



UNIVERSIDAD NACIONAL AUTÓNOMA
DE MÉXICO

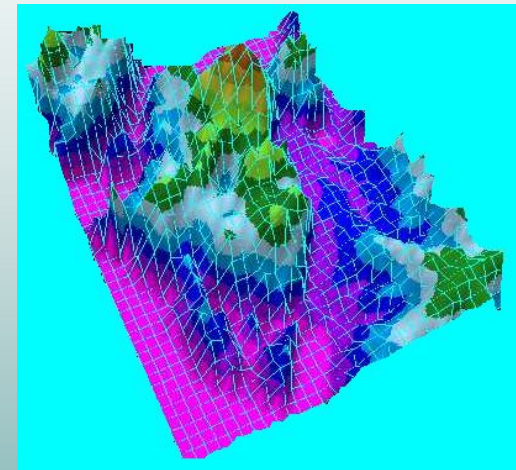
GEOESTADÍSTICA APLICADA

**Tema: Análisis Exploratorio
de Datos**

Instructores:

Dr. Martín A. Díaz Viera (mdiazv@imp.mx)

Dr. Ricardo Casar González (rcasar@imp.mx)



2009

Análisis Exploratorio de Datos

- *¿Qué es el AED?*
- *Importancia del AED*
- *Etapas de cualquier AED*
- *Herramientas del AED*
- *Estadística univariada*
- *Estadística bivariada*
- *Estadística multivariada*
- *Regresión lineal y mínimos cuadrados*

¿Qué es el AED?

- *Es un conjunto de técnicas estadísticas y gráficas que permiten establecer un buen entendimiento básico del comportamiento de los datos y de las relaciones existentes entre las variables que se estudian.*

Importancia del AED

- *El análisis exploratorio de datos (AED) es un paso previo e indispensable para la aplicación exitosa de cualquier método estadístico.*
- *En particular permite la detección de fallos en el diseño y toma de datos, el tratamiento y/o la evaluación de datos ausentes, la identificación de valores atípicos y la comprobación de los supuestos requeridos por parte de las técnicas geoestadísticas.*

Etapas de un AED

- Realizar un examen gráfico de la naturaleza de las variables individuales y un análisis descriptivo numérico que permita cuantificar algunos aspectos gráficos de los datos.
- Realizar un examen gráfico de las relaciones entre las variables y un análisis descriptivo numérico que cuantifique el grado de interrelación existente entre ellas.
- Evaluar algunos supuestos básicos subyacentes a muchas técnicas estadísticas, por ejemplo, normalidad, linealidad y homocedasticidad.
- Identificar los posibles valores atípicos (*outliers*) y evaluar el impacto potencial que puedan ejercer en análisis estadísticos posteriores.
- Evaluar, el impacto potencial que pueden tener los datos ausentes (*missing*) sobre la representatividad de los datos analizados.

Herramientas del AED

- *Estadística univariada*
- *Estadística bivariada*
- *Regresión lineal y mínimos cuadrados*

Estadística univariada

Variable Aleatoria (V.A.): Es una variable **Z** que puede tomar una serie de valores o realizaciones (z_i) cada una de las cuales tienen asociadas una probabilidad de ocurrencia (p_i).

Ejemplo: Al lanzar un dado puede resultar $\{1, 2, 3, 4, 5 \text{ o } 6\}$ con una probabilidad de ocurrencia igual a $1/6$.

Las probabilidades cumplen las condiciones:

$$a) \ p_i \geq 0, \quad \forall i$$

$$b) \ \sum_i p_i = 1$$

Estadística univariada

Variable Aleatoria Discreta: cuando el número de ocurrencias es finito o contable, se conoce como variable aleatoria discreta.

Ejemplo: tipos de facies en un yacimiento.

Variable Aleatoria Continua: si el número de ocurrencias posibles es infinito.

Ejemplo: el valor de la porosidad de un medio se encuentra en el intervalo $[0,100\%]$.

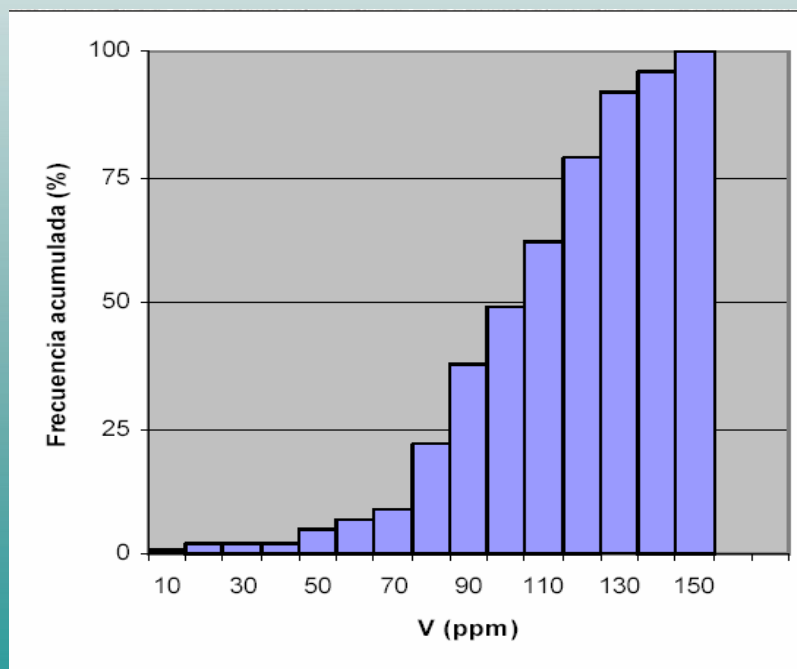
Estadística univariada

Función de Distribución de Probabilidad (FDP)

La **FDP** caracteriza completamente a la **VA**.

Se define como: $F(z) = \Pr\{Z \leq z\} \in [0,1]$

Su gráfica es el histograma acumulativo



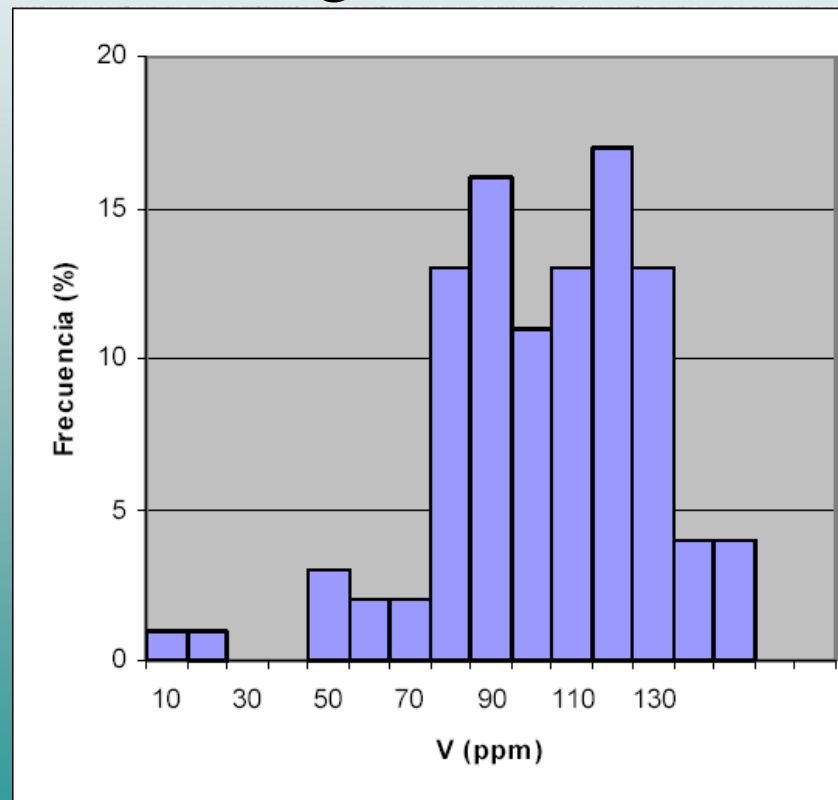
Estadística univariada

Función de Densidad de Probabilidad (fdp).

Se define como:

$$f(z) = \frac{dF(z)}{dz}$$

Su gráfica es el histograma.



Estadística univariada

Percentiles o cuantiles de una distribución .

- El percentil de una distribución $F(z)$ es el valor z_p de la V.A. que corresponde a un valor p de probabilidad acumulada, es decir:

$$F(z_p) = p$$

- Si existe la función inversa se puede expresar como:

$$z_p = F^{-1}(p)$$

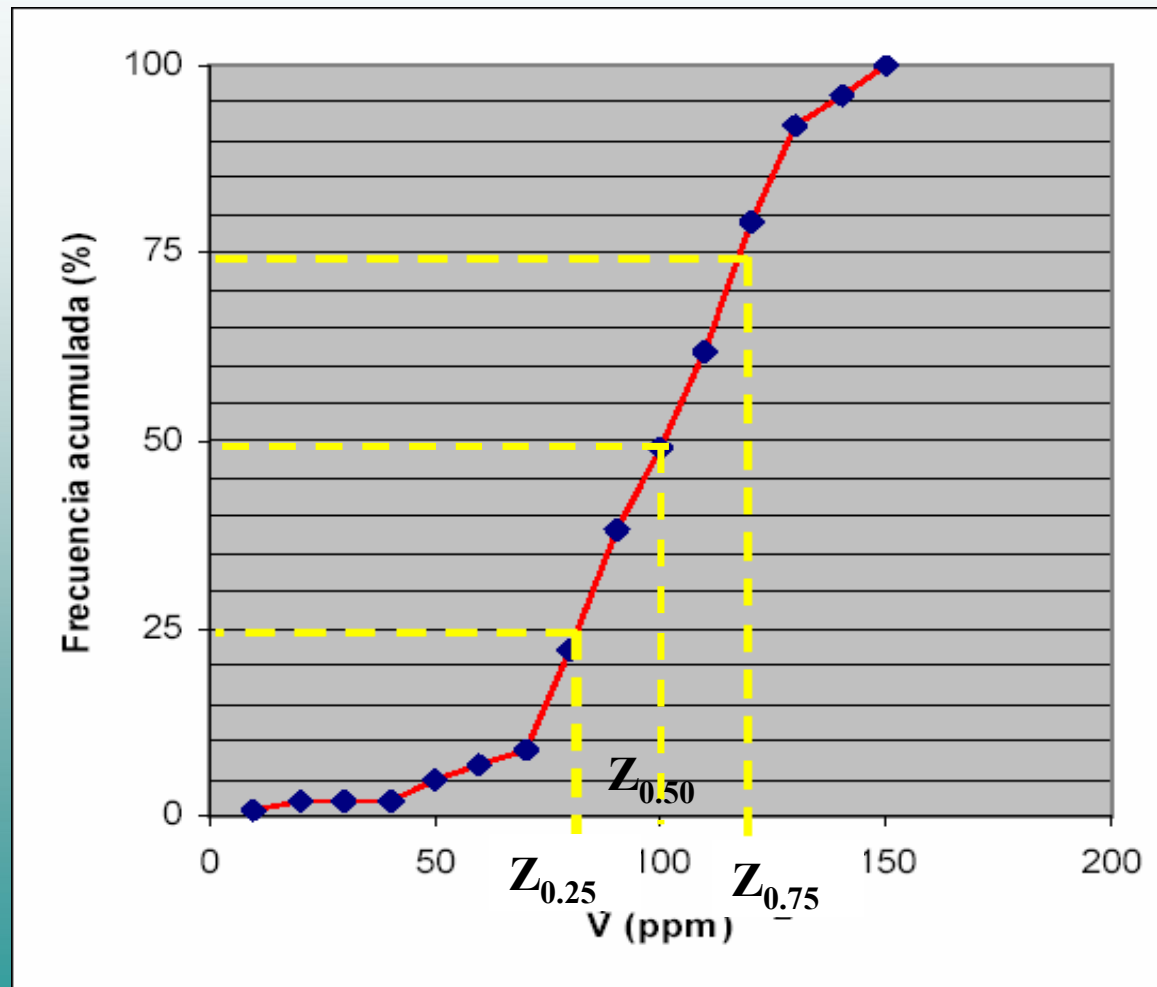
Estadística univariada

Algunos cuantiles de interés:

- **Mediana**, $p=0.5$ $M = F^{-1}(0.5)$
- **Cuartiles**
- (primer cuartil o inferior) $p=0.25$ $z_{0.25} = F^{-1}(0.25)$
- (tercer cuartil o superior) $p=0.75$ $z_{0.75} = F^{-1}(0.75)$
- Rango o intervalo intercuartil (IR) $[z_{0.25}, z_{0.75}]$

Estadística univariada

Ejemplo de cuartiles y rango intercuartil



Estadística univariada

Valor esperado o esperanza matemática de una VA.

