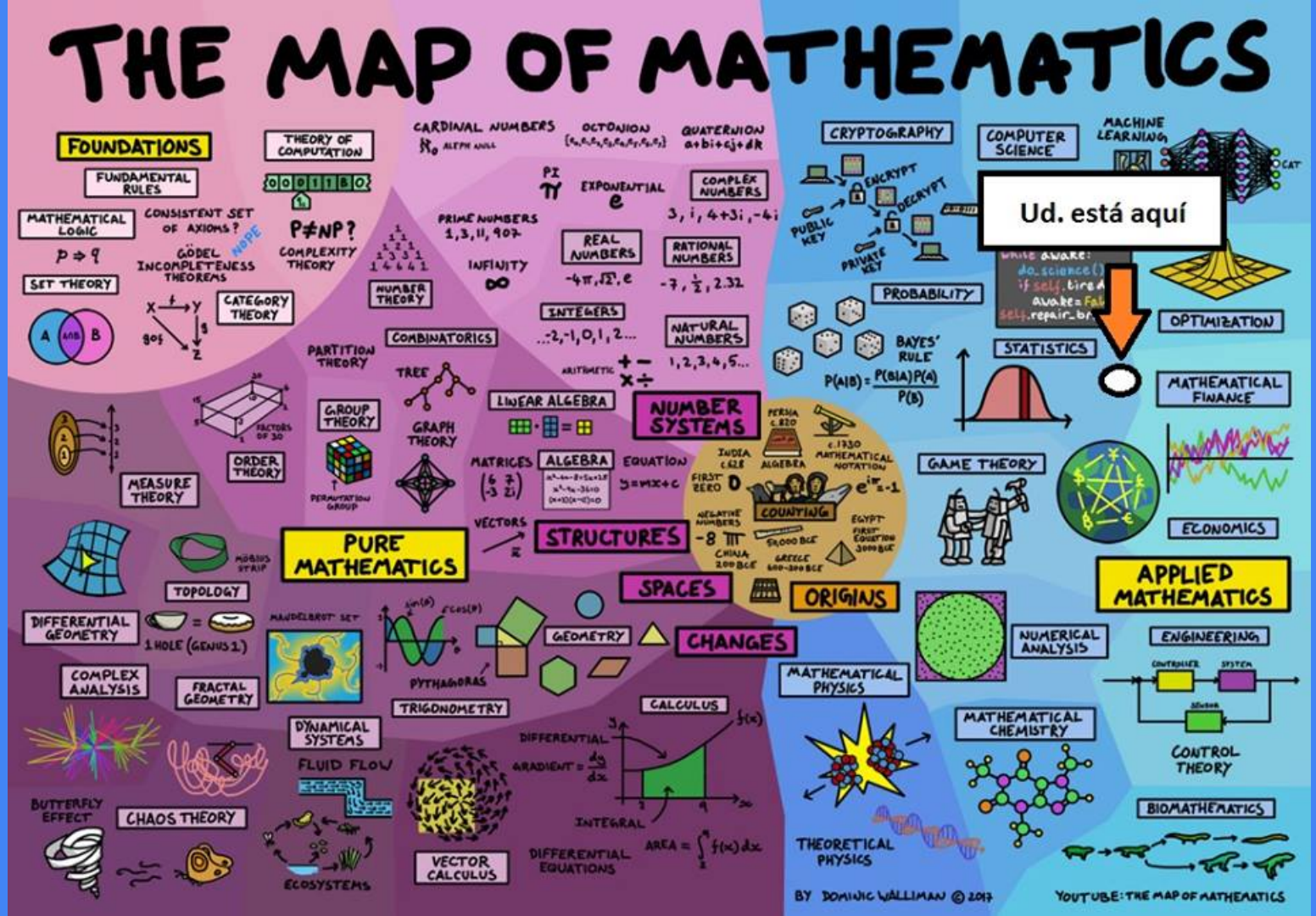
Algunas herramientas estadísticas para el estudio de la Calidad del aire

Sol Represa
13/06/2017



Domain of Science
https://youtu.be/0mJ-4B-mS-Y

¿ Para qué sirve la estadística?

- Para descubrir patrones -> Estadística descriptiva o deductiva

- Para confirmar una hipótesis -> Estadística inferencial o inductiva

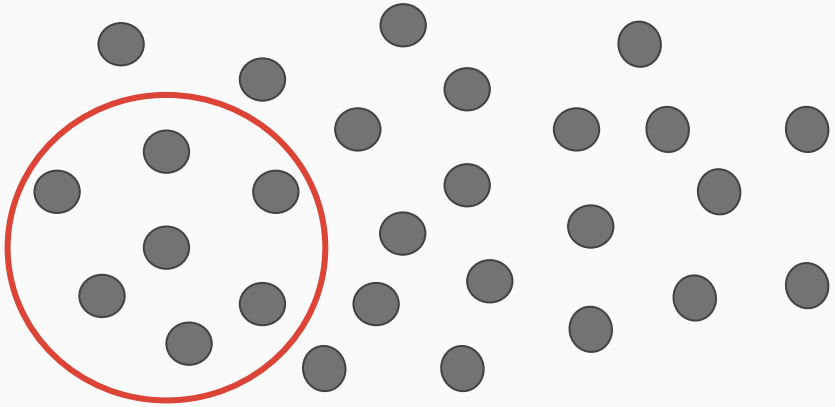
Estadística descriptiva

Utiliza **técnicas de análisis exploratorio** que nos permiten conocer los datos.

- Medidas de tendencia central o posición
- Medidas de dispersión
- Medidas de distribución

Caso de estudio 1

Quiero conocer las características de una población
¿Qué hago?



Población
μ, σ, σ^2

Muestra
$\overline{x}, s, s^2$

N=20

76, 42, 90, 30, 56, 73, 69, 15, 47, 76,
11, 46, 70, 43, 67, 58, 50, 65, 58, 25

Obtenemos los estadísticos

N=20

76, 42, 90, 30, 56, 73, 69, 15, 47, 76,
11, 46, 70, 43, 67, 58, 50, 65, 58, 25

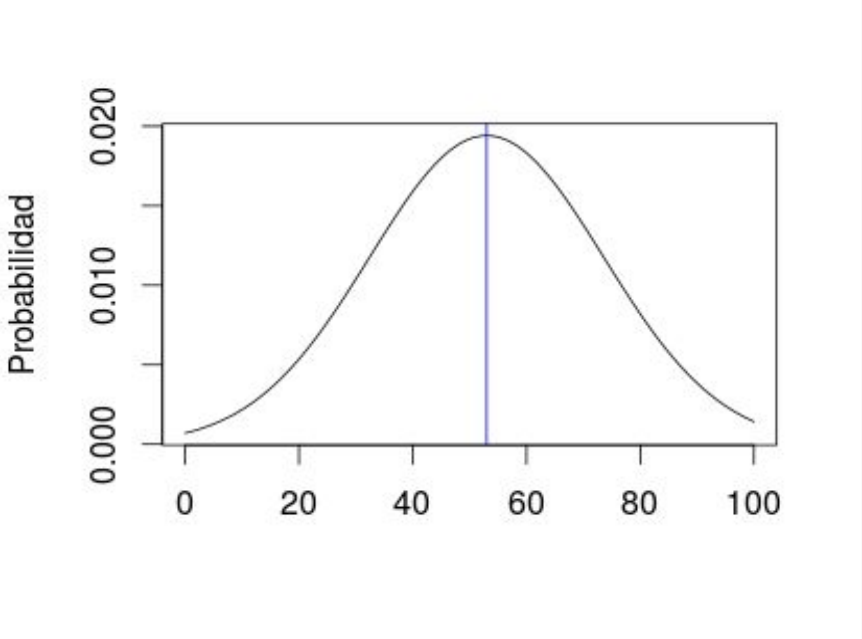
Mediana = 57
Media = 53.35
S = 21.3

De la población?

