

# MAT-042: Estadísticas de resumen

**Felipe Osorio**

<http://fosorios.mat.utfsm.cl>

Departamento de Matemática, UTFSM



Desde 1988 el SIMCE evalúa los **resultados de aprendizaje** de los estudiantes del sistema de educación chileno.

## Objetivos:

- ▶ Describir el **comportamiento del aprendizaje** de los estudiantes.
- ▶ Determinar si existe diferencias significativas entre el **tipo de dependencia** (municipal, subvencionado, particular).

## Características del problema:

- ▶ Mediciones de un mismo individuo (estudiante) **a través del tiempo** (4<sup>º</sup> y 8<sup>º</sup> básico, 2<sup>º</sup> medio).<sup>1</sup>
- ▶ Datos disponibles para los años 2007, 2011 y 2013, pruebas de Lenguaje y Matemáticas.
- ▶ Aproximadamente **133K estudiantes** para ser analizados (base de datos de **mediano porte**).

---

<sup>1</sup>Conocido como: **datos con estructura longitudinal**.



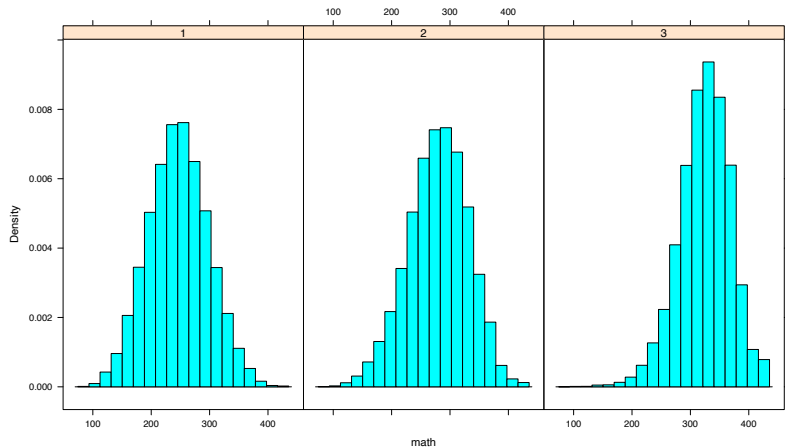


Figure: histograma puntajes matemática.

<sup>2</sup>colegios, 1: municipales, 2: subvencionados y 3: particulares.

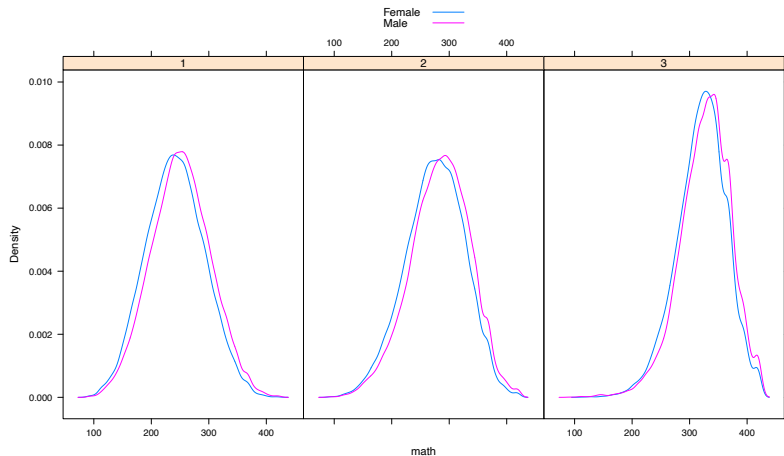


Figure: densidad puntajes matemática, organizados por Sexo.



