MAT-032: Probabilidad y Estadística Comercial

Felipe Osorio

http://fosorios.mat.utfsm.cl

Departamento de Matemática, UTFSM

Información

Horario:

Clases: Lunes, bloque 16-20 (19:05-22:20 hrs.) via Zoom
Taller: Martes, bloque 16-20 (19:05-22:20 hrs.) a cargo de prof. Enzo Hernández.
Ayudante: Fabián Ramírez

Contacto:

E-mail: felipe.osorios@usm.cl.
Web: http://fosorios.mat.utfsm.cl/teaching.html y **AULA**

Evaluación:

Se realizará **3 Certámenes**, **Controles** y **Tareas**.

Ponderaciones:

Sea $\overline{C}$, $\overline{Q}$ y $\overline{T}$ el promedio de certámenes, controles, y tareas, respectivamente. De este modo, la nota de presentación (NP) es dada por:

$$NP = 0.8\,\overline{C} + 0.1\,\overline{Q} + 0.1\,\overline{T}.$$

Criterio de aprobación

Criterio de aprobación:

Aquellos estudiantes que obtengan NP mayor o igual a 55 y **todos** los certámenes sobre 40, aprobarán la asignatura con nota final, $NF = NP$.

Criterio para rendir global:

En caso contrario, y siempre que $NP \geq 45$, los estudiantes podrán rendir el certamen global (CG), en cuyo caso la nota final es calculada como sigue:

$$NF = 0.6 \cdot NP + 0.4 \cdot CG.$$

Reglas adicionales

- Se llevará un control de asistencia.
- Se puede realizar preguntas sobre la materia en cualquier momento.
- Los alumnos deben apagar/silenciar sus teléfonos celulares durante clases.
- Conversaciones sobre asuntos ajenos a la clase no serán tolerados. Otros estudiantes tiene derecho a asistir clases en silencio.
- Al enviar algún e-mail al profesor, identificar el código de la asignatura en el asunto (MAT206).
- E-mail será el canal de comunicación oficial entre el profesor y los estudiantes.

Reglas: sobre las pruebas

- Es derecho del estudiante conocer la pauta de corrección la que será publicada en la página web del curso.
- El uso de lápiz grafito es aceptado. Sin embargo, inhabilita al estudiante de pedir recorrección.
- Pedidos de recorrección deben ser argumentados por escrito.
- En modalidad online, Certámenes, Controles y Tareas deben ser enviados en formato **PDF**.[1]
- Cualquier tipo de fraude en prueba (copia, WhatsApp, suplantación, etc.) implicará la reprobación de los involucrados.[2]

[1] En un único archivo, orientado en una dirección legible.
[2] Puede implicar la apertura de un proceso disciplinario.

Orientaciones de estudio

- Mantener la frecuencia de estudio de inicio a final del semestre. El ideal es estudiar el contenido luego de **cada** clase.
- Estudiar primeramente el contenido dado en clases, buscando apoyo en las referencias bibliográficas.
- Las referencias son fuentes de ejemplos y ejercicios. Resuelva una buena cantidad de ejercicios. No deje esto para la víspera de la prueba.
- Buscar las referencias bibliográficas al inicio del semestre, dando preferencia a las principales y complementarias.

Programa del curso

- Introducción y conceptos básicos.
- Estadística descriptiva.
- Cálculo de probabilidades.
- Variables aleatorias.
- Inferencia estadística.

Bibliografía

Canavos, G. (1990).
Probabilidad y Estadística, Aplicaciones y Métodos.
McGraw-Hill Latinoamericana.

Meyer, P.L. (1976)
Probabilidad y Aplicaciones Estadísticas.
Fondo Educativo Interamericano.

Newbold, O., Carlson, W.L., Thorne, B. (2008).
Estadística para Administración y Economía (6ta Ed.).
Prentice Hall, Madrid.

Wackerly, D., Mendenhall, W., Scheaffer, R. (2008).
Estadística Matemática con Aplicaciones.
Cengage Learning.

Motivación mediante ejemplos

- ¿Existe competencia en el mercado de AFPs chileno?
- Modelando rentabilidades de acciones en el mercado chileno usando el CAPM[3].
- Ideas sobre el proceso de modelación.
- Algunos conceptos preliminares.

[3] CAPM: Capital asset pricing model

Ideas subyacentes

"Todos los modelos son errados, pero algunos son útiles."

– George Box.

"Aunque puede parecer una paradoja, toda la ciencia exacta está dominada por la idea de aproximación."

– Bertrand Russell.

Principio KISS: "Keep It Simple, Stupid."

– Clarence "Kelly" Johnson.

Ideas subyacentes

"Todos los modelos son errados, pero algunos son útiles."

– George Box.

"Aunque puede parecer una paradoja, toda la ciencia exacta está dominada por la idea de aproximación."

– Bertrand Russell.

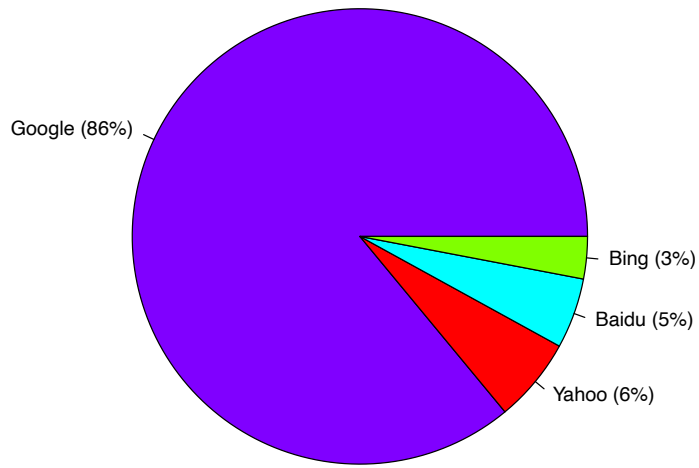
Principio KISS: "Keep It Short and Simple."

– Clarence "Kelly" Johnson.

El éxito de Google: Aplicar el principio KISS[4]

Evolución de Yahoo vs. Google:

Cuota de mercado de los motores de búsqueda:



[4] En estadística este se conoce como Principio de Parsimonia.

Esto NO es una crítica al sistema de AFP...

Administradoras de Fondos de Pensiones (AFP) de Chile

Aplicación:

El sistema de AFP (o de capitalización individual) chileno está en vigor desde 1980.

Ahorros de los contribuyentes son administrados en un sistema de multifondos.

Existe 5 tipos de fondos (A, B, C, D y E) divididos por la proporción del portfólio que es invertido en títulos de renta variable.

El fondo A tiene la mayor proporción de inversión en renta variable, la que disminuye progresivamente para los fondos B, C, D y E.

Conjunto de datos:

Rentabilidades mensuales de AFPs: Cuprum, Habitat, PlanVital y ProVida en el periodo de agosto/2005 a diciembre/2013.

Datos fueron obtenidos desde el sitio web de la superintendencia de pensiones (www.spensiones.cl)

Conjunto de datos con 101 observaciones y 4 variables (solamente datos del Fondo D).

Obs. 28, 30, 35, 38, 39, 42, 66, 73 y 75 son identificadas como outliers.

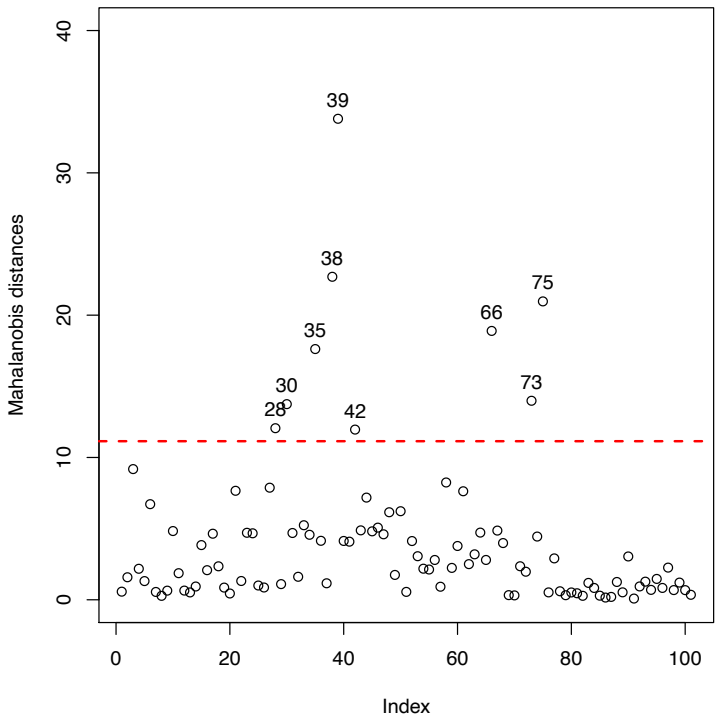
QQ-plot de distancias transformadas revelan la presencia de colas pesadas.

Identificando observaciones atípicas

En mercados emergentes como el chileno suele ocurrir periodos con alta volatilidad.

Existe una bateria de procedimientos para detectar observaciones que presentan un comportamiento es aberrante/atípico.

Este tipo de observaciones puede tener un efecto nefasto sobre la inferencia estadística.

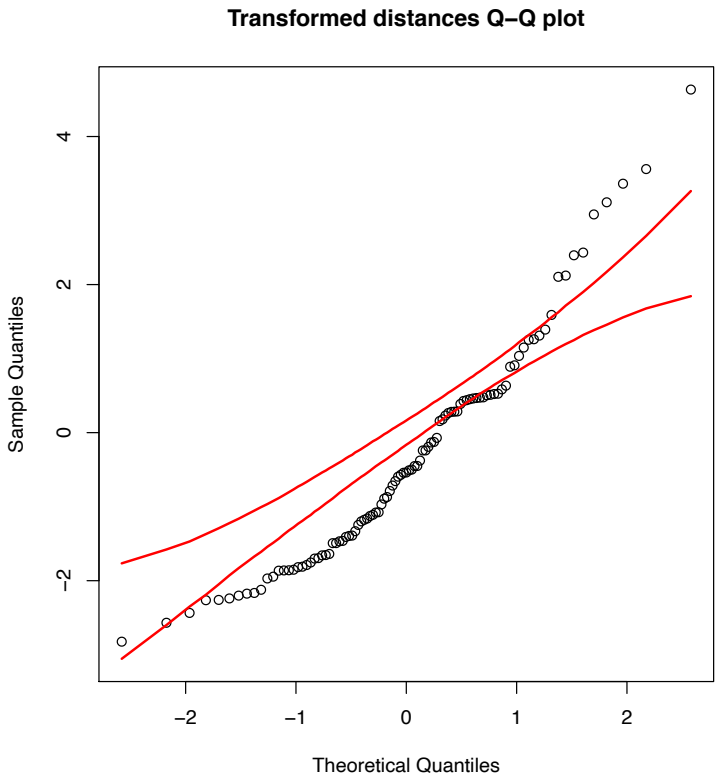


Evaluando los supuestos distribucionales

El supuesto de normalidad es habitual en este tipo de problemas.

