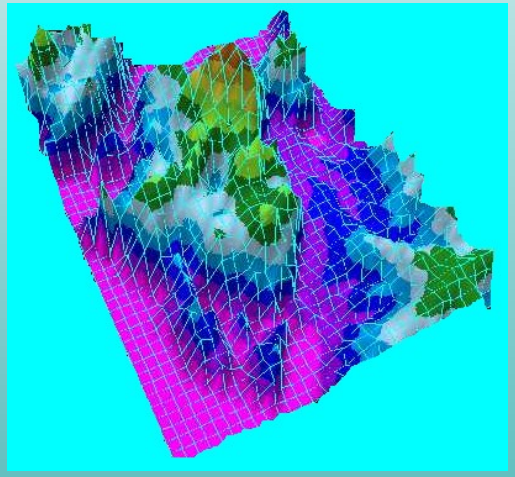
UNIVERSIDAD NACIONAL AUTÓNOMA
DE MÉXICO

GEOESTADÍSTICA APLICADA



Tema: Análisis Exploratorio de Datos

Instructores:

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

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

2020

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 multivariada*
- *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 o 6} con una probabilidad de ocurrencia igual a 1/6.
- Las probabilidades cumplen las condiciones:

$$a)\ p_i \geq 0, \qquad \forall i \qquad\qquad b)\ \sum_i p_i = 1$$

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%].

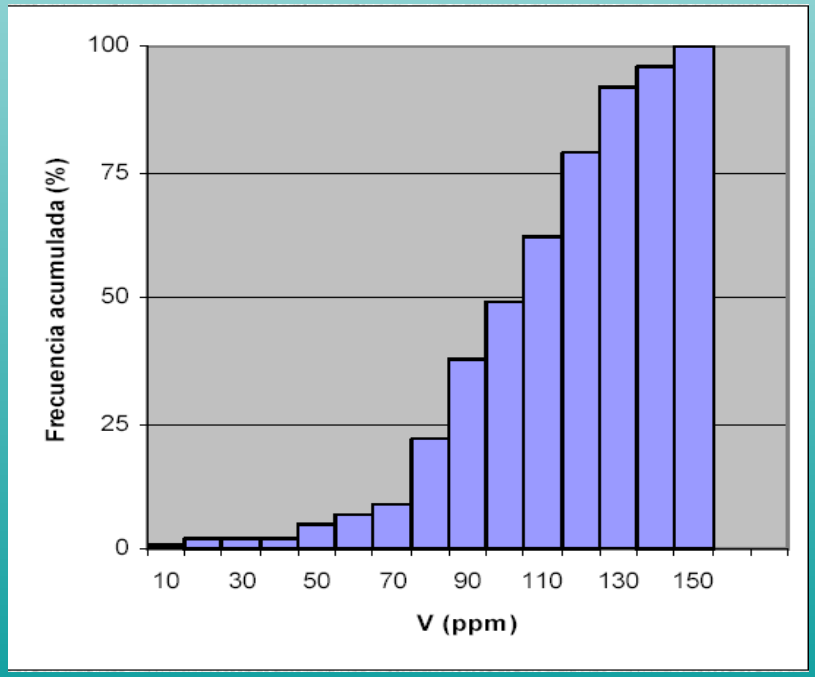
Estadística univariada

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



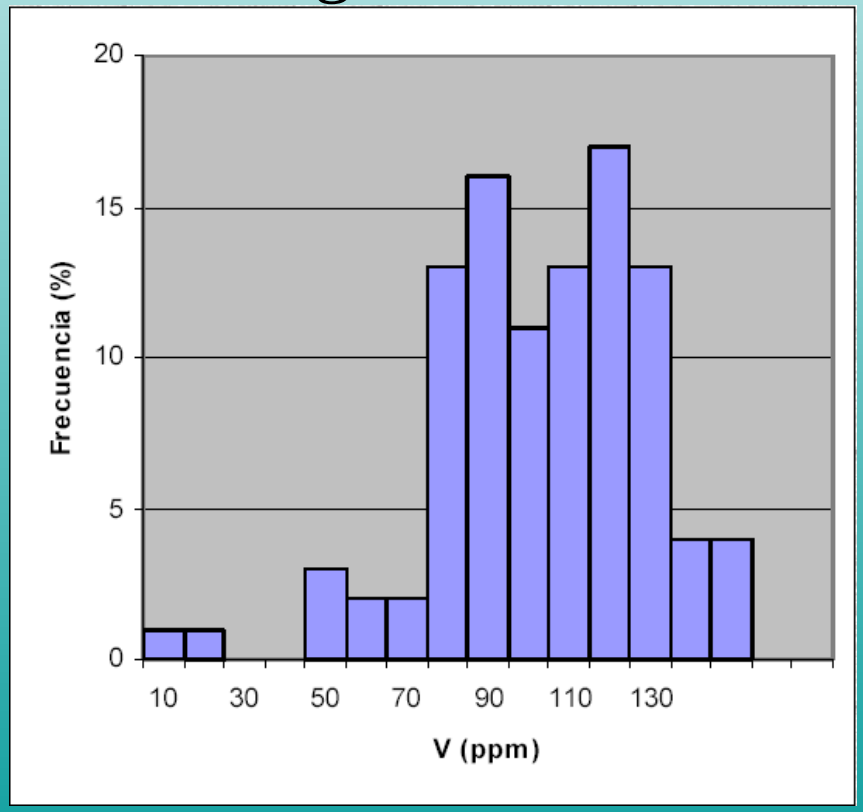
Estadística univariada

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:

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

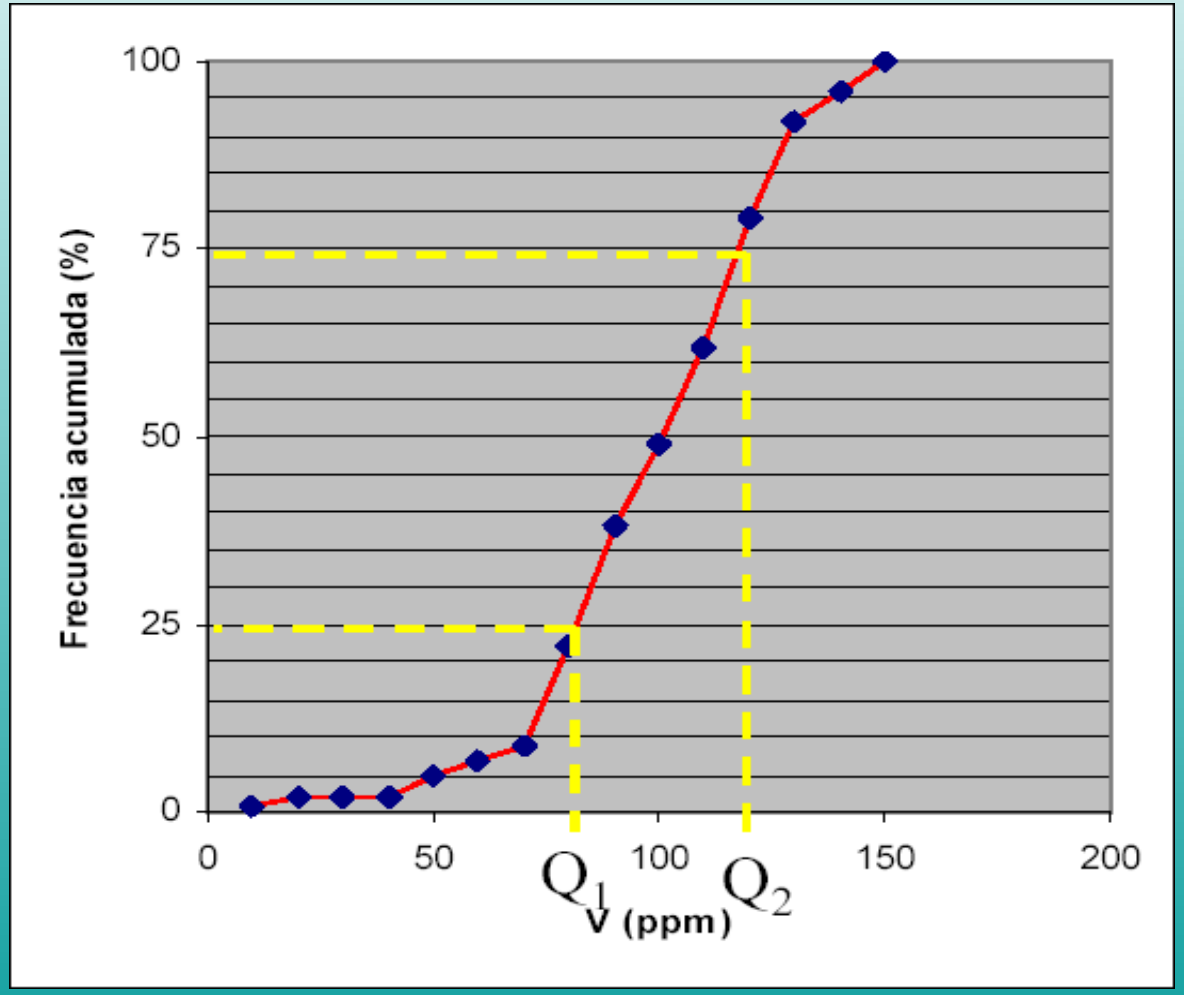
Estadística univariada

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) $\left[z_{0.25}, z_{0.75} \right]$

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:

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

Se calcula como el promedio de todas las observaciones de la variable Z

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

Es muy sensible a los valores atípicos (*outliers*)

Estadística univariada

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

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

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

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

Estadística univariada

Varianza de una VA (2do momento centrado)

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

Estadística univariada

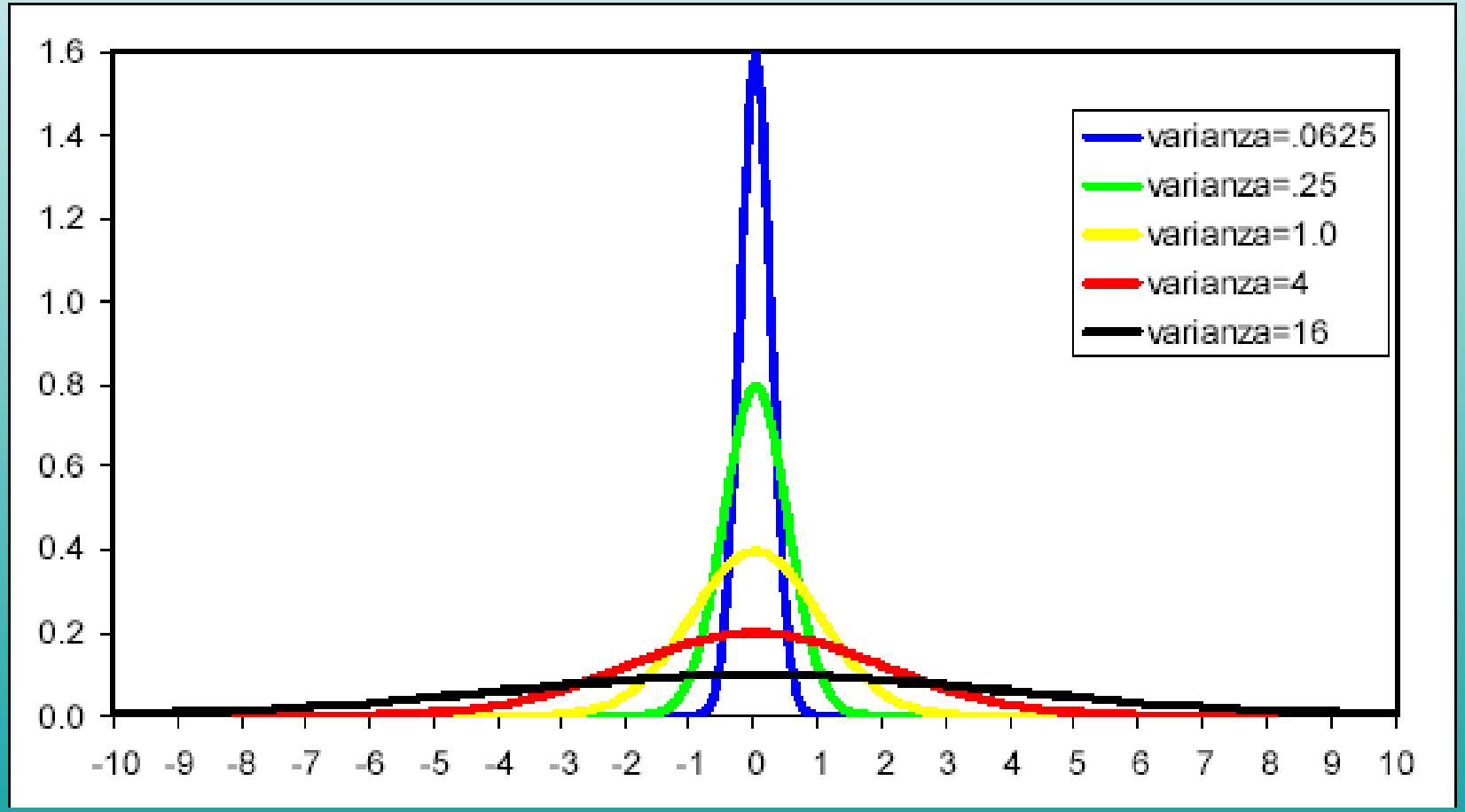
- **Distribución Normal o Gaussiana.**
- Esta distribución está completamente caracterizada por sus dos parámetros: media y varianza y se designa mediante $N\left(m,\sigma^2\right)$
- La *fdp* normal o Gaussina está dada por

