Tema 2
Nociones elementales de inferencia estadística

José R. Berrendero

Departamento de Matemáticas
Universidad Autónoma de Madrid

Estructura de este tema

- Conceptos básicos de probabilidad
- La distribución normal
- Estimación de una media poblacional
- Estimación de una proporción poblacional
- Intervalos de confianza para una media y una proporción poblacional

Parámetro y estimador

Un **parámetro** es un número que describe alguna característica de interés de una **población**. En la práctica, siempre tiene un valor desconocido.

Los parámetros que nos van a interesar en este curso son:

- La *media poblacional* (μ) de una variable.
- La *varianza poblacional* (σ^2) de una variable.
- La *proporción poblacional* (p) de individuos que presentan cierta característica.

La población no es conocida en su totalidad, pero suponemos que se dispone de una **muestra** $x_1, \ldots, x_n$.

Un **estimador** es una cantidad que se puede calcular con los datos muestrales y que aproxima el valor de un parámetro de interés.

Ejemplos

Ejemplo 1: Se seleccionan aleatoriamente 200 personas de una ciudad y se les pregunta si han seguido alguna dieta en los últimos cinco años. De las personas seleccionadas 20 responden afirmativamente.

Ejemplo 2: Se seleccionan aleatoriamente 200 personas de una ciudad y se mide su índice de masa corporal (IMC). La media de los IMC medidos es de 22.3.

Determina en los ejemplos anteriores un parámetro poblacional de interés y su correspondiente estimador.

Variables aleatorias

Un **experimento aleatorio** puede dar resultados diferentes aunque se repita bajo condiciones aparentemente idénticas.

Una **variable aleatoria** (v.a.) representa el resultado de un experimento aleatorio.

En el ejemplo 1, el experimento consiste en seleccionar a una persona aleatoriamente y preguntarle si ha seguido o no una dieta. Una variable aleatoria que representa el resultado es:

$$X = \begin{cases} 1, & \text{si la respuesta es afirmativa;} \\ 0, & \text{si la respuesta es negativa.} \end{cases}$$

Probabilidad de un suceso

Al repetir muchas veces un experimento sus posibles resultados suelen presentar un comportamiento regular a largo plazo. Por ejemplo, al tirar muchas veces una moneda, la fracción de veces que sale cara se aproximará al 50%.

La **probabilidad de un suceso** puede interpretarse como el valor al que converge la frecuencia relativa de veces que ocurre ese suceso al aumentar el número de veces que se repite el experimento.

Algunas **propiedades**:

- La probabilidad de un suceso siempre es un valor entre 0 y 1.
- La probabilidad de que un suceso no ocurra es 1 menos la probabilidad de que ocurra: $\mathrm{P}(A^c) = 1 - \mathrm{P}(A)$.
- Dados dos sucesos, la probabilidad de que ocurra alguno de los dos es la suma de sus probabilidades menos la probabilidad de que ocurran los dos a la vez:

$$\mathrm{P}(A \text{ o } B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \text{ y } B).$$

Distribución de una v.a. discreta

La **distribución de una v.a.** X está determinada por los valores que puede tomar X y la probabilidad con la que los toma.

Una v.a. es **discreta** si toma un número finito de valores. Para determinar su distribución basta escribir una lista de los valores que toma y sus respectivas probabilidades.

Ejemplos:

- Si X es la v.a. del ejemplo 1, entonces su distribución es:

Valores	0	1
Probabilidades		

- ¿Cuál es la distribución de la v.a. que representa el resultado de tirar un dado?

Media y varianza de una v.a. discreta

Sea X una v.a. discreta con distribución:

Valores	x_1	$\cdots$	x_k
Probabilidades	p_1	$\cdots$	p_k

La **media o esperanza** de X es:

$$\mu = \mathrm{E}(X) = p_1 x_1 + \cdots + p_k x_k$$

La **varianza** de X es:

$$\sigma^2 = \mathrm{Var}(X) = p_1(x_1 - \mu)^2 + \cdots + p_k(x_k - \mu)^2,$$

donde $\mu = \mathrm{E}(X)$.

La **desviación típica** de X es $\sigma = \sqrt{\mathrm{Var}(X)}$

Ejemplos: calcula la media y la varianza de X

(a) X es la v.a. que representa el resultado de tirar un dado.

(b) X tiene distribución

Valores	0	1
Probabilidades	0.25	0.75

(c) X tiene distribución

Valores	0	1
Probabilidades	$1-p$	p

(d) Sea X el número de aleteos por segundo de una cierta especie de mariposa cuando vuela. Su distribución es:

Valores	6	7	8	9	10
Probabilidades	0.05	0.1	0.6	0.15	

Calcula $P(X \geq 8)$, $\mathrm{E}(X)$ y $\mathrm{Var}(X)$.

La distribución de Bernoulli

Una v.a. de Bernoulli (con parámetro p) es aquella que toma el valor 1 con probabilidad p y el valor 0 con probabilidad $1 - p$.

Siempre que examinamos a n individuos de una población para ver si presentan o no cierta característica tenemos una muestra $x_1, \ldots, x_n$ de variables de Bernoulli.

La media y la varianza son las que se han calculado en el ejemplo anterior.

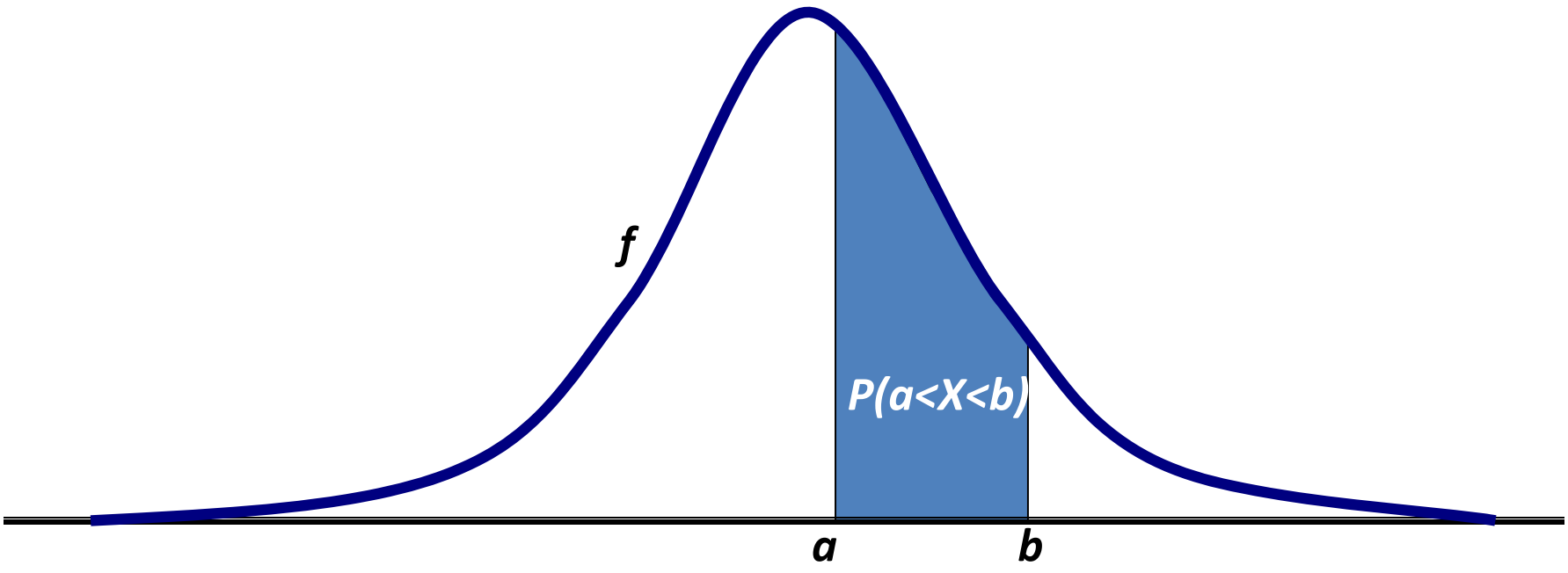
Distribución de una v.a. continua

Una v.a. es **continua** si puede tomar cualquier valor en un intervalo de números.

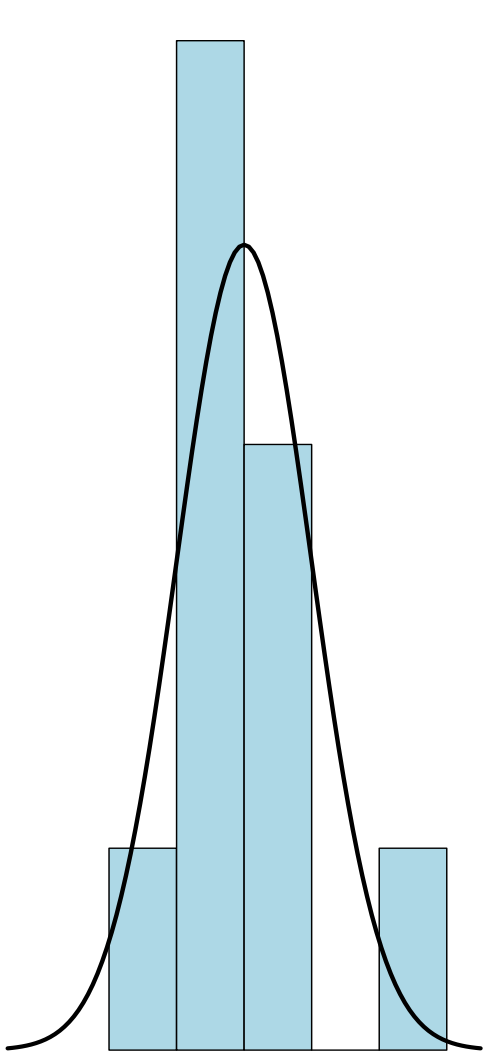
Para determinar la distribución de una v.a. continua no se puede hacer una lista de todas las probabilidades ya que la variable puede tomar infinitos valores.

Se utiliza una **función de densidad** $f \geq 0$ de manera que la probabilidad de un intervalo (a, b), que escribimos $\mathrm{P}(a < X < b)$, es el área entre a y b bajo la función de densidad.