Es el valor más probable que puede tomar una VA. Se conoce también como valor medio o media. Se define como:

$$m = E[Z] = \int_{-\infty}^{+\infty} z dF(z) = \int_{-\infty}^{+\infty} z f(z) dz$$

Su estimador más común es el promedio de todas las observaciones de la variable Z

$$m^* = \frac{1}{N} \sum_{i=1}^N z_i$$

Este estimador es muy sensible a los valores atípicos (*outliers*)

Estadística univariada

- **Momento de orden r de una FDP**

$$m_r = E[Z^r] = \int_{-\infty}^{+\infty} z^r dF(z) = \int_{-\infty}^{+\infty} z^r f(z) dz$$

- **Momento central de orden r de una FDP**

$$\mu_r = E[(Z - m)^r] = \int_{-\infty}^{+\infty} (z - m)^r dF(z) = \int_{-\infty}^{+\infty} (z - m)^r f(z) dz$$

Estadística univariada

Varianza de una VA (2do momento central)

- Se define como $\sigma^2 = \text{Var}[Z] = E[(Z - m)^2] \geq 0$
- Y caracteriza la dispersión de la distribución alrededor de la media.
- Su estimador es $(\sigma^2)^* = \frac{1}{N-1} \sum_{i=1}^N (z_i - m)^2$

Estadística univariada

- **Distribución Normal o Gaussiana.**
- Esta distribución está completamente caracterizada por sus dos parámetros: media y varianza y se designa mediante $N(m, \sigma^2)$

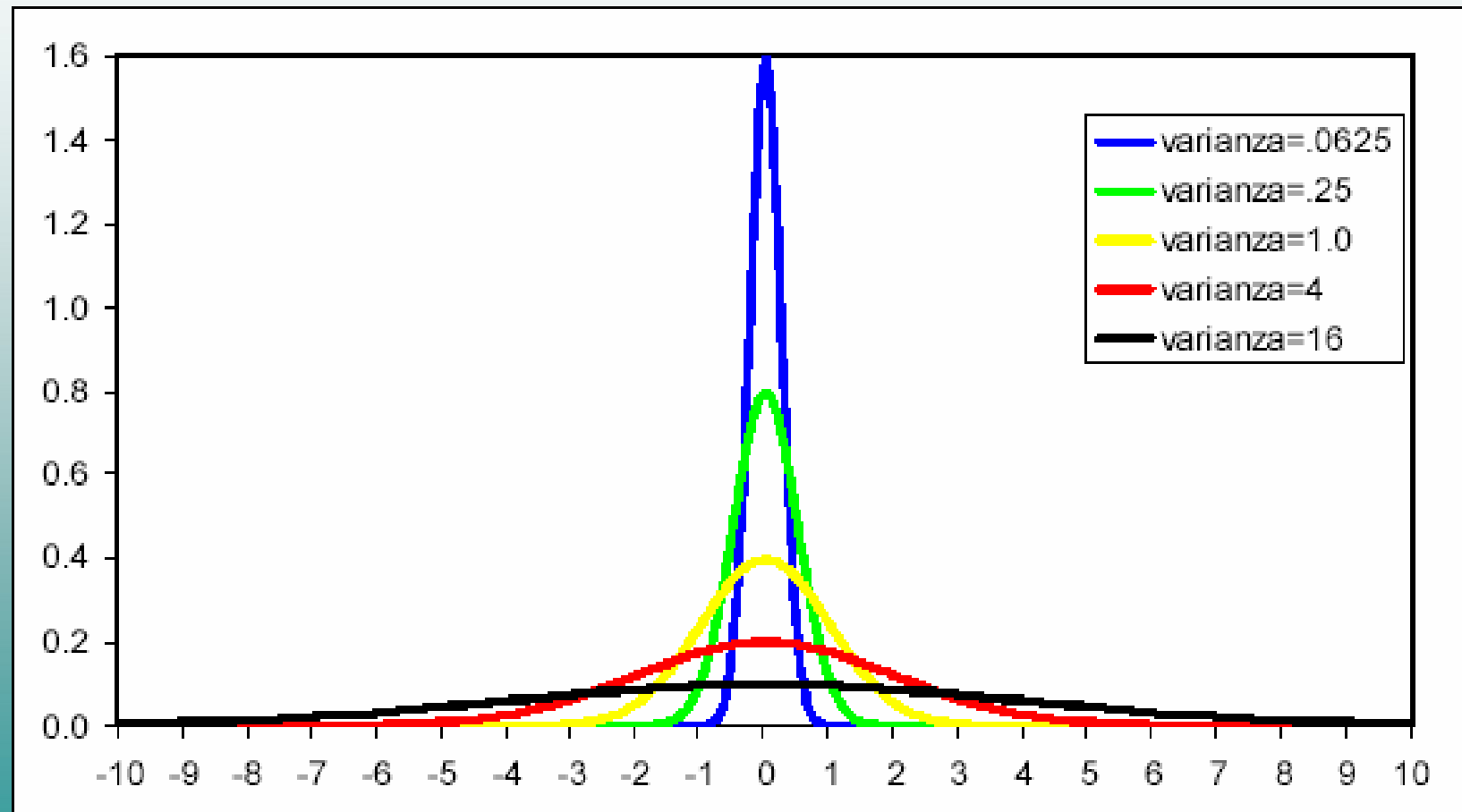
- La *fdp* normal o Gaussina está dada por

$$g(z) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{z-m}{\sigma}\right)^2\right]$$

- Es simétrica respecto a la media

Estadística univariada

Ejemplos de distribuciones Gaussianas



Estadística univariada

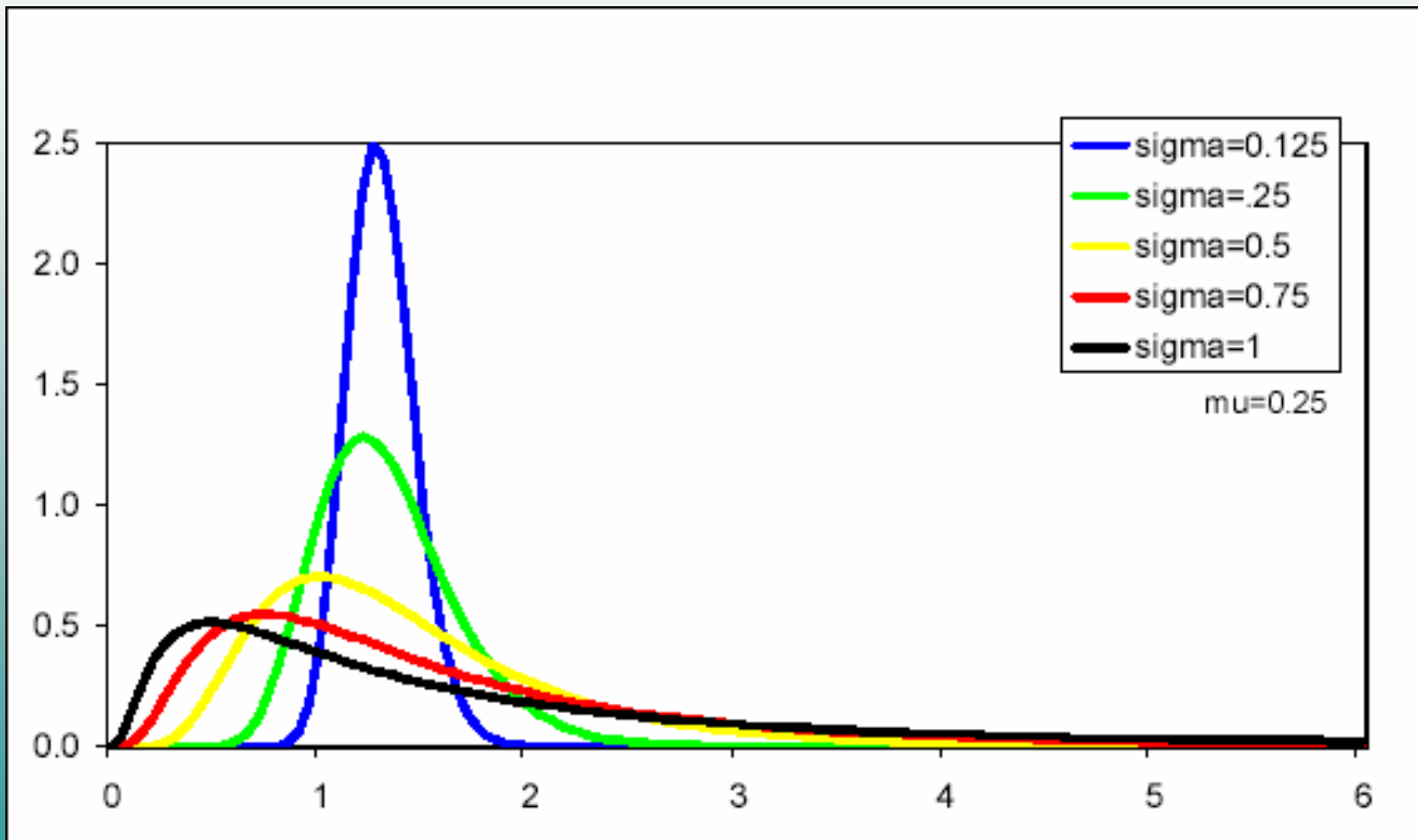
- **Distribución LogNormal**
- Una VA positiva Y se dice que tiene una distribución lognormal si su logaritmo $\ln(Y)$ esta normalmente distribuido.

$$Y > 0 \rightarrow \log N(m, \sigma^2), \text{ si } X = \ln Y \rightarrow N(\alpha, \beta^2)$$

- Muchas distribuciones experimentales en Ciencias de la Tierra tienden a ser asimétricas y la mayoría de las variables toman valores no negativos.

Estadística univariada

Ejemplos de distribuciones Lognormales



Estadística univariada

- **Desviación Estándar**

$$\sigma = \sqrt{\text{Var}[Z]}$$

- **Coeficiente de variación (dispersión relativa)**

$$CV = \sigma / m$$

- **Coeficiente de simetría (medida de la simetría)**

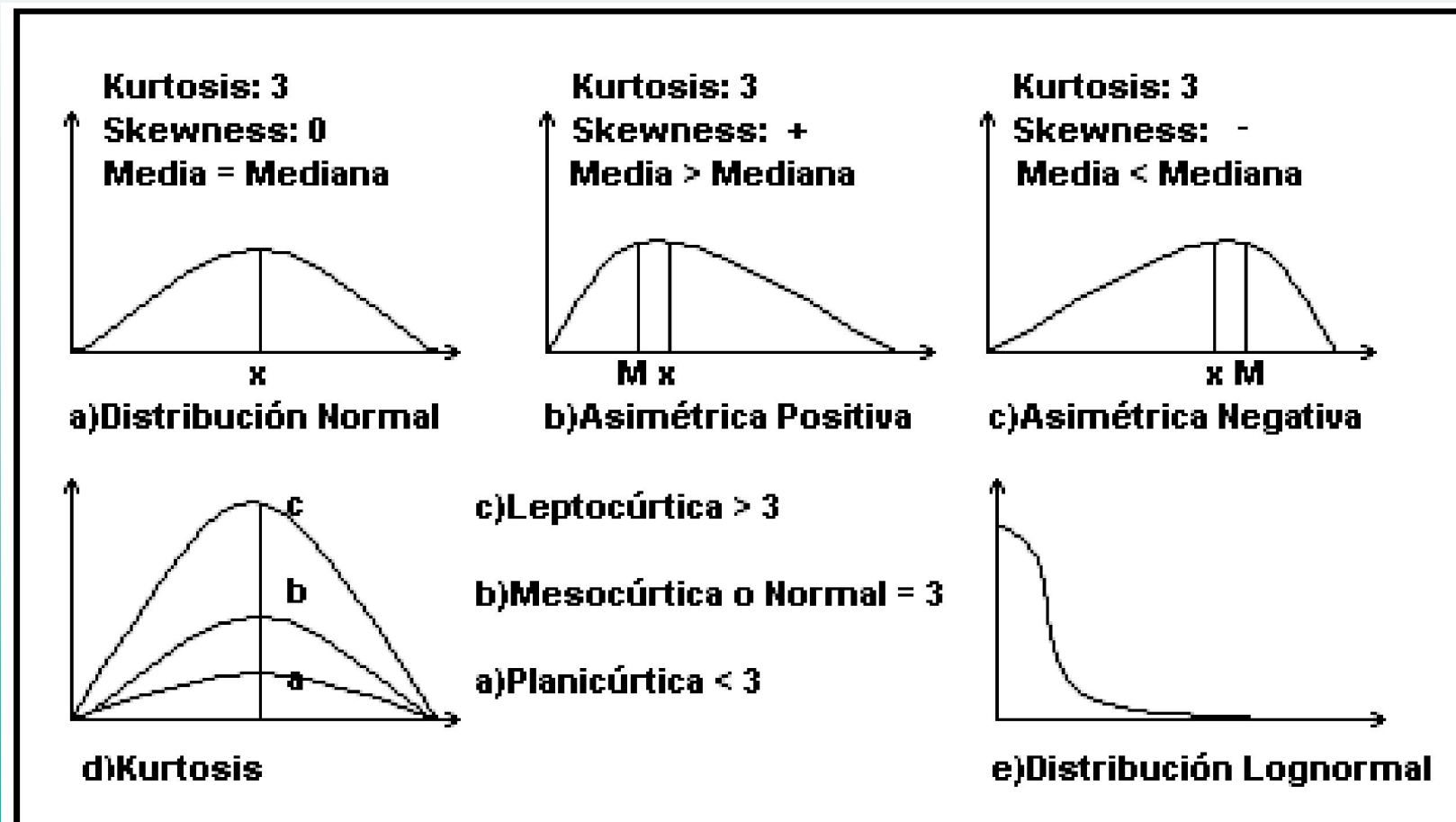
$$\alpha_1 = \frac{\mu_3}{\mu_2^{3/2}}$$