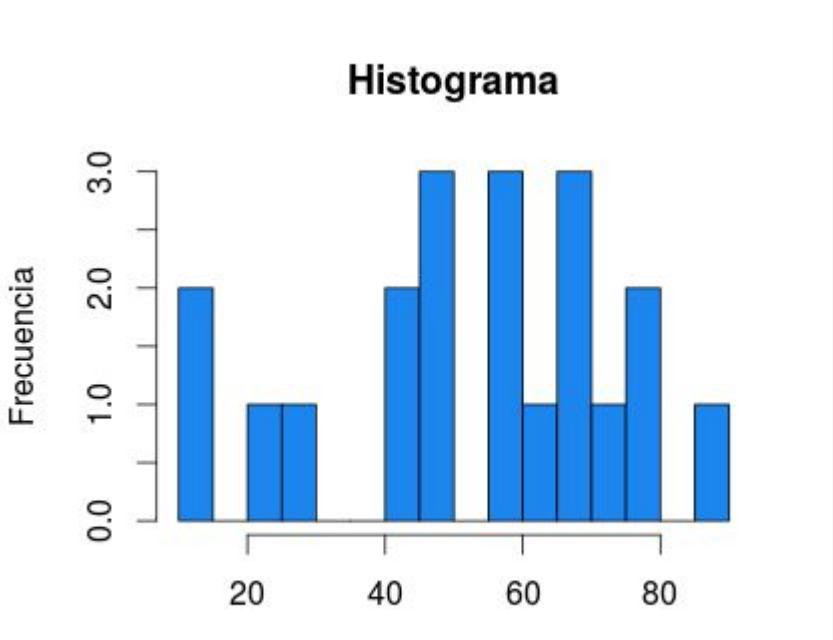
$$P(\bar{x} - T_{(n-1);\,1-\alpha/2}.\sqrt{(S^2/n)} \leq \bar{x} - \mu \leq \bar{x} + T_{(n-1);\,1-\alpha/2}.\sqrt{(S^2/n)}\ ) = (1-\alpha)$$

$$T_{(n-1);\,1-\alpha/2}=2.09 \qquad \alpha = 0.05$$

Expectativa

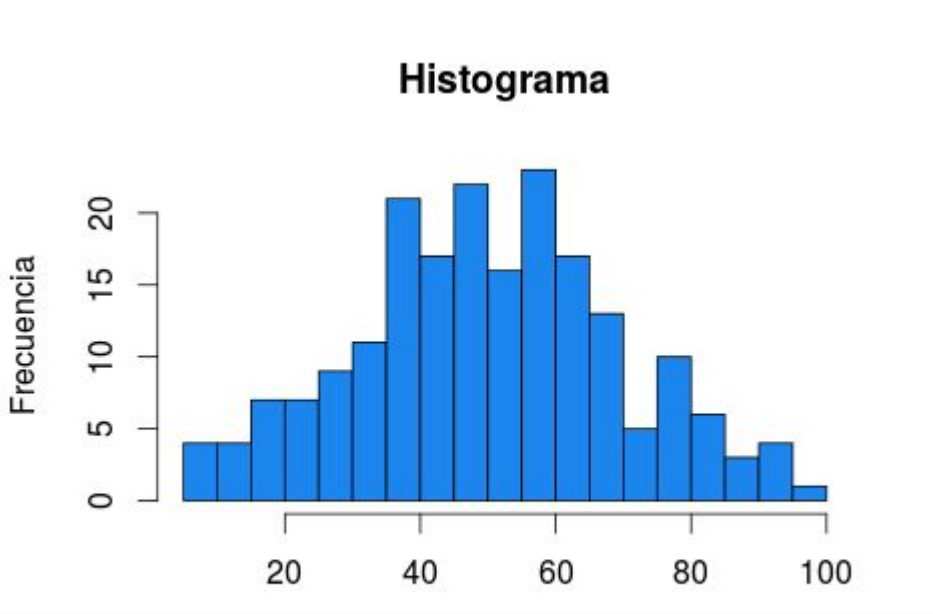


Realidad



N=200

81 57 54 62 59 44 58 47 62 70 70 16 40
35 46 59 36 95 10 51 12 63 24 71 37 48
71 48 66 77 90 36 44 25 21 45 84 14 64
51 46 65 47 38 53 15 30 39 62 56 71 17
58 32 89 52 100 40 59 47 51 50 78 28 8
41 46 43 31 77 21 61 63 46 43 50 35 49
44 62 78 34 55 68 24 44 43 61 37 68 82
78 32 91 69 76 84 44 82 73 57 54 40 44
61 40 29 8 39 10 47 47 38 69 59 78 59 54
69 48 45 19 31 29 17 79 25 38 73 67 53
39 65 64 78 83 51 64 58 49 55 53 57 20
57 58 42 70 47 60 59 37 37 58 48 27 59
50 93 21 95 50 30 58 62 52 51 42 55 67
42 37 68 67 30 39 20 32 37 60 43 38 30
31 46 77 47 58 31 88 41 61 57 35 20 14
27 38 61 60



¿Con qué N caracterizo a mi población?

...o de cuánto tiene que ser mi N para tener una amplitud del intervalo de confianza A?

$$LI = \bar{X} - T_{(n-1);\,1-\alpha/2} \cdot \sqrt{(S^2/n)}$$

$$LS = \bar{X} + T_{(n-1);\,1-\alpha/2} \cdot \sqrt{(S^2/n)}\ )$$

$$A = LS - LI$$

Amplitud del intervalo de confianza

$S = 21.3$

$T_{(n-1);\,1-\alpha/2} = 2.09$

Reemplazo y despejo n:

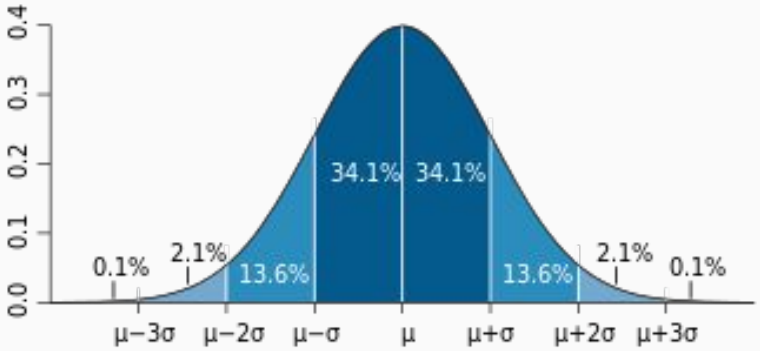
$$n \geq (2 \,.\, T_{(n-1);\,1-\alpha/2} \,.\, S / A)^2$$

La importancia de conocer la distribución de probabilidad

¿La distribución es normal?