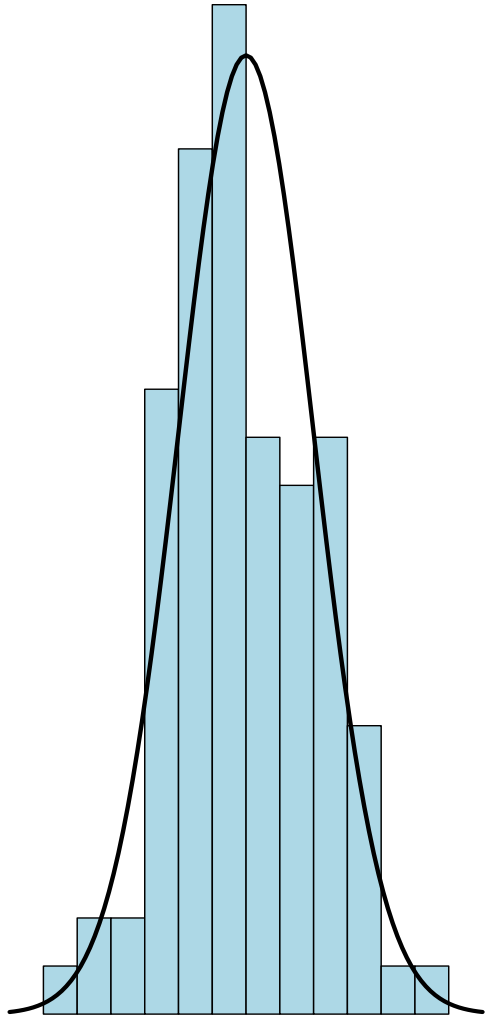
Densidad y probabilidad de un suceso



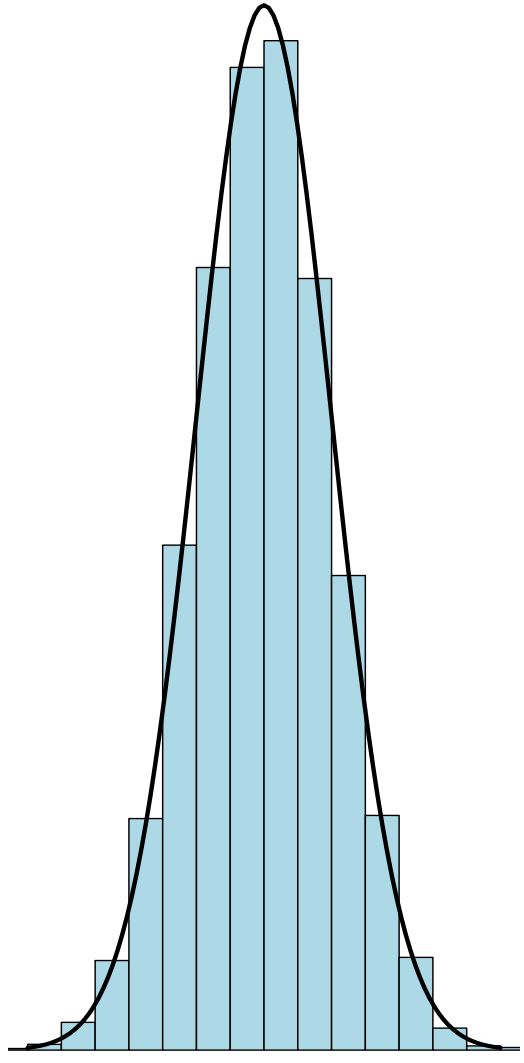
La densidad como límite de histogramas



n=10

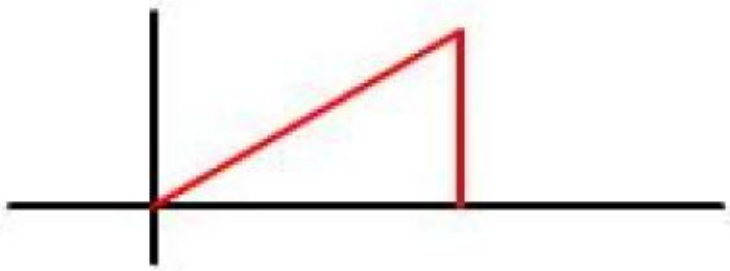


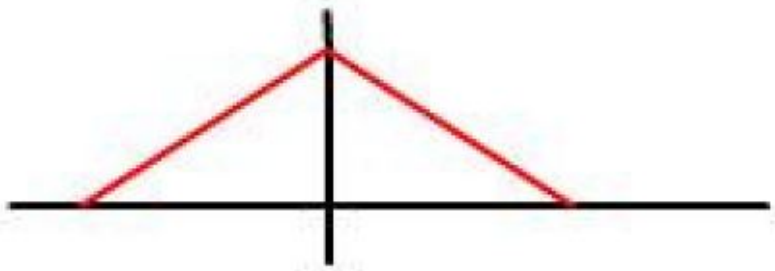
n=100

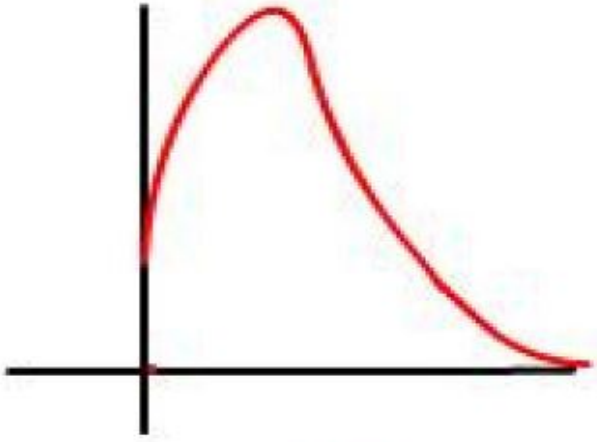


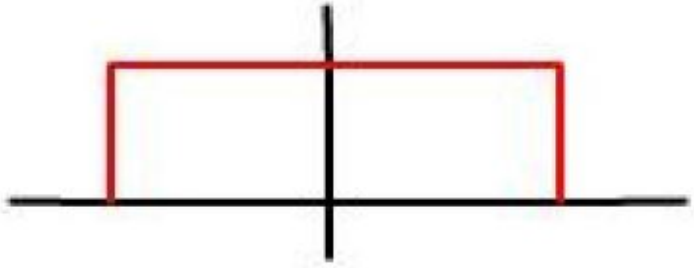
n=10000

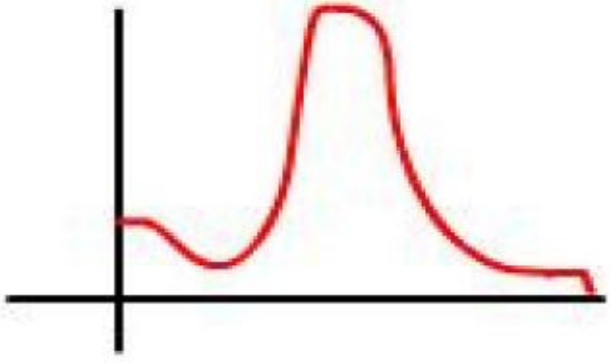
Algunas funciones de densidad

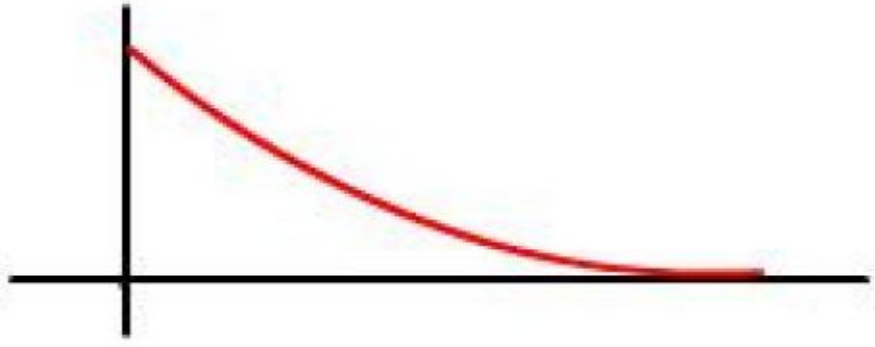












Cuestiones

De las seis funciones de densidad para una v.a. X, indica cuáles verifican cada una de las condiciones siguientes:

- X sólo toma valores positivos.
- X puede tomar cualquier valor positivo.
- La probabilidad de que X tome valores en (a, b) es la misma que la de que tome valores en $(-b, -a)$.
- La v.a. X es un modelo adecuado para el tiempo de vida humana.

Ejemplo

La proporción X de cierto ingrediente en un producto es una v.a. con función de densidad

$$f(x) = \begin{cases} 2x & \text{si } 0 \leq x \leq 1 \\ 0 & \text{si } x \notin [0,1]. \end{cases}$$

Dibuja la densidad y calcula las siguientes probabilidades:

- $P(0 < X < 0.5)$
- $P(X > 3)$
- $P(X < 0.75)$

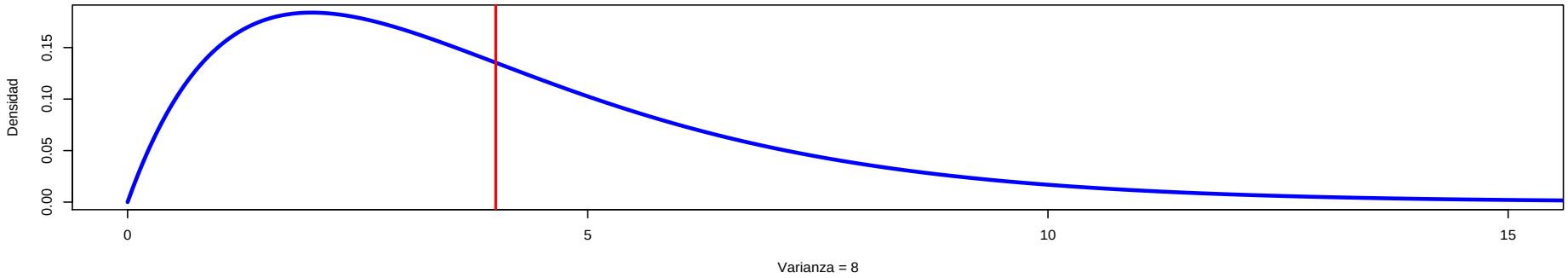
Media y varianza de una v.a. continua

Las v.a. continuas también tienen su media, varianza y desviación típica.

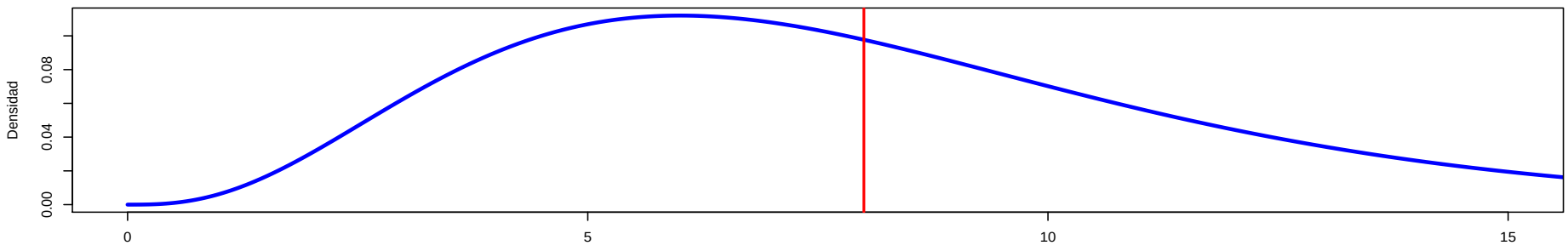
La media de X es el promedio de los valores que toma X ponderado por la probabilidad con la que los toma. En el caso continuo es necesario expresar este promedio como una integral.

La varianza de X es la media de las desviaciones al cuadrado de los valores que toma X respecto a la media de X. También puede expresarse como una integral.

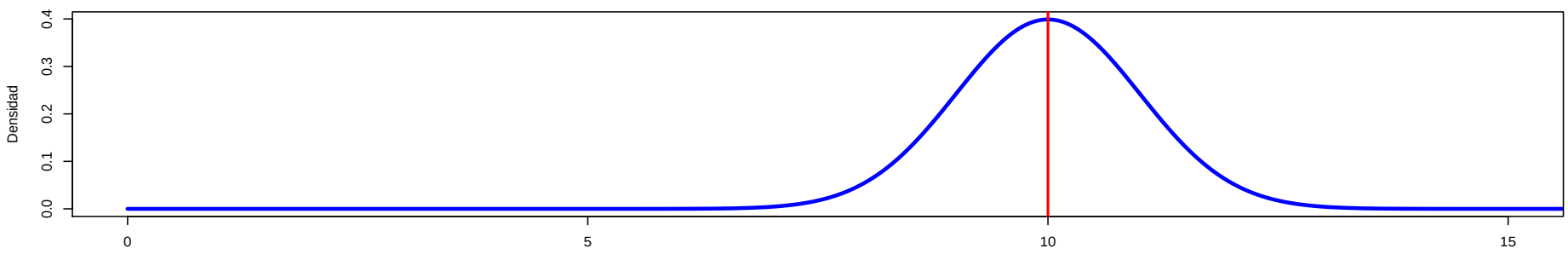
Densidades, media y varianza



Varianza = 8



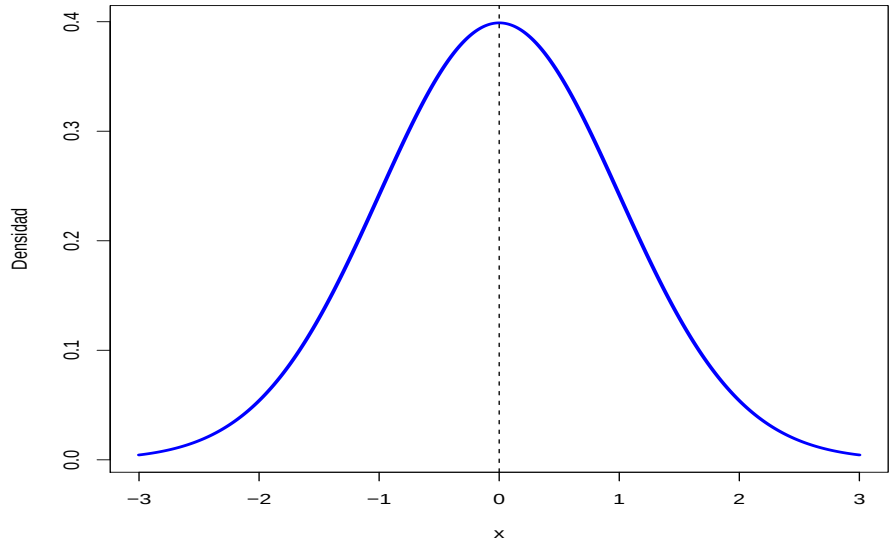
Varianza = 16



Varianza = 1

Distribución normal

Muchos histogramas tienen la siguiente forma aproximada:

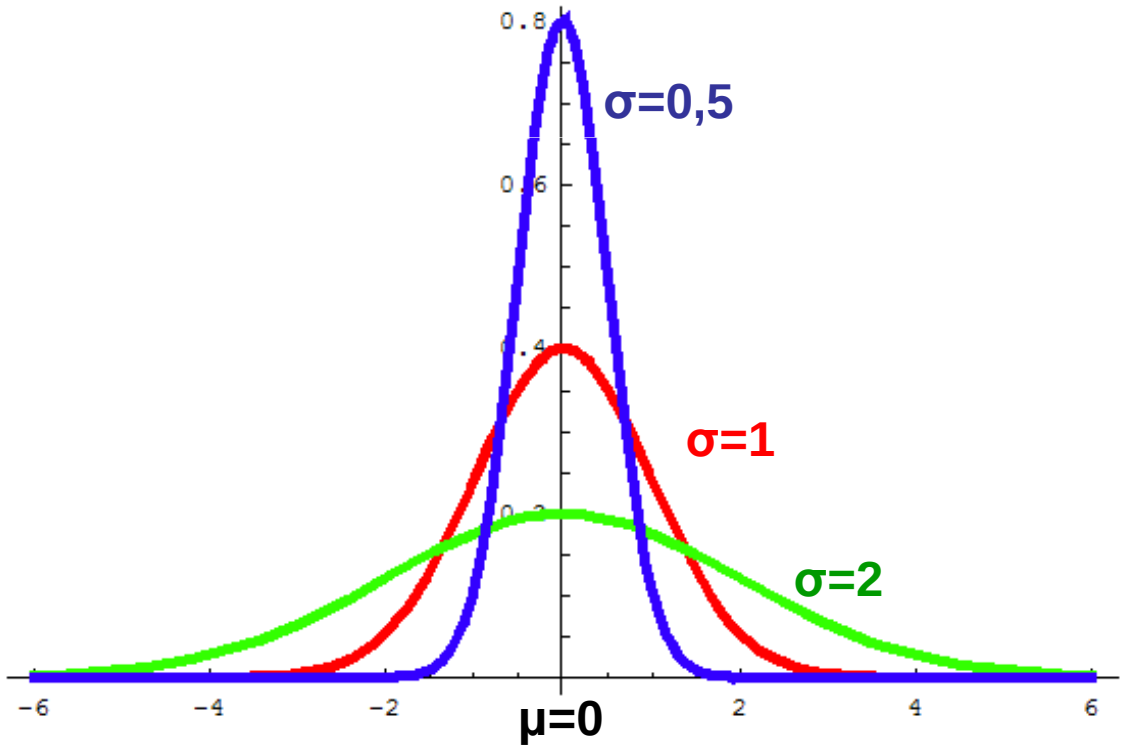


- Simétrica alrededor de un valor central μ.
- A medida que los valores se alejan del centro las frecuencias disminuyen rápidamente.
- La dispersión viene dada por la desviación típica poblacional σ. Los puntos de inflexión se sitúan en los valores $\mu - \sigma$ y $\mu + \sigma$.