- **Coeficiente de curtosis (medida del achatamiento)**

$$\alpha_2 = \frac{\mu_4}{\mu_2^2} - 3$$

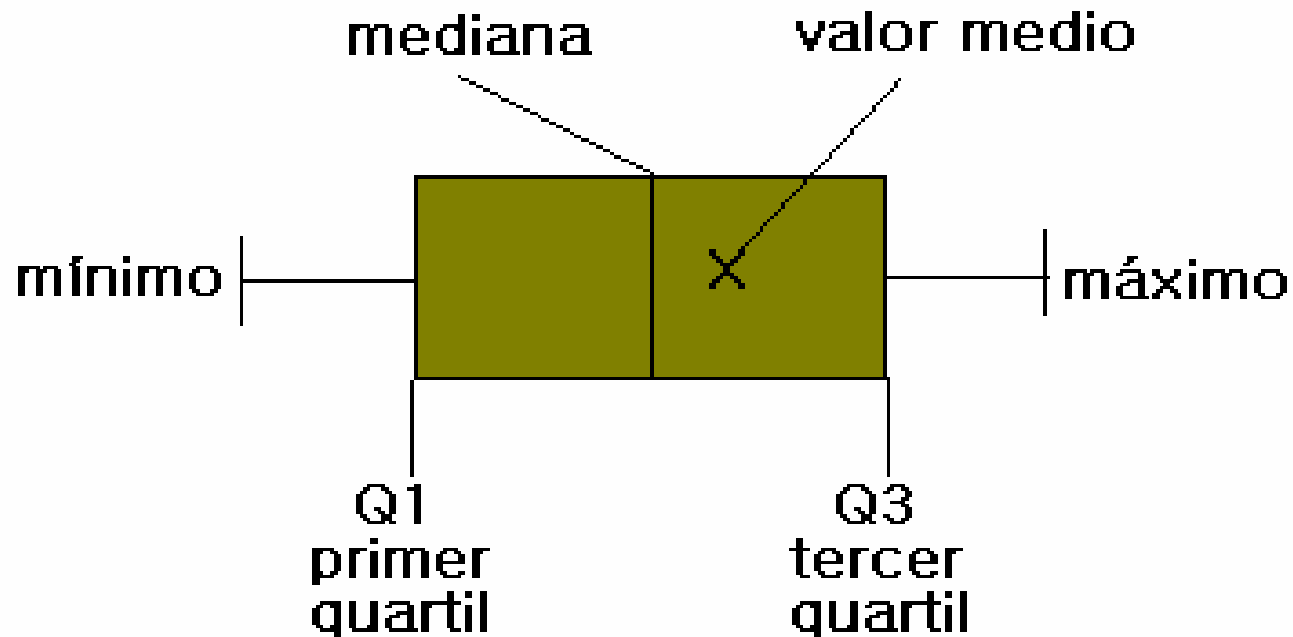
Estadística univariada

Simetría y Curtosis de una distribución



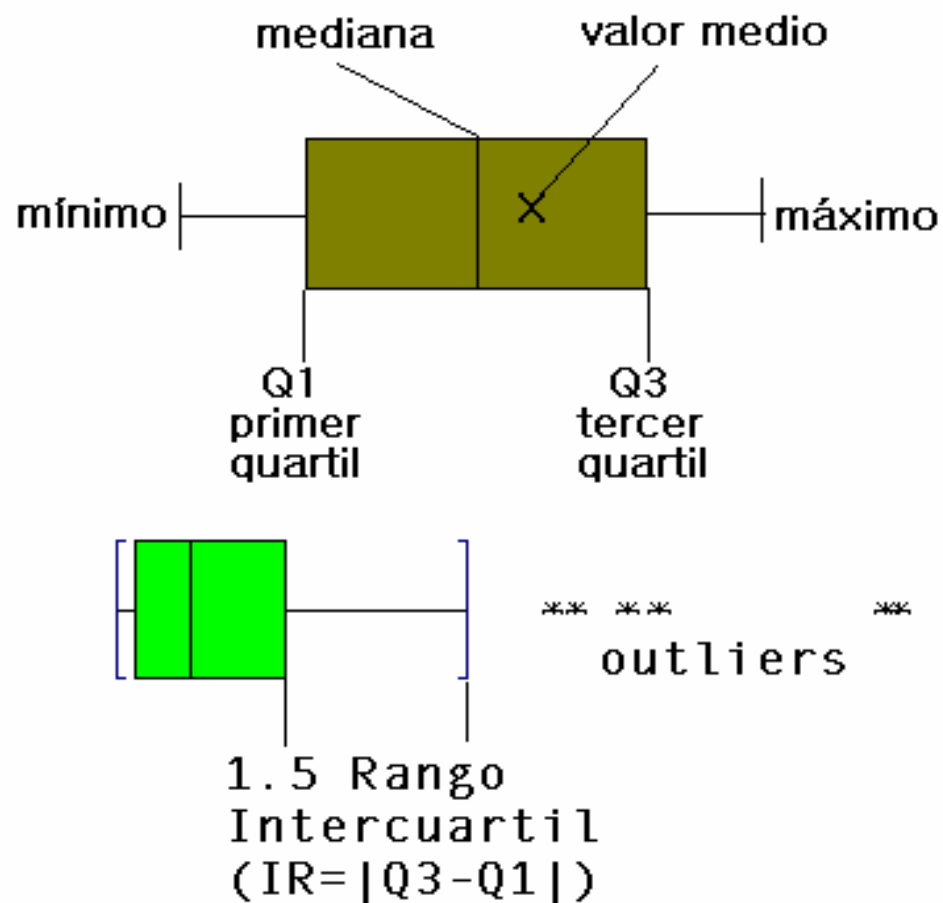
Estadística univariada

BOX PLOT



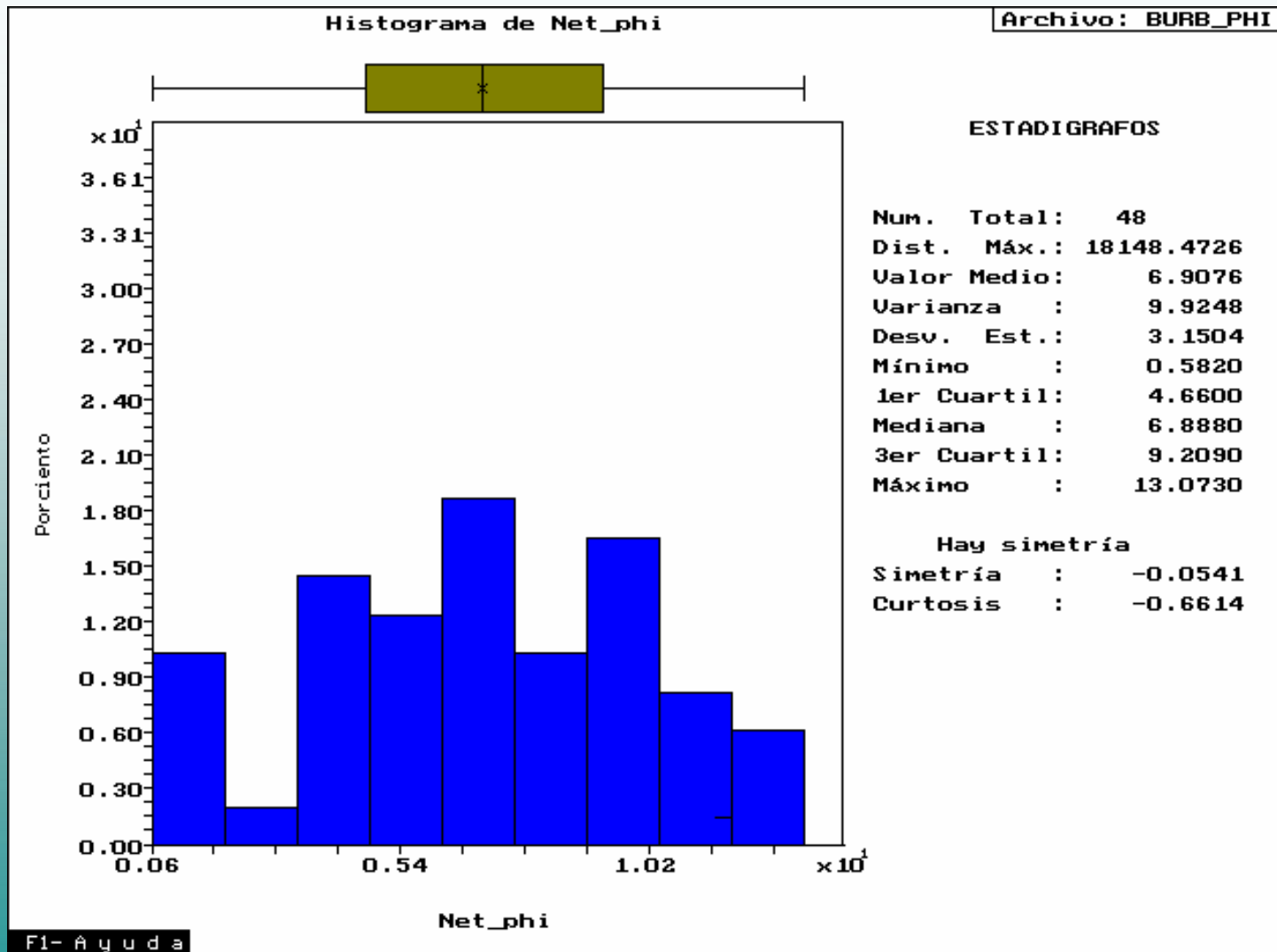
Estadística univariada

BOX PLOT



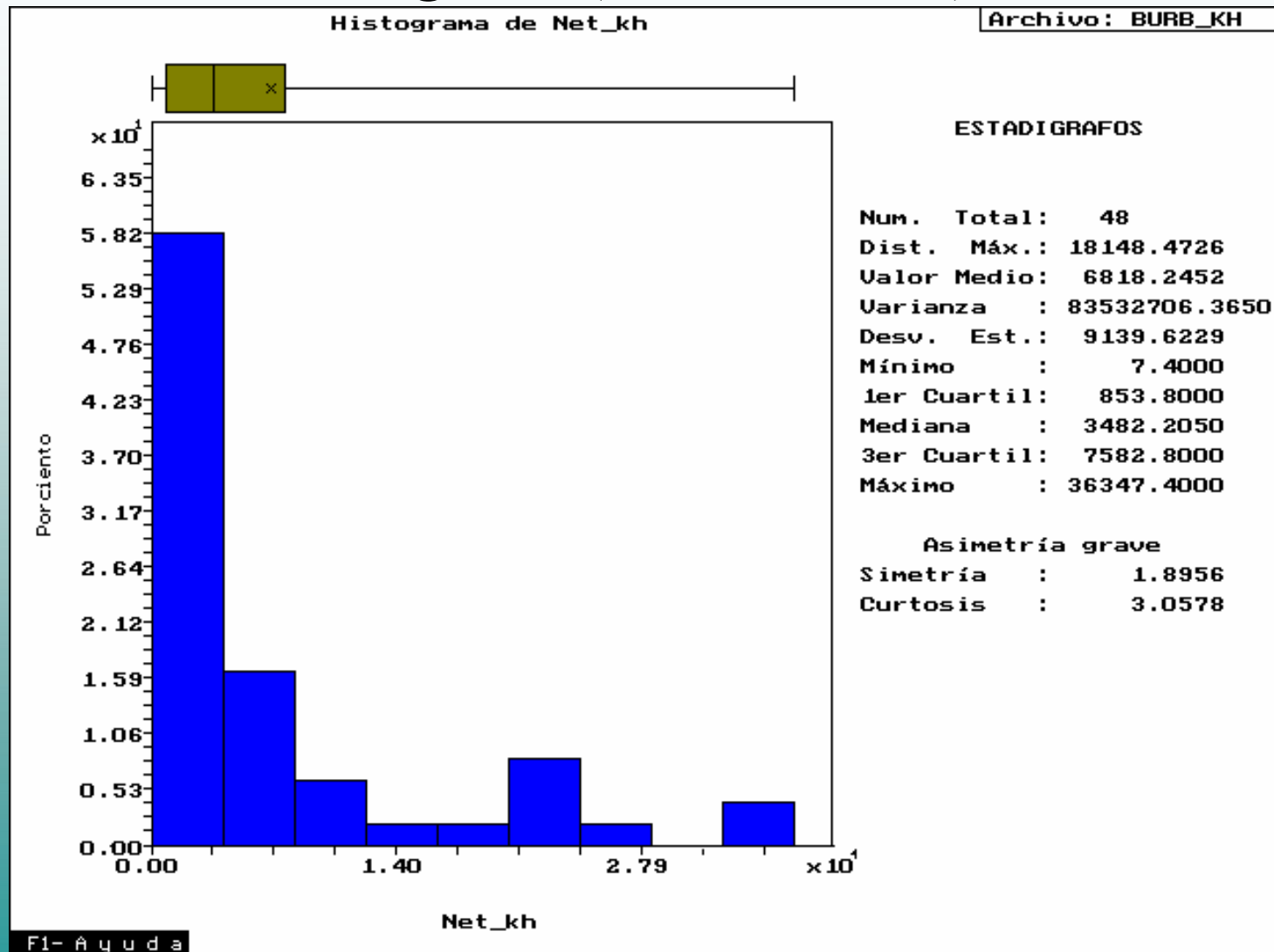
Estadística univariada

Histograma (Porosidad)



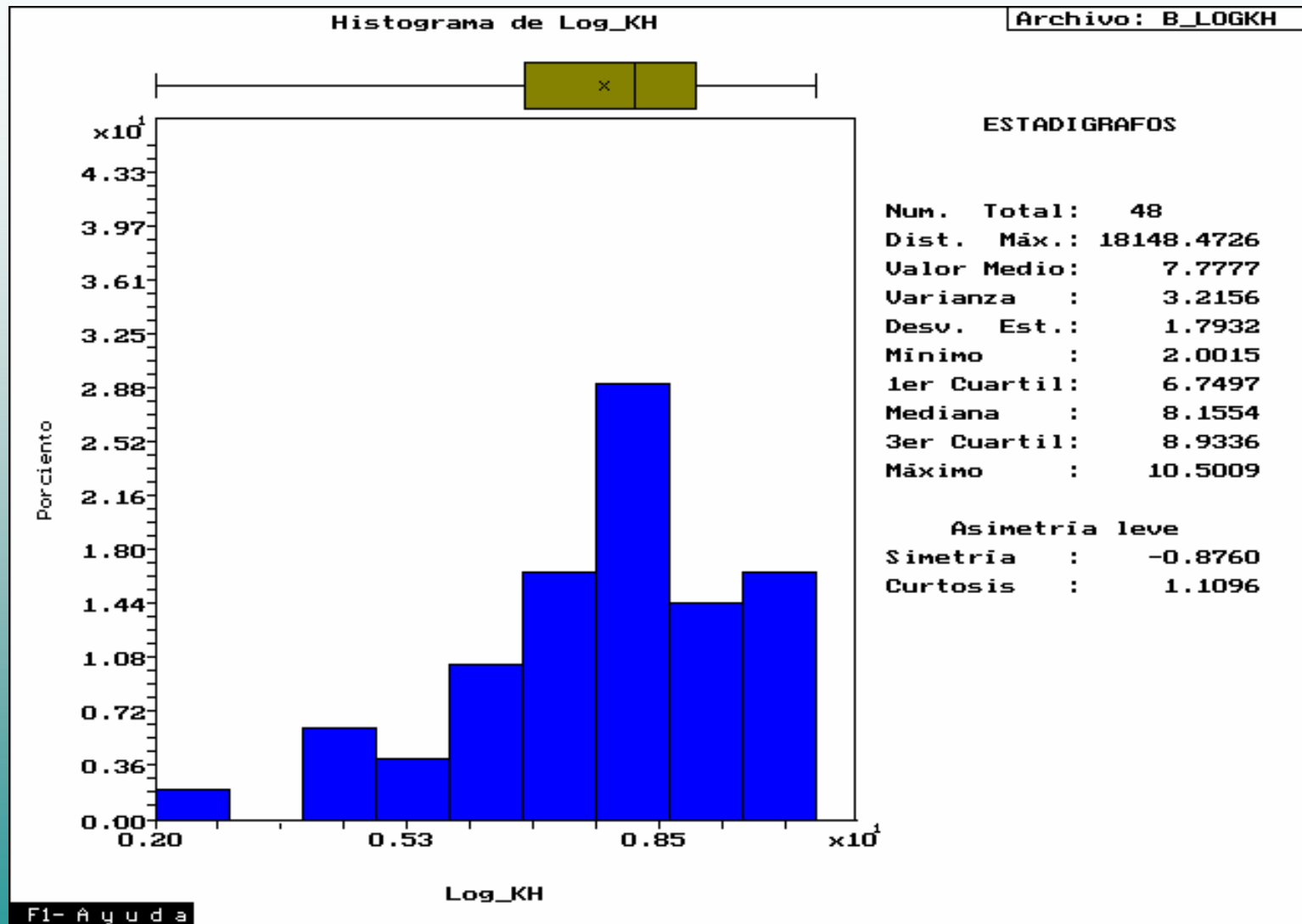
Estadística univariada

Histograma (Permeabilidad)