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

- Es simétrica respecto a la media

Estadística univariada

Ejemplos de distribuciones Gaussianas



Estadística univariada

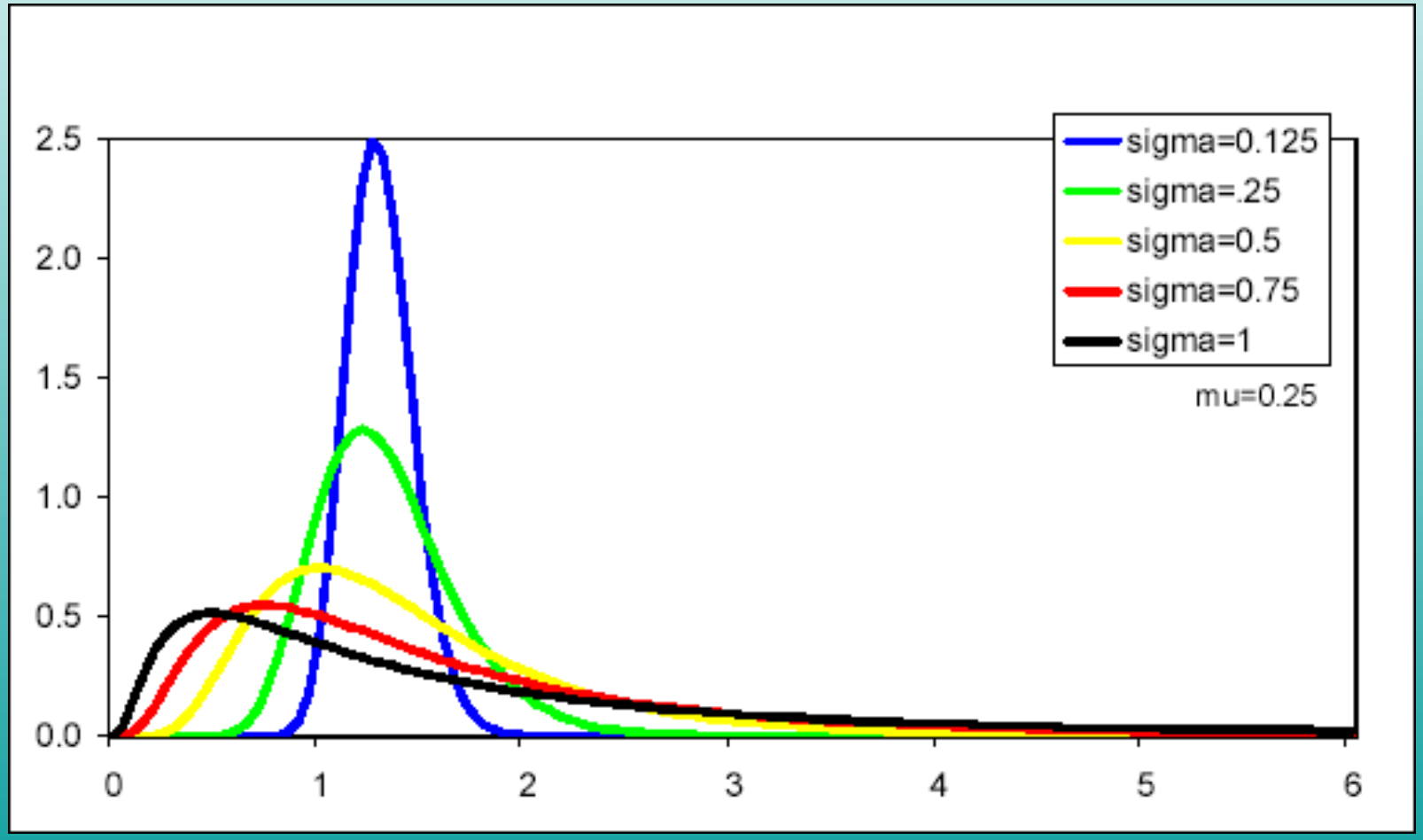
- **Distribución LogNormal**
- Una VA positiva ***Y*** se dice que tiene una distribución lognormal si su logaritmo ln(***Y***) esta normalmente distribuido.

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

- 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** $\sigma = \sqrt{\mathrm{Var}[Z]}$
- **Coeficiente de variación (dispersión relativa)**

$$CV = \sigma / m$$

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

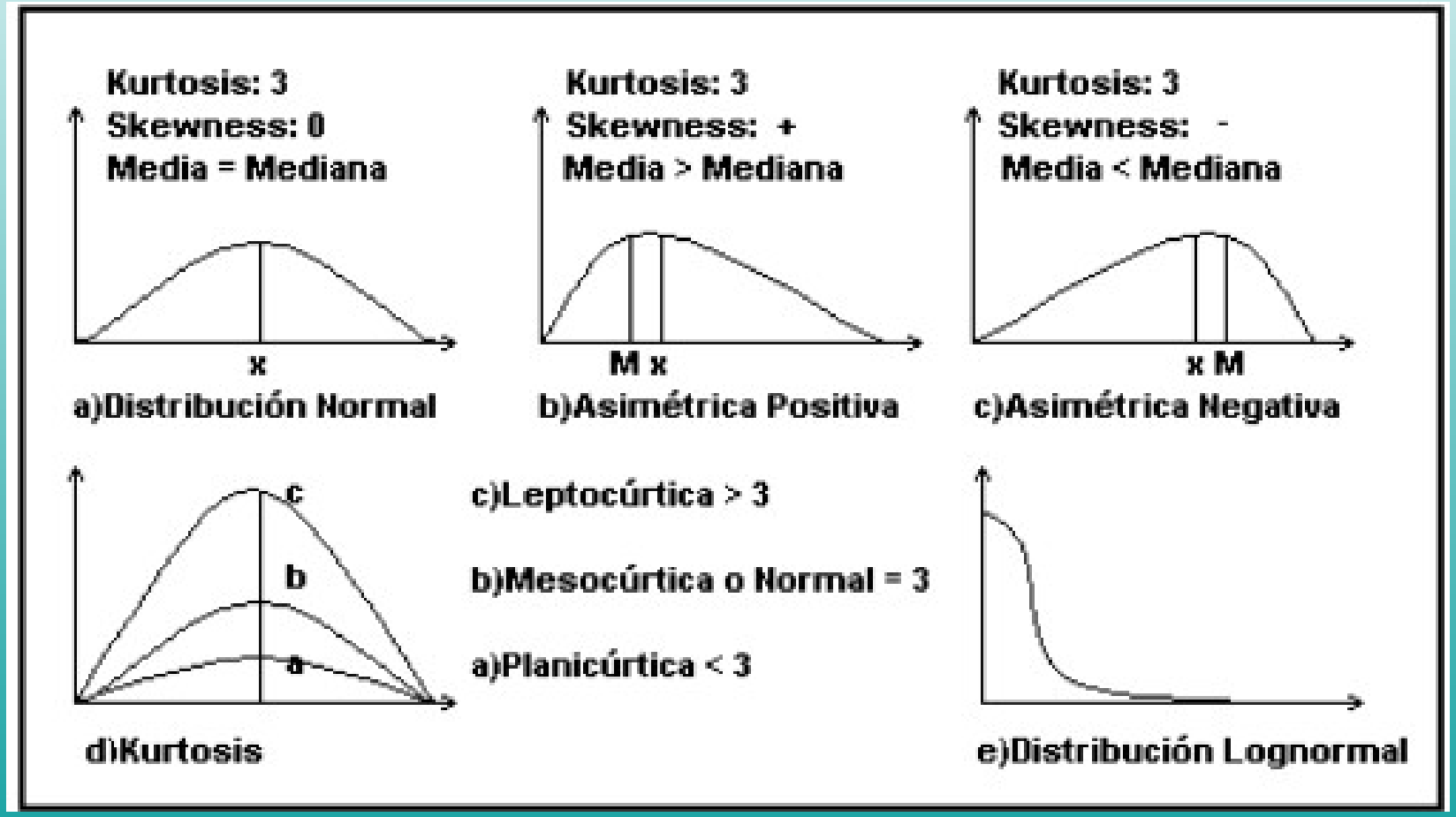
$$\alpha_1 = \frac{\mu_3}{\mu_2^{3/2}}$$

