# MAT-042: Estadísticas de resumen

**Felipe Osorio**

http://fosorios.mat.utfsm.cl

Departamento de Matemática, UTFSM

# Investigación Estadística

**Método Científico:**

Observación sistemática, medición y experimentación, y la formulación, análisis y modificación de las hipótesis.

- Etapas de una Investigación Estadística.
  - Formulación del problema.
  - Diseño del experimento.
  - Experimentación y recolección de datos.
  - Tabulación y descripción de los resultados.
  - Inferencia estadística.

# Conceptos básicos

- Población: Conjunto de entidades (indivíduos, elementos) desde los que se desea extraer información, i.e. hacer inferencias.
- Muestra: Es un subconjunto de la población, seleccionada de acuerdo a una regla o plan.
- Variable: Características o atributos de los elementos que conforman la población.
  - Categóricas: Partición en dos o más clases (variables discretas o factores).
    - Binarias: sólo dos categorías (masc/fem, fumador/no fumador).
    - Nominal: no existe orden entre categorías (país de origen).
    - Ordinal: categorías con un orden natural (leve/moderado/grave).
  - Cuantitativas: Pueden adoptar infinitos valores sobre un conjunto ($\mathbb{R}$).

# Tipos de muestreo

Ejemplo: Datos del SIMCE.

- Regiones geográficas.
- Niveles educacionales: 2º, 4º, 8º básico; 2º medio.
- Dependencia: Municipal, Subvencionado, Particular.
- Área: Urbano, Rural.

# Tipos de muestreo

- Muestreo Aleatorio Simple (m.a.s.)

  Todas las muestras posibles de tamaño $n$ desde una población de tamaño $N$ tienen la misma probabilidad de ser escogida.

- Muestreo Estratificado

  Se emplea cuando la población está agrupada en varios grupos homogéneos o estratos. Luego, se obtiene una m.a.s. desde cada estrato.

- Muestreo Sistemático

  En este caso las unidades de la población están ordenadas y se selecciona la muestra aprovechando este ordenamiento.

- Muestreo por Conglomerado

  Se emplea cuando la población está dividida en grupos pequeños. Consiste en una m.a.s. de conglomerados y luego se censa cada uno de éstos.

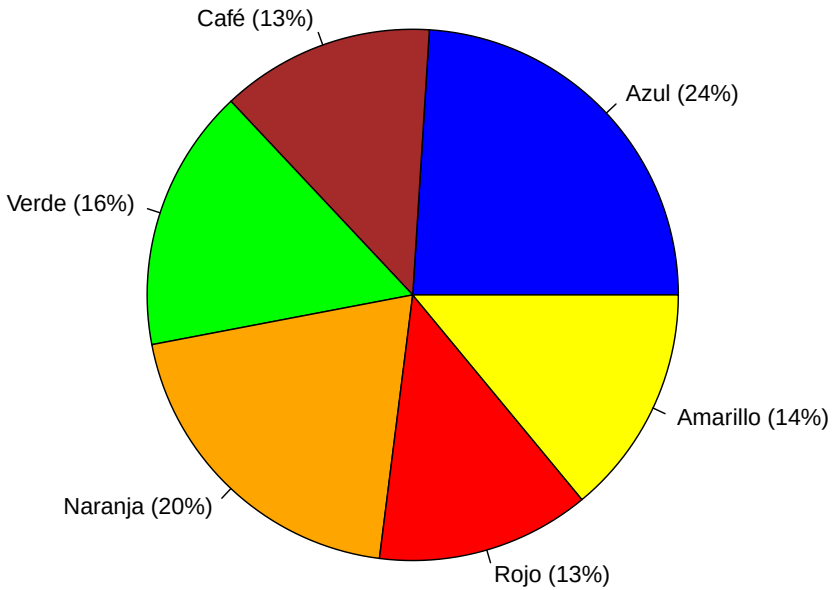
- Muestreo en dos Etapas

  Primero se selecciona una muestra de unidades primarias y luego se realiza un muestreo desde cada muestra escogida.

# Gráfico Circular

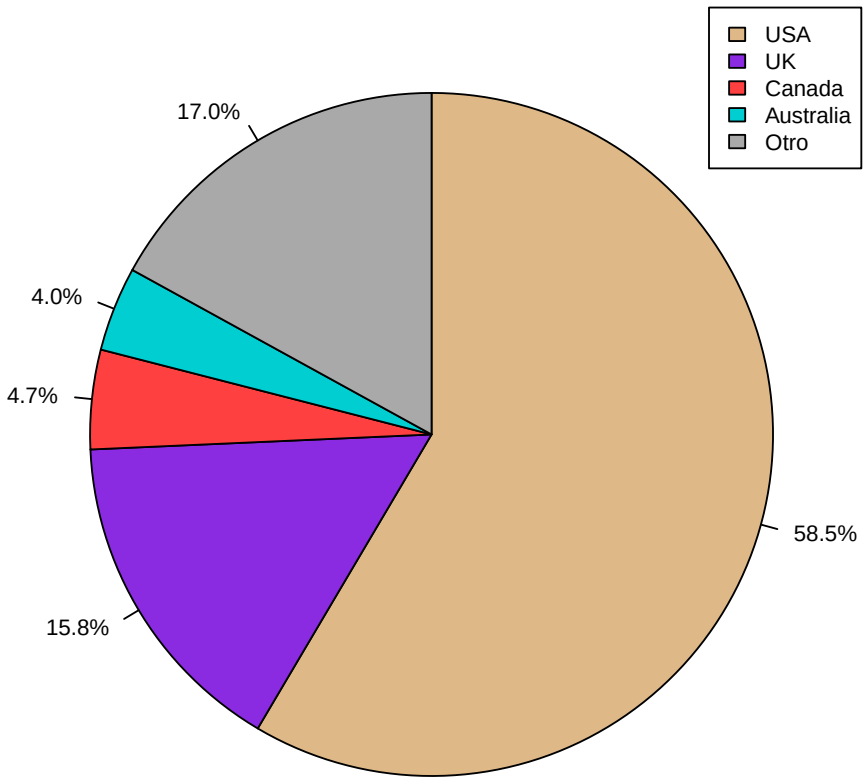
Gráfico circular: se usa para representar magnitudes en frecuencias o porcentajes. El largo de arco (i.e. área) de cada sector es proporcional a la cantidad que representa.

Ejemplo: Distribución de colores en bolsitas de M&M (chocolate de leche)