Estadística univariada

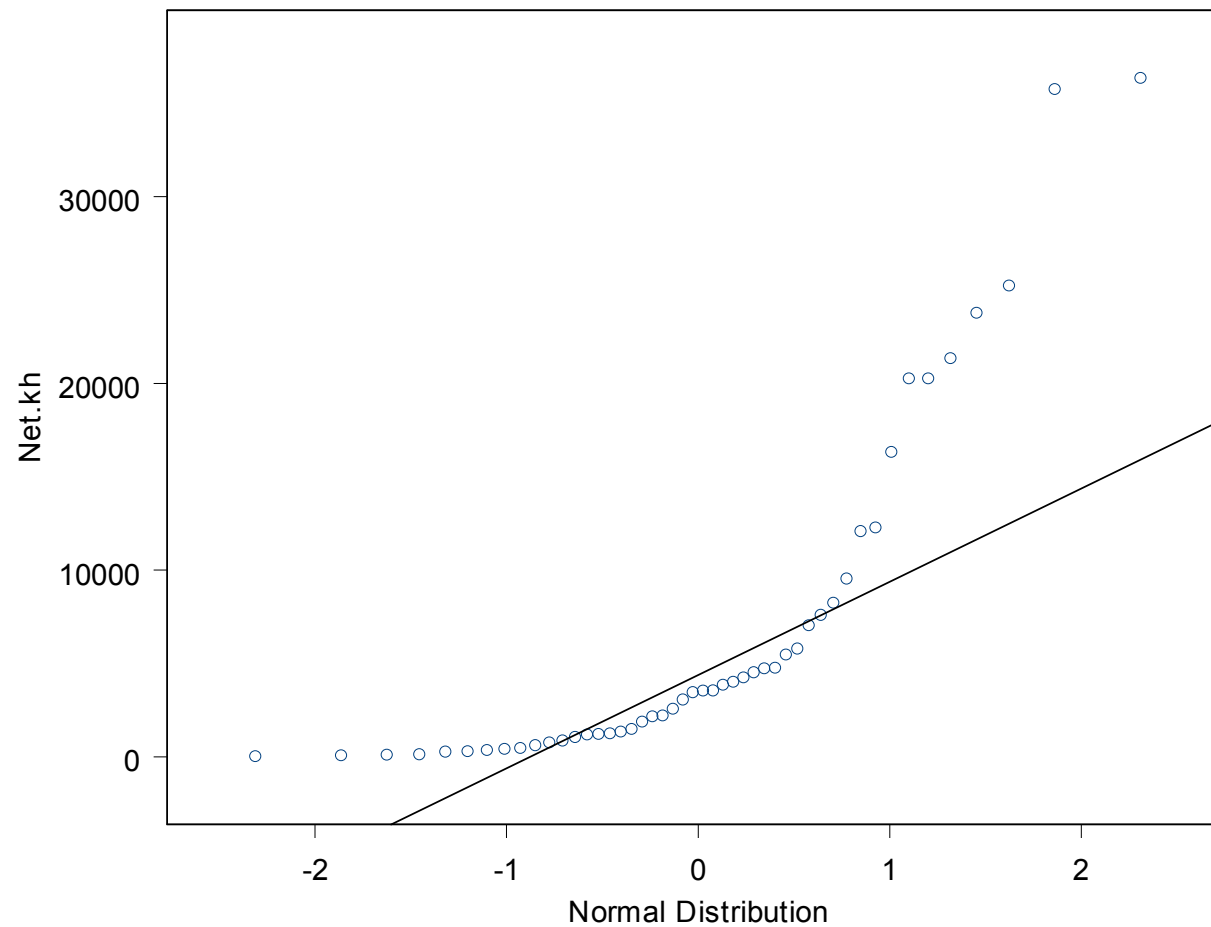
Transformación logarítmica de la Permeabilidad