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

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

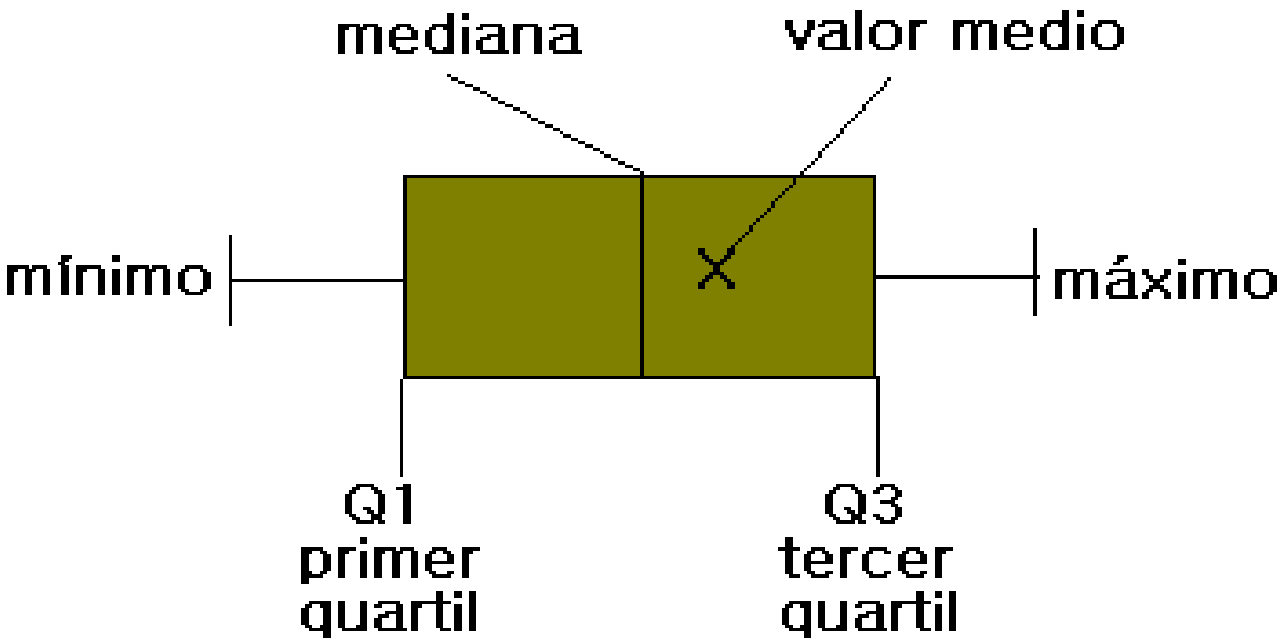
Estadística univariada

Simetría y Curtosis de una distribución



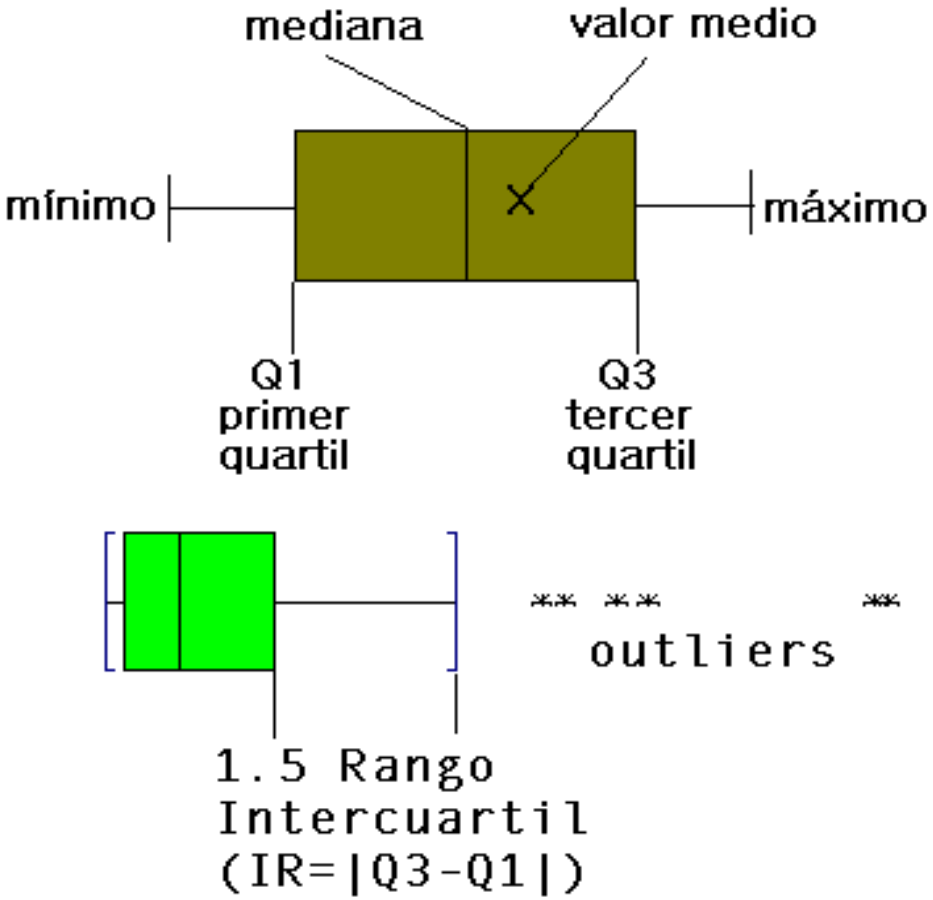
Estadística univariada

BOX PLOT



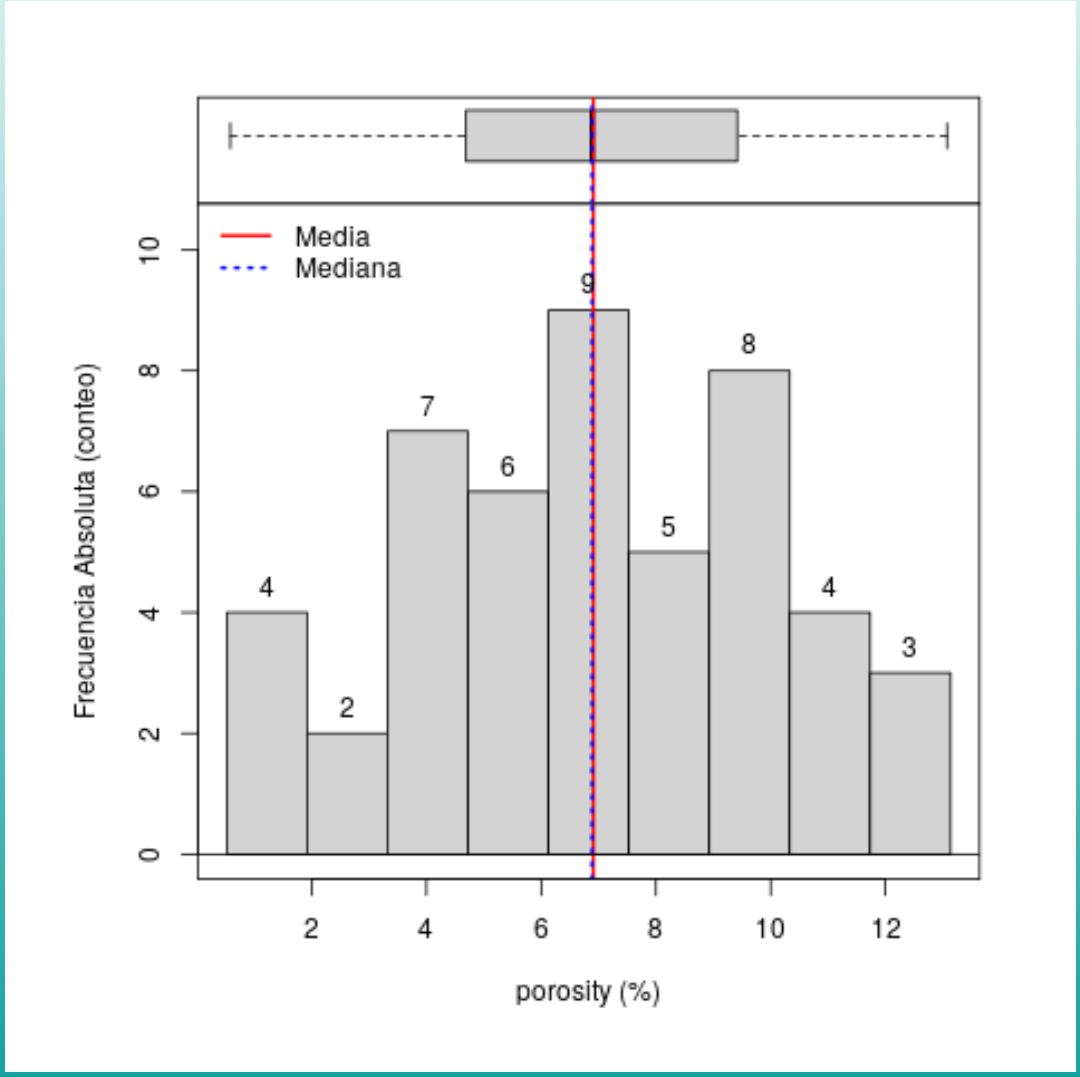
Estadística univariada

BOX PLOT



Estadística univariada

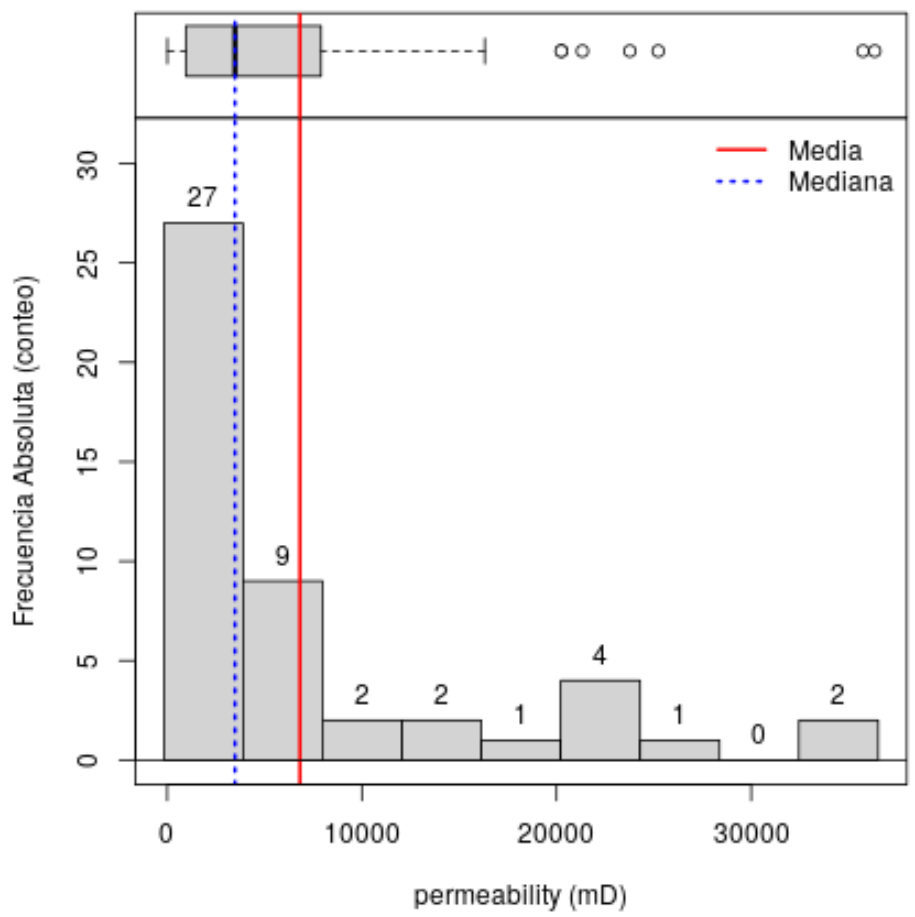
Histograma (Porosidad)



No_muestras	48
Minimo	0.58
Cuartil_1er	4.70
Mediana	6.88
Media	6.90
Cuartil_3er	9.31
Maximo	13.07
Rango	12.49
Rango_Intercuartil	4.61
Varianza	9.92476
Desv_Estandar	3.15036
Simetria	-0.05236

Estadística univariada

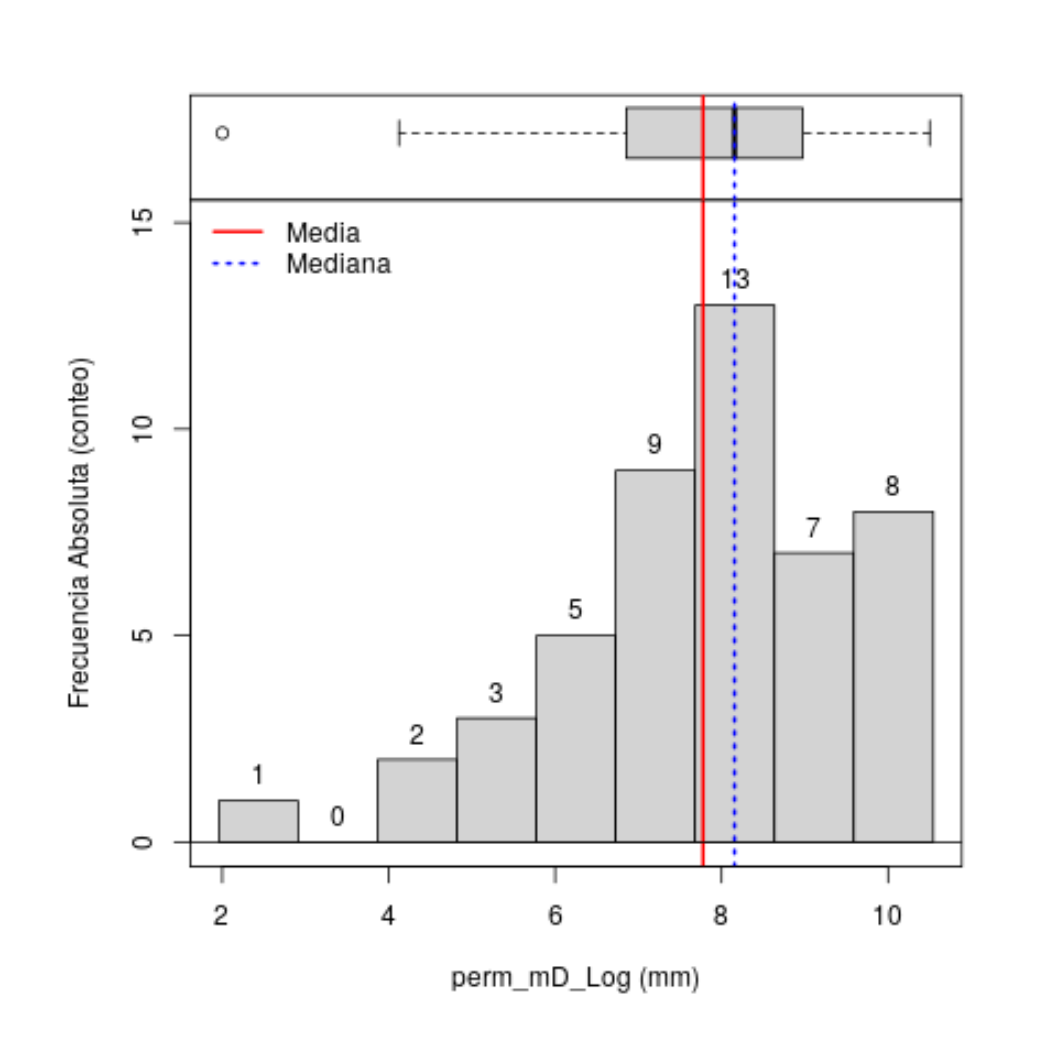
Histograma (Permeabilidad)



No_muestras	48
Minimo	7.4
Cuartil_1er	1002.45
Mediana	3482.205
Media	6818.24521
Cuartil_3er	7743.625
Maximo	36347.4
Rango	36340
Rango_Intercuartil	6741.175
Varianza	83532706.36
Desv_Estandar	9139.62288
Simetria	1.83579
Curtosis	5.62603

