

Módulo 0: Nociones matemáticas básicas

0.4. Estadística Inferencial

Rafael Zambrano

rafazamb@gmail.com

Introducción

4. ESTADÍSTICA INFERENCIAL

4.1 Variables aleatorias

4.2 Distribuciones de probabilidad

4.3 Distribución normal

4.4 Intervalos de confianza

4.5 Error absoluto y tamaño de la muestra

4.6 Modelos de regresión lineal

4. Estadística Inferencial

4.1. Variables Aleatorias

Variables aleatorias (VA)

- Supongamos el experimento de elegir al azar un universitario de España. El espacio muestral está formado por todos los distintos universitarios: $E = \{\omega_i\}_{i=1}^N$
- Nuestro interés no está en el individuo ω_i sino en un valor asociado a él que denotamos por $X(\omega_i) = x$, por ejemplo, su edad
- A la función que asocia a un resultado un valor numérico se le llama **variable aleatoria**

$$\begin{aligned} X: E &\rightarrow \mathbb{R} \\ \omega &\rightarrow X(\omega) = x \end{aligned}$$

Variables aleatorias discretas

- Toma un número finito de valores que podemos numerar, por ejemplo, el número de habitantes por casa en una ciudad.
- Si denotamos por $D = \{x_1, x_2, \dots\}$ el conjunto de valores para la variable, las probabilidades $P(X = x_i)$ con $i = 1, \dots, N$ reciben el nombre de **función de probabilidad** de la variable aleatoria X
- Ejemplo:

x	0.000	1.000	2.000	3.000	4.000	5.000	6.000	7.000	8.000	9.000	10.000
$P(X = x)$	0.107	0.268	0.302	0.201	0.088	0.026	0.006	0.001	0.000	0.000	0.000

$$P(X < 3) = P(X = 0) + P(X = 1) + P(X = 2) = 0.107 + 0.268 + 0.302 = 0.677$$

Variables aleatorias continuas

- Pueden tomar un número infinito de valores, por ejemplo, la concentración de azúcar de un refresco
- De estos valores aleatorios, nos interesa conocer la probabilidad que tenemos de que el valor que observemos esté entre dos números:

$$P(a < X \leq b) = \int_a^b f(x) d(x)$$

- La función $f(x)$ recibe el nombre de **función densidad de probabilidad** de la variable aleatoria X , y nos informa sobre la cantidad de probabilidad que hay en un intervalo determinado. Cumple las siguientes propiedades:

$$1) f(x) \geq 0$$

$$2) \int f(x) = 1$$

Media de una variable aleatoria

- **VA discreta:** $E(X) = \mu_X = \sum_{i=1}^N x_i P(X = x_i)$

➤ Ejemplo: Se conoce la distribución de probabilidad del número de mascotas por casa en Leganés

x	1	2	3	4
$P(X = x)$	0.64	0.21	0.12	0.03

$$E(X) = 1 \times 0.64 + 2 \times 0.21 + 3 \times 0.12 + 4 \times 0.03 = 1.81$$

- **VA continua:** $E(X) = \mu_X = \int_{-\infty}^{\infty} x \cdot f(x) dx$

4. Estadística Inferencial

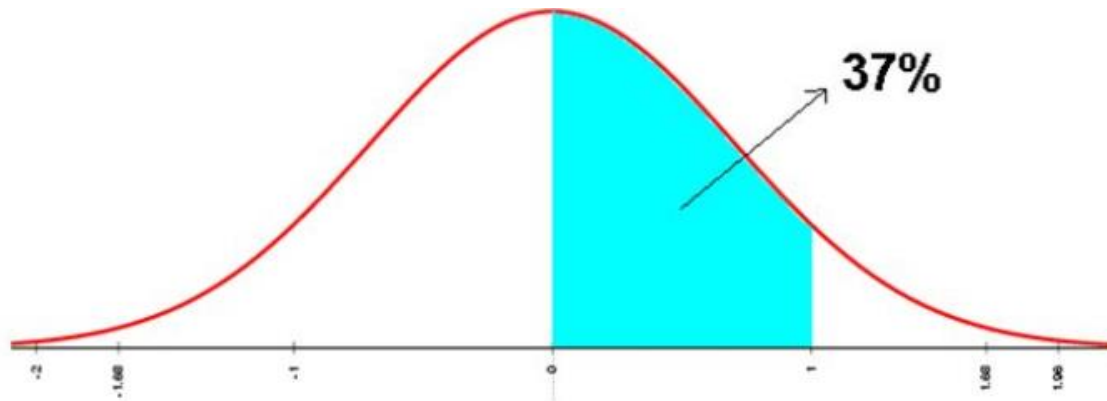
4.2. Distribuciones de probabilidad

Distribución de probabilidad

- La distribución de probabilidad describe las probabilidades de los posibles valores de una variable aleatoria
- Si la VA es discreta, le corresponderá una distribución discreta
- Si la VA es continua (puede tomar cualquier valor dentro de un intervalo), la distribución será continua.

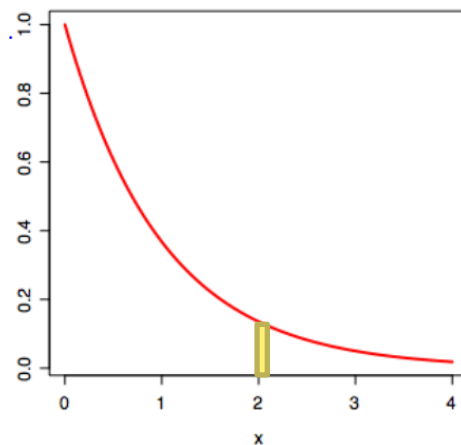
Función densidad de probabilidad (fdp)

- La **función de densidad de probabilidad**, función de densidad, o, simplemente, densidad de una **variable aleatoria continua** describe la probabilidad relativa según la cual dicha variable aleatoria tomará determinado valor



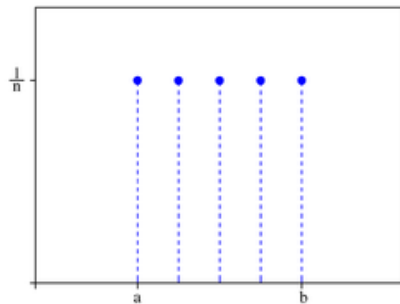
Función densidad de probabilidad (fdp)

- Ejemplo: Una especie de bacteria típicamente vive entre 0 y 4 horas. ¿Cuál es la probabilidad de que una bacteria viva *exactamente* 2 horas?
- La respuesta es 0%. Muchas bacterias vivirán *aproximadamente* 2 horas, pero es improbable que dada una bacteria ésta viva *exactamente* 2.000000 horas
- En lugar de eso, la pregunta debería ser: ¿Cuál es la probabilidad de que la bacteria muera entre 2 y 2.01 horas?