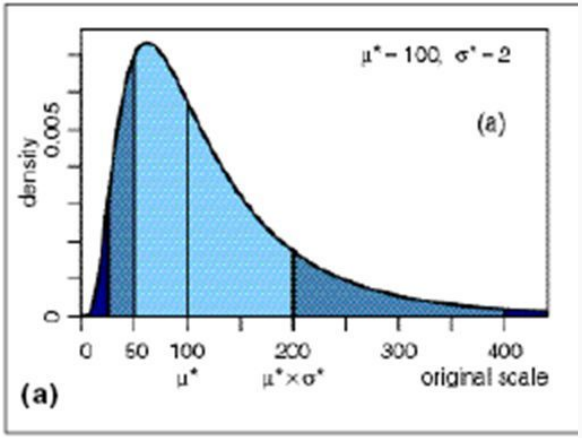
(Test Shapiro-Wilk)

Sí



No

Ej. LogNormal

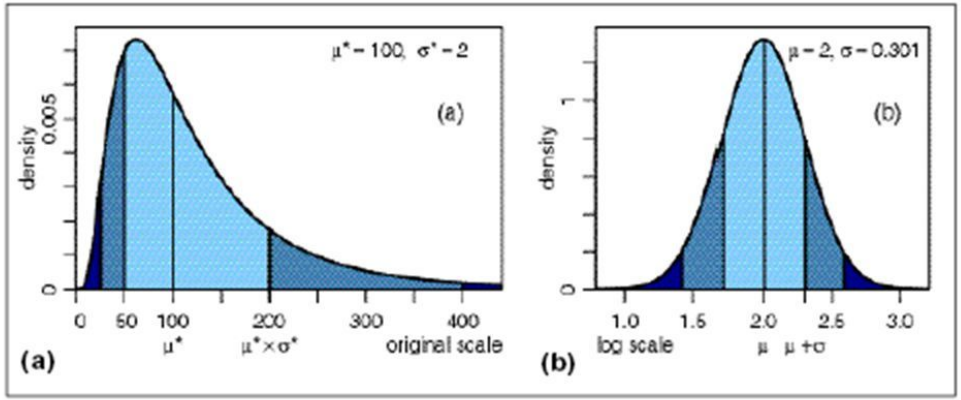


La importancia de conocer la distribución de probabilidad

¿Puedo transformarla?

Sí

Ej. LogNormal



Sin transformación

Con transformación

No

↓

Métodos no paramétricos

Si tenemos valores extremos, qué hacemos?

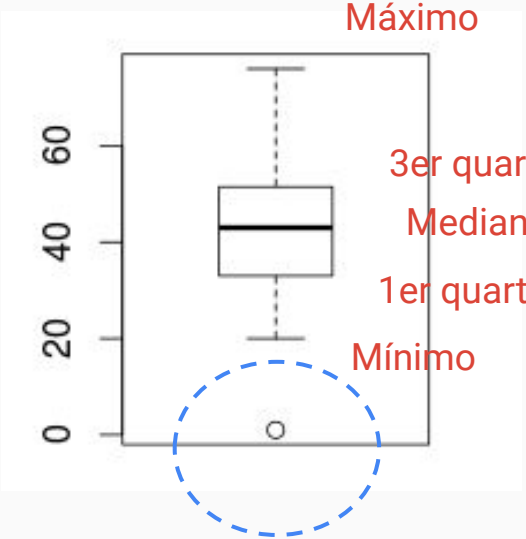
La importancia de conocer la distribución de probabilidad

Tenemos valores extremos, qué hacemos?

¿La distribución es normal?

Sí

Los analizamos para ver si podemos descartarlos.



No

¡Los aguantamos!

Estadísticos robustos

Valores extremos (Outliers)

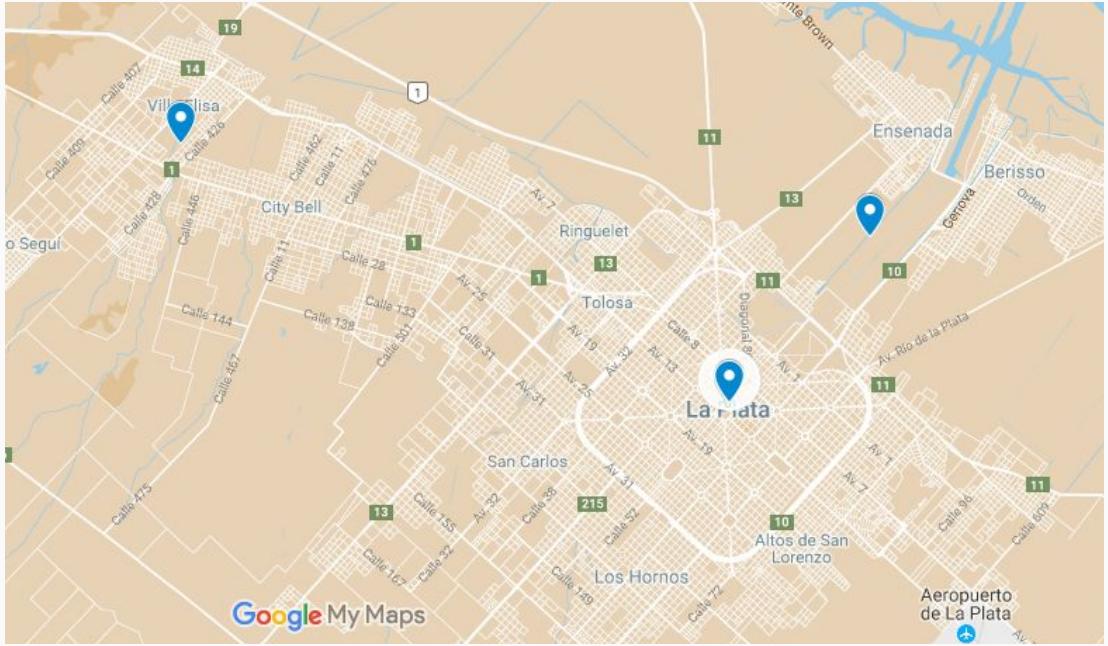
Tamaño de muestra	Test	¿Asume normalidad?	¿Múltiples Outliers?
n ≤ 25	Test de valores extremos	Sí	No / Sí
n ≤ 50	Test de discordancia	Sí	No
n ≥ 25	Test de Rosner	Sí	Sí
n ≥ 50	Test de Walsh	No	Sí

Fuente: "Guidance for Data Quality Assessment. Practical Methods for Data Analysis. EPA QA-G9"
https://www.epa.gov/sites/production/files/2015-06/documents/g9-final.pdf