Estadística univariada

Transformación logarítmica de la Permeabilidad

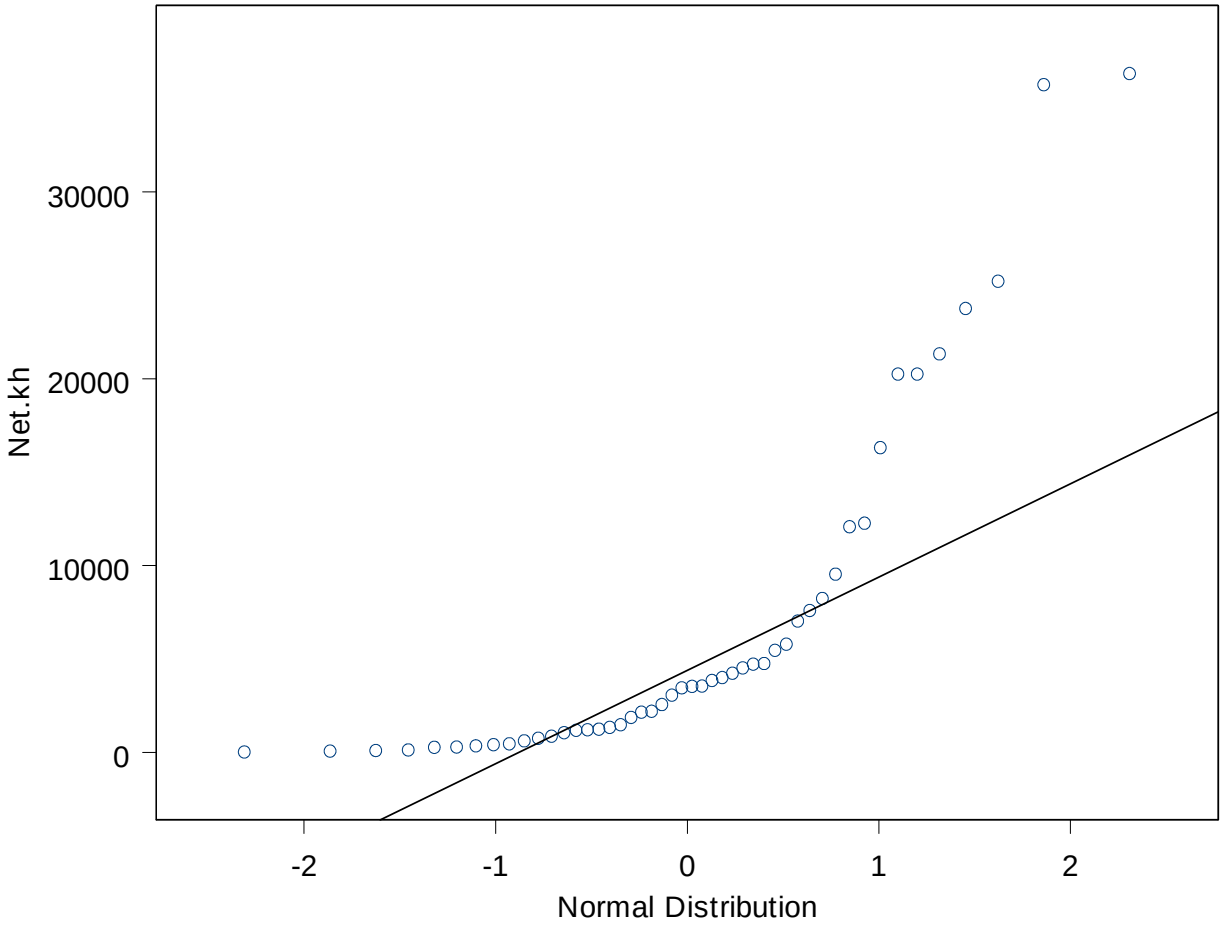


No_muestras	48
Minimo	2.0015
Cuartil_1er	6.9063
Mediana	8.1554
Media	7.7777
Cuartil_3er	8.954
Maximo	10.5009
Rango	8.4994
Rango_Intercuartil	2.0477
Varianza	3.2156
Desv_Estandar	1.7932
Simetria	-0.8484

Estadística univariada

Q-Q Plot de la Permeabilidad

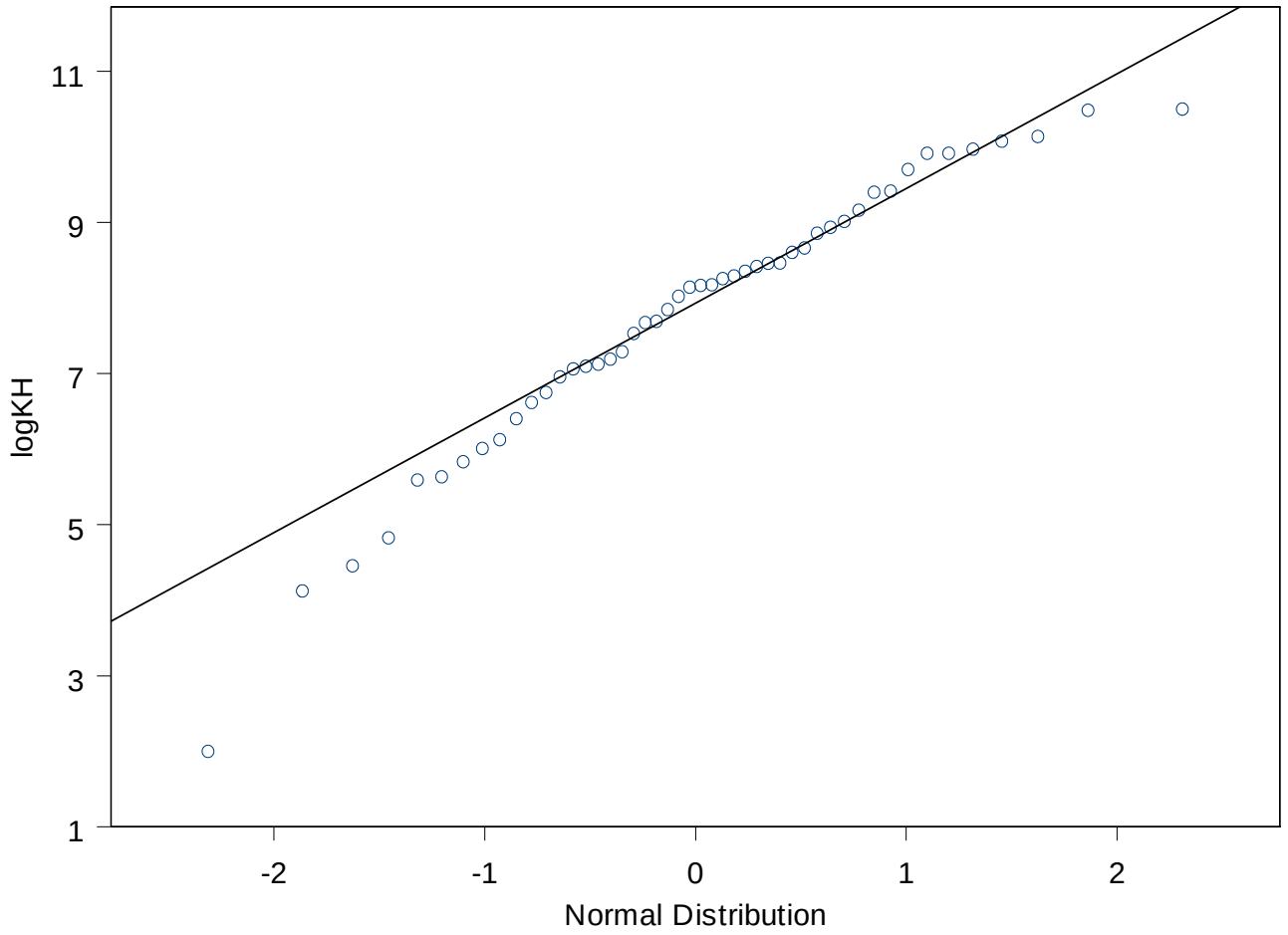
Antes de transformar



Estadística univariada

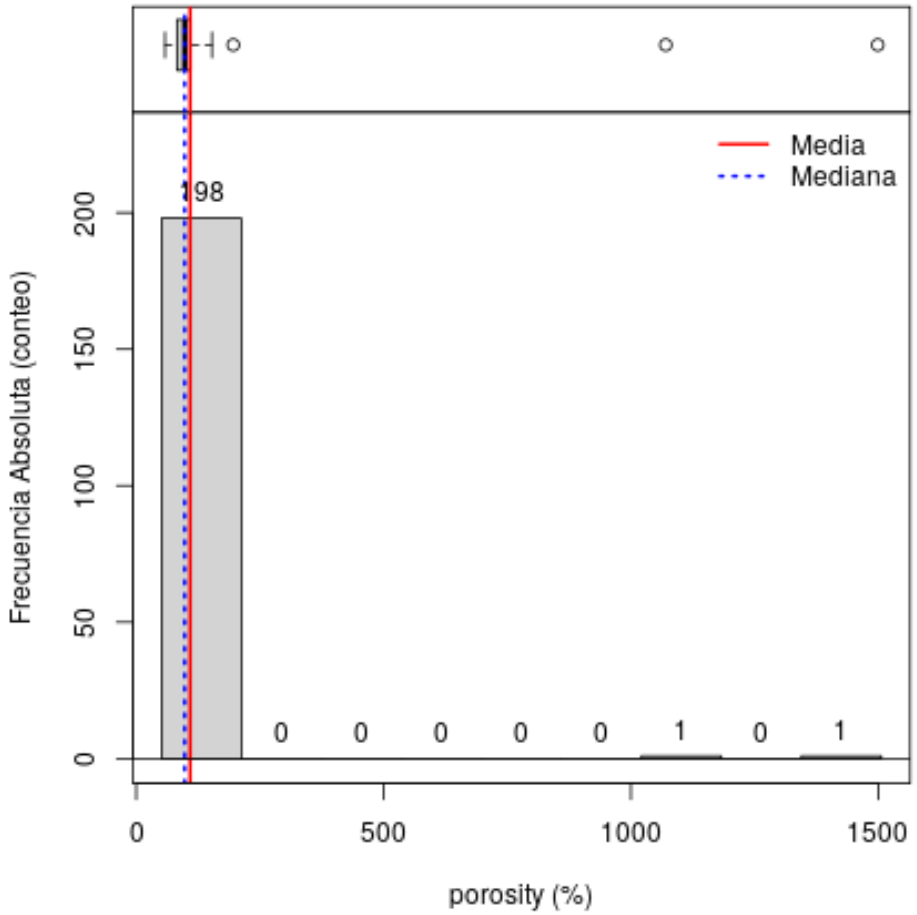
Q-Q Plot de la Permeabilidad

Después de transformar



Estadística univariada

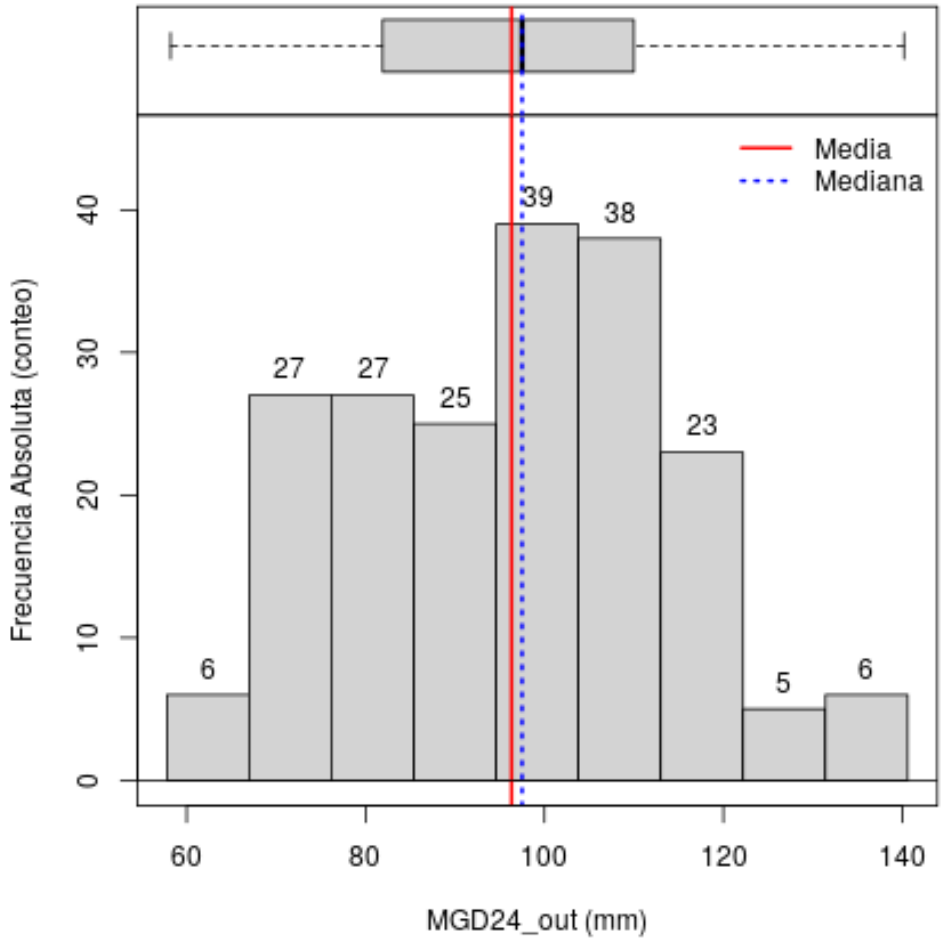
Con valores atípicos (outliers)



