sigma}{\sqrt{n}}$. Se demuestra fácilmente que la probabilidad de que una media esté fuera del intervalo $\mu + 3 \frac{\sigma}{\sqrt{n}}$ y $\mu - 3 \frac{\sigma}{\sqrt{n}}$ es de 0,003 (esto significa que un valor fuera de este intervalo, si el proceso funcionase correctamente, se puede dar sólo con una probabilidad de 0,003). Por tanto, cuando se dé un valor fuera del intervalo, pensaremos que no es casualidad y que el problema es que la variable no se comporta como suponíamos.

Ejemplo de realización de un control de calidad

Consideremos una máquina que llena tarros de yogur. Supongamos que, de media, cada tarro contiene 125 gramos de yogur con una desviación típica de 1,5 gramos. Todas las semanas hacemos un control de la máquina: analizamos una muestra de treinta tarros y calculamos la media de cada uno. En este ejemplo el error estándar es:

$$\frac{1,5}{\sqrt{30}} = 0,274$$

Por tanto, los límites de control serán:

$$\begin{aligned} 125 + 3 \cdot 0,274 &= 125,82 \\ 125 - 3 \cdot 0,274 &= 124,18 \end{aligned}$$

Así pues, si la media de las muestras semanales de tamaño 30 está entre estos dos valores, consideraremos que todo está correcto, mientras que si es inferior a 124,18 o superior a 125,82 supondremos que hay alguna anomalía en el proceso de producción, y habrá que revisarlo.

Por cierto, fijaos en que para hacer este control de calidad sólo se desperdician treinta yogures a la semana.

3. Resumen

En esta sesión hemos presentado un resultado fundamental de la estadística, el teorema del límite central. Lo hemos desarrollado a partir del estudio de una proporción. Hemos acabado viendo una de sus aplicaciones más habituales, la realización de un control de calidad.

Ejercicios

1. En un experimento de laboratorio se mide el tiempo de una reacción química. Se ha repetido el experimento 98 veces y se obtiene que la media de los 98 experimentos es de 5 segundos con una desviación de 0,05 segundos. ¿Cuál es la probabilidad de que la media poblacional μ difiera de la media muestral en menos de 0,01 segundos?
2. Se establece un control de calidad para un proceso de producción de balas. Se ha dispuesto que cuando el proceso está bajo control, el diámetro de las balas es de 1 cm, con una desviación típica de 0,003 cm. Cada hora se toman muestras de nueve balas y se miden sus diámetros. Los diámetros de media de diez muestras sucesivas, en centímetros, son:

1,0006	0,9997	0,9992	1,0012	1,0008
1,0012	1,0018	1,0016	1,0020	1,0022

Estableced cuáles son los límites de control y explicad qué podéis concluir sobre el proceso de producción en estos instantes.

Solucionario

1. Dado que la muestra es grande, por el teorema del límite central podemos suponer que la distribución de la media es una normal de media μ y desviación típica el error estándar. Por tanto, la probabilidad que nos preguntan, que es:

$$\begin{aligned}
 P(|\bar{X} - \mu| < 0,01) &= P(-0,01 < \bar{X} - \mu < 0,01) = P\left(-\frac{0,01}{\frac{0,05}{\sqrt{98}}} < \frac{\bar{X} - \mu}{\frac{0,05}{\sqrt{98}}} < \frac{0,01}{\frac{0,05}{\sqrt{98}}}\right) = \\
 &= P\left(-1,98 < \frac{\bar{X} - \mu}{\frac{0,05}{\sqrt{98}}} < 1,98\right)
 \end{aligned}$$

se puede aproximar por la probabilidad de una distribución normal estándar Z :

$$P(-1,98 < Z < 1,98) = 1 - 2 \cdot 0,0239 = 0,9522$$

Por tanto, la probabilidad que nos piden es de 0,9522.

2. Observamos que la media $\mu = 1$ y que el error estándar es:

$$\frac{\sigma}{\sqrt{n}} = \frac{0,003}{\sqrt{9}} = 0,001$$

Por tanto, los límites de control serán 1,003 y 0,997. Observemos que absolutamente todas las medias que hemos obtenido de las sucesivas muestras están dentro del intervalo formado por los dos límites de control. Es decir, no hay ningún dato superior a 1,003 ni ningún dato inferior a 0,997. Por tanto, podemos concluir que el proceso de control ha sido correcto durante el tiempo que lo hemos analizado, y que no hemos detectado ninguna anomalía.

