LECTURAS MFIA

LIBRO 1

# Estadística

Antonio Garre, MFIA
Enrique Castellanos, FRM, MFIA

2

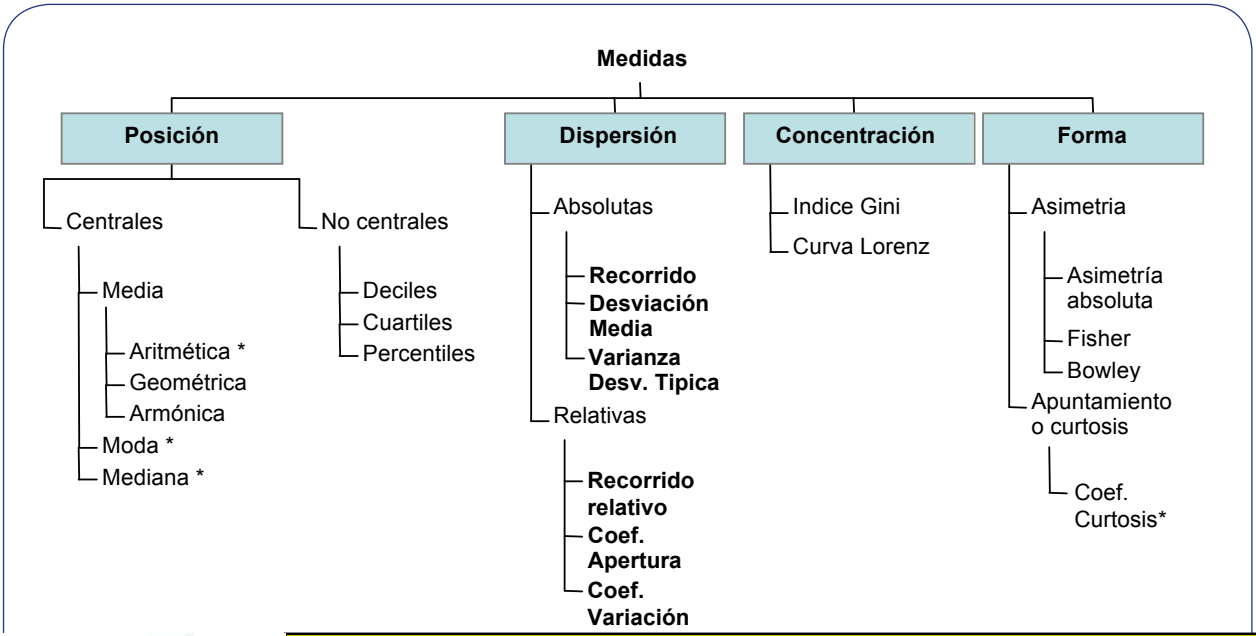
## 1. INTRODUCCIÓN

La **Estadística descriptiva** tiene como objetivo: recolección, descripción, visualización y resumen de los datos obtenidos, de una población o una muestra, originados a partir de los fenómenos en estudio. Ejemplos básicos de descriptores: media, desviación estándar, covarianza.

El objetivo de la Estadística Descriptiva es reducir una serie de datos a unos pocos coeficientes que contengan la mayor parte de la información relevante, con el fin de descubrir regularidades estadísticas en el colectivo analizado.

Por otro lado la **Estadística inferencial**, tiene como objetivo la generación de modelos, inferencias y predicciones asociadas a los fenómenos en cuestión teniendo en cuenta lo aleatorio e incertidumbre en las observaciones. Con el fin de poder estimar valores futuros y poder establecer, con ellos, conclusiones generales y estimar características de las futuras observaciones a partir de los datos de la muestra, o describir el grado de asociación entre variables: regresión.

**CUADRO 1: PRINCIPALES MEDIDAS DE LA ESTADÍSTICA DESCRIPTIVA.**
FUENTE: ELABORACIÓN PROPIA




Pueden ser de tendencia central o no central.

**Medidas de dispersión:** Miden la dispersión de los valores de la muestra respecto de la tendencia central, y por tanto, su representatividad como síntesis de toda la información.

**Medidas de forma:** Tratan de expresar de forma numérica la forma que tiene la distribución de frecuencias.

**Medidas de concentración:** Miden la concentración de la distribución de frecuencias.

## 2. MEDIDAS DE POSICIÓN CENTRALES Y NO CENTRALES: MEDIA, MEDIANA, MODA

### MEDIDAS DE POSICIÓN CENTRALES

Buscan lo valores en torno a los que se agrupan los datos de una distribución de frecuencias. Tratan de identificar el punto alrededor del cual se centran los datos, que se tomará como representativo de todo el conjunto.

