ANÁLISIS GEOESTADÍSTICO

- Origen de la "Geoestadística"
- Geoestadística: definición y objeto
- Datos geográficos y análisis estadístico
- Conceptos básicos de Estadística
- Técnicas básicas de Estadística para el Análisis Exploratorio de Datos

Concepción González García (2008)

Imagen de la NASA

Origen de la "Geoestadística"

Geoestadística (i)

- La Geoestadística tiene su origen en la búsqueda, exploración y evaluación de yacimientos minerales útiles.

- Se ha consolidado y desarrollado en los últimos 30 años como ciencia aplicada casi exclusivamente en el campo minero.

- La gran diversidad de formas en que se presentan los datos ha llevado a la utilización de técnicas matemáticas y estadísticas para resolver un único problema: ***estimar valores desconocidos a partir de los conocidos, para la estimación y caracterización de los recursos y reservas.***

Origen de la "Geoestadística"

Geoestadística (ii)

- Las investigaciones han buscado los métodos más eficientes que proporcionen la mayor información posible de los datos disponibles.

- Mediante el mejor estimador que minimice la varianza del error de estimación (error cuadrático medio) surge la Geoestadística por los trabajos de G. Matheron en la Escuela Superior de Minas de París (1949)

- Entre los métodos más recientes se pueden citar los "geomatemáticos": El Inverso de la Distancia, Triangulación, Splines, etc.

Origen de la "Geoestadística"

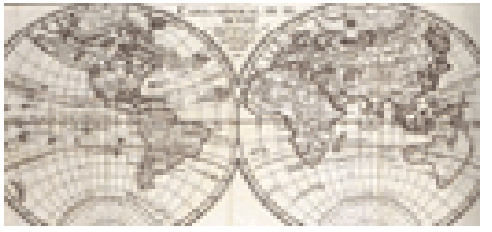
Geoestadística (antecedentes)

- Sichel (1947), 1949) observó la naturaleza asimétrica de la distribución del contenido de oro en las minas surafricanas, la equiparó a una distribución de probabilidad lognormal y desarrolló las fórmulas básicas para esta distribución.

D.G. Krige (1951) desarrolló la aplicación del análisis de regresión entre muestras y bloques de mena (Mineral metalífero, principalmente el de hierro, tal como se extrae del criadero y antes de limpiarlo).

De la minería, las técnicas geoestadísticas, se han exportado a más campos como la hidrología, física del suelo, ciencias de la tierra y más recientemente a la gestión ambiental y al procesado de imágenes de satélite.

Geoestadística :
Definición y Objeto (i)



✓La geoestadística es una rama de la estadística que trata fenómenos espaciales (Journel & Huijbregts, 1978).

✓Su interés primordial es la estimación, predicción y simulación de dichos fenómenos (Myers, 1987).

- Se reconoce como una rama de la estadística tradicional, que parte de la observación de que la variabilidad o continuidad espacial de las variables distribuidas en el espacio tienen una estructura particular que se estudia mediante las depencias entre ellas.

Geoestadística :
Definición y Objeto (ii)

› Matheron (1970) denominó a estas variables dependientes entre si, **variables regionalizadas**, además de elaborar su teoría. [Journel y Huijbregts (1978), David (1977) y de Fouquet (1996)].

▪En resumen, la aplicación de la teoría de los procesos estocásticos a los problemas de evaluación de reservas de distintos tipos de materias primas minerales y en general a las ciencias naturales en el análisis de datos distribuidos espacial y temporalmente dio origen a lo que hoy se conoce como Geoestadística.

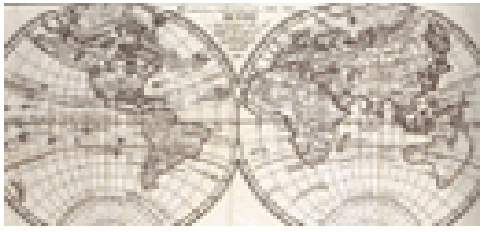
Datos geográficos y análisis estadístico

Los SIG actuales incluyen posibilidades de exploración y análisis de datos.

Las técnicas más elementales son de Estadística descriptiva (Análisis Exploratorio de Datos, EDA).

La Estadística Descriptiva: para una, dos y hasta 3 variables, permite resumir conjuntos de valores y visualizar estructuras de distribuciones de probabilidad.

Datos geográficos y análisis estadístico



Características de los datos geográficos:
en un punto, además de sus coordenadas, se dispone de información "multivariante" (altitud, precipitación, profundidad del suelo, tipo de vegetación,…)

El denominado **Análisis exploratorio espacial de datos** (ESDA), es una ampliación y desarrollo del EDA.
El ESDA incluye, junto a técnicas exploratorias, muchas ideas tomadas del Análisis espacial o Estadística espacial.

Datos geográficos y análisis estadístico

•Existen algunas dificultades fundamentales para que las técnicas estadísticas convencionales manejen correctamente datos geográficos:

•El empleo de las Técnicas clásicas de Inferencia Estadística, suponen, en los datos de partida :

> la independencia de las observaciones

> la distribución en curva de Gauss (distribución Normal)

lo cual a menudo no se cumple en datos geográficos.

Conceptos básicos de Estadística

Revisión de Técnicas estadísticas

- Muestreo y análisis Exploratorio de datos

- **Conceptos de Inferencia Estadística** paramétrica:
  Una variable: Estimador, propiedades, intervalos de confianza y tests de hipótesis.
  Dos ó más variables: modelos lineales (regresión, Análisis de la varianza)

- **Conceptos de procesos estocásticos** (variables dependientes, medidas de dependencia espacial)

Conceptos básicos de Estadística

Muestreo y análisis Exploratorio de datos

- Población (Universo) y Muestra.
- Muestreo (Obtención de datos)
- Variables y tipos

- Antes de comenzar un estudio geoestadístico se deben discutir todos los elementos que aporten conocimientos del problema a resolver, fenómeno en estudio, organización y verificación de la información disponible y finalmente realizar el análisis exploratorio de los datos.

Conceptos básicos de Estadística

- Población estadística o universo es el conjunto de referencia sobre el cual van a recaer las observaciones.

- Muestra: es el subconjunto de la población en el que se mide una o más variables de interés.

-a partir de este subconjunto se obtienen conclusiones sobre las características de la población.

- la muestra debe ser representativa, en el sentido de que las conclusiones obtenidas deben servir para el total de la población.

Unidad muestral: elementos de la población, no solapados en los que se mide. Cada elemento de la población pertenecerá a una y sólo una unidad muestral.

Estadística básica

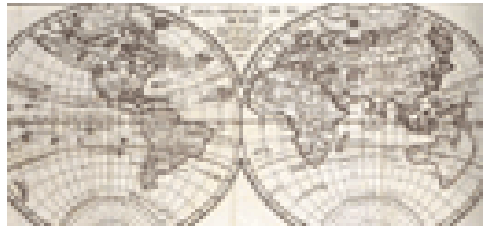
Tipos de muestras

- Muestra probabilística: se elige mediante ciertas reglas, de manera que la probabilidad de selección de cada unidad es conocida de antemano.

- Muestra no probabilística: no se rige por las reglas matemáticas de la probabilidad.