Es decir, suponga $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ una muestra aleatoria desde $\mathrm{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Usando test de hipótesis y técnicas gráficas se concluye que el supuesto de normalidad no es soportado por los datos.



Análisis multivariado usando la distribución t de Student

Características del problema:

- AFPs invierten esencialmente en la misma cartera de inversiones.
- Mercados emergentes suelen presentar alta volatilidad.
- Los datos son bien modelados usando la distribución t multivariada.

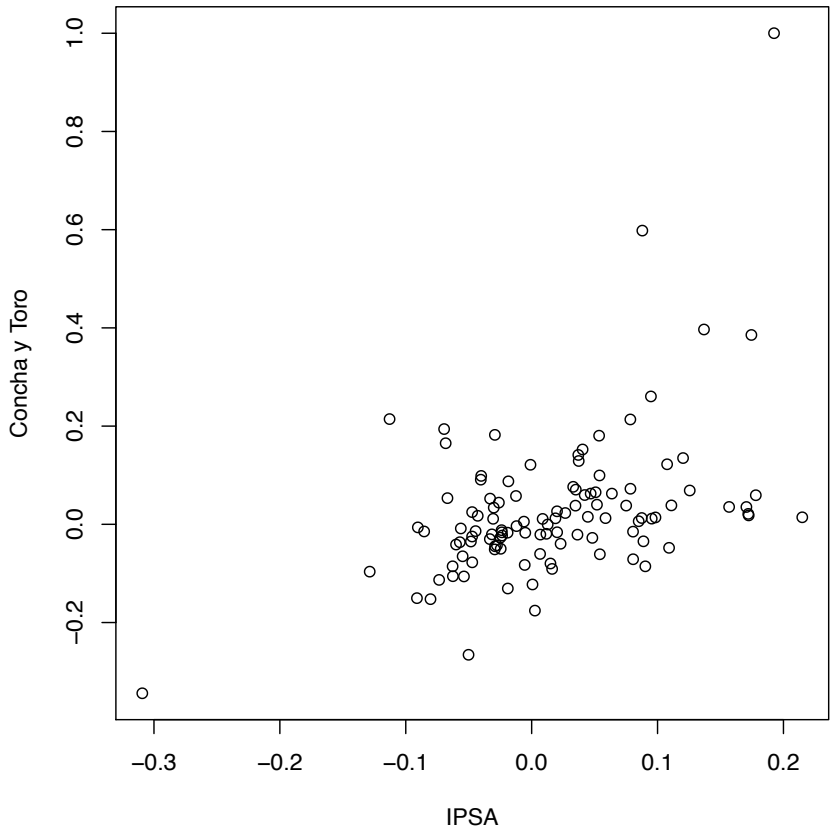
Conclusiones:

- Aparentemente, no existe competencia en el mercado de AFP.
- Cálculo óptimo de los porcentajes de inversión en los distintos fondos.
- Test para evaluar la igualdad entre razones de Sharpe.[5]

[5] Trabajo en desarrollo junto al prof. Manuel Galea (PUC)

Datos de Concha y Toro (Osorio y Galea, 2006)[6]

Rentabilidades mensuales de Concha y Toro vs. IPSA, ajustados por bonos de interés del Banco Central entre marzo/1990 a abril/1999.



[6] Statistical Papers **47**, 31-38

Datos de Concha y Toro

Modelo CAPM (Valoración de Activos de Capital), Sharpe (1964)[7]

$$\mathrm{E}(r) = r_f + \beta(\mathrm{E}(r_m) - r_f),$$

usando datos observados, podemos escribir

$$R_t = \alpha + \beta \times IPSA_t + \epsilon, \qquad t = 1, \ldots, T.$$

Características del problema:

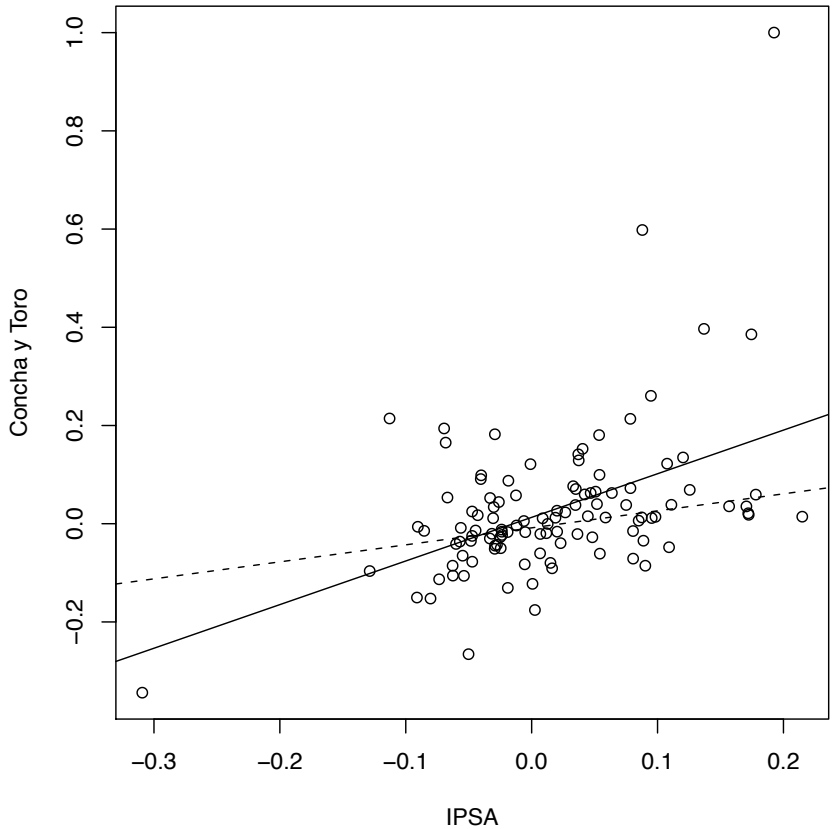
- Relación lineal entre las variables.
- Posibles periodos de alta volatilidad.

Hipótesis de interés:

- $H_0 : \beta > 1$ (Amante del riesgo).
- $H_0 : \beta = 1$ (Neutral al riesgo).
- $H_0 : \beta < 1$ (Averso al riesgo).

[7] Journal of Finance **19**, 425-442

Datos de Concha y Toro



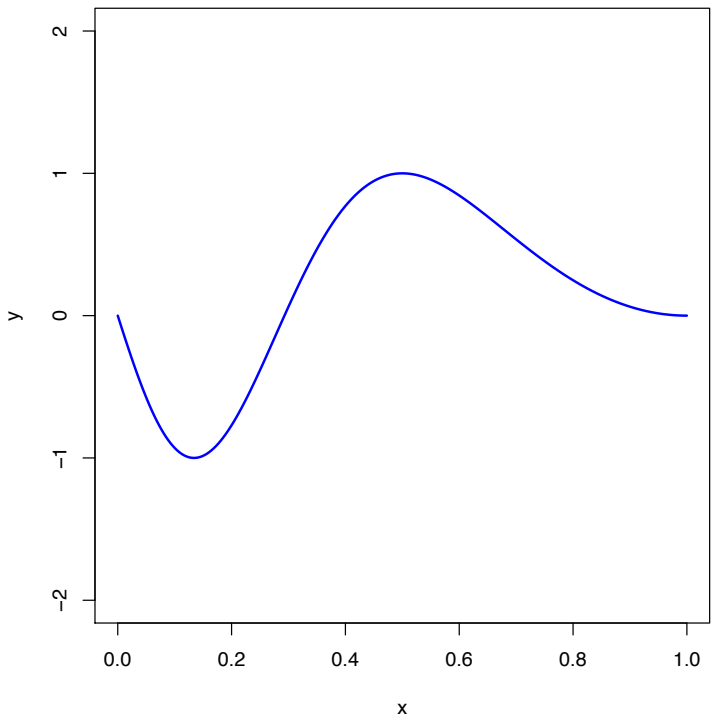
Ajuste usando errores normales (—) y Cauchy (– –). ($\widehat{\beta} = 0.89$ y $\widehat{\beta} = 0.35$, respectivamente).