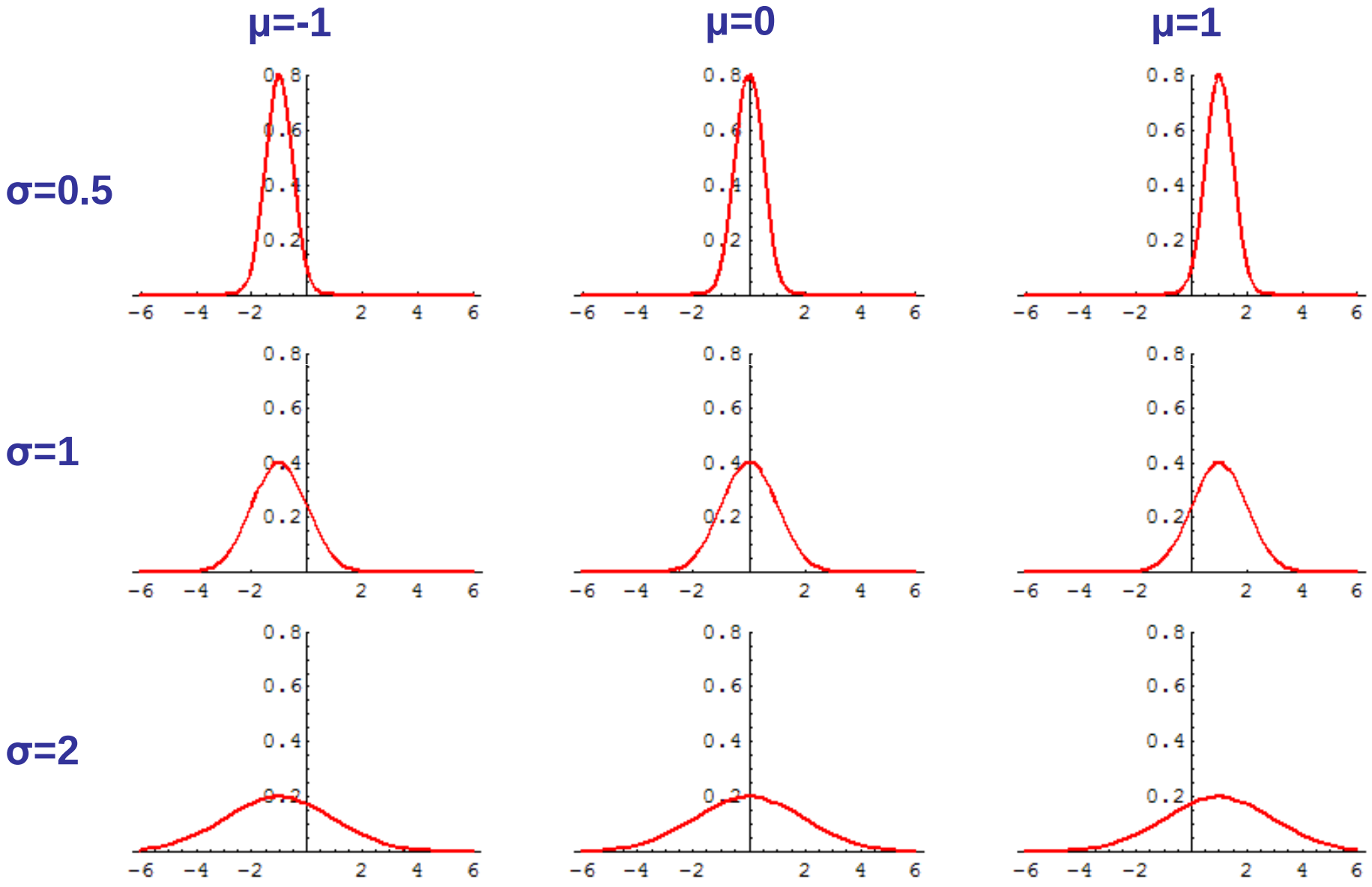
Distribución normal: definición

La v.a. continua X sigue una **distribución normal** $\mathrm{N}(\mu, \sigma)$ de parámetros μ y σ ($\sigma > 0$), si su densidad es

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$$



La curva de densidad normal según μ y σ



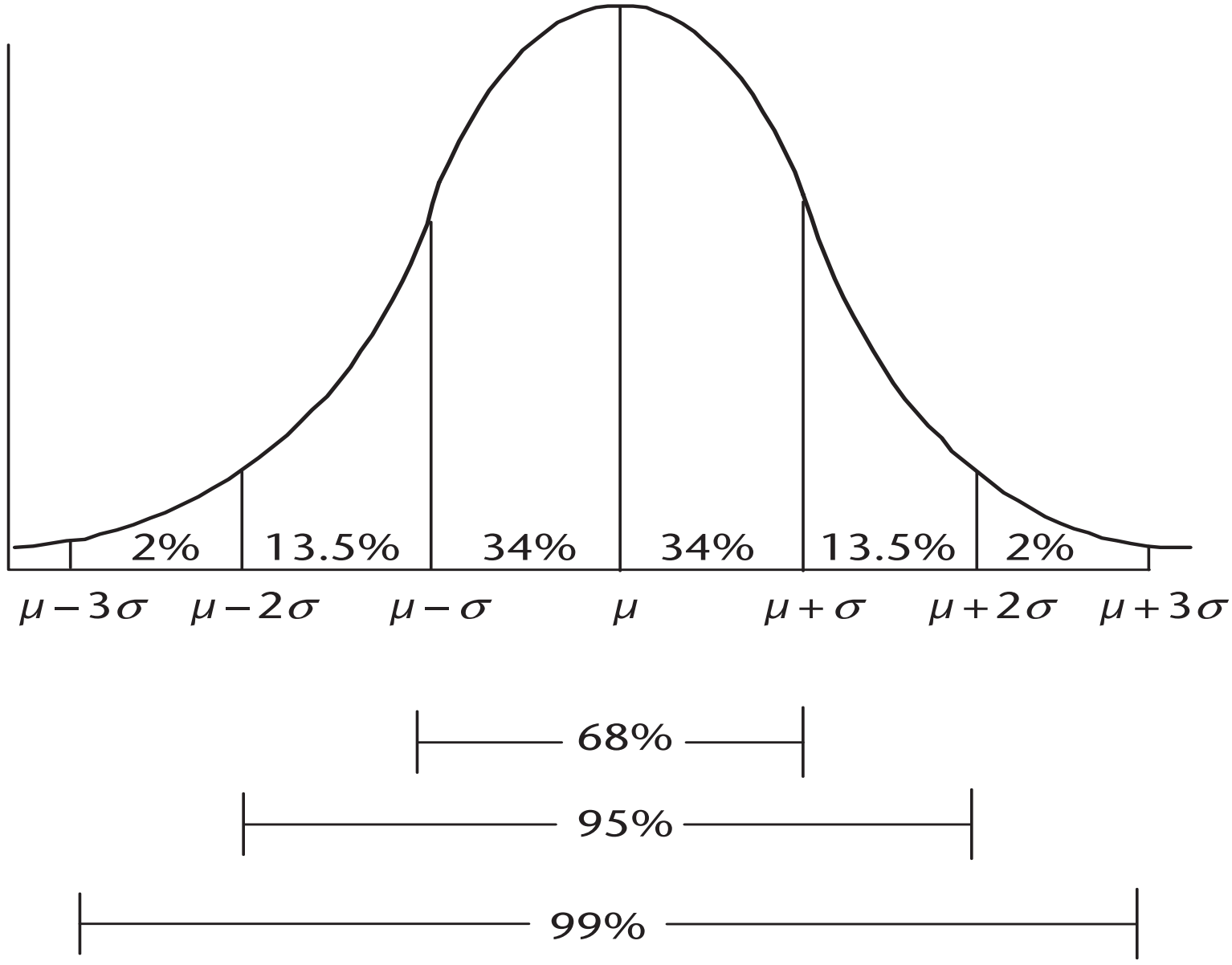
Propiedades importantes de una población normal

(1) Regla 68-95-99

En una población con distribución N(μ, σ):

- Aproximadamente el 68% de los datos está entre $\mu - \sigma$ y $\mu + \sigma$.
- Aproximadamente el 95% de los datos está entre $\mu - 2\sigma$ y $\mu + 2\sigma$.
- Más del 99% de los datos está entre $\mu - 3\sigma$ y $\mu + 3\sigma$.

Propiedades importantes de una población normal



Ejemplo

Sea X la v.a. que representa la cantidad diaria de kcal que toma una persona elegida al azar en una población. Se sabe que la población es normal con media $\mu = 2500$ kcal y desviación típica $\sigma = 100$ kcal. Usando las propiedades anteriores da respuestas aproximadas a las preguntas siguientes:

- ¿Cuál es la probabilidad de que X esté entre 2300 y 2700 kcal?
- ¿Cuál es la probabilidad de que X sea mayor que 2700 kcal?
- ¿Cuál es la probabilidad de que X sea mayor que 2500 kcal?
- ¿Cuál es la probabilidad de que X sea mayor que 2300 kcal?

Propiedades importantes de una población normal

(2) Estandarización

Si una v.a. X tiene distribución $\mathrm{N}(\mu, \sigma)$, entonces la variable estandarizada

$$Z = \frac{X - \mu}{\sigma}$$

tiene distribución $\mathrm{N}(0, 1)$ (normal estándar).

Como consecuencia, solo necesitamos tablas para la normal estándar.

Tablas de la distribución normal estándar

Desv. normal x	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641
0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010

Ejercicios para usar las tablas

Z una v.a. normal estándar. Mira en las tablas para hacer los ejercicios siguientes:

$P(Z > 1)$	$P(Z > c) = 0.025$
$P(Z < -1)$	$z_{0.05}$
$P(-1 < Z < 1)$	$z_{0.95}$
$P(-2 < Z < 1)$	$z_{0.1}$

X es una v.a. normal con $\mu = 1$ y $\sigma = 2$. Mira en las tablas para hacer los ejercicios siguientes:

$P(X > 3)$	$P(X > c) = 0.96$

Propiedades importantes de una población normal

(3) Producto de una normal por una constante

Si $a \in \mathbb{R}$ y $X \equiv \mathrm{N}(\mu, \sigma)$, entonces:

$$aX \equiv \mathrm{N}(a\mu, |a|\sigma)$$

Ejemplo: Si $X \equiv \mathrm{N}(5, 1)$, determina la distribución de
(a) $-X$
(b) $10X$
(c) $-10X$
(d) $X/2$

Propiedades importantes de una población normal

(4) Suma de v.a. normales independientes

Si $X_1 \equiv \mathrm{N}(\mu_1, \sigma_1)$ y $X_2 \equiv \mathrm{N}(\mu_2, \sigma_2)$ independientes, entonces

$$X_1 + X_2 \equiv \mathrm{N}(\mu_1 + \mu_2, \sqrt{\sigma_1^2 + \sigma_2^2})$$

Ejemplo: Si $X_1 \equiv \mathrm{N}(5, \sigma = 1)$ y $X_2 \equiv \mathrm{N}(5, \sigma = 1)$ y ambas variables son independientes, determina la distribución de
(a) $X_1 + X_2$
(b) $X_1 - X_2$
(c) $2X_1 + X_2$
(d) $(X_1 + X_2)/2$

Estimación de la media poblacional

Para estimar la media de una población, μ, el estimador más natural es la media muestral $\bar{x}$.