No_muestras	200
Minimo	58.2
Cuartil_1er	82.25
Mediana	97.85
Media	108.9925
Cuartil_3er	110.325
Maximo	1499
Rango	1440.8
Rango_Intercuartil	28.075
Varianza	14873.08823
Desv_Estandar	121.95527
Simetria	9.92162
Curtosis	104.73871

Estadística univariada

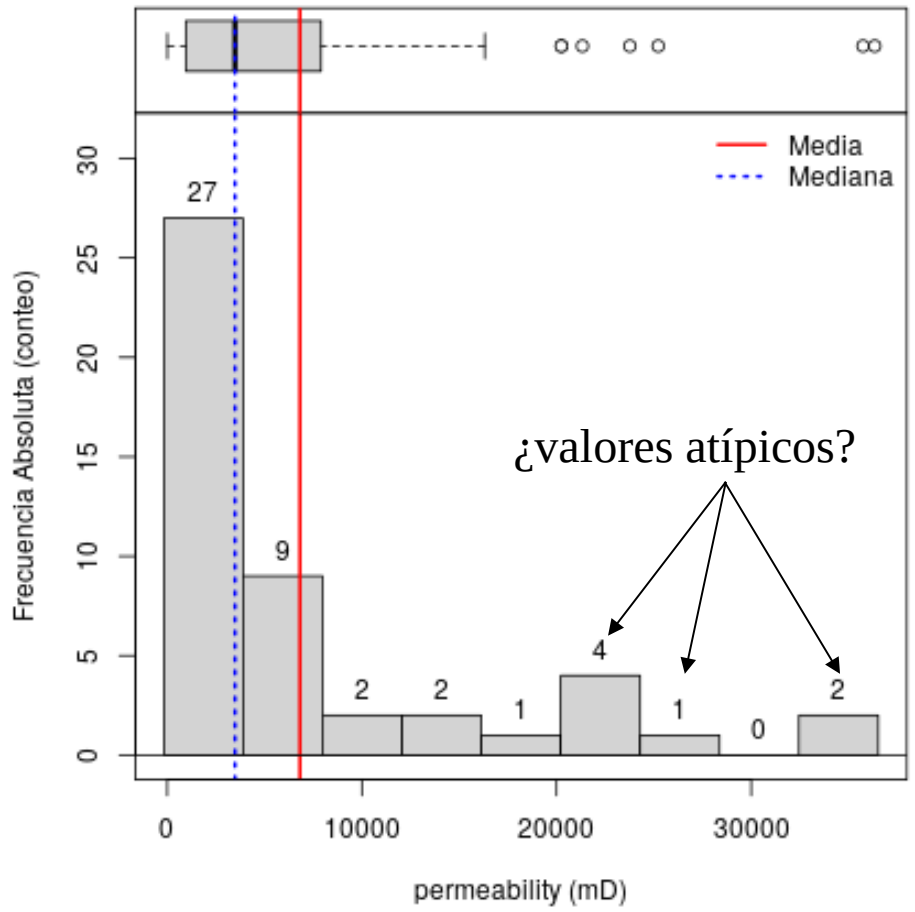
Sin valores atípicos (outliers)



No_muestras	196
Minimo	58.2
Cuartil_1er	82
Mediana	97.5
Media	96.3265
Cuartil_3er	110
Maximo	140.2
Rango	82
Rango_Intercuartil	28
Varianza	319.7503
Desv_Estandar	17.8816
Simetria	0.0291
Curtosis	2.3889

Estadística univariada

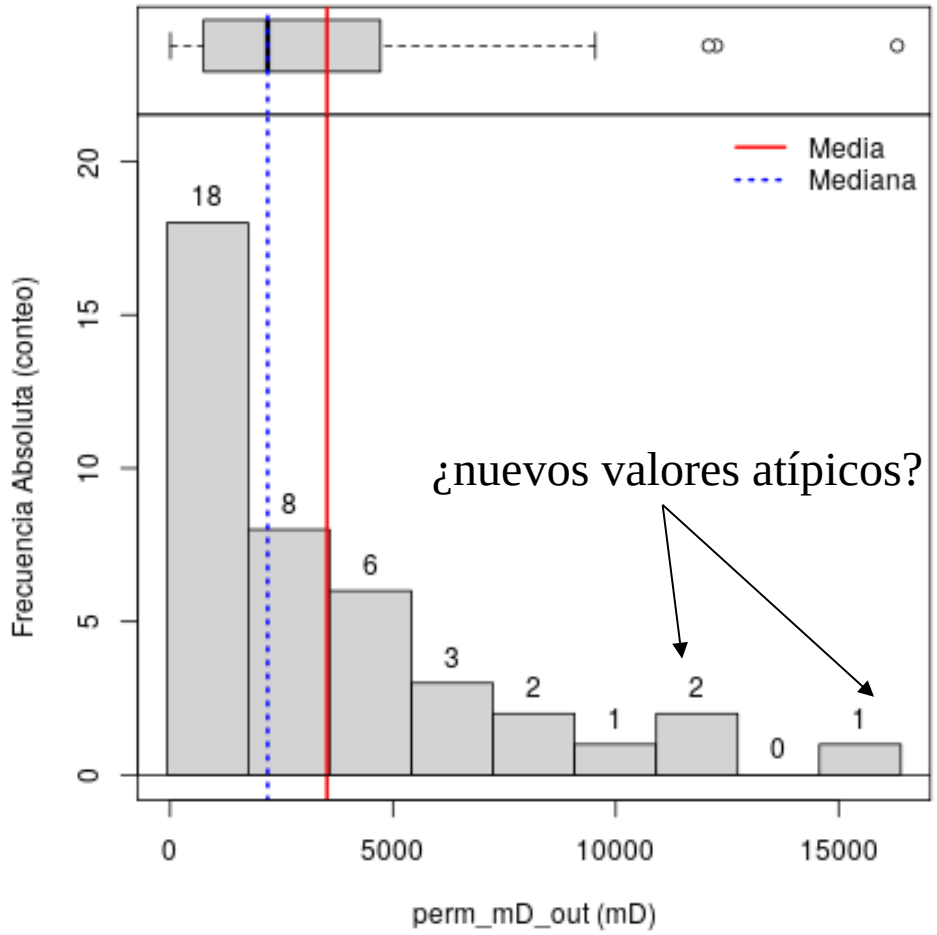
¿Serán valores atípicos?



No_muestras	48
Minimo	7.4
Cuartil_1er	1002.45
Mediana	3482.205
Media	6818.24521
Cuartil_3er	7743.625
Maximo	36347.4
Rango	36340
Rango_Intercuartil	6741.175
Varianza	83532706.36
Desv_Estandar	9139.62288
Simetria	1.83579
Curtosis	5.62603

Estadística univariada

Después de eliminar los valores atípicos



No_muestras	41
Minimo	7.4
Cuartil_1er	748
Mediana	2188.7
Media	3521.0285
Cuartil_3er	4720.5
Maximo	16315.9
Rango	16308.5
Rango_Intercuartil	3972.5
Varianza	14353741.71
Desv_Estandar	3788.6332
Simetria	1.5704
Curtosis	5.1874

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 **X** y **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 **X** y **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

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

- Se calcula como

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

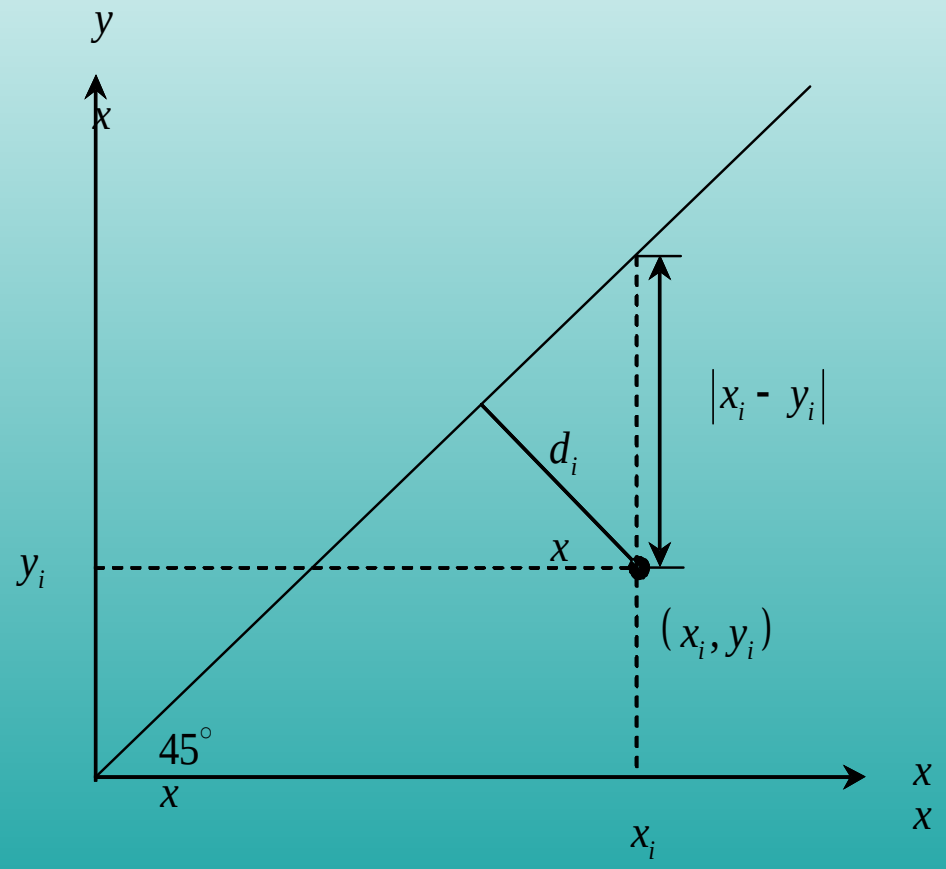
Estadística bivariada

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

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

- 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:

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

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

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

Estadística bivariada

- **Coeficiente de correlación de rango de Spearman**
- Se define como:

$$\rho = 1 - \frac{6 \sum D^2}{N(N^2 - 1)}$$

donde D es la diferencia entre los estadísticos de orden de x - y. N es el número de parejas de datos.

Oscila entre -1 y +1, indicándonos asociaciones negativas o positivas respectivamente, cero, significa no correlación pero no independencia.

Estadística bivariada

- **Coeficiente de correlación de rango de Kendall**
- Se define como:

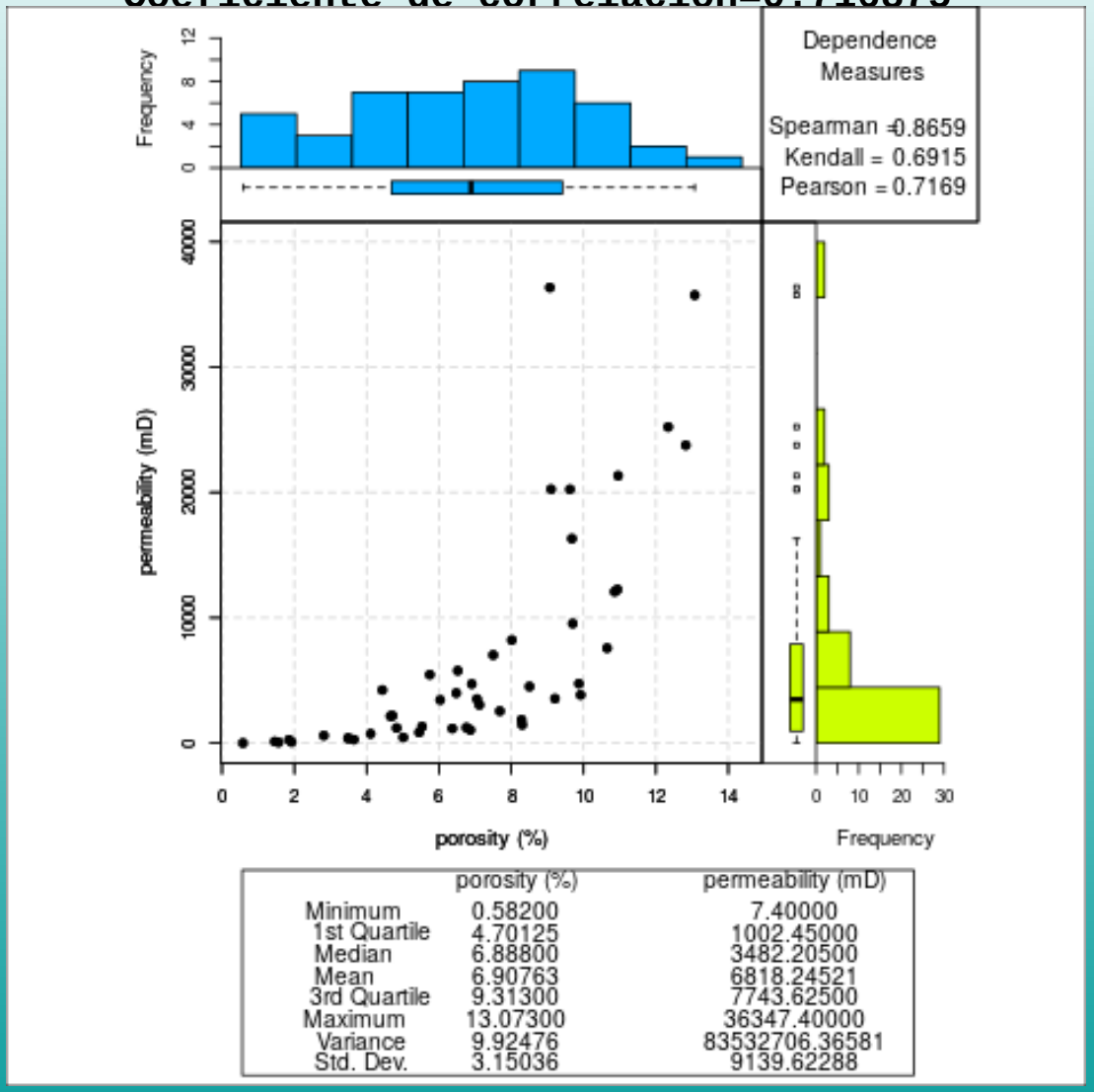
$$\tau = \frac{(\text{número de pares concordantes}) - (\text{número de pares discordantes})}{\binom{n}{2}}.$$

- Un par es concordante si el orden de ambos está de acuerdo de lo contrario se dice que son discordantes.
- Si X y Y son independientes, entonces esperaríamos que el coeficiente sea aproximadamente cero.