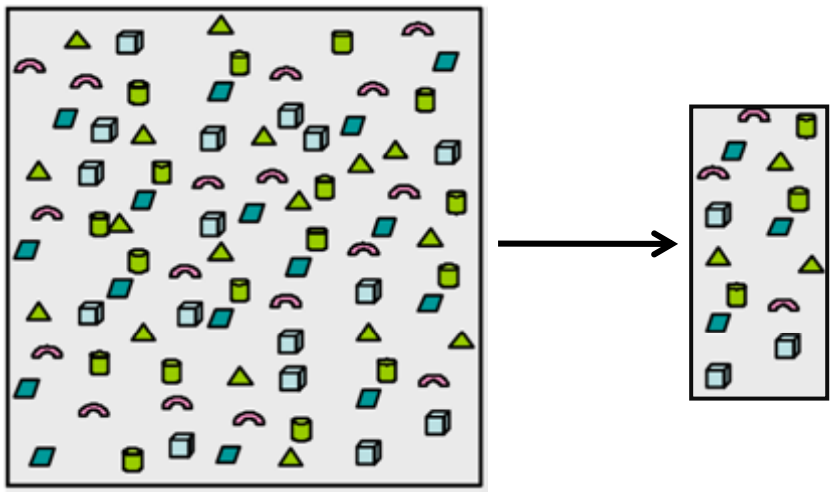
- en las muestras probabilísticas es posible calcular la magnitud del error muestral,
- no es factible hacerlo en el caso de las muestras no probabilísticas (puntos de fácil acceso, estaciones de medición de la calidad del aire en una ciudad)

Estadística básica



Métodos de muestreo

➢ Muestreo aleatorio simple: todos los componentes o unidades de la población tienen la misma probabilidad de ser seleccionados. Es la modalidad más elemental de m. probabilistico.



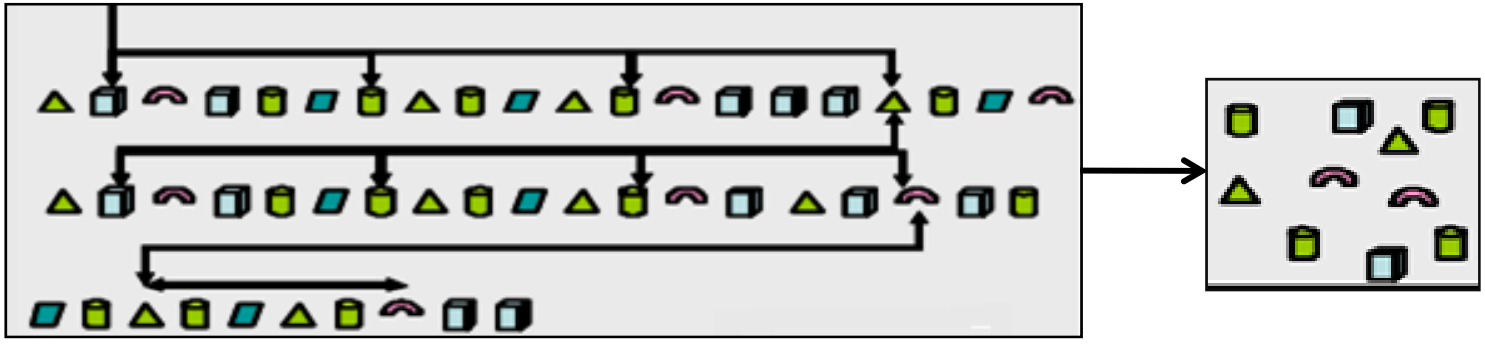
Representación gráfica del muestreo aleatorio simple

Estadística básica

Métodos de muestreo

➢ Muestreo sistemático:

Se selecciona al azar un punto de partida y un intervalo muestral. Así si el punto de partida fuera el 11 y el intervalo el 6 se elegirían el 11, 16, 21, 16 hasta recorrer toda la población.



Representación gráfica del muestreo sistemático

Estadística básica

Métodos de muestreo

- Muestreo estratificado (i):

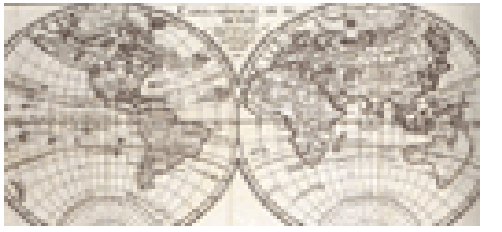
-la población en estudio se sub- divide en estratos o subpoblaciones que tienen cierta homogeneidad en el terreno y en cada estrato se realiza un muestreo aleatorio simple (o sistemático).

-requisito principal para aplicar este método de muestreo: conocimiento previo de información que permita subdividir la población,

Por ejemplo: división que se puede realizar con base en la topografía, los horizontes del suelo, la mancha del contaminante, los cambios de color en el suelo, el crecimiento irregular de las plantas, etc.

Estadística básica

Métodos de muestreo



- Muestreo estratificado (ii):

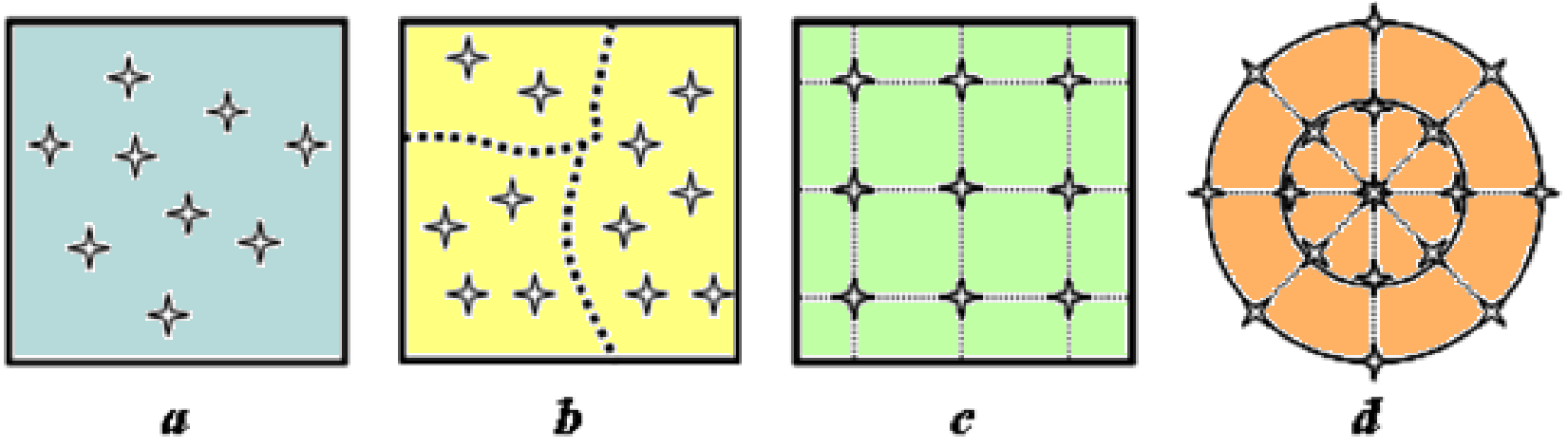
- garantiza que los puntos de muestreo se encuentren repartidos más uniformemente en toda la zona en función del tamaño del estrato;

- permite conocer de forma independiente las características particulares de cada estrato

-recomendable para áreas mayores de diez hectáreas y cuando el terreno no es homogéneo (Mason 1992, Valencia y Hernández 2002).