Estadística univariada

Q-Q Plot de la Permeabilidad

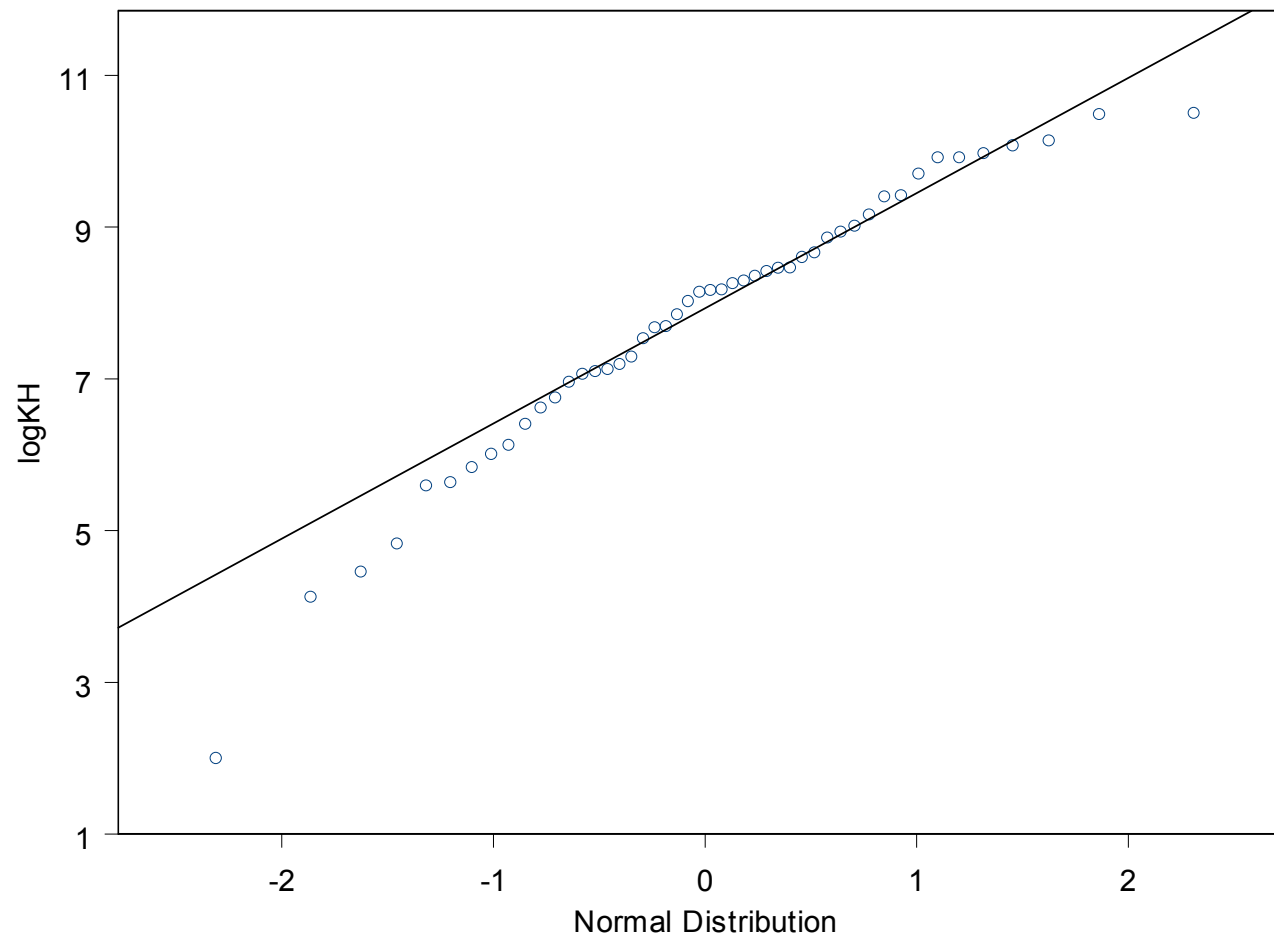
Antes de transformar



Estadística univariada

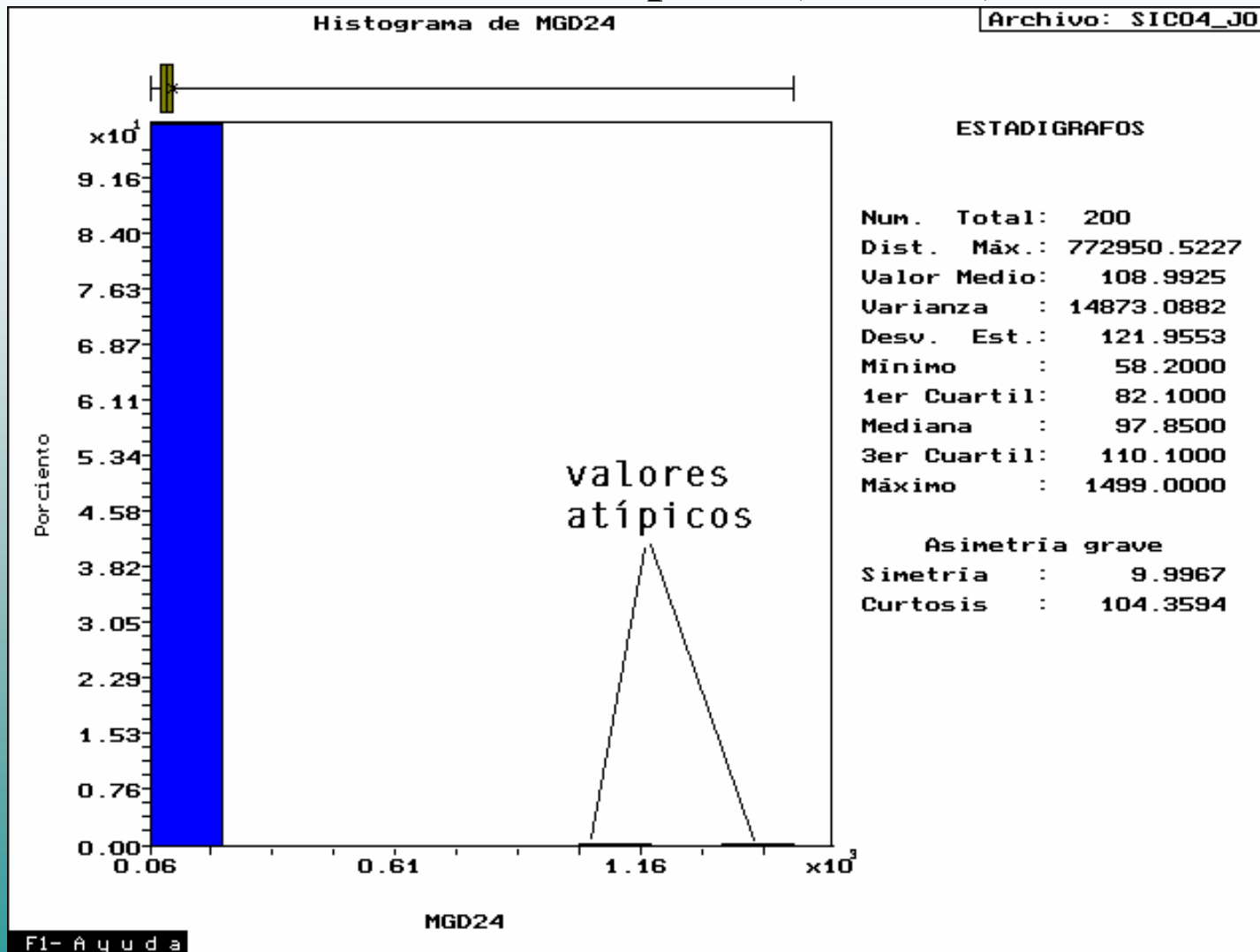
Q-Q Plot de la Permeabilidad

Después de transformar



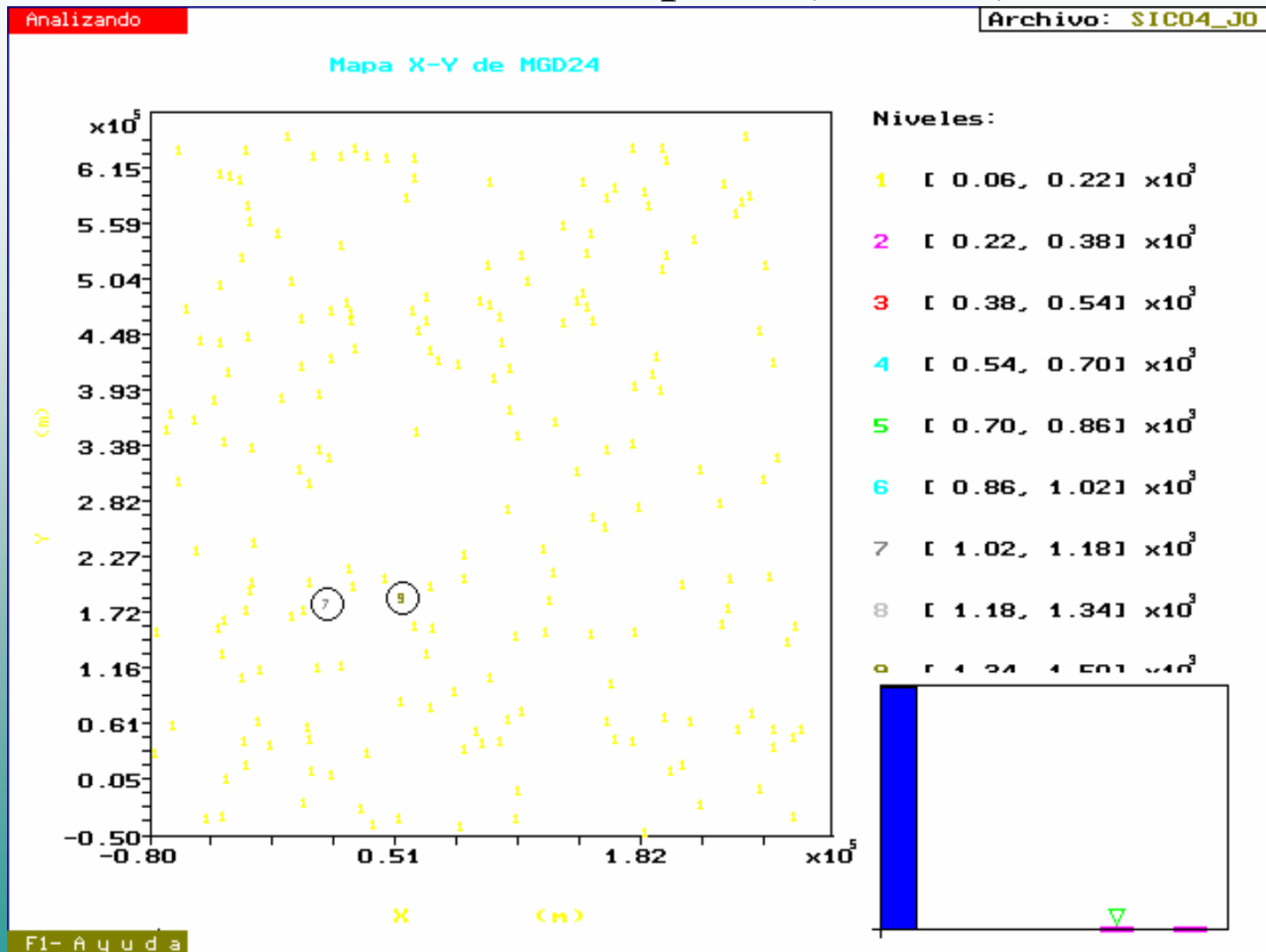
Estadística univariada

Con valores atípicos (outliers)



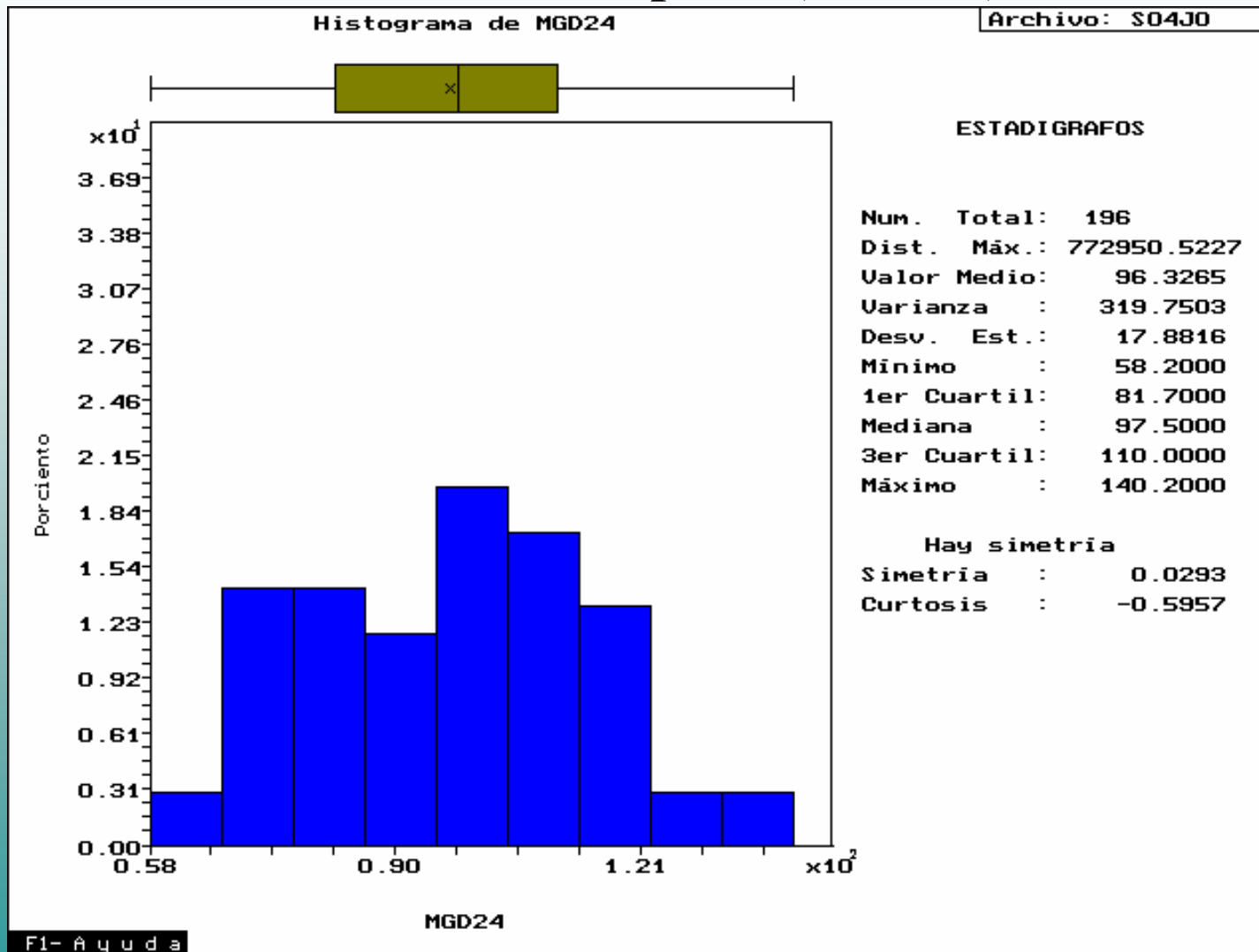
Estadística univariada

Con valores atípicos (outliers)



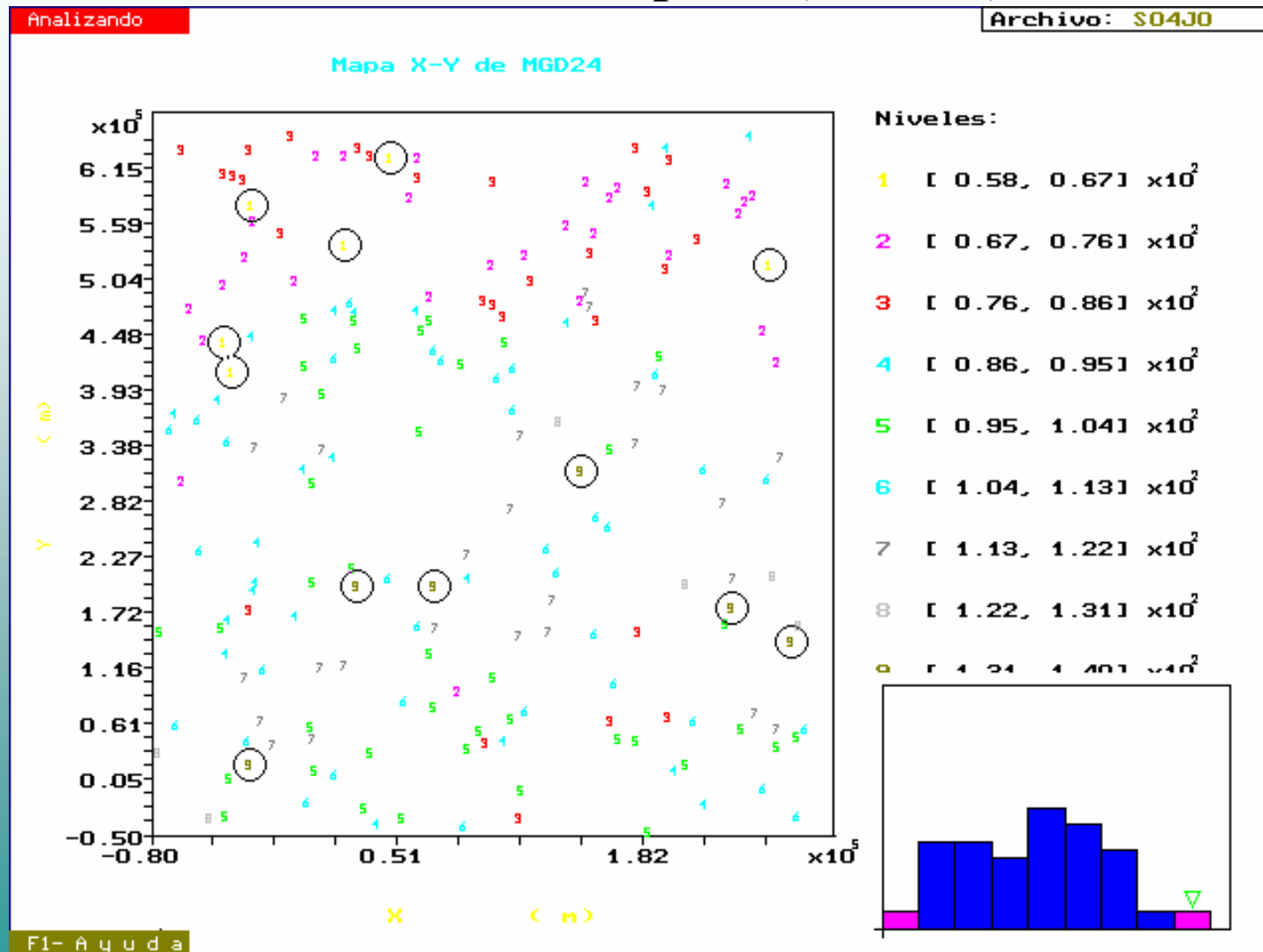
Estadística univariada

Sin valores atípicos (outliers)



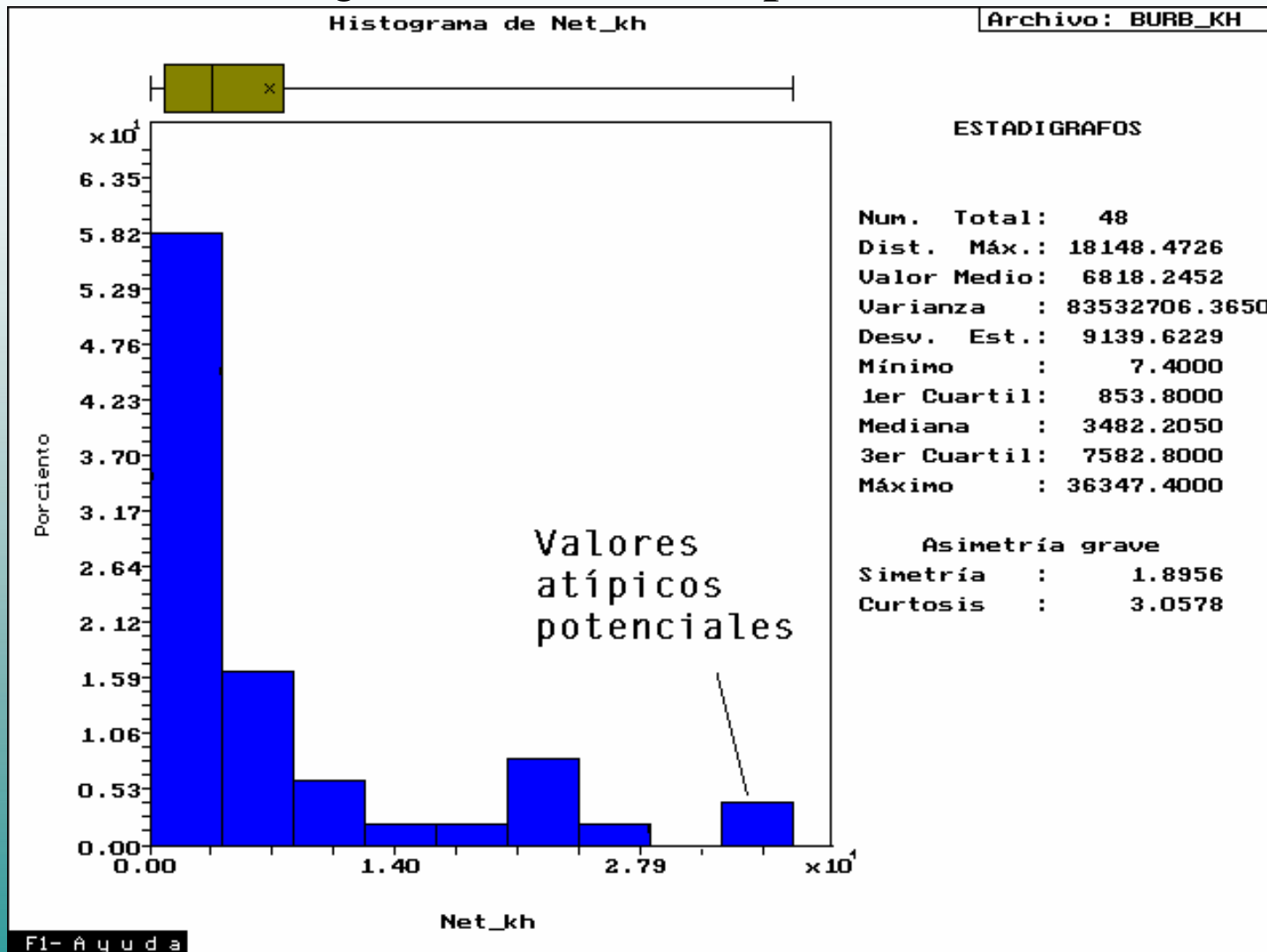
Estadística univariada

Sin valores atípicos (outliers)