**Media aritmética**

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_i + \ldots x_n}{N} = \frac{\sum_{i=1}^{n} x_i}{N}$$

Propiedades de la media aritmética:

- La suma algebraica de las desviaciones de un conjunto de números respecto de su media aritmética es cero.
- La suma de los cuadrados de las desviaciones de un conjunto de números, xi, respecto de un cierto número a es mínima si y solo si a = $\bar{x}$
- Esta propiedad nos servirá cuando veamos el tema de la regresión

**Media geométrica**

El empleo más frecuente de la media geométrica es el de promediar porcentajes, tasas, números índices, etc., es decir, en los casos en los que se supone que la variable presenta variaciones acumulativas. Es menos sensible a los valores extremos que la media aritmética, pero su cálculo es más complicado y difícil de interpretar. A veces puede ni existir, cuando trabajamos con datos negativos.

$$G = \sqrt[N]{x_1^{n_1}.x_2^{n_2}....x_n^{n_n}} \qquad G = \sqrt[N]{(1+x_1)^{n_1}.(1+x_2)^{n_2}...(1+x_n)^{n_n}} - 1$$

**Ejemplo**

Tras depositar en un plan de pensiones una cantidad de 1000 euros. El tipo de interés con el que me remuneran esta inversión es del 1% este año, 3% el segundo año y 4% el tercer año. ¿Cuál ha sido el tipo de interés medio?

$$G = \sqrt[N]{(1+i_1)^{n_1} \bullet (1+i_2)^{n_1} \bullet ..... \bullet (1+i_N)^{n_1}} - 1$$

**Media armónica**

La media armónica de N observaciones es la inversa de la media de las inversas de las observaciones y la denotaremos por H. Al igual que en el caso de la media geométrica su utilización es bastante poco frecuente.

$$H = \frac{N}{\sum_{i=1}^{n} \frac{1}{x_i} \bullet n_i}$$

Relación entre la media aritmética, geométrica y armónica

**H<= G <= Xmed**

**Ejemplo**

Tenemos una muestra de datos: 3, 6 ,2, 11, 14.

Ordeno de menor a mayor: 2,3,6,11,14.

El 6 es el valor mediano, puesto que deja dos valores a su izquierda y dos a su derecha.

**Moda**

Es el valor que ocurre con mayor frecuencia, el que más se repite.

**Ejemplo**

La moda es 6 pues su frecuencia es la más alta.

| X | n |
|---|---|
| 1 | 3 |
| 2 | 4 |
| 6 | 10 |
| 7 | 7 |

## MEDIDAS DE POSICIÓN NO CENTRALES (*CUANTILES*)

Las medidas de posición no centrales permiten conocer otros puntos característicos de la distribución que no son los valores centrales, y se denominan cuantiles. Los cuantiles son los valores de la distribución que la dividen en n partes iguales, o intervalos que contienen el mismo número de valores, y pueden denominarse:

**Cuartiles:** son 3 valores que dividen la serie de datos ordenada en 4 partes iguales, en las que cada una contiene el 25% de los datos.

## 3. MEDIDAS DE DISPERSIÓN, VOLATILIDAD, ASIMETRÍA Y CURTOSIS

Las medidas de dispersión en estadística nos informan sobre cuánto se alejan del centro los valores de la distribución. Algunas de las medidas de dispersión más destacadas son:

**Recorrido**

Es la diferencia entre el mayor y el menor valor de la distribución. No es una medida muy significativa en la mayoría de los casos, pero es muy fácil de calcular. Es útil para distribuciones muy uniformes.

$$Re = \text{Valor max} - \text{Valor min}$$

**Recorrido Intercuartilico**

Es la diferencia entre el tercer cuartil y el primero. Nos indica que en un intervalo de longitud RI están comprendidos el 50% de los valores centrales. Si RI es pequeño podemos intuir una pequeña dispersión.

$$RI = C3 - C1$$

**Desviación**

Es la diferencia que se observa entre el valor de la variable y una medida de posición central (media o mediana , generalmente). Una primera solución, es calcular la Media de las desviaciones. Pero esto no es correcto, puesto que la suma de las desviaciones siempre va a ser 0. Las desviaciones positivas se compensan con las negativas.

Entonces, lo que se nos ocurre es calcular la media de las desviaciones, pero en valor absoluto, para que no se compensen las desviaciones positivas con las negativas. Es lo que se denomina Desviación Absoluta media.

respecto a la mediana

$$D_{Me} = \sum_{i=1}^{n} |x_i - Me| \cdot \frac{n_i}{N}$$

## VARIANZA

La varianza es una medida de dispersión en la que las desviaciones con respecto a la media se elevan al cuadrado, para calcular posteriormente su media. De esta manera se evita que desviaciones positivas y negativas se compensen, pero también se está dando más importancia a las desviaciones grandes (tanto positivas como negativas) frente a las pequeñas. Esta medida de dispersión es la más utilizada.

$$S^2 = \sum_{i=1}^{n} (X_i - \bar{X})^2 \cdot \frac{n_i}{N}$$

### Propiedades de la Varianza:

1. Si le sumamos a todas las observaciones un mismo número, la varianza no se ve afectada. (No le afectan cambios de origen).
2. Si multiplicamos o dividimos todas las observaciones por un mismo número, la varianza queda multiplicada o dividida por dicho numero al cuadrado. (Le afectan cambios de escala).
3. Derivado de las anteriores propiedades, si podemos distinguir varias submuestras dentro de la muestra, la varianza global no es la suma de las varianzas de las submuestras. (No linealidad de la varianza).
4. Siempre es un valor positivo.