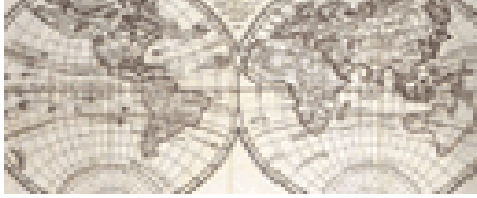
Estadística básica

Métodos de muestreo



Esquemas de tipos de muestreo:

a) aleatorio simple;
b) aleatorio estratificado;
c) sistemático rejilla rectangular;
d) sistemático rejilla polar



Otros Métodos de muestreo

Muestreo por conglomerados

En poblaciones muy extensas, donde la localización y medición de la muestra seleccionada supone grandes desplazamientos se suelen agrupar las unidades elementales en **conglomerados o unidades primarias**

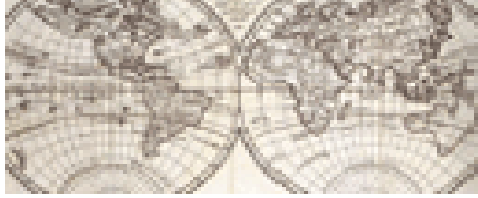
Características del conglomerado:

> Conjunto de unidades muestrales elementales.

> Heterogeneidad de la variable a medir

> El número total de conglomerados en la población es conocido

Estadística básica



Muestreo por conglomerados

Características:

División previa de la población en conglomerados o "áreas convenientes", de las cuales se selecciona un cierto número para la muestra

Ventajas:

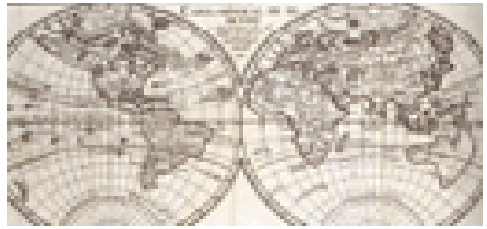
Ahorro de costes y tiempo al efectuar visitas a las unidades seleccionadas.

Disminución de necesidad de desplazamientos al concentrar unidades elementales.

Inconvenientes:

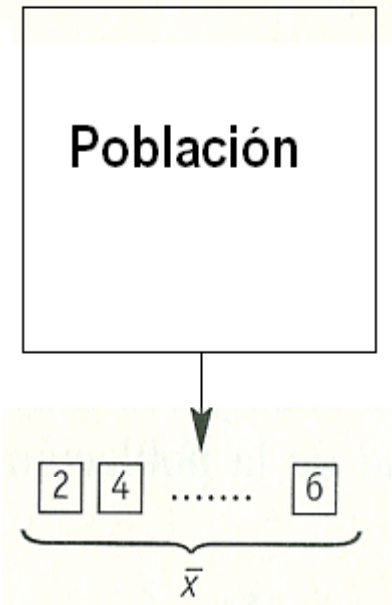
Menor precisión en las estimaciones, sobre todo con conglomerados de gran tamaño

Conceptos de Estadística básica

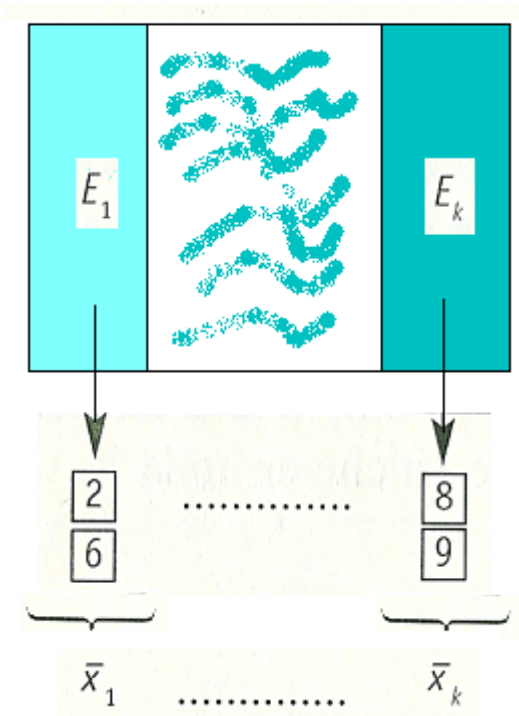


Diferencias entre tipos de muestreo

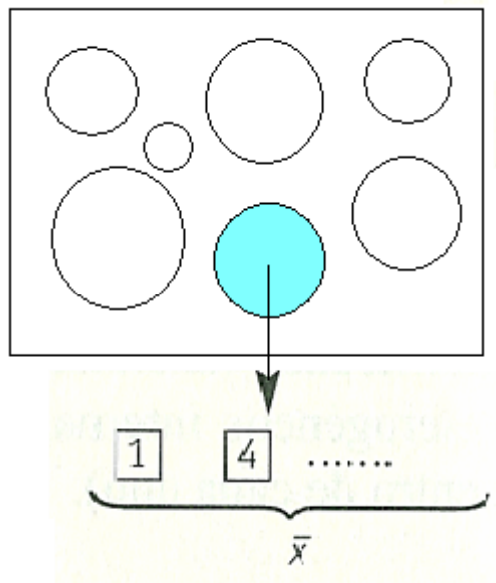
ALEATORIO



ESTRATIFICADO



CONGLOMERADOS



(Adaptado de Peña, 2001

Estadística básica

Variables y Tipos de variables

➢ Variable: cada una de las características de los elementos de una población y que varían de una unidad a otra.

✓Variables cualitativas (o categóricas): aquellas que no tienen medida numérica; se representan por categorías o atributos (tipo de suelo, de vegetación, textura,…).

✓Variables cuantitativas: las que pueden expresarse numéricamente (temperatura, precipitación, profundidad suelo, altitud, pendiente, ….)

Estadística básica

Variables cuantitativas

❖Variables discretas: son el resultado de contar y sólo toman valores enteros (número de puntos, de cuadrículas, de píxeles).

❖ Variables continuas: son el resultado de medir, y pueden contener decimales (temperatura, profundidad, altura). Se pueden subdividir a voluntad. Pueden tomar, entonces, cualquier valor de un determinado intervalo

Estadística básica

Estadística Descriptiva

Objetivo: conocer la información disponible.

Cálculos estadísticos o estadística descriptiva. Permiten determinar si la distribución de los datos es normal, lognormal, o si no se ajustan a una distribución estadística conocida. Implica tener conocimiento de:

- **Número de casos:** representado por "n", es el número de valores muestreados del fenómeno en estudio, los datos representados por xi, i = 1, . . . , n.

Frecuencia de cada x_i nº de veces que aparece el mismo valor medido.

Estadística básica

Distribuciones de frecuencias

Los valores de cada xi medidos y su frecuencia de aparición en los n datos se conoce como la distribución de la variable estudiada.

Valores resumen: Medidas de posición

- **Media:** Es la media aritmética de la distribución,

$$\overline{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Estadística básica

Valores resumen: Medidas de posición

- **Moda:** Es el valor más frecuente de la distribución

- **Mediana:** Es el valor para el cual la mitad de los datos son menores y la otra mitad están por encima de este valor.

La mediana es también llamada percentil 50

Ordenando los datos en orden ascendente podemos calcular la mediana como.

$$M = \begin{cases} X_{(n+1)/2} & \text{si n es impar.} \\ (X_{n/2} + X_{n/2+1})/2 & \text{si n es par.} \end{cases}$$

Estadística básica

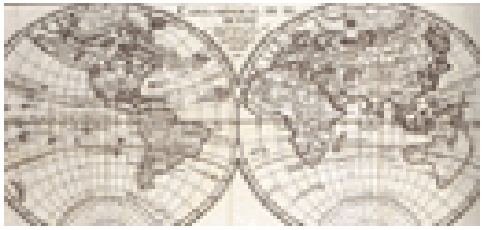
Valores resumen: Medidas de posición

- **Cuartiles,** donde Q1 = percentil 25, Q2 = Mediana y Q3 = percentil 75.

