MAT-042: Conceptos básicos

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)

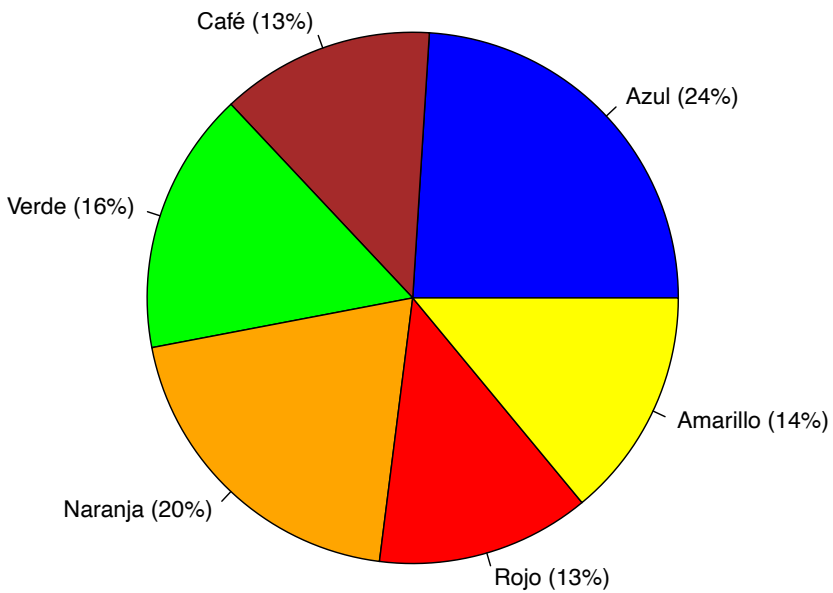


Gráfico Circular

Ejemplo: Proporción de población anglófona en el mundo

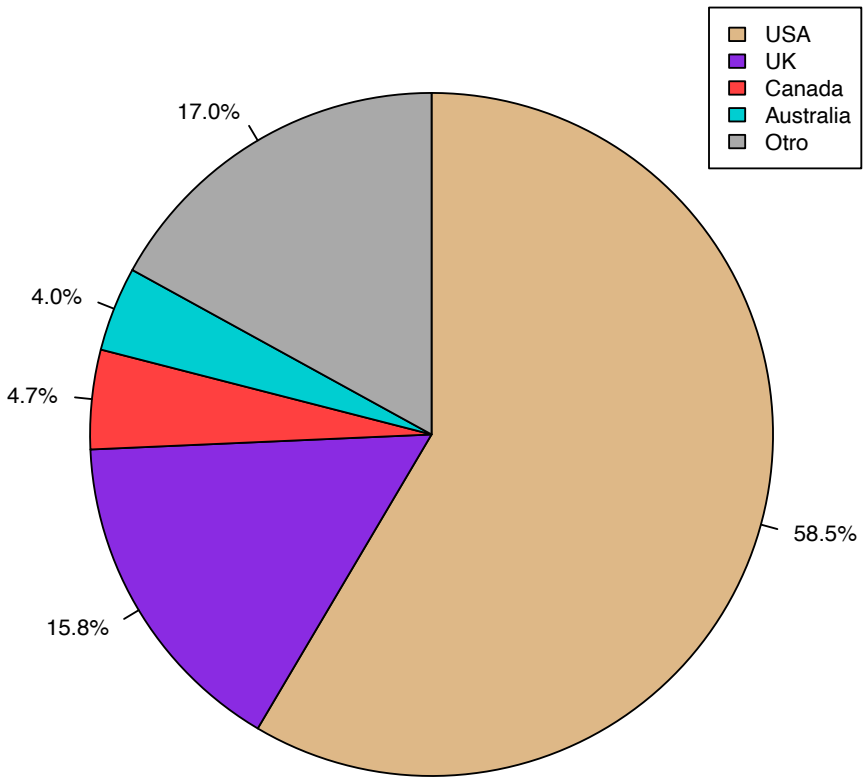


Gráfico Circular

Limitación: Gráfico circular de un arcoiris.