Caso de estudio 2

Queremos determinar si existen diferencias en las mediciones de 3 sitios:

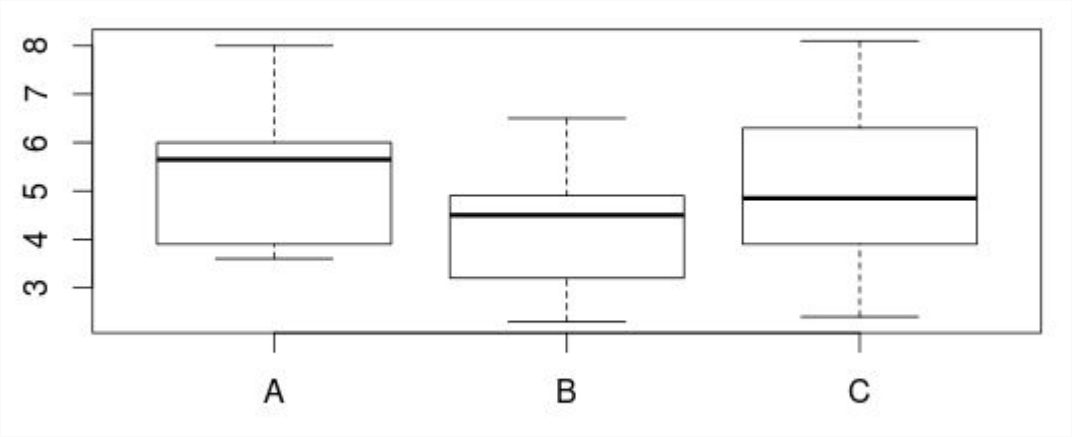
¿Cómo procesamos los datos?



Comparamos distintos puntos

A: 3.8, 6.8, 8.0, 3.6, 3.9, 5.9, 6.0, 5.7, 5.6, 4.5
B: 4.2, 4.8, 4.8, 2.3, 6.5, 4.9, 3.6, 2.4, 3.2, 4.9
C: 3.9, 4.5, 8.1, 5.7, 3.6, 2.4, 6.3, 4.6, 5.1, 7.2

¿La distribución es normal?



$H_0: \mu_1 = \mu_2$ $H_A: \mu \neq \mu$

¿Conozco la desviación de la población?

Sí

$$Z_{prueba} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

$$t_{prueba} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

No

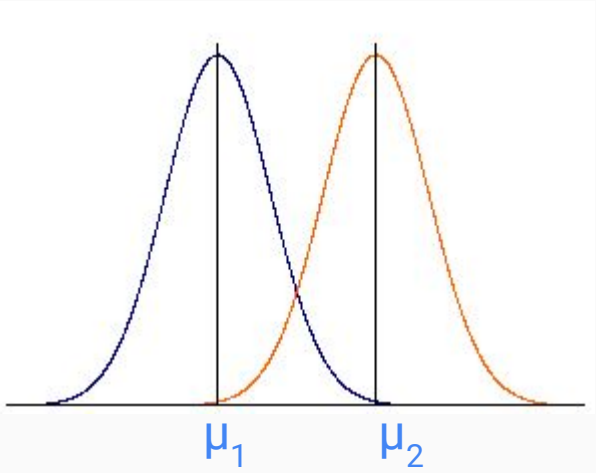
Comparación de poblaciones normales

A: 3.8, 6.8, 8.0, 3.6, 3.9, 5.9, 6.0, 5.7, 5.6, 4.5
B: 4.2, 4.8, 4.8, 2.3, 6.5, 4.9, 3.6, 2.4, 3.2, 4.9
C: 3.9, 4.5, 8.1, 5.7, 3.6, 2.4, 6.3, 4.6, 5.1, 7.2

$T_{(n-1);\ 1-\alpha/2}=2.31$ $\alpha=0.05$

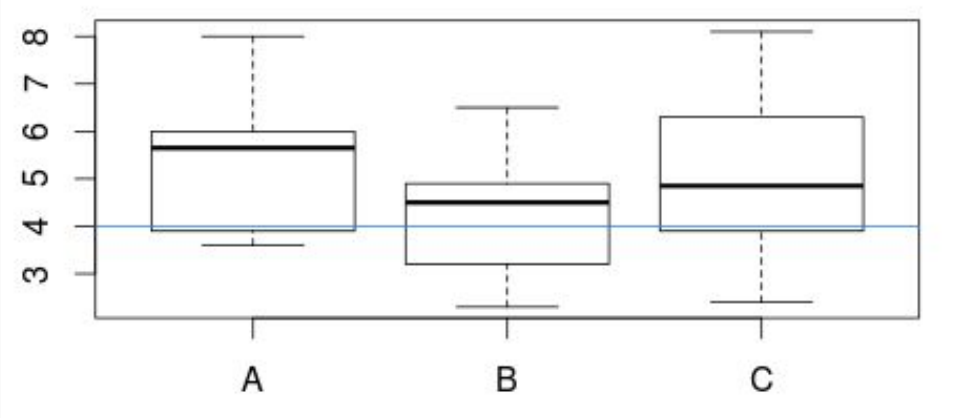
A	B	C
Median :5.650	Median :4.500	Median :4.85
Mean :5.380	Mean :4.160	Mean :5.14
Sd: 1.43	Sd: 1.30	Sd: 1.73

Si t_{prueba} > T entonces hay pruebas suficientes para descartar H_0



.. si hubiese tenido n distintos:

$$t_{prueba}=\frac{\bar{x}_1-\bar{x}_2}{\sqrt{\left[\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}\right]\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$



$H_0: \mu \leq a$ $H_A: \mu > a$

A	B	C
Median :5.650	Median :4.500	Median :4.85
Mean :5.380	Mean :4.160	Mean :5.14
Sd: 1.43	Sd: 1.30	Sd: 1.73

$T_{(n-1);\ 1-\alpha/2}=2.31$ $\alpha = 0.05$

Si $t_{prueba} > T$ entonces hay pruebas suficientes para descartar H_0

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

Análisis (de series) temporales

¿Qué es una serie de tiempo?

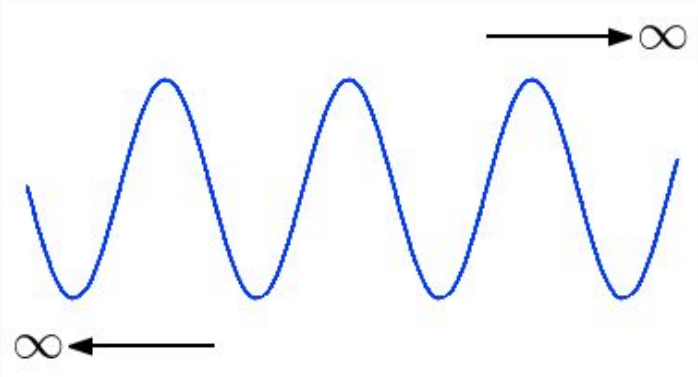
Es una secuencia de datos medidos en determinados momentos x-espaciados y ordenados cronológicamente.

¿Qué NO es una serie de tiempo?