El problema del modelado

Considere la función

$$Y = \text{sen}\{2\pi(1-x)^2\},$$

cuyo gráfico es dado por:

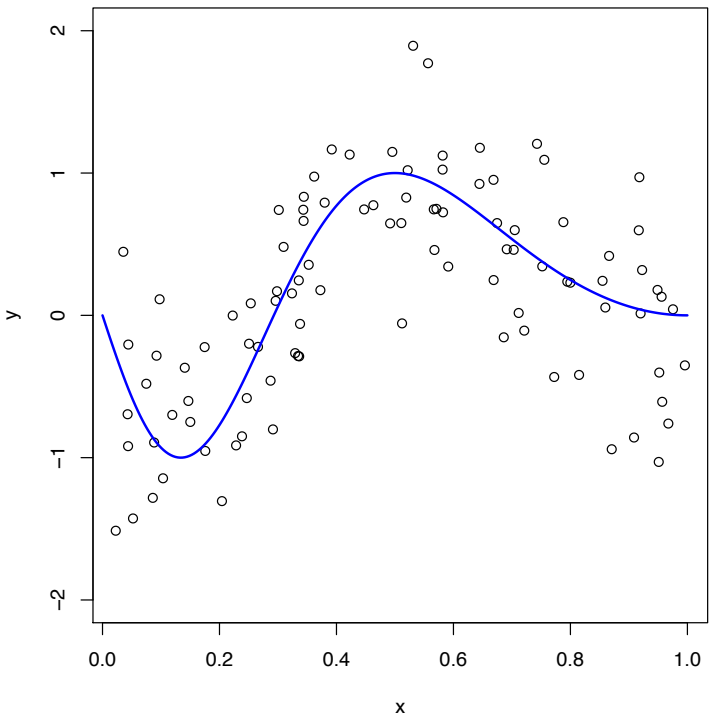


El problema del modelado

Suponga que "generamos" datos, usando

$$Y_i = \text{sen}\{2\pi(1 - x_i)^2\} + \sigma\epsilon_i, \qquad i = 1, \ldots, 100,$$

donde $x_i \sim \mathcal{U}(0, 1)$, $\epsilon_i \sim \mathcal{N}(0, 1)$ y $\sigma = 1/2$,

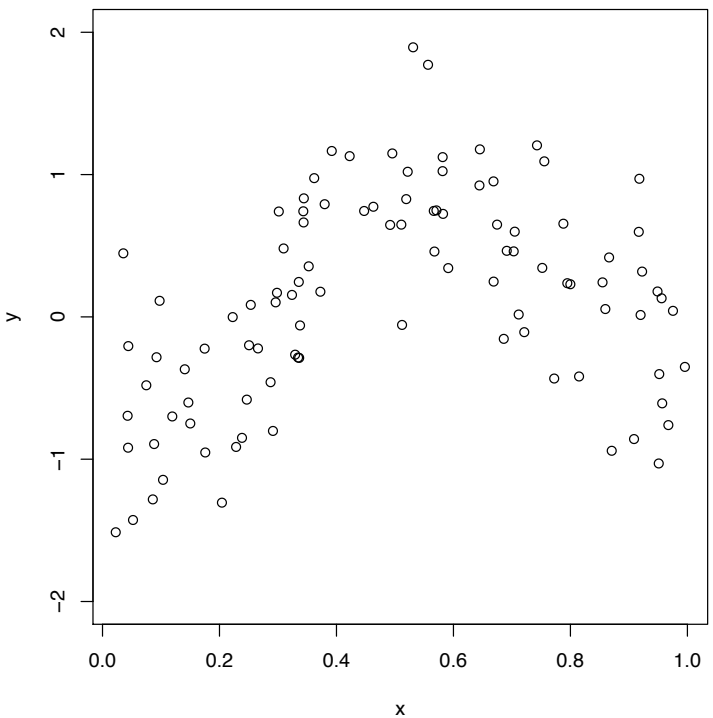


El problema del modelado

Lamentablemente, en la práctica **sólo** disponemos de los datos observados:

$$(x_1, Y_1), (x_2, Y_2), \ldots, (x_{100}, Y_{100}),$$

el primer paso es hacer un análisis exploratorio:

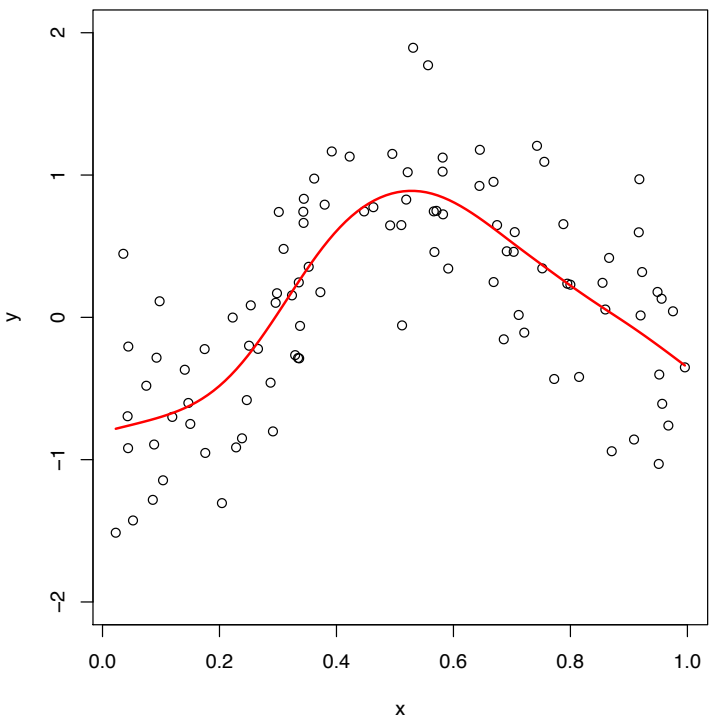


El problema del modelado

El analista propone el modelo:

$$Y_i = g(x_i) + \epsilon_i, \qquad i = 1, \ldots, 100,$$

y su objetivo es "estimar" la función $g(\cdot)$ desde los datos, obteniendo

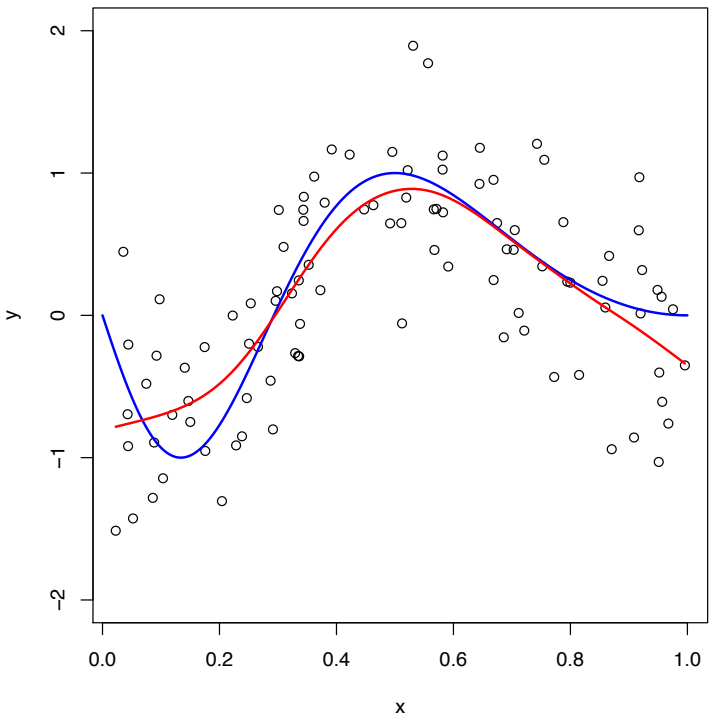


El problema del modelado

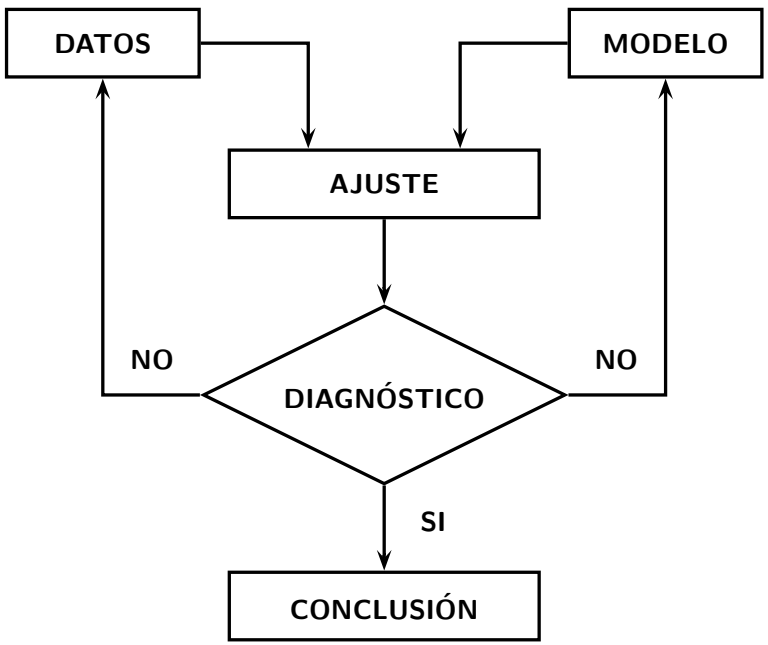
En Estadística se estudia teóricamente, la "bondad del modelo" comparando

$$\widehat{Y} = \widehat{g}(x), \qquad \text{v.s.} \qquad Y = \text{sen}\{2\pi(1-x)^2\},$$

esto es, el modelo ajustado v.s. el modelo subyacente (verdadero).



Esquema de Modelación Estadística



Recolección de datos: Muestreo.

Análisis exploratorio de datos.

Análisis Multivariado.

Técnicas de Regresión.

Series de Tiempo, entre (muchas) otras.

Inferencia Estadística.

Bondad de ajuste, técnicas gráficas.

Análisis de Sensibilidad.

Comunique sus resultados!

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 tienen 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.

  Suponga que se desea una muestra de tamaño n desde una población de tamaño N y considere K el entero más cercano a N/n. Se escoge un numero al azar entre 1 y K, luego se selecciona una observación cada K observaciones hasta obtener una muestra de tamaño n.