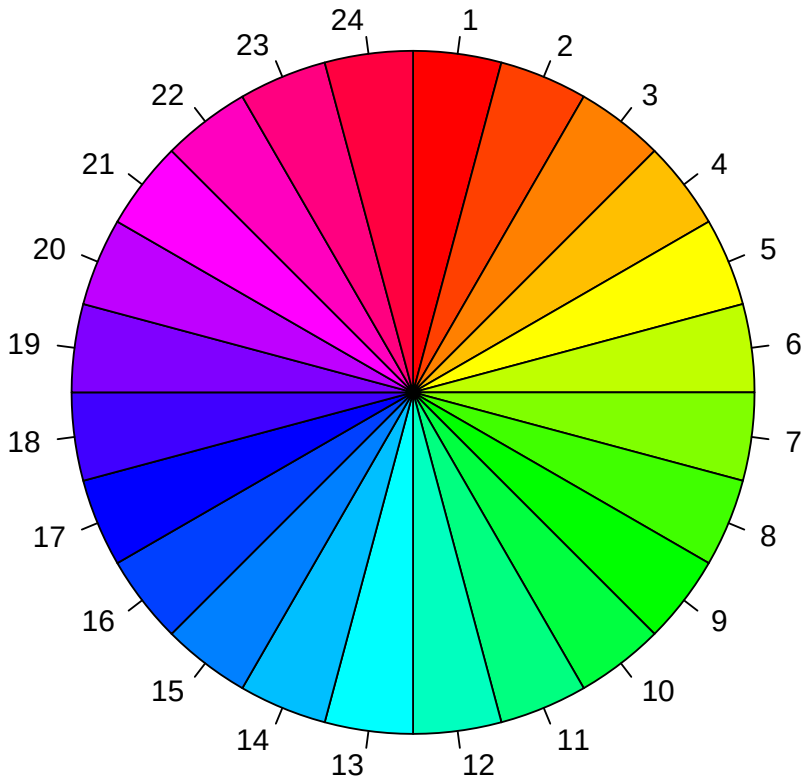
# Gráfico Circular

Ejemplo: Proporción de población anglófona en el mundo

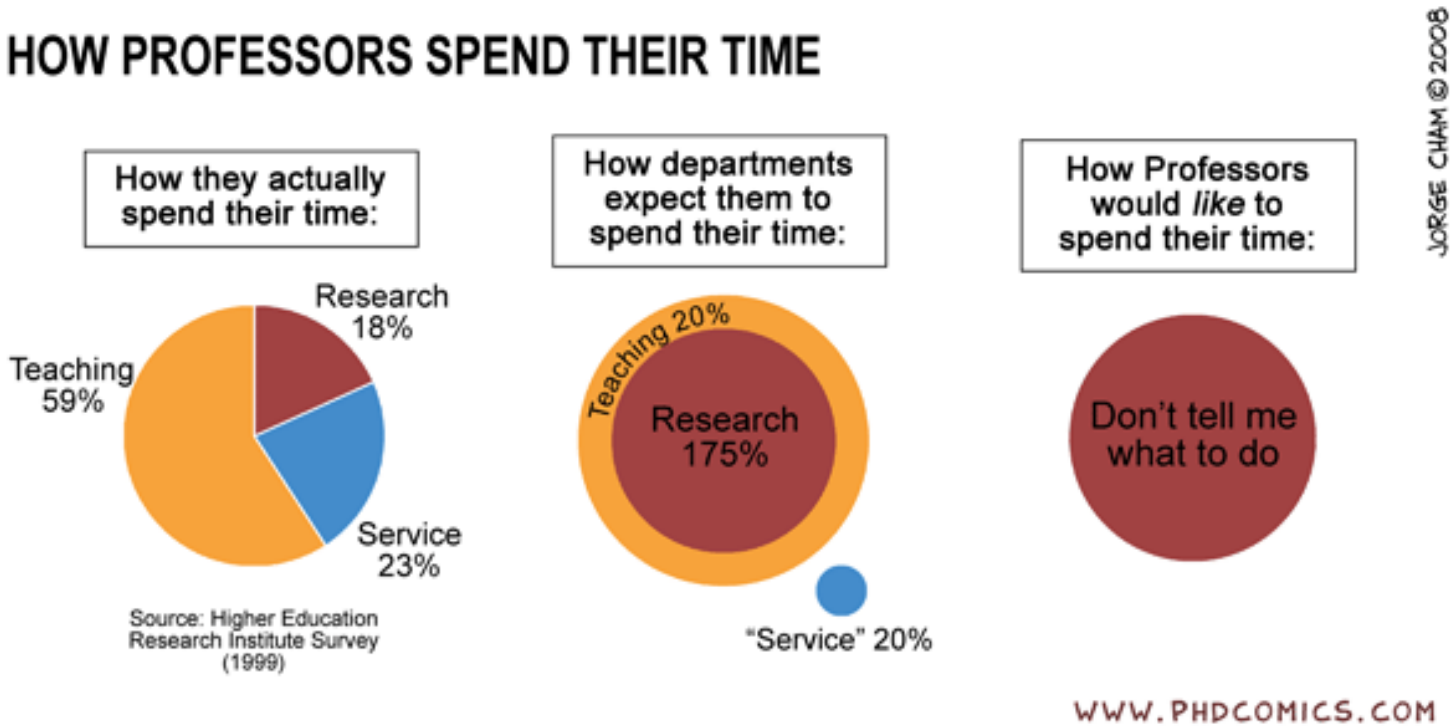


# Gráfico Circular

Limitación: Gráfico circular de un arcoiris.



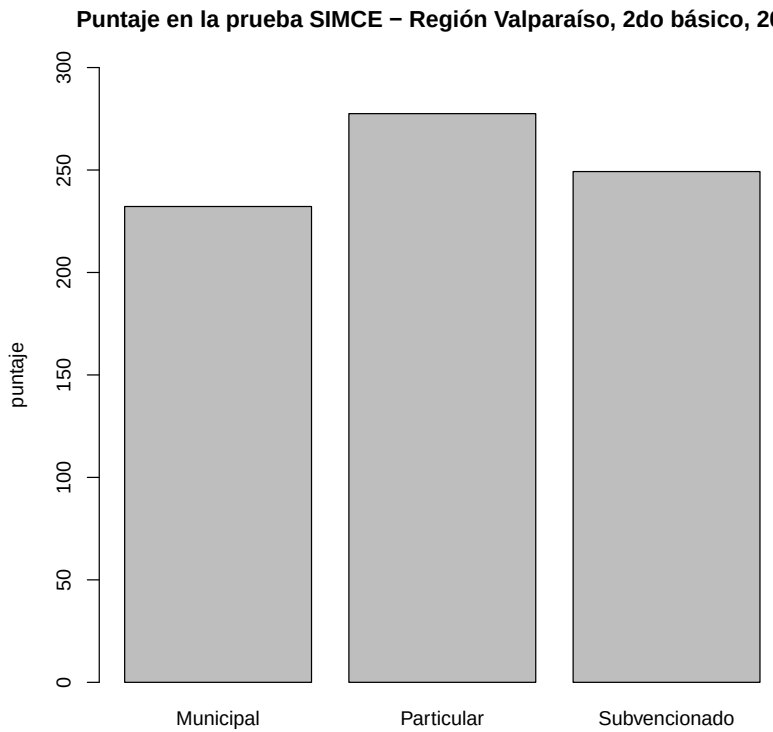
# Gráfico Circular



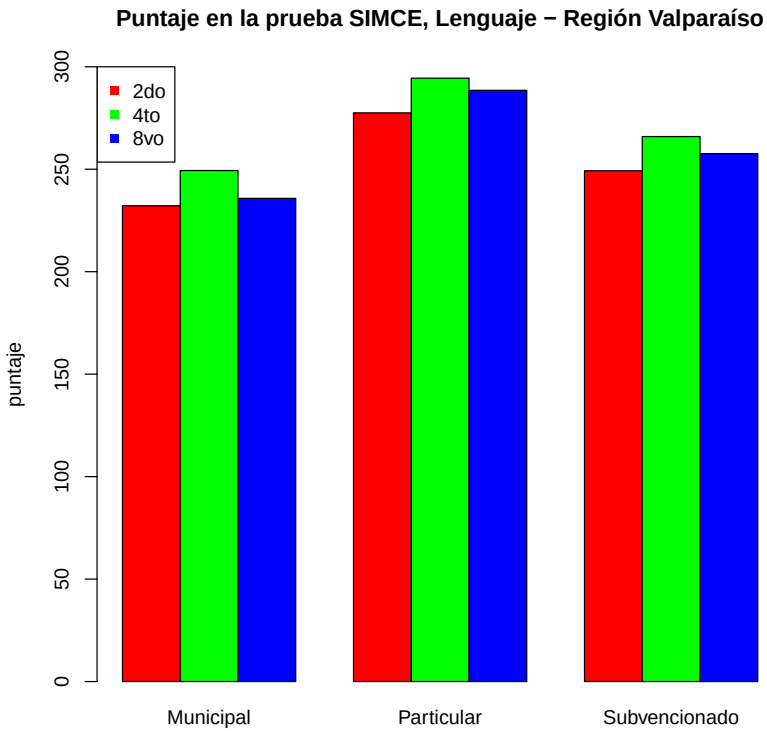
# Gráfico de Barras

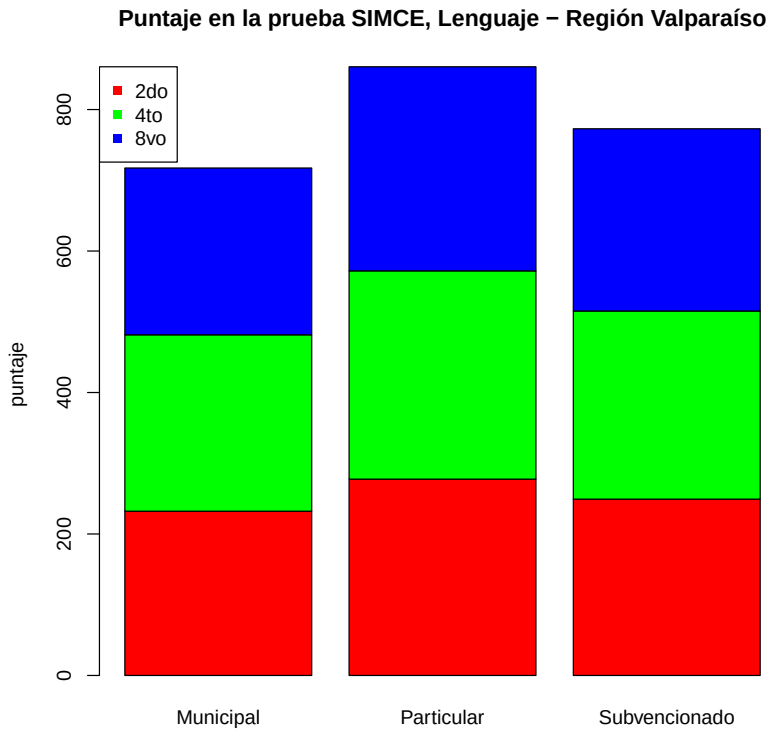
Gráfico de barras (bloques): la magnitud de la variable es representada por la altura de un rectángulo. Permite una mejor comparación que juzgando áreas relativas.

Ejemplo: Promedio prueba SIMCE Lenguaje en 631 establecimientos de Valparaíso.