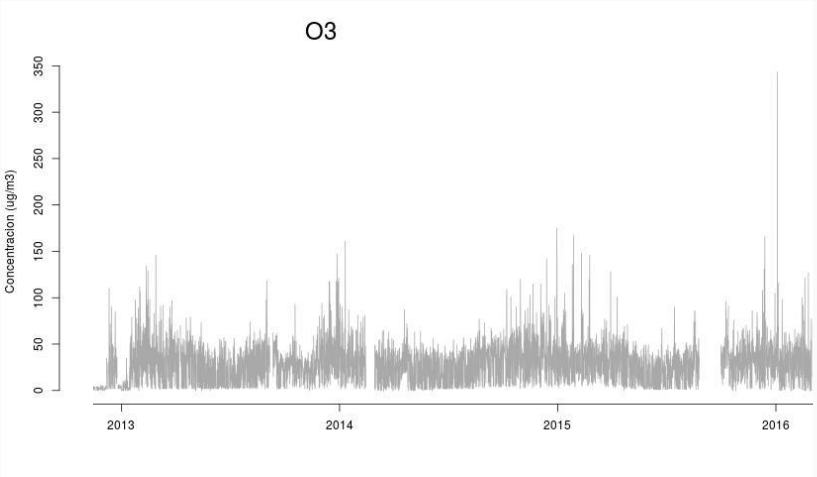
Un conjunto de datos medidos en distintos momentos y que no se encuentran ordenados cronológicamente.

Análisis de series de tiempo

Expectativa



Realidad



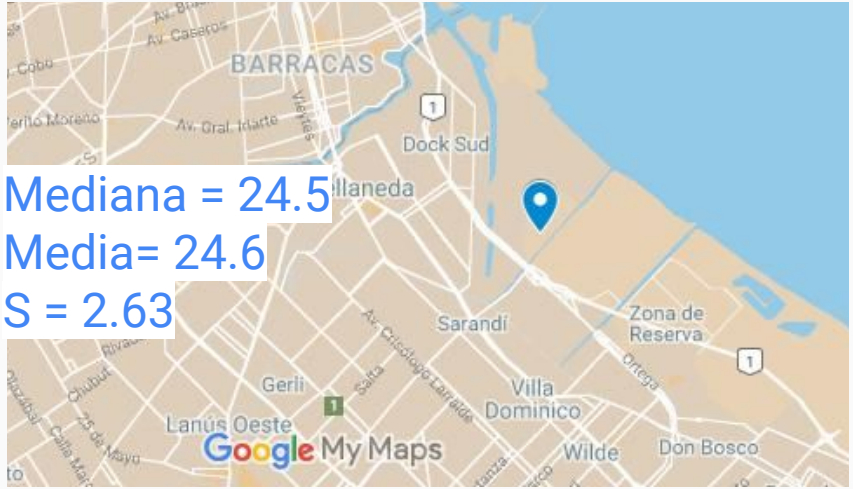
Un análisis temporal
no es un análisis longitudinal

Caso de estudio 3

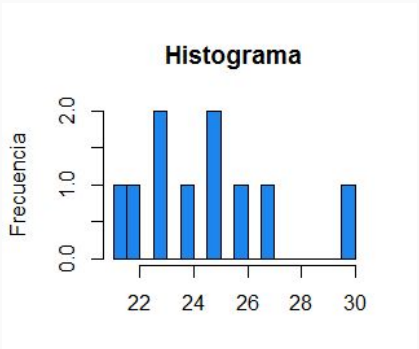
Tengo un sistema de monitoreo contínuo que me genera un dato por hora...

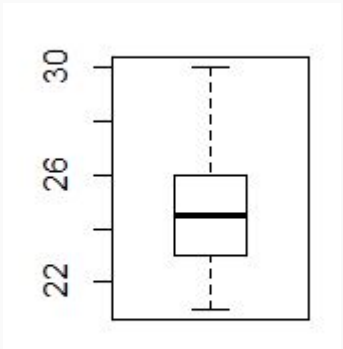
¿Qué me dicen los datos?

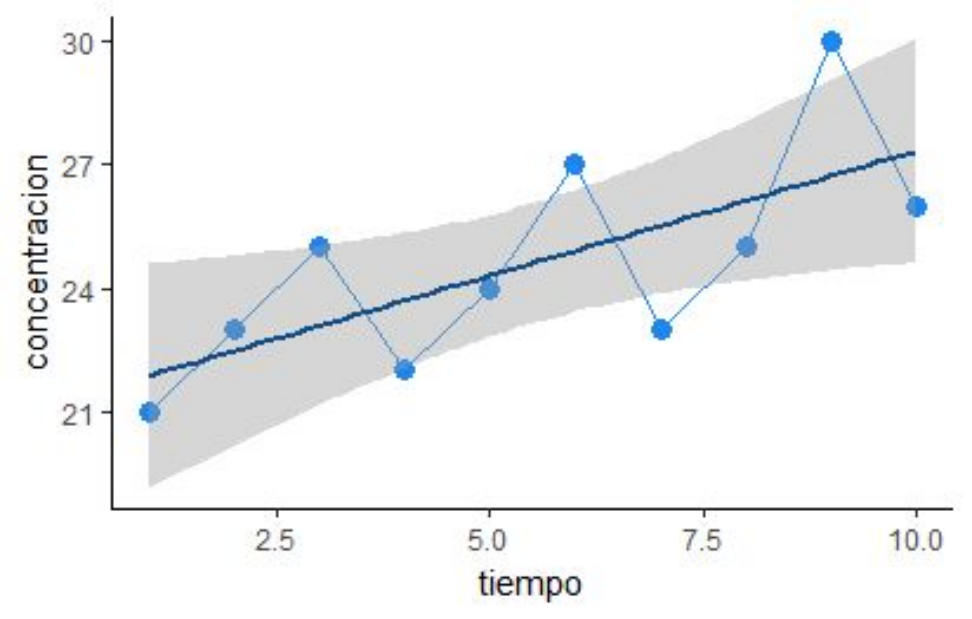
t	[C]
1	21
2	23
3	25
4	22
5	24
6	27
7	23
8	25
9	30
10	26