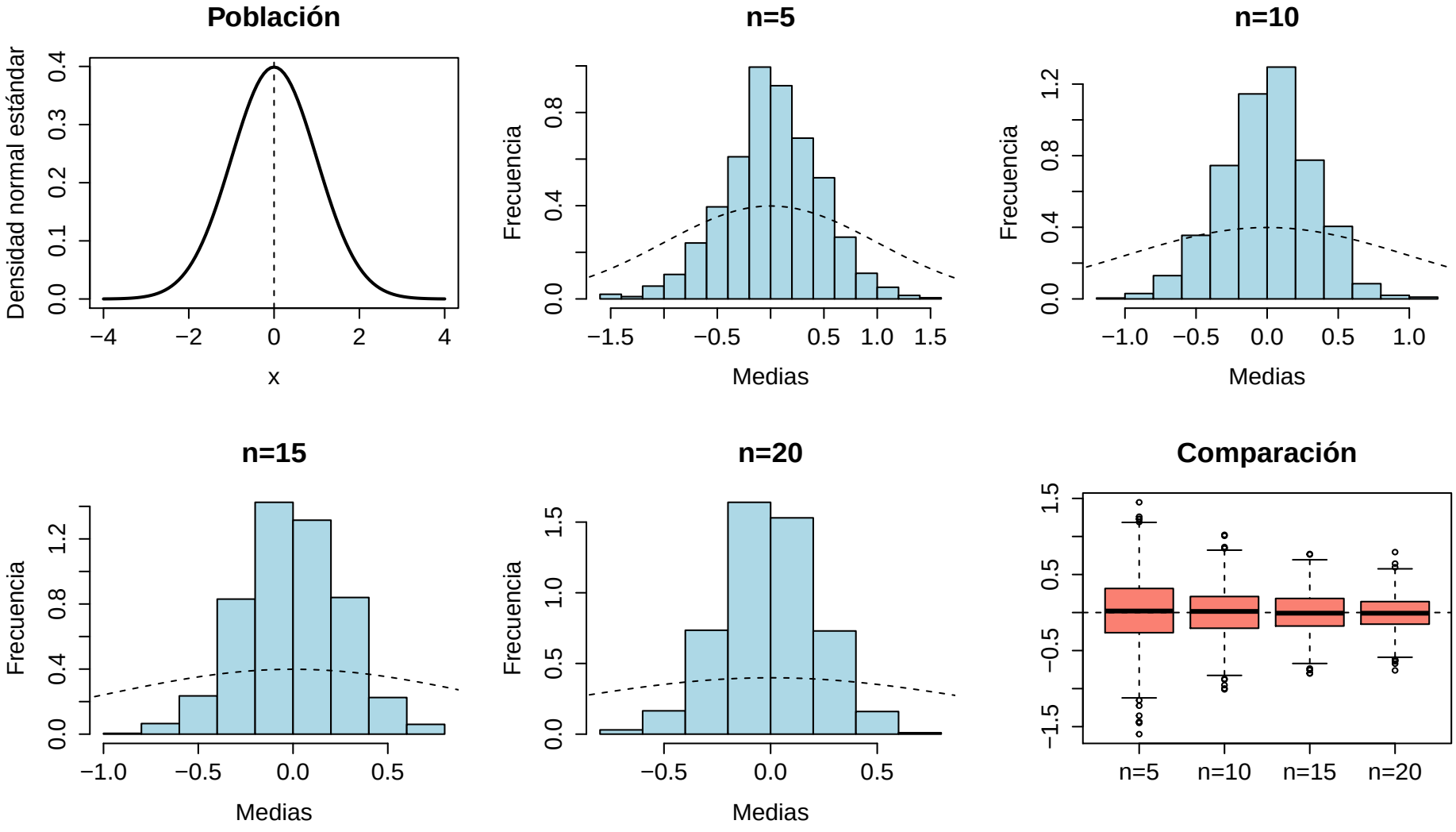
¿Cuál es la calidad de la estimación?

Un estimador es una variable aleatoria ya que su valor depende de la muestra concreta de la que se dispone y la selección de la muestra es aleatoria.

La precisión de la estimación se mide analizando lo que ocurriría si dispusiéramos de muchas muestras y pudiéramos evaluar la media para cada una de ellas.

Tenemos que estudiar la distribución de $\bar{x}$.

Distribución de la media muestral en una población normal



Distribución de la media muestral en una población normal

Si $x_1, \ldots, x_n$ son datos independientes procedentes de un población normal de media μ y desviación típica σ,

$$\bar{x} \equiv \mathrm{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

Observaciones:

- Para cualquier n, el valor esperado de $\bar{x}$ coincide con la media de la población.
- Al aumentar n, la desviación típica disminuye y la probabilidad de obtener valores de $\bar{x}$ cercanos a μ aumenta.

Ejemplos

- Un laboratorio pesa el filtro de una mina de carbón para medir la cantidad de polvo ambiental en la mina. Debido a imprecisiones en los aparatos, las medidas tienen distribución normal con media el verdadero peso μ mg, que es desconocido, y desviación típica $\sigma = 0.08$ mg.

  (a) Se calcula la media de 3 medidas realizadas con el filtro:

  $$\bar{x} = \frac{x_1 + x_2 + x_3}{3}.$$

  ¿Cuál es la distribución de $\bar{x}$?

  (b) ¿Cuál es la probabilidad de que $\bar{x}$ diste de μ menos de 0.05 mg?

  (c) ¿Cuál es la probabilidad de que una única medida del filtro diste de μ menos de 0.05 mg?

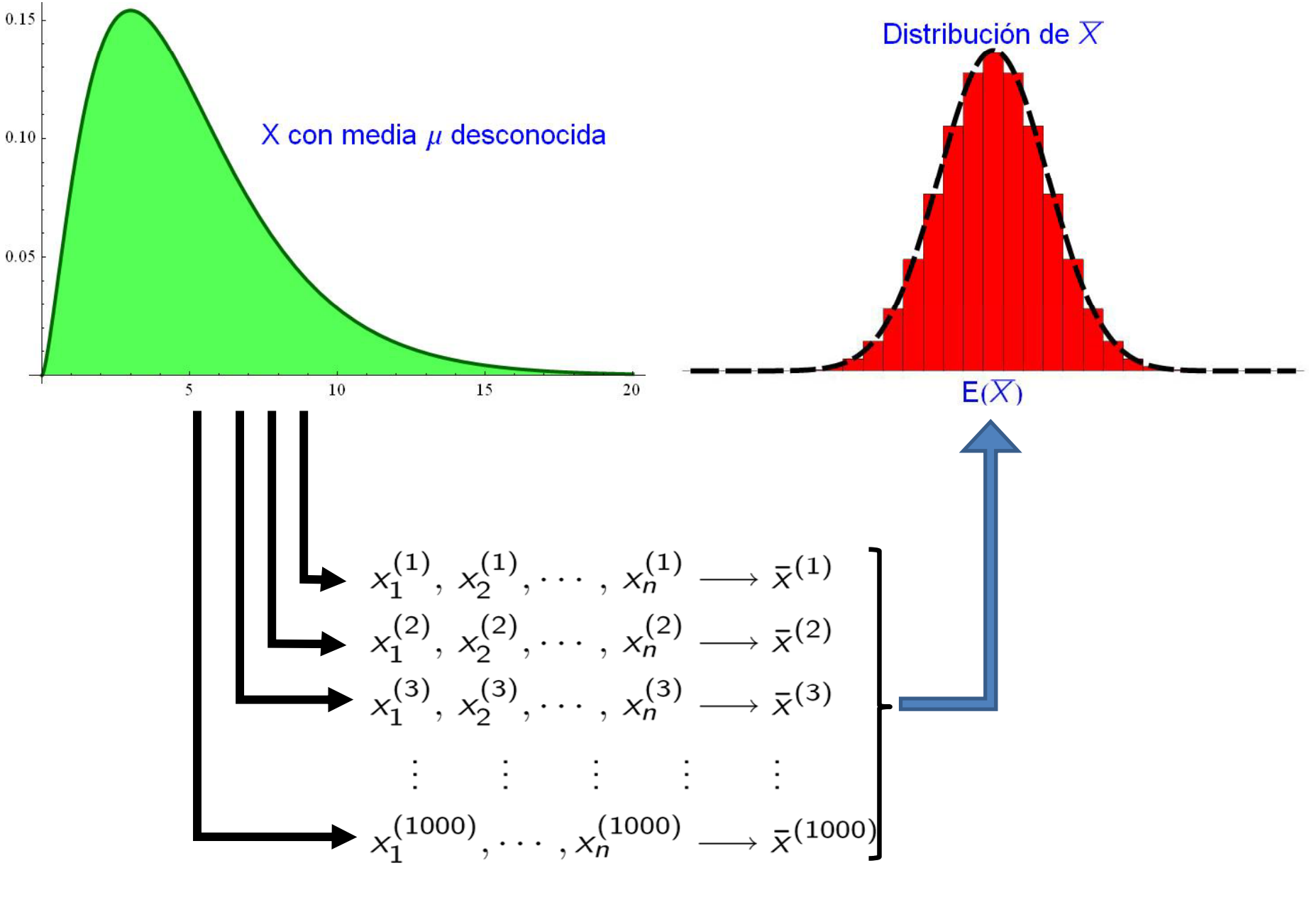
- De acuerdo con la Organización Mundial de la Salud un individuo tiene sobrepeso si su índice de masa corporal (IMC) es superior a 25. Se sabe que el IMC de una población es una variable con distribución normal de media $\mu = 26$ y desviación típica $\sigma = 6$.

  (a) Calcula la probabilidad de que un individuo seleccionado al azar en esta población presente sobrepeso.

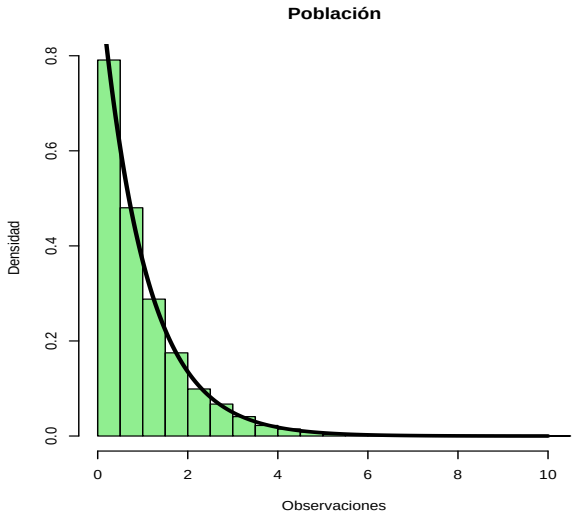
  (b) Calcula el valor x tal que el IMC del 25% de la población es menor que x.

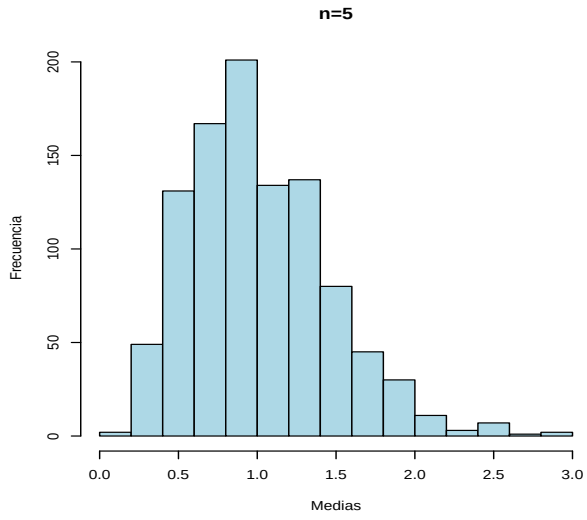
  (c) Si se seleccionan aleatoriamente 100 individuos y se calcula la media de sus IMC, ¿cuál es la probabilidad de que esta media sea superior a 25.5?

