Teorema del límite central

2023-11-15

ÍNDICE

La distribución de la media muestral **2**
1.1 Distribución de la media muestral para variables normales 2
1.1.1 Caso de desviación típica poblacional conocida 2
1.2 Caso de desviación típica poblacional desconocida. La t de Student 4
1.3 Ejercicios 7
El teorema del límite central **9**
2.1 Aproximación de la binomial a la normal 9

1 La distribución de la media muestral

Supongamos que queremos estudiar la media de la altura de unos estudiantes. De 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. También veremos que si la variable no es normal, pero la muestra es lo bastante grande, la media también será aproximadamente normal.

1.1 Distribución de la media muestral para variables normales

Supongamos que tenemos una muestra $x_1, ..., x_n$ de una variable aleatoria normal. La media se define como:

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

Esta media depende de la muestra. Normalmente tendremos solo 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 $\overline{X}$.

Deberemos distinguir dos casos: cuando la deviación típica de la variable que medimos es conocida y cuando es desconocida.

1.1.1 Caso de desviación típica poblacional conocida

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*.

Supongamos que en un estudio anterior se había demostrado que las alturas de los estudiantes 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 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.

Pensemos ahora que tenemos una muestra de cien estudiantes. Hacemos diez grupos de diez estudiantes y hacemos la media aritmética para cada grupo. Obtenemos diez valores, correspondientes a las diez medias $\overline{x}_1, ..., \overline{x}_{10}$. 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 $\overline{X}$ y obtenemos que:

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

sigue una distribución normal estándar.

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 el 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 $\frac{11}{\sqrt{10}} = 3.48\ cm$. Observamos que cuanto mayor es la muestra, menor resulta la desviación típica y, por tanto, hay menor 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:

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

Ejemplo de error estándar de una media muestral

Consideremos las alturas de los estudiantes. 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:

$$P(\overline{X} < 170) = P\left(\frac{\overline{X} - 172}{0.635} < \frac{-2}{0.635}\right) = P(Z < -3.149) = 0.0008$$

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 $|a - b| < k$):

$$P(|\overline{X} - \mu| < 1) = P(-1 < \overline{X} - \mu < 1) = P\left(-\frac{1}{0.635} < \frac{\overline{X} - \mu}{0.635} < \frac{1}{0.635}\right) = P(-1.57 < Z < 1.57)$$

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.

Hay que observar 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 300 estudiantes, sabemos que la diferencia entre su media y la media poblacional μ (que quizás no conozcamos) será menor de 1 cm con una probabiliad del 0.8836.

c) Consideremos ahora el problema inverso. Supongamos que desconocemos la media μ de la altura de los trescientos estudiantes 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:

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

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:

$$0.95 = P\left(-1.96 < \frac{\overline{X}-\mu}{\frac{11}{\sqrt{n}}} < 1.96\right) = P\left(-1.96\frac{11}{\sqrt{n}} < \overline{X}-\mu < 1.96\frac{11}{\sqrt{n}}\right)$$

Y si imponemos que la diferencia $\overline{X}-\mu$ 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. 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, 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

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**:

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

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

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 a 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 $n > 100$.

Observamos que cuando conocemos el valor auténtico de σ, la variable $\overline{X}$ 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 $(0, 1)$.