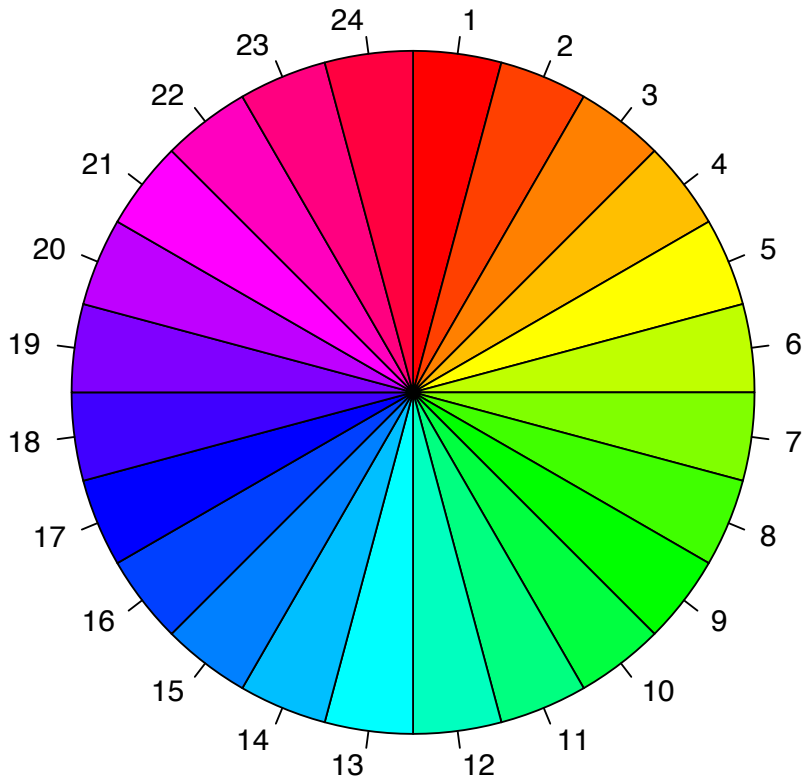


Gráfico Circular

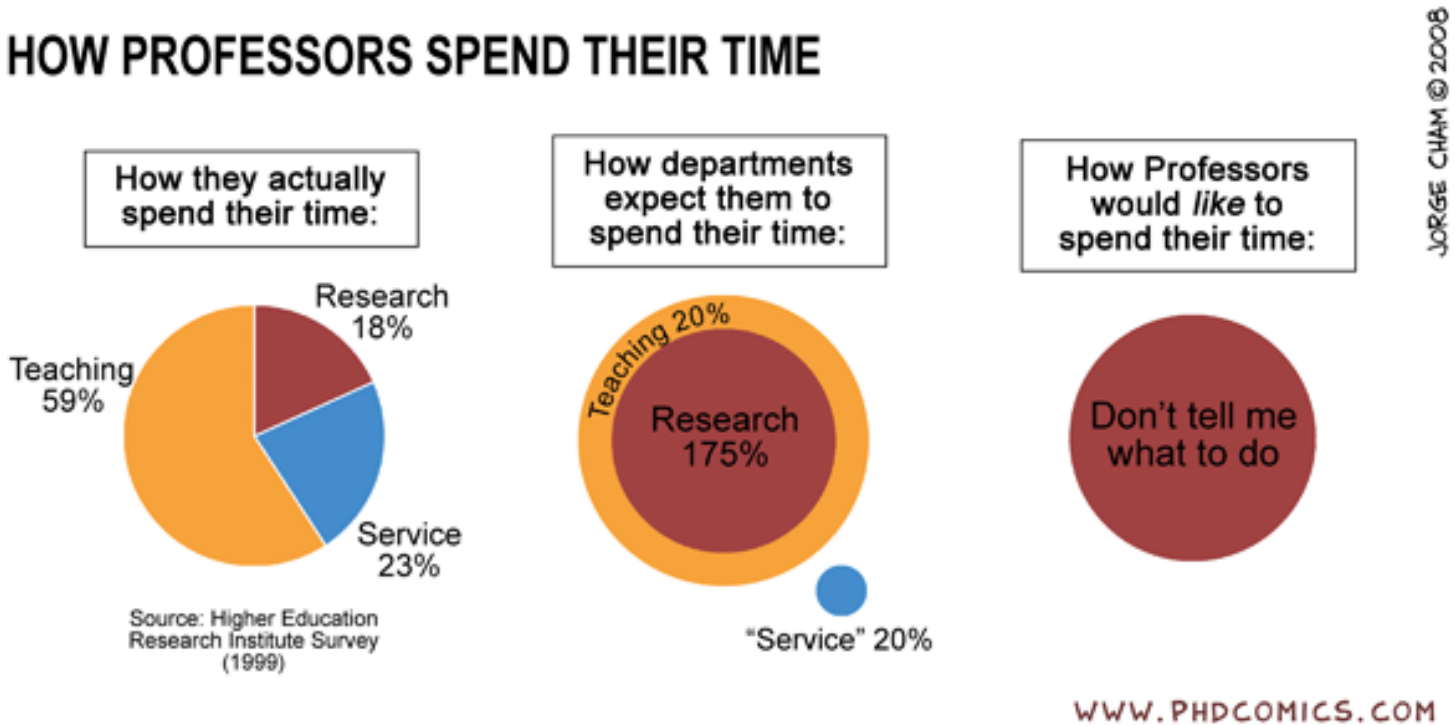


Gráfico de Barras

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.

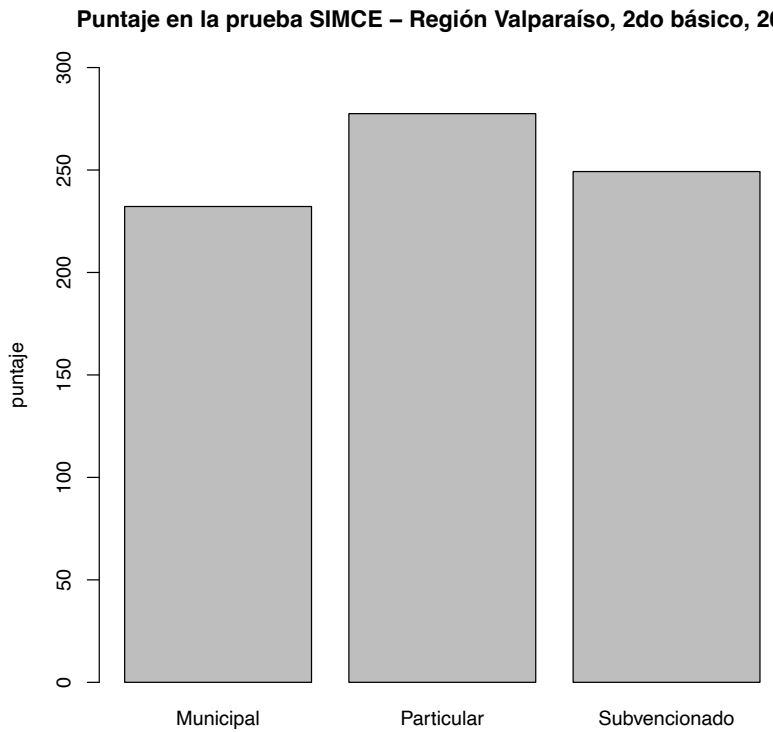
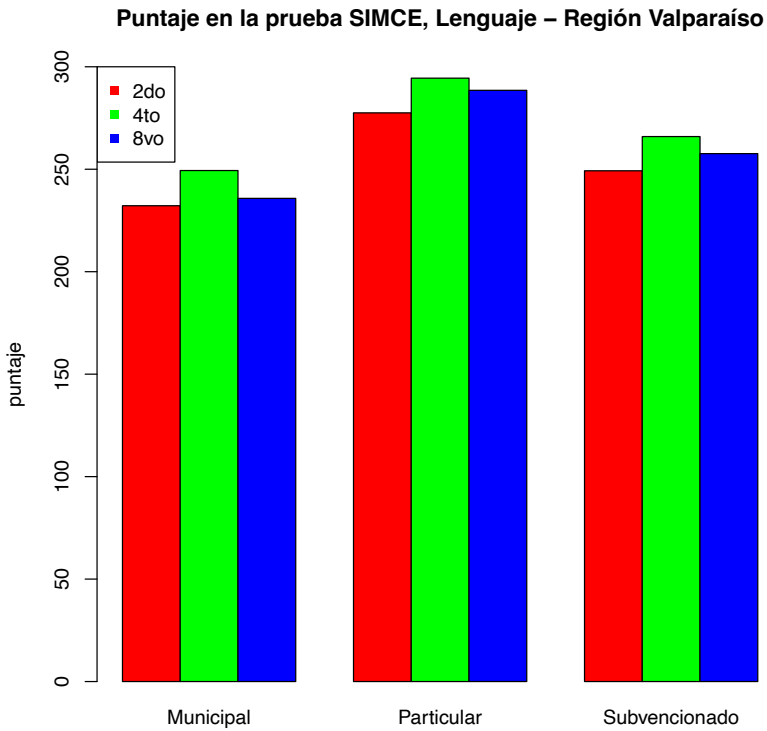


