Ciencia de Datos Aplicada a Ciencias de la Tierra

Tema: Conceptos básicos de estadística

Martín A. Díaz-Viera[1], Farhid M. Elisea Guerrero[2]

[1)] *mdiazv@imp.mx*, [2)] *felisea@imp.mx*

25 de enero de 2024

Contenido I

Contenido II

- Análisis de regresión lineal
- Análisis de los residuos

Estadística univariada

Estadística

Univariada

Variable Aleatoria

- **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:

$$a) p_i \geq 0 \qquad \forall i \qquad b) \sum_i p_i = 1 \tag{1}$$

Variable Aleatoria

- **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 %].

Función de Distribución de Probabilidad (FDP)

- La **FDP** caracteriza completamente a la **VA** en términos de probabilidad acumulada
- Se define como:

$$F(z) = Pr\{Z \leq z\} \in [0, 1] \qquad (2)$$

- Su gráfica es el histograma acumulativo

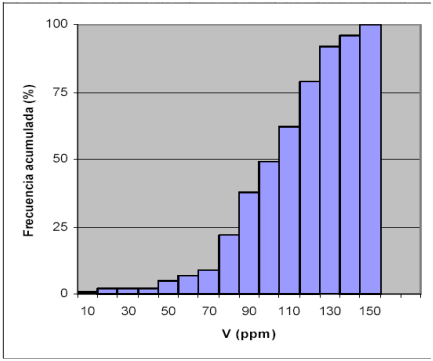


Figura 1: Histograma acumulativo.

Función de Densidad de Probabilidad (fdp)

- La **fdp** caracteriza completamente a la **VA** en términos de densidad de probabilidad
- Se define como:

$$f(z) = \frac{dF(z)}{dz} \qquad (3)$$

- Su gráfica es el histograma

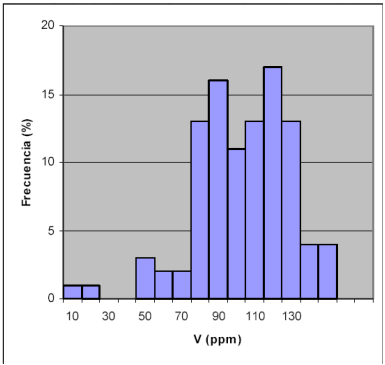


Figura 2: Histograma.

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:

$$F(z_p) = p \tag{4}$$

- Si existe la función inversa se puede expresar como:

$$z_p = F^{-1}(p) \tag{5}$$

Percentiles o cuantiles de una distribución

Algunos cuantiles de interés:

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

Cuartiles y rango intercuartil

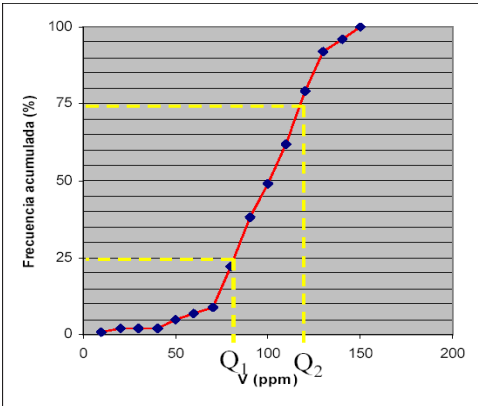


Figura 3: Ejemplo de cuartiles y rango intercuartil.

Valor esperado o esperanza matemática de una VA.

- El valor esperado de una VA, se conoce también como valor medio o media.
- Se define como:

$$m = E[Z] = \int_{-\infty}^{+\infty} z dF(z) \tag{6}$$

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

$$m^* = \frac{1}{N} \sum_{i=1}^{N} z_i \tag{7}$$

- Es muy sensible a los valores atípicos (outliers)

Momentos de una distribución de probabilidad

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

$$m_r = E[Z^r] = \int_{-\infty}^{+\infty} z^r dF(z) = \int_{-\infty}^{+\infty} z^r f(z) dz \tag{8}$$

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

$$\mu_r = E[(Z - m)^r] = \int_{-\infty}^{+\infty} (z - m)^r dF(z) = \int_{-\infty}^{+\infty} (z - m)^r f(z) dz \tag{9}$$

Medidas de tendencia central

- Es un **único valor** con el que se pretende describir un conjunto de datos
- A través de la identificación de la **posición central** del mismo
- A veces se denominan **medidas de localización**
- Forman parte de un **resumen estadístico**.