Estadística bivariada

Antes de transformar

Coeficiente de correlación=0.716875

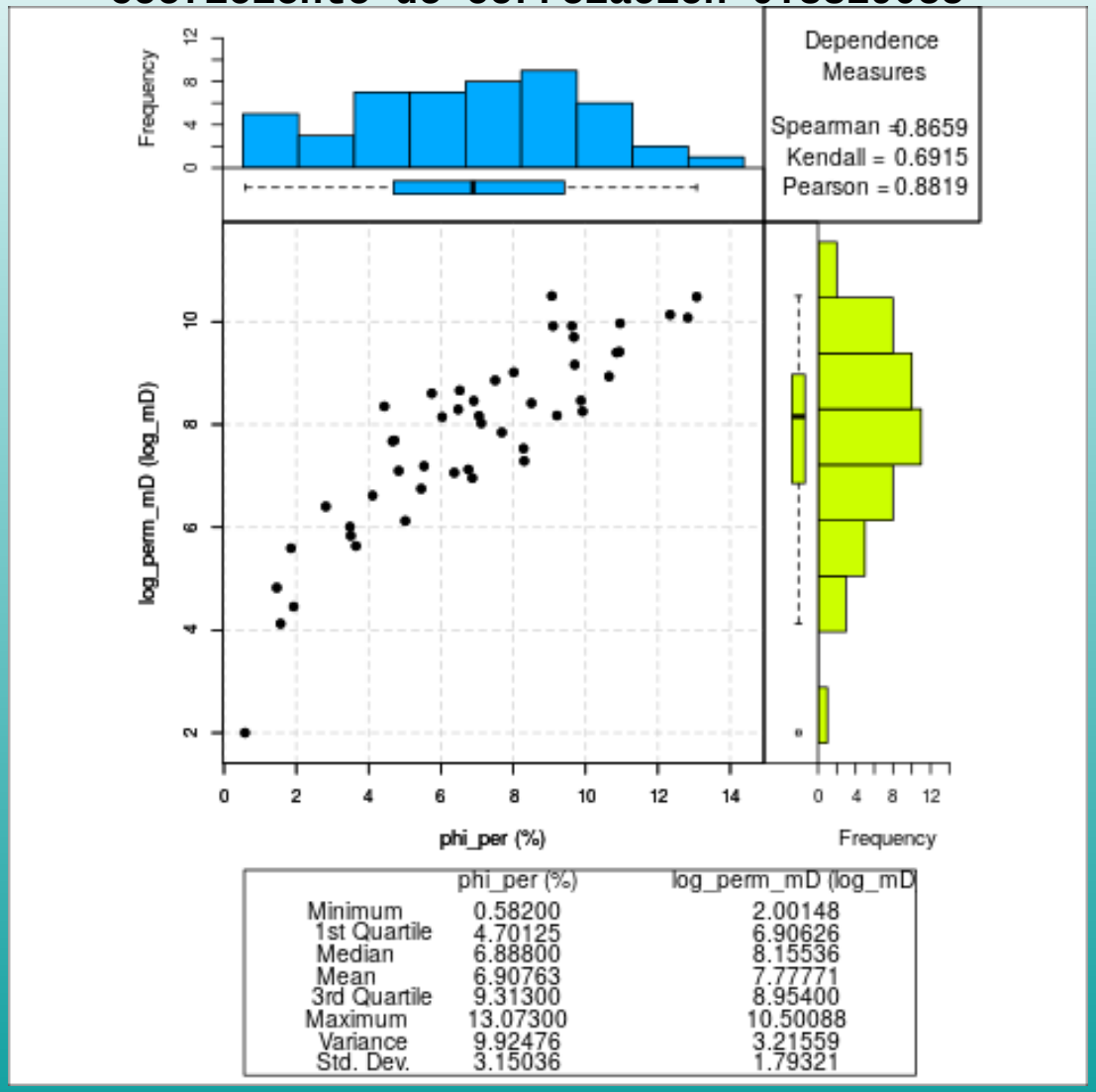


	porosity (%)	permeability (mD)
Minimum	0.58200	7.40000
1st Quartile	4.70125	1002.45000
Median	6.88800	3482.20500
Mean	6.90763	6818.24521
3rd Quartile	9.31300	7743.62500
Maximum	13.07300	36347.40000
Variance	9.92476	83532706.36581
Std. Dev.	3.15036	9139.62288

Estadística bivariada

Después de transformar

Coeficiente de correlación=0.8819055



	phi_per (%)	log_perm_mD (log_mD)
Minimum	0.58200	2.00148
1st Quartile	4.70125	6.90626
Median	6.88800	8.15536
Mean	6.90763	7.77771
3rd Quartile	9.31300	8.95400
Maximum	13.07300	10.50088
Variance	9.92476	3.21559
Std. Dev.	3.15036	1.79321

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

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.

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

- 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***

1. Es igual al coeficiente de correlación

1. Representa la proporción de varianza explicada por la regresión lineal.

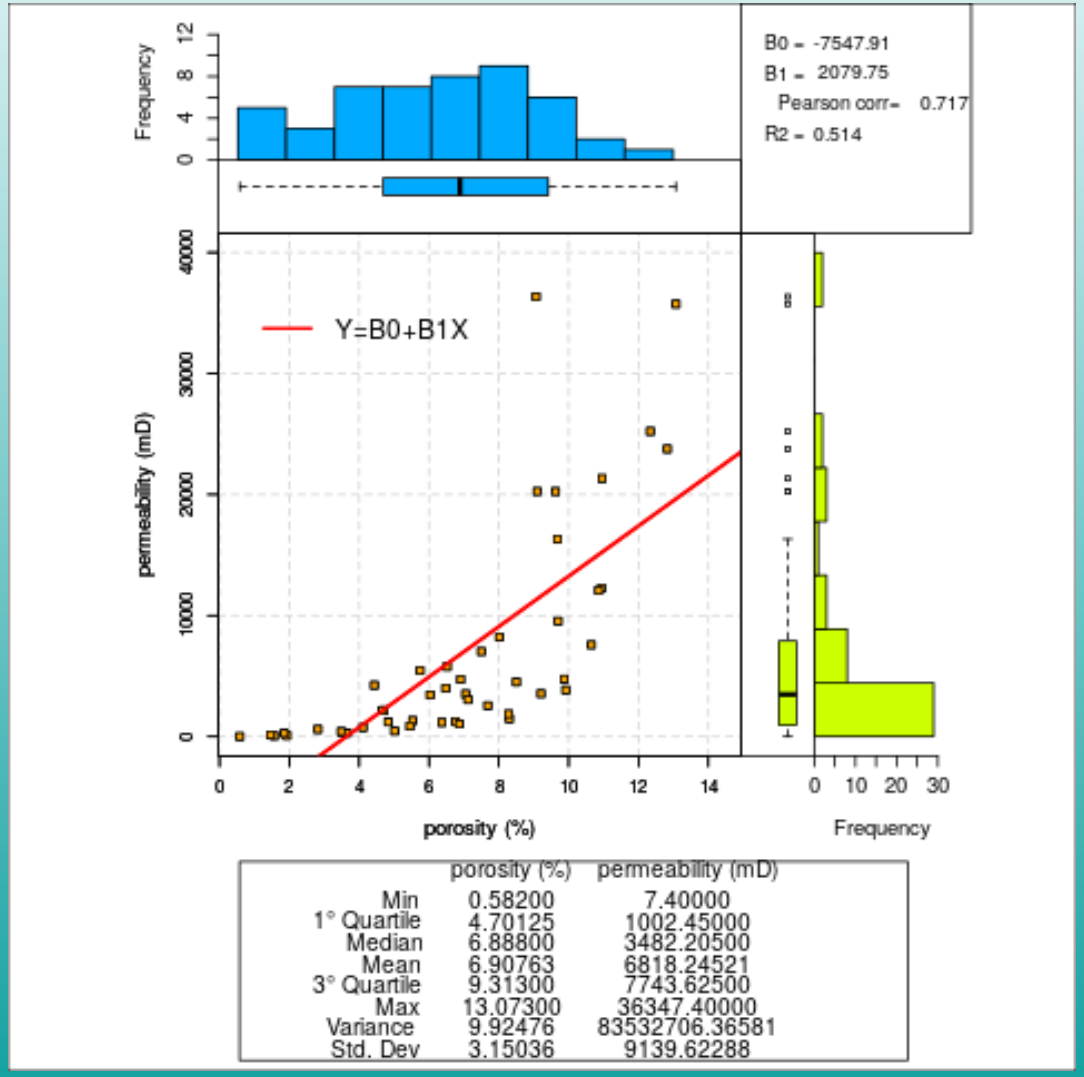
Mínimos Cuadrados Ordinarios (MCO)

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.