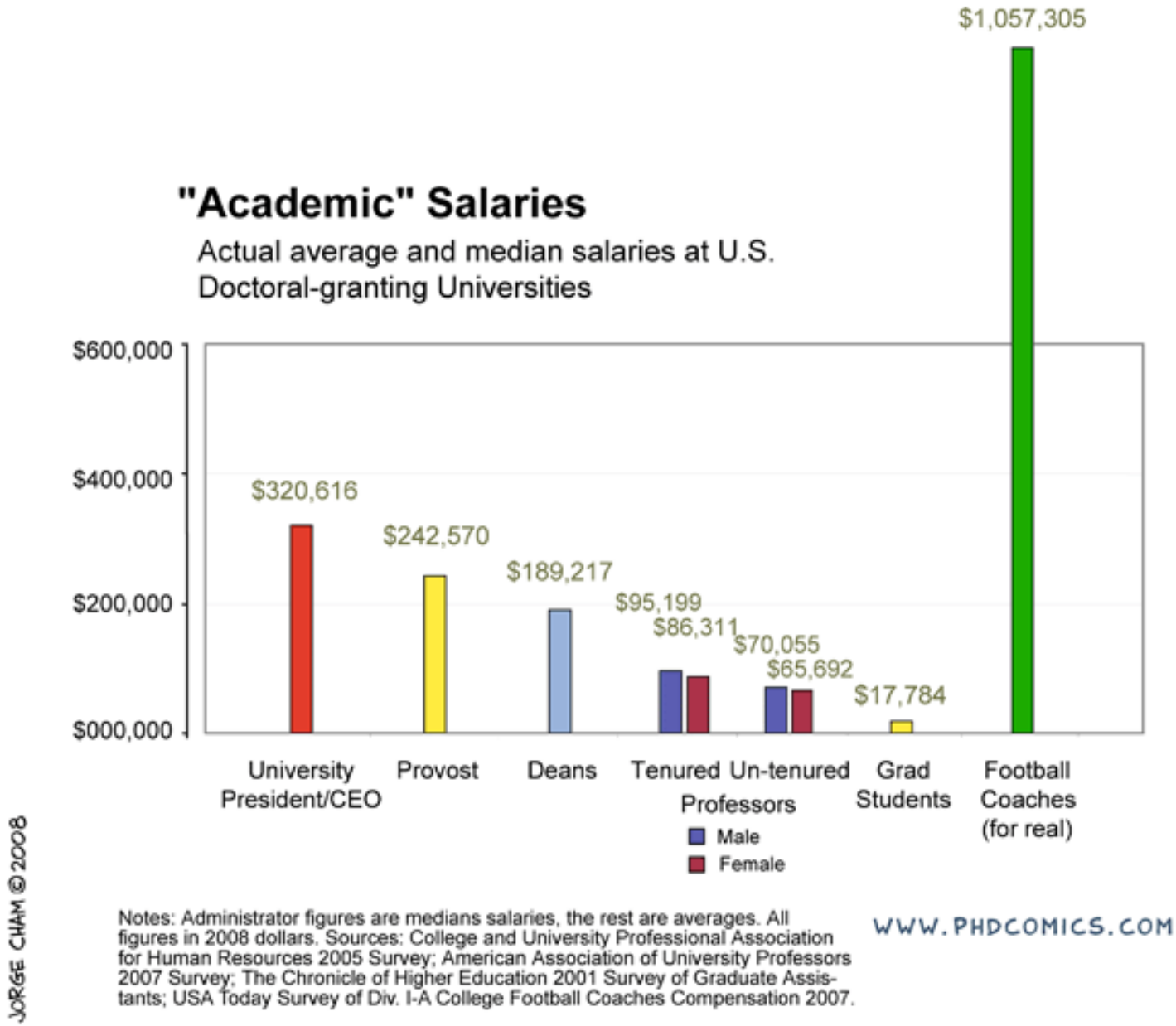


# Gráfico de Barras



# Gráfico de Barras



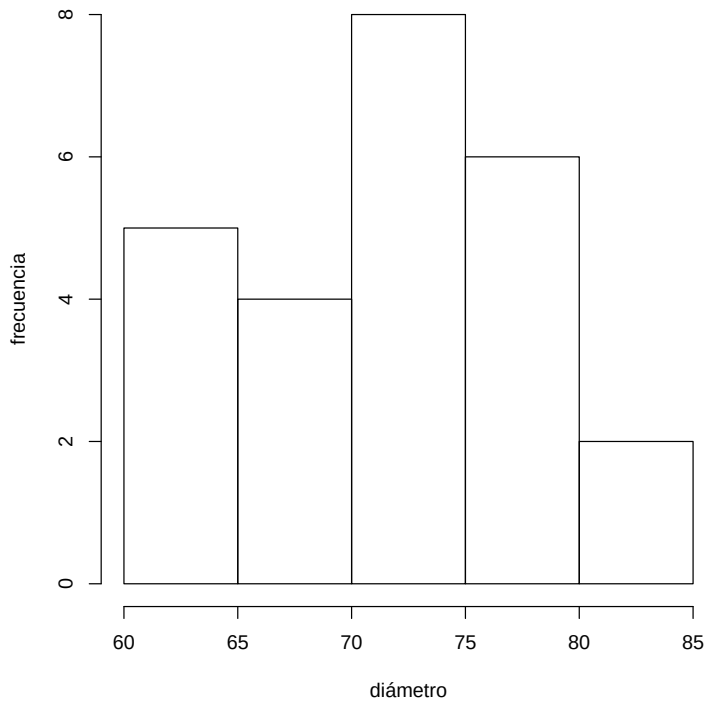
# Histograma

Histograma: usa rectángulos para visualizar frecuencias y proporciones. Se debe:

- dividir el rango de los datos en "bins"
- contar el número de observaciones en cada clase
- dibujar rectángulos representando las frecuencias o porcentajes

Ejemplo: Diámetro (mm) de pernos producidos por una máquina en un día.

| | | | | | | |
|---|---|---|---|---|---|---|
| 72 | 61 | 76 | 76 | 67 | 67 | 77 |
| 77 | 72 | 69 | 62 | 71 | 67 | 63 |
| 71 | 81 | 64 | 72 | 73 | 72 | 78 |
| 73 | 76 | 65 | 84 | | | |

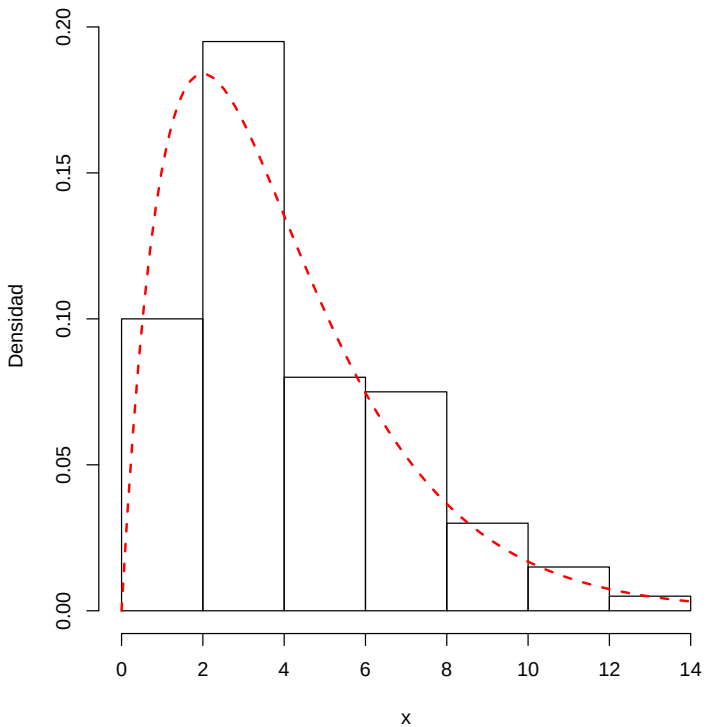


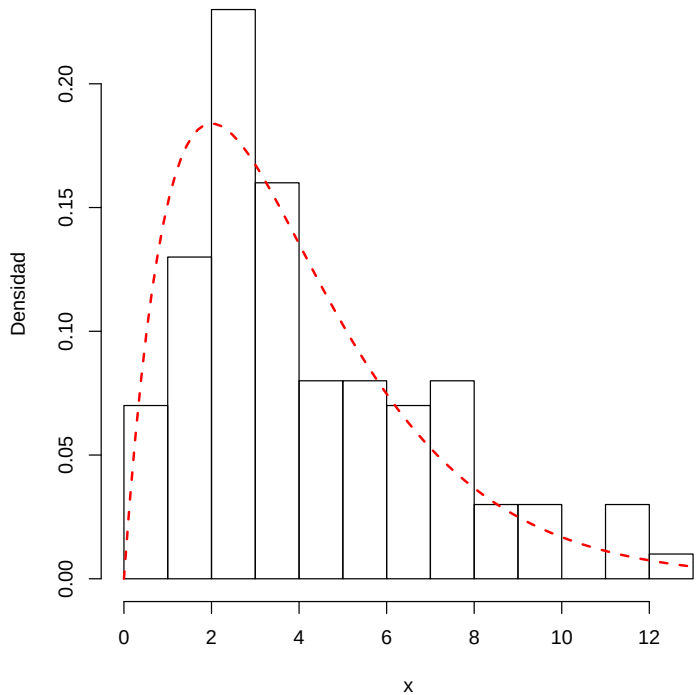
# Histograma

Reglas para escoger el número de bins:

- Raíz cuadrada:
$$k = \sqrt{n},$$
- Regla de Sturges:
$$k = 1 + 3.3 \log_{10}(n),$$
- Regla de Rice:
$$k = \lceil 2n^{1/3} \rceil,$$
- En general es posible considerar
$$k = \left\lceil \frac{\max\{x\} - \min\{x\}}{h} \right\rceil,$$
donde $h$ es el "ancho de ventana".
- Otros tipos de reglas: Doane, Scott, Freedman-Diaconis.

# Histograma

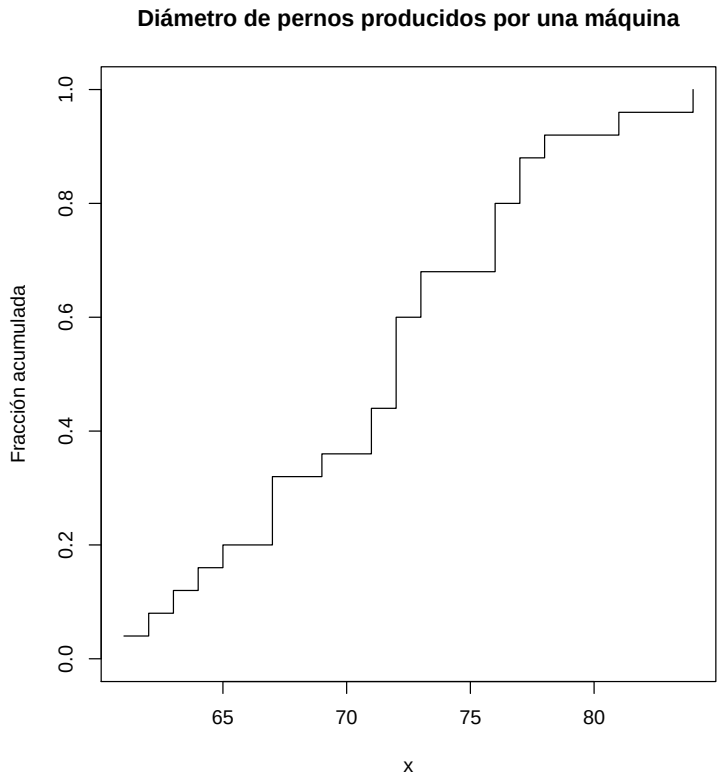

Densidad
0.00
0.05
0.10
0.15
0.20
0
2
4
6
8
10
12
14
x


Densidad
0.00
0.05
0.10
0.15
0.20
0
2
4
6
8
10
12
x

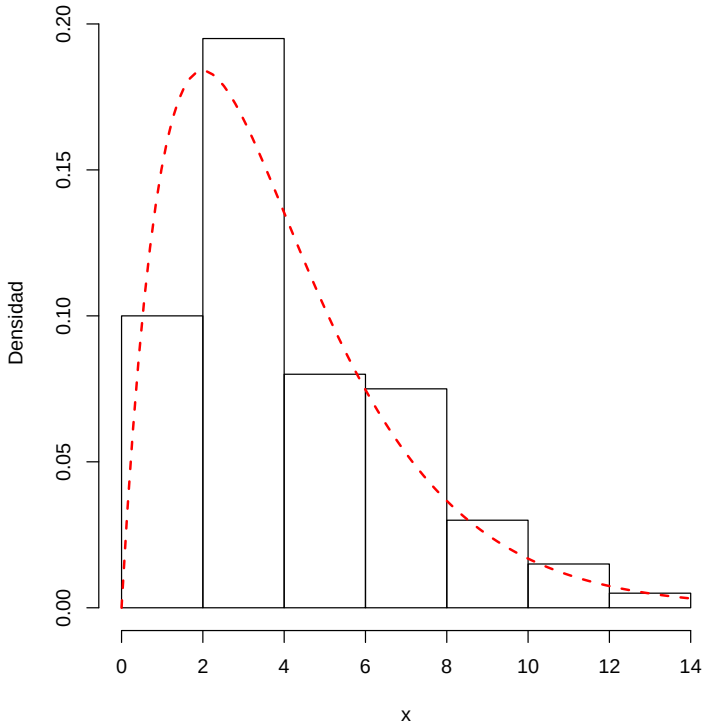
# Función de distribución acumulada empírica (ojiva)