Mediana = 24.5
Media= 24.6
S = 2.63







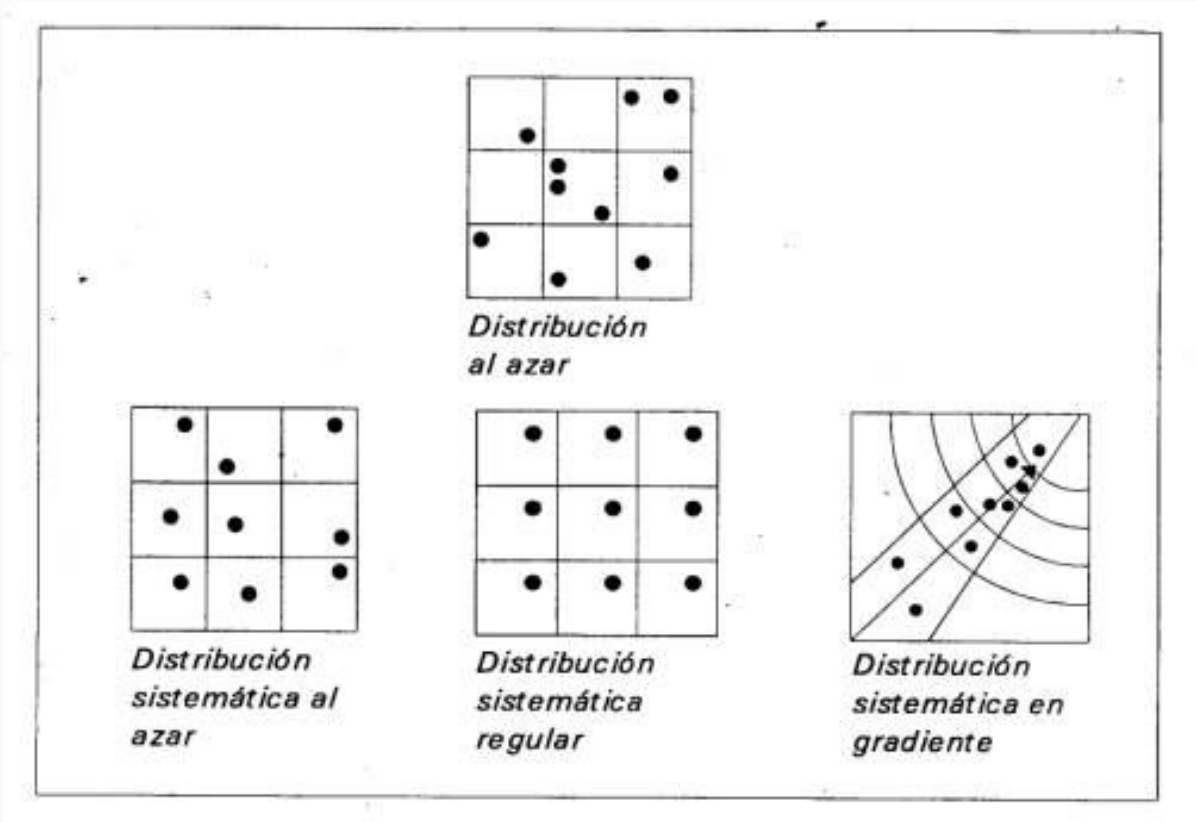
- Tendencia
- Estacionalidad
- Autocorrelación (o “la memoria de los datos”)

Análisis espacial

Muchas de las variables ambientales tienen una dimensión espacial.

Es decir, varían en el espacio.

Tipos de muestreo



Análisis espacial

- Métodos determinísticos:

  Modelos de dispersión, distribución y transporte

- Métodos estocásticos:

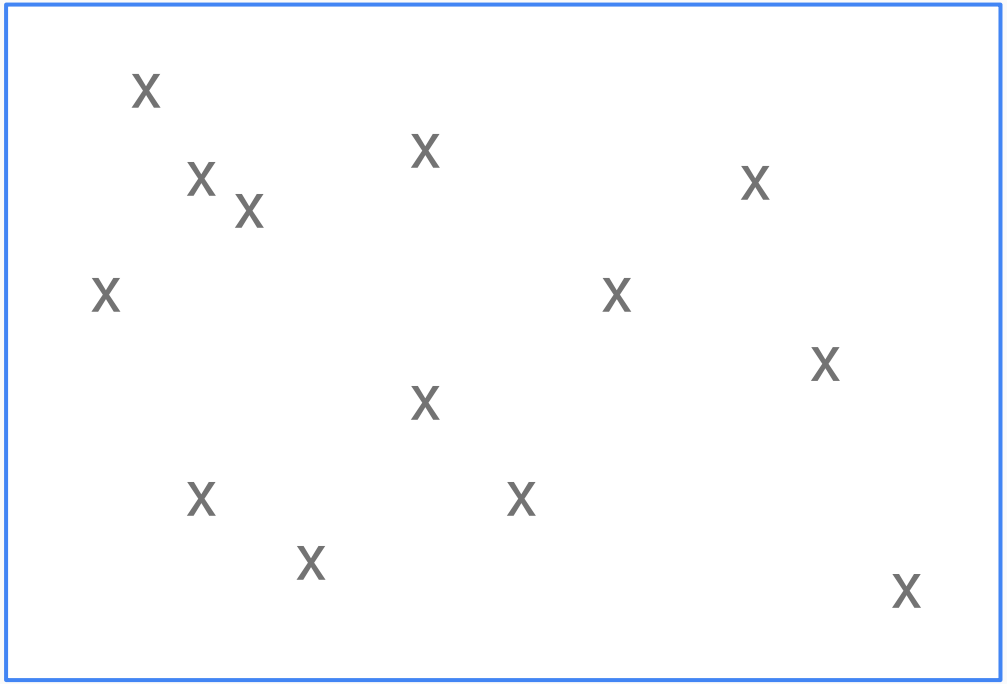
  Geoestadística

  - Métodos de interpolación espacial
  - Métodos de simulación espacial
  - Modelos de regresión múltiple

Caso de estudio 4

Queremos describir el comportamiento de una variable en una región y para ellos diseñamos un muestreo en distintos puntos.

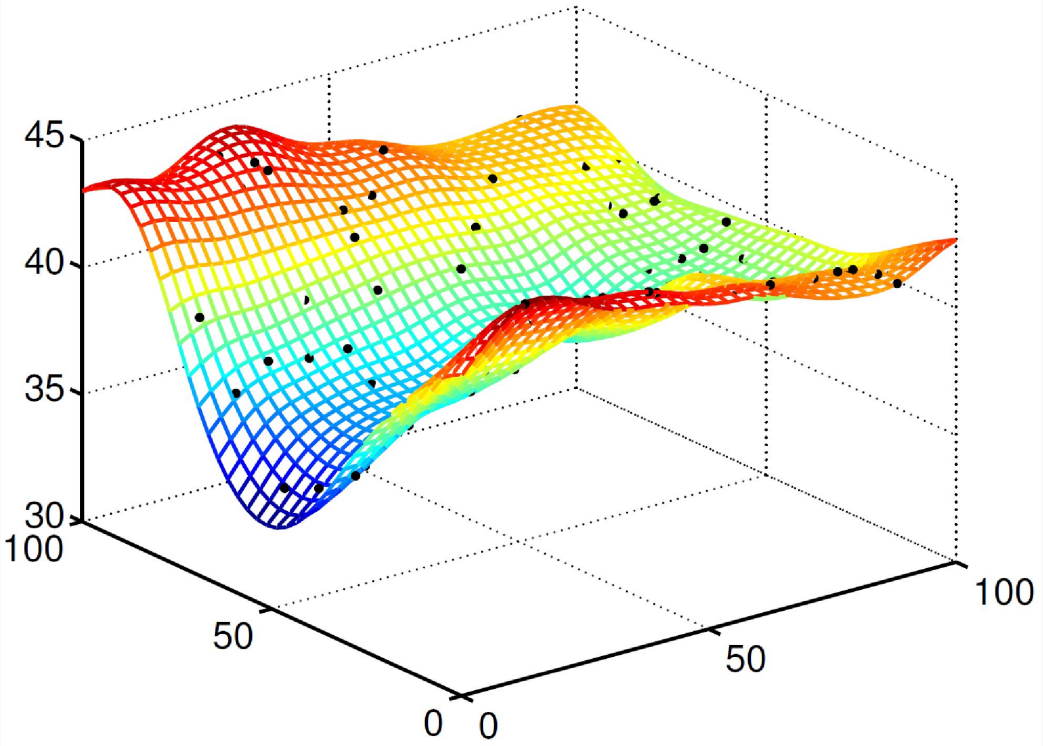
Medimos.. y ahora?