Medidas de tendencia central

- La **media** es la esperanza matemática, también conocida como media aritmética o promedio.
- La **mediana** es el valor que divide a la distribución en dos partes que representan el 50 %.
- La **moda** es el valor más frecuente de nuestro conjunto de datos, pero puede existir más de una moda.

Medidas de tendencia central

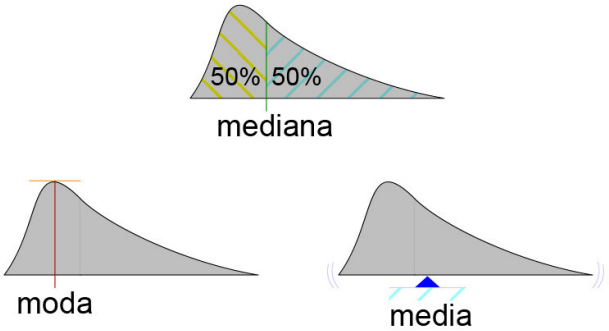


Figura 4: Comparación entre media, mediana y moda.

Medidas de dispersión

- Permiten medir el grado de dispersión de un conjunto de datos numéricos.
- Es necesario considerar un valor central de los datos como punto de referencia.
- Como tal valor central se toma usualmente a la media.
- Cualquier otra medida de localización puede usarse como valor central

Varianza de una VA

- **Varianza de una VA (2do. momento centrado)**
  - Se define como

$$\sigma^2 = Var[Z] = E[(Z - m)^2] \geq 0 \tag{10}$$

  - Caracteriza la dispersión de la distribución respecto al valor medio
  - Se estima

$$(\sigma^2)^* = \frac{1}{N-1} \sum_{i=1}^{N} (z_i - m)^2 \tag{11}$$

Otras medidas de dispersión

- Desviación estándar

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

- Coeficiente de variación (dispersión relativa)

$$CV = \frac{\sigma}{m}$$

- Desviación absoluta media alrededor de la media

$$mAD = E[|Z - m|]$$

- Desviación absoluta media alrededor de la mediana

$$MAD = E[|Z - M|]$$

Medidas de simetría/asimetría

- Signo de simetría $sign(m - M)$
  - cuando >0, es decir m>M se dice que hay asimetría positiva
  - cuando <0, es decir m<M se dice que hay asimetría negativa
- Coeficiente de simetría (medida de la simetría)

$$\alpha_1 = \frac{\mu_3}{\sqrt[3]{\mu_2^2}}$$

- Coeficiente de curtosis (medida del achatamiento)

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

Distribución Normal o Gaussiana

- Esta distribución está completamente caracterizada por sus dos parámetros: media m y varianza σ^2. Se designa mediante

$$N(m, \sigma^2) \tag{12}$$

- La *fdp* normal o Gaussina está dada por

$$g(z) = \frac{1}{\sigma\sqrt{2\pi}} exp\left[-\frac{1}{2}\left(\frac{z-m}{\sigma}\right)^2\right] \tag{13}$$

Distribución Normal o Gaussiana

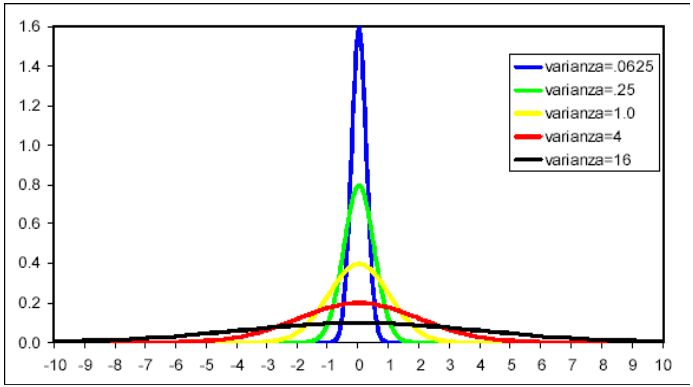


Figura 5: Ejemplos de distribuciones Gaussianas.