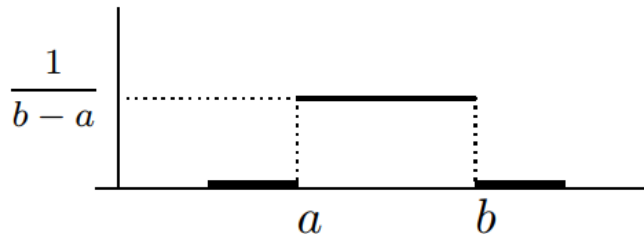
Distribuciones discretas más comunes

- **Distribución binomial:** describe el número de aciertos en experimentos con posibles resultados binarios con probabilidad de acierto p y probabilidad de fallo $q = 1 - p$.
 - Para representar que una VA sigue una distribución binomial: $X \sim B(n, p)$
 - Ejemplos: n° de caras al lanzar 20 veces una moneda $X \sim B(20, 0.5)$
- **Distribución uniforme:** Asume un número finito de valores con la misma probabilidad
 - La probabilidad de cada resultado x_i es $p(x_i) = \frac{1}{n}$
 - Ejemplo: En un dado, todos los resultados tienen la probabilidad $1/6$

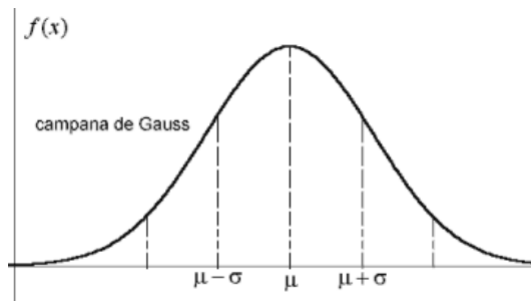


Distribuciones continuas más comunes

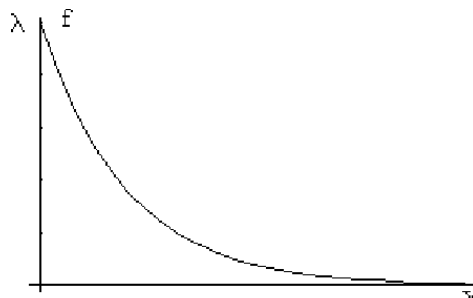
- Distribución uniforme



- Distribución normal

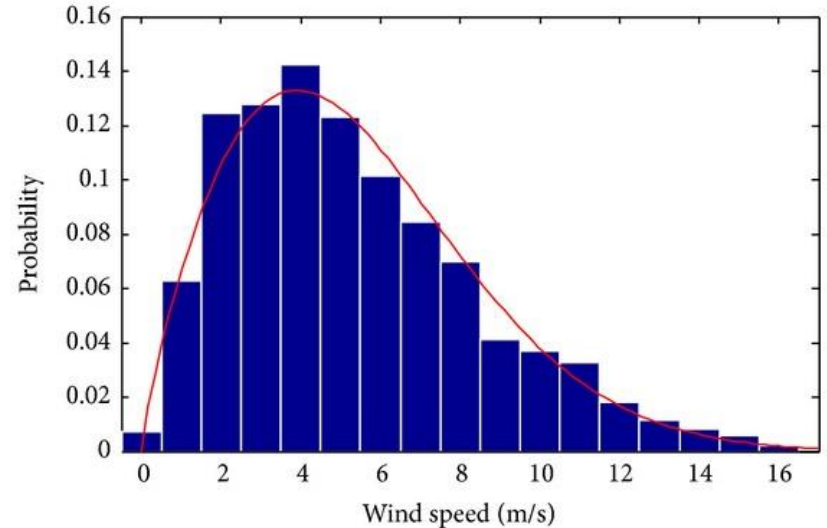
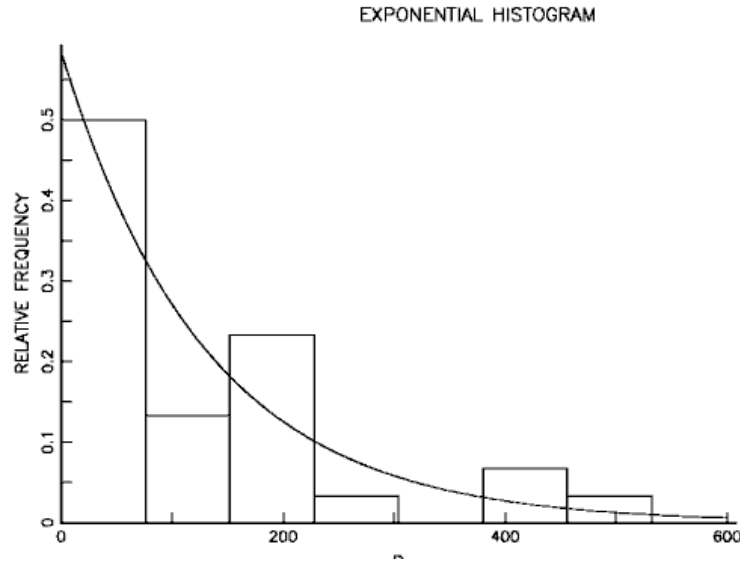


- Distribución exponencial



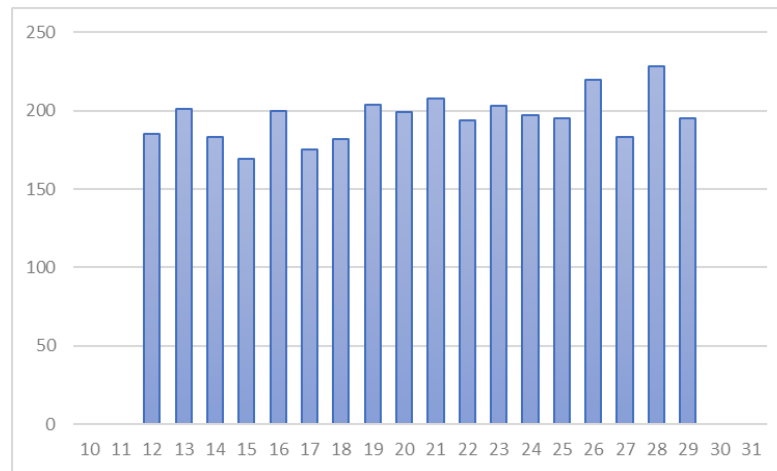
Histogramas y fdps

- Los histogramas dan una idea aproximada de la distribución de los datos, y a menudo se utilizan para estimar la función densidad de probabilidad



Distribuciones para la asignación de valores aleatorios

- A menudo, se realizan experimentos escogiendo números aleatorios, donde se especifica la distribución que deben seguir estos números
- Ejemplo: generar 100 números que sigan una distribución uniforme entre 12 y 29 (en Excel: `ALEATORIO.ENTRE(12;29)`)