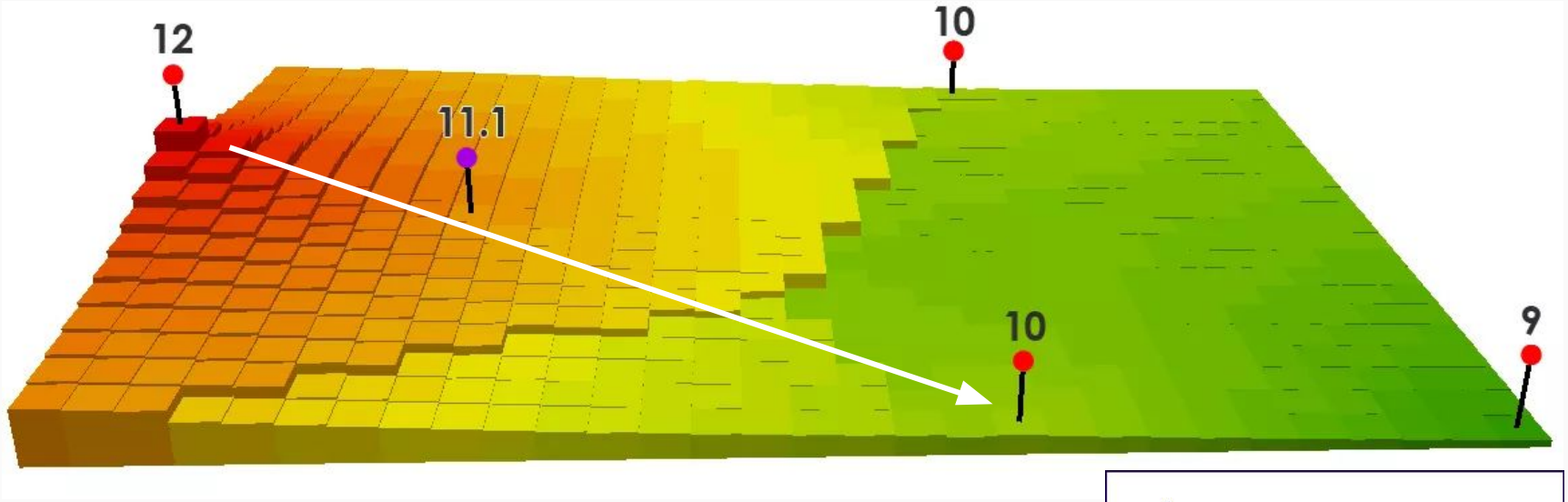
..es posible conocer lo que sucede entre los puntos medidos?

Caso 4

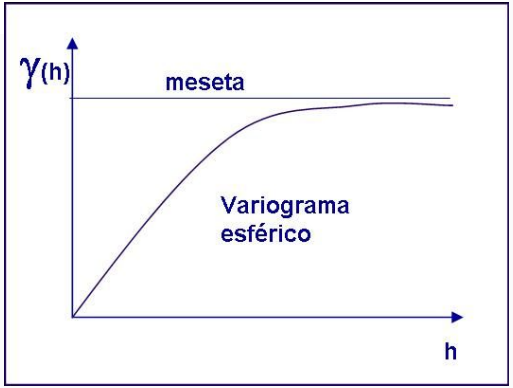
Queremos describir el comportamiento de una variable en una región y para ellos diseñamos un monitoreo en distintos puntos. Medimos.. y ahora?