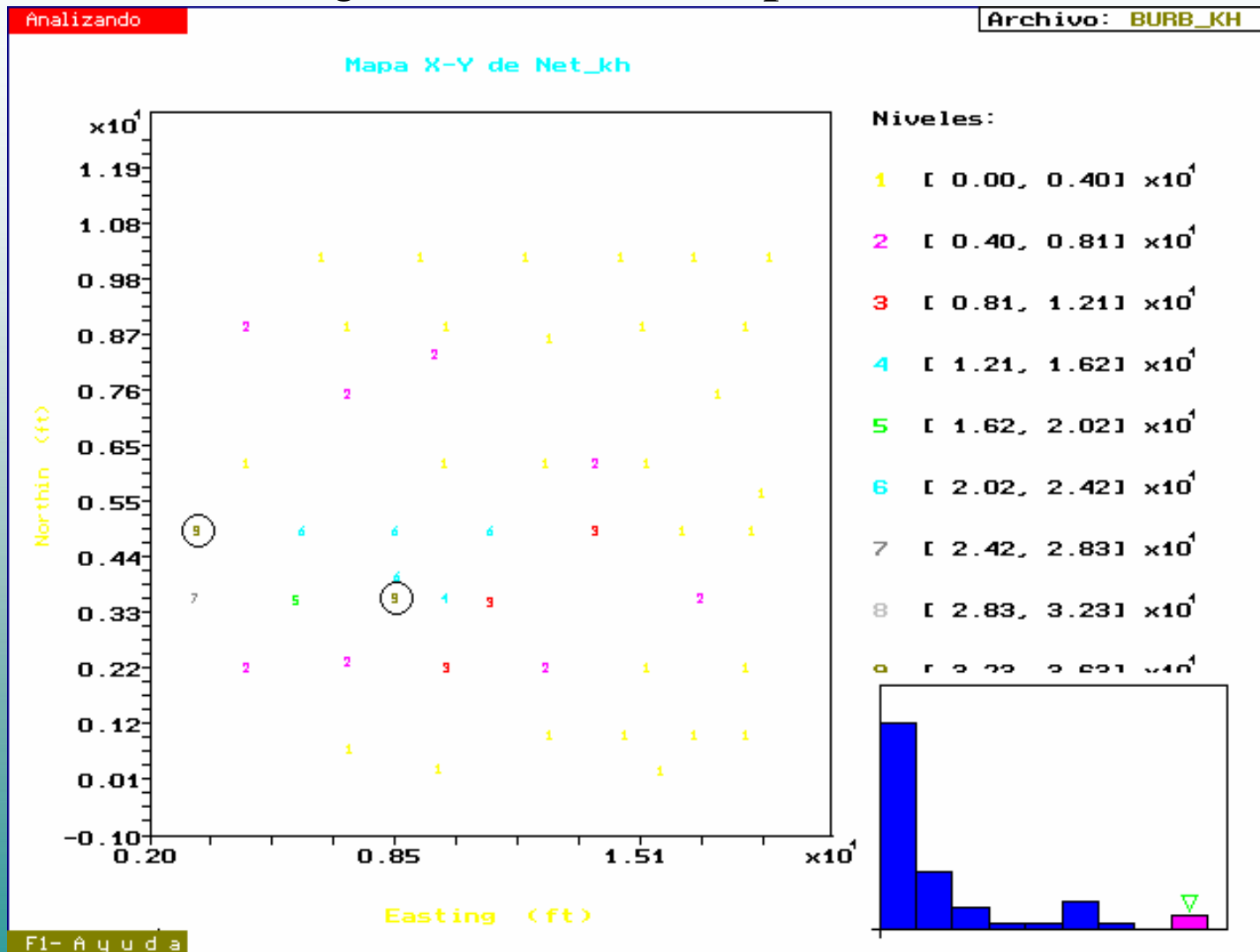
Estadística univariada

¿Serán valores atípicos?



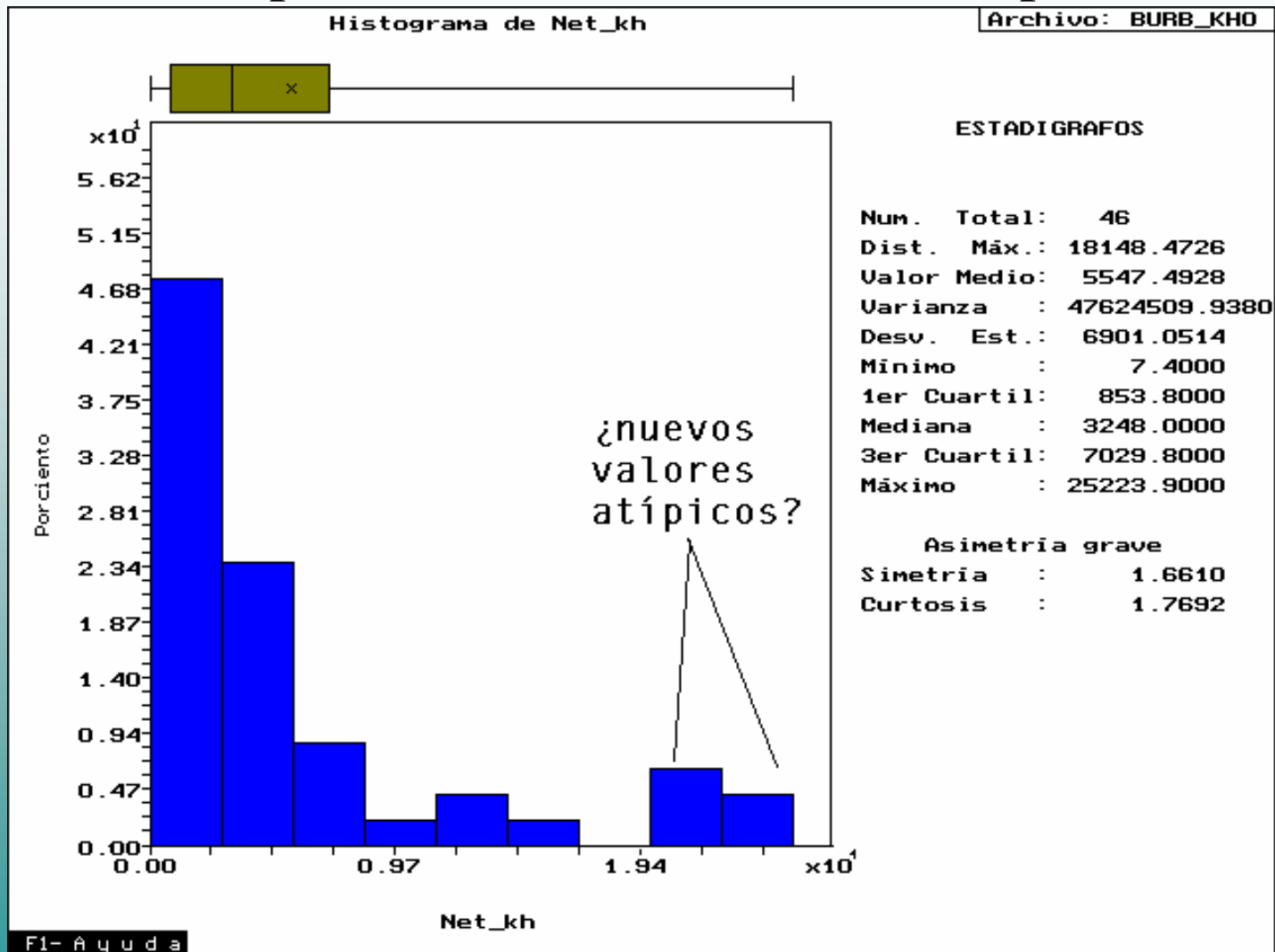
Estadística univariada

¿Serán valores atípicos?



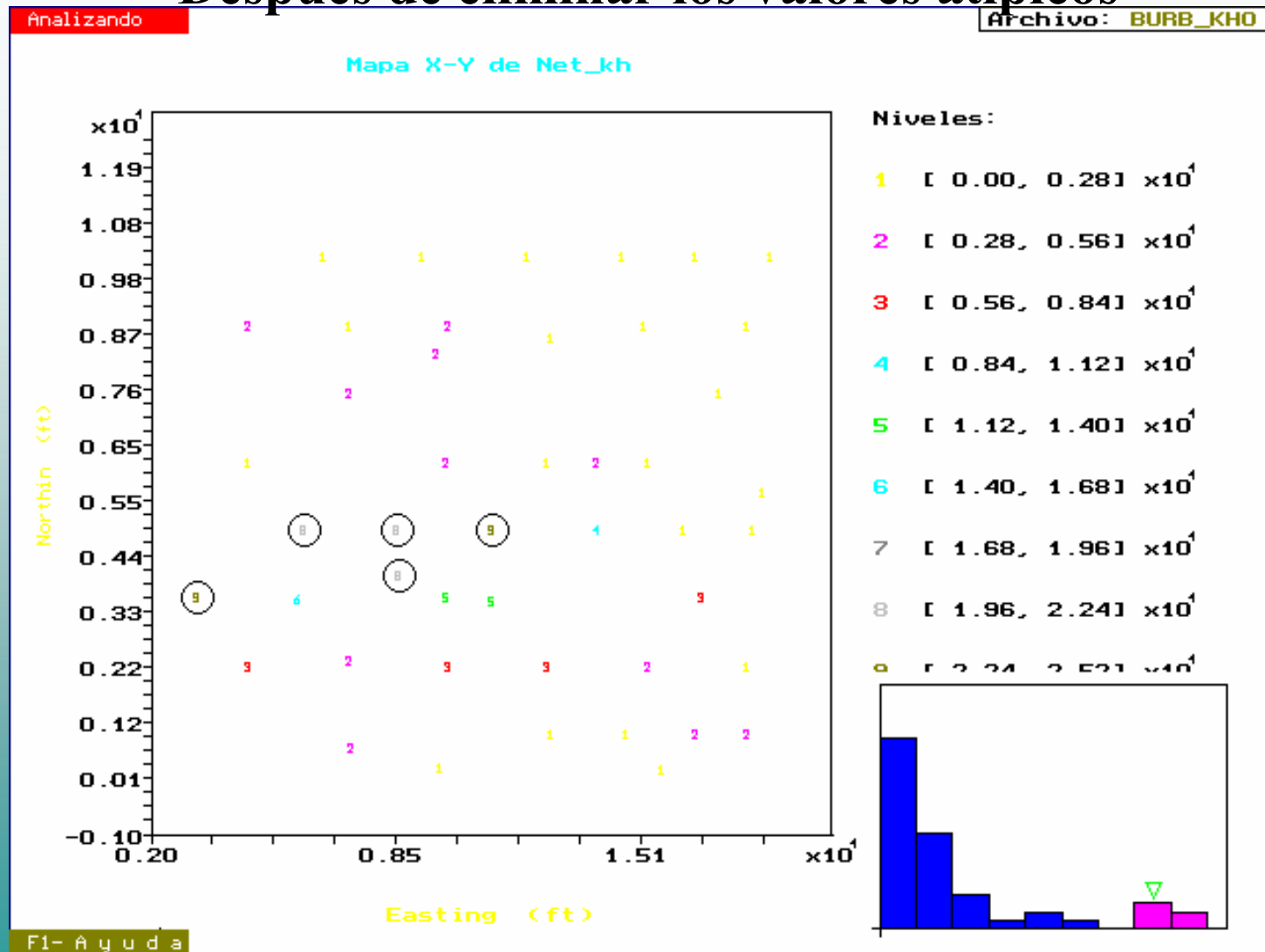
Estadística univariada

Después de eliminar los valores atípicos



Estadística univariada

Después de eliminar los valores atípicos



Estadística bivariada

- Hasta el momento, sólo hemos considerado a las variables aleatorias por separado, sin que exista ninguna interrelación entre éstas.
- En muchos campos de aplicación y en particular, en las Ciencias de la Tierra, es frecuentemente más importante conocer el patrón de dependencia que relaciona a una variable aleatoria X (porosidad) con otra variable aleatoria Y (permeabilidad).
- Por lo que le dedicaremos especial atención al análisis conjunto de dos variables aleatorias, conocido como análisis bivariado.

Estadística bivariada

Función de Distribución de Probabilidad Bivariada

- La distribución de probabilidad conjunta de un par de variables aleatorias \mathbf{X} y \mathbf{Y} se define como:

$$F_{XY}(x, y) = \Pr \{ X \leq x, Y \leq y \}$$

- En la práctica se estima mediante la proporción de pares de valores de \mathbf{X} y \mathbf{Y} que se encuentran por debajo del umbral x, y respectivamente.

Estadística bivariada

- **Diagrama de Dispersión (Scattergram)**
- El equivalente bivariado del histograma es el diagrama de dispersión o scattergram, donde cada par (x_i, y_i) es un punto.
- El grado de dependencia entre dos variables aleatorias X y Y puede ser caracterizado por el diagrama de dispersión alrededor de cualquier línea de regresión.