Gráfico de Barras



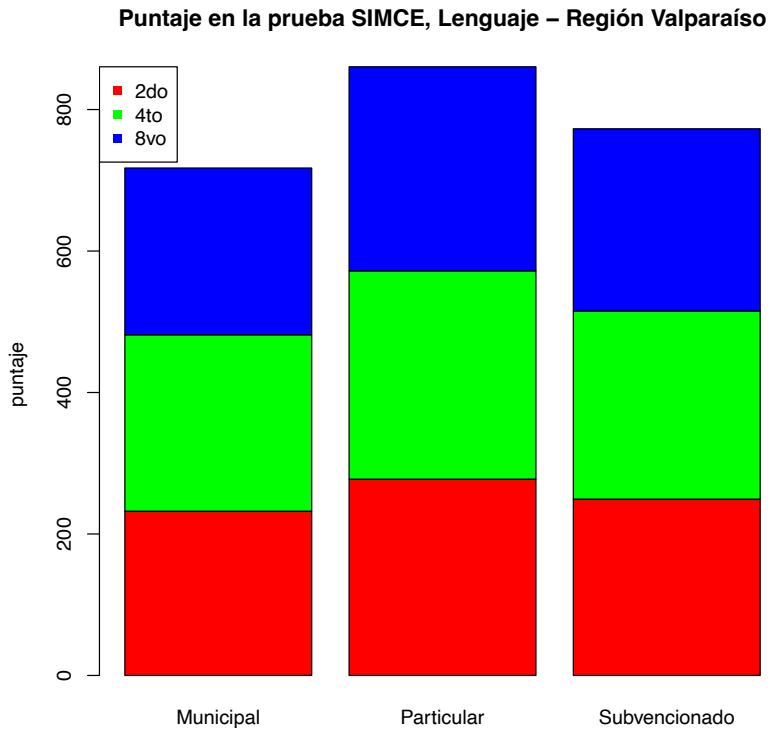
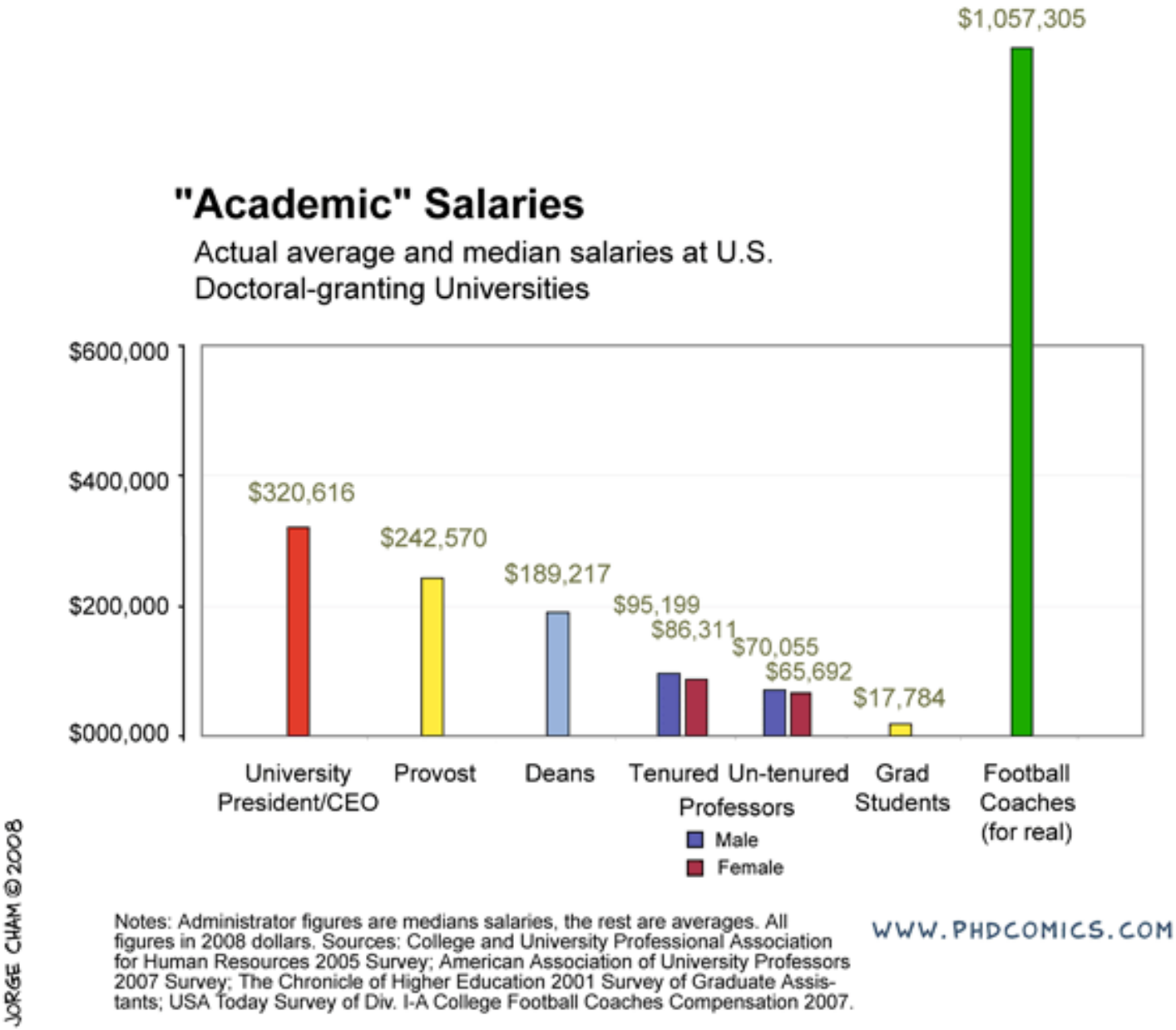