Distribución LogNormal

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

$$Y > 0 \rightarrow logN(m, \sigma^2), \qquad si \quad X = ln(Y) \rightarrow N(\alpha, \beta^2) \tag{14}$$

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

Ejemplos de distribuciones Lognormales

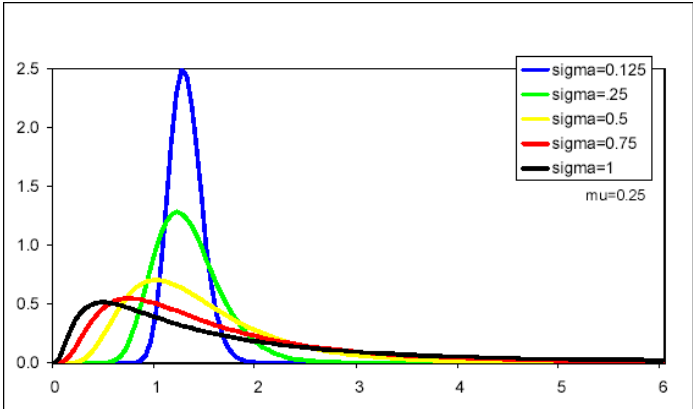


Figura 6: Ejemplos de distribuciones Lognormales.

Simetría y curtosis de una distribución

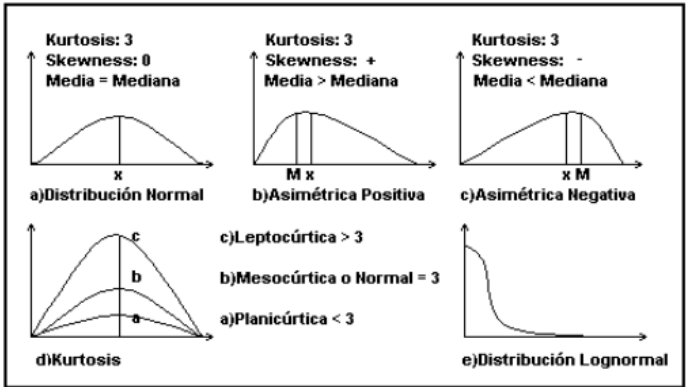


Figura 7: Simetría y curtosis de una distribución.

Gráfica de cajas (boxplot) sin valores atípicos

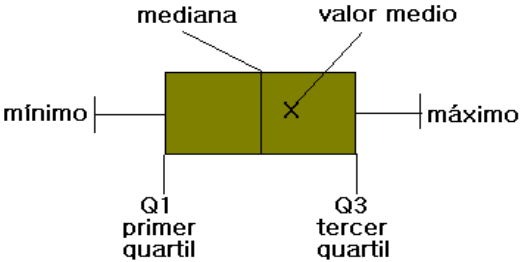


Figura 8: Gráfica de cajas sin valores atípicos.

Gráfica de cajas (boxplot) con valores atípicos

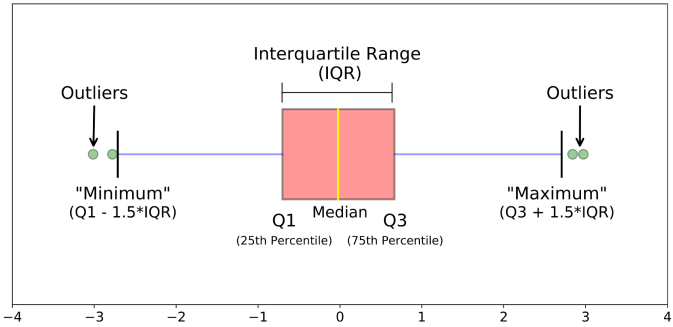


Figura 9: Gráfica de cajas con valores atípicos.

Estadística univariada

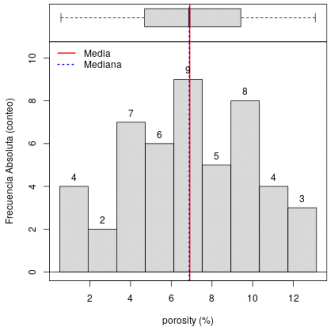


Figura 10: Histograma de la porosidad.