- **Deciles** si los datos se dividen en 10.

De forma general estas medidas se pueden calcular por: [p(n+1)/100] ésima observación de los datos ordenados ascendentemente, donde p es el **percentil** que se desea calcular.

Valores resumen: Medidas de posición



Interpretación de los Cuartiles

Se forman cuatro grupos con igual cantidad de datos

Un cuarto de los datos toman valores iguales o inferiores a 2,2

La mitad de los datos (dos cuartos) toman valores iguales o inferiores a 3,6

Tres cuartos de los datos toman valores iguales o inferiores a 7,6

Q_2= Me=3,6

Q_3=7,6

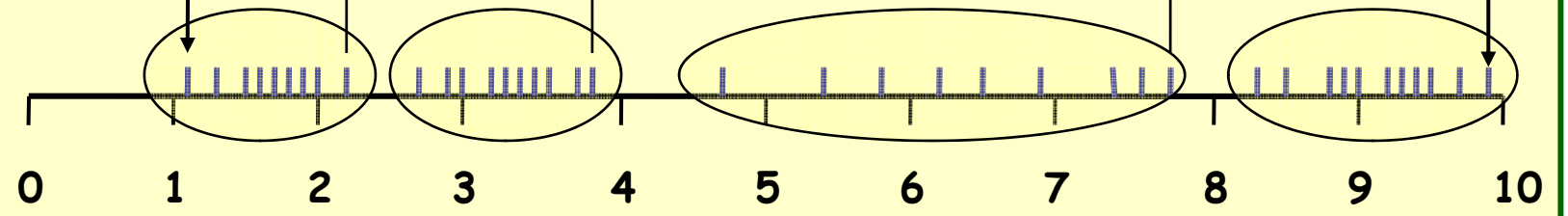
Q_1=2,2

Rango Intercuartil

$RI=Q_3-Q_1$

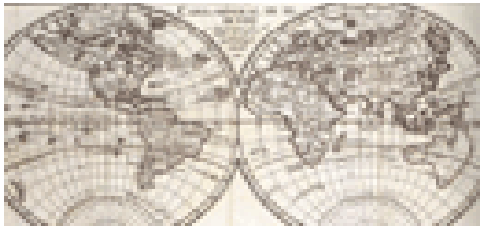
x_{min}=1,1

x_{max}=9,9



Escala graduada de la variable en estudio

Estadística básica



Valores resumen: Dispersión

- **Rango de la distribución:** Es la diferencia entre el valor máximo y el mínimo observados.

- **Varianza:** Describe la variabilidad de la distribución. Es la medida de la desviación o dispersión de la distribución.

$$\sigma^2_{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \overline{x}_n\right)^2$$

Se divide por (n-1) y no por n y se representa por S^2 cuando se calcula con una muestra observada porque proporciona mejor estimación de la varianza de la población. (estimación insesgada)

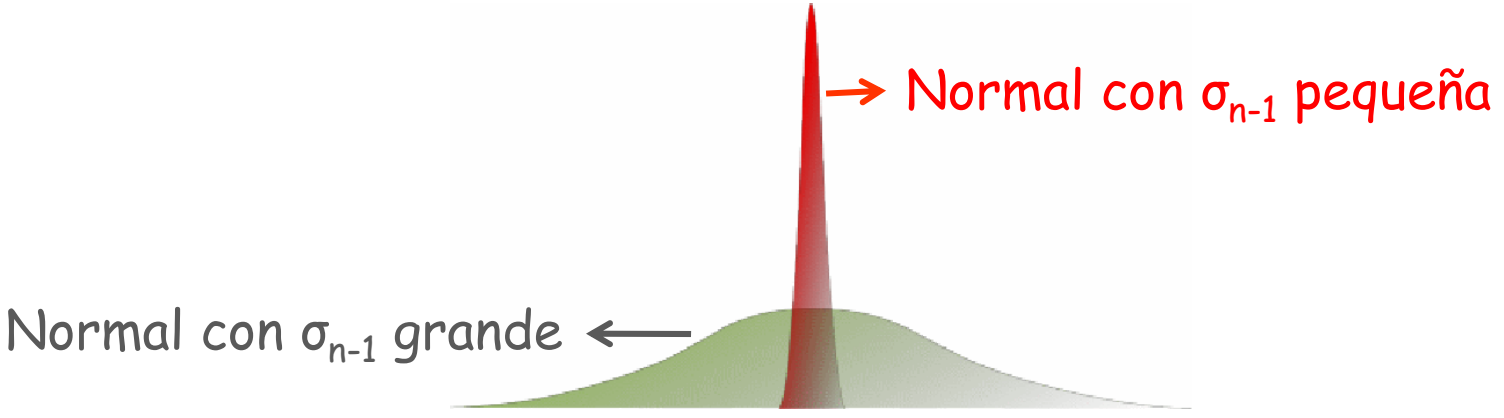
Esto significa que si un experimento fuera repetido muchas veces se podría esperar que el promedio de los valores así obtenidos para S^2 (valor muestral) igualaría a σ^2.

Estadística básica

Valores resumen

- **Desviación estándar:** También describe dispersión de la distribución. Es la raíz de la medida de desviación alrededor de la media, $\sqrt{\sigma^{2}_{n-1}}$

En las mismas unidades de medida que la variable estudiada.



Estadística básica

Valores resumen

- **Error estándar:** que se comete al estimar la media de la variable medida con los "n" observaciones de la muestra. A mayor tamaño muestral menor error,

$$\varepsilon = \sqrt{\sigma^{2}_{n-1}/n}$$

- **Coeficiente de variación:** Es una medida de la variación relativa de los datos en porcentaje,

$$CV\,\% = \frac{\sigma_{n-1}}{\overline{X}_{n}} 100$$

Estadística básica

Valores resumen: De forma

- **Coeficiente de asimetría** (de Fisher):

Describe la simetría de la distribución relativa a la distribución normal.

$$\alpha_3 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{X}_n\right)^3 \Big/ \sigma^3$$

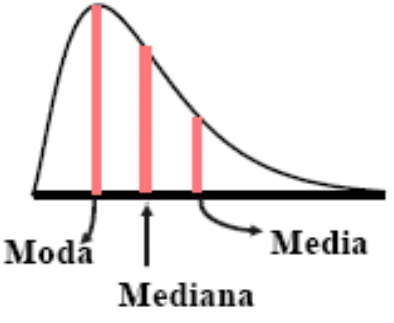
Estadística básica

Valores resumen: De forma

- **Coeficiente de asimetría** (cont):

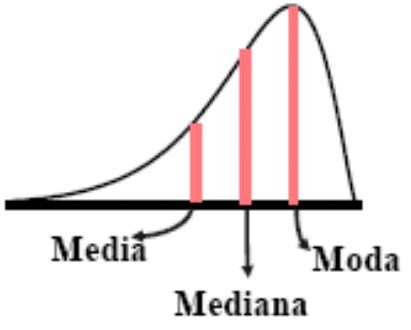
Media
Mediana
Moda
Simétrica

$\alpha_3 = 0$



Asimétrica hacia
la derecha

$\alpha_3 < 0$



Asimétrica hacia
la izquierda

$\alpha_3 > 0$

Asimetría negativa = mayor concentración de valores a la izquierda de la media.