## DESVIACIÓN TÍPICA O VOLATILIDAD

El problema que presenta esta medida es que si la variable tiene unidades (por ej. Euros, metros,...), la varianza al elevar al cuadrado, vendría expresada en unidades al cuadrado. Y esto,

### Coeficiente de Variación de Pearson

Es el más utilizado para comparar dos distribuciones con distintas unidades de medida o con distintas medias. Es el cociente entre la desviación típica y la media aritmética.

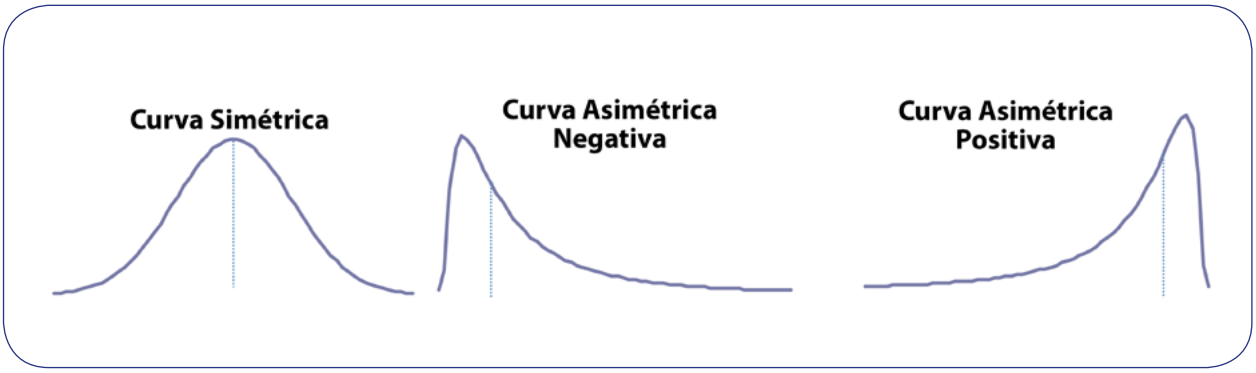
$$CV = \frac{S}{\bar{X}} \cdot 100$$

Representa el número de veces que S contiene a la media aritmética. Cuanto mayor sea, más veces contendrá S a la media, luego menos representatividad tendrá esta. Si CV=0, la media alcanza su máxima representatividad (no existe dispersión). Si la media es 0, el CV tiende a infinito. Entonces no es válida esta medida.

## ASIMETRÍA

El concepto de asimetría se refiere a si la curva que forman los valores de la serie con respecto a un valor central (media aritmética). Ver gráfico 1.

**GRÁFICO 1: DIFERENTES SIMETRÍAS EN CURVAS**

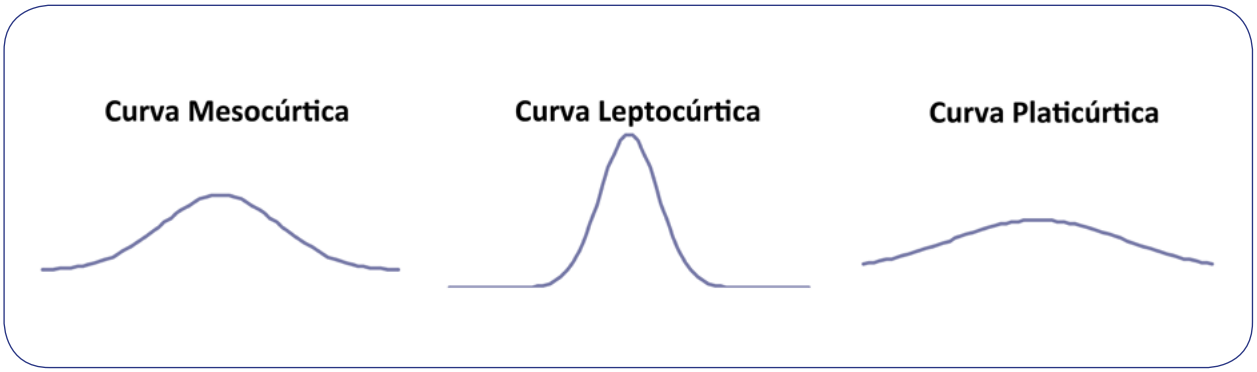



### Coeficiente de Asimetría de Fisher

### CURTOSIS

El Coeficiente de Curtosis analiza el grado de concentración que presentan los valores alrededor de la zona central de la distribución. Ver gráfico 2.