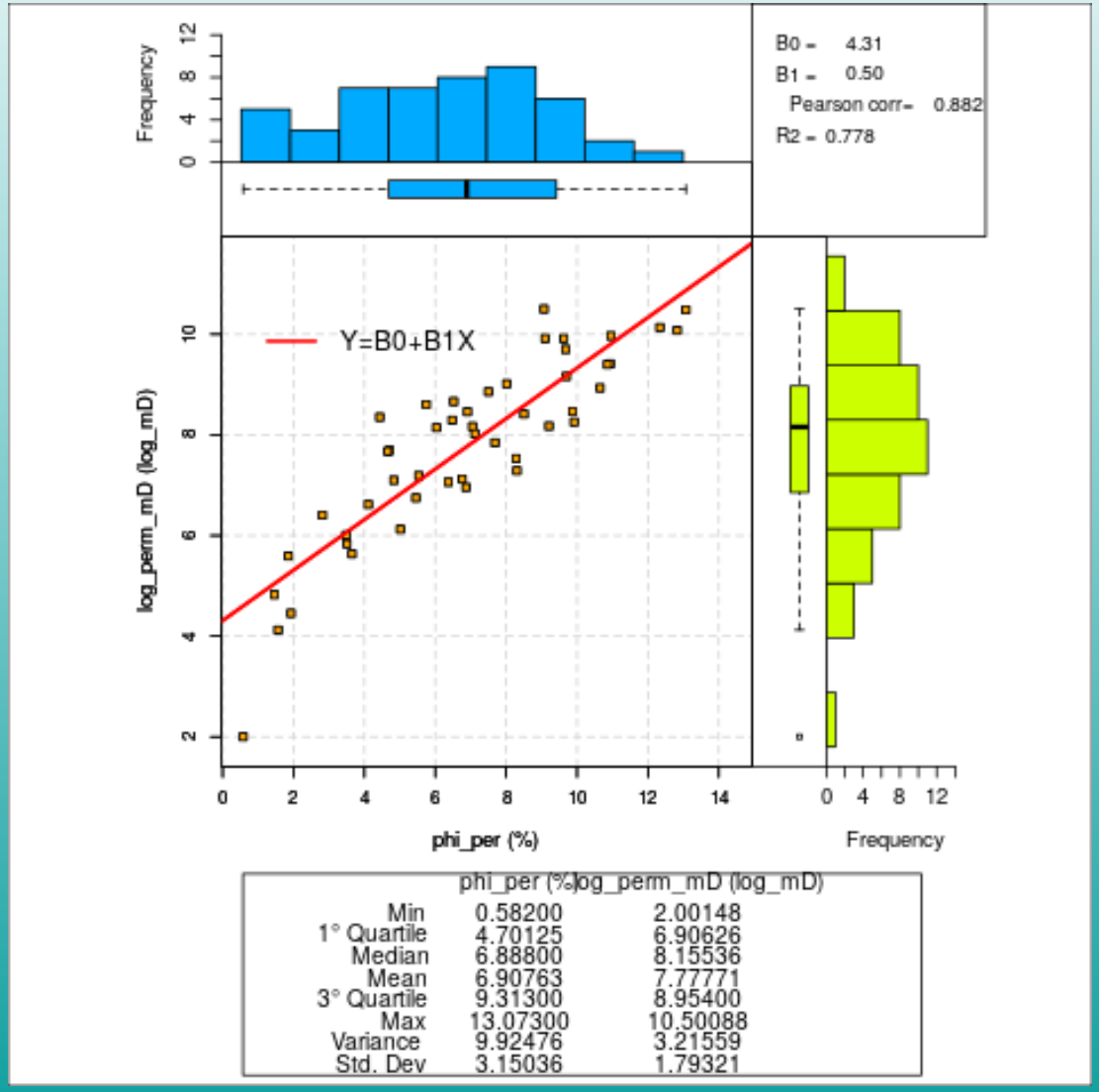
Regresión lineal

Antes de transformar



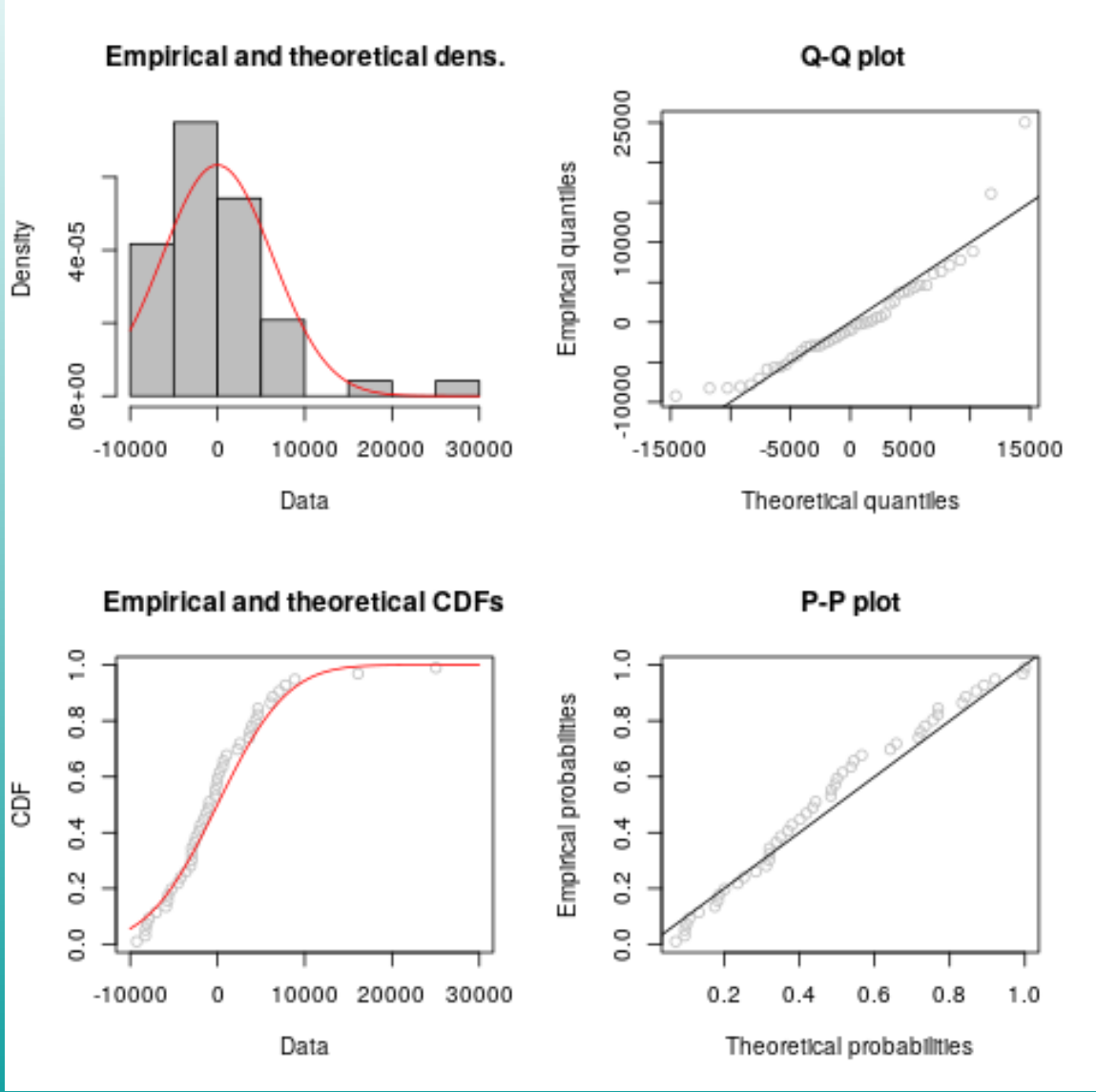
	porosity (%)	permeability (mD)
Min	0.58200	7.40000
1° Quartile	4.70125	1002.45000
Median	6.88800	3482.20500
Mean	6.90763	6818.24521
3° Quartile	9.31300	7743.62500
Max	13.07300	36347.40000
Variance	9.92476	83532706.36581
Std. Dev	3.15036	9139.62288

Regresión lineal
Después de transformar

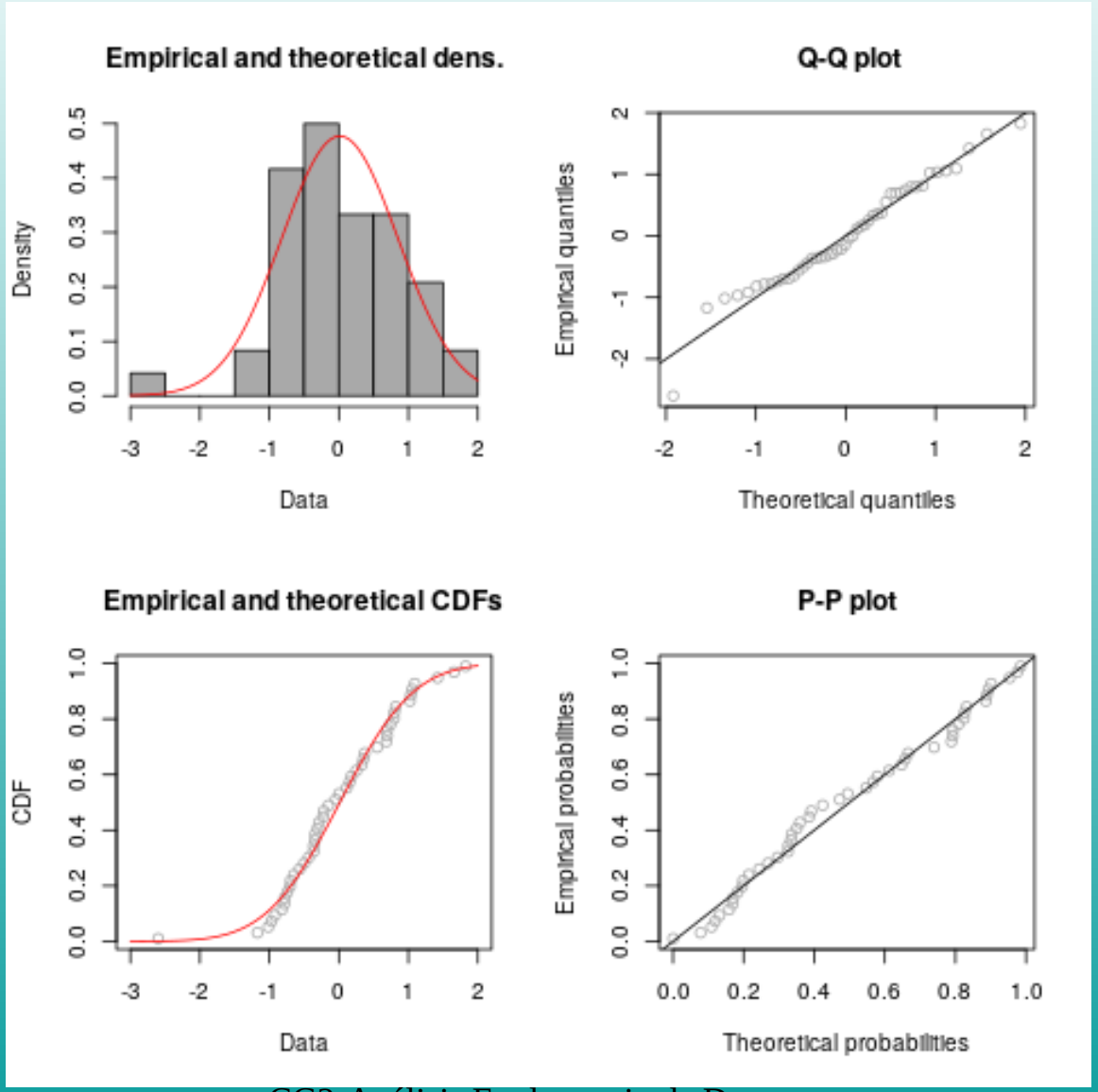


	phi_per (%)	log_perm_mD (log_mD)
Min	0.58200	2.00148
1° Quartile	4.70125	6.90626
Median	6.88800	8.15536
Mean	6.90763	7.77771
3° Quartile	9.31300	8.95400
Max	13.07300	10.50088
Variance	9.92476	3.21559
Std. Dev	3.15036	1.79321

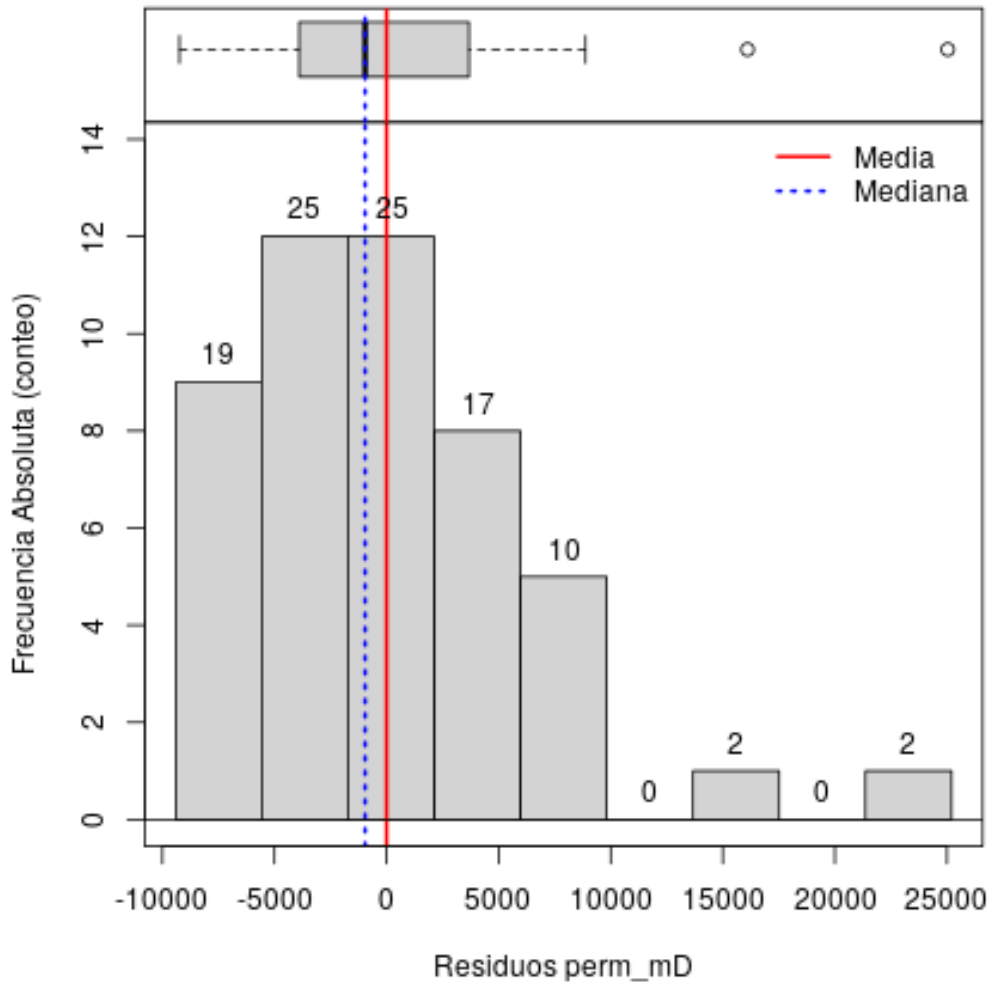
Análisis de los residuos
Antes de transformar