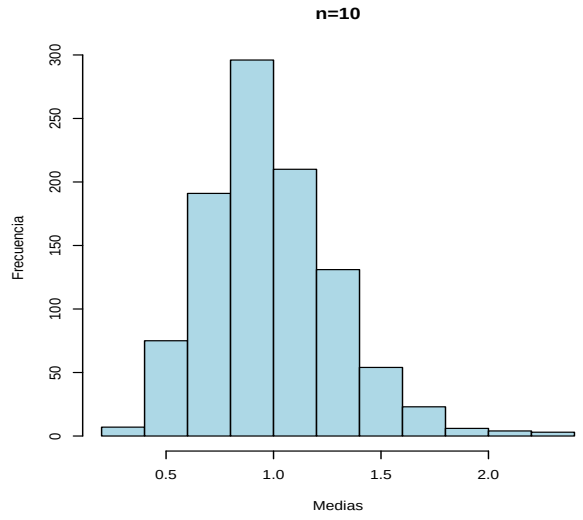
Distribución de la media muestral

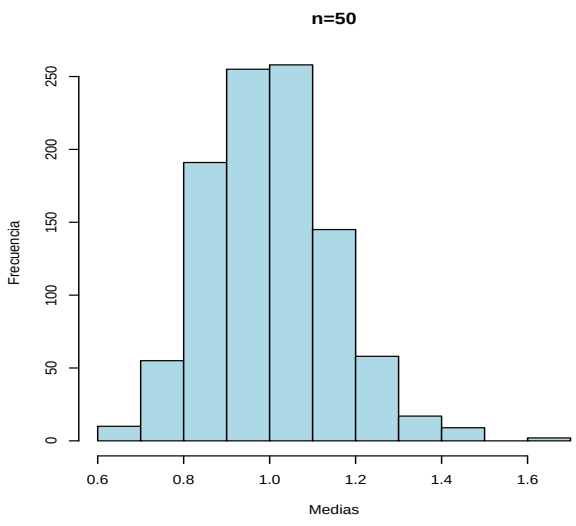


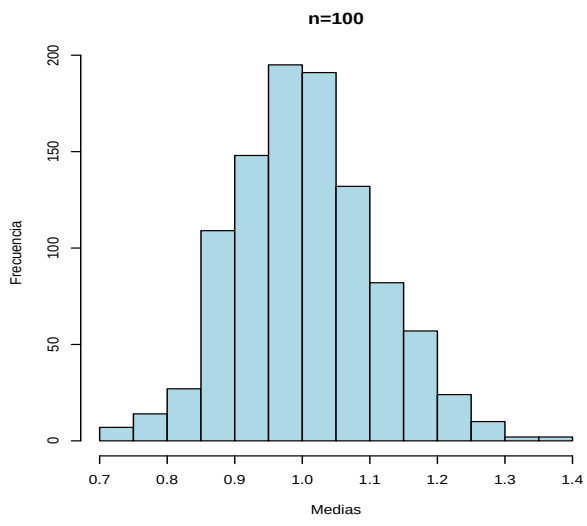
Distribución de la media muestral

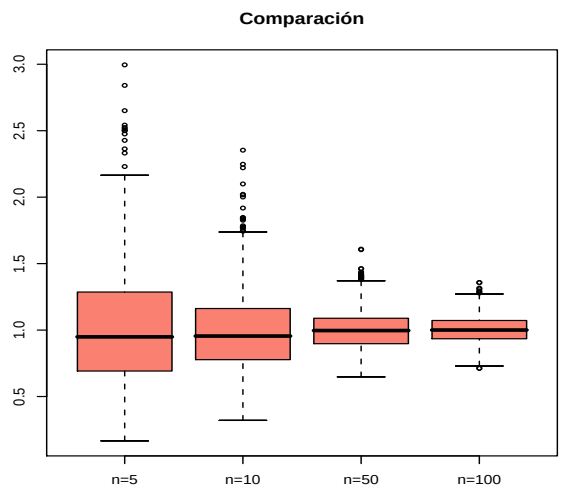












Distribución de la media muestral

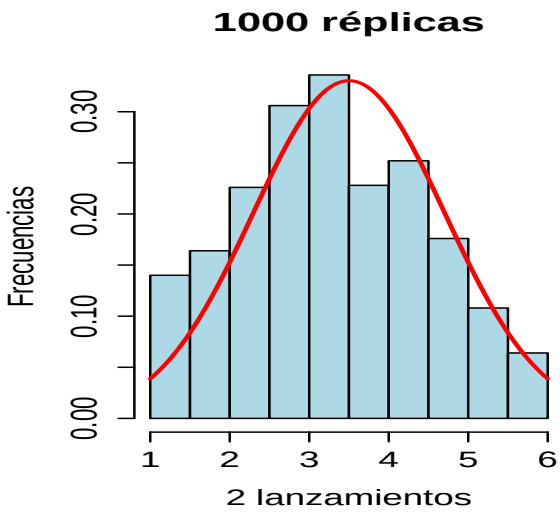
Teorema central del límite: Sea $\bar{x}$ la media de una muestra de tamaño n de una población con media μ y desviación típica σ. Entonces, si n es grande la distribución de los valores que toma $\bar{x}$ es aproximadamente normal de media μ y desviación típica $\sigma/\sqrt{n}$

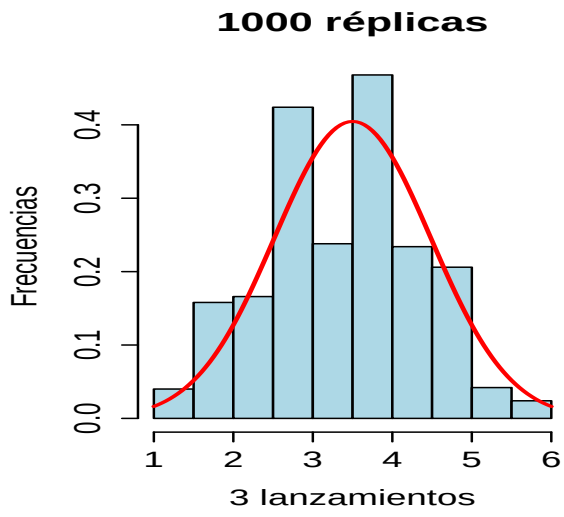
En notación matemática, podemos escribir:

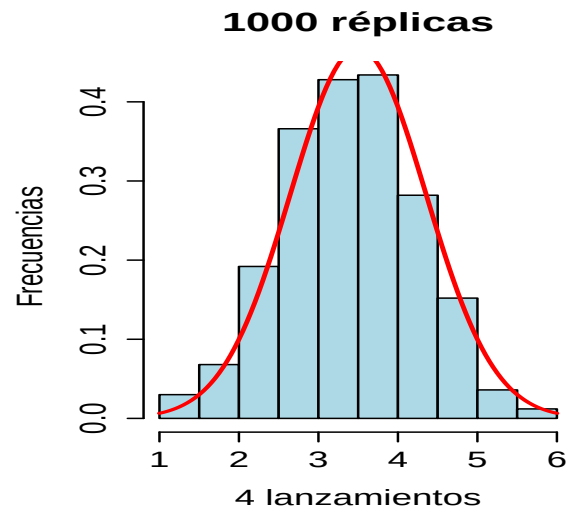
$$\bar{x} \cong \mathrm{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

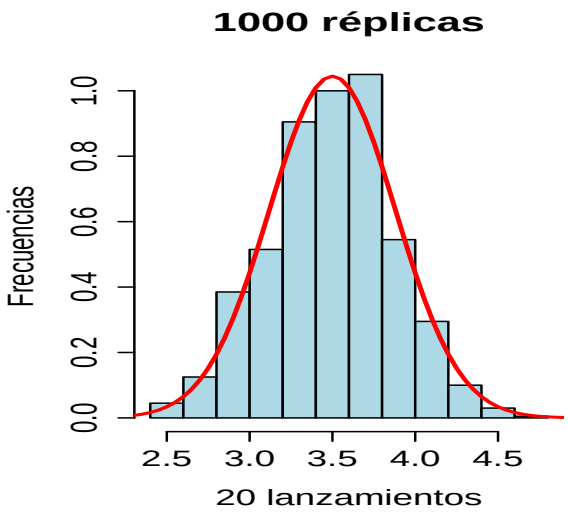
Si la población de partida es normal, el resultado anterior es cierto **de forma exacta para cualquier tamaño muestral n.**

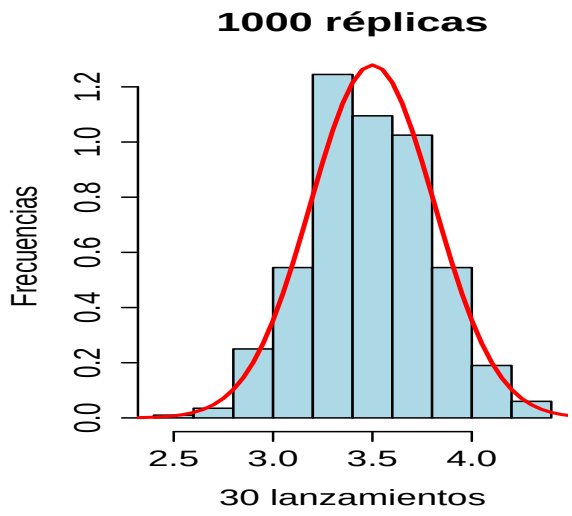
Simulación del promedio al lanzar un dado

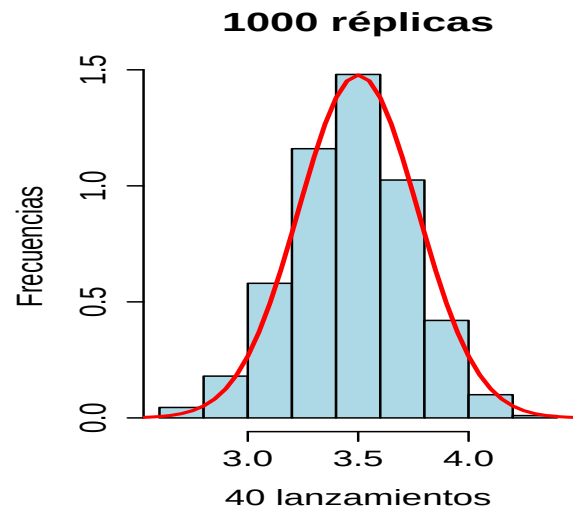












Ejemplos

- El tiempo de espera de los estudiantes de la UAM hasta que llega el tren a la estación de Cantoblanco es una variable aleatoria con media y desviación típica de 10 minutos. Si se obtiene el promedio de los tiempos de espera de 100 estudiantes (que llegan a la estación en días y horas diferentes, de manera que los tiempos se pueden considerar independientes), calcula la probabilidad aproximada de que este promedio sea superior a 11 minutos.

- El peso de los huevos producidos por una gallina tiene distribución aproximadamente normal de media $\mu = 65$ g y desviación típica $\sigma = 5$ g. ¿Cuál es la probabilidad de que una docena de huevos pese entre 750 y 825 g?

Error típico de la media muestral

El **error típico** de un estimador es un estimador de su desviación típica.

La desviación típica de la media es $\sigma/\sqrt{n}$, pero en la práctica σ es un parámetro poblacional desconocido.

Resulta natural estimar σ^2 con la cuasivarianza muestral:

$$S^2 = \frac{(x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n-1}.$$

Se divide $n-1$ ya que puede demostrarse que al dividir por n el estimador tiene una tendencia sistemática a infraestimar σ^2.

El error típico de la media muestral es

$$\frac{S}{\sqrt{n}}$$

Error típico de la media muestral

¿Sabes distinguir entre los conceptos siguientes? ¿Qué notación estamos usando para cada uno de ellos?

- La varianza de la población
- La desviación típica de la población.
- La varianza de la media muestral.
- La desviación típica de la media muestral.

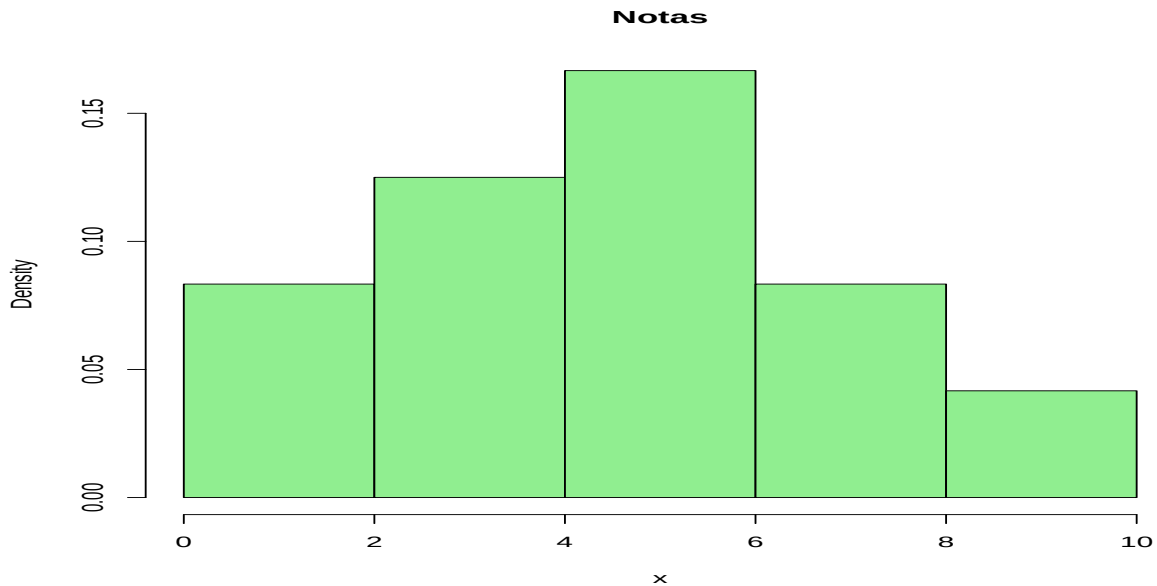
- La cuasivarianza muestral.
- La cuasidesviación típica muestral.
- El error típico de la media muestral.

¿En qué se diferencian los cuatro primeros de los tres últimos?