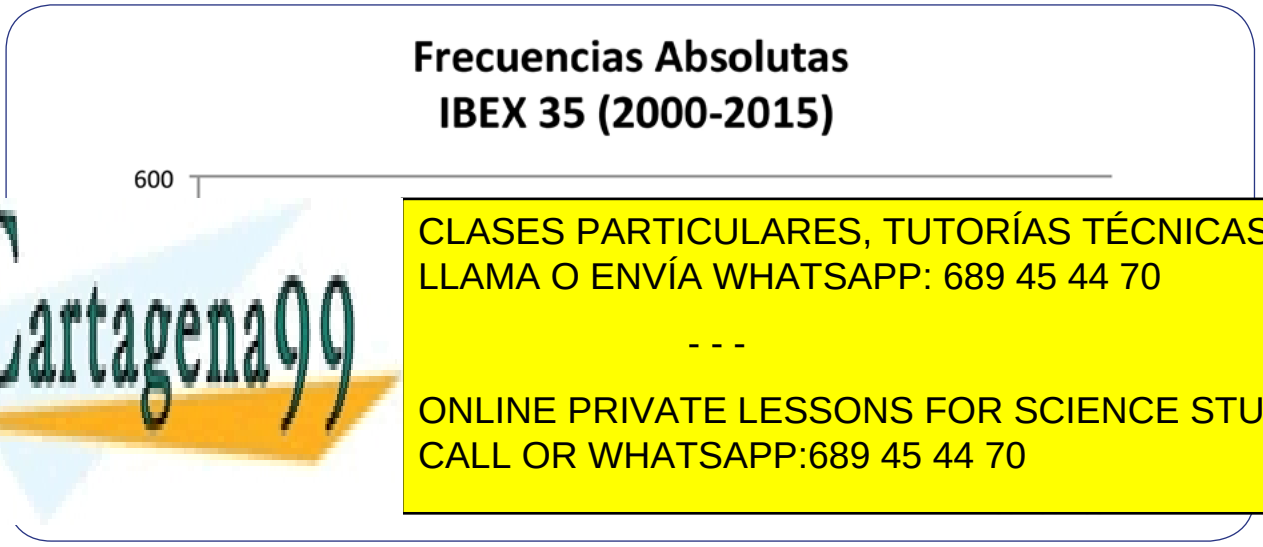
**GRÁFICO 2: CURVAS CON DIFERENTES CURTOSIS**




## 4. HISTOGRAMA DE FRECUENCIAS

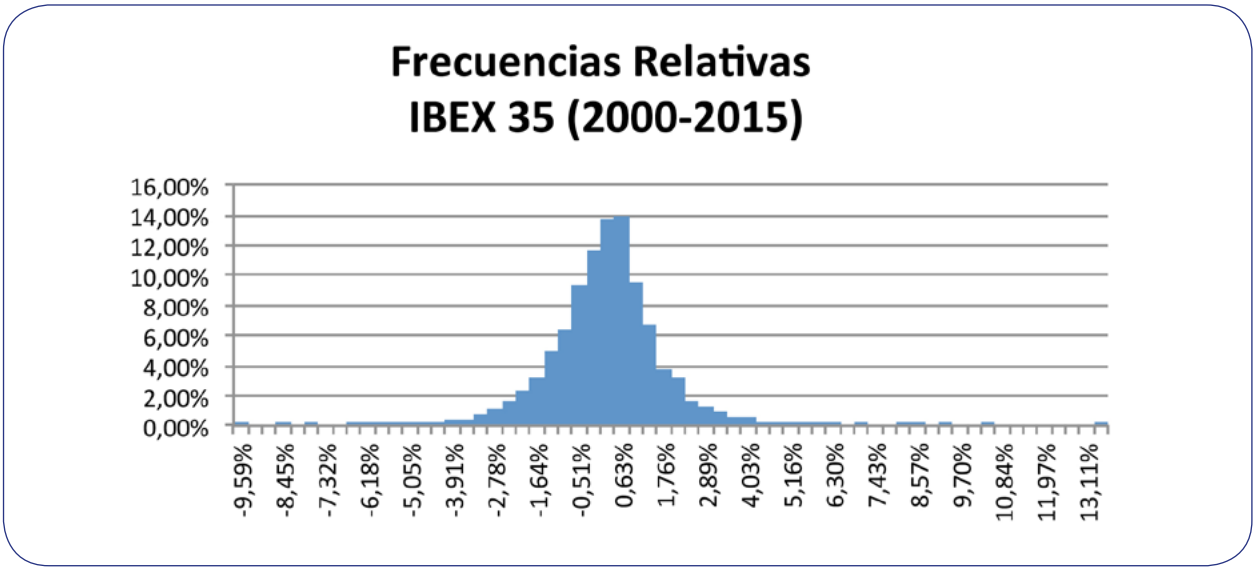
Cuando se trata de analizar la dispersión que presentan unos datos, la representación gráfica más adecuada es el histograma. Para realizar un histograma se marcan una serie de intervalos sobre un eje horizontal, y sobre cada intervalo se coloca un rectángulo de altura proporcional al número de observaciones (frecuencia absoluta) que caen dentro de dicho intervalo. De esta manera el histograma de frecuencias resulta muy útil para representar gráficamente la distribución de frecuencias.