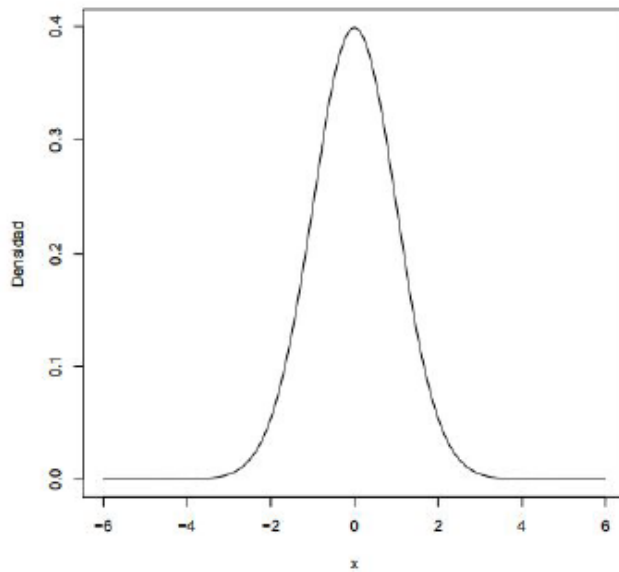
4. Estadística Inferencial

4.3. Distribución normal

Distribución normal

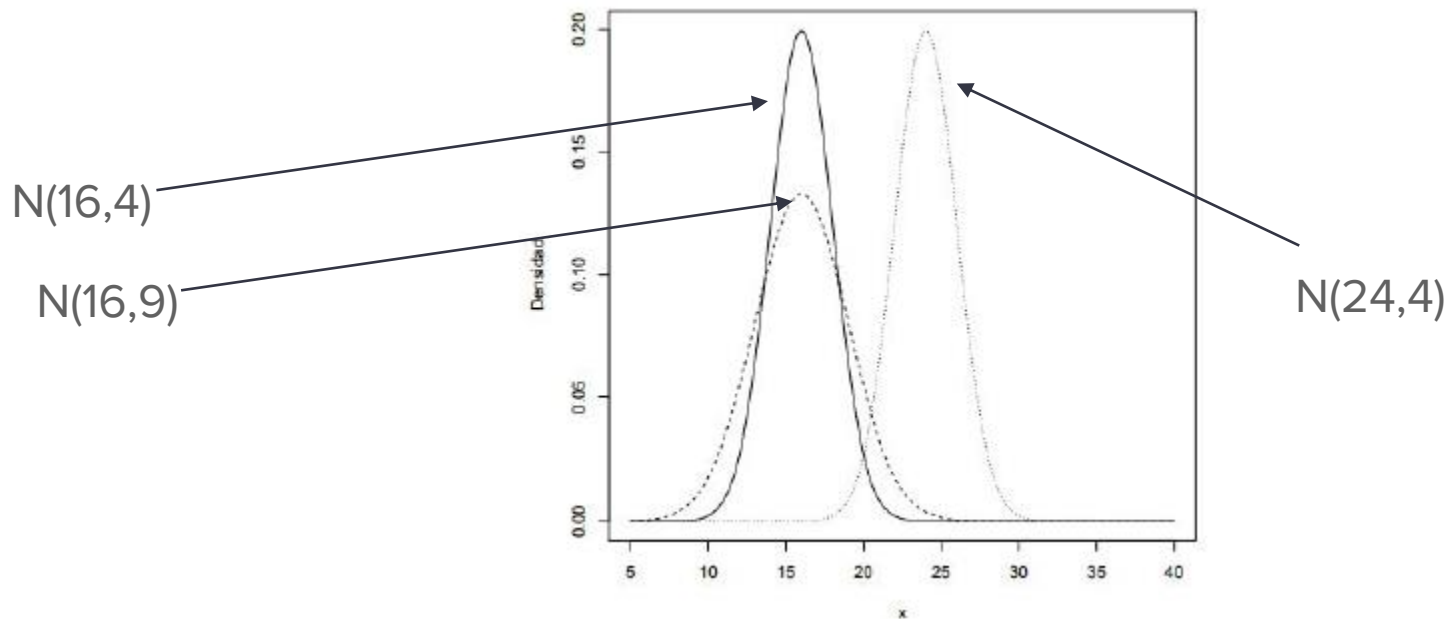
- Una variable aleatoria X se dice que sigue una distribución normal con media μ y varianza σ^2 (o, simplemente, que es una variable aleatoria normal) y se denota con $X \sim N(\mu, \sigma)$ si su función de densidad viene dada por

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2} \frac{(x-\mu)^2}{\sigma^2}}$$



Distribución normal

- Ejemplo: ¿A cuál de las siguientes curvas corresponden las distribuciones normales $N(16,4)$, $N(24,4)$ y $N(16,9)$?



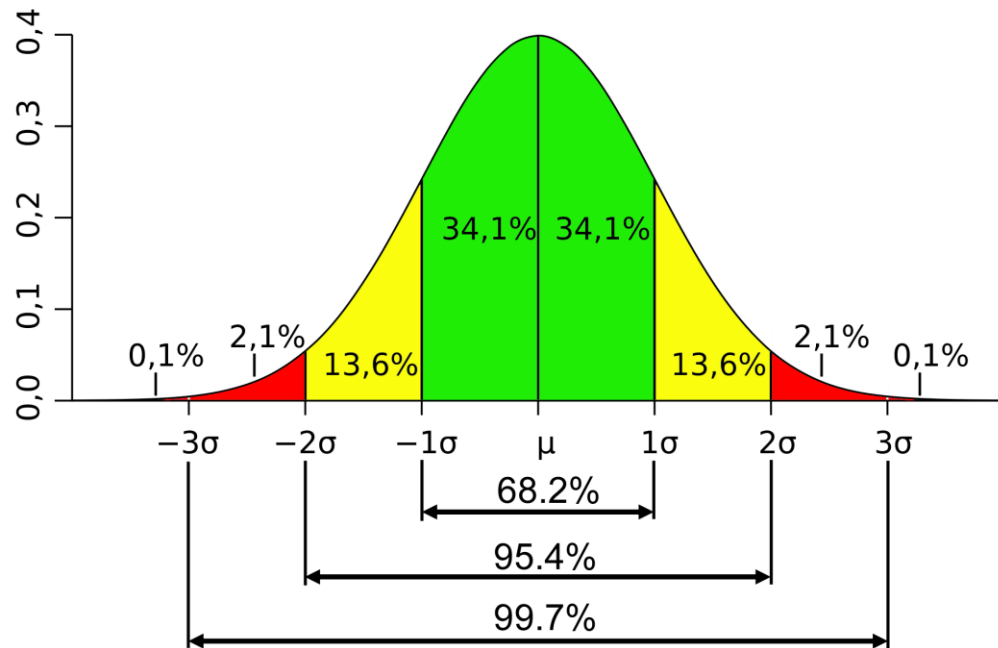
Distribución normal

Muchos fenómenos físicos se pueden modelar de manera adecuada a través de esta distribución.



Distribución normal

Se pueden conocer las proporciones de datos/probabilidades en función de la desviación estándar:



Distribución normal

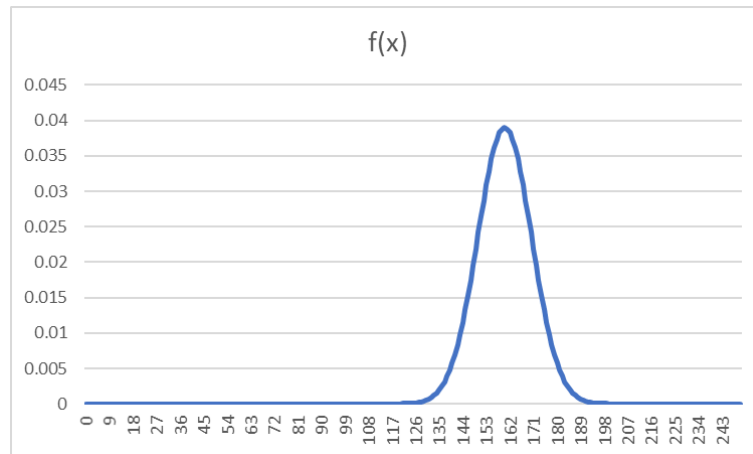
Ejemplo: Supongamos la VA de la altura de los universitarios españoles, la cual se distribuye de forma normal con media 160cm y desviación estándar de 10.23

Suponiendo que hay 1500 personas en el estudio, vamos a generar aleatoriamente estos valores

En Excel: `DISTR.NORM(x,mu,sigma,FALSO)`

¿Cuál es la probabilidad de que la altura
Esté entre 159 y 162cm?