Análisis de los residuos
Después de transformar

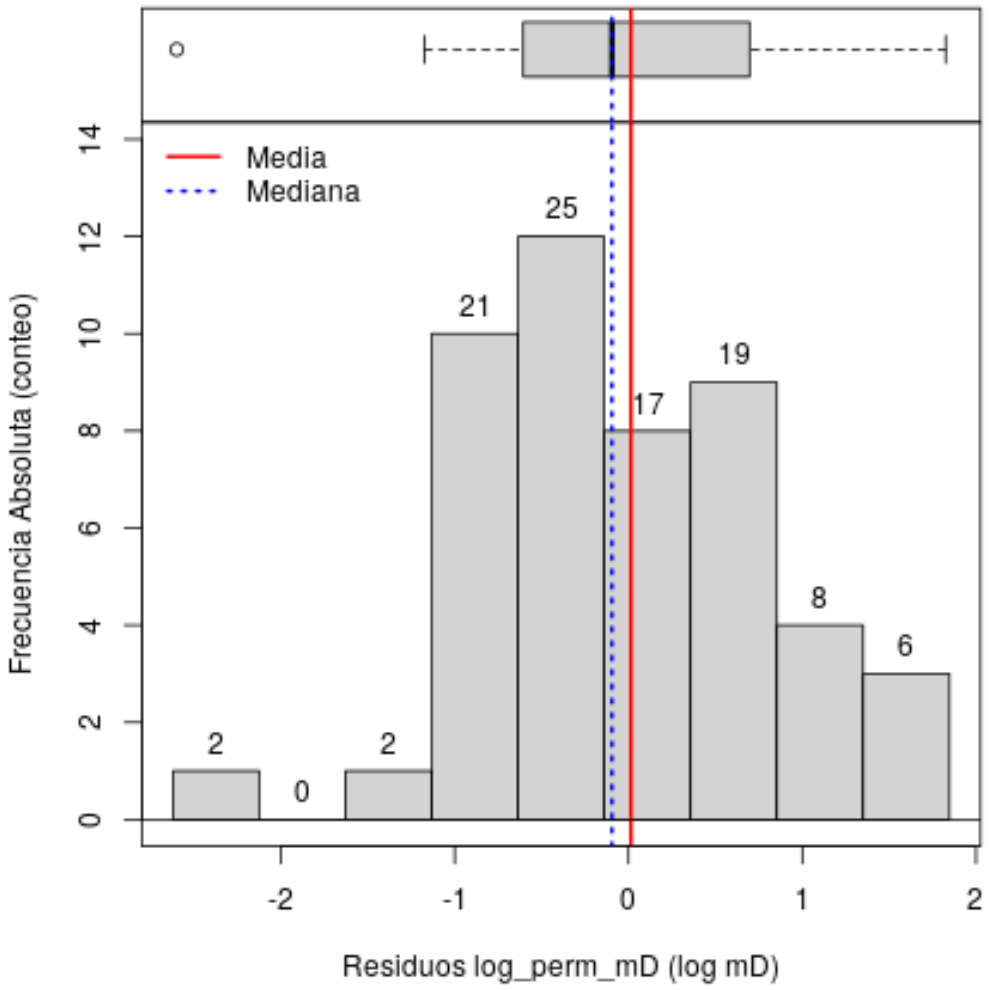


Análisis de los residuos
Antes de transformar



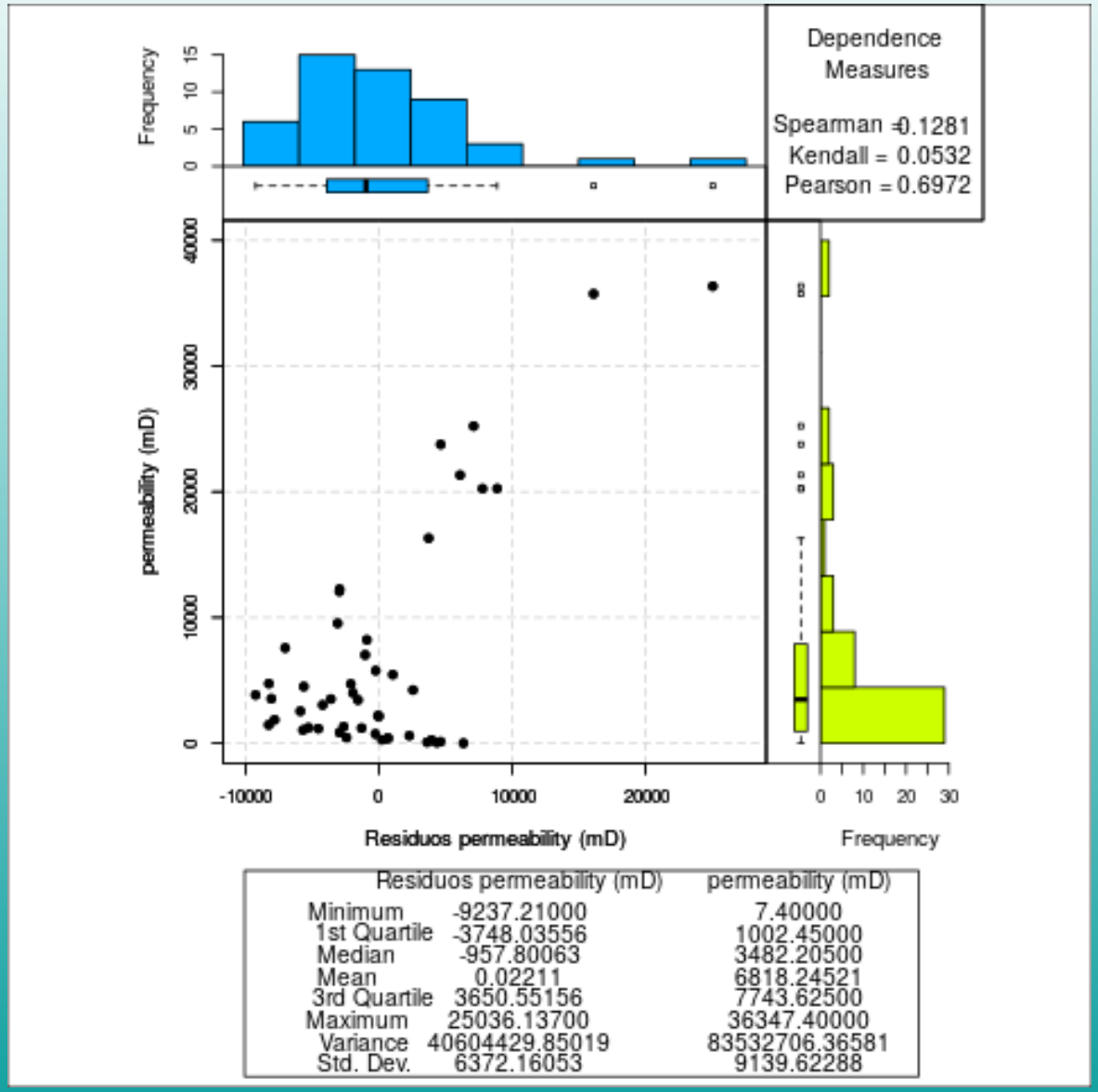
No_muestras	48
Minimo	-9237.21
Cuartil_1er	-3748.0356
Mediana	-957.8006
Media	0.0221
Cuartil_3er	3650.5516
Maximo	25036.137
Rango	34273.347
Rango_Intercuartil	7398.5871
Varianza	40604429.85
Desv_Estandar	6372.1605
Simetria	1.4966
Curtosis	6.7186

Análisis de los residuos
Después de transformar



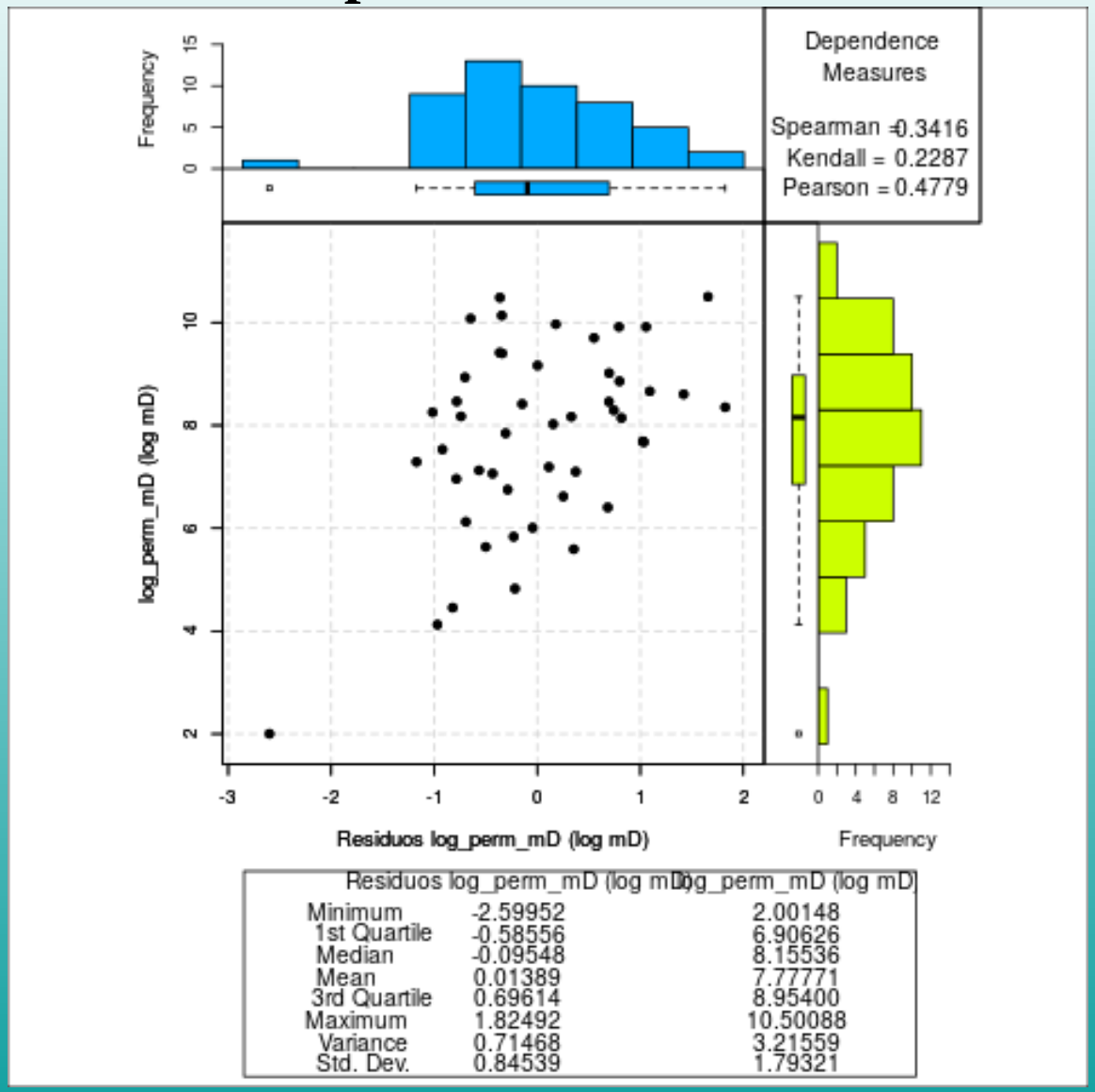
No_muestras	48
Minimo	-2.5995
Cuartil_1er	-0.5856
Mediana	-0.0955
Media	0.0139
Cuartil_3er	0.6961
Maximo	1.8249
Rango	4.4244
Rango_Intercuartil	1.2817
Varianza	0.7147
Desv_Estandar	0.8454
Simetria	-0.1914
Curtosis	3.5273

Análisis de los residuos
Antes de transformar



	Residuos permeability (mD)	permeability (mD)
Minimum	-9237.21000	7.40000
1st Quartile	-3748.03556	1002.45000
Median	-957.80063	3482.20500
Mean	0.02211	6818.24521
3rd Quartile	3650.55156	7743.62500
Maximum	25036.13700	36347.40000
Variance	40604429.85019	83532706.36581
Std. Dev.	6372.16053	9139.62288

Análisis de los residuos
Después de transformar



	Residuos log_perm_mD (log mD)	log_perm_mD (log mD)
Minimum	-2.59952	2.00148
1st Quartile	-0.58556	6.90626
Median	-0.09548	8.15536
Mean	0.01389	7.77771
3rd Quartile	0.69614	8.95400
Maximum	1.82492	10.50088
Variance	0.71468	3.21559
Std. Dev.	0.84539	1.79321