**GRAFICO 3: HISTOGRAMA DE FRECUENCIAS ABSOLUTAS DE IBEX 35® ENTRE EL AÑO 2000 Y NOVIEMBRE DE 2015.** FUENTE: ELABORACIÓN PROPIA



Si se pretende comparar varios histogramas construidos con distinto número de datos, es preferible que las alturas de los rectángulos sean proporcionales al porcentaje de observaciones en cada intervalo, o al tanto uno por uno (frecuencia relativa). Utilizando la frecuencia relativa en el eje de ordenadas también se facilita la comparación entre el histograma obtenido y un determinado modelo teórico representado por una función densidad de probabilidad. En este caso se considera que la frecuencia relativa es proporcional al área definida por cada columna. Puede interpretarse la función densidad de probabilidad como la representación del histograma cuando el número de observaciones tiende a infinito y la anchura de los rectángulos tiende a cero.

**GRÁFICO 4: HISTOGRAMA DE FRECUENCIAS RELATIVAS DE IBEX 35® ENTRE EL AÑO 2000 Y NOVIEMBRE DE 2015.** FUENTE: ELABORACIÓN PROPIA




El rango nos informa de la extensión de la variabilidad de los datos:

$$Rango = Max(X_i) - Min(X_{i)}$$

El número de clases indica la cantidad de rectángulos a realizar:

$$Nº\,Clases = \sqrt{N}\ (\text{se redondea})$$

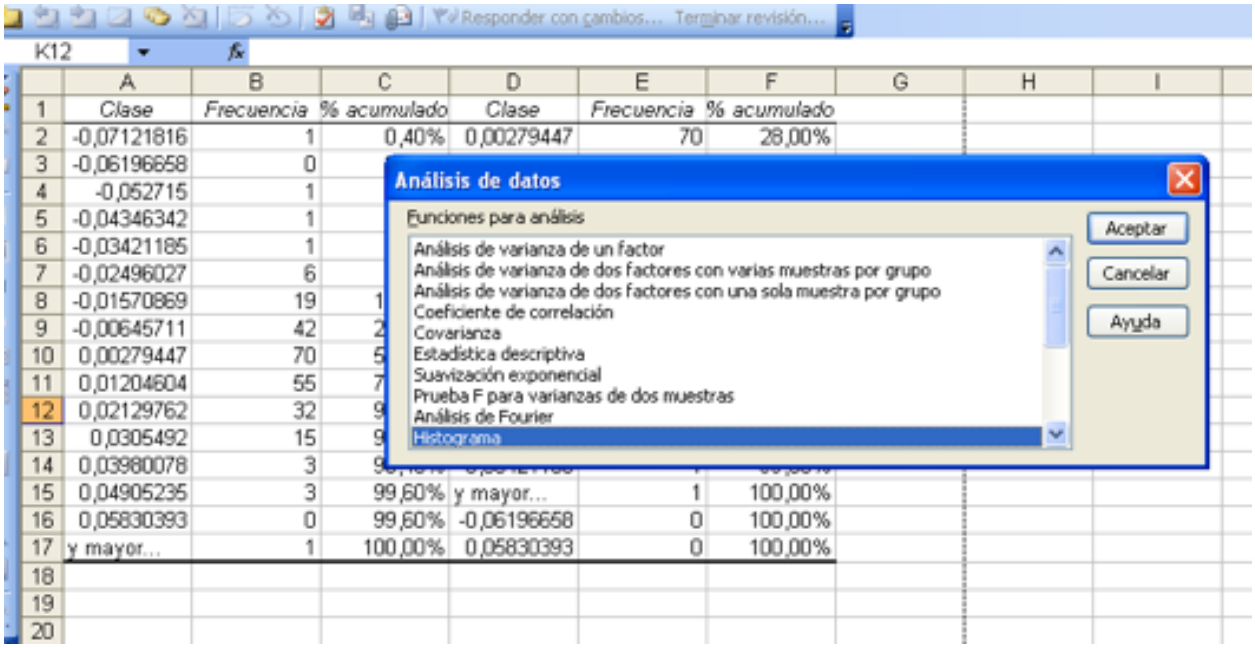
Si el número de clases por la longitud supera el rango se ajusta dividiendo la diferencia entre dos y ampliando la primera y última clase.