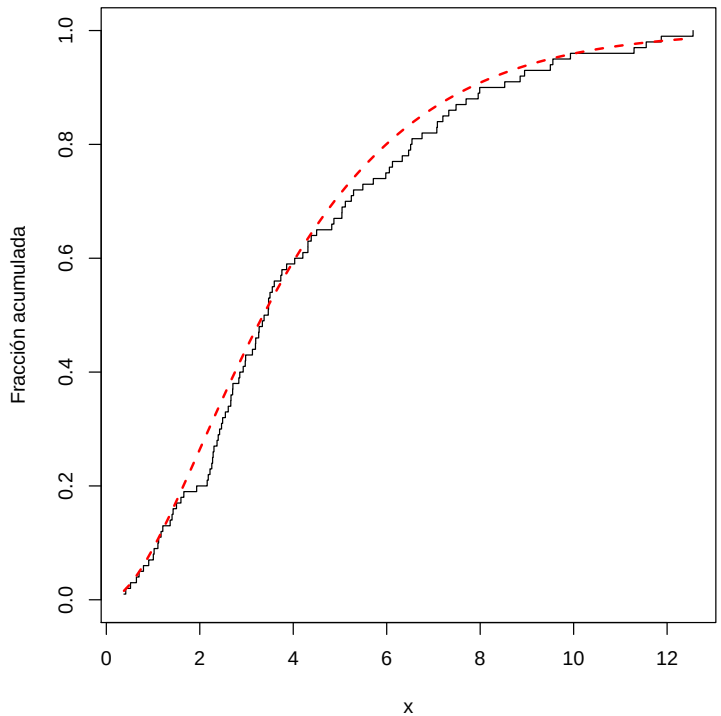
$F_n(x)$, CDF empírica, atribuye a cada valor de $x$, la fracción de datos menor o igual a $x$, es decir:

$$F_n(x) = \frac{1}{n}\{\#\text{ elementos} \leq x\}$$
$$= \frac{1}{n}\sum_{i=1}^{n} I_{\{x_i \leq x\}}.$$



# Función de distribución acumulada empírica



# Diagrama de dispersión (scatterplot)

Se utilizan cuando tenemos pares de observaciones

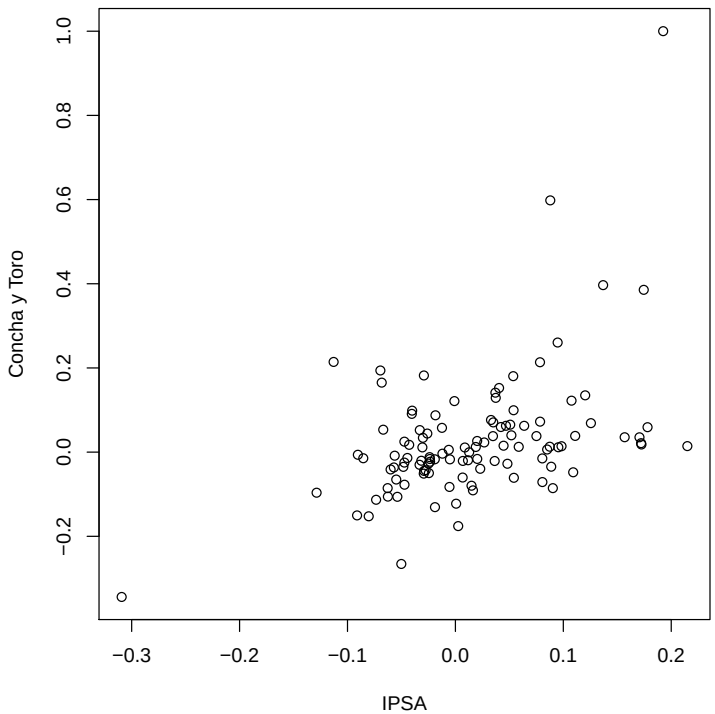
$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n),$$

que pueden ser descritos por alguna función

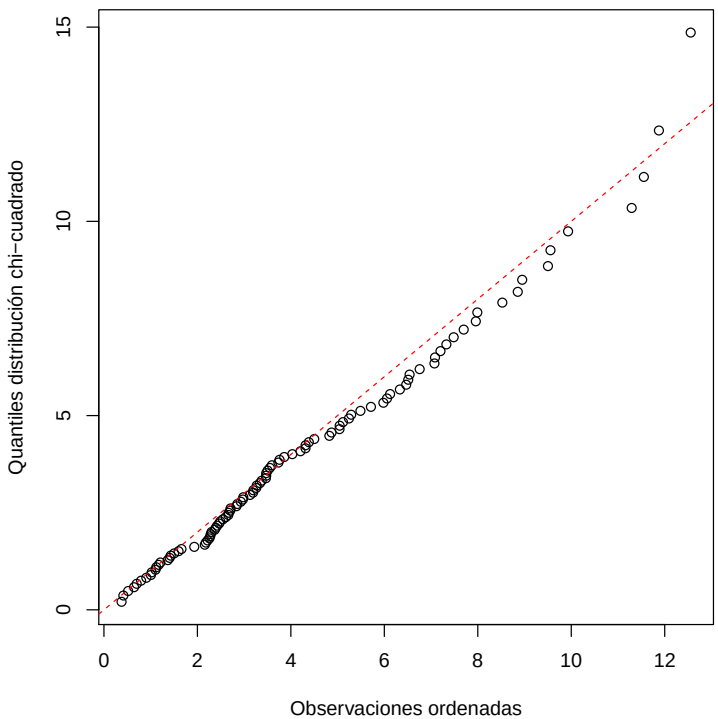
$$Y = f(x).$$

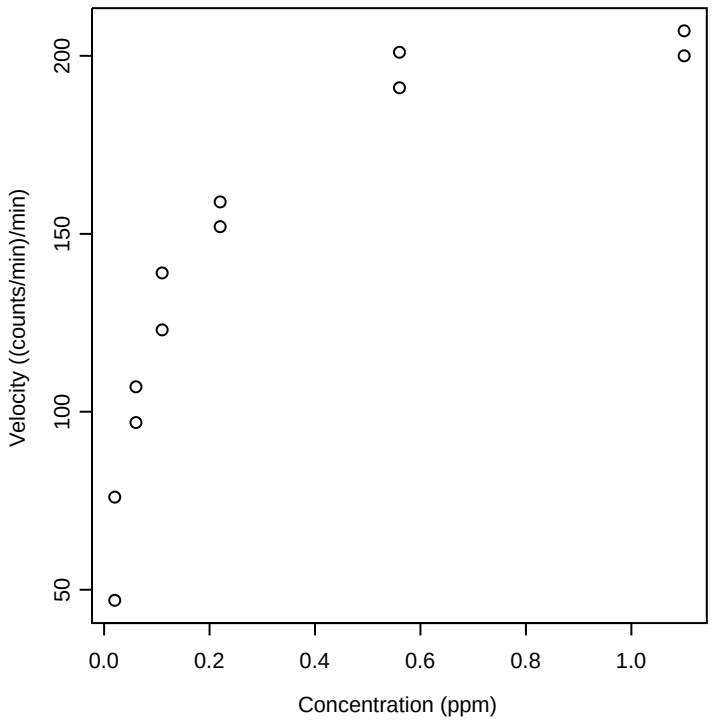
Permiten identificar:

- relaciones funcionales
- agrupaciones
- direcciones de asociación



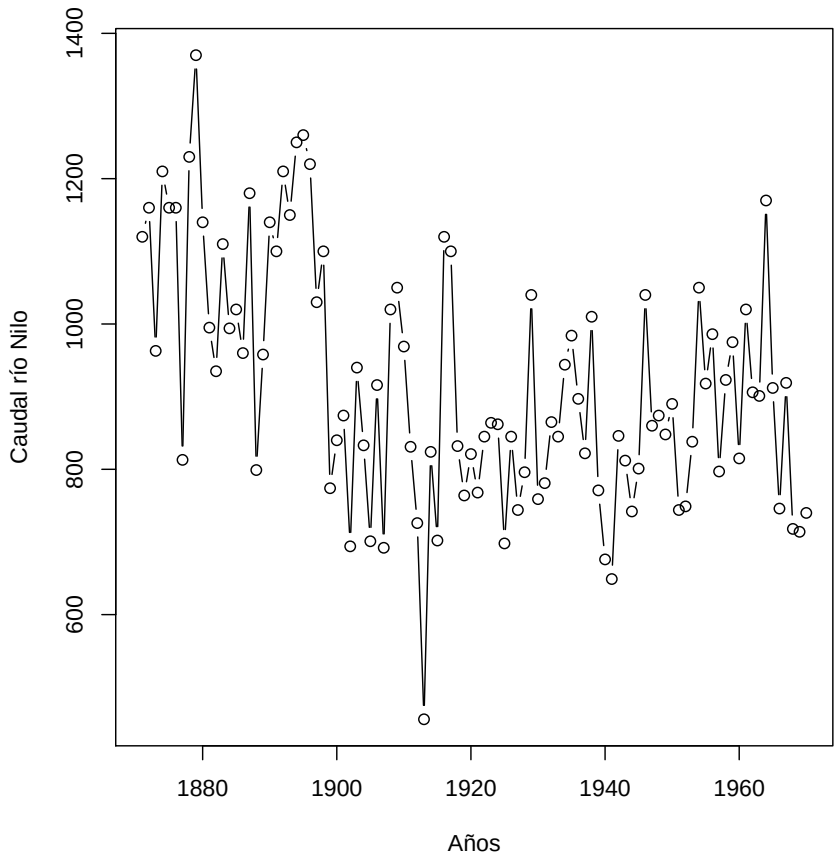
# Diagrama de dispersión (scatterplot)