Tipos de muestreo

- 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)

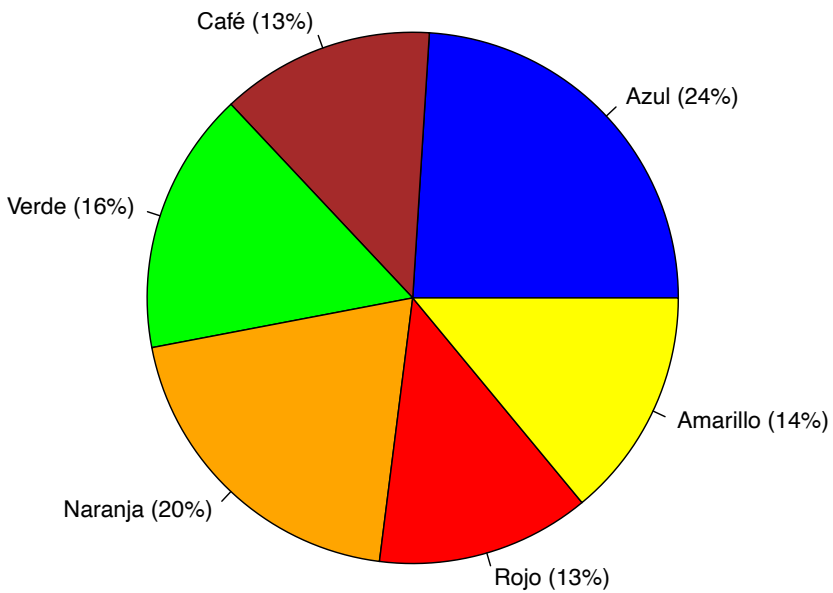


Gráfico Circular

Ejemplo: Proporción de población anglófona en el mundo

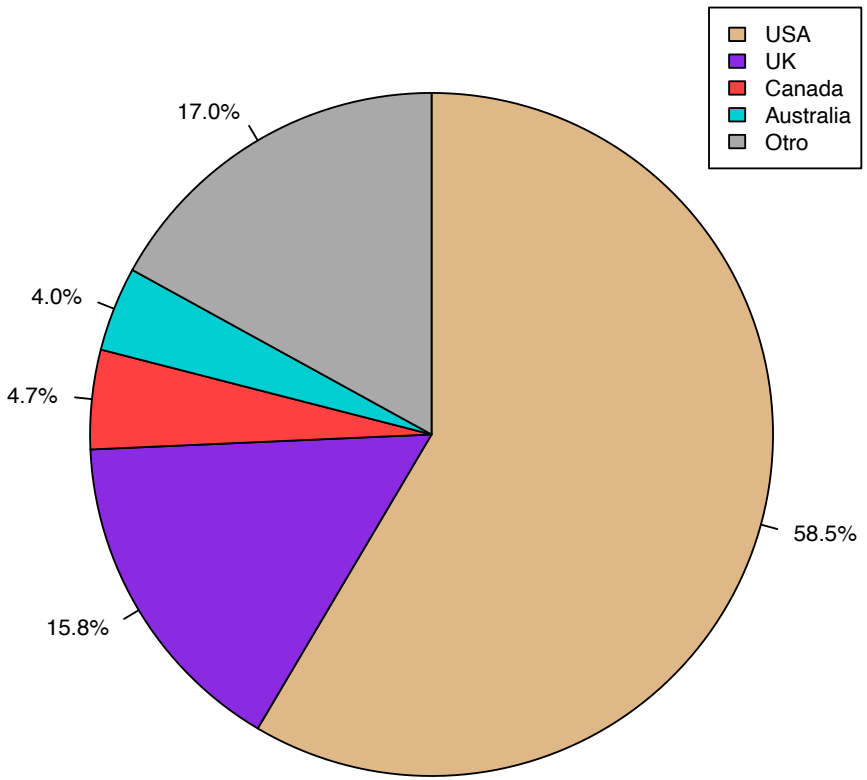


Gráfico Circular

Limitación: Gráfico circular de un arcoiris.

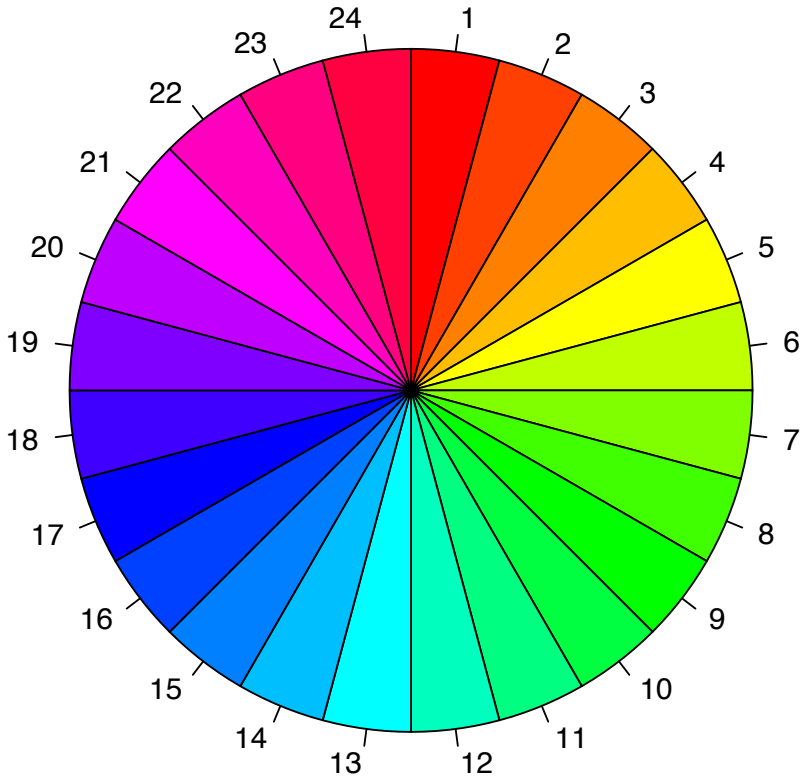


Gráfico Circular

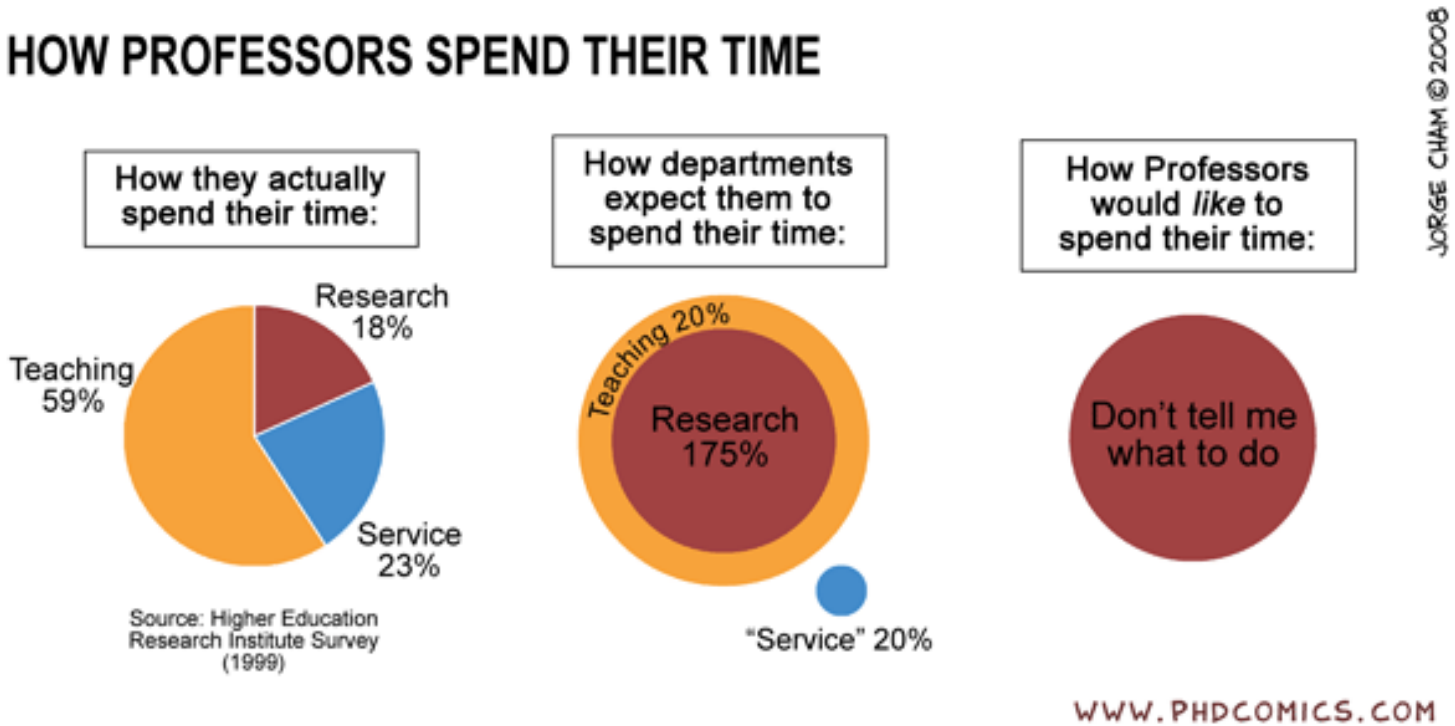


Gráfico de Barras

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.

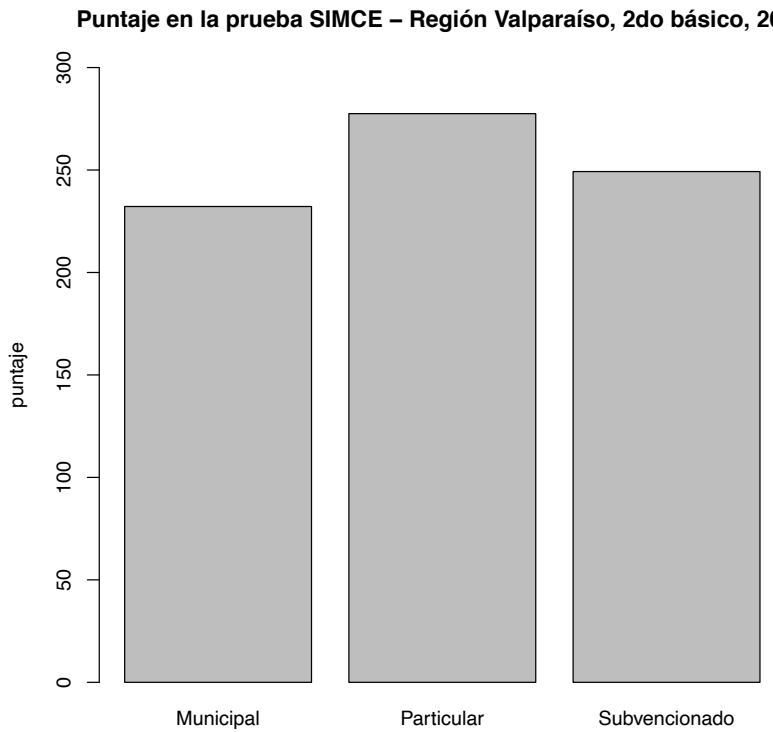
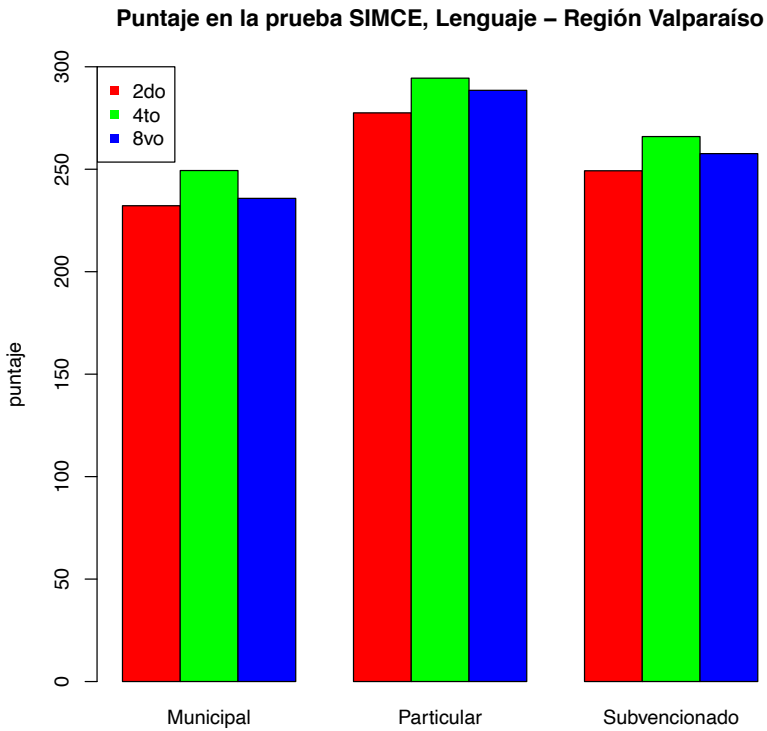