En Excel podemos utilizar la función "frecuencia" o por medio del "análisis de datos"

**CUADRO 2: EJEMPLO EN EXCEL DE CÓMO UTILIZAR LA FUNCIÓN FRECUENCIA.**
FUENTE: ELABORACIÓN PROPIA

| TEF | 251 | | | |
|---|---|---|---|---|
| 16,35 | | | | |
| 16,08 | -1,67% | MAX | 6,76% | |
| 16,20 | 0,74% | MIN | -7,12% | |
| 16,26 | 0,37% | Rango | 13,88% | |
| 16,00 | -1,61% | Clases | 16 | |
| 15,97 | -0,19% | Longitud | 0,87% | |
| 16,03 | 0,38% | | | |
| 16,05 | 0,12% | | | |
| 16,25 | 1,24% | | | |
| 16,26 | 0,06% | | | |
| 16,10 | -0,99% | | -7,12% | =+FRECUENCIA(C3:C253;E12:E28) |
| 16,16 | 0,37% | | -6,25% | 0,00 |
| 16,05 | -0,68% | | -5,39% | 1,00 |
| 16,22 | 1,05% | | -4,52% | 1,00 |
| 16,10 | -0,74% | | -3,65% | 1,00 |
| 16,27 | 1,05% | | -2,79% | 1,00 |
| 16,25 | -0,12% | | -1,92% | 17,00 |
| 16,25 | 0,00% | | -1,05% | 30,00 |
| 16,38 | 0,80% | | -0,18% | 59,00 |
| 16,35 | -0,18% | | 0,68% | 65,00 |
| 16,41 | 0,37% | | 1,55% | 39,00 |
| 16,66 | 1,51% | | 2,42% | 24,00 |
| 16,55 | -0,66% | | 3,29% | 5,00 |
| 16,47 | -0,48% | | 4,15% | 4,00 |
| 16,49 | 0,12% | | 5,02% | 2,00 |
| 16,60 | 0,66% | | 5,89% | 0,00 |
| 16,51 | -0,54% | | 6,77% | 1,00 |
| 16,63 | 0,72% | | | |

**CUADRO 3: FUNCIÓN DE "ANÁLISIS DE DATOS" DE EXCEL**




| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 | *Clase* | *Frecuencia* | *% acumulado* | *Clase* | *Frecuencia* | *% acumulado* |
| 2 | -0,07121816 | 1 | 0,40% | 0,00279447 | 70 | 28,00% |
| 3 | -0,06196658 | 0 | | | | |
| 4 | -0,052715 | 1 | | | | |
| 5 | -0,04346342 | 1 | | | | |
| 6 | -0,03421185 | 1 | | | | |
| 7 | -0,02496027 | 6 | | | | |
| 8 | -0,01570869 | 19 | 1 | | | |
| 9 | -0,00645711 | 42 | 2 | | | |
| 10 | 0,00279447 | 70 | 5 | | | |
| 11 | 0,01204604 | 55 | 7 | | | |
| 12 | 0,02129762 | 32 | 9 | | | |
| 13 | 0,0305492 | 15 | 9 | | | |
| 14 | 0,03980078 | 3 | 9 | | | |
| 15 | 0,04905235 | 3 | 99,60% | y mayor... | 1 | 100,00% |
| 16 | 0,05830393 | 0 | 99,60% | -0,06196658 | 0 | 100,00% |
| 17 | y mayor... | 1 | 100,00% | 0,05830393 | 0 | 100,00% |
| 18 | | | | | | |
| 19 | | | | | | |
| 20 | | | | | | |

## 5. ESTADÍSTICOS DE DOS VARIABLES Y CORRELACIÓN

En las secciones anteriores se ha tratado el análisis descriptivo de una sola variable, el siguiente paso que debemos dar es conocer cómo se pueden comportar dos variables analizadas de forma conjunta. Para ello debemos conocer la covarianza o varianza conjunta entre dos variables aleatorias.

### COVARIANZA

Se representa por Cov(X,Y) y es la media aritmética de los productos de las diferencias entre las observaciones de cada una de las variables y su valor medio.

$$cov_{x,y} = \frac{1}{n}\sum_{j=1}^{n}(x_j - \bar{x})(y_j - \bar{y})$$

- **covarianza negativa**, significa que existe un comovimiento entre las variables en dirección contraria.
- **valor de la covarianza nulo**, significaría que existe independencia entre los movimientos de ambas series, o no existencia de una relación lineal entre las variables.