Ejemplo con una población pequeña

- Población: Los 12 alumnos de una clase.
- Variable: Nota que un alumno obtiene en un examen

Estudiante	1	2	3	4	5	6	7	8	9	10	11	12
Nota	1	0	3	10	8	7	5	5	5	6	4	3



Parámetros poblacionales

- Media poblacional:

$$\mu = \frac{1+0+3+10+8+7+5+5+5+6+4+3}{12} = 4.75$$

- Varianza poblacional:

$$\sigma^2 = \frac{(1-4.75)^2 + (0-4.75)^2 + \cdots + (3-4.75)^2}{12} = 7.3542$$

- Desviación típica poblacional:

$$\sigma = \sqrt{7.3542} = 2.7119$$

Una muestra de tamaño $n = 4$

- Una posible muestra de tamaño 4 es:

Estudiante	1	2	3	4	5	6	7	8	9	10	11	12
Nota	1	0	3	10	8	7	5	5	5	6	4	3

$$x_1 = 4, \quad x_2 = 3, \quad x_3 = 5, \quad x_4 = 6$$

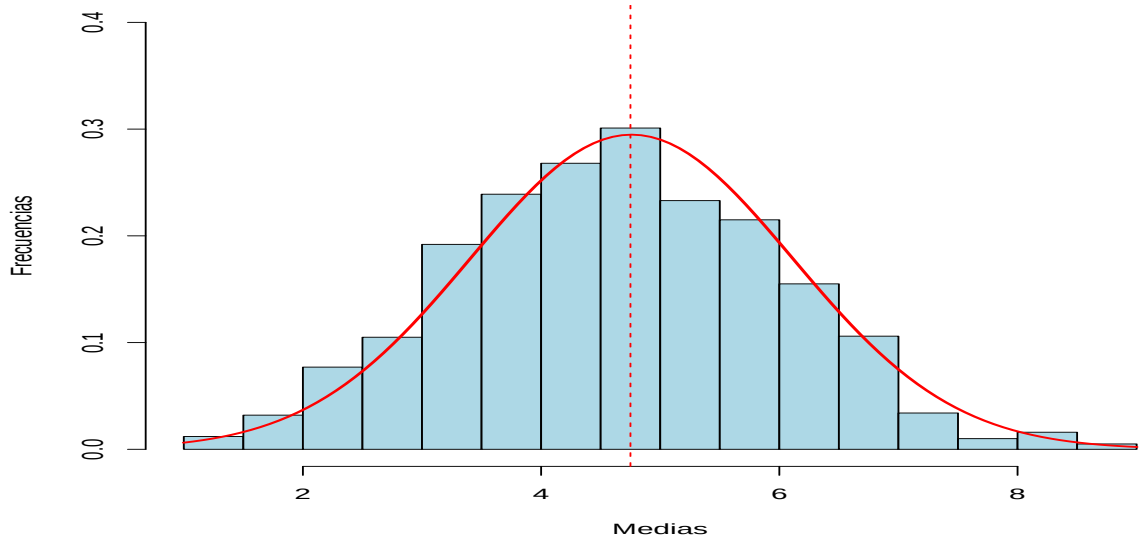
- A partir de estos datos, un estimador de μ (que sería útil si no conociéramos μ) es:

$$\hat{\mu} = \bar{x} = \frac{4 + 3 + 5 + 6}{4} = 4.5$$

- ¿Cómo se evalúa la precisión de $\bar{x}$, sin conocer μ?

2000 muestras de tamaño 4

- Extraemos 2000 muestras de tamaño 4.
- Todos los valores son equiprobables y se extraen con reemplazamiento (muestreo aleatorio simple).
- Un histograma de las correspondientes 2000 medias muestrales:



Características de la distribución de $\bar{x}$

- Las propiedades de $\bar{x}$ como estimador de μ se corresponden con las propiedades del histograma anterior.
- La forma del histograma es la de una distribución normal.
- Los valores de $\bar{x}$ se centran alrededor del verdadero valor de μ. El estimador es **centrado o insesgado**.
- La desviación típica de $\bar{x}$ es menor que σ. Se puede demostrar que la desviación típica de $\bar{x}$ es:

$$\frac{\sigma}{\sqrt{n}} = \frac{2.7119}{2} \approx 1.356.$$

Conclusiones de las observaciones anteriores

- Como $\bar{x}$ es insesgado, no hay tendencia sistemática a infraestimar o sobreestimar el valor de μ.

- Como $\bar{x} \cong \mathrm{N}(\mu, \sigma/\sqrt{n})$, con probabilidad aproximada 0.95 el error cometido al estimar μ mediante $\bar{x}$ es menor o igual que $2 \times \sigma/\sqrt{n} \approx 2.7119$

- Es decir, que podemos tener bastante confianza en que el valor de μ se encuentra en el intervalo:

$$[4.5 \mp 2.7119]$$

- Como en la práctica σ^2 es desconocida se usa S^2 en su lugar:

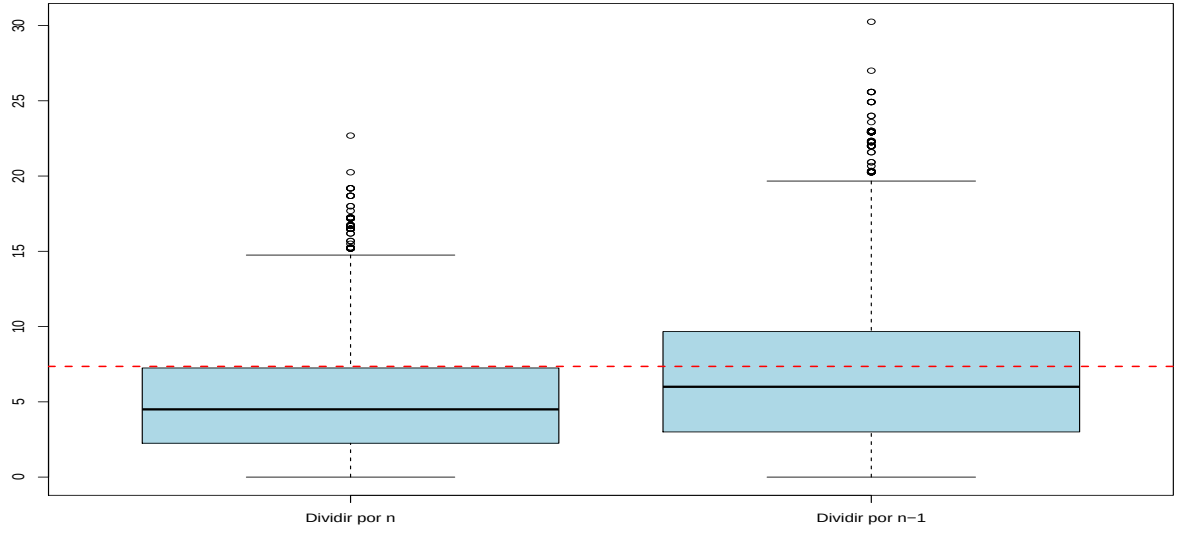
$$S^2 = \frac{(4-4.5)^2 + (3-4.5)^2 + (5-4.5)^2 + (6-4.5)^2}{3} = 1.666.$$

¿Por qué se divide por $n-1$ en lugar de n?

- Puede comprobarse que la varianza muestral (dividiendo por n) presenta una tendencia sistemática a infraestimar σ^2.

- Para corregir este sesgo se incrementa ligeramente el valor del estimador dividiendo por $n-1$ en lugar de n.

- Diagramas de cajas de las 2000 varianzas y cuasivarianzas muestrales. La línea roja corresponde a $\sigma^2 = 7.3542$.



Estimación de una proporción poblacional

Queremos estimar la proporción p de personas en una población que han seguido una dieta en los últimos 5 años. Para ello, preguntamos a 100 personas y definimos

$$x_i = \begin{cases} 0, & \text{si la persona } i \text{ no ha seguido una dieta;} \\ 1, & \text{si la persona } i \text{ ha seguido una dieta.} \end{cases}$$

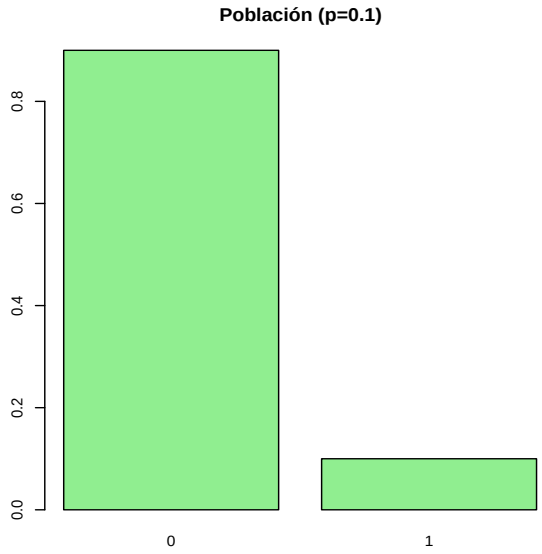
Obtenemos los siguientes datos:

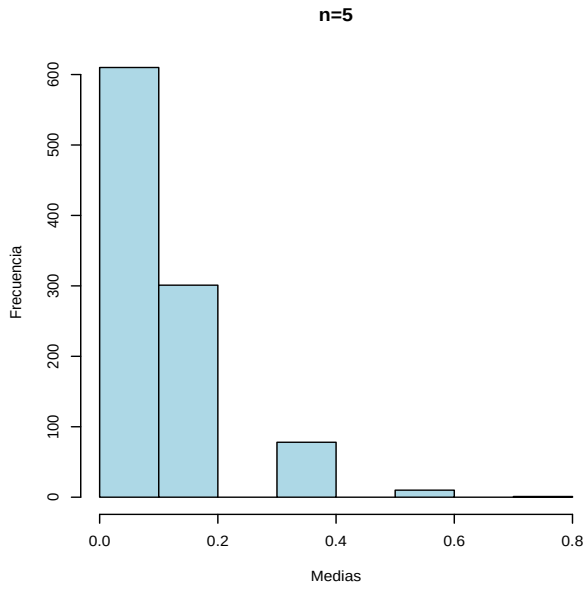
$$1,\ 0,\ 0,\ 1,\ 1,\ 0,\ 0,\ 0,\ 1,\ 0$$

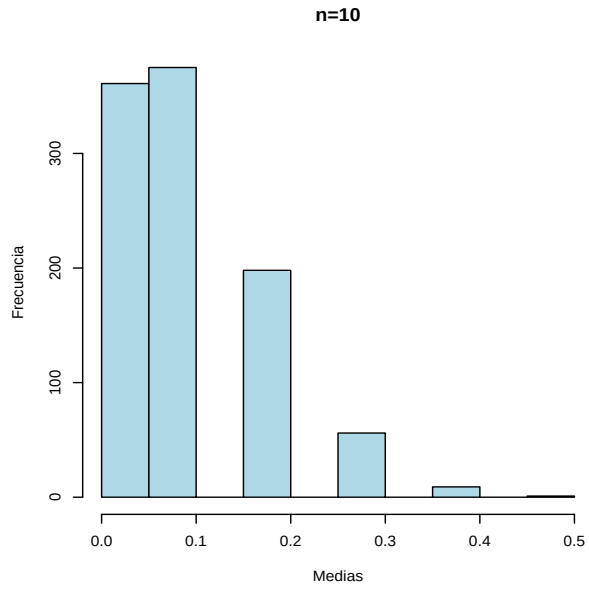
Estos datos son 10 observaciones de una v.a. de Bernoulli con parámetro p.

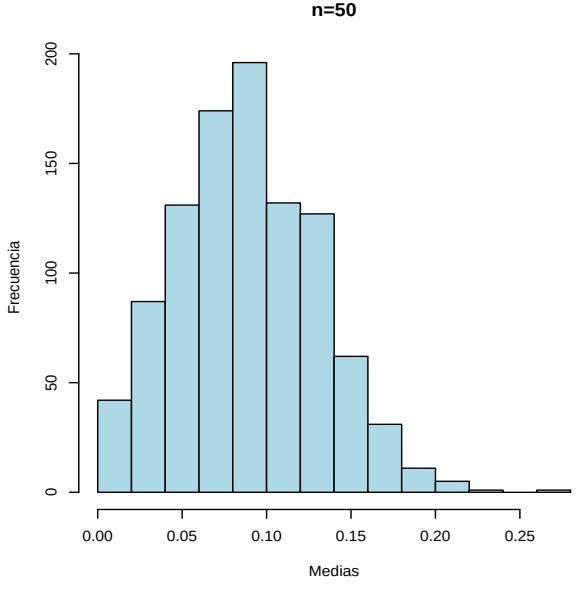
¿Cuál es el estimador más natural de p?

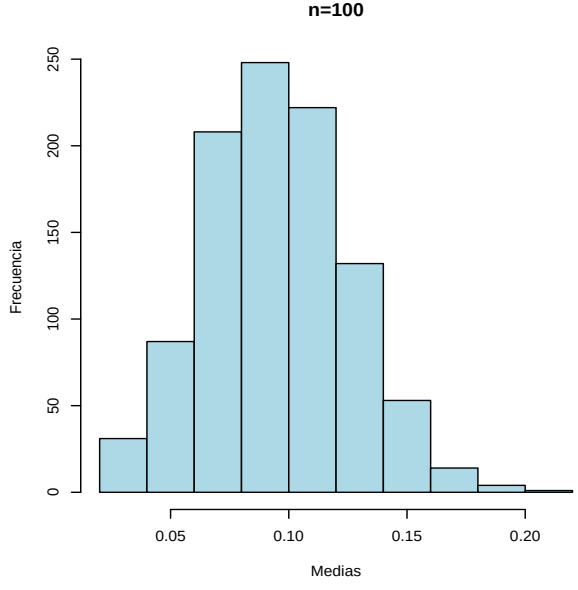
Distribución de la proporción muestral

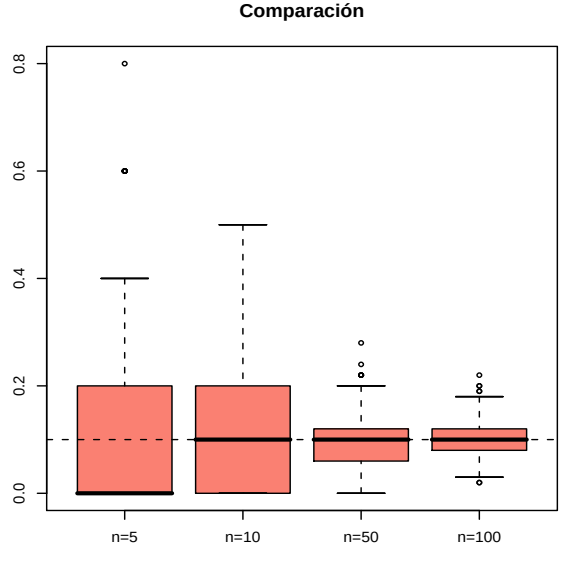












Distribución de la proporción muestral

Según el TCL, ¿cómo se distribuye aproximadamente la proporción muestral $\hat{p}$?