Gráfico de Barras



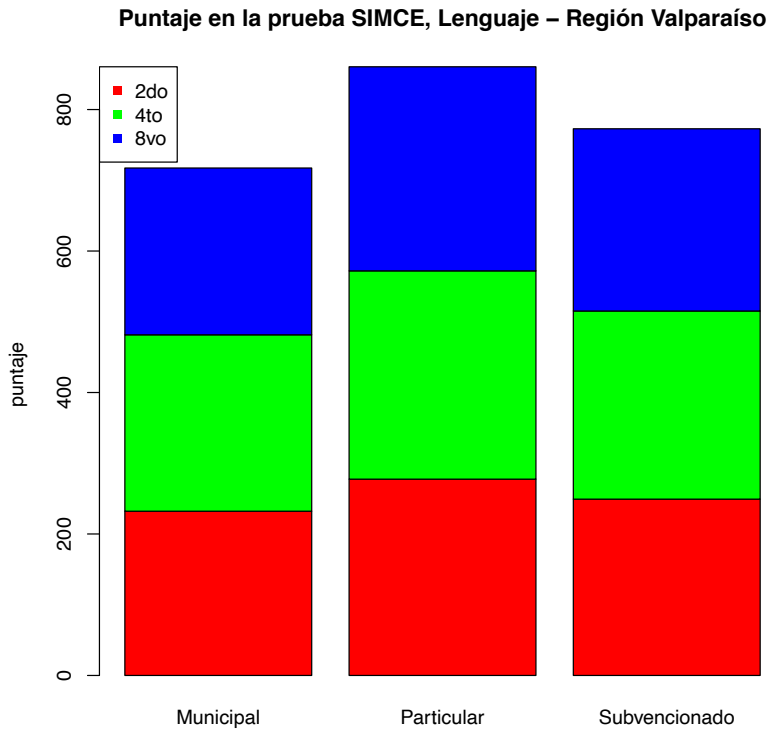
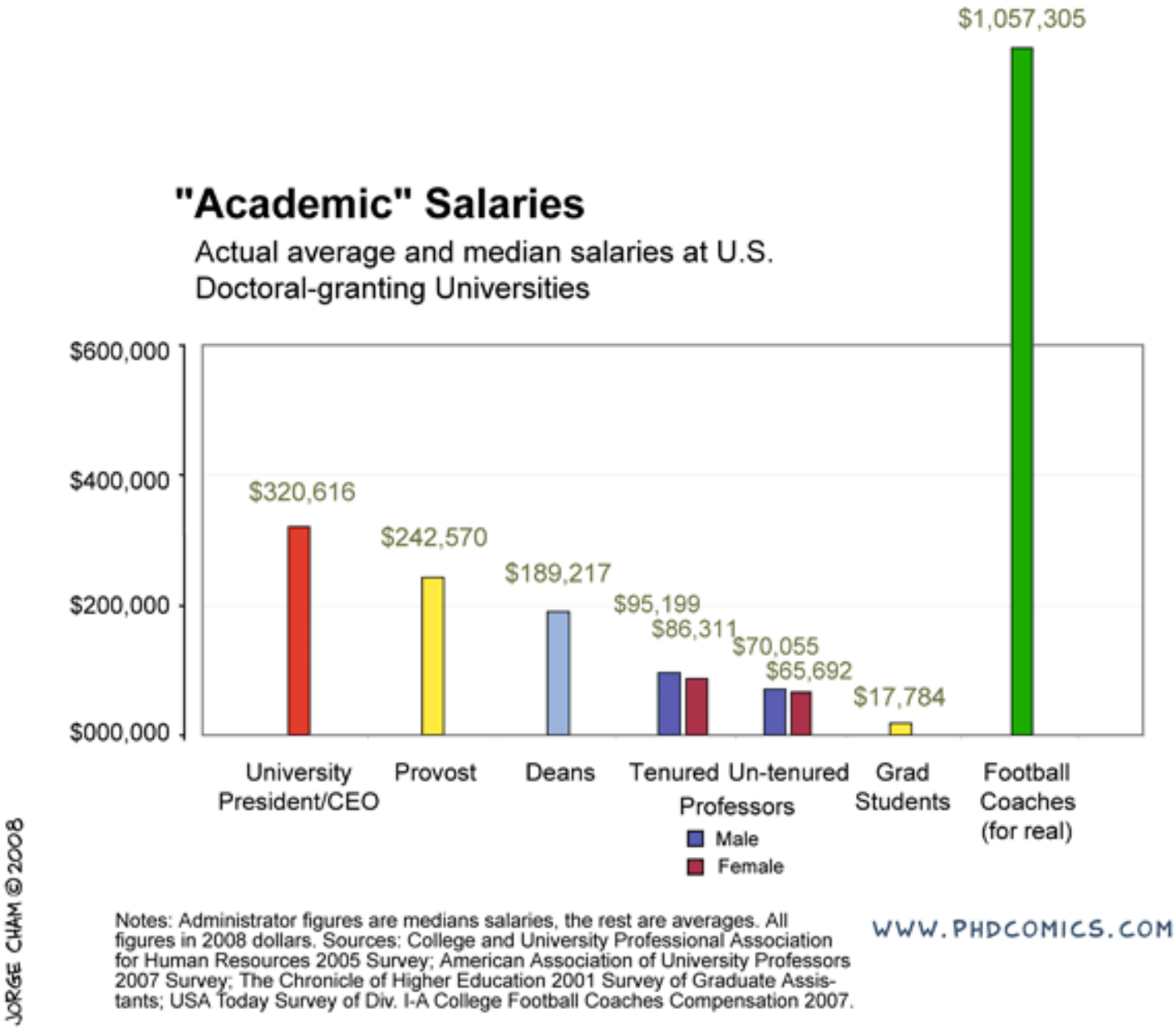