Gráfico de Barras



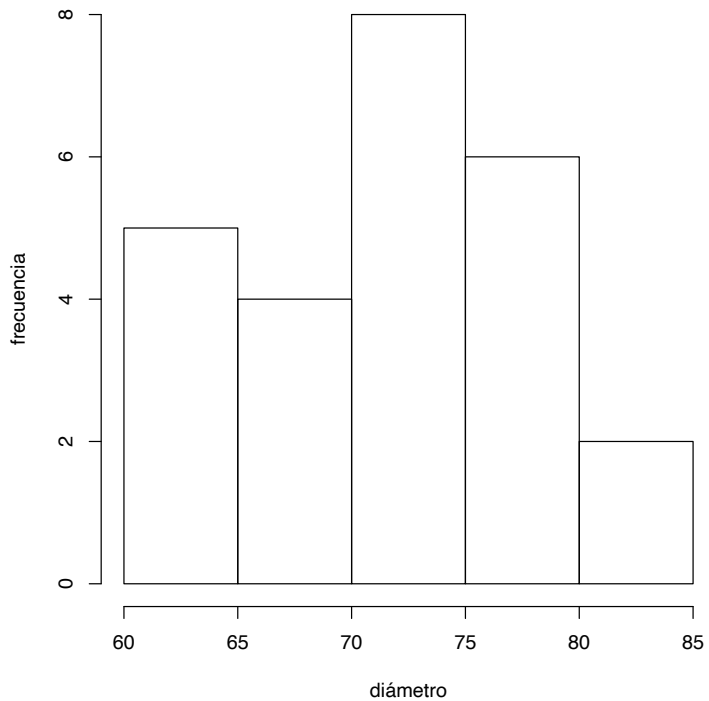
Histograma

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			

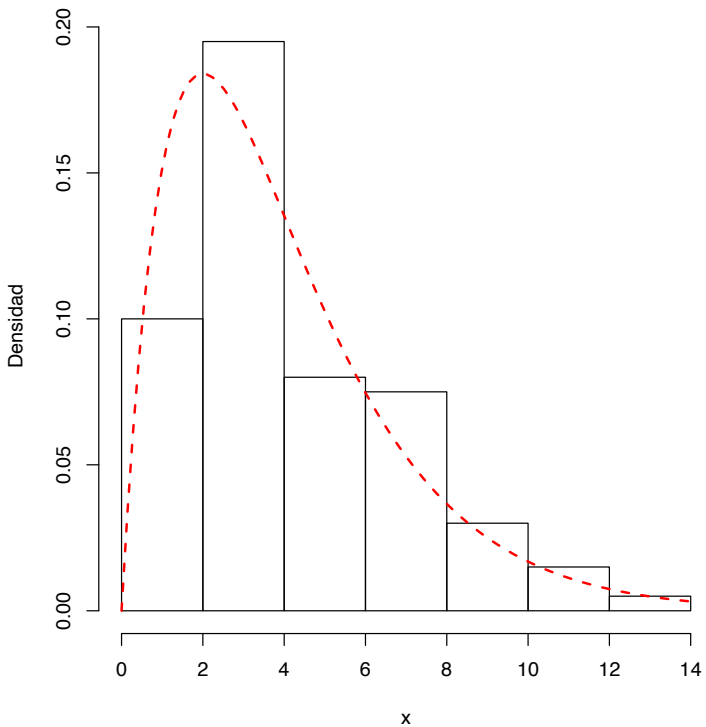


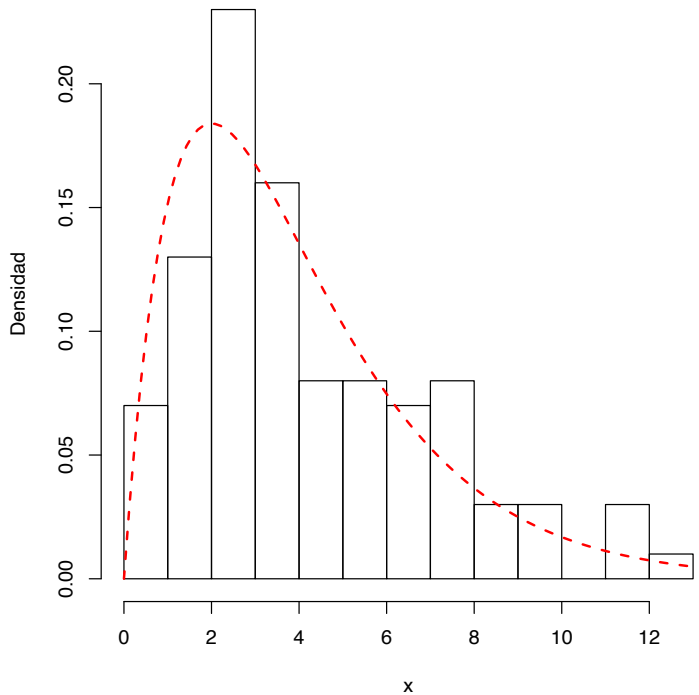
Histograma

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

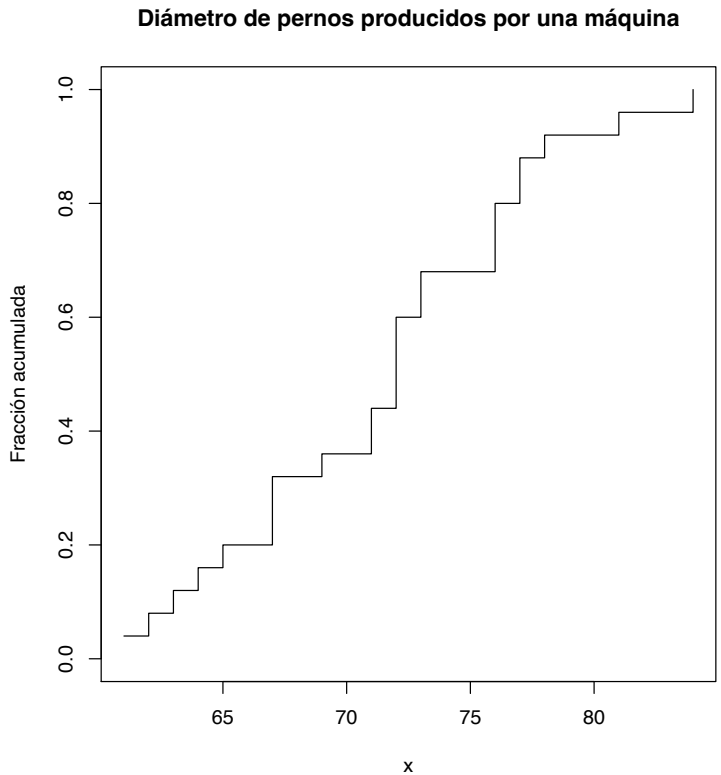

Densidad
0.00
0.05
0.10
0.15
0.20
0
2
4
6
8
10
12
14
x


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