Estadígrafo	Valor
Muestras	48
Mínimo	0.58
1º cuartil	4.70
Mediana	6.88
Media	6.90
3º cuartil	9.31
Máximo	13.07
Rango	12.49
Rango intercuartil	4.61
Varianza	9.92
Desviación estándar	3.15
Simetría	-0.05
Curtosis	2.28

Tabla 1: Estadística básica.

Estadística univariada

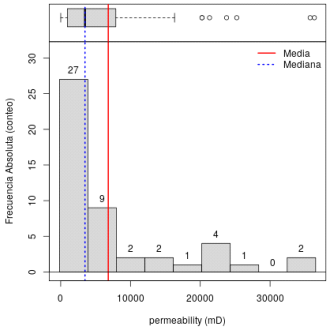


Figura 11: Histograma de la permeabilidad.

Estadígrafo	Valor
Muestras	48
Mínimo	7.40
1º cuartil	1,002.45
Mediana	3,482.20
Media	6,818.24
3º cuartil	7,743.62
Máximo	36,347.4
Rango	36,340
Rango intercuartil	6,741.17
Varianza	83,532,706.36
Desviación estándar	9,139.62
Simetría	1.83
Curtosis	5.62

Tabla 2: Estadística básica.

Estadística univariada

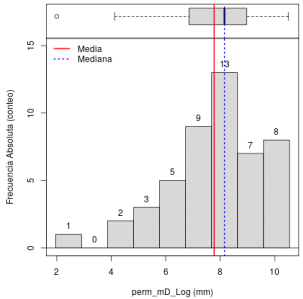


Figura 12: Transformación logarítmica de la Permeabilidad.

Estadígrafo	Valor
Muestras	48
Mínimo	2.00
1º cuartil	6.91
Mediana	8.15
Media	7.77
3º cuartil	8.95
Máximo	10.50
Rango	8.49
Rango intercuartil	2.04
Varianza	3.21
Desviación estándar	1.79
Simetría	-0.84
Curtosis	3.87

Tabla 3: Estadística básica.

Estadística univariada

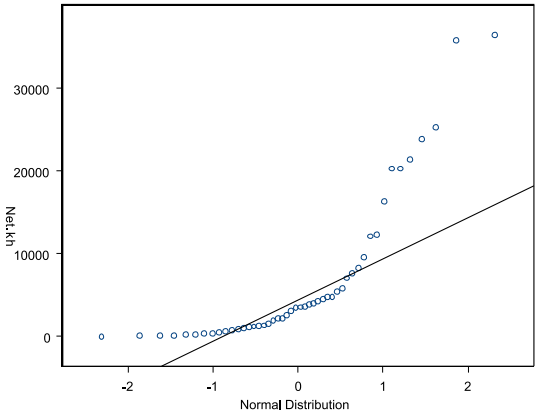


Figura 13: Q-Q plot de la permeabilidad antes de transformar.

Estadística univariada

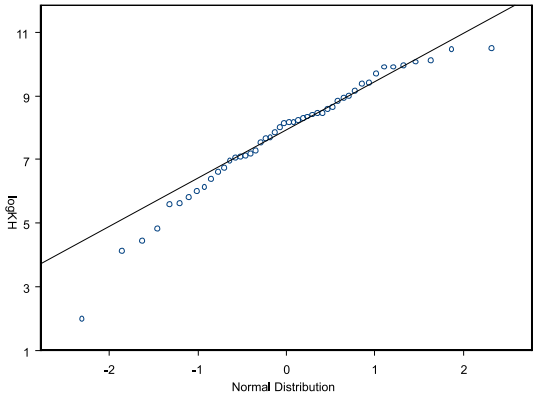


Figura 14: Q-Q plot de la permeabilidad después de transformar.