Gráfico de Barras



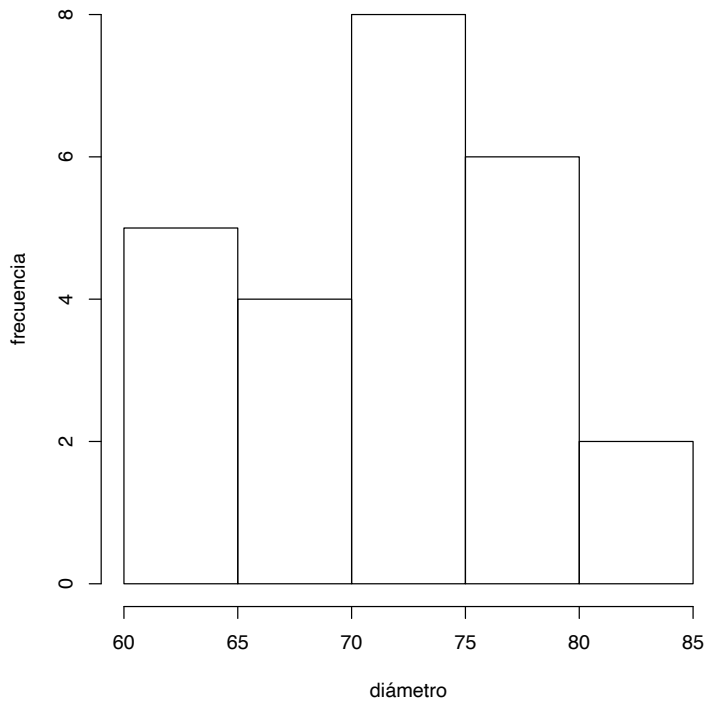
Histograma

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			

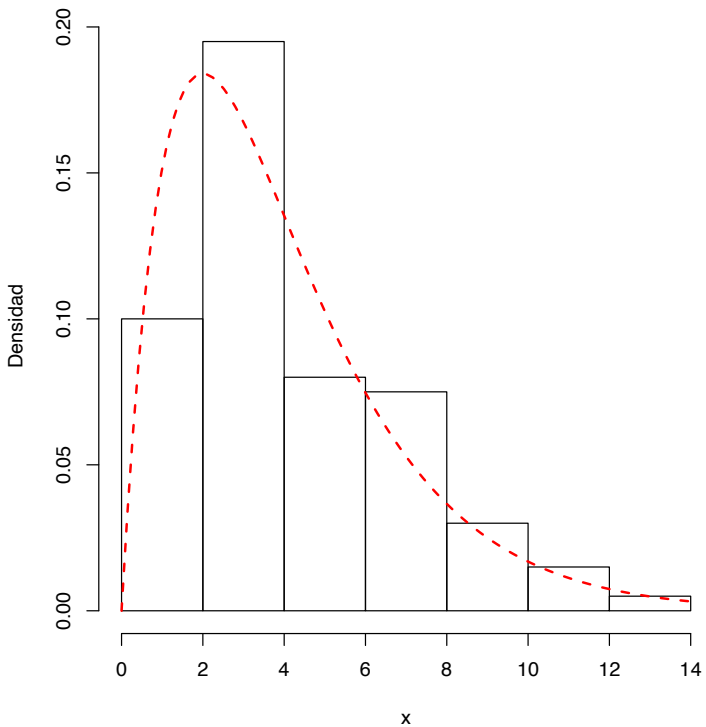


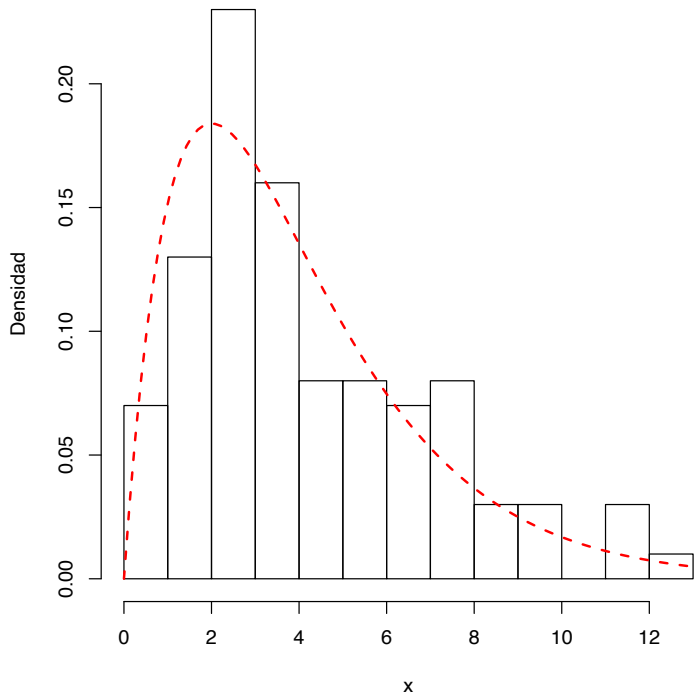
Histograma

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

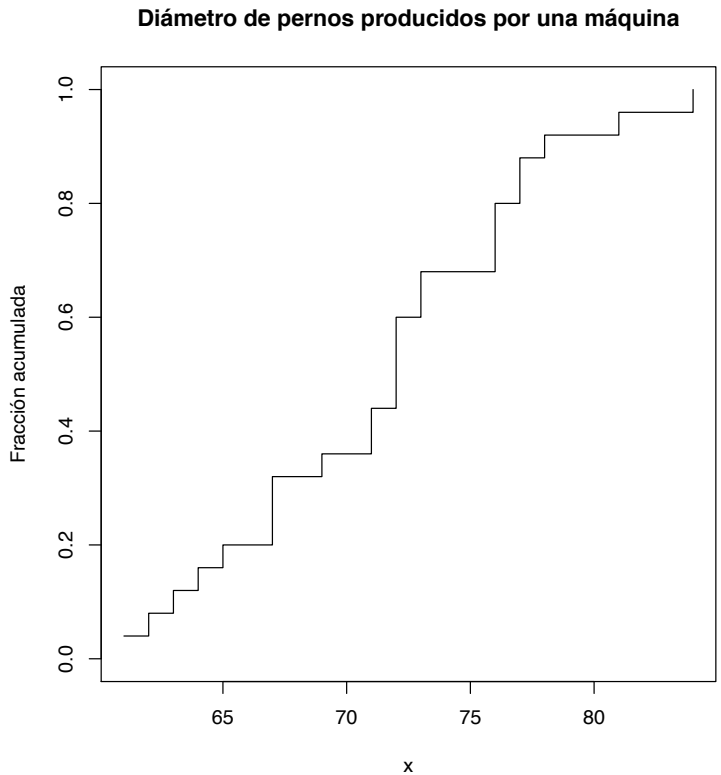
Función de distribución acumulada empírica (ojiva)

$F_n(x)$, cdf empírica, atribuye a cada valor de x, la fracción de datos menos o igual a x, i.e,

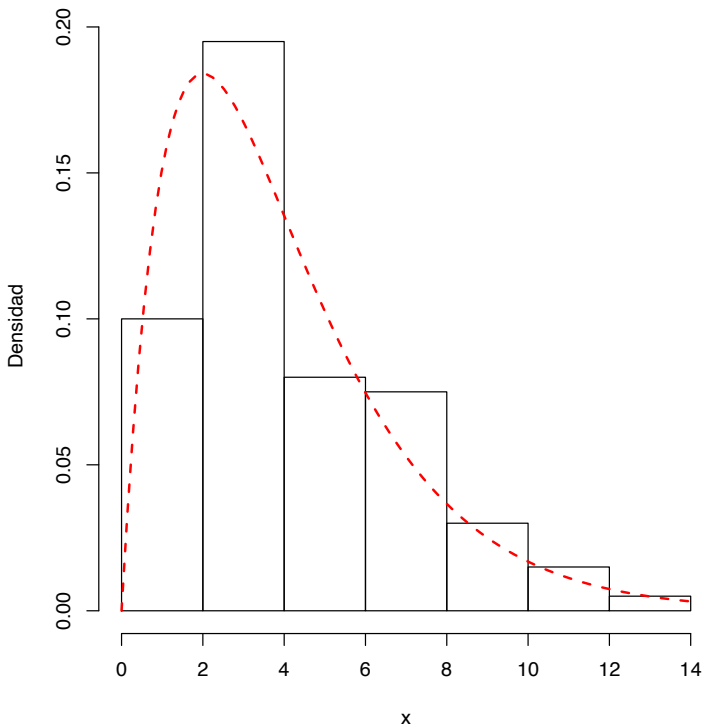
$$F_n(x) = \frac{1}{n} \# \text{ elementos} \leq x$$
$$= \frac{1}{n} \sum_{i=1}^{n} I_{\{x_i \leq x\}}.$$

Función de sobrevivencia:

$$S_n(x) = 1 - F_n(x).$$



Función de distribución acumulada empírica



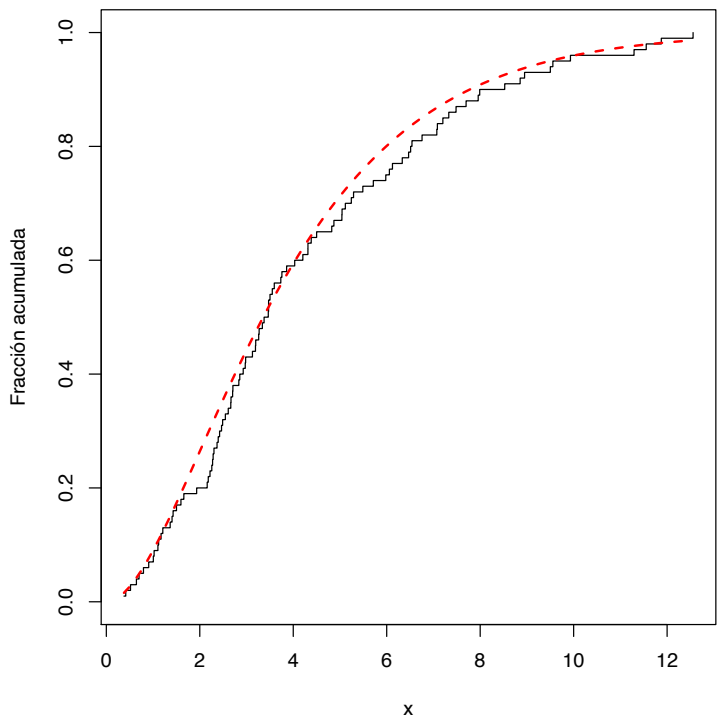


Diagrama de dispersión (scatterplot)

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

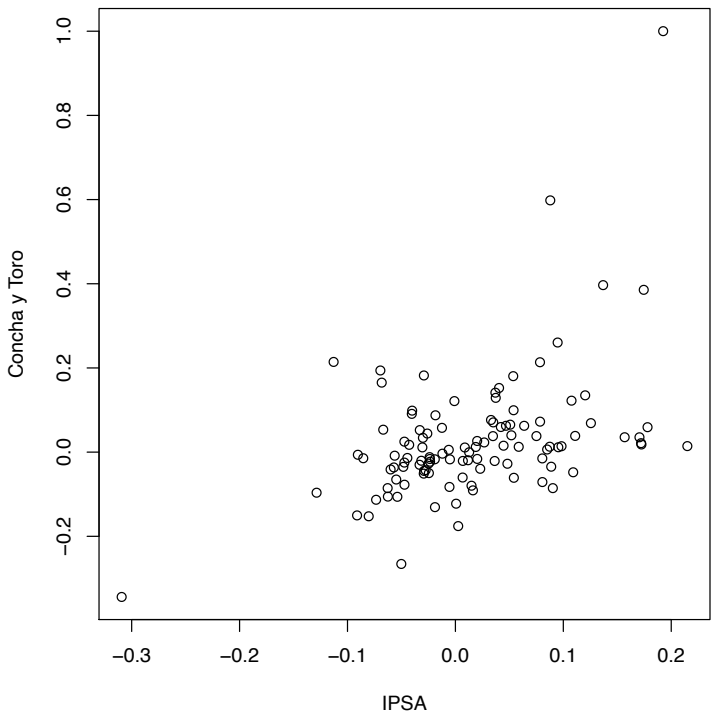
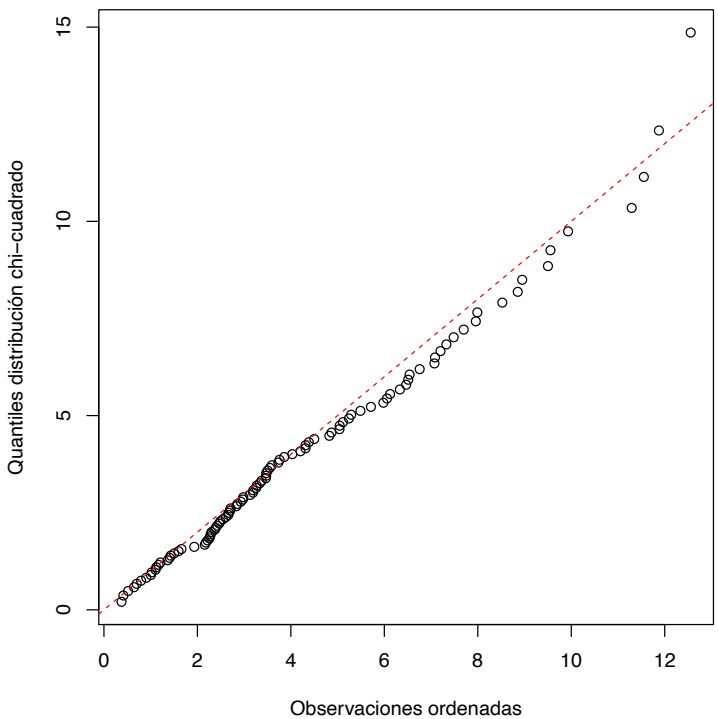
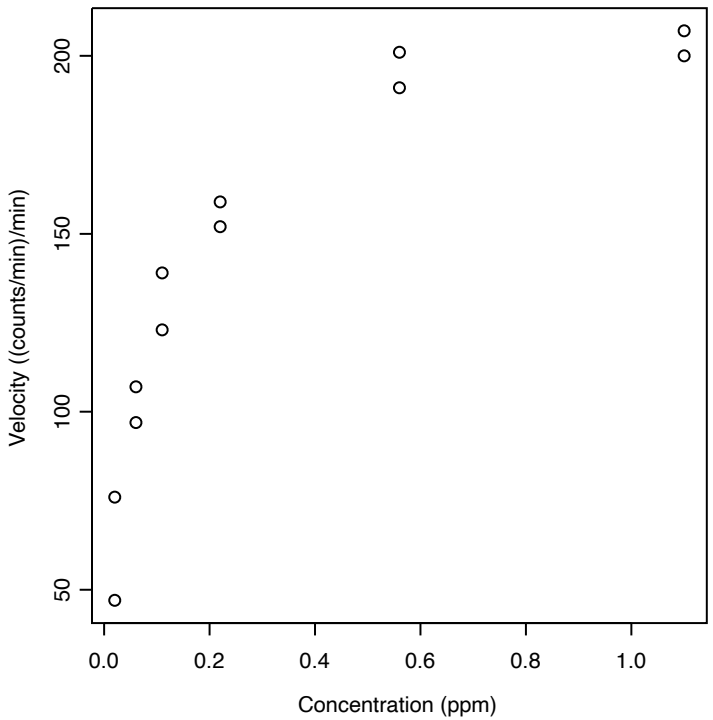