Propiedades matemáticas de la covarianza

$$Cov(X,X) = Var(X)$$
$$Cov(X,Y) = Cov(Y,X)$$

Sean a,b valores numéricos

$$Var(aX) = a2Var(X)$$
$$Var(a+X) = Var(X)$$
$$Cov(aX+b, cY+d) = abCov(X,Y)$$

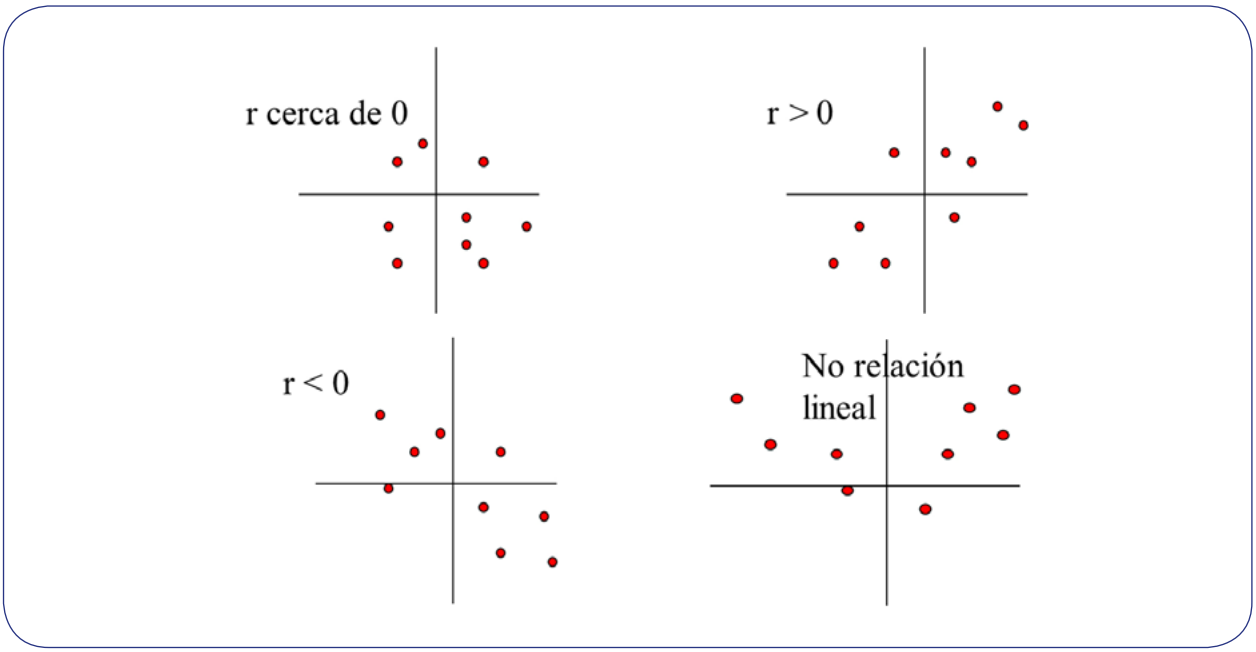
**CORRELACIÓN**

La correlación es al grado de asociación lineal entre dos variables, midiendo la intensidad de la misma mediante el coeficiente de correlación. El coeficiente de correlación $\rho$ exige que tanto la variable X como la Y sean variables aleatorias normales y que además la distribución conjunta de ambas variables siga una ley normal bivariante.

$$\rho = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{s_{xy}}{s_x s_y}$$

Propiedades del coeficiente de correlación:

**GRÁFICO 5: DIFERENTES TIPOS DE CORRELACIÓN**




## MATRIZ DE VARIANZAS-COVARIANZAS

La matriz de Varianzas-Covarianzas es un concepto muy útil para calcular la varianza de un conjunto de activos.

Si sólo hubiese dos activos:

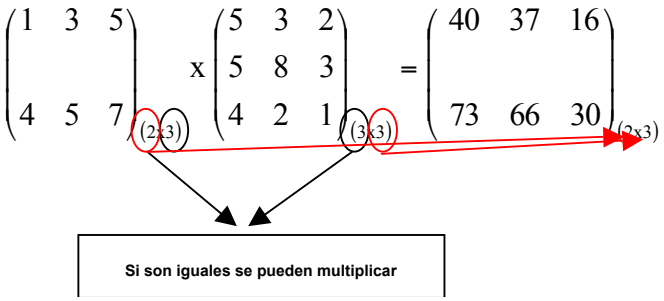
$$\sigma_c^2 = w_a^2 \cdot \sigma_a^2 + w_b^2 \cdot \sigma_b^2 + 2 w_a w_b Cov_{a,b}$$

Donde:

$W_a$ / $W_b$ = es el peso del activo A/B
$\sigma^2_a$ / $\sigma^2_b$ = es la varianza del activo A/B
$Cov_{a,b}$ = es la covarianza entre los activos A y B

Un pequeño apunte sobre la multiplicación de matrices. Existen matrices cuadradas (m x m), rectangulares (m x n) y vectores fila o columna.