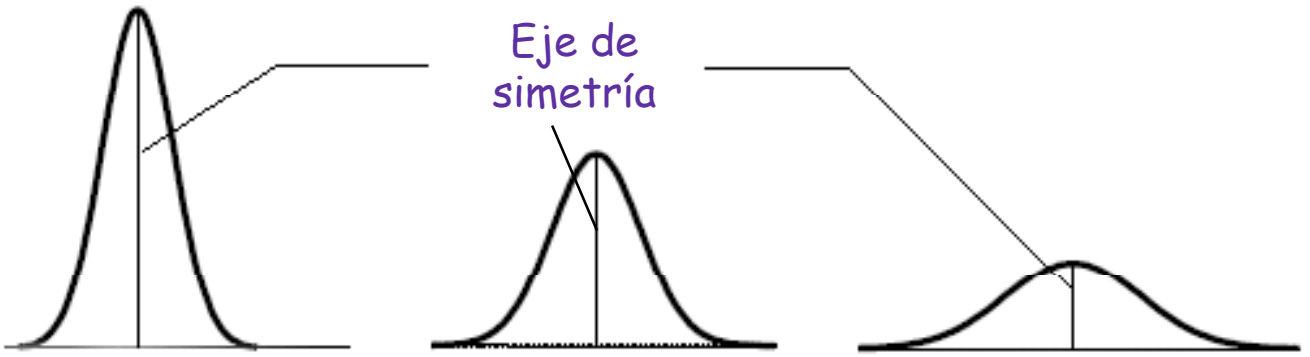
Asimetría positiva = mayor concentración de valores a la derecha de la media.

Estadística básica

Valores resumen: De forma

- **Curtosis** (o apuntamiento): Describe el grado de esbeltez de la distribución, en relación a una distribución normal,

$$\alpha_4 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{X}_n\right)^4 \Big/ \sigma^4$$



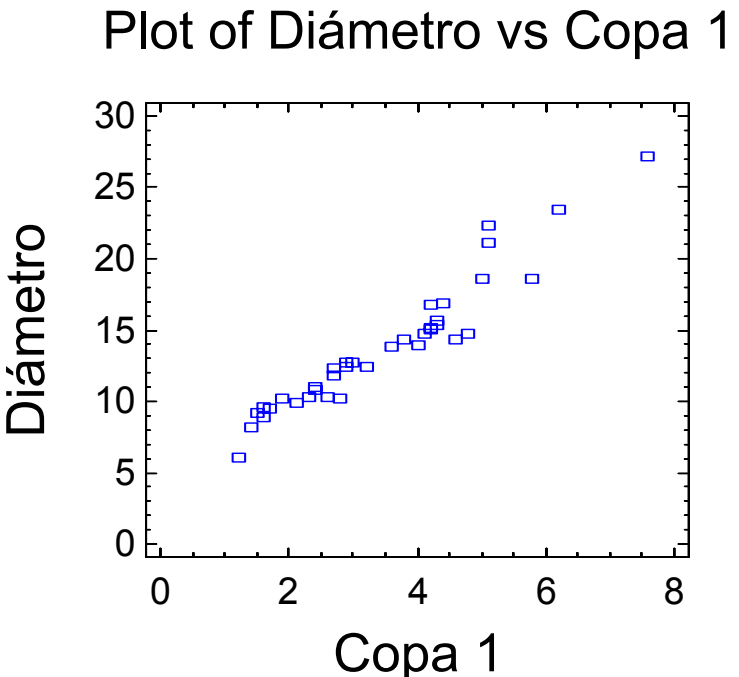
$\alpha_4 > 3$ $\alpha_4 = 3$ $\alpha_4 < 3$

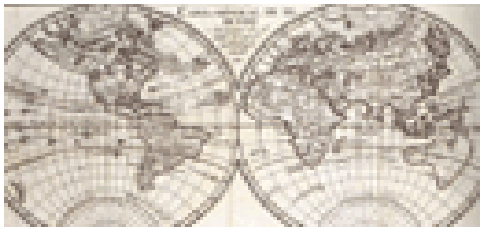
Estadística básica

Gráficos estadísticos

Permiten ilustrar y entender las distribuciones de los datos, identificar datos errados, valores extremos, tendencias en la variación de los datos, relaciones entre variables,...

Gráfico de dispersión
(scatterplot X-Y)



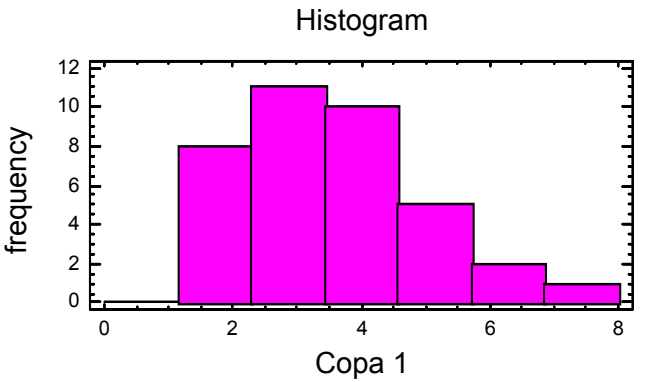


Exploración de datos

Gráficos estadísticos

Gráficos descriptivos para una variable (i)

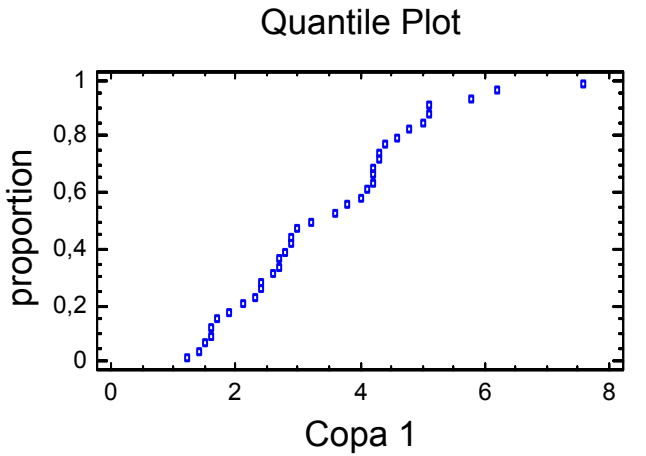
- Histogramas



- Gráficos de cuantiles:

```
Percentiles for Copa 1

1,0% = 1,2
5,0% = 1,4
10,0% = 1,6
25,0% = 2,4
50,0% = 3,2
75,0% = 4,3
90,0% = 5,1
```



Exploración de datos

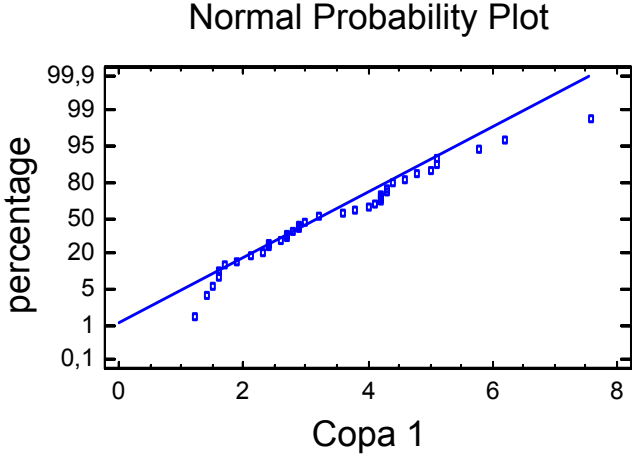
Gráficos estadísticos

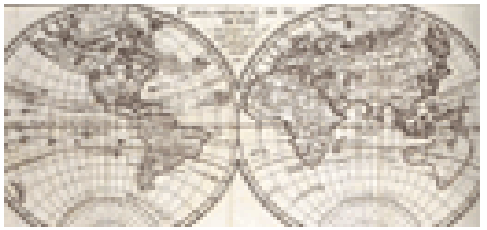
Gráficos descriptivos para una variable (ii)

- Gráfico de cuantiles para verificar el ajuste de los datos a la distribución Normal: (Q-Q Normal)

Eje vertical: valores de la función de distribución de la Normal.
Recta: gráfico de los valores de la variable con los valores de probabilidad acumulada de ocurrencia según la distribución Normal.