Estadística univariada

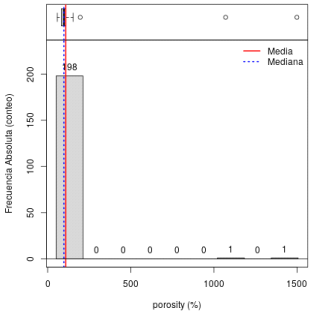


Figura 15: Con valores atípicos (outliers).

Estadígrafo	Valor
Muestras	200
Mínimo	58.2
1º cuartil	82.25
Mediana	97.85
Media	108.9925
3º cuartil	110.325
Máximo	1499
Rango	1440.8
Rango intercuartil	28.075
Varianza	14873.08823
Desviación estándar	121.95527
Simetría	9.92162
Curtosis	104.73871

Tabla 4: Estadística básica.

Estadística univariada

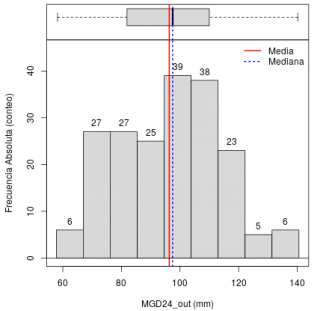


Figura 16: Sin valores atípicos (outliers).

Estadígrafo	Valor
Muestras	196
Mínimo	58.2
1º cuartil	82
Mediana	97.5
Media	96.3265
3º cuartil	110
Máximo	140.2
Rango	82
Rango intercuartil	28
Varianza	319.7503
Desviación estándar	17.8816
Simetría	0.0291
Curtosis	2.3889

Tabla 5: Estadística básica.

Estadística univariada

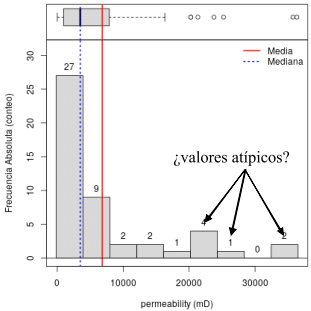


Figura 17: ¿Serán valores atípicos?.

Estadígrafo	Valor
Muestras	48
Mínimo	7.4
1º cuartil	1002.45
Mediana	3482.205
Media	6818.24521
3º cuartil	7743.625
Máximo	36347.4
Rango	36340
Rango intercuartil	6741.175
Varianza	83532706.36
Desviación estándar	9139.62288
Simetría	1.83579
Curtosis	5.62603

Tabla 6: Estadística básica.

Estadística univariada

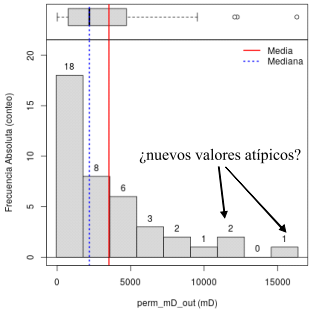


Figura 18: Después de eliminar los valores atípicos.

Estadígrafo	Valor
Muestras	41
Mínimo	7.4
1º cuartil	748
Mediana	2188.7
Media	3521.0285
3º cuartil	4720.5
Máximo	16315.9
Rango	16308.5
Rango intercuartil	3972.5
Varianza	14353741.71
Desviación estándar	3788.6332
Simetría	1.5704
Curtosis	5.1874

Tabla 7: Estadística básica.

Estadística bivariada

Estadística

Bivariada

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.

Función de Distribución de Probabilidad Bivariada

- La distribución de probabilidad conjunta de un par de variables aleatorias **X** y **Y** se define como:

$$F_{XY}(x, y) = Pr\{X \leq x, Y \leq y\} \tag{15}$$

- 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.

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.

Covarianza

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

$$Cov(X, Y) = \sigma_{XY} = E\{(X - m_X)(Y - m_Y)\} \tag{16}$$

- Se estima como

$$\sigma^*_{XY} = \frac{1}{N}\sum_{i=1}^{N}(x_i - m_X)(y_i - m_Y) = \frac{1}{N}\sum_{i=1}^{N} x_i y_i - m_X m_Y \tag{17}$$

Semivarianza

- Se define como

$$\gamma_{XY} = \frac{1}{2}E[(X - Y)^2] \tag{18}$$

- Se interpreta como el momento de inercia del diagrama de dispersión con respecto a una línea con pendiente de 45^o.
- Se estima como

$$\gamma^*_{XY} = \frac{1}{N}\sum_{i=1}^{N}[d_i]^2 = \frac{1}{2N}\sum_{i=1}^{N}[x_i - y_i]^2 \tag{19}$$

- Permite caracterizar la carencia de dependencia.