Estadística bivariada

- **Covarianza**
- Se define la covarianza de manera análoga a los momentos centrales univariados, como

$$Cov(X, Y) = \sigma_{XY} = E \{ (X - m_X)(Y - m_Y) \}$$

- Se calcula como

$$\sigma_{XY} = \frac{1}{N} \sum_{i=1}^N (x_i - m_X)(y_i - m_Y) = \frac{1}{N} \sum_{i=1}^N x_i y_i - m_X m_Y$$

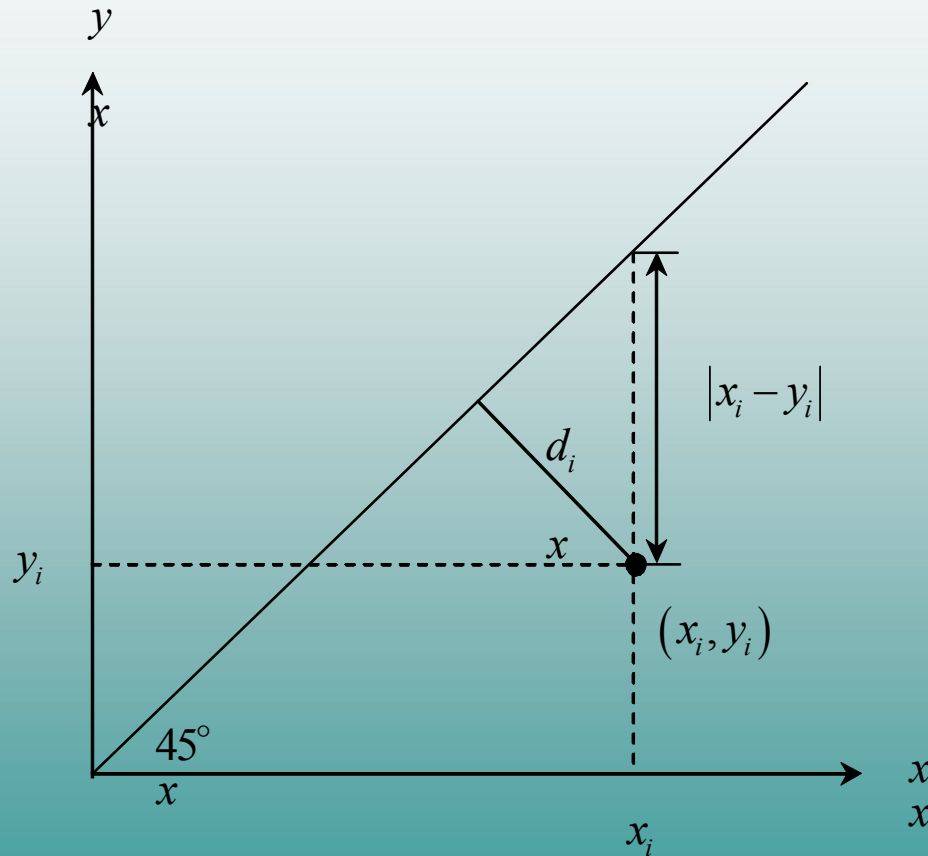
Estadística bivariada

- **Semivariograma**
- Es el momento de inercia del diagrama de dispersión con respecto a una línea con pendiente de 45° y se define como

$$\gamma_{XY} = \frac{1}{N} \sum_{i=1}^N [d_i]^2 = \frac{1}{2N} \sum_{i=1}^N [x_i - y_i]^2$$

- Permite caracterizar la carencia de dependencia

Estadística bivariada



Semivariograma

Mientras mayor sea el valor del semivariograma más dispersos estarán los valores en el diagrama de dispersión y menor será la dependencia entre las dos variables aleatorias.

Estadística bivariada

- **Coeficiente de correlación lineal de Pearson**
- Se define como:

$$r_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{Cov\{X, Y\}}{\sqrt{Var\{X\} Var\{Y\}}} \in [-1, 1]$$

- Caracteriza el grado de dependencia lineal entre dos variables aleatorias.
- Por ejemplo si $Y=aX+b$, entonces se cumple que:

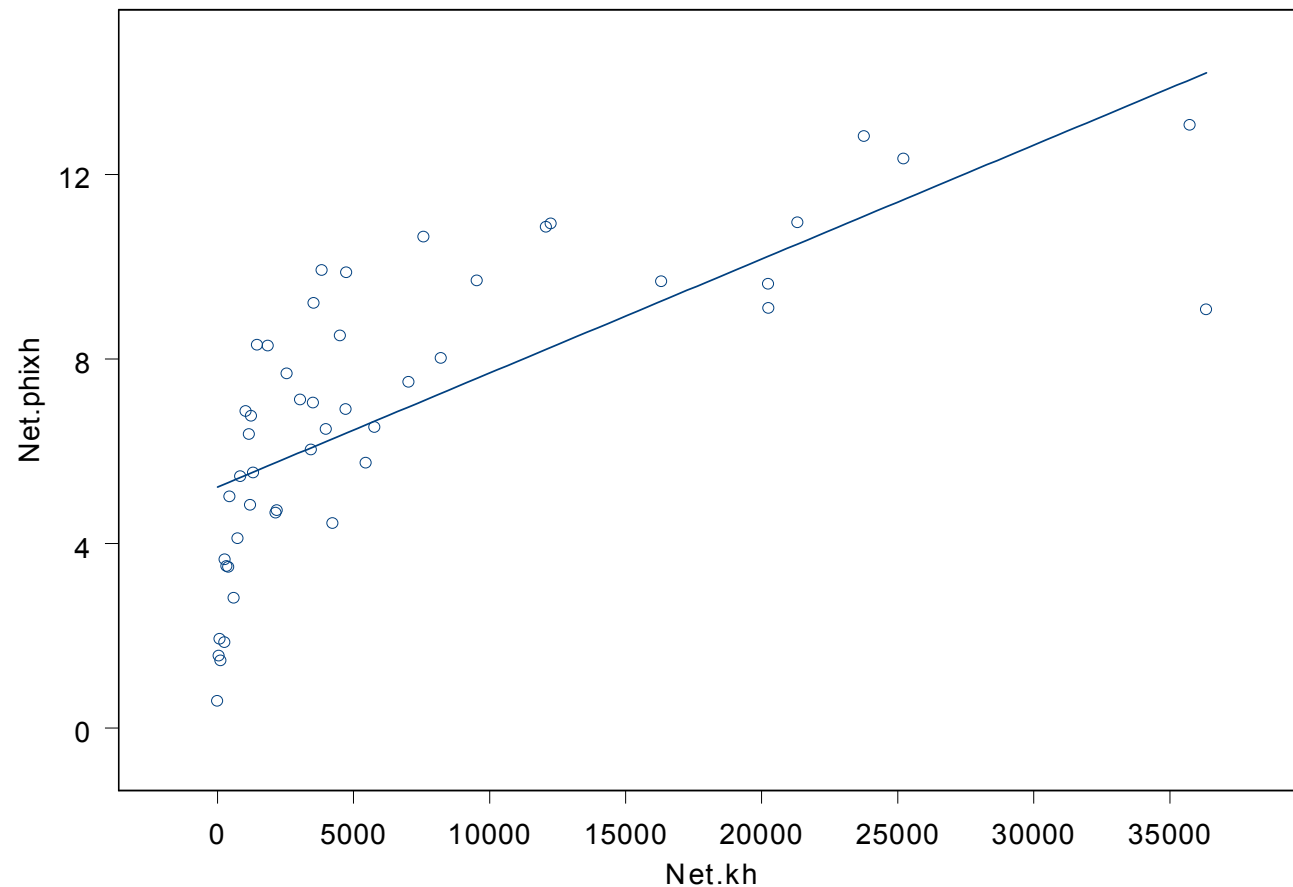
$$r_{XY} = \begin{cases} 1, & \text{para } a > 0 \\ -1, & \text{para } a < 0 \end{cases}$$

Estadística bivariada

Diagrama de Dispersión (Scattergram)

Coeficiente de correlación=0.716875

Antes de transformar

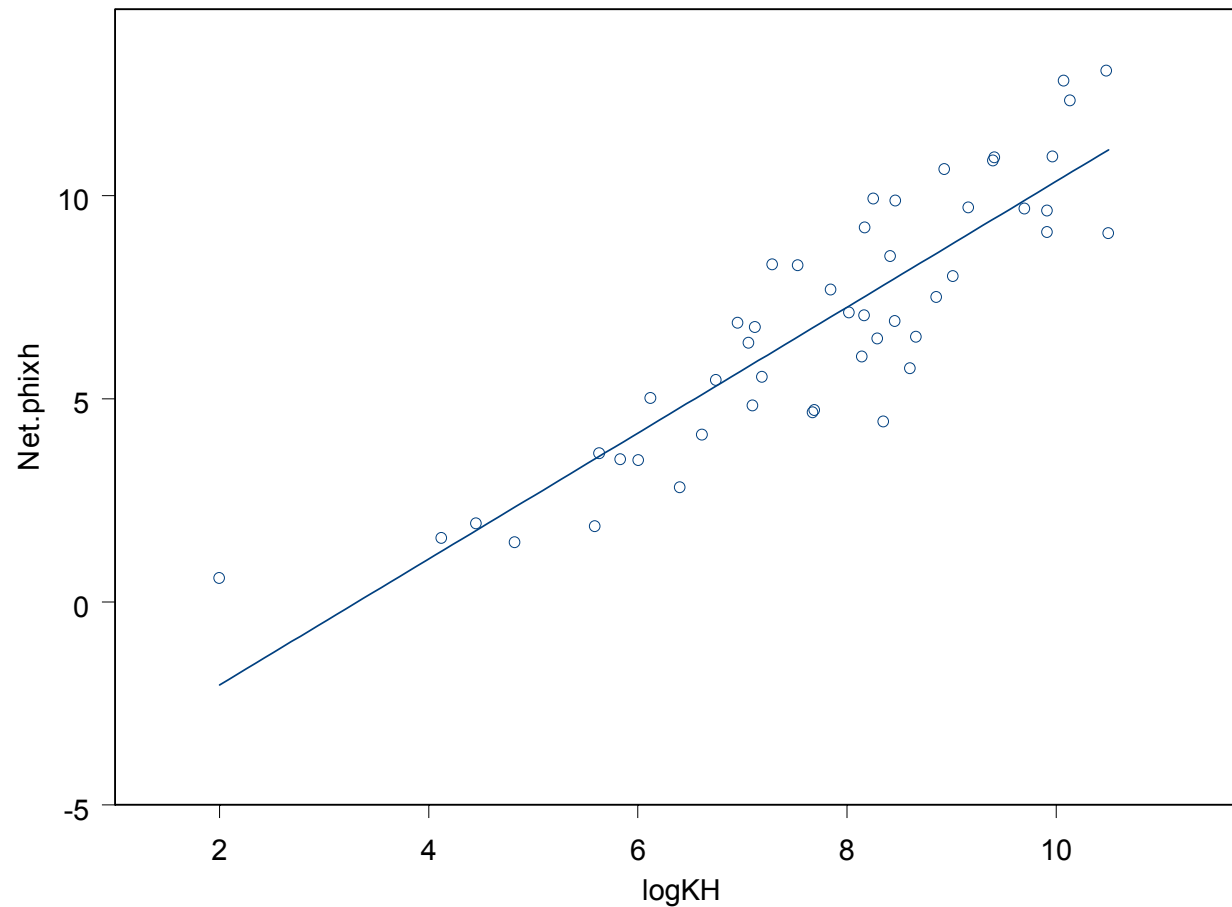


Estadística bivariada

Diagrama de Dispersión (Scattergram)

Coeficiente de correlación=0.8819055

Después de transformar



Estadística multivariada

Existen muchas técnicas multivariadas:

- Análisis de Regresión
- Análisis de Conglomerados
- Análisis de Componentes Principales
- Análisis Factorial
- Análisis Discriminante, etc

Regresión lineal y Mínimos cuadrados

- La *regresión* trata de establecer relaciones funcionales entre variables aleatorias.
- En particular la *regresión lineal* consiste en establecer una relación descrita mediante una recta.
- Los *modelos de regresión* nos permiten hacer predicciones o pronósticos a partir del modelo establecido.
- El método que se emplea para estimar los parámetros del modelo de regresión es el de los *Mínimos Cuadrados*

Regresión lineal

- Dados N valores de dos v.a. X y Y .

Suponemos que:

1. X es una variable independiente
2. Y depende de X en forma lineal

Modelo lineal: $Y = \beta_0 + \beta_1 X$

Donde

$$y_i = \beta_0 + \beta_1 x_i + e_i, \quad i=1, \dots, N$$

β_0, β_1 – son los parámetros del modelo

e_i – errores o residuos del modelo

Regresión lineal

- Condiciones que deben cumplir los residuos

$$E\{e_i\} = 0, \quad (\text{valor esperado cero})$$

$$Var\{e_i\} = \sigma_e^2, \quad (\text{varianza constante})$$

$$Cov\{e_i, e_j\} = 0, \quad \forall i \neq j, \quad (\text{no correlacionados})$$

$$e \sim N(0, \sigma_e^2), \quad (\text{distribución normal})$$

Mínimos Cuadrados Ordinarios (MCO)

- *Mínimos Cuadrados Ordinarios* consiste en hallar los parámetros del modelo de manera que la suma de los cuadrados de los errores sea mínima.

$$SCR = \sum_{i=1}^N e_i^2 = \sum_{i=1}^N [y_i - \hat{y}_i]^2 = \sum_{i=1}^N \left[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \right]^2$$

- Sistema de ecuaciones a resolver

$$\frac{\partial SCR}{\partial \beta_0} = 0, \quad \frac{\partial SCR}{\partial \beta_1} = 0$$

Mínimos Cuadrados Ordinarios (MCO)

Coeficiente de determinación R^2

- Para los modelos lineales
 1. Mide el *grado de la bondad del ajuste*
 2. Es igual al coeficiente de correlación lineal al cuadrado.
 3. Representa la proporción de varianza explicada por la regresión lineal.