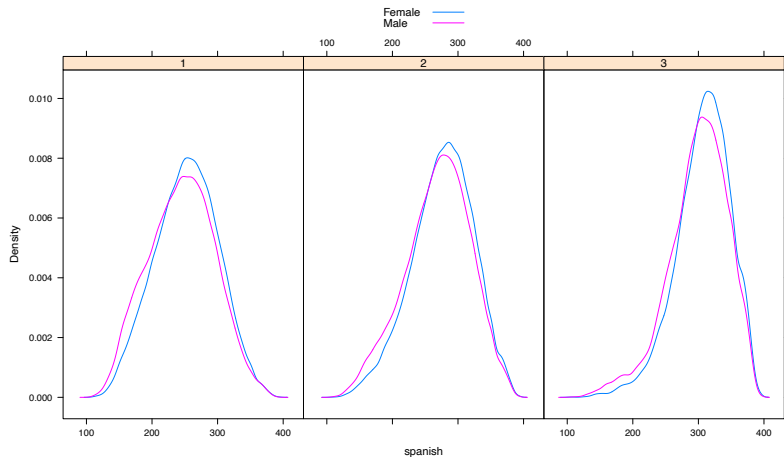


Figure: densidad puntajes lenguaje, organizados por Sexo.



# Datos del SIMCE

## Base de datos con aproximadamente 133K individuos

> SIMCE

	Sex	type	math04	math08	math10	spa04	spa08	spa10
1	Male	1	338.86	303.94	372.51	342.74	327.92	317.38
2	Female	2	301.98	256.04	324.65	298.30	263.12	322.40
3	Female	1	258.45	263.44	225.95	192.59	206.72	216.66
4	Male	2	233.13	323.76	288.60	268.91	274.84	251.44
5	Male	1	284.17	276.37	293.11	236.55	261.67	283.78
6	Male	1	248.64	259.76	210.17	254.34	252.15	280.53

...

132947	Female	2	211.78	254.21	246.78	244.97	286.21	269.24
132948	Female	3	285.18	315.25	354.90	303.95	341.67	315.81
132949	Male	1	259.05	232.65	224.18	305.65	195.92	217.71



*Para pensar:*

- ▶ ¿Cómo resumir la información del total de 133K datos para cada una de las 8 variables?<sup>3</sup>
- ▶ ¿Podemos usar, digamos unas pocas cantidades para describir esta información?

---

<sup>3</sup>Es decir un poco más de 1 millón de registros.



## Ingredientes:

Conjunto de  $n$  observaciones  $\{x_1, x_2, \dots, x_n\}$  conocidas como **muestra**.

En general, nuestro interés recaerá en resúmenes de la información a través de una **estadística**, digamos  $T = T(x_1, \dots, x_n)$ .

En esta clase consideraremos 3 tipos de **estadísticas de resumen**<sup>4</sup> para una muestra  $\mathbf{x} = (x_1, \dots, x_n)^\top$ ,

- ▶ medidas de posición.
- ▶ medidas de dispersión.
- ▶ medidas de forma (asimetría y curtosis).

---

<sup>4</sup>En ocasiones escribiremos  $T = T(\mathbf{x})$ .





## Definición 3 (Media muestral o promedio):

Sea  $x_1, \dots, x_n$  valores muestrales. Se define el **promedio** o **media muestral** como:

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + \dots + x_n) = \frac{1}{n} \sum_{i=1}^n x_i.$$

Suponga que la observación  $i$ -ésima, digamos  $x_i$ , se repite  $n_i$  veces. Entonces tenemos

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

donde  $f_i = n_i/n$  es la frecuencia relativa. En general, considere “pesos”  $\omega_1, \dots, \omega_n$  asociados a las observaciones  $x_1, \dots, x_n$ . En este caso,

$$\bar{x} = \frac{1}{\sum_{j=1}^n \omega_j} \sum_{i=1}^n \omega_i x_i.$$



## Ejemplo:

Considere el conjunto de datos  $x = \{1, 2, 2, 2, 3, 3, 8\}$ . Tenemos  $n = 7$ , y

$$\begin{aligned}\sum_{i=1}^7 x_i &= 1 + 2 + 2 + 2 + 3 + 3 + 8 \\ &= 1 + 2 \cdot 3 + 2 \cdot 3 + 8 = 21,\end{aligned}$$

así  $\bar{x} = 21/7 = 3$ . Note también que el gráfico de **tallo y hoja**, adopta la forma:

1		*		
2		*	*	*
3		*	*	
4				
5				
6				
7				
8		*		



### *Ejemplo (datos de accidentes):*

Suponga el siguiente conjunto de datos:

Número de accidentes ( $x_i$ )	Frecuencia ( $n_i$ )	$n_i x_i$
0	55	0
1	14	14
2	5	10
3	2	6
4	0	0
Total	76	30

De este modo,  $\bar{x} = 30/76 = 0.3947$  es el número promedio de accidentes.



## Definición 4 (Estadísticos de orden):

Sea  $x_1, \dots, x_n$  una muestra. Entonces los valores ordenados

$$x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)},$$

se denominan **estadísticos de orden**. Algunas estadísticas de orden son: el **mínimo** muestral  $x_{(1)}$ , el **máximo** muestral  $x_{(n)}$ .

## Definición 5 (Mediana):

Sea  $x_{(1)}, \dots, x_{(n)}$  observaciones ordenadas. La **mediana** es definida como:

$$\text{me}(\mathbf{x}) = \begin{cases} x_{(n+1)/2}, & n \text{ es impar,} \\ (x_{(n/2)} + x_{(n/2+1)})/2, & n \text{ es par.} \end{cases}$$



### Observación:

Sea  $f(x)$  cualquier función de números reales.<sup>5</sup> Entonces podemos definir

$$\bar{f} = \frac{1}{n} \sum_{i=1}^n f(x_i) = \frac{1}{n} (f(x_1) + \cdots + f(x_n)).$$

### Caso particular (media geométrica):

Suponga  $x_1, \dots, x_n$  números positivos y  $f(x) = \log(x)$ . Entonces la **media geométrica**  $G$  es dada por:

$$\log G = \frac{1}{n} (\log x_1 + \cdots + \log x_n) = \frac{1}{n} \sum_{i=1}^n \log x_i.$$

Es decir,

$$G = (x_1 \cdot x_2 \cdots x_n)^{1/n} = \left( \prod_{i=1}^n x_i \right)^{1/n}.$$

---

<sup>5</sup>Por ejemplo,  $f(x) = x^2$  lleva a la media cuadrática, mientras que  $f(x) = 1/x$  es la media armónica.



## Datos del SIMCE: Puntajes de matemáticas

```
## sólo puntajes de matemáticas
> MATH
math04 math08 math10
1 338.86 303.94 372.51
2 301.98 256.04 324.65
3 258.45 263.44 225.95
4 233.13 323.76 288.60
5 284.17 276.37 293.11
6 248.64 259.76 210.17
...

> x <- MATH$math04 # análogamente x <- MATH[,1]
> mean(x)          # promedio
[1] 261.5766
> median(x)        # mediana
[1] 263.96
> library(fastmatrix) # https://faosorios.github.io/fastmatrix
> geomean(x)       # media geométrica
[1] 256.0357
# alternativamente: exp(mean(log(x)))

> apply(MATH, 2, mean) # para todas la variables
  math04  math08  math10
261.5766 269.6779 276.6267
```



## Medidas de dispersión

Considere los siguientes conjuntos de datos:

$$D_1 = \{10, 20, 30\}, \quad D_2 = \{5, 5, 20, 35, 35\}, \quad D_3 = \{20, 20, 20\},$$

Tenemos los gráficos de tallo-y-hoja:

Datos  $D_1$ :

5		
10		*
15		
20		*
25		
30		*
35		

Datos  $D_2$ :

5		*	*
10			
15			
20		*	
25			
30			
35		*	*

Datos  $D_3$ :