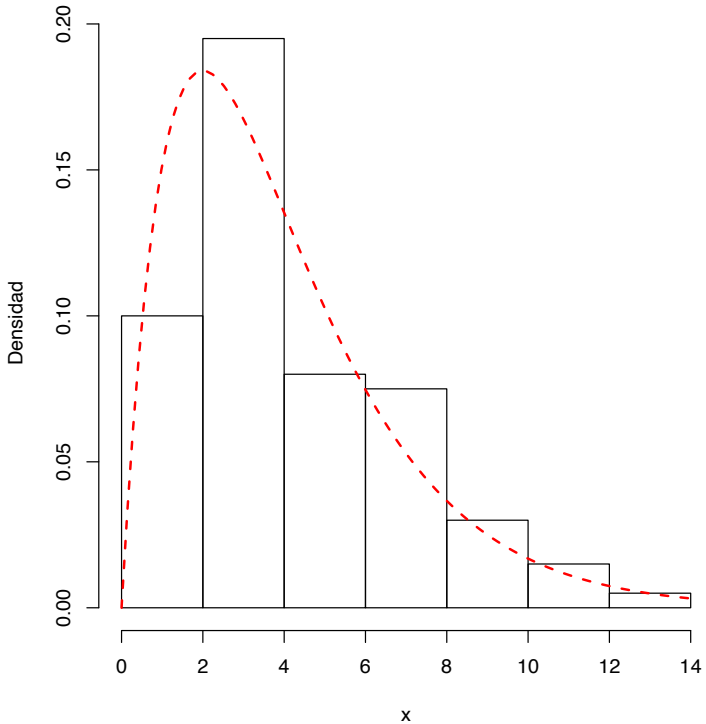
Función de distribución acumulada empírica (ojiva)

$F_n(x)$, 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



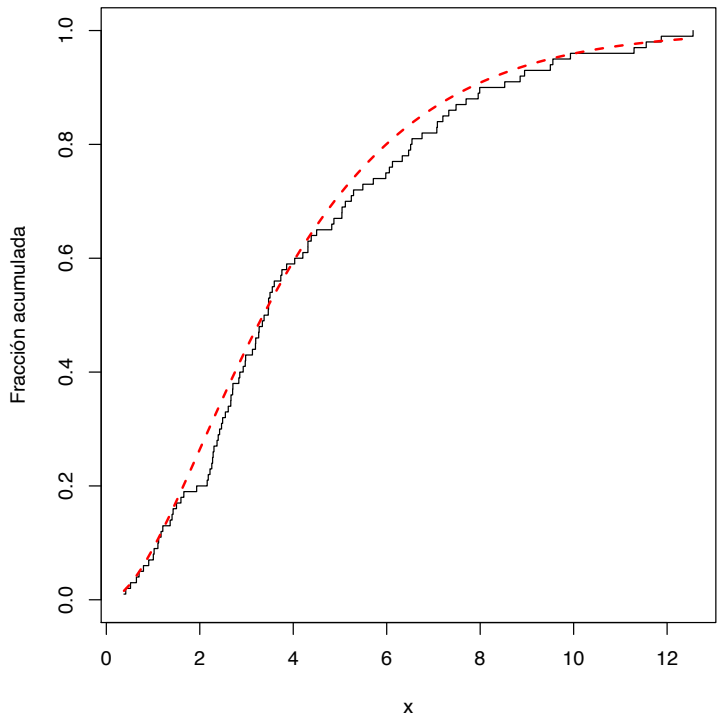


Diagrama de dispersión (scatterplot)

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

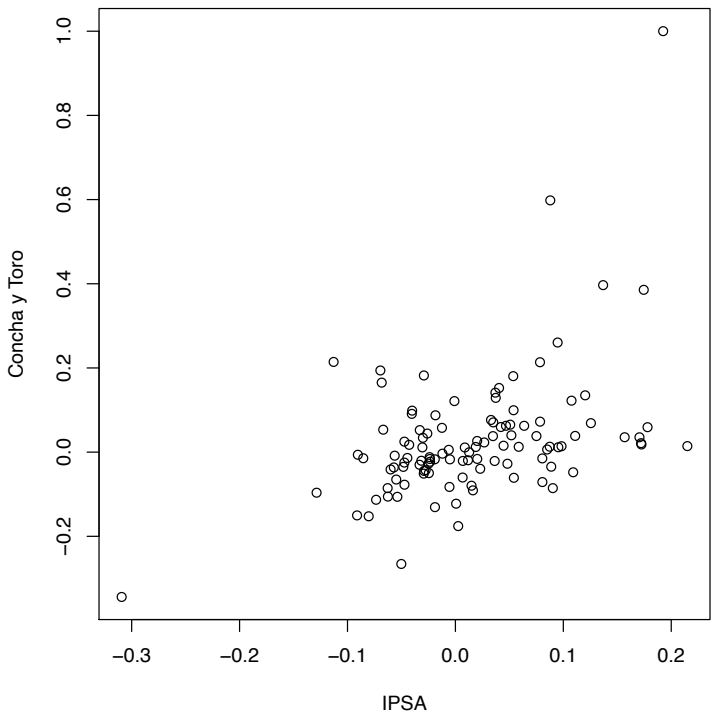
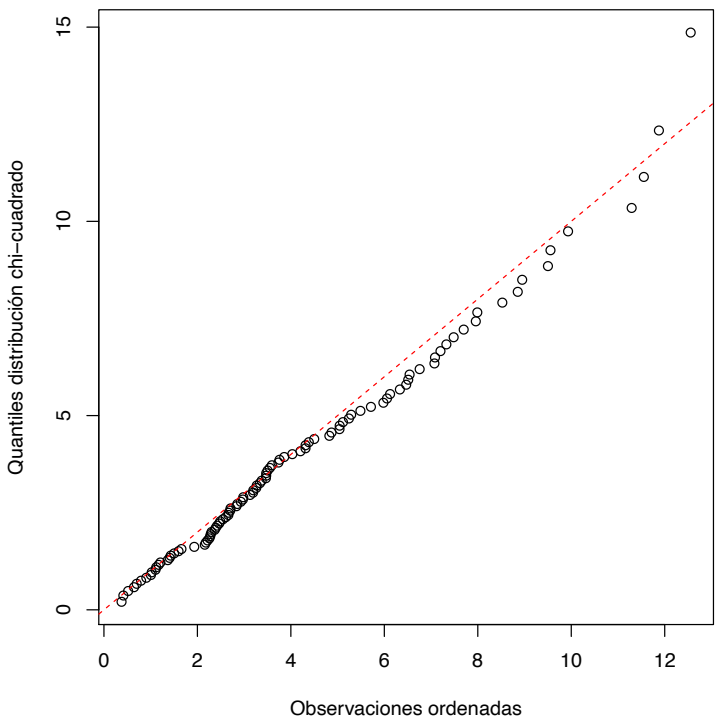
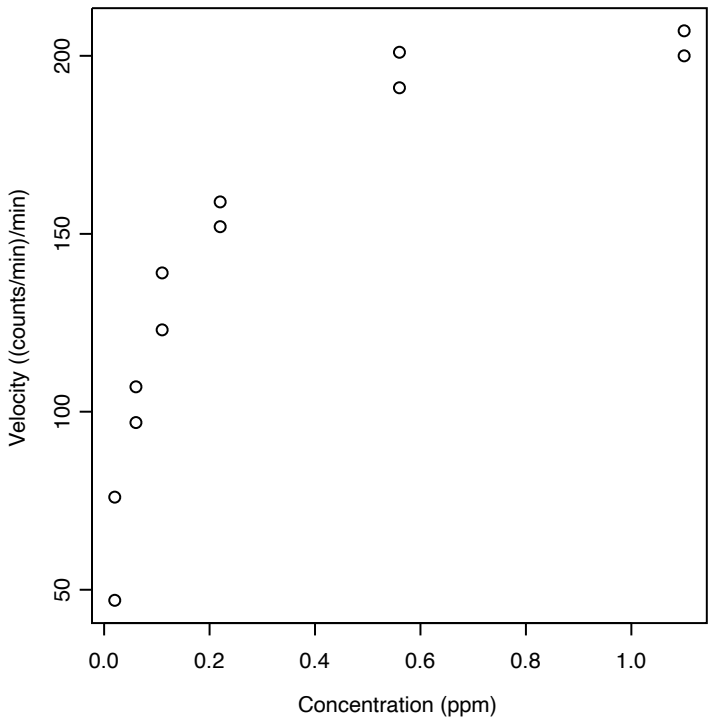