Solución: 15.49%



Distribución normal

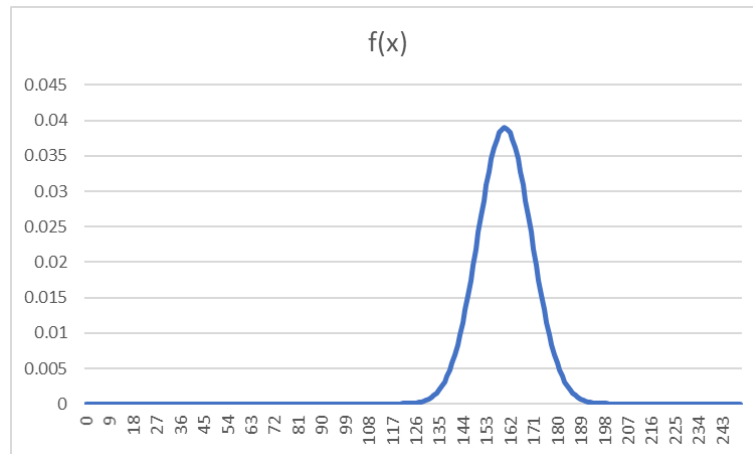
Ejemplo: Supongamos la VA de la altura de los universitarios españoles, la cual se distribuye de forma normal con media 160cm y desviación estándar de 10.23

Suponiendo que hay 1500 personas en el estudio, vamos a generar aleatoriamente estos valores

En Excel: `DISTR.NORM(x,mu,sigma,FALSO)`

¿Cuál es la probabilidad de que la altura
Esté entre 159 y 162cm?

Solución: 15.49%

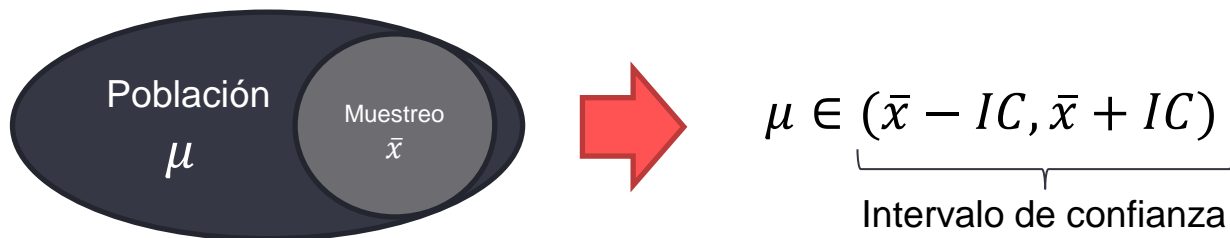


4. Estadística Inferencial

4.4. Intervalos de confianza

Intervalos de confianza

- El intervalo de confianza nos da una idea del “margen de error” al realizar un muestreo
- El intervalo de confianza nos da un rango en el que podemos estar seguros con cierta probabilidad (normalmente del 95%) de que la media **real** de la población estará en ese rango.



Intervalos de confianza

- Ejemplo: Disponemos de 100.000 clavos y queremos conocer la longitud media de cada uno de ellos. Para ello, se realiza un muestreo de 100 clavos y se calcula la media de ellos.



100.000 clavos con media
de longitud μ



100 clavos con media de
longitud \bar{x}

μ estará
comprendido entre
 $\bar{x} \pm IC$
con un 95% de
probabilidad

Error estándar de la muestra

- Es la desviación estándar de todas las posibles muestras escogidas en una población

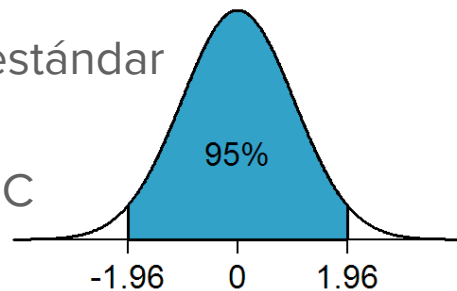
$$SE = \frac{s}{\sqrt{n}}$$

donde s es la desviación estándar de la muestra

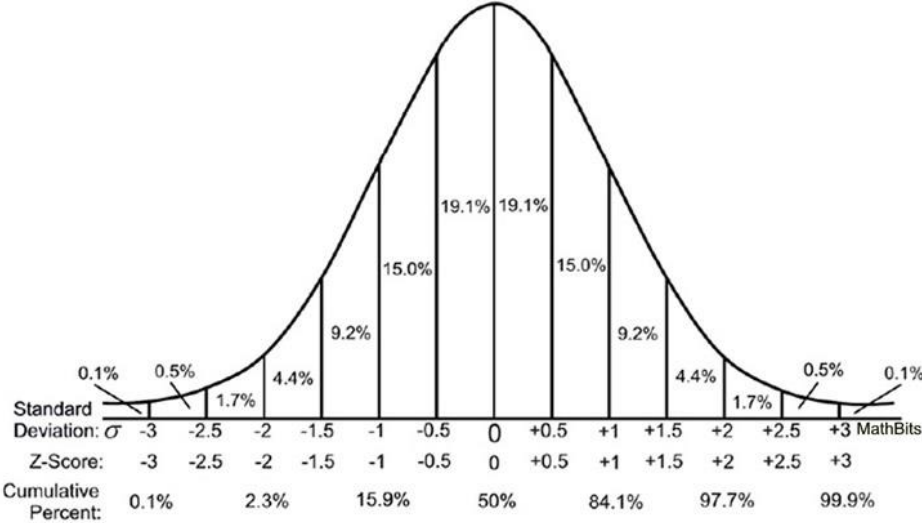
- El intervalo de confianza se calcula como $IC = \pm 1.96 \cdot SE$

Factor multiplicador

- El valor $z = 1.96$ proviene del 95% de la distribución normal estándar
- Cuanto más se quiera aumentar la confianza, mayor será el IC



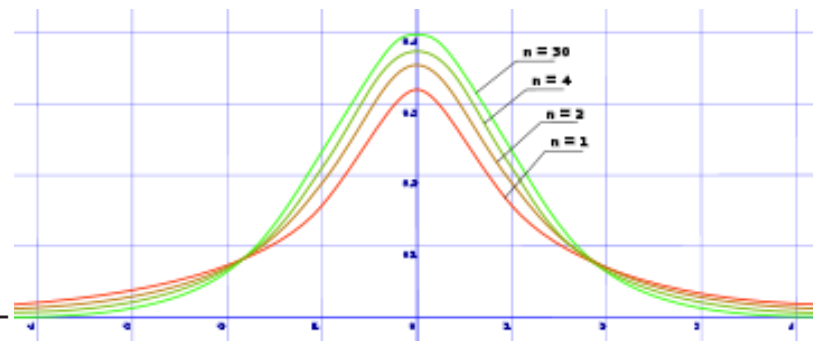
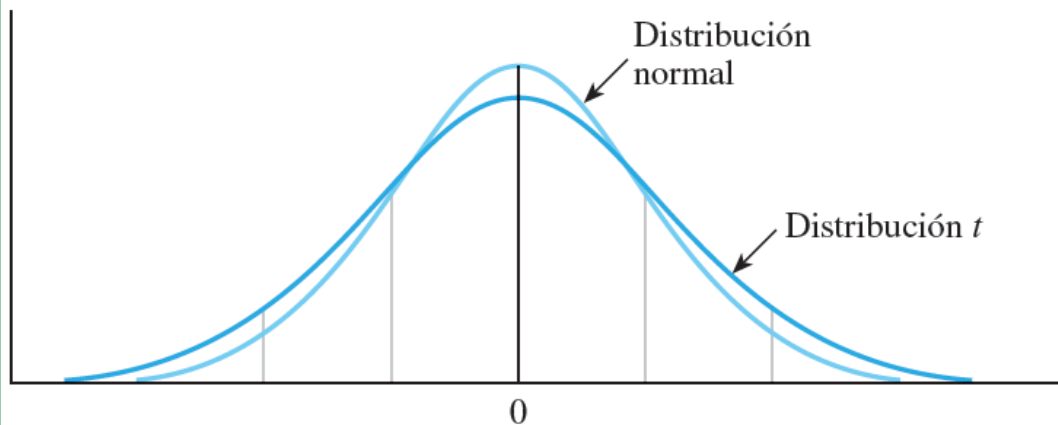
Error estándar de la muestra