5				
10				
15				
20		*	*	*
25				
30				
35				



## Medidas de dispersión

Sea  $\bar{x}_j$  y  $me_j$  el promedio y la mediana asociada al conjunto de datos  $D_j$  ( $j = 1, 2, 3$ ).  
Entonces,

$$\bar{x}_1 = \frac{1}{3}(10 + 20 + 30) = \frac{60}{3} = 20,$$

$$\bar{x}_2 = \frac{1}{5}(2 \cdot 5 + 20 + 2 \cdot 35) = \frac{100}{5} = 20,$$

$$\bar{x}_3 = \frac{3 \cdot 20}{3} = 20.$$

Además,  $me_j = 20$  para todo  $j$ .

### *Observación:*

Es decir, tenemos tres configuraciones de datos con **valores centrales idénticos**.





Sean  $Q_1$  y  $Q_3$  las medianas de la mitad inferior y superior de los datos, conocidos como el 1er y 3er **cuartiles**, respectivamente. Esto permite definir:

$$IQR = Q_3 - Q_1,$$

que es conocido como **rango intercuartílico**.

También podemos considerar el **rango** de la muestra como:

$$R = \max\{x_i\}_{i=1}^n - \min\{x_i\}_{i=1}^n = x_{(n)} - x_{(1)}.$$

Algunos software estadísticos (**R/S-Plus**, **Stata**, entre otros) reportan:

$$x_{(1)}, Q_1, \text{me}, Q_3, x_{(n)}.$$



## Medidas de dispersión

Considere subdividir los datos ordenados  $x_{(1)}, \dots, x_{(n)}$  en secciones de 100%, llamados **percentiles**. Entonces el percentil de orden  $j$  ( $1 \leq j \leq 100$ ) está dado por:

$$P_j = x_{(j(n+1)/100)}.$$

Note que  $Q_1 = P_{25}$ , la mediana (o 2º cuartil,  $Q_2$ ) es  $me = P_{50}$  y  $Q_3 = P_{75}$ .

### Ejemplo:

Considere el conjunto de datos  $x = \{4, 7, 18, 1, 7, 13, 2\}$  y suponga que deseamos calcular el rango intercuartílico  $IQR$ .

Primeramente es necesario ordenar el conjunto de datos:

$$\{x_{(1)}, x_{(2)}, x_{(3)}, x_{(4)}, x_{(5)}, x_{(6)}, x_{(7)}\} = \{1, 2, 4, 7, 7, 13, 18\}.$$

Disponemos de  $n = 7$  datos, luego para obtener el 1er y 3er cuartiles podemos usar

$$Q_1 = P_{25} = x_{(25 \cdot (7+1)/100)} = x_{(1.8/4)} = x_{(2)} = 2,$$

$$Q_3 = P_{75} = x_{(75 \cdot (7+1)/100)} = x_{(3.8/4)} = x_{(6)} = 13.$$

De este modo,  $IQR = Q_3 - Q_1 = 13 - 2 = 11$ .



## Definición 6 (Varianza muestral):

Considere  $x_1, \dots, x_n$  valores observados, se define su **varianza muestral** como:

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

*Observación:*

$s = \sqrt{s^2}$  se denomina **desviación estándar**.



## Observación:

Otras medidas de dispersión:

- ▶ Desviación absoluta en torno de la media:

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

- ▶ Desviación absoluta en torno de la mediana:

$$\frac{1}{n} \sum_{i=1}^n |x_i - \text{me}(\mathbf{x})|.$$

- ▶  $r$ -ésimo momento centrado en torno de  $a$ :<sup>6</sup>

$$m_r(a) = \frac{1}{n} \sum_{i=1}^n (x_i - a)^r.$$

---

<sup>6</sup>Para  $r = 2$  y  $a = \bar{x}$  obtenemos la varianza.