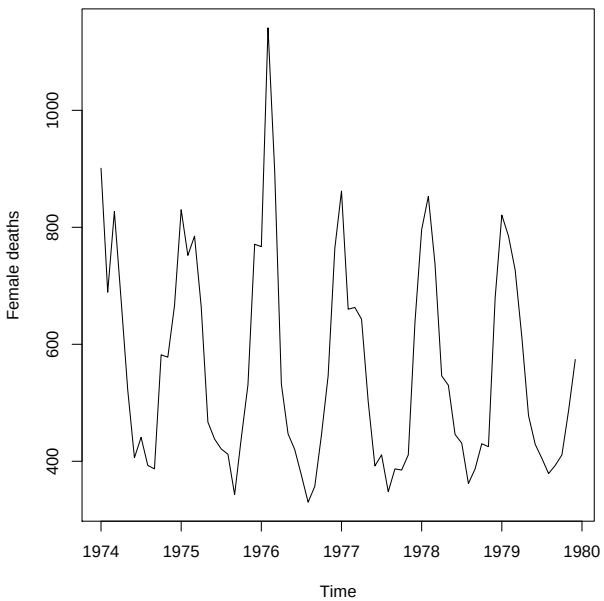



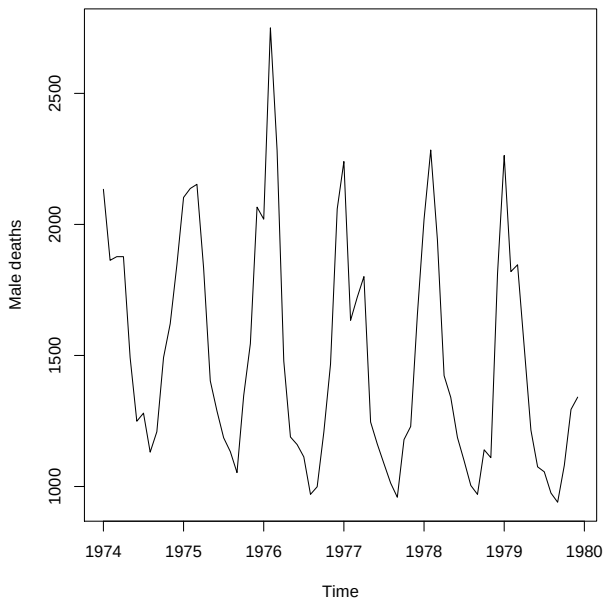
# Series de tiempo

Algunos conjuntos de datos tienen un ordenamiento natural, por ejemplo el tiempo.



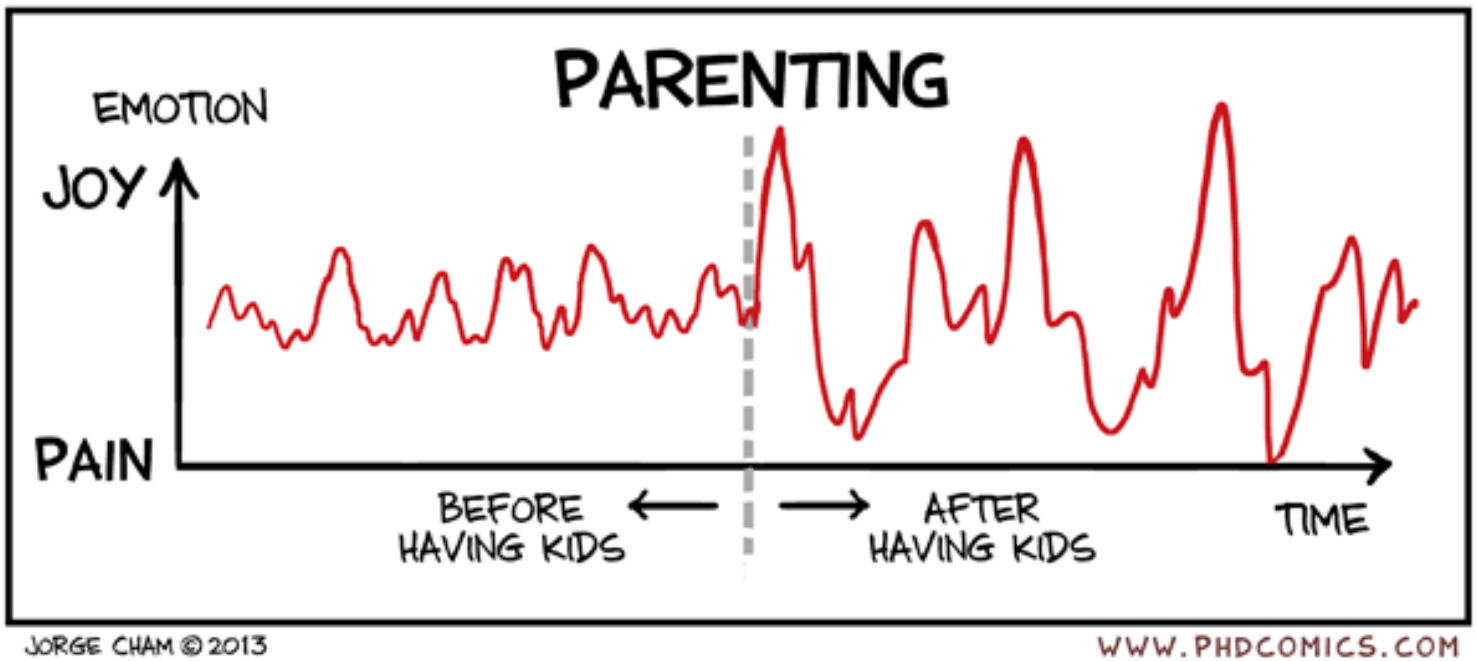
# Series de tiempo




Muertes por bronquitis, enfisema y asma en UK, 1974-1979.

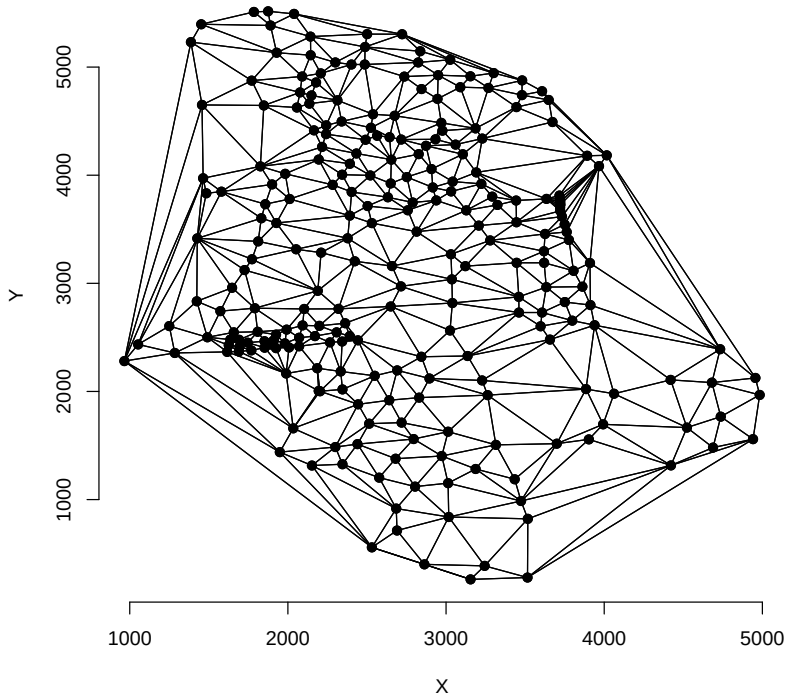
# Series de tiempo



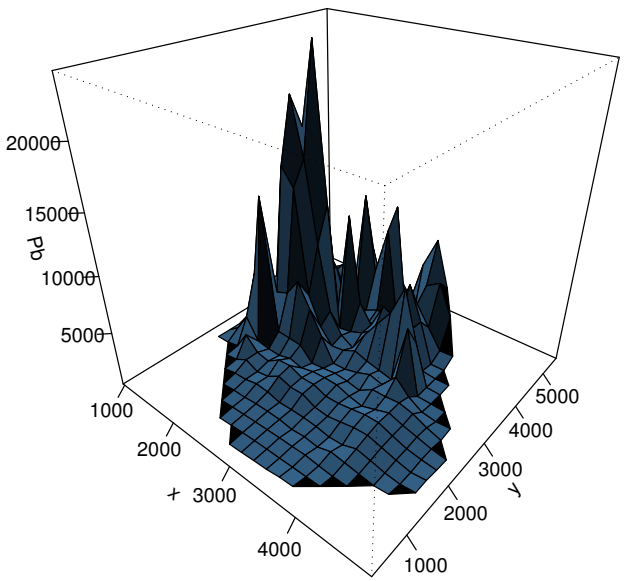

Análisis de puntos de cambio (Chen y Gupta, 2011).