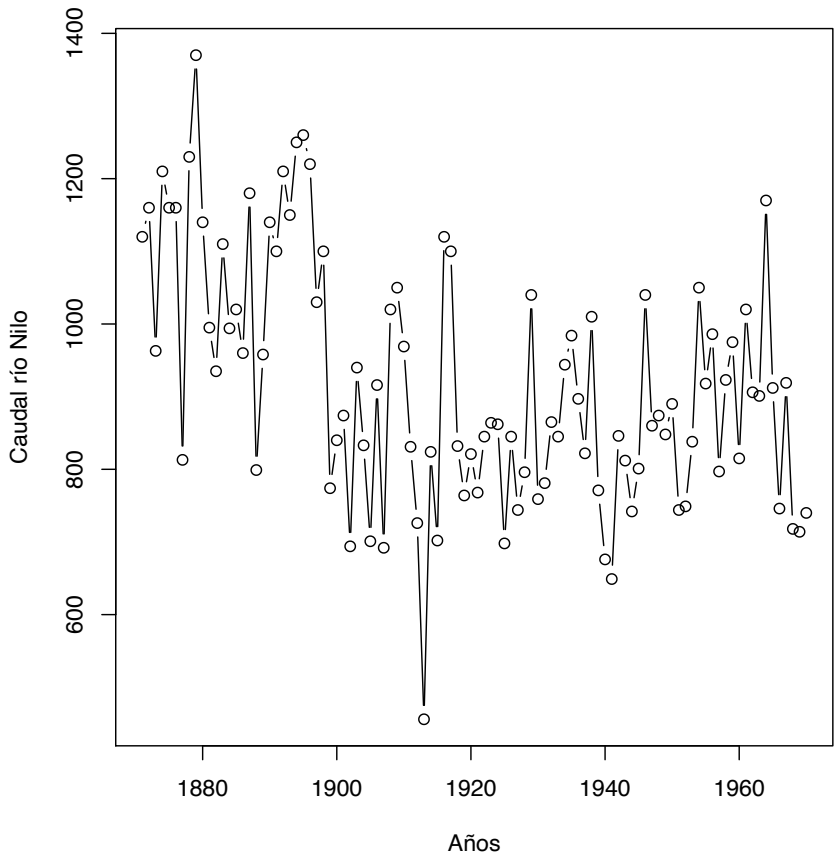
Diagrama de dispersión (scatterplot)