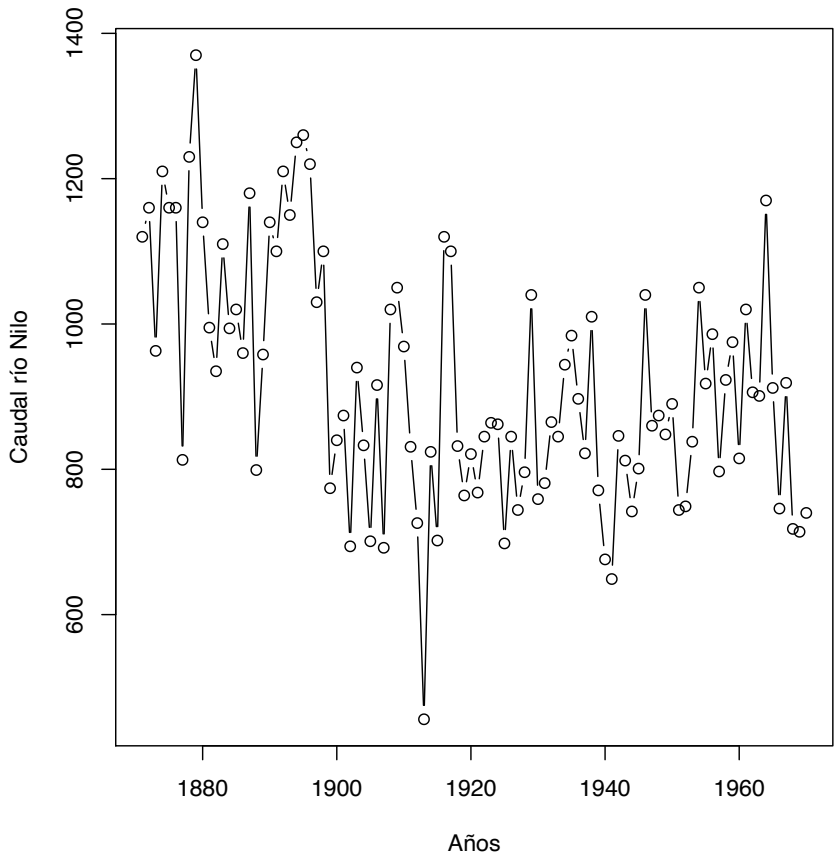
Diagrama de dispersión (scatterplot)