Confidence Level	z*- value
80%	1.28
85%	1.44
90%	1.64
95%	1.96
98%	2.33
99%	2.58

Distribución de t-student

- La distribución t - student es una distribución de probabilidad que surge del problema de estimar la media de una población normalmente distribuida cuando el tamaño de la muestra es pequeña ($n < 30$)



- En función de los **grados de libertad** (n° de muestras -1) se tienen diferentes fdp
- Los valores suelen consultarse en tablas

Ejemplo 1

- Al salir de una película en el cine, se entrevistan a 11 personas para saber qué puntuación entre 0 y 10 le darían a la película que acaban de ver. Se quiere conocer la media muestral el intervalo de confianza. Las puntuaciones fueron:

2,6,5,5,6,3,7,4,1,1,7

1. Calculamos la media muestral $\bar{x} = \frac{\sum x_i}{n} = 4.27$
2. Calculamos la desviación estándar de la muestra $s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}} = 2.24$
3. Calculamos el error estándar de la media $SE = \frac{s}{\sqrt{n}} = \frac{2.24}{\sqrt{11}} = 0.675$
4. Busco en la tabla de t-student el factor multiplicador con $n = 11$ (10 grados de libertad) y $p = 0.95$, obteniendo $t = 1.8125$
5. Intervalo de confianza: $(\bar{x} - t \cdot SE, \bar{x} + t \cdot SE) = (4.27 - 0.675 \cdot 1.8125, 4.27 + 0.675 \cdot 1.8125)$
 $= (3.05, 5.49)$ (La nota media real de la película está entre esos valores con un 95% de confianza)

Ejemplo 2

- Se quiere conocer la altura de los estudiantes de cursos de análisis de datos con una confianza del 99%, teniendo las siguientes muestras:

180,165,176,165,169,179,168,176,191,178,173,157,175,179,169,185,168,
170,166,178,177,180,168,179,173,162,175,175,180,167

1. $\bar{x} = 173.43$

2. $s = 7.24$

3. $SE = \frac{s}{\sqrt{n}} = \frac{7,24}{\sqrt{30}} = 1.32$

4. $z \cdot SE = 2.58 \cdot 1.32 = 3.40$

5. *Intervalo de confianza:* $(\bar{x} - z \cdot SE, \bar{x} + z \cdot SE) = (170,176)$

Confidence Level	z*- value
80%	1.28
85%	1.44
90%	1.64
95%	1.96
98%	2.33
99%	2.58

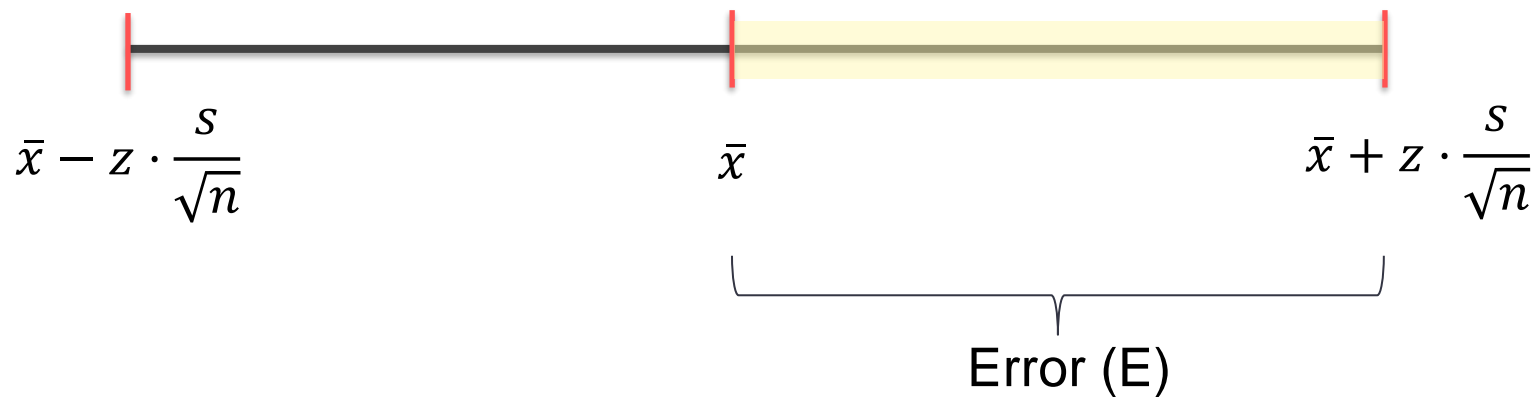
4. Estadística Inferencial

4.5. Error absoluto y tamaño de la muestra

Error absoluto y tamaño de la muestra

- ¿Cuántos datos hemos de tener para que nuestro estudio tenga validez? Es una pregunta muy genérica y sin respuesta
- ¿Cuántos datos necesitamos que al estimar una media poblacional el error máximo que cometemos sea menor que una cantidad que previamente especificamos?
- En Estadística nunca podemos afirmar con seguridad nada. Siempre hacemos afirmaciones basadas en la probabilidad
- El error absoluto es la mitad de la longitud del intervalo de confianza

Estimar una media



$$E = z \cdot \frac{s}{\sqrt{n}} \Rightarrow n = \left(\frac{z \cdot s}{E} \right)^2$$

“Para asumir un cierto error acerca de la media, con un grado de confianza determinado, necesito n muestras”

Estimar una media

- Ejemplo: deseamos conocer la media del nivel de azúcar en un refresco, con una seguridad del 95% y una precisión de ± 3 mg/dl y tenemos información bibliográfica de que la varianza es de 250 mg/dl

$$n = \left(\frac{z \cdot s}{E} \right)^2 = \left(\frac{1.96 \cdot \sqrt{250}}{3} \right)^2 = 106.7$$

- Necesitaría tomar al menos 107 muestras para mantener esa precisión

Estimar una proporción (población total desconocida)

$$n = \left(\frac{z}{E}\right)^2 \cdot p \cdot (1 - p)$$

- Si deseamos estimar una proporción, debemos tener una idea aproximada del parámetro que queremos medir (en este caso una proporción). En caso de no tener dicha información utilizaremos el valor $p=0.5$ (50%), que maximiza el tamaño muestral.
- Ejemplo: Sabiendo que un 5% de la población tiene diabetes, ¿a cuántas personas habría que examinar para conocer la proporción de diabetes con una precisión del 3% y una confianza del 95%?
 - $n = \left(\frac{1.96}{0.03}\right)^2 \cdot 0.05 \cdot 0.95 = 203$