¿Cuál es la desviación típica de $\hat{p}$?

¿Cuál es el error típico de $\hat{p}$?

¿Cuál es el máximo (mínimo) valor posible de este error típico?

¿En qué situación se va a dar ese valor?

Calcula el error típico de $\hat{p}$ para los datos de la encuesta sobre la dieta.

Intervalos de confianza

Un **intervalo de confianza (IC)** para un parámetro es un intervalo, calculado a partir de la muestra, que contiene al parámetro con un alto grado de seguridad.

La **fórmula general** de los intervalos que vamos a estudiar es:

$$[\text{Estimador} \mp \text{Margen de error}]$$

El **centro** del intervalo es el estimador del parámetro en el que estamos interesados.

El **margen de error** depende

- de la precisión del estimador utilizado,
- del grado de seguridad con el que queremos que el intervalo contenga al parámetro (el nivel de confianza).

IC para la media de una población normal (varianza conocida)

Queremos estimar el contenido medio en grasas (en g/100 g) de la carne de cerdo, μ. Para ello disponemos de una muestra de 12 piezas de carne para la que el contenido medio es $\bar{x} = 24.93$.

Esto significa que $\mu \approx 24.93$. Por supuesto, $\mu \neq 24.93$. Si tomáramos otras 12 piezas distintas nos habría resultado una estimación de μ diferente.

Un IC es una forma de precisar qué significa $\mu \approx 24.93$.

Suponemos que la población es normal y que la desviación típica de la población es conocida y vale $\sigma = 0.25$.

Como $\bar{x} \equiv \mathrm{N}(\mu, 0.25/\sqrt{12})$, sabemos qué valores podríamos esperar si tomáramos muchas muestras de tamaño 12.

Aproximadamente para el 95% de las muestras de tamaño 12 se cumple:

$$-0.072 \times 1.96 < \bar{x} - \mu < 0.072 \times 1.96.$$

Las desigualdades anteriores son equivalentes a:

$$\bar{x} - 0.072 \times 1.96 < \mu < \bar{x} + 0.072 \times 1.96.$$

Aproximadamente para el 95% de las muestras de tamaño 12 se cumple que $\mu \in [\bar{x} \mp 0.1411]$.

Confiamos (con un nivel del 95%) en que la única muestra de la que disponemos sea una de las que verifican la condición.

Decimos que $[24.93 \mp 0.1411]$ es un IC para μ de nivel 95%.

Cuestiones:

- Con los mismos datos del ejemplo anterior calcula dos intervalos cuyos niveles de confianza sean 90% y 99%.
- Se ha obtenido $\bar{x} = 24.93$ pero la muestra era de 36 piezas en lugar de 12. Calcula un intervalo de nivel 95%.
- Se ha obtenido $\bar{x} = 24.93$ con una muestra de 36 piezas pero $\sigma = 1$ en lugar de $\sigma = 0.25$. Calcula un intervalo de nivel 95%.

Fórmula general: Un IC con nivel de confianza $1 - \alpha$ para la media de una población normal con σ conocida viene dado por:

$$\left[\bar{x} \mp z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

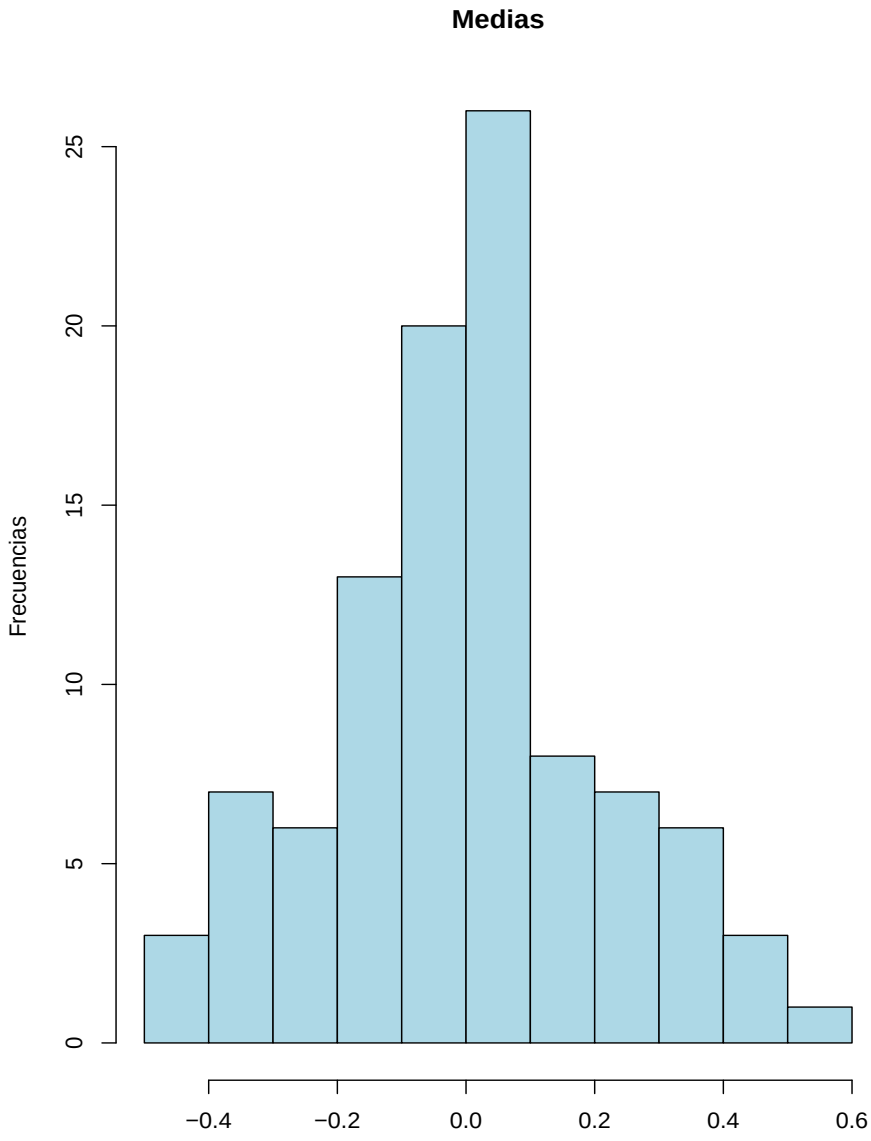
Interpretación del nivel de confianza

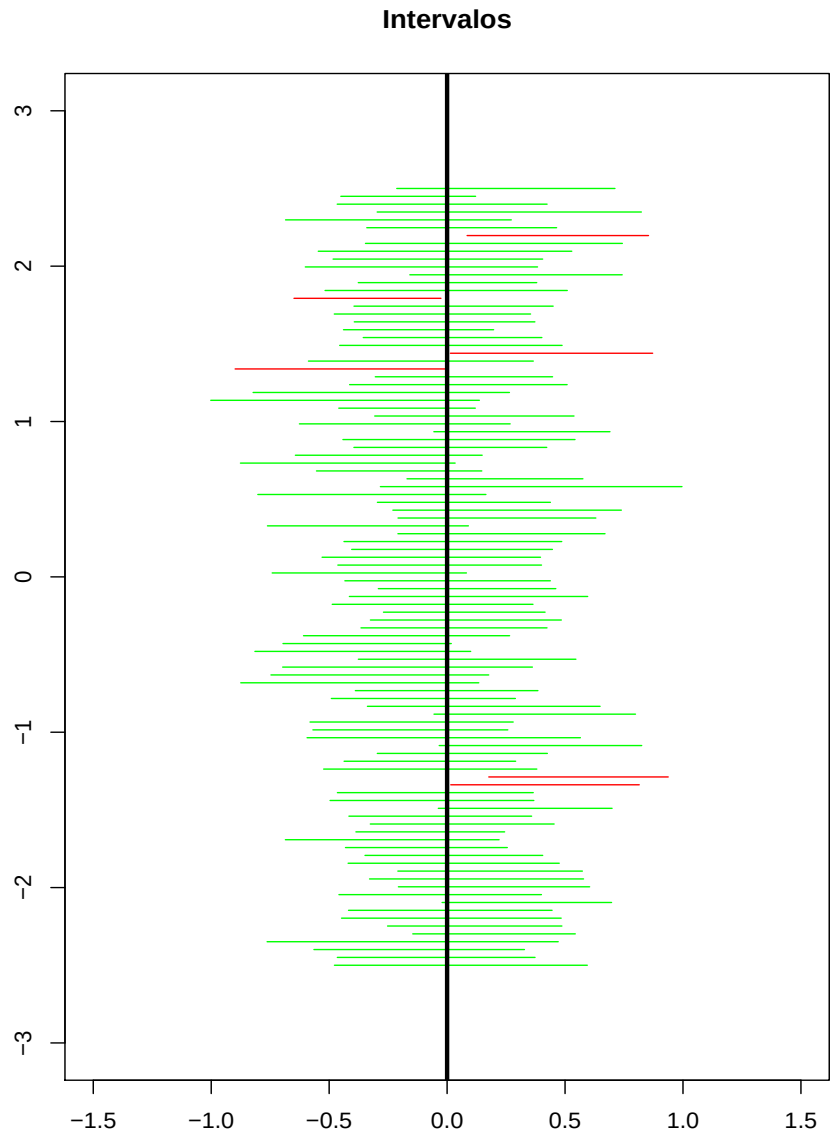
- Población: normal con media $\mu = 0$ y $\sigma = 1$.
- Se extraen 100 muestras de tamaño $n = 20$.
- Para cada muestra se calcula $\bar{x}$ y el intervalo de confianza para μ de nivel 95% (suponemos varianza poblacional conocida):

$$[\bar{x} \mp z_{0.025}\sigma/\sqrt{n}].$$

- Se representa un histograma de las 100 medias obtenidas, así como los 100 intervalos (en verde si contienen el valor 0 y en rojo si no).

Interpretación del nivel de confianza





Si σ no es conocida y la población no es normal

Como no conocemos σ, sustituimos en la fórmula σ por su estimador s calculado a partir de la muestra.

Debido al TCL, cuando el tamaño muestral n es suficientemente grande la fórmula sigue dando un intervalo de confianza aproximadamente válido:

$$\left[\bar{x} \mp z_{\alpha/2}\frac{s}{\sqrt{n}}\right].$$

El nivel de confianza ya no es exactamente $1-\alpha$. Este nivel es aproximado.

Margen de error

Al radio del intervalo se le suele llamar **margen de error**, E. En la situación anterior:

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

El margen de error depende de:

- El **nivel de confianza** deseado, a través de $z_{\alpha/2}$. Se suele tomar $\alpha = 0.05$ lo que da $z_{0.025} = 1.96 \approx 2$.
- La **heterogeneidad de la población**, medida a través de s.
- El **tamaño muestral** n.

Si σ no es conocida y la población es normal

- Cuando la población es normal y σ no es conocida, es posible dar un IC exacto incluso cuando el tamaño muestral es pequeño.

- Para ello, basta mirar en unas tablas distintas. En lugar de buscar $z_{\alpha/2}$ en las tablas de la normal, buscamos $t_{n-1,\alpha/2}$ en las tablas de la distribución t de Student. La fórmula del IC queda

$$\left[\bar{x} \mp t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}\right].$$

Distribución t de Student

- La distribución t de Student con $n-1$ grados de libertad (t_{n-1}) es la distribución de

$$\frac{\bar{x}-\mu}{s/\sqrt{n}}$$

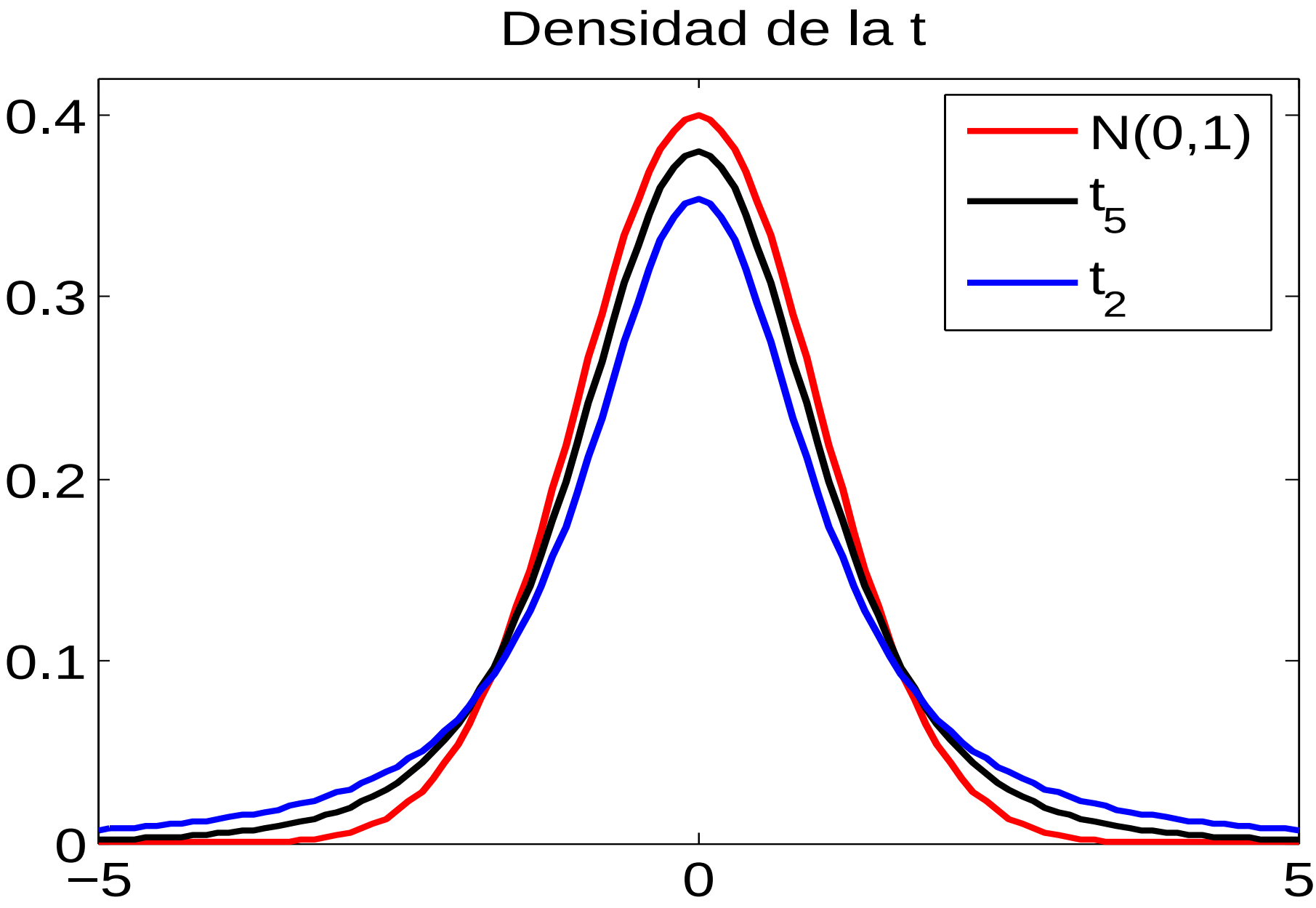
en una población normal.