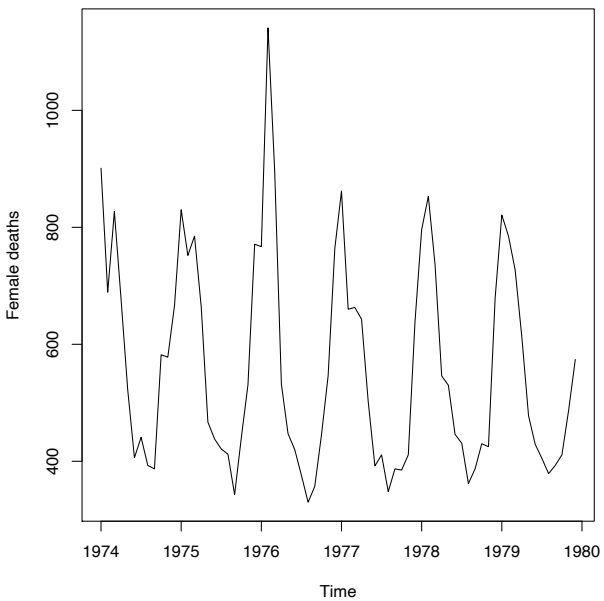


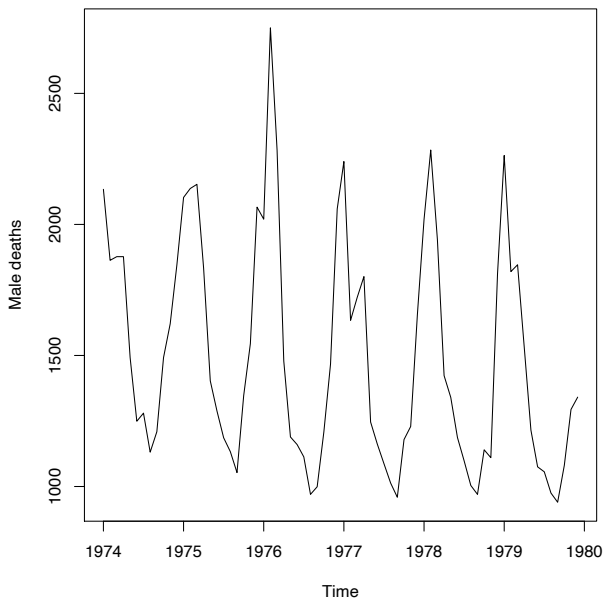
Series de tiempo

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



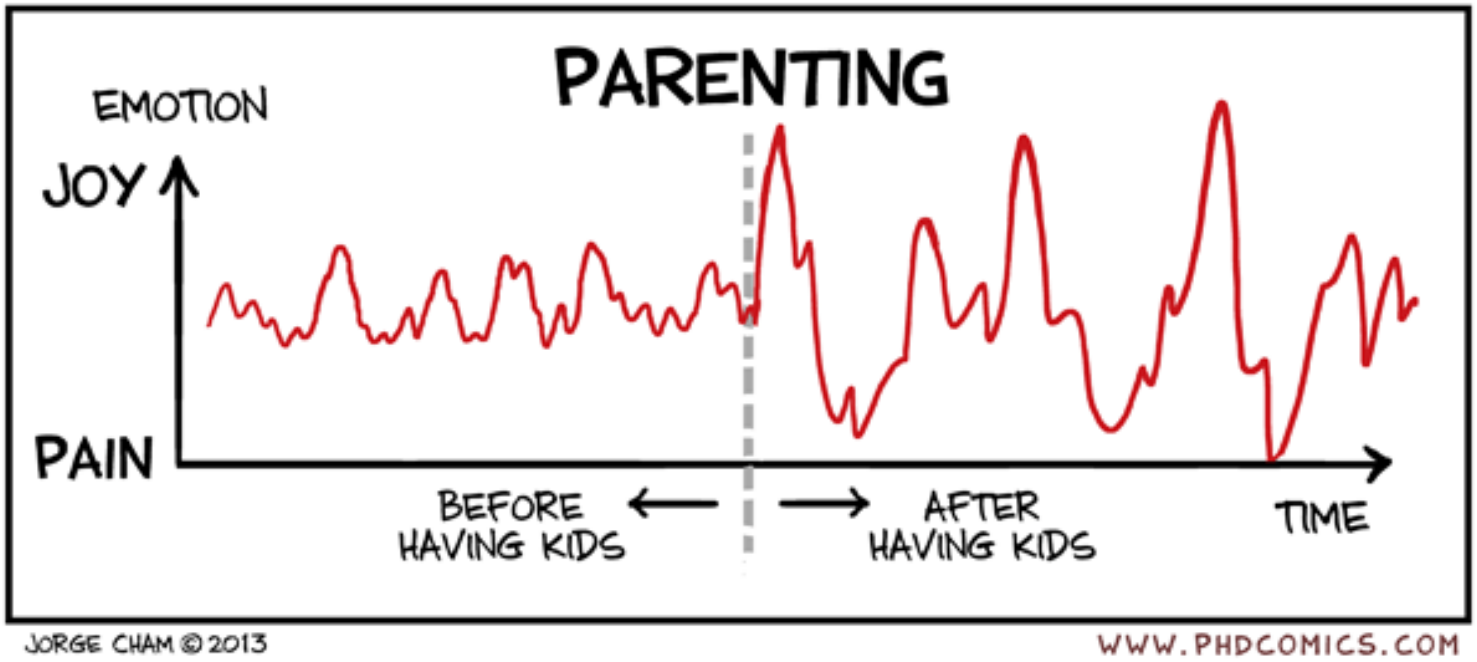
Series de tiempo





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

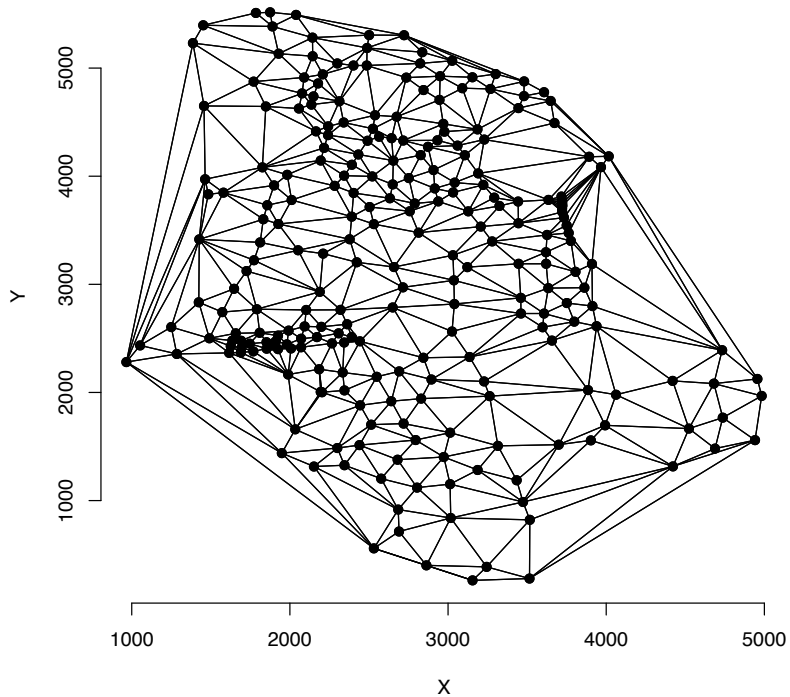
Series de tiempo



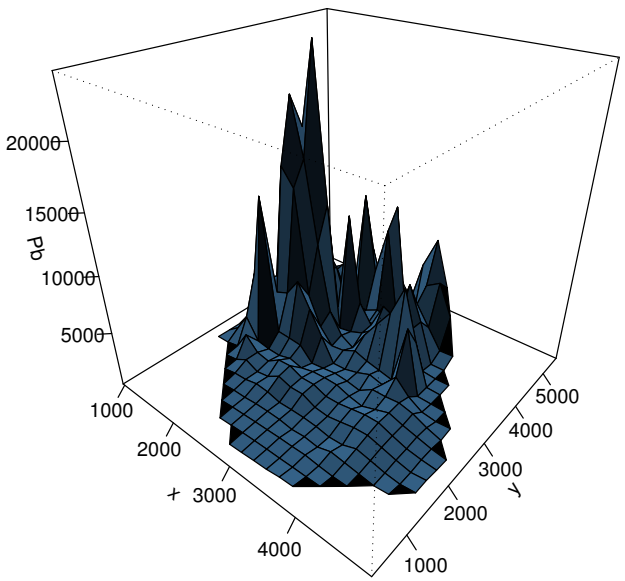
Análisis de puntos de cambio (Chen y Gupta, 2011).