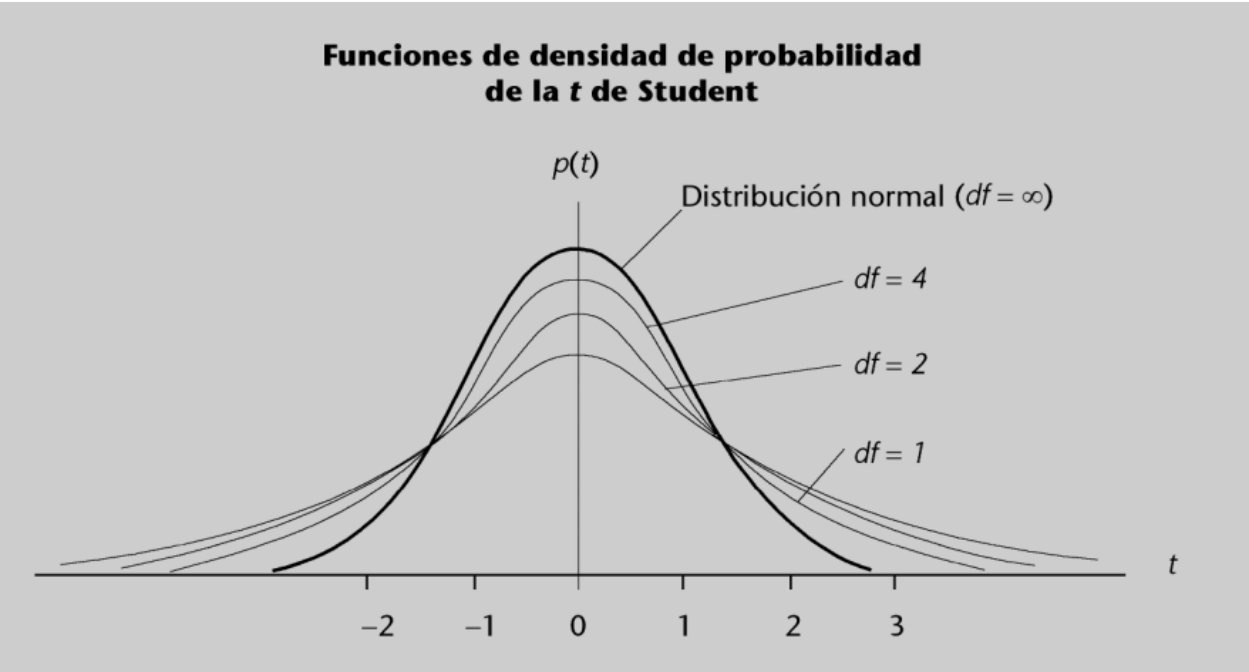

Figure 1: Funciones de densidad de probabilidad de la t de Student

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

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

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:

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

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 probabilidaddes 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 solo 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:

$$s = 4.19$$

Podemos utilizar el hecho de que $(\overline{X} - \mu)/(s/\sqrt{n})$ 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:

$$P(|\overline{X}-\mu| > 3) = 1-P(-3 < \overline{X}-\mu < 3) = 1-P\left(-\frac{3}{\frac{4.19}{\sqrt{10}}} < \frac{\overline{X}-\mu}{\frac{4.19}{\sqrt{10}}} < \frac{3}{\frac{4.19}{\sqrt{10}}}\right) = 1-P(-2.26 < t_9 < 2.26)$$

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 solo 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.

1.3 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?

   a) Llamamos X_A al gasto anual. Puesto que el gasto mensual X_M sigue una ley normal de media 3000 y desviación típica $\sqrt{500}$ y:

   $$12 \cdot 3000 = 36000 \text{ y } \sqrt{12 \cdot 500} = 77.4597$$

   sabemos que $\dfrac{X_A - 36000}{77.4597}$ sigue una distribución normal estándar.

   Por tanto, la probabilidad de que la familia Robles gaste menos de 37000 pesos es:

   $$P(X_A < 37000) = P\left(\frac{X_A - 36000}{77.4597} < \frac{37000 - 36000}{77.4597}\right) = P(Z < 12.9099)$$

   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.

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

   $$P(X_A < G) = P\left(\frac{X_A - 36000}{77.4597} < \frac{G - 36000}{77.4597}\right) = 0.99$$

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

   $$\frac{G - 36000}{77.4597}$$

   debería ser igual a 2.33 y, por tanto, si resolvemos la ecuación:

   $$\frac{G - 36000}{77.4597} = 2.33$$

obtenemos que es preciso que $G = 36180.4811$ para tener una seguridad del 99 % de que la familia no gastará más de lo que gana.

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 mentido cuando nos han dicho su peso?

   Observamos que la diferencia entre la media de nuestros datos y el valor poblacional es de $77.57 - 72 = 5.57$. 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:

   $$P(|\overline{X} - \mu| \geq 5.57)$$

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

   $$P(|\overline{X} - \mu| < 5.57) = P(-5.57 < \overline{X} - \mu < 5.57) = P\left(-\frac{5.57}{\frac{3.5}{\sqrt{4}}} < \frac{\overline{X} - \mu}{\frac{3.5}{\sqrt{4}}} < \frac{5.57}{\frac{3.5}{\sqrt{4}}}\right) =$$

   $$= P(-3.18 < t_3 < 3.18) = 1 - 2P(t_3 \geq 3.18) = 1 - 0.05 = 0.95$$

   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 (solo tenemos la desviación típica de la muestra). Por tanto:

   $$P(|\overline{X} - \mu| \geq 5.57) = 1 - P(|\overline{X} - \mu| < 5.57) = 0.05$$

   Así pues, parece que nos han mentido, ya que la probabilidad de que la diferencia entre las medidas de los pesos que nos han dicho y 72 es muy pequeña, del orden de 0.05.

   Todos estos cálculos se pueden hacer con las tablas de la t de Student.