- La forma de la densidad de t_n es similar a la de la normal. Es simétrica alrededor de cero.

- Sin embargo, la distribución t_n da más probabilidad a valores lejanos al centro.

- Si n es grande $t_n \cong \mathrm{N}(0,1)$.

Función de densidad de la distribución t-Student



Tablas de la distribución t-Student

r \ α	0,25	0,2	0,15	0,1	0,05	0,025	0,01	0,005	0,0005
1	1,000	1,376	1,963	3,078	6,314	12,706	31,821	63,656	636,578
2	0,816	1,061	1,386	1,886	2,920	4,303	6,965	9,925	31,600
3	0,765	0,978	1,250	1,638	2,353	3,182	4,541	5,841	12,924
4	0,741	0,941	1,190	1,533	2,132	2,776	3,747	4,604	8,610
5	0,727	0,920	1,156	1,476	2,015	2,571	3,365	4,032	6,869
6	0,718	0,906	1,134	1,440	1,943	2,447	3,143	3,707	5,959
7	0,711	0,896	1,119	1,415	1,895	2,365	2,998	3,499	5,408
8	0,706	0,889	1,108	1,397	1,860	2,306	2,896	3,355	5,041
9	0,703	0,883	1,100	1,383	1,833	2,262	2,821	3,250	4,781
10	0,700	0,879	1,093	1,372	1,812	2,228	2,764	3,169	4,587
11	0,697	0,876	1,088	1,363	1,796	2,201	2,718	3,106	4,437
12	0,695	0,873	1,083	1,356	1,782	2,179	2,681	3,055	4,318
13	0,694	0,870	1,079	1,350	1,771	2,160	2,650	3,012	4,221
14	0,692	0,868	1,076	1,345	1,761	2,145	2,624	2,977	4,140
15	0,691	0,866	1,074	1,341	1,753	2,131	2,602	2,947	4,073
16	0,690	0,865	1,071	1,337	1,746	2,120	2,583	2,921	4,015
17	0,689	0,863	1,069	1,333	1,740	2,110	2,567	2,898	3,965
18	0,688	0,862	1,067	1,330	1,734	2,101	2,552	2,878	3,922
19	0,688	0,861	1,066	1,328	1,729	2,093	2,539	2,861	3,883
20	0,687	0,860	1,064	1,325	1,725	2,086	2,528	2,845	3,850
21	0,686	0,859	1,063	1,323	1,721	2,080	2,518	2,831	3,819
22	0,686	0,858	1,061	1,321	1,717	2,074	2,508	2,819	3,792
23	0,685	0,858	1,060	1,319	1,714	2,069	2,500	2,807	3,768
24	0,685	0,857	1,059	1,318	1,711	2,064	2,492	2,797	3,745
25	0,684	0,856	1,058	1,316	1,708	2,060	2,485	2,787	3,725
26	0,684	0,856	1,058	1,315	1,706	2,056	2,479	2,779	3,707
27	0,684	0,855	1,057	1,314	1,703	2,052	2,473	2,771	3,689
28	0,683	0,855	1,056	1,313	1,701	2,048	2,467	2,763	3,674
29	0,683	0,854	1,055	1,311	1,699	2,045	2,462	2,756	3,660
30	0,683	0,854	1,055	1,310	1,697	2,042	2,457	2,750	3,646
40	0,681	0,851	1,050	1,303	1,684	2,021	2,423	2,704	3,551
60	0,679	0,848	1,045	1,296	1,671	2,000	2,390	2,660	3,460
120	0,677	0,845	1,041	1,289	1,658	1,980	2,358	2,617	3,373
∞	0,674	0,842	1,036	1,282	1,645	1,960	2,326	2,576	3,290

¿En qué tablas hay que mirar?

Para calcular intervalos para la media:

	Pob. normal	**Pob. no normal**
n **grande**	t Student	normal
n **pequeño**	t Student	?

Observaciones:

- El tamaño necesario de n depende de cuánto se parezca la población a la normal. Un tamaño $n > 40$ suele ser suficiente.
- Para valores grandes de n la dist. normal y la t son muy parecidas. El resultado será similar.
- El signo de interrogación significa que hay que usar métodos distintos en función de la población. No trataremos estos casos en este curso.

Un ejemplo resuelto

El envenenamiento por DDT causa temblores y convulsiones. En un estudio se ha administrado una dosis de DDT a 4 ratones y se ha medido posteriormente en cada uno el *periodo absolutamente refractario*, es decir, el tiempo que tardan sus nervios en recuperarse tras un estímulo. Las 4 medidas en milisegundos son:

1.7 1.6 1.8 1.9

(a) Estima el *periodo absolutamente refractario* medio μ para toda la población de ratones de la misma cepa sujeta al mismo tratamiento con DDT.

(b) Calcula el error típico de la estimación anterior.

(c) Calcula un intervalo de confianza para μ con nivel de confianza 90%. (Se supone normalidad).

(d) Calcula otro intervalo, pero ahora con un nivel del 95%

(a) La estimación de μ es la media muestral:

$$\bar{x} = \frac{1.7 + 1.6 + 1.8 + 1.9}{4} = 1.75.$$

(b) Para calcular el error típico, primero hay que calcular la varianza muestral:

$$s^2 = \frac{(1.7 - 1.75)^2 + (1.6 - 1.75)^2 + (1.8 - 1.75)^2 + (1.9 - 1.75)^2}{3}$$

Por lo tanto $s^2 \approx 0.017$ y $s = \sqrt{0.017} \approx 0.13$.

El error típico es $s/\sqrt{n} = 0.13/2 = 0.065$.

(c) Como $t_{3,0.05} = 2.353$, un I.C. con nivel de confianza $1 - \alpha = 0.90$ es

$$[1.75 \mp 2.353 \times 0.065] = [1.597\,,\, 1.903].$$

Podemos afirmar que $1.597 < \mu < 1.903$ con un nivel de confianza del 90%.

(d) Como $t_{3,0.025} = 3.182$, un I.C. con nivel de confianza $1 - \alpha = 0.95$ es

$$[1.75 \mp 3.182 \times 0.065] = [1.543\,,\, 1.957].$$

Podemos afirmar que $1.543 < \mu < 1.957$ con un nivel de confianza del 95%.

Cuestiones

- En un informe leemos que un intervalo de confianza para la puntuación media de los estudiantes en un test de inglés es (267.8, 276.2).
  (a) Verdadero o falso: El 95% de los estudiantes han tenido puntuaciones entre 267.8 y 276.2
  (b) ¿Cuál fue la puntuación media de los estudiantes de la muestra utilizada para calcular el intervalo?

- Mirando en las tablas de la distribución **t-Student**, determina un valor c tal que la probabilidad de que una distribución **normal estándar** sea mayor que c es 0.2

IC para una proporción

Las ideas para construir un IC en este caso son exactamente las mismas.

Sabemos que para la distribución de Bernoulli $\sigma = \sqrt{p(1-p)}$ que se puede estimar mediante $\hat{\sigma} = \sqrt{\hat{p}(1-\hat{p})}$.

La fórmula del intervalo queda:

$$\left[\hat{p} \mp z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$$

y es válida para n grande, ya que se basa en el TCL.

El margen de error en este caso es

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

Un ejemplo resuelto

En una encuesta para estudiar la preocupación de la población por su alimentación, se ha preguntado a 965 personas si han seguido alguna dieta en los últimos 5 años. De ellas, 406 han respondido afirmativamente. Con esta información:

(a) Estima la proporción p de la población que ha seguido alguna dieta en los últimos 5 años.

(b) Calcula el error típico del estimador anterior.

(c) Calcula un intervalo de confianza para p con un nivel de confianza del 95%

(d) Si para un nuevo estudio se desea estimar p con un margen de error de $\mp 1\%$ y un nivel de confianza del 95%, ¿a cuántas personas hay que entrevistar aproximadamente?

(a) El estimador de p a partir de los datos disponibles es la proporción muestral $\hat{p} = 406/965 = 0.421$.

(b) El error típico de este estimador es

$$\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.421 \times (1-0.421)}{965}} = 0.0159$$

(c) Como $z_{0.025} = 1.96$, un I.C. con nivel de confianza $1 - \alpha = 0.95$ es

$$[0.421 \mp 1.96 \times 0.0159] = [0.39\,,\, 0.45].$$

Podemos afirmar que $0.39 < p < 0.45$ con un nivel de confianza del 95%.

(d) Para calcular n despejamos en la ecuación:

$$1.96 \times \sqrt{\frac{0.421 \times (1 - 0.421)}{n}} = 0.01$$

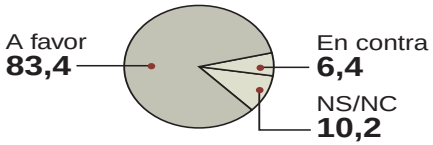
De aquí obtenemos:

$$n = \frac{0.421 \times (1 - 0.421) \times 1.96^2}{0.01^2} = 9364.246 \approx 9365.$$

Ficha técnica de una encuesta

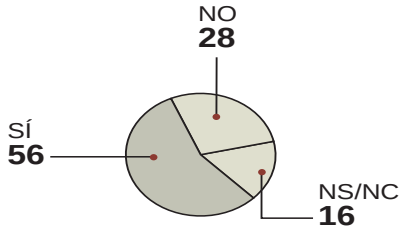
¿Está a favor o en contra de que se modifique la Constitución para abolir la preferencia del hombre sobre la mujer en la sucesión al trono?

En %



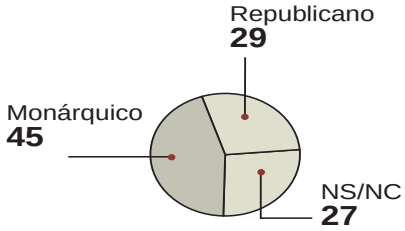
¿Es partidario de que esa reforma se realice cuanto antes?

En %



¿Usted se considera monárquico o republicano?

En %



FICHA TÉCNICA

Realización del trabajo de campo: la encuesta ha sido realizada por el Instituto Opina los días 7 y 8 de noviembre de 2005. **Ámbito geográfico:** España. **Recogida de información:** mediante entrevista telefónica asistida por ordenador (CATI). **Universo de análisis:** población mayor de 18 años residente en hogares con teléfono. **Tamaño de la muestra:** 1.000 entrevistas proporcionales. **Error muestral:** el margen de error para el total de la muestra es de ± 3,10% para un margen de confianza del 95% y bajo el supuesto de máxima indeterminación (p=q=50%). **Procedimiento de muestreo:** selección polietápica del entrevistado: unidades primarias de muestreo (municipios) seleccionadas de forma aleatoria proporcional para cada provincia. Unidades secundarias (hogares) mediante la selección aleatoria de números de teléfono. Unidades últimas (individuos) según cuotas cruzadas de sexo, edad y recuerdo de voto en las elecciones generales de 2004.

Explicación

- El margen de error del intervalo de confianza de una proporción verifica

$$\epsilon = z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq z_{\alpha/2}\sqrt{\frac{1}{4n}}$$

ya que el caso $p = q = 1/2$ ($q = 1 - p$) es el más desfavorable.

- Según la ficha técnica, $n = 1000$ y $1 - \alpha = 0.95$ ($z_{0.025} = 1.96$), por lo que en el caso más desfavorable:

$$\epsilon = 1.96\sqrt{\frac{1}{4000}} \approx 0.031.$$

- El valor que da la fórmula es consistente con el margen de error de $\mp 3.10\%$ para los porcentajes estimados en el sondeo.