Semivarianza

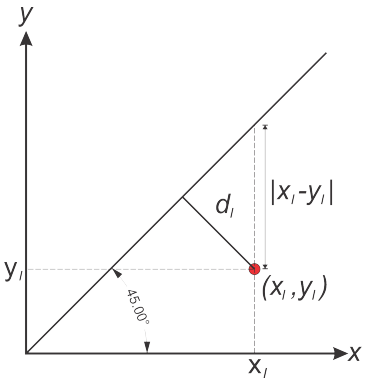


Figura 19: Semivarianza.

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

Observe que por el teorema de Pitágoras tenemos:

$$2[d_i]^2 = [x_i - y_i]^2 \qquad (20)$$

Coeficiente de correlación lineal de Pearson

- Se define como:

$$r_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{Cov\{X, Y\}}{\sqrt{Var\{X\}\, Var\{Y\}}} \in [-1, 1] \tag{21}$$

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

$$r_{XY} = \begin{cases} 1 & \text{para } a > 0 \\ -1 & \text{para } a < 0 \end{cases} \tag{22}$$

Coeficiente de correlación de rango: ρ de Spearman

- Mide el grado de relación monótona entre las variables.
- Se define como el coeficiente de correlación de Pearson entre las variables de rango como sigue:

$$\rho_s = r_{\mathrm{R}(X),\mathrm{R}(Y)} = \frac{Cov(R(X), R(Y))}{\sigma_{\mathrm{R}(X)}\sigma_{\mathrm{R}(Y)}}, \tag{23}$$

donde

- $\mathrm{R}(X_i)$, $\mathrm{R}(Y_i)$ son los rangos de las observaciones X_i, Y_i,
- $r_{\mathrm{R}(X),\mathrm{R}(Y)}$ denota el coeficiente de correlación de Pearson habitual, pero aplicado a las variables de rango,
- $Cov(\mathrm{R}(X), \mathrm{R}(Y))$ es la covarianza de las variables de rango y
- $\sigma_{\mathrm{R}(X)}, \sigma_{\mathrm{R}(Y)}$ son las desviaciones estándar de las variables de rango.

Coeficiente de correlación de rango: ρ de Spearman

- Se puede calcular mediante la siguiente expresión:

$$\rho_s^* = 1 - \frac{6 \sum_i D_i^2}{N(N^2 - 1)} \tag{24}$$

donde $D_i = \mathrm{R}(X_i) - \mathrm{R}(Y_i)$ es la diferencia entre los dos rangos de cada observación y N es el número de observaciones.

- Para calcular ρ, las parejas de datos X y Y se ordenan y son reemplazados por su respectivo orden donde D es la diferencia $X \check{} Y$ entre los estadísticos de orden 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.
- Menos sensible a los valores atípicos que Pearson.

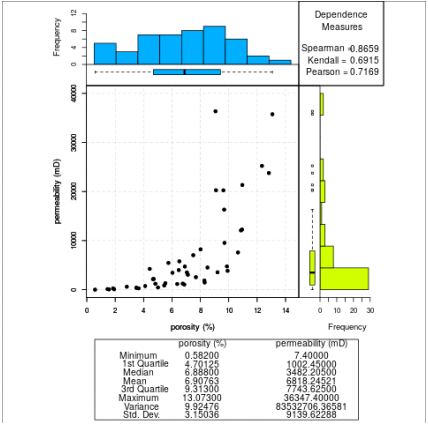
Coeficiente de correlación de rango: τ de Kendall

- Se define como:

$$\tau = \frac{\text{(número de pares concordantes)-(número de pares discordantes)}}{\binom{n}{2}} \tag{25}$$

- 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.
- Menos sensible a los valores atípicos que Pearson.