La proximidad de los valores observados a la recta indica que los datos se pueden considerar con distribución Normal



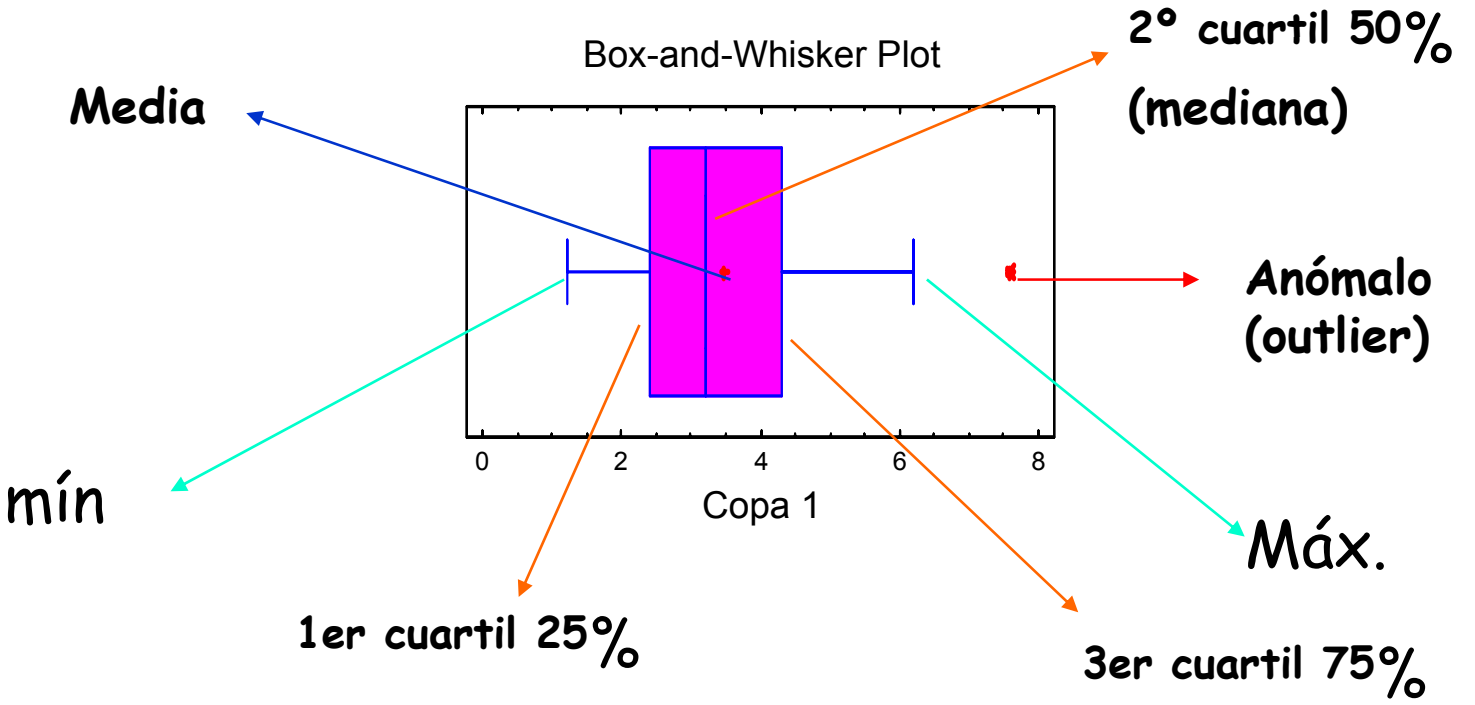


Exploración de datos

Gráficos estadísticos

Gráficos descriptivos para una variable (iii)

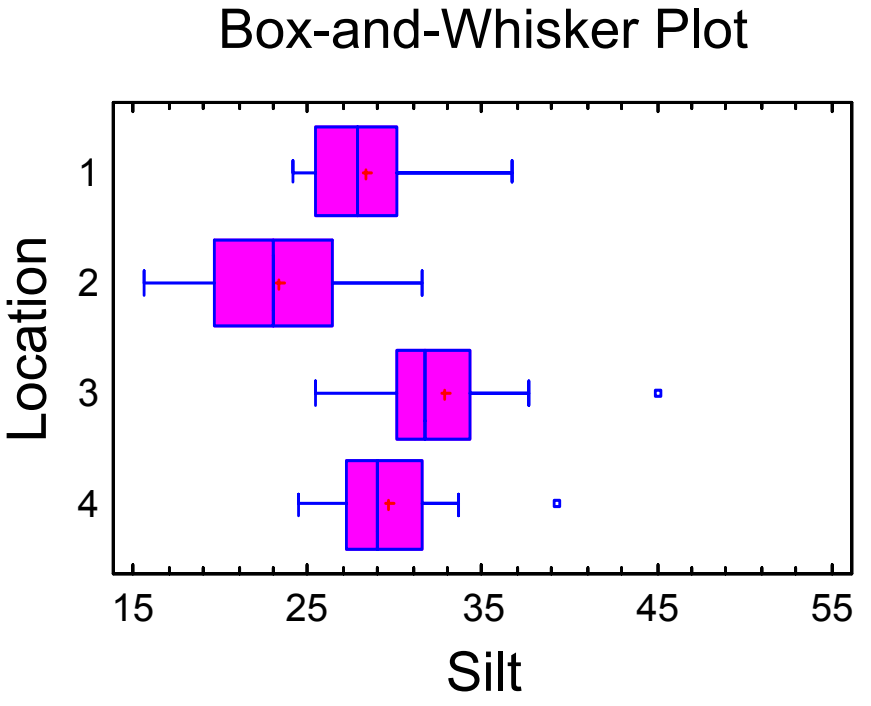
Gráficos de cajas (box-plot)



Exploración de datos

Gráficos estadísticos

Comparación gráfica de la variable silt (sedimento) en los distintos puntos de muestreo (1 a 4):



Transformaciones

Para modelos de interpolación del tipo regresión, las hipótesis requieren, entre otras condiciones:

Normalidad de los datos

Homogeneidad en la varianza

Si en el análisis exploratorio no se observa simetría en el histograma y con un contraste de bondad de ajuste (prueba chi-cuadrado o Kolmogorov-Smirnov) se confirma la falta de normalidad, se tendrá que recurrir a algún tipo de transformación "normalizante" de los datos.