Mínimos Cuadrados Ordinarios (MCO)

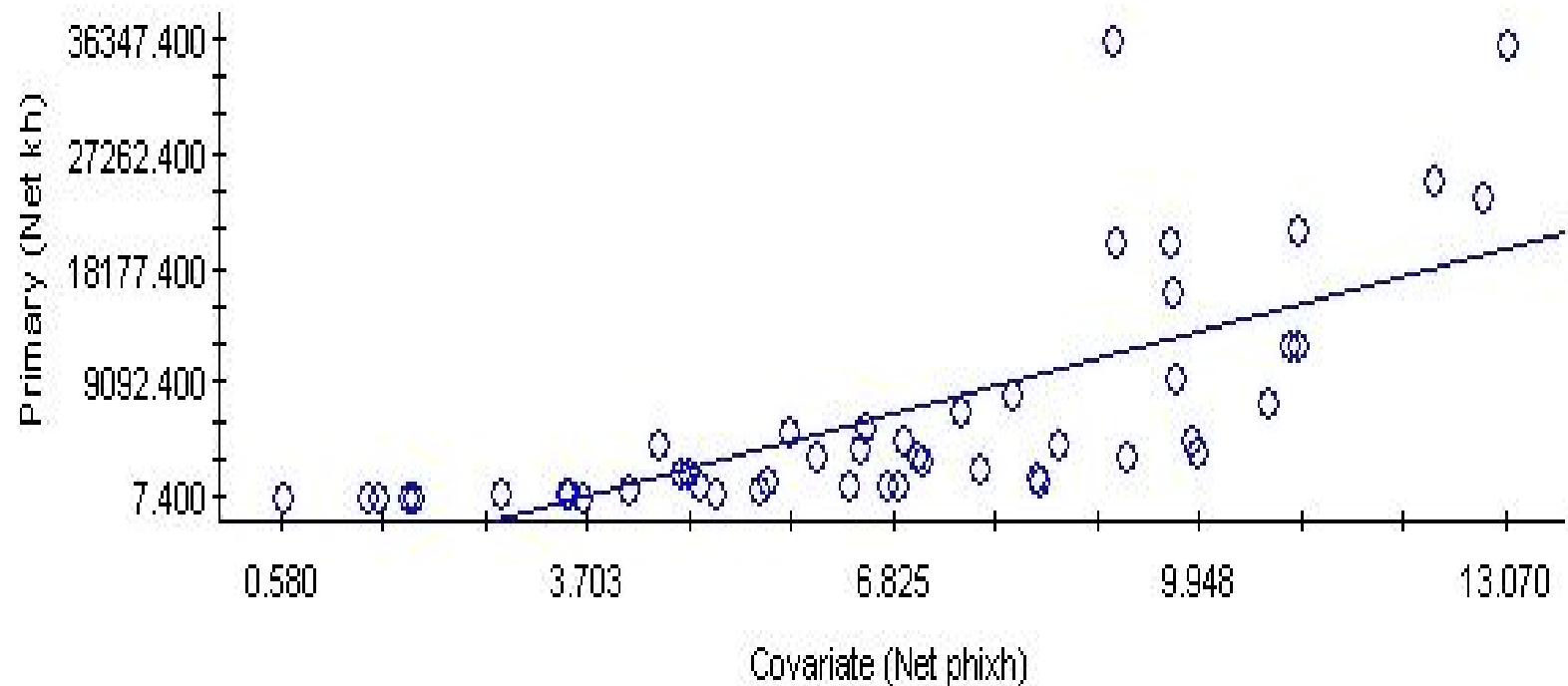
Criterios de la bondad del ajuste

- Si $R^2 \approx 1$, el ajuste es bueno (Y se puede calcular de modo bastante aproximado a partir de X y viceversa).
- Si $R^2 \approx 0$, las variables X y Y no están relacionadas (linealmente al menos), por tanto no tiene sentido hacer un ajuste lineal.
- Sin embargo no es seguro que las dos variables no posean ninguna relación en el caso $r=0$, ya que si bien el ajuste lineal puede no ser procedente, tal vez otro tipo de ajuste sí lo sea.

Regresión lineal

Y=Permeabilidad, X=Porosidad

Antes de transformar

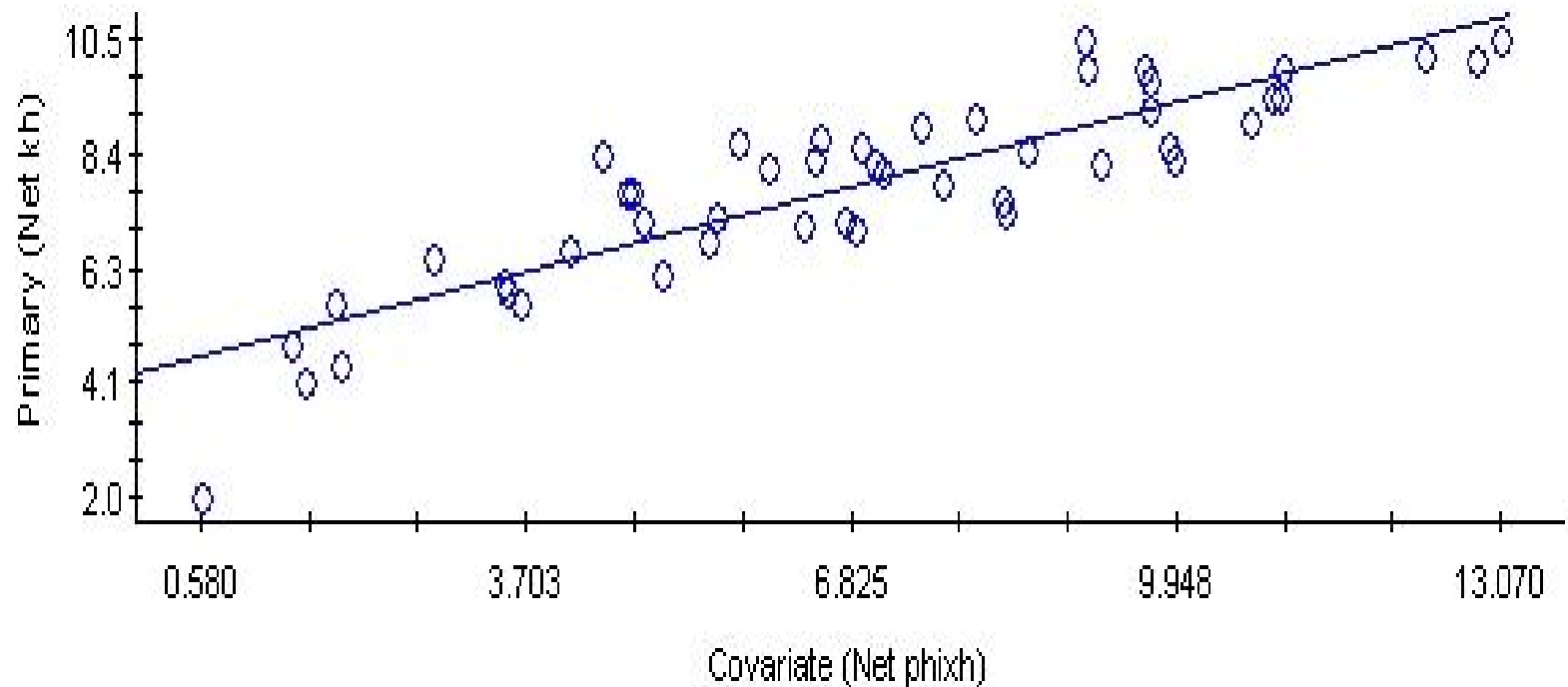


Regression coefficient = 2080.0 (SE = 298.1, $r^2=0.514$, y intercept = -7550.6945)

Regresión lineal

Y=logPermeabilidad, X=Porosidad

Después de transformar

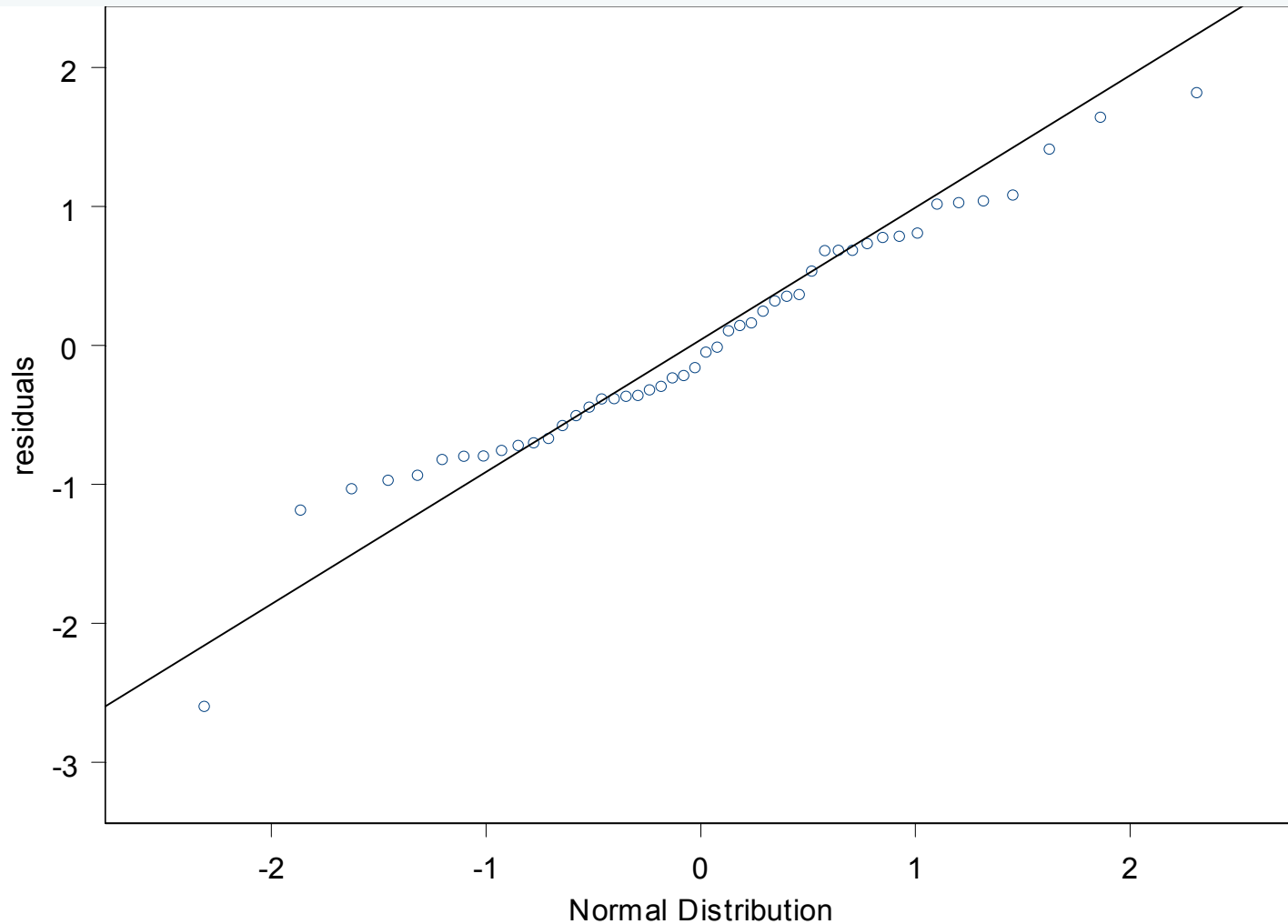


Regression coefficient = 0.50 (SE = 0.04, $r^2 = 0.778$, y intercept = 4.31, n = 48)

Regresión lineal

$Y = \log \text{Permeabilidad}$, $X = \text{Porosidad}$

Análisis de los residuos

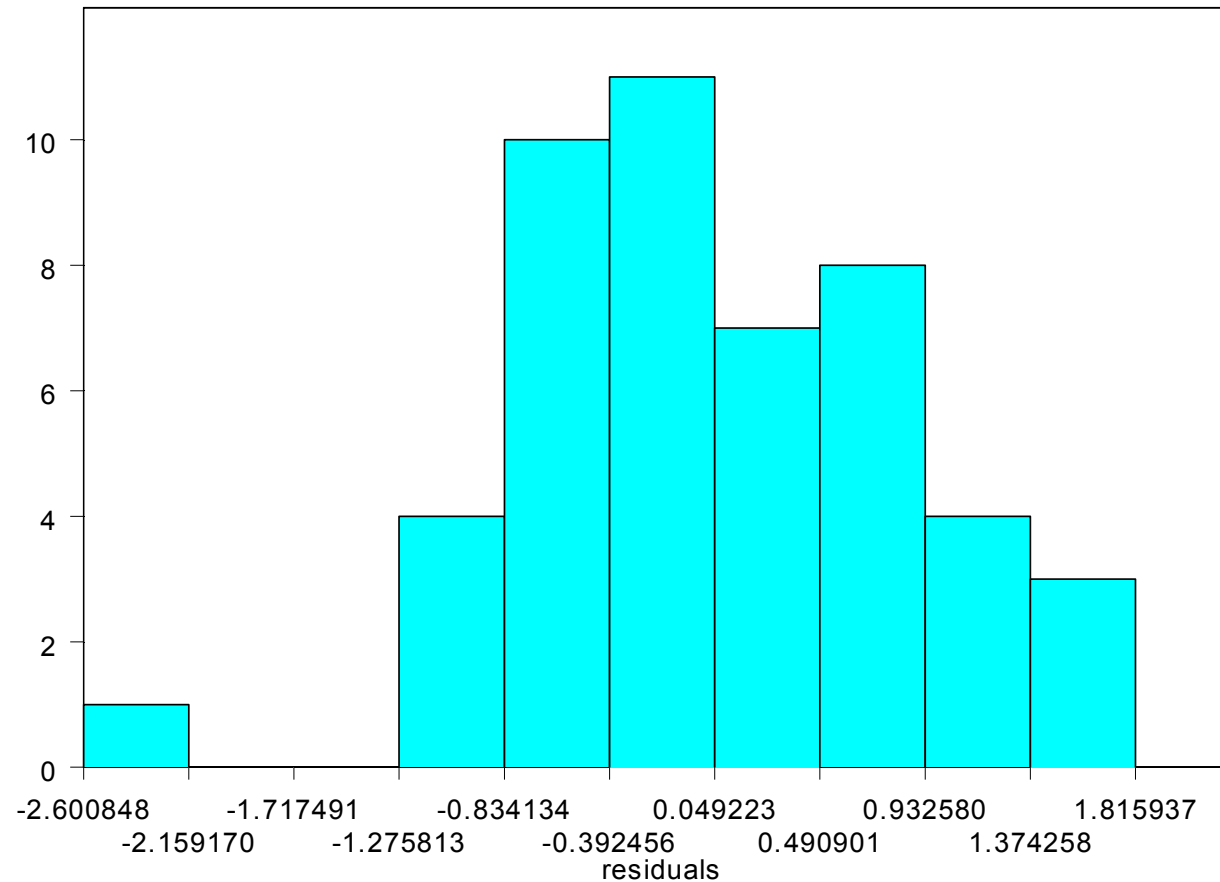


Regresión lineal

Y=logPermeabilidad, X=Porosidad

Análisis de los residuos

Estadística Residuos



Min: -2.600848e+000
1st Qu.: -6.021758e-001
Mean: 1.156482e-018
Median: -1.075592e-001
3rd Qu.: 6.815479e-001
Max: 1.815937e+000
Total N: 4.800000e+001
Variance: 7.146409e-001
Std Dev.: 8.453644e-001
SE Mean: 1.220178e-001
Skewness: 1.878733e-001
Kurtosis: 6.868942e-001