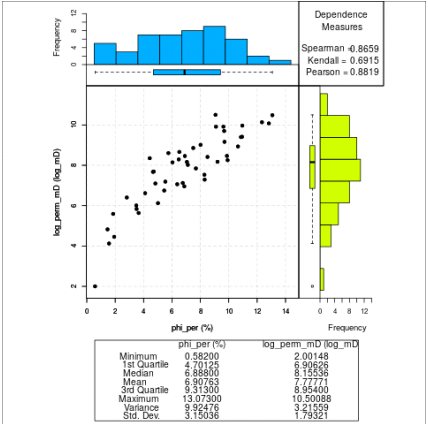
Antes de transformar



	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

Figura 20: Coeficiente de correlación lineal = 0.71.

Después de transformar



	phi_per (%)	log_perm_mD (log_mD)
Minimum	0.58200	2.00148
1st Quartile	4.70125	6.90626
Median	6.88800	8.15336
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

Figura 21: Coeficiente de correlación lineal = 0.88.

Regresión lineal

Regresión

Lineal

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 I

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:

$$Y = \beta_0 + \beta_1 X \tag{26}$$

Donde

$$y_i = \beta_0 + \beta_1 x_i + e_i, \qquad i = 1, \ldots, N \tag{27}$$

- β_0, β_1 son los parámetros del modelo
- e_i son los errores o residuos del modelo

Mínimos Cuadrados Ordinarios (MCO)

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

$$SCR = \sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} [y_i - \hat{y}_i]^2 = \sum_{i=1}^{N} \left[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \right]^2 \tag{28}$$

- Sistema de ecuaciones a resolver

$$\frac{\partial SCR}{\partial \beta_0} = 0, \frac{\partial SCR}{\partial \beta_1} = 0 \tag{29}$$

Coeficiente de determinación R^2

- El coeficiente de determinación se define como:

$$R^2 = 1 - \frac{SCR}{SCT} \tag{30}$$

donde $SCT = \sum_{i=1}^{N} [y_i - m_y^*]^2$ es la suma de cuadrados total (proporcional a la varianza de los datos)

- 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.

Criterios de la bondad del ajuste

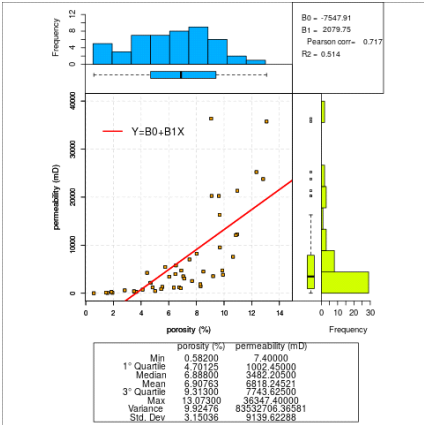
- Si $R^2 \approx 1$, el ajuste es bueno (Y se puede calcular de modo bastante aproximado a partir de X y viceversa).
- Si $R^2 \approx 0$, las variables X y Y no están relacionadas (linealmente al menos), por tanto no tiene sentido hacer un ajuste lineal.
- Sin embargo no es seguro que las dos variables 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

- Condiciones que deben cumplir los residuos

1. Valor esperado cero: $E\{e_i\} = 0$
2. Varianza constante: $Var\{e_i\} = \sigma_e^2$
3. No correlacionados: $Cov\{e_i, e_j\} = 0, \qquad \forall i \neq j$
4. Distribución normal: $e \sim N(0, \sigma_e^2)$

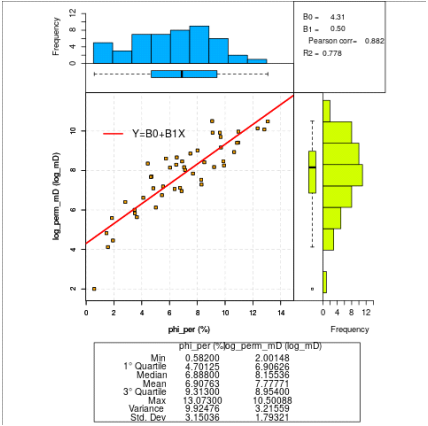
Antes de transformar



	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	7745.62500
Max	13.07300	36347.40000
Variance	9.92476	83532706.36581
Std. Dev	3.15036	9139.62288

Figura 22: Permeabilidad vs. porosidad antes de transformar.