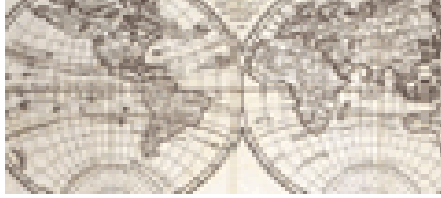
Transformaciones Box-Cox o de potencia

$$T(X) = Y = \begin{cases} \dfrac{X^{\lambda} - 1}{\lambda} & \lambda \neq 0 \\ \ln X & \lambda = 0 \end{cases}$$

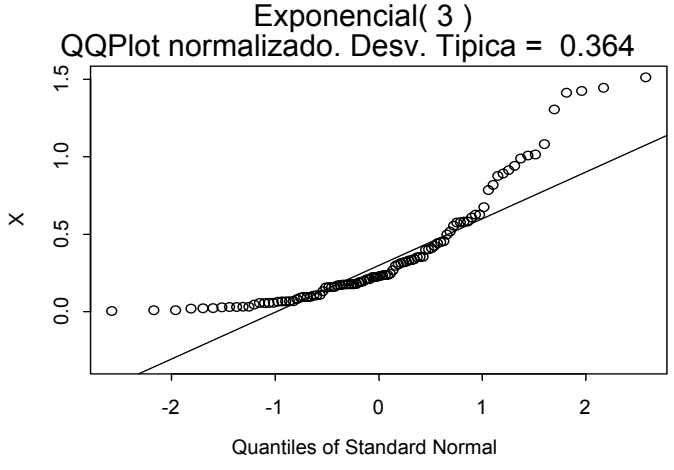
- $\lambda=2$, $Y=X^2$
- $\lambda=1/2$, $Y=X^{1/2}$
- Se busca que la variable transformada se parezca a una distribución normal

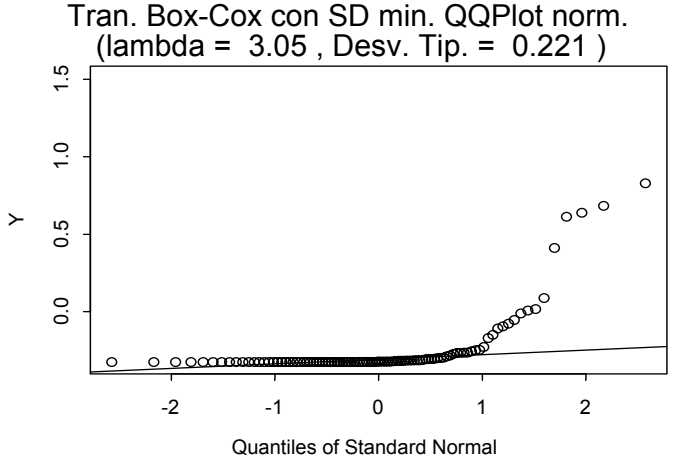
$$Y \equiv X^{(\lambda)} \sim N(\mu, \sigma^2)$$

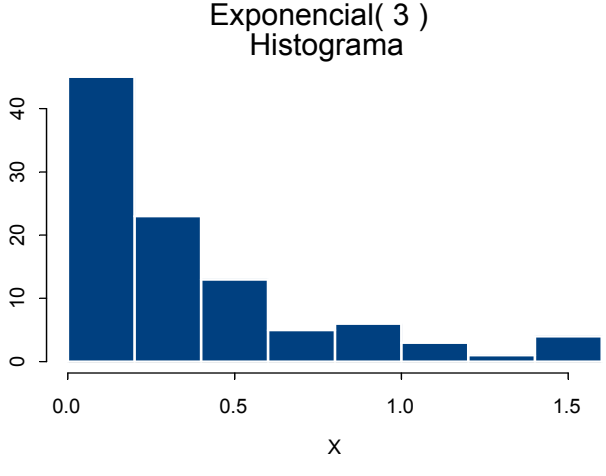
Ejemplo: X ~ Exp(3)

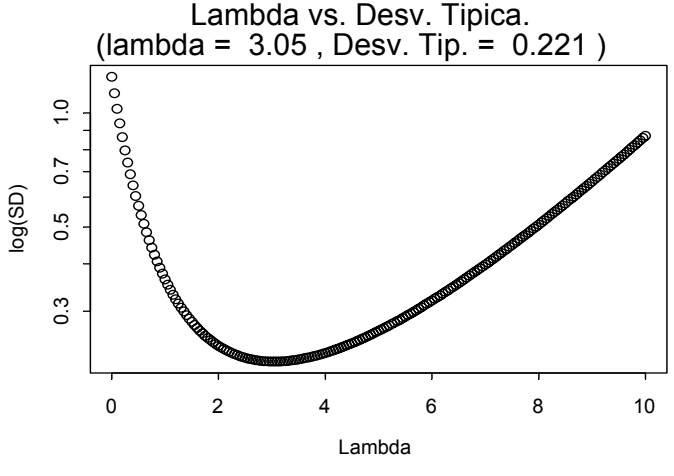


- Rango: [0, 10] pasos de 0.05.
- La mejor fue $\lambda = 3.05$

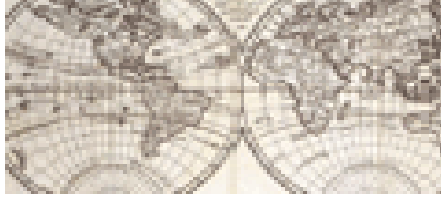




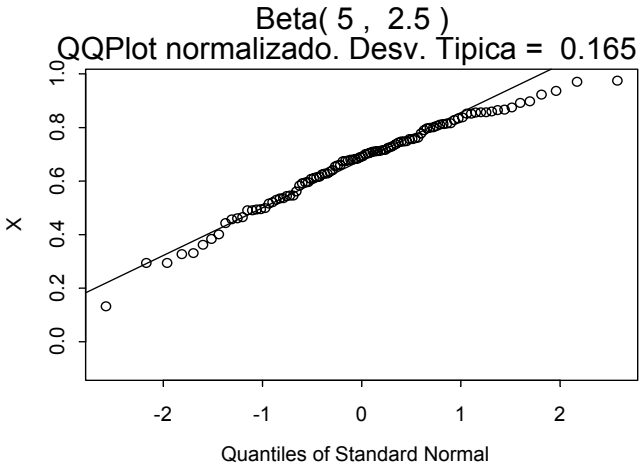


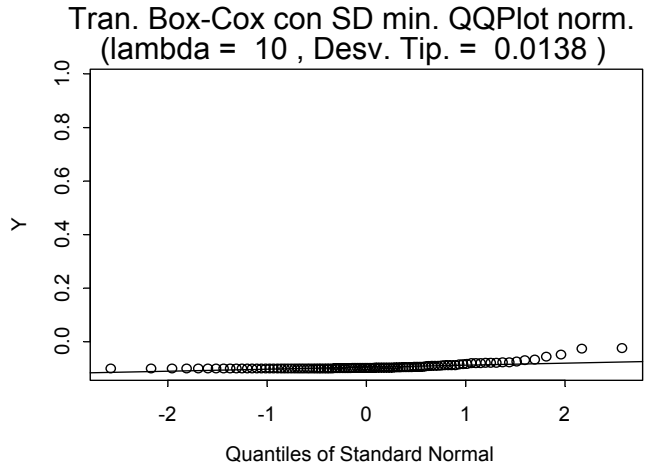


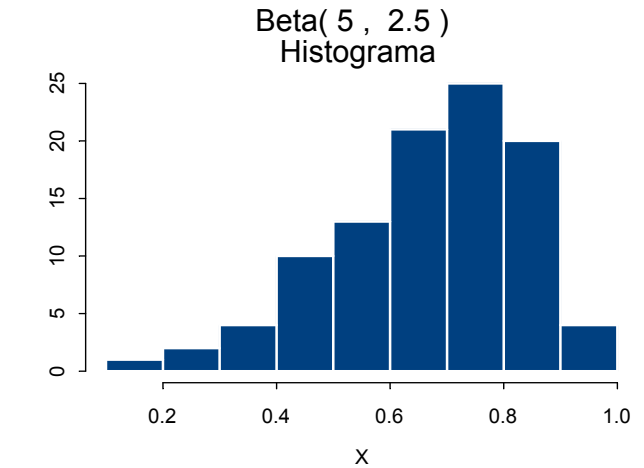
Ejemplo: X ~ Beta(5, 2.5)