## Propiedades:

(a)

$$\sum_{i=1}^n (x_i - \bar{x}) = 0.$$

(b)

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

(c)  $\bar{x}$  es el valor que minimiza la función:

$$S(a) = \sum_{i=1}^n (x_i - a)^2.$$

(d) Sea  $x_1, \dots, x_n$  y considere la transformación:

$$y_i = a x_i + b, \quad i = 1, \dots, n.$$

Entonces  $\bar{y} = a\bar{x} + b$  y  $s_y^2 = a^2 s_x^2$ .



(a) En efecto,

$$\sum_{i=1}^n (x_i - \bar{x}) = \sum_{i=1}^n x_i - \sum_{i=1}^n \bar{x} = \sum_{i=1}^n x_i - n\bar{x} = n\bar{x} - n\bar{x} = 0.$$

(b) (Fórmula de Köning)

$$\begin{aligned}\sum_{i=1}^n (x_i - \bar{x})^2 &= \sum_{i=1}^n (x_i^2 - 2x_i\bar{x} + \bar{x}^2) = \sum_{i=1}^n x_i^2 - 2\bar{x} \sum_{i=1}^n x_i + n\bar{x}^2 \\ &= \sum_{i=1}^n x_i^2 - n\bar{x}^2 = \sum_{i=1}^n x_i^2 - \frac{1}{n} \left( \sum_{i=1}^n x_i \right)^2\end{aligned}$$



- (c)  $\bar{x}$  es el valor que minimiza la función  $S(a) = \sum_{i=1}^n (x_i - a)^2$ . En efecto, note que

$$\frac{d}{da} S(a) = \sum_{i=1}^n \frac{d}{da} (x_i - a)^2 = -2 \sum_{i=1}^n (x_i - a),$$

resolviendo la condición de primer orden, tenemos

$$\sum_{i=1}^n (x_i - \hat{a}) = 0,$$

desde donde sigue que  $\hat{a} = \bar{x}$ . Además

$$\frac{d^2}{da^2} S(a) = -2 \sum_{i=1}^n \frac{d}{da} (x_i - a) = 2n,$$

y como la segunda derivada es positiva (para cualquier valor de  $n$ ), obtenemos que  $\bar{x}$  es máximo global.



## Propiedades del promedio y varianza muestrales

- (d) Sea  $x_1, \dots, x_n$  y considere la transformación,  $y_i = ax_i + b$ , para  $i = 1, \dots, n$ .  
Entonces

$$\begin{aligned}\bar{y} &= \frac{1}{n} \sum_{i=1}^n y_i = \frac{1}{n} \sum_{i=1}^n (ax_i + b) = \frac{1}{n} \left( a \sum_{i=1}^n x_i + nb \right) \\ &= a \left( \frac{1}{n} \sum_{i=1}^n x_i \right) + b = a\bar{x} + b.\end{aligned}$$

Mientras que

$$s_y^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2,$$

como  $y_i - \bar{y} = ax_i + b - (a\bar{x} + b) = a(x_i - \bar{x})$ , sigue que

$$\begin{aligned}s_y^2 &= \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2 = \frac{1}{n-1} \sum_{i=1}^n \{a(x_i - \bar{x})\}^2 \\ &= a^2 \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = a^2 s_x^2.\end{aligned}$$





### Observación:

Un caso particular de importancia es la **estandarización** del conjunto de datos  $x_1, \dots, x_n$ , definida como:

$$z_i = \frac{x_i - \bar{x}}{s}, \quad i = 1, \dots, n.$$

Entonces,<sup>7</sup>

$$\bar{z} = 0 \quad \text{y} \quad s_z^2 = 1.$$

---

<sup>7</sup>Basta hacer  $a = 1/s$  y  $b = \bar{x}/s$  en la Propiedad (d).



### Definición 7 (Coeficiente de variación):

Este coeficiente es una medida que compara la desviación estándar con el promedio de una muestra y es definido como

$$CV = s/\bar{x}, \quad \bar{x} \neq 0.$$

El coeficiente es particularmente útil para **comparar dos o más muestras** (o grupos).