# Otros tipos de gráficos

Coordenadas de las muestras de suelo

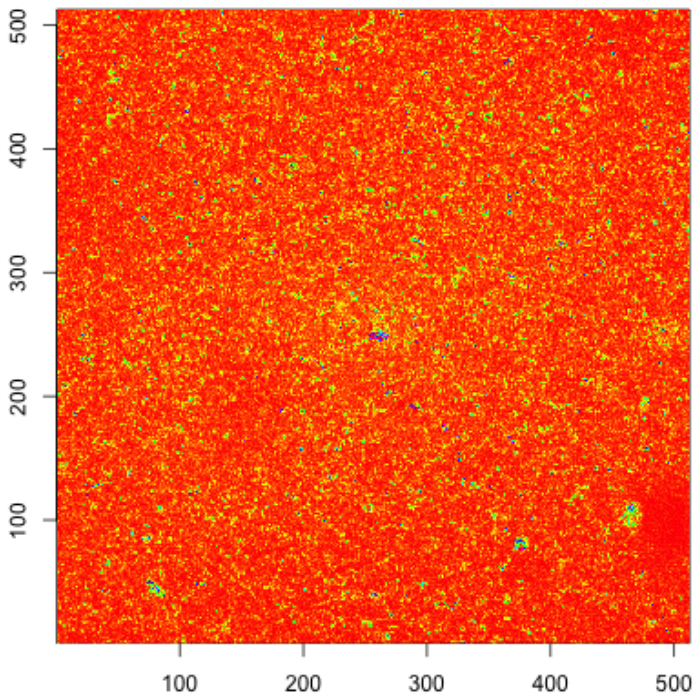



Concentración de Plomo

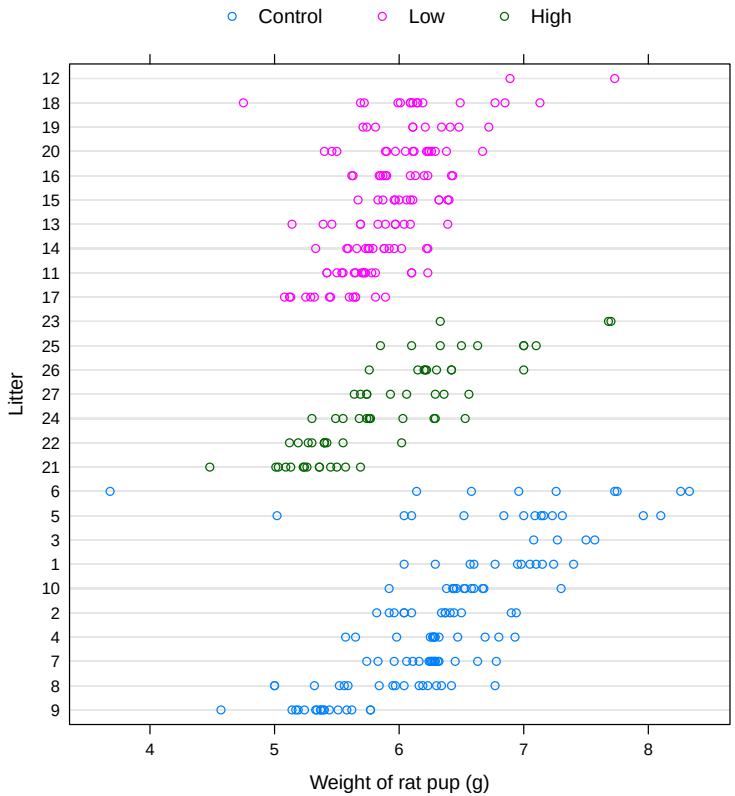


# Otros tipos de gráficos

Flamabilidad de nanotubos

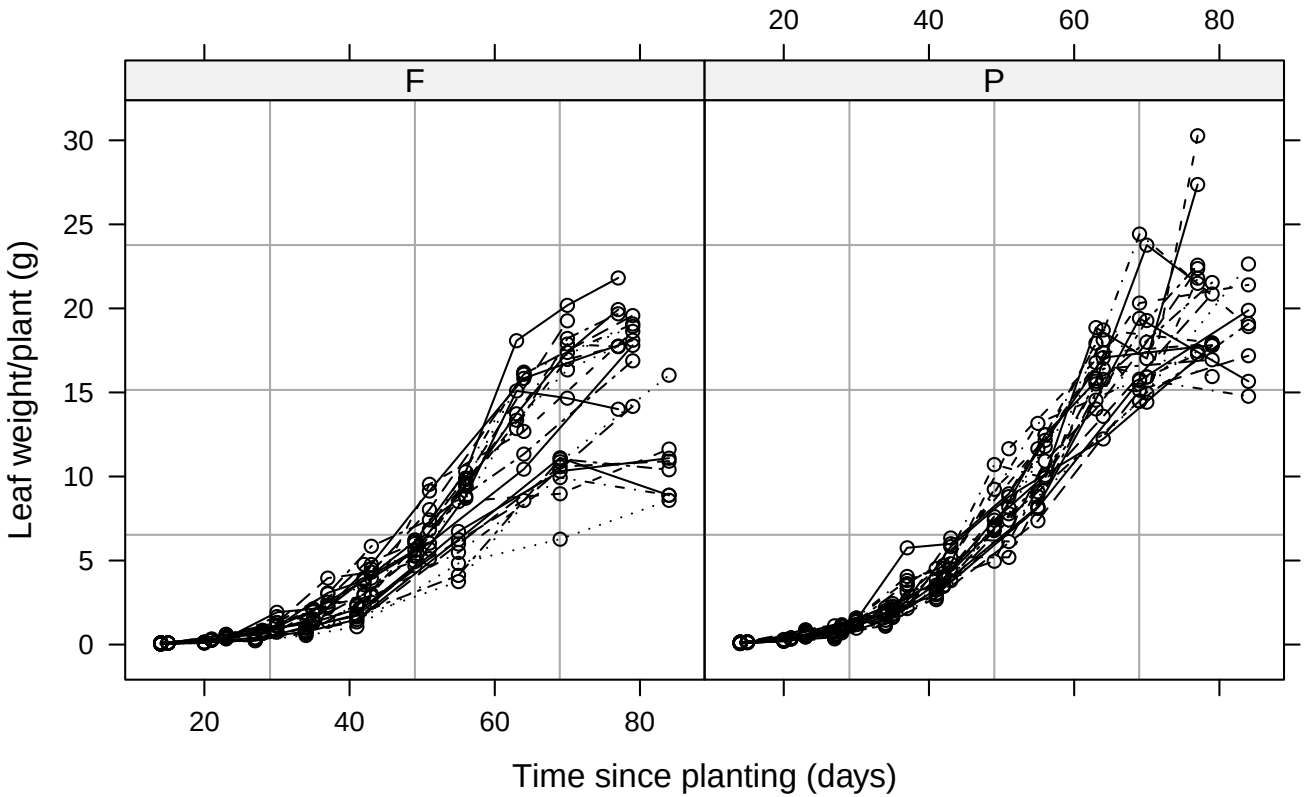



Estudio reproductivo con roedores



# Otros tipos de gráficos

Comparación del patrón de crecimiento de dos genotipos de plantas de soya (Davidian y Giltinan, 1995). Variedad comercial (F), Variedad experimental (P).



# Referencias adicionales sobre gráficos en Estadística



Chambers, J.M., Cleveland, W.S., Kleiner, B., and Tukey, P.A. (1983).
Graphical Methods for Data Analysis.
Wadsworth & Brooks/Cole.


Cleveland, W.S. (1993).
Visualizing Data.
Hobart Press, Summit.


Murrell, P. (2005).
R Graphics.
Chapman & Hall/CRC Press.


Sarkar, D. (2008).
Lattice: Multivariate Data Visualization with R.
Springer. URL: `http://lmdvr.r-forge.r-project.org`


Wilkinson, L. (2005).
The Grammar of Graphics, 2nd edition.
Springer.


# Datos del SIMCE

Desde 1988 el SIMCE evalua 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 (4º y 8º básico, 2º medio).[1]
- 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).

[1] Conocido como: datos con estructura longitudinal.

# Datos del SIMCE[2]

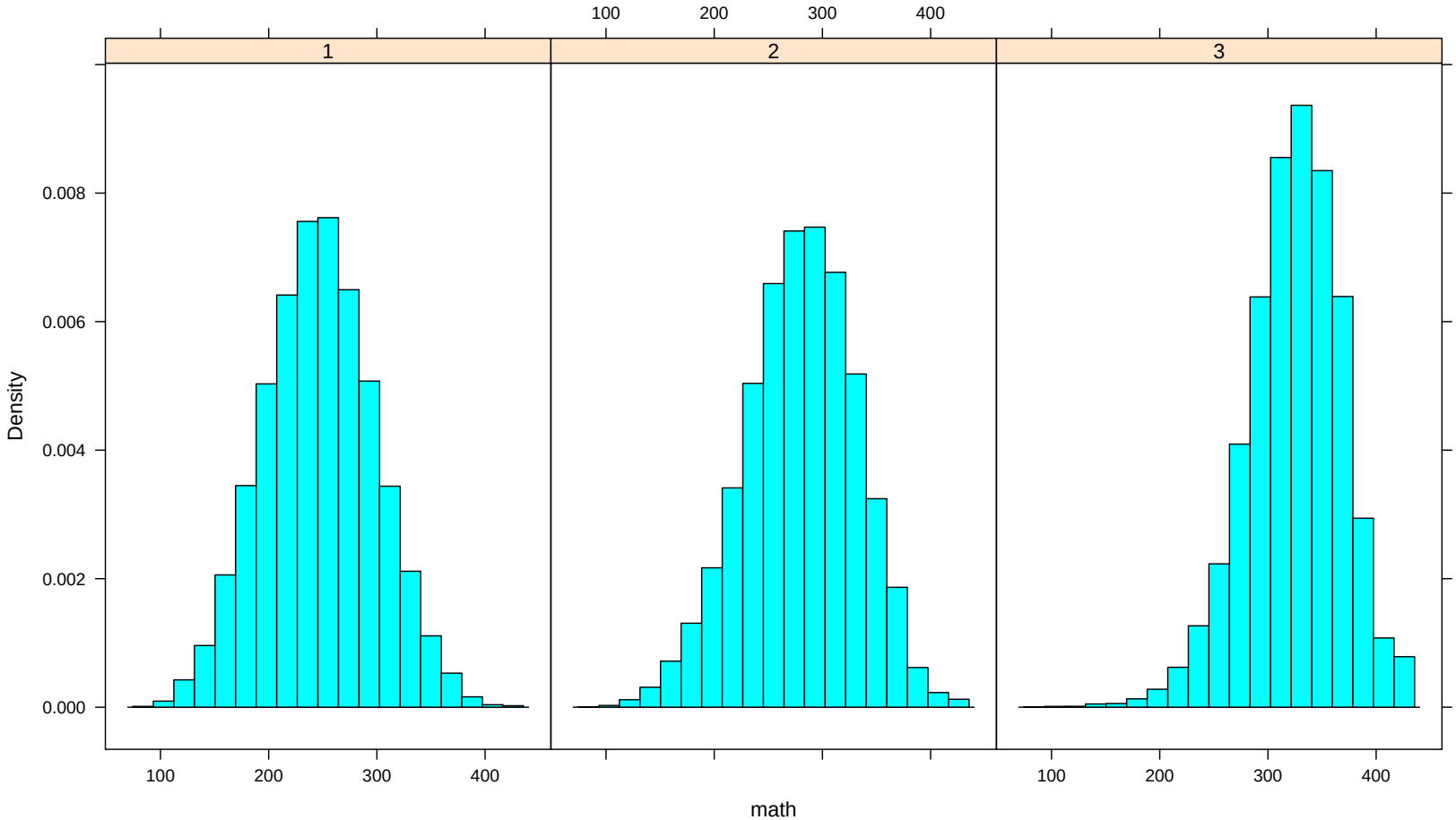



Figure: histograma puntajes matemática.