Análisis espacial

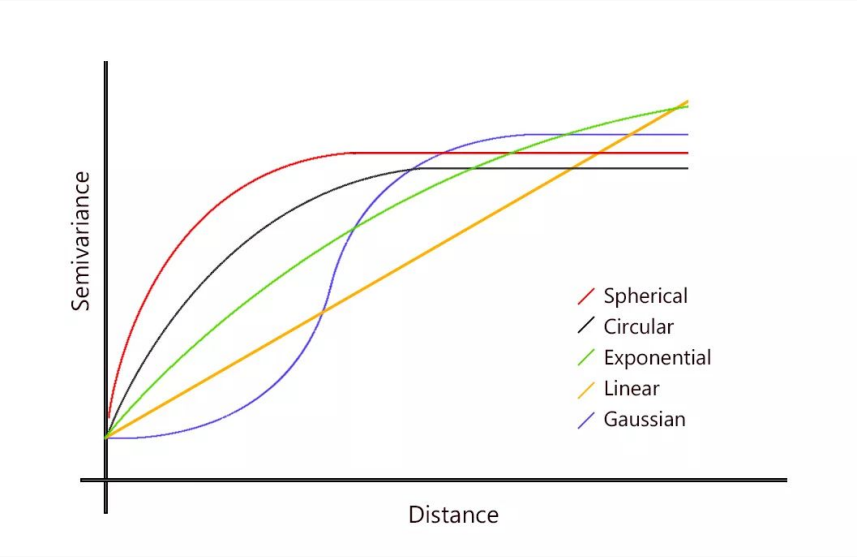


A medida que me alejo en x,
me alejo también en z

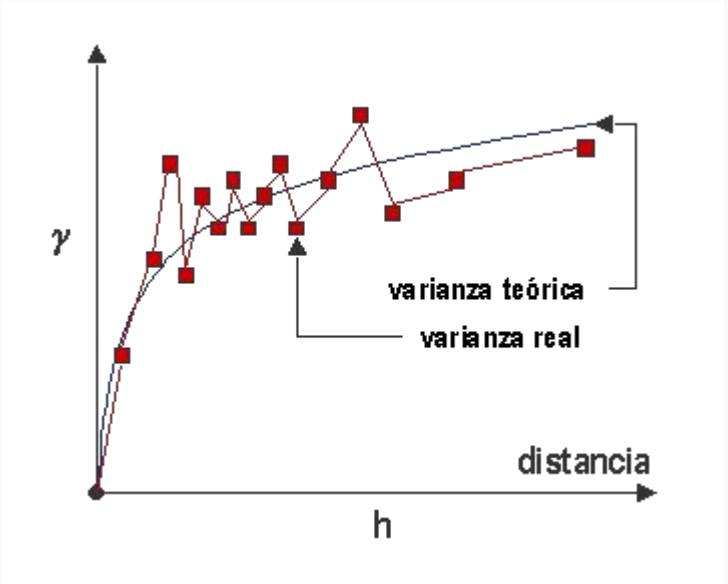


Análisis espacial

Expectativa



Realidad

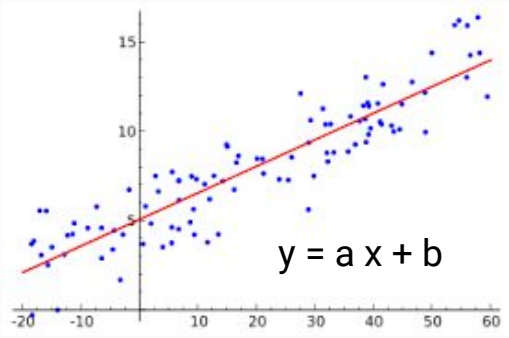


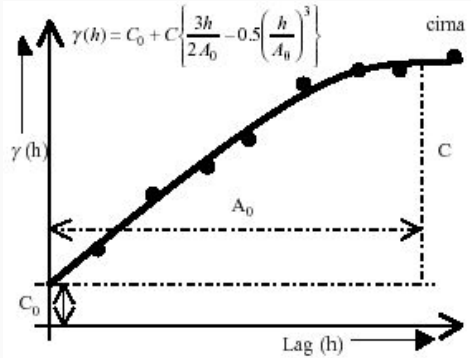
Semivariogramas teóricos

Análisis espacial

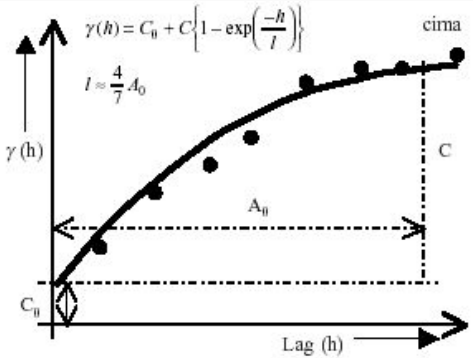
Así como en la regresión lineal generamos la mejor recta;

en el **kriging** generamos la mejor curva a partir del **semivariograma** teórico.

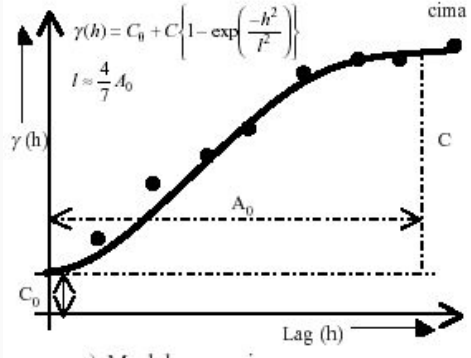




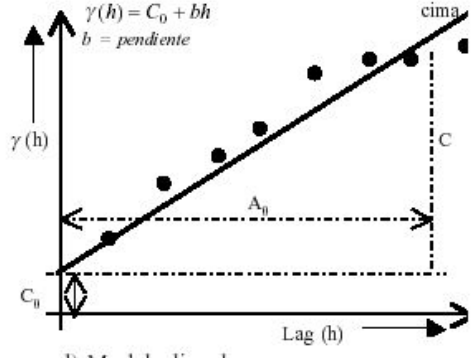
a) Modelo esférico



b) Modelo exponencial

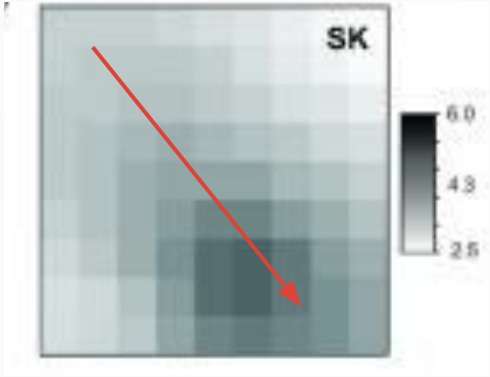


c) Modelo gaussiano



d) Modelo lineal

Análisis espacial



Si varía de forma distinta espacialmente?

-> Lo corrijo!

Cómo puedo evaluar el ajuste del modelo? -> Midiendo el error!

$$MPE = \frac{1}{n}\sum_{j=1}^{n}(Z_{oi} - Z_{pi})$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(Z_{oi} - Z_{pi})^2}$$

$$R_I = \frac{RMSE_R - RMSE_{RK}}{RMSE_R} \times 100\%$$

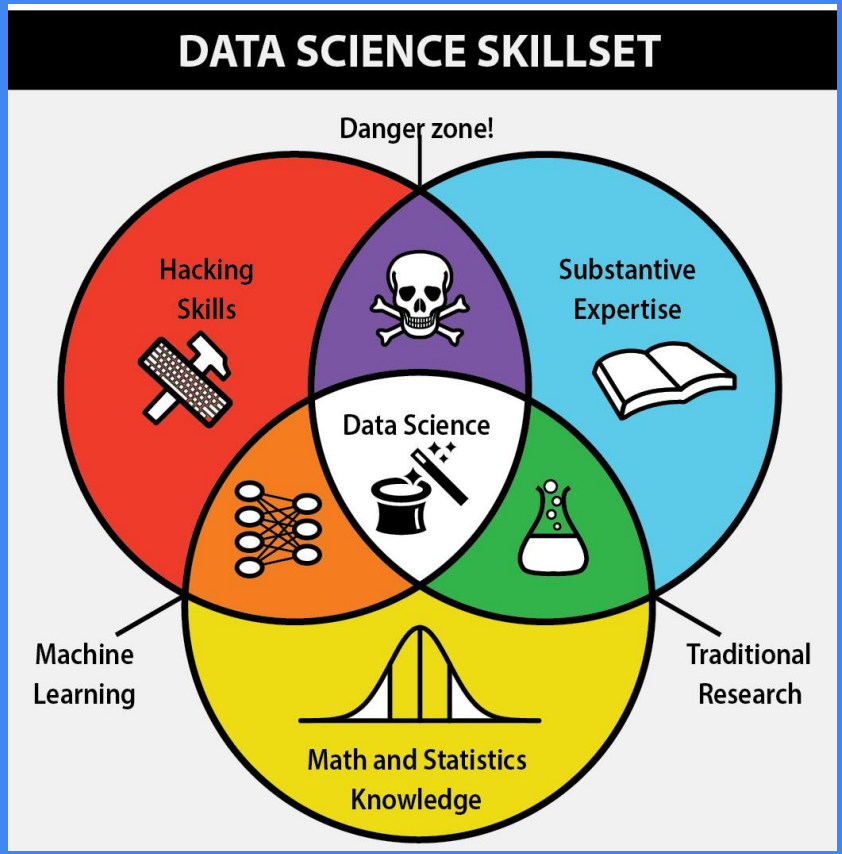
No todo es

describir

También podemos

predecir

"Todo lo relacionado con la ciencia está cambiando debido al impacto de la tecnología de la información y el diluvio de datos"
Jim Gray



Design: Natalia Bilenko, modified from Drew Conway;
Book: MTchemik; network: Qwertyus.

sol.represa@gmail.com

Centro de Investigaciones del Medio Ambiente
CONICET - UNLP