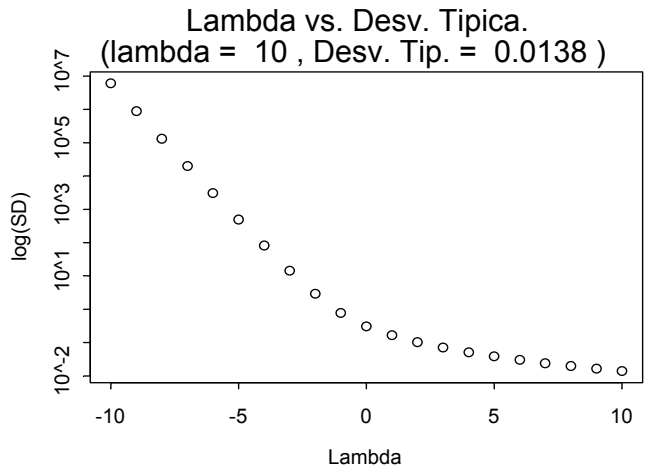


- Rango: [-10, 10] pasos de 1.
- La mejor fue $\lambda >= 10$

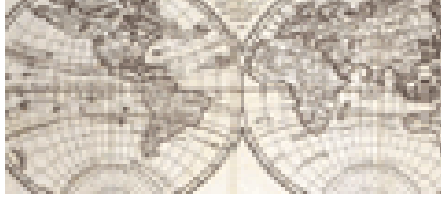




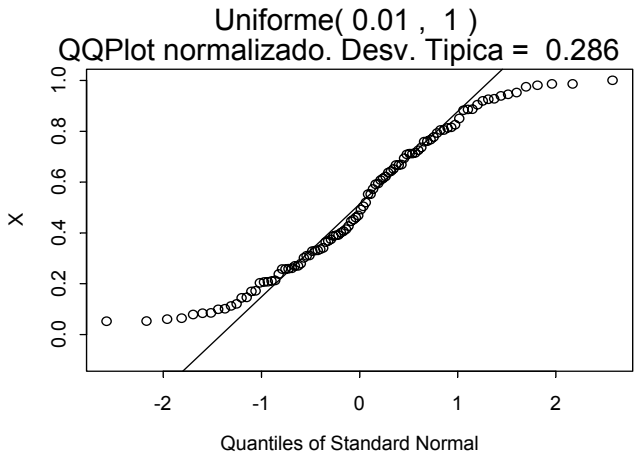


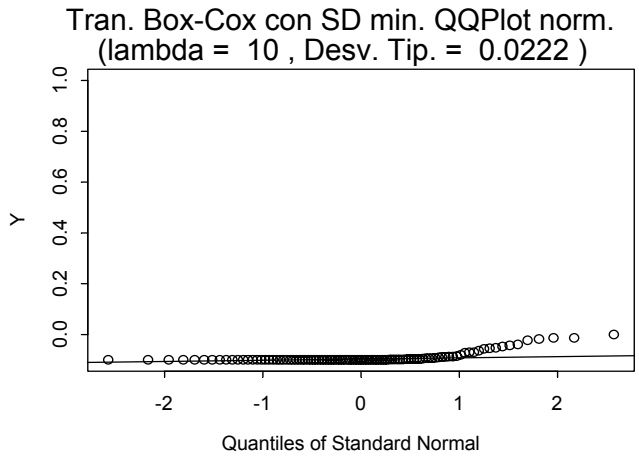


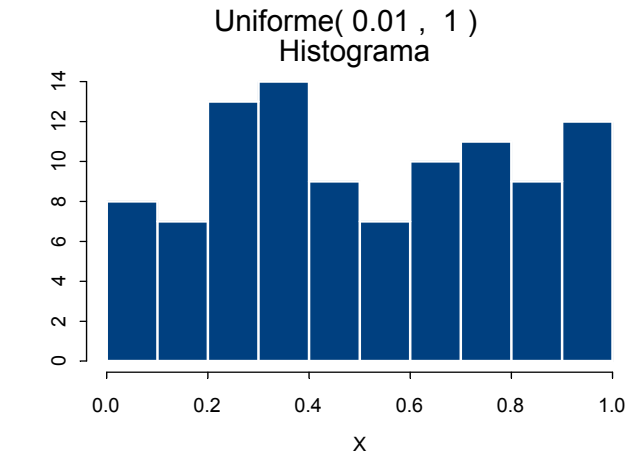
Ejemplo: X ~ U(0.01, 1)

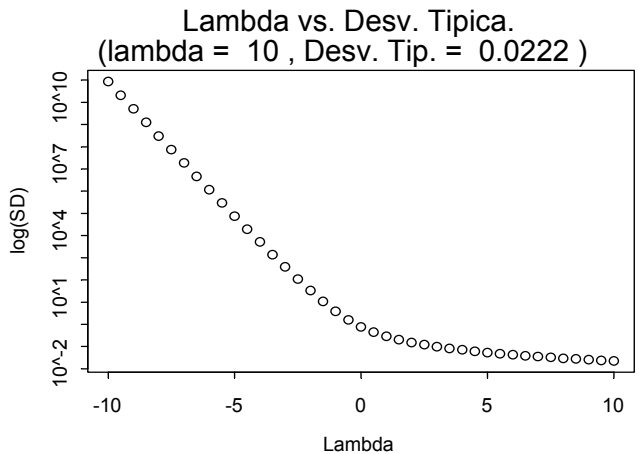


- Rango: [-10, 10] pasos de 0.5.
- La mejor fue $\lambda >= 10$

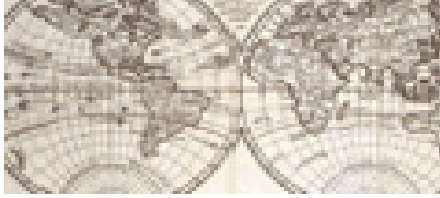








REFERENCIAS - ENLACES WEB



http://descargas.cervantesvirtual.com/servlet/SirveObras/46860175104026839600080/006458_8.pdf
Cap.7: Sistemas de Información Geográfica: Pasado, presente y futuro (tesis doctoral)

www.geogra.uah.es/~joaquin/curso-quito/**SIG**-OdelT.pdf

http://ares.unimet.edu.ve/postgrado/mpi002/Estadistica%20Descriptiva/256,1, Estadística Descriptiva

ttp://www.elagrimensor.net/elearning/lecturas/sig-capitulo%206.pdf
Interpolación a partir de mapas e isolíneas (aplicaciones estadísticas a datos geográficos, diseños de muestreo.,…en región de Murcia)

http://www.ine.gob.mx/ueajei/publicaciones/libros/459/cap3.html
Diseños de muestreo para suelos. Ejemplo de sistemático en contaminación de suelos.

www.monografías .com. Elementos de Geoestadística. CUADOR GIL, J.Q. Universidad de Pinar del Río (Cuba).