2 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.

Ahora 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.

2.1 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:

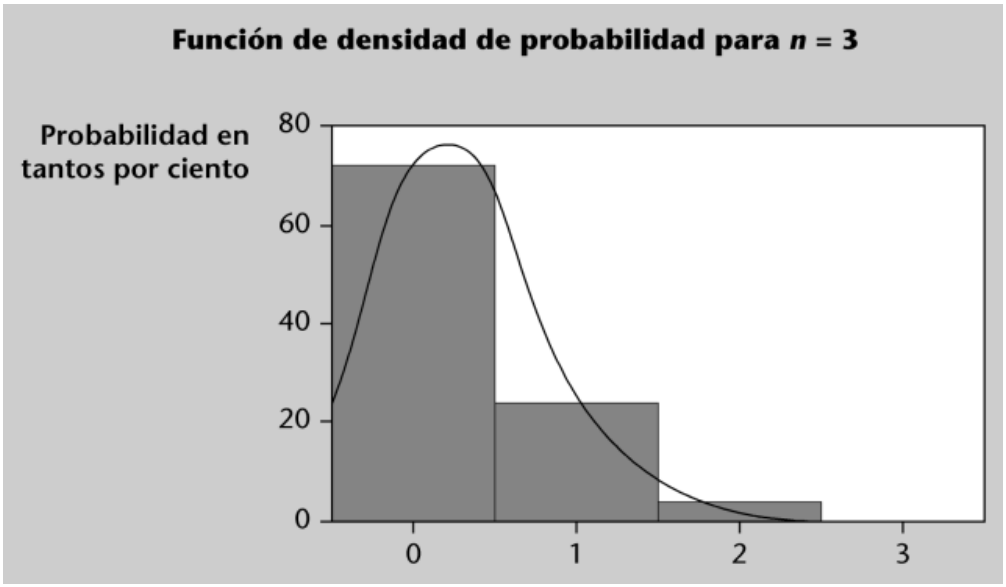

Figure 2: Función de densidad de probabilidad para $n = 3$

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

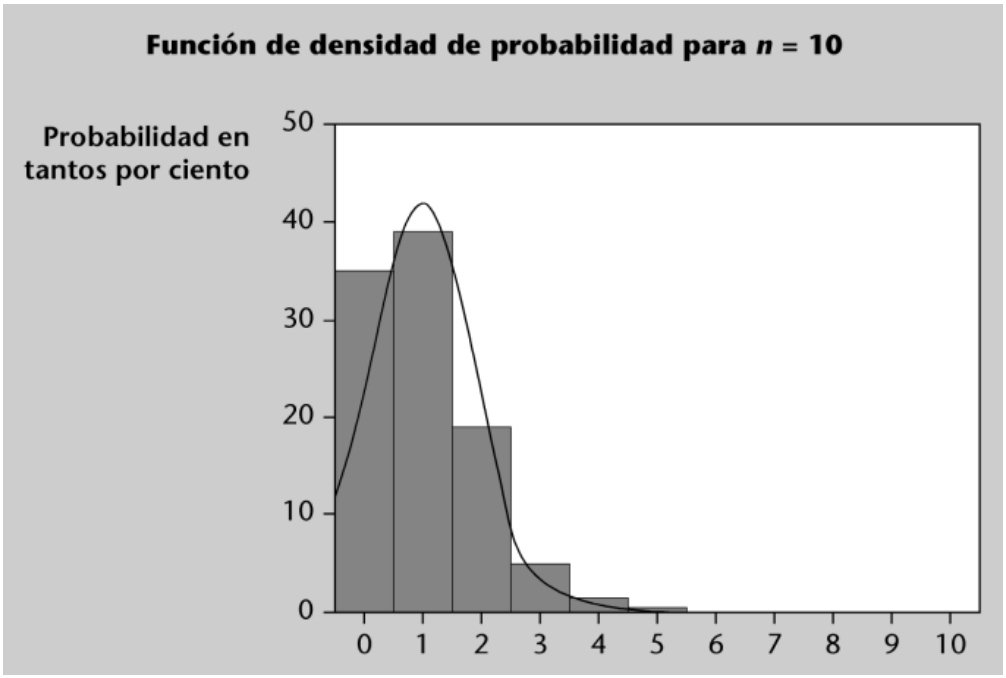

Figure 3: Función de densidad de probabilidad para $n = 10$

Si tomamos $n = 100$, el gráfico es:

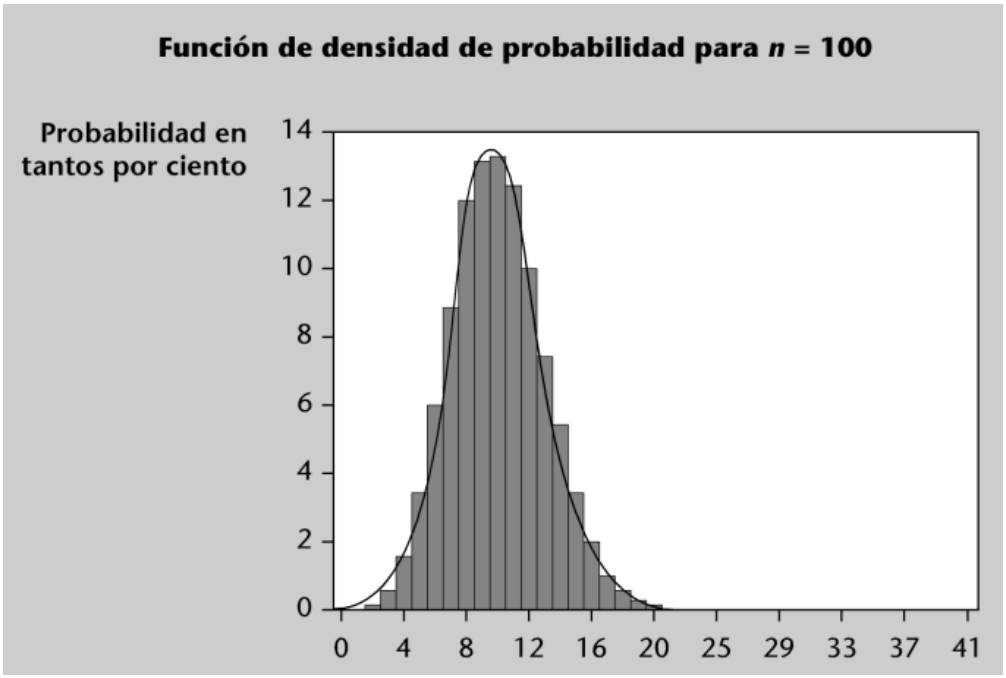

Figure 4: Función de densidad de probabilidad para $n = 100$

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

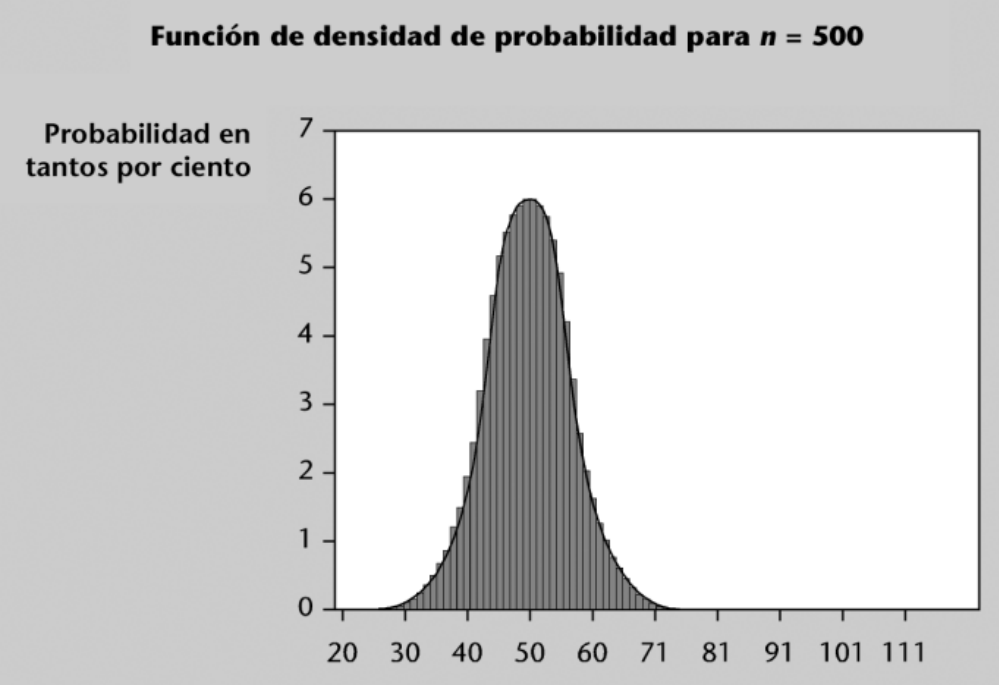

Figure 5: Función de densidad de probabilidad para $n = 500$

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.