Estimar una proporción (población total conocida)

$$n = \frac{N \cdot z^2 \cdot p \cdot (1 - p)}{E^2 \cdot (N - 1) + z^2 \cdot p \cdot (1 - p)}$$

donde N es el tamaño de la población

- Ejemplo: ¿A cuántas personas tendría que estudiar de una población de 15.000 habitantes para conocer la prevalencia de diabetes?

$$\text{➤ } n = \frac{15000 \cdot 1.96^2 \cdot 0.05 \cdot 0.95}{0.03^2 (15000 - 1) + 1.96^2 \cdot 0.05 \cdot 0.95} = 200$$

Estimar una proporción (población total conocida)

Tamaño de la población	Tamaño de la muestra por margen de error		
	$\pm 3\%$	$\pm 5\%$	$\pm 10\%$
500	345	220	80
1000	525	285	90
3000	810	350	100
5000	910	370	100
10 000	1000	385	100
100 000	1100	400	100

Calculadora online

- <https://www.netquest.com/es/gracias-calculadora-muestra>

5000

TAMAÑO DEL UNIVERSO

Número de personas que componen la población a estudiar.

50

HETEROGENEIDAD %

Es la diversidad del universo. Lo habitual suele ser 50%.

5

MARGEN DE ERROR

Menor margen de error requiere mayores muestras.

95

NIVEL DE CONFIANZA

Cuanto mayor sea el nivel de confianza, mayor tendrá que ser la muestra (95% - 99%).

El tamaño de muestra que necesitas es...

357

El resultado anterior debe interpretarse así:

Si encuestas a 357 personas, el 95% de las veces el dato que quieres medir estará en el intervalo $\pm 5\%$ respecto al dato que observes en la encuesta.

4. Estadística Inferencial

4.6. Modelos de regresión

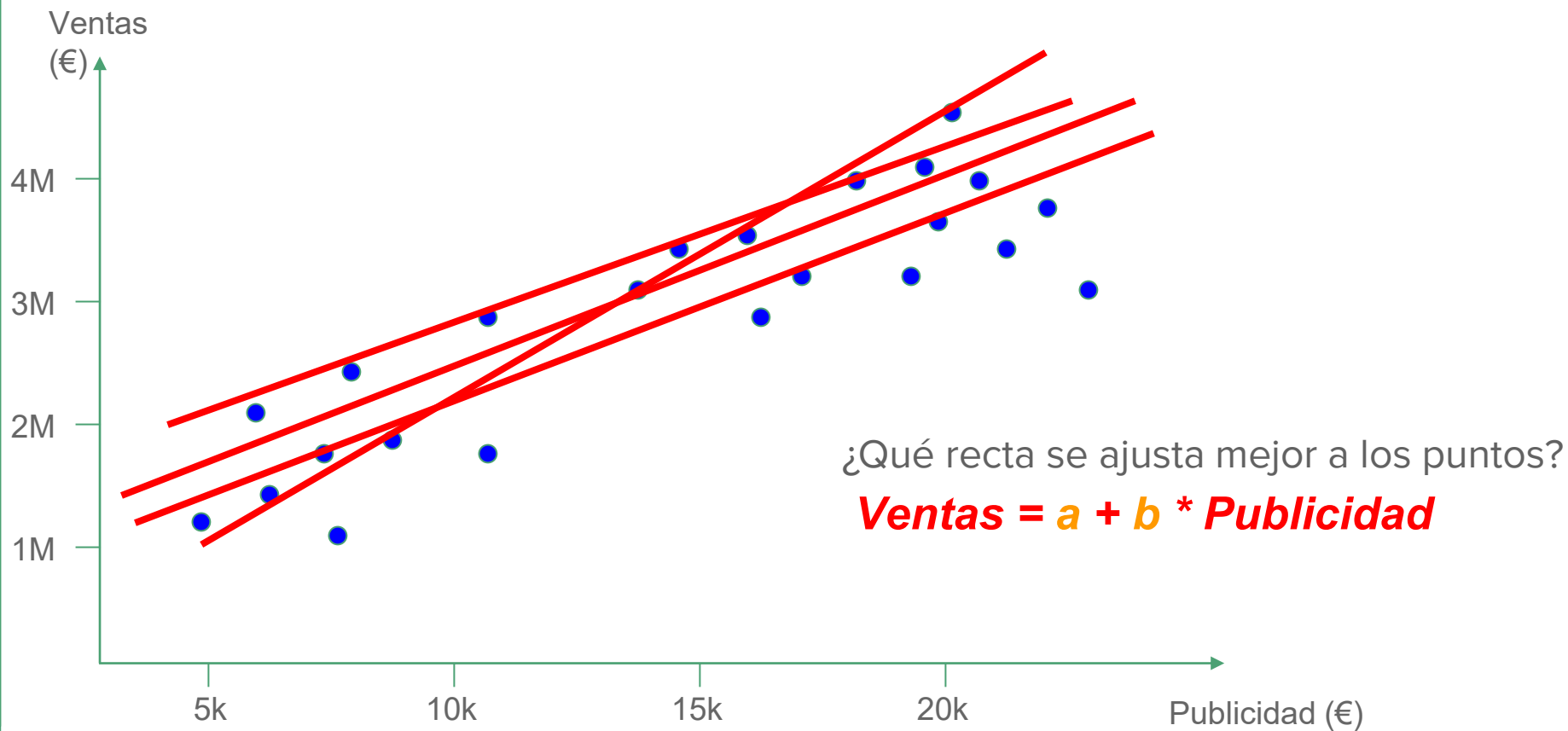
Análisis de regresión

- Proceso estadístico para estimar las relaciones entre variables. Incluye muchas técnicas para el modelado y análisis de diversas variables, cuando la atención se centra en la relación entre una variable dependiente y una o más variables independientes (o predictoras)

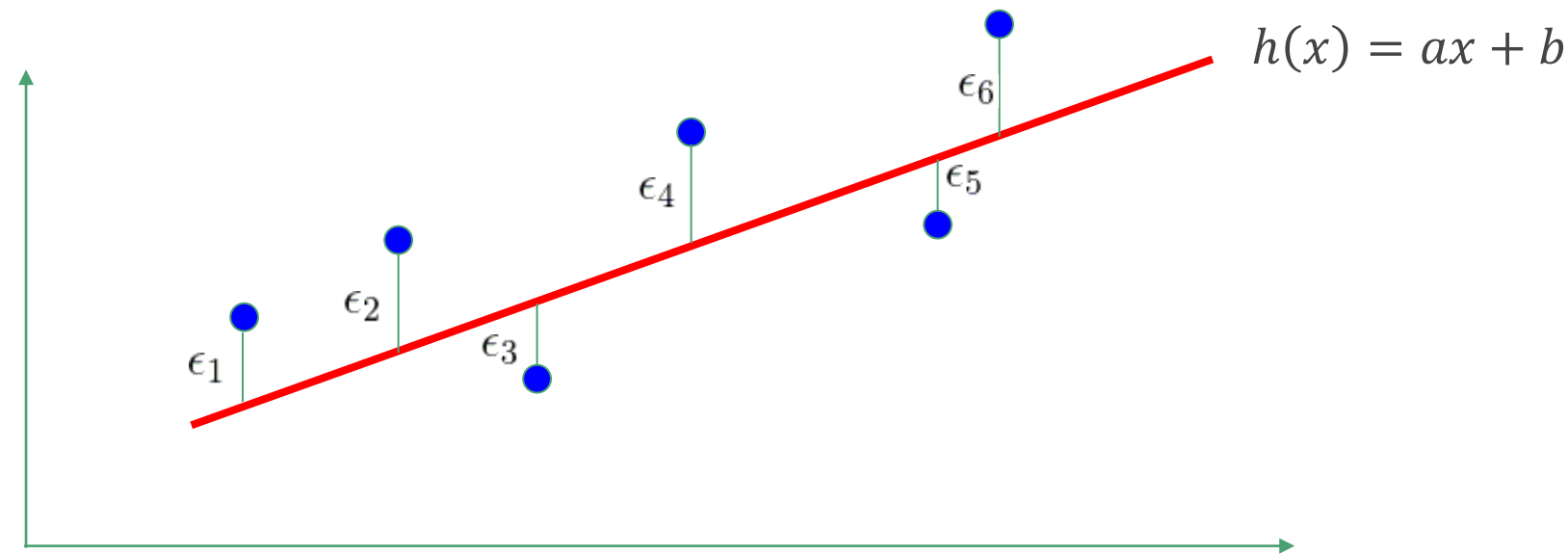
Tamaño	Habitaciones	Baños	Distancia metro	Antigüedad	Precio
53	1	1	120	30	120.000
67	2	1	240	25	164.000
125	4	2	10	7	250.000
84	3	1	500	12	198.000
...