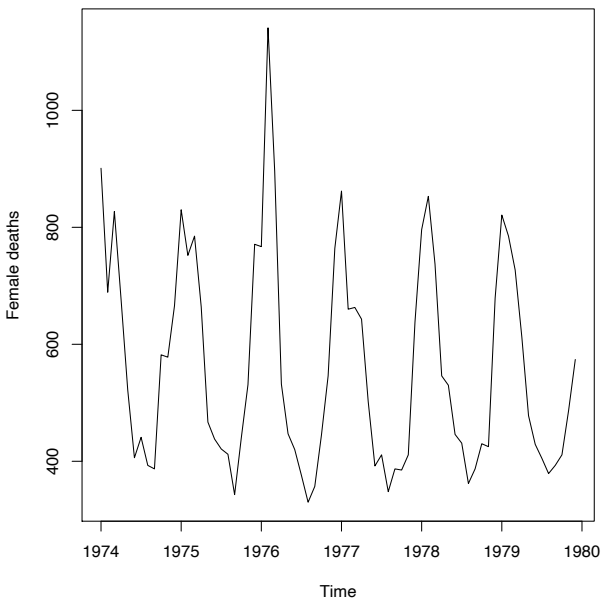


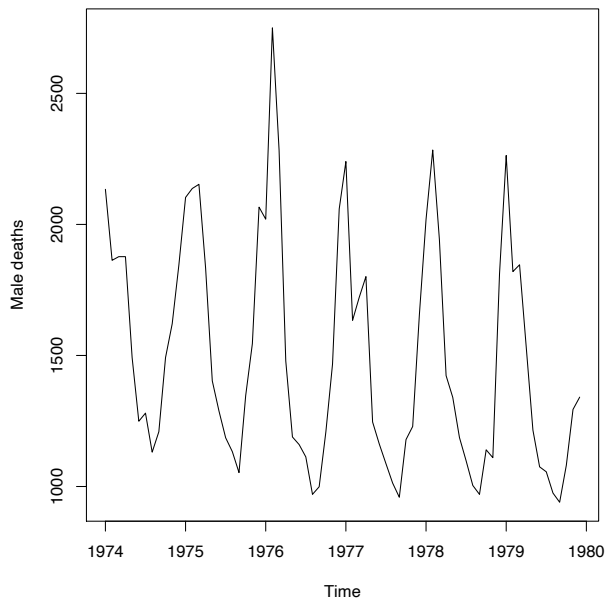
Series de tiempo

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



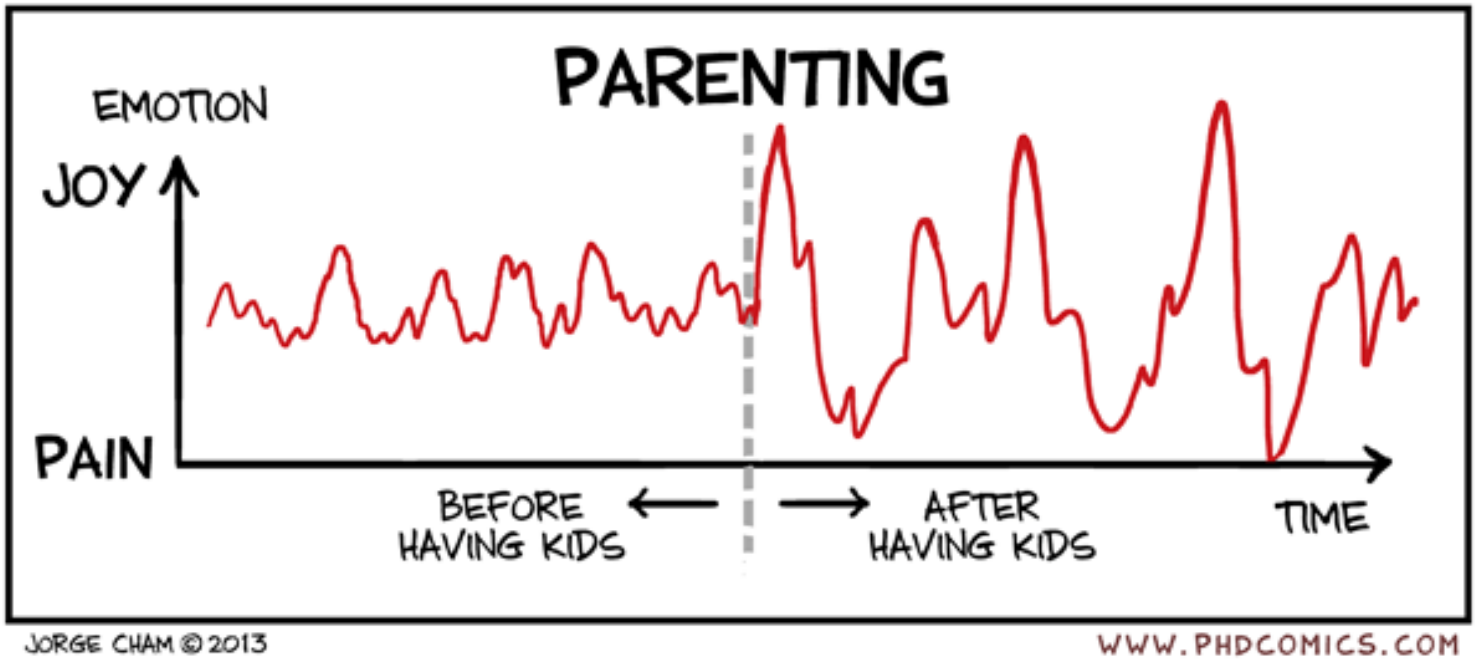
Series de tiempo





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

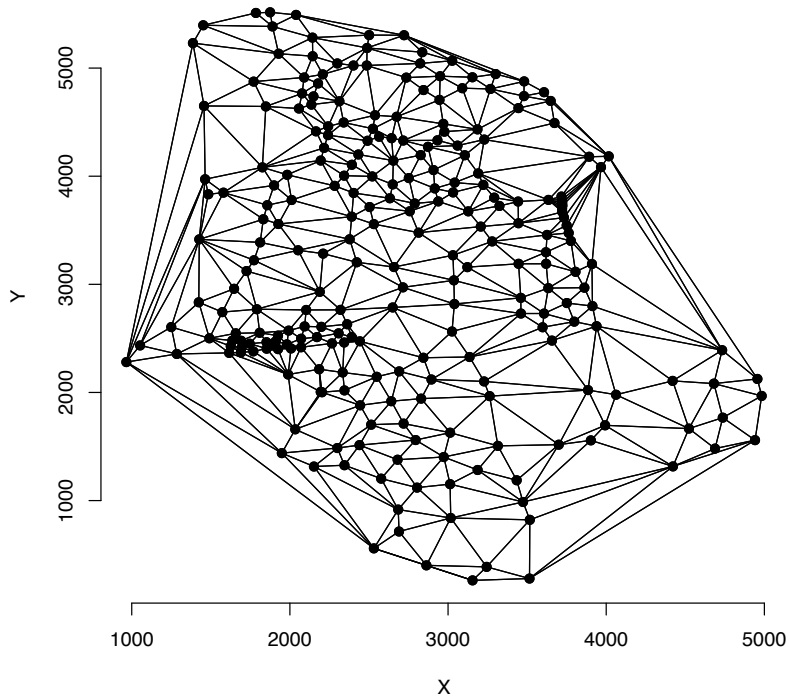
Series de tiempo



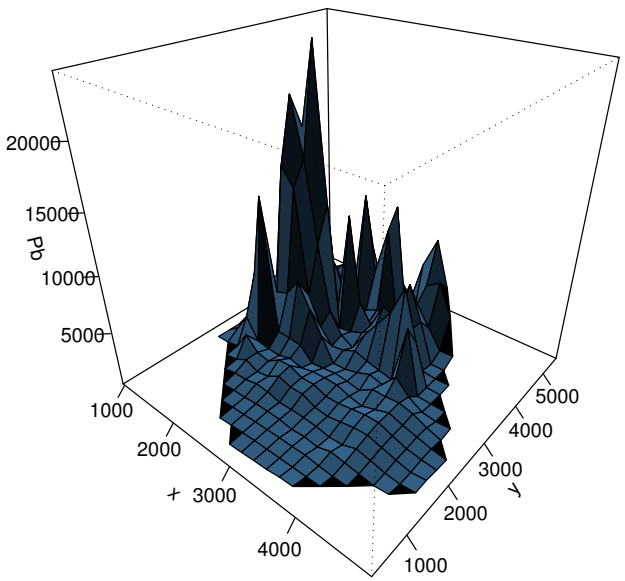
Análisis de puntos de cambio (Chen y Gupta, 2011).

Otros tipos de gráficos

Coordenadas de las muestras de suelo

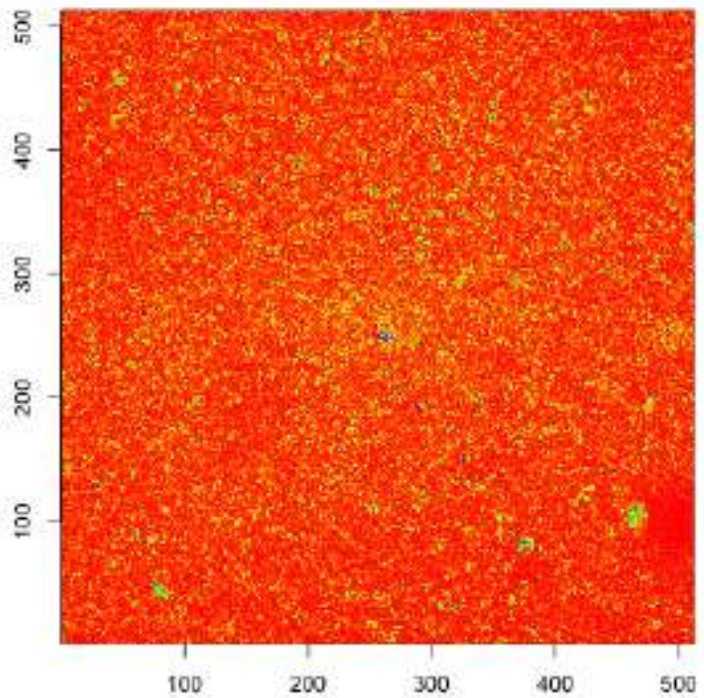


Concentración de Plomo

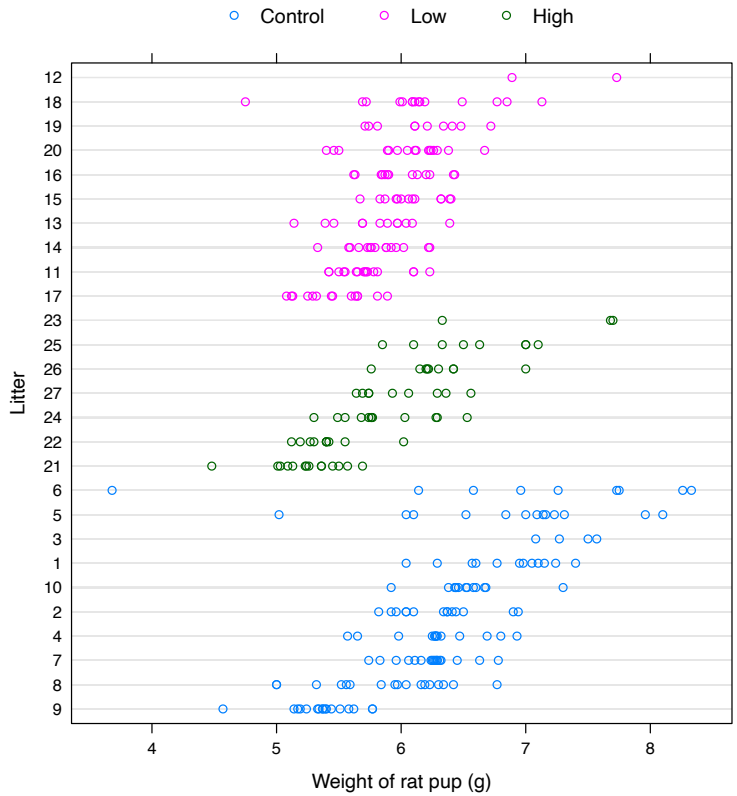


Otros tipos de gráficos

Flamabilidad de nanotubos

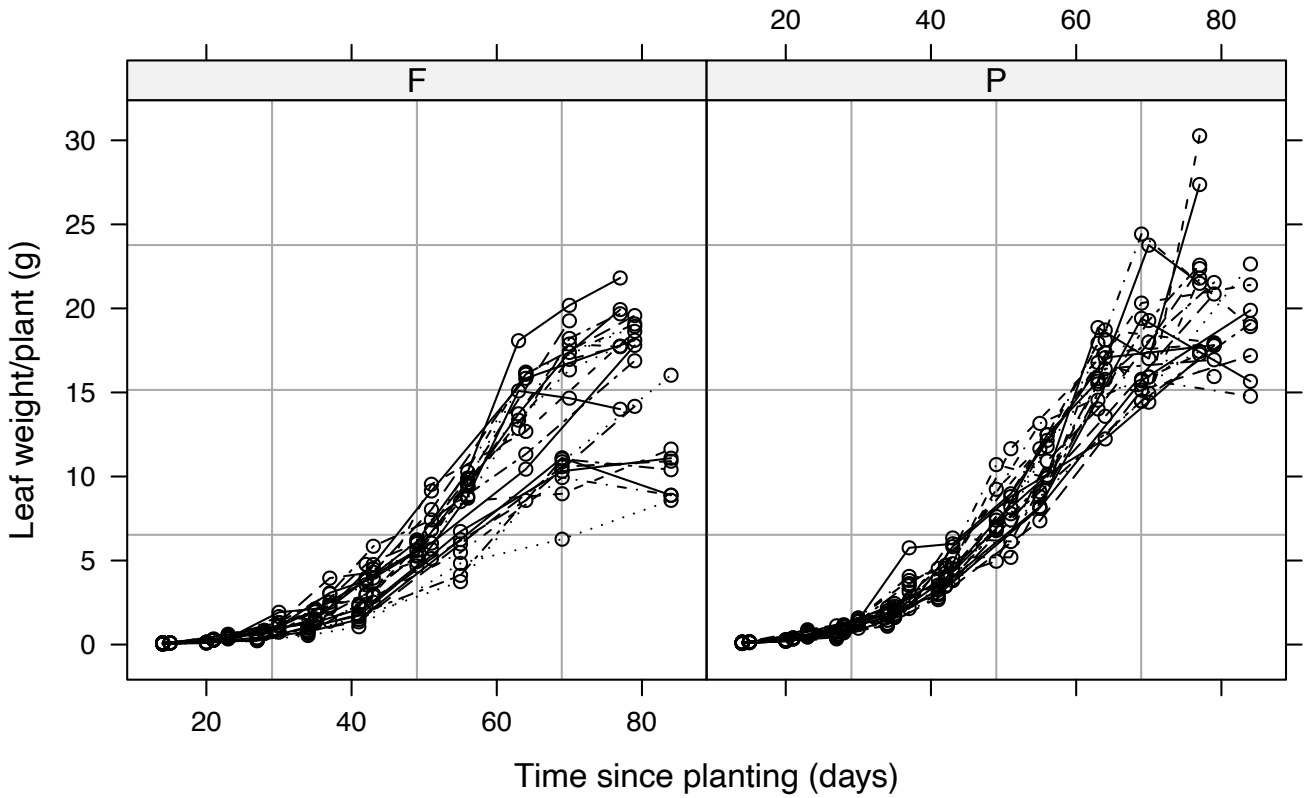


Estudio reproductivo con roedores



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.