[2] colegios, 1: municipales, 2: subvencionados y 3: particules.

# Datos del SIMCE

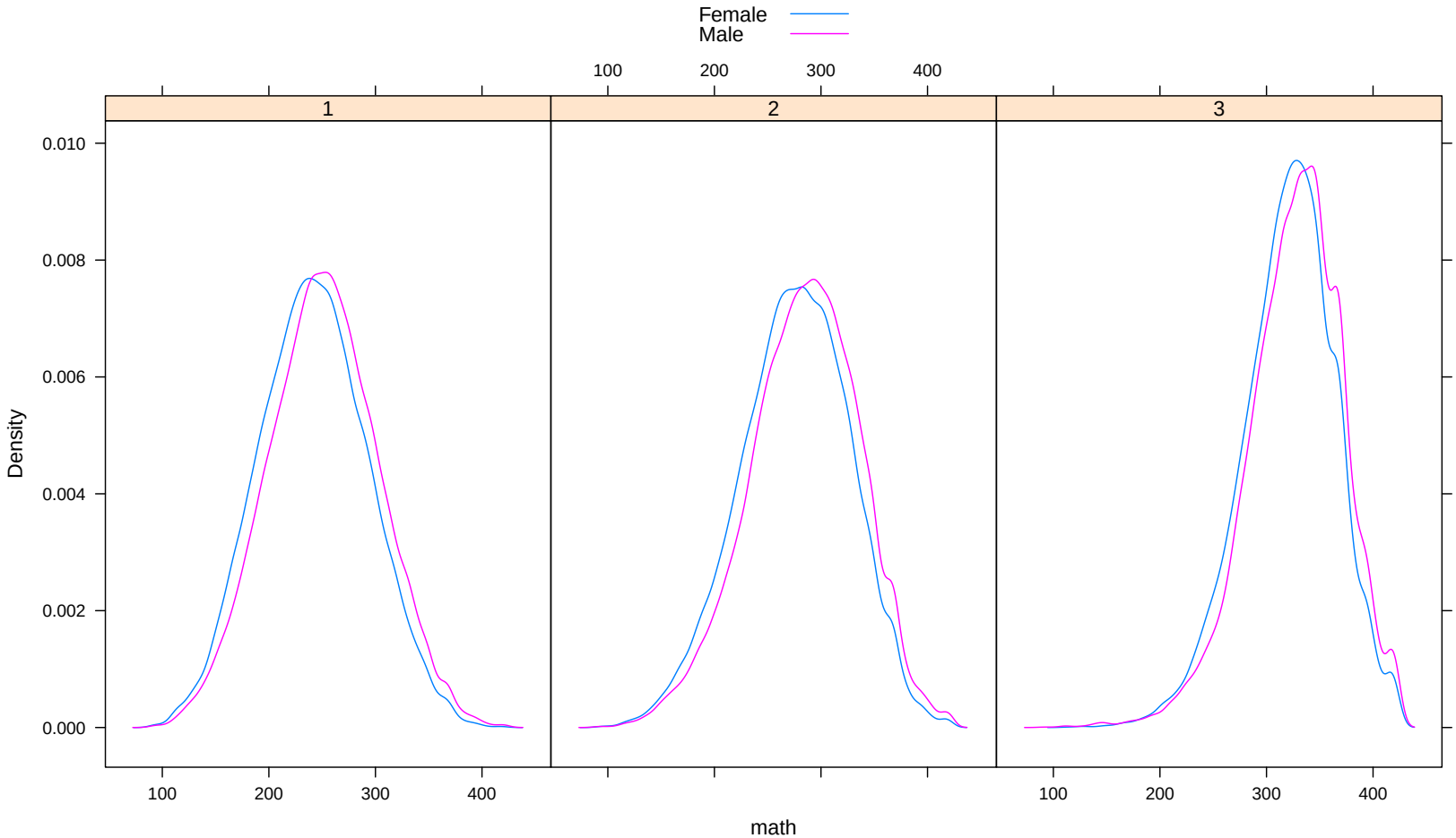



Figure: densidad puntajes matemática, organizados por Sexo.

# Datos del SIMCE

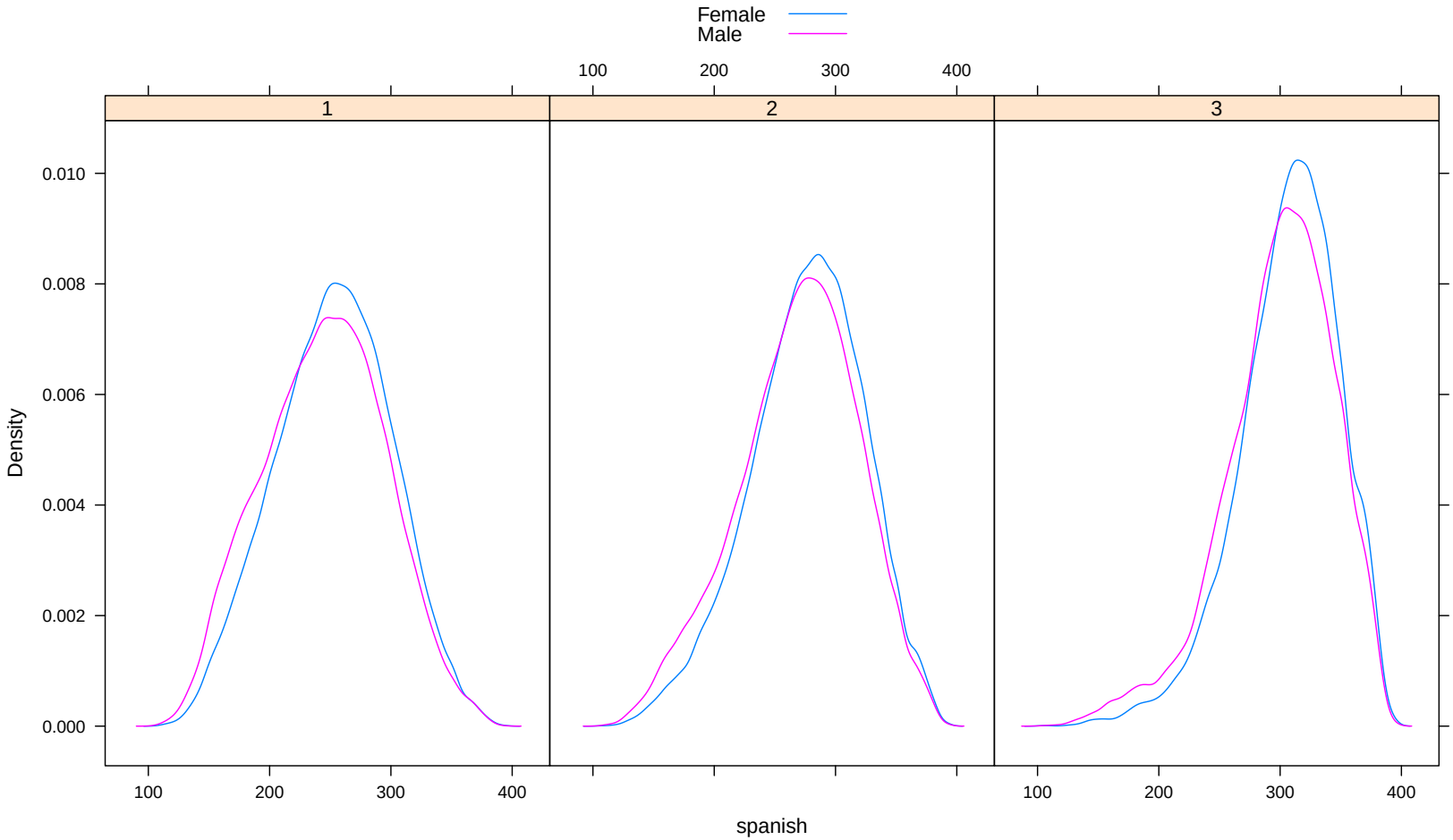



Figure: densidad puntajes lenguaje, organizados por Sexo.