$$\text{Precio} = -18.000 + 589 \times \text{Tamaño} + 22158 \times \text{Habitaciones} + 7714 \times \text{Baños} - 700 \times \text{Antigüedad}$$

Regresión Lineal



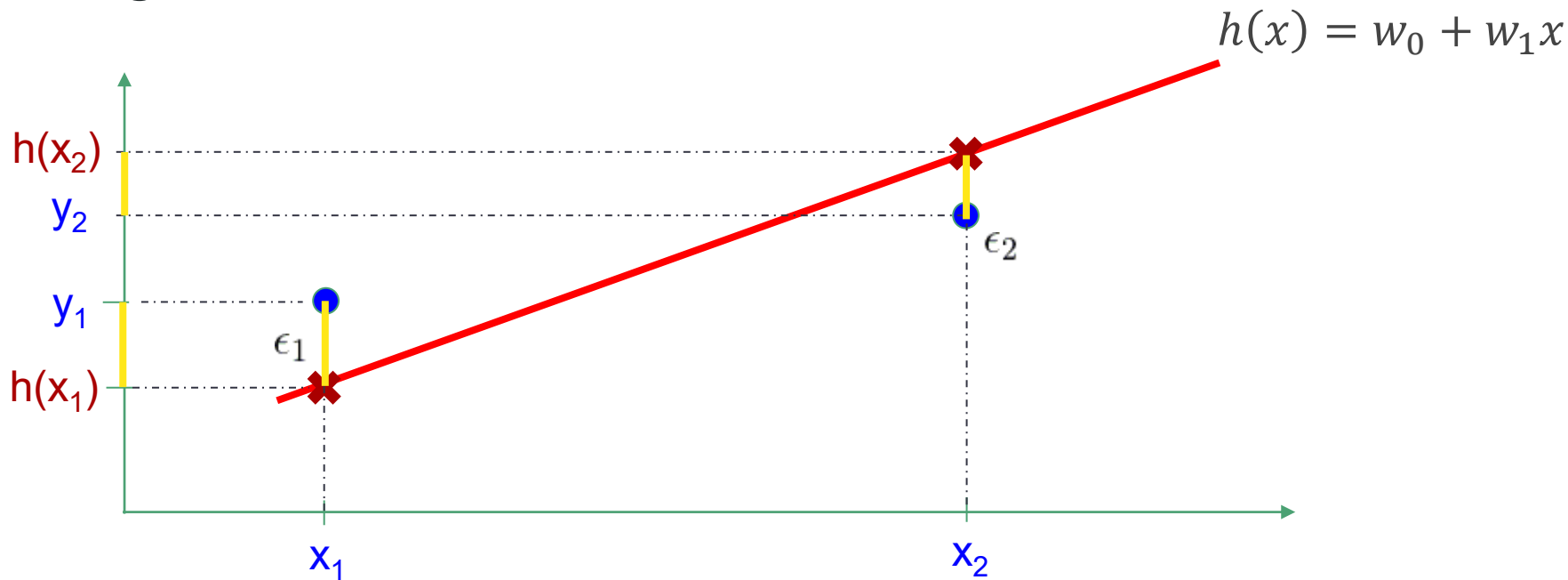
Regresión Lineal



Se busca la recta que minimice
$$\sum_n \epsilon_n^2 = \epsilon_1^2 + \epsilon_2^2 + \epsilon_3^2 + \epsilon_4^2 + \epsilon_5^2 + \epsilon_6^2 + \dots + \epsilon_N^2$$

Es decir, se minimiza la distancia entre la recta y todos los puntos, para obtener los coeficientes a y b

Regresión Lineal



$$\epsilon_1 = y_1 - h(x_1)$$

$$\epsilon_2 = h(x_2) - y_2$$

$$\Rightarrow E = \frac{1}{N} \sum_{n=1}^N (h(x_n) - y_n)^2$$

Función de error
(error cuadrático medio)

Regresión lineal

Pares entrada-salida: $(x_i, y_i), i = 1, \dots, N$

Hipótesis: $h(x) = w_0 + w_1 x$

Parámetros (pesos): w_0, w_1

Function de costo: $E_{in}(w_0, w_1) = \frac{1}{N} \sum_{n=1}^N (h(x_n) - y_n)^2$

Objetivo: $\underset{w_0, w_1}{\text{minimizar}} \quad E_{in}(w_0, w_1)$

Regresión lineal

Tamaño x_1	Habitaciones x_2	Baños x_3	Distancia metro x_4	Antigüedad x_5	Precio y
53	1	1	120	30	120.000
67	2	1	240	25	164.000
125	4	2	10	7	250.000
84	3	1	500	12	198.000
...

$$h(x) = w_0 + w_1x_1 + w_2x_2 + \dots + w_dx_d$$


Por conveniencia de notación, definimos $x_0 = 1 \Rightarrow h(x) = \sum_{i=0}^d w_i x_i = \mathbf{w}^T \mathbf{x}$

Regresión lineal

N ejemplos $(x_1, y_1), \dots, (x_N, y_N)$; d variables

$$\mathbf{x}_i = \begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_d \end{bmatrix}, \quad X = \begin{pmatrix} \text{---} \mathbf{x}_1^T \text{---} \\ \text{---} \mathbf{x}_2^T \text{---} \\ \vdots \\ \text{---} \mathbf{x}_N^T \text{---} \end{pmatrix}, \quad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}$$

$$E_{in}(\mathbf{w}) = \frac{1}{N} \sum_{n=1}^N (\mathbf{w}^T \mathbf{x}_n - y_n)^2 = \frac{1}{N} \|X\mathbf{w} - \mathbf{y}\|^2$$


 $h(x_n)$

Regresión lineal

Tamaño x_1	Habitaciones x_2	Baños x_3	Distancia metro x_4	Antigüedad x_5	Precio y
53	1	1	120	30	120.000
67	2	1	240	25	164.000
125	4	2	10	7	250.000
84	3	1	500	12	198.000
...