Después de transformar



	phi_per (%)	log_perm_mD (log_mD)
Min	0.58200	2.00148
1° Quartile	4.70125	6.90826
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

Figura 23: Permeabilidad vs. porosidad después de transformar.

Análisis de los residuos

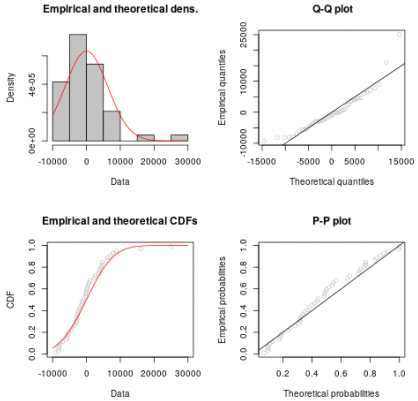


Figura 24: Residuos antes de transformar.

Análisis de los residuos

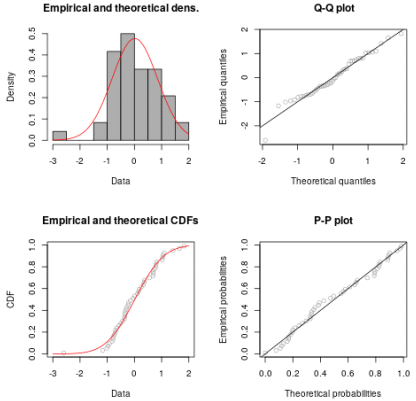


Figura 25: Residuos después de transformar.

Análisis de los residuos

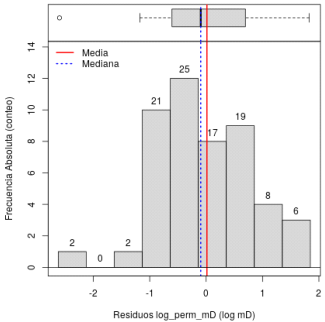
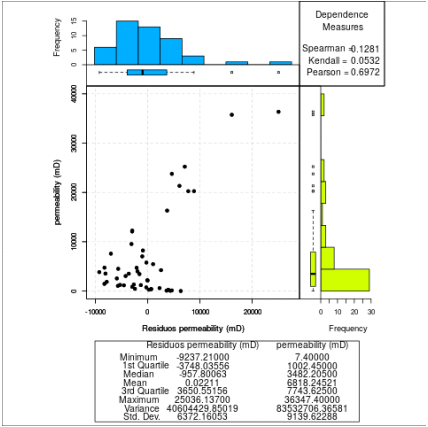


Figura 26: Histograma de los residuos después de transformar.

Estadígrafo	Valor
Muestras	48
Mínimo	-2.5995
1º cuartil	-0.5856
Mediana	-0.0955
Media	0.0139
3º cuartil	0.6961
Máximo	1.8249
Rango	4.4244
Rango intercuartil	1.2817
Varianza	0.7147
Desviación estándar	0.8454
Simetría	-0.1914
Curtosis	3.5273

Tabla 8: Estadística básica.

Análisis de los residuos



	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

Figura 27: Correlación de la permeabilidad vs. los residuos antes de transformar.

Análisis de los residuos

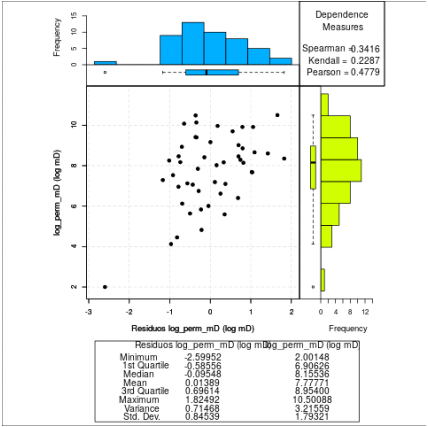


Figura 28: Correlación de la permeabilidad vs. los residuos después de transformar.

Siguiente tema:

Análisis Exploratorio de Datos