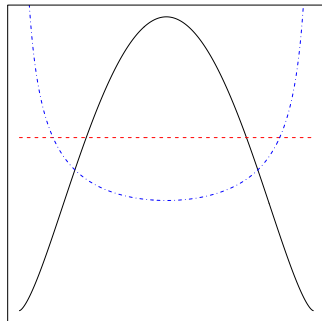
#### *Observación:*

Un valor pequeño para el CV está asociado a una muestra homogénea.

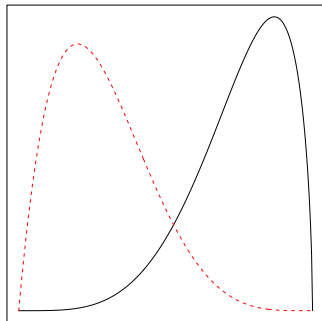
#### *Observación:*

En Econometría  $1/CV$  es conocido como la **razón de Sharpe**.





(a) distribuciones simétricas



(b) asimetría negativa (-), positiva (- -)

### Definición 8 (Coeficiente de asimetría):

Considere  $m_3$  el tercer momento muestral, entonces se define el **coeficiente de asimetría** (o sesgo) como:

$$b_1 = \frac{m_3}{s^3} = \frac{1}{n} \sum_{i=1}^n \left( \frac{x_i - \bar{x}}{s} \right)^3.$$

#### Observación:

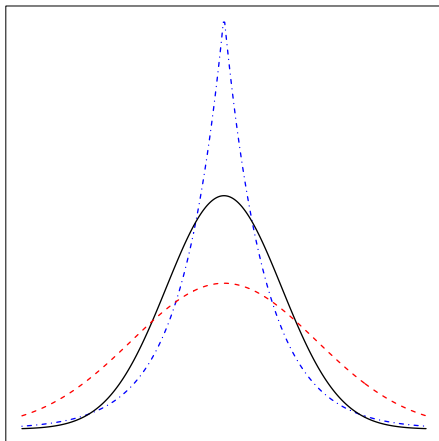
- ▶ Si  $b_1 = 0$  la distribución es simétrica con relación a  $\bar{x}$ .
- ▶ Si  $b_1 > 0$  la distribución tiene **sesgo positivo**. En caso contrario, decimos que tiene **sesgo negativo**.

#### Observación:

Se han definido diversos índices de simetría, por ejemplo la **medida de sesgo de Galton**:

$$b_G = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{Q_3 - Q_1}.$$





(a) Distribución leptocúrtica (- · -), mesocúrtica (—) y platicúrtica (— —)

### Definición 9 (Coeficiente de curtosis):

Considere  $m_4$  el cuarto momento muestral, entonces se define el **coeficiente de curtosis**<sup>8</sup> como:

$$b_2 = \frac{m_4}{s^4} - 3 = \left\{ \frac{1}{n} \sum_{i=1}^n \left( \frac{x_i - \bar{x}}{s} \right)^4 \right\} - 3.$$

#### *Observación:*

El término  $-3$  hace que  $b_2 = 0$  cuando los datos siguen una **distribución normal**.

---

<sup>8</sup>también conocido como **exceso de curtosis**



## Datos del SIMCE: Puntajes de matemáticas<sup>9</sup>

```
> z <- quantile(x)
> z
      0%      25%      50%      75%     100%
 87.74 226.32 263.96 299.29 369.55

> sd(x) # desviación estándar
[1] 51.79042
> var(x) # varianza
[1] 2682.247

> library(fastmatrix) # https://faosorios.github.io/fastmatrix
> moments(x)
$second
[1] 2682.227
$third
[1] -30409.6
$fourth
[1] 18784749
$skewness
[1] -0.2189084
$kurtosis
[1] -0.3889947
```

---

<sup>9</sup> $n = 132\,793$  observaciones, así que  $(n-1)/n = 0.9999925$ .



## Gráfico de cajón con bigotes (boxplot)