$$X = \begin{bmatrix} 1 & 53 & 1 & 1 & 120 & 30 \\ 1 & 67 & 2 & 1 & 240 & 25 \\ 1 & 125 & 4 & 2 & 10 & 7 \\ 1 & 85 & 3 & 1 & 500 & 12 \end{bmatrix}, y = \begin{bmatrix} 120000 \\ 164000 \\ 250000 \\ 198000 \end{bmatrix}$$

Objetivo: encontrar $\mathbf{w} = \begin{bmatrix} w_0 \\ w_1 \\ w_2 \\ w_3 \\ w_4 \\ w_5 \end{bmatrix}$

Análisis de regresión

$$\text{Minimizar } E_{in}(\mathbf{w}) = \frac{1}{N} |\mathbf{X}\mathbf{w} - \mathbf{y}|^2$$

$$\nabla E_{in}(\mathbf{w}) = \frac{2}{N} \mathbf{X}^T (\mathbf{X}\mathbf{w} - \mathbf{y}) = \mathbf{0}$$

$$\mathbf{X}^T \mathbf{X}\mathbf{w} - \mathbf{X}^T \mathbf{y} = \mathbf{0}$$

$$\mathbf{X}^T \mathbf{X}\mathbf{w} = \mathbf{X}^T \mathbf{y}$$

$$(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{X}\mathbf{w} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

$$\mathbf{I} \cdot \mathbf{w} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

$$\Rightarrow \boxed{\mathbf{w} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}}$$

Tamaño x_1	Habitaciones x_2	Precio y
53	1	120.000
67	2	164.000
125	4	250.000
84	3	198.000

$$\mathbf{X} = \begin{bmatrix} 1 & 53 & 1 \\ 1 & 67 & 2 \\ 1 & 125 & 4 \\ 1 & 85 & 3 \end{bmatrix}, \mathbf{y} = \begin{bmatrix} 120000 \\ 164000 \\ 250000 \\ 198000 \end{bmatrix}$$

$$\mathbf{X}^T = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 53 & 67 & 125 & 85 \\ 1 & 2 & 4 & 3 \end{bmatrix}$$

$$\Rightarrow \mathbf{w} = \begin{bmatrix} 66995.7 \\ 416.8 \\ 32645.8 \end{bmatrix}$$

$$\text{Precio} = 66995.7 + 416.8 \times \text{Tamaño} + 32645.8 \times \text{Habitaciones}$$

Resultados de la regresión

Call:

```
lm(formula = beneficio ~ gasto, data = datos1)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-5.8540	-1.9686	-0.5407	1.5360	14.1982

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-3.89578	0.71948	-5.415	4.61e-07 ***
gasto	1.19303	0.07974	14.961	< 2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.024 on 95 degrees of freedom

Multiple R-squared: 0.702, Adjusted R-squared: 0.6989

F-statistic: 223.8 on 1 and 95 DF, p-value: < 2.2e-16

Coefficientes de la recta de regresión:
 $y = -3.89578 + 1.19303 x$

el coeficiente R^2
(entre 0 y 1) evalúa
cómo el modelo se
ajusta a los datos

El valor p indica si la
regresión es
significativa ($p < 0.05$)

Significancia de la correlación

- Para probar la significancia del coeficiente de correlación, podemos usar el estadístico el cual se distribuye siguiendo una distribución t-student.

$$t = \frac{\rho}{\sqrt{\frac{1 - \rho^2}{n - 2}}}$$

- El coeficiente de correlación se considera estadísticamente significativo si el valor de t supera al valor de la distribución t-student con n-2 grados de libertad.
- El valor-p nos ofrece una medida de la significancia de la correlación. Cuanto menor sea, más significativa será la relación

Significancia de la correlación

- Ejemplo: Imaginemos que observamos un coeficiente de correlación de 0.50 entre la temperatura anual y el consumo de pan en 10 ciudades diferentes. ¿Esta relación es “fuerte”?
- El coeficiente de determinación (r -cuadrado=0.25) significa que el 25% de la varianza del consumo de pan se “explica” por la temperatura anual
- ¿Es esta relación significativa?
- Calculamos el estadístico y lo comparamos con la distribución t-student

$$t = \frac{\rho}{\sqrt{\frac{1-\rho^2}{n-2}}} = \frac{0.5}{\sqrt{\frac{1-0.5^2}{10-2}}} = 1.63$$

- El valor de t-student para 8 grados de libertad (10-2) y confianza del 95% es de 1.86
- Por tanto, no podemos rechazar la hipótesis nula ni afirmar que exista relación estadísticamente significativa entre ambas variables

Conclusiones

4. ESTADÍSTICA INFERENCIAL

4.1 Variables aleatorias

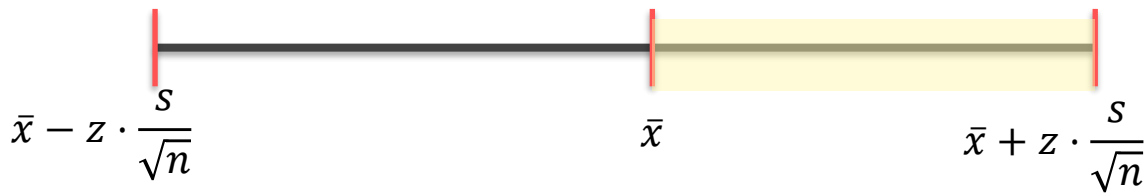
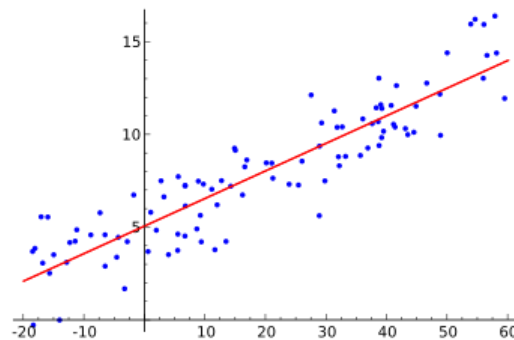
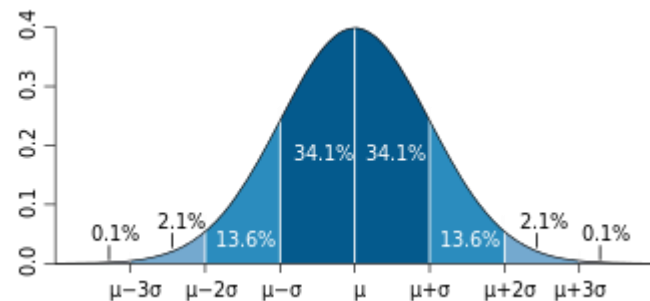
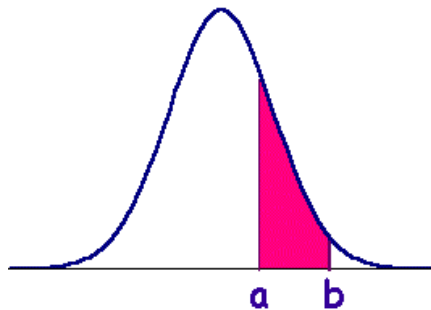
4.2 Distribuciones de probabilidad

4.3 Distribución normal

4.4 Intervalos de confianza

4.5 Error absoluto y tamaño de la muestra

4.6 Modelos de regresión lineal



¡Gracias!

Contacto: Rafael Zambrano

rafazamb@gmail.com