# 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
```

# Datos del SIMCE

*Para pensar:*

- ¿Cómo resumir la información del total de 133K datos para cada una de las 8 variables?[3]

- ¿Podemos usar, digamos unas pocas cantidades para describir esta información?

[3] Es decir un poco más de 1 millón de registros.

# Estadísticas de resumen

## Ingredientes:

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

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

En esta clase consideraremos 3 tipos de estadísticas de resumen[4] para una muestra $\boldsymbol{x} = (x_1, \ldots, x_n)^\top$,

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

[4] En ocasiones escribiremos $T = T(\boldsymbol{x})$.

**Definición 3 (Media muestral o promedio):**

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

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

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

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

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

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

*Ejemplo:*

Considere el conjunto de datos $\boldsymbol{x} = \{1, 2, 2, 2, 3, 3, 8\}$. 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 $\overline{x} = 21/7 = 3$. 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, $\overline{x} = 30/76 = 0.3947$ es el número promedio de accidentes.

### Definición 4 (Estadísticos de orden):

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

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

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

### Definición 5 (Mediana):

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

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

*Observación:*

Sea $f(x)$ cualquier función de números reales.[5] Entonces podemos definir

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

*Caso particular (media geométrica):*

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

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

Es decir,

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

[5] 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:

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

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 | |

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

$$\overline{x}_1 = \frac{1}{3}(10 + 20 + 30) = \frac{60}{3} = 20,$$
$$\overline{x}_2 = \frac{1}{5}(2 \cdot 5 + 20 + 2 \cdot 35) = \frac{100}{5} = 20,$$
$$\overline{x}_3 = \frac{3 \cdot 20}{3} = 20.$$

Además, $\text{me}_j = 20$ para todo $j$.

*Observación:*

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

# Medidas de dispersión

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 $x_{(1)}, \ldots, x_{(n)}$ 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 $\boldsymbol{x} = \{4, 7, 18, 1, 7, 13, 2\}$ 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

$$Q_1 = P_{25} = x_{(25\cdot(7+1)/100)} = x_{(1\cdot 8/4)} = x_{(2)} = 2,$$
$$Q_3 = P_{75} = x_{(75\cdot(7+1)/100)} = x_{(3\cdot 8/4)} = x_{(6)} = 13.$$

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

**Definición 6 (Varianza muestral):**

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

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

*Observación:*

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

*Observación:*

Otras medidas de dispersión:

- Desviación absoluta en torno de la media:

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

- Desviación absoluta en torno de la mediana:

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

- $r$-ésimo momento centrado en torno de $a$:[6]

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

[6] Para $r=2$ y $a=\overline{x}$ obtenemos la varianza.

# Propiedades del promedio y varianza muestrales

**Propiedades:**

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

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

(c) $\overline{x}$ es el valor que minimiza la función:
$$S(a) = \sum_{i=1}^{n}(x_i - a)^2.$$

(d) Sea $x_1, \ldots, x_n$ y considere la transformación:
$$y_i = a\, x_i + b, \qquad i = 1, \ldots, n.$$
Entonces $\overline{y} = a\,\overline{x} + b$ y $s_y^2 = a^2 s_x^2$.

# Propiedades del promedio y varianza muestrales

(a) En efecto,

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

(b) (Fórmula de Köning)

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

# Propiedades del promedio y varianza muestrales

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

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

resolviendo la condición de primer orden, tenemos

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

desde donde sigue que $\widehat{a} = \overline{x}$. Además

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

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

# Propiedades del promedio y varianza muestrales

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

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\sum_{i=1}^{n}(ax_i + b) = \frac{1}{n}\Big(a\sum_{i=1}^{n} x_i + b\Big)$$

$$= a\Big(\frac{1}{n}\sum_{i=1}^{n} x_i\Big) + b = a\overline{x} + b.$$

Mientras que

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

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

$$s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2 = \frac{1}{n-1}\sum_{i=1}^{n}\{a(x_i - \overline{x})\}^2$$

$$= a^2\,\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 = a^2 s_x^2.$$

*Observación:*

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

$$z_i = \frac{x_i - \overline{x}}{s}, \qquad i = 1, \ldots, n.$$

Entonces,[7]

$$\overline{z} = 0 \qquad \text{y} \qquad s_z^2 = 1.$$

[7] Basta hacer $a = 1/s$ y $b = \overline{x}/s$ 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

$$\mathrm{CV} = s/\overline{x}, \qquad \overline{x} \neq 0.$$

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

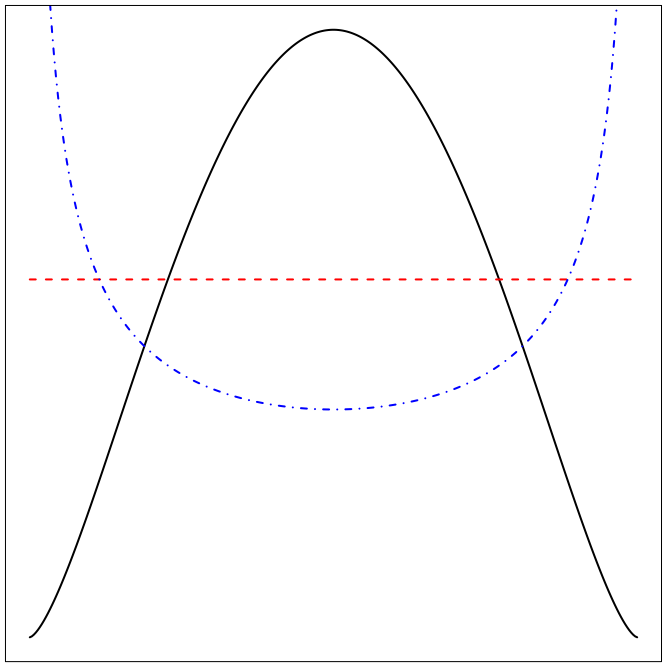
*Observación:*

Un valor pequeño para el $\mathrm{CV}$ está asociado a una muestra homogénea.

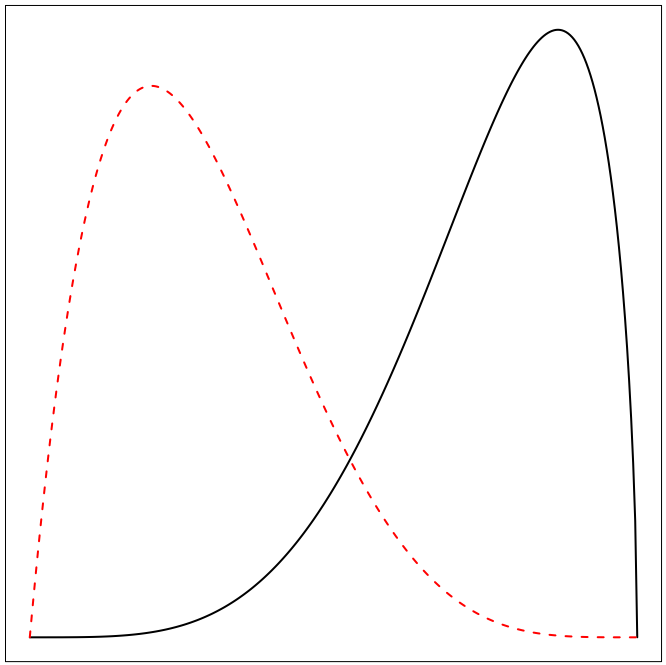
*Observación:*

En Econometría $1/\mathrm{CV}$ es conocido como la razón de Sharpe.

# Medidas de forma



(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:

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

*Observación:*

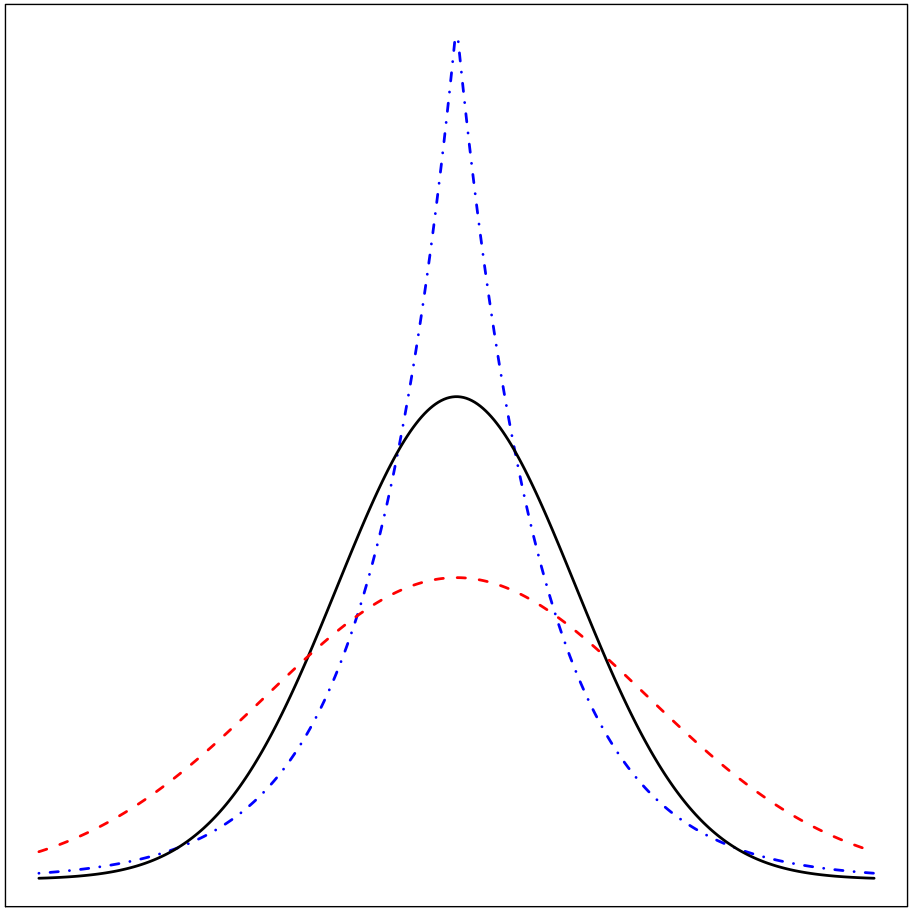
- Si $b_1 = 0$ la distribución es simétrica con relación a $\overline{x}$.
- Si $b_1 > 0$ la distribución tiene sesgo positivo. En caso contrario, decimos que tiene sesgo negativo.

*Observación:*

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

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

# Medidas de forma



(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[8] como:

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

*Observación:*

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

[8] también conocido como exceso de curtosis

# Datos del SIMCE: Puntajes de matemáticas[9]

```
> 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
```

[9] $n = 132\,793$ observaciones, así que $(n-1)/n = 0.9999925$.

# Gráfico de cajón con bigotes (boxplot)