Otros tipos de gráficos

Coordenadas de las muestras de suelo

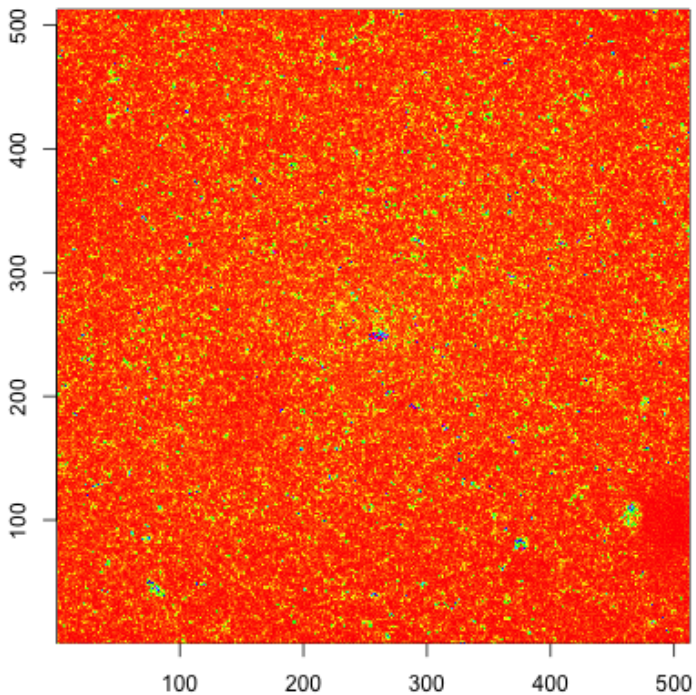


Concentración de Plomo

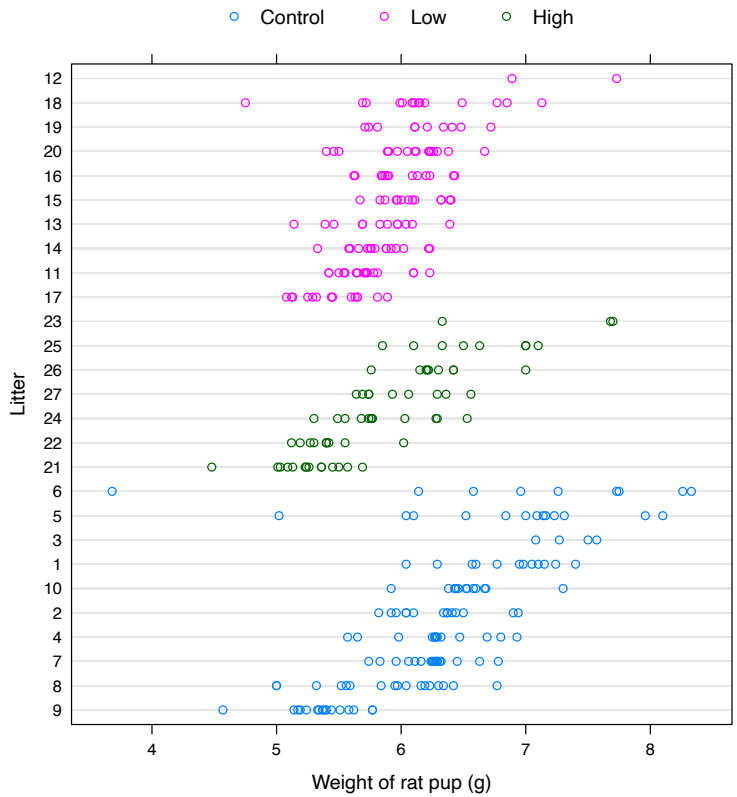


Otros tipos de gráficos

Flamabilidad de nanotubos

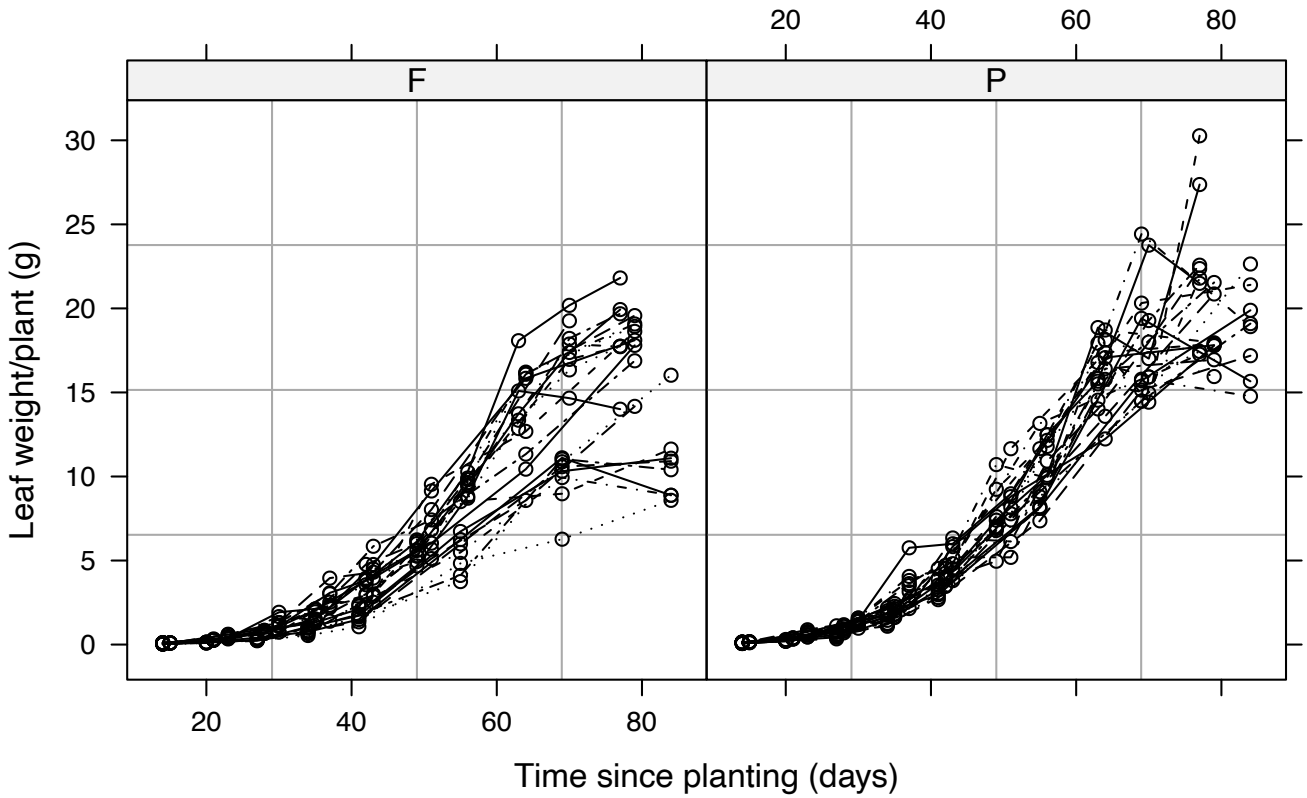


Estudio reproductivo con roedores



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.