Dos matrices, aunque no sean cuadradas se pueden multiplicar cuando la primera matriz tenga un numero de columnas n, que coincida con el número de filas de la segunda matriz. (m x n)




**Ejemplo**

Pongamos un ejemplo. Supongamos una cartera con tres activos: Telefónica, Mediaset y Santander con los siguientes datos:

- Desviación Típica TEF.=25,23%
- Desviación Típica TL5.=27,74%
- Desviación Típica SAN.=27,94%
- La correlación entre TEF y TL5=48,23%
- La correlación entre TEFy SAN=72,63%
- La correlación entre SAN y TL5=59,42%

Si suponemos que la ponderación de la cartera es la siguiente:

- TEF=30%
- TL5=30%
- SAN=40%

**CUADRO 4: EJEMPLO DE CÁLCULO DE LA VOLATILIDAD DE UNA CARTERA.**
FUENTE: ELABORACIÓN PROPIA

| | | | Σ | | | X' |
|---|---|---|---|---|---|---|
| X | | | 0,0002526 | 0,0001340 | 0,0002032 | 0,3000 |
| 0,3000 | 0,3000 | 0,4000 | 0,0001340 | 0,0003054 | 0,0001828 | 0,3000 |
| | | | 0,0002032 | 0,0001828 | 0,0003097 | 0,4000 |

$$\sigma_{Cartera} = \sqrt{X\Sigma X'}$$

$$X = W_{TEF} \quad W_{TL5} \quad W_{SAN}$$

$$\Sigma = \begin{matrix} \sigma^2_{TEF} & \sigma_{TEF-TL5} & \sigma_{TEF-SAN} \\ \sigma_{TEF-TL5} & \sigma^2_{TL5} & \sigma_{TL5-SAN} \\ \sigma_{TEF-SAN} & \sigma_{TL5-SAN} & \sigma^2_{SAN} \end{matrix}$$

$$X' = \begin{matrix} W_{TEF} \\ W_{TL5} \\ W_{SAN} \end{matrix}$$

| XΣ | | |
|---|---|---|
| 0,01972% | 0,02049% | 0,02397% |

| XΣX' |
|---|
| 0,02165% |

| Desviación Típica Cartera |
|---|
| 23,36% |

## 6. CONTRASTES DE HIPÓTESIS

Al aplicar la Estadística inferencial trabajamos en un marco de cierta aleatoriedad e incertidumbre en las observaciones. Toda investigación en el marco de la estadística diferencial suele barajar al menos dos opciones: una teoría o hipótesis que defiende el investigador, y otra la cual queremos invalidar.

Los contrastes de hipótesis es la herramienta estadística que disponemos para ello. A la primera teoría se la representa por $H_1$, y se la denomina hipótesis de investigación o alternativa, mientras que a la segunda se la representa como $H_0$, la cual es la hipótesis nula:

## CONTRASTES DE HIPÓTESIS PARAMÉTRICAS

Por lo general estamos interesados en saber si un determinado parámetro θ, que desconocemos, está en una determinada región del espacio paramétrico Θ, es decir, el contraste de hipótesis será:

$$\mathbf{H}_0\text{: } \theta \in \mathbf{\Theta}_0.$$
$$\mathbf{H}_1\text{: } \theta \in \mathbf{\Theta}_1.$$

donde es evidente que $\mathbf{\Theta} = \mathbf{\Theta}_0 \cup \mathbf{\Theta}_1$.

Para saber cuál de las dos hipótesis es la correcta formularemos un test de contraste de hipótesis:

Llamaremos test para contrastar la hipótesis nula $\mathbf{H}_0\text{: } \theta \in \mathbf{\Theta}_0$ frente a la hipótesis alternativa $\mathbf{H}_1\text{: } \theta \in \mathbf{\Theta}_1$ al test que consiste en decidir, para cada posible muestra obtenida, si aceptamos o rechazamos $H_0$. Por lo tanto un test consistirá en dividir el espacio muestral (conjunto de todas las posibles muestras) en dos regiones:

Una **región crítica *R***, en la cual rechazamos $\mathbf{H_0}$.
Una **región o zona de aceptación *A*,** en la cual aceptamos $\mathbf{H_0}$.

**GRÁFICO 6: ZONA DE ACEPTACIÓN DE TEST**

Dichas opciones llevan emparejados los siguientes errores en el caso de que no hayamos elegido bien:

- **Error de tipo I**: Hemos rechazado $\mathbf{H_0}$ cuando en realidad es cierta.
- **Error de tipo II**: Hemos rechazado $\mathbf{H_1}$ cuando en realidad es cierta.

**Ejemplo**

En 1969 se calculó que en EE.UU. un 8% del contenido de las basuras era metal. Debido al incremento de los procesos de reciclaje se espera que esta cifra se haya reducido. Se realiza un experimento para verificar esta suposición. Se pide:

a) Construir un contaste de hipótesis.
b) Explicar, en términos prácticos, los errores de tipo I y tipo II.

Llamaremos **nivel de significación de un test** o **tamaño de un test con región crítica *R*** para contrastar $\mathbf{H_0}$: $\theta \in \Theta_0$ frente a $\mathbf{H_1}$: $\theta \in \Theta_1$ al valor

$$\alpha = \max_{\theta \in \Theta_0} P_\theta(R)$$

es decir, a la máxima probabilidad de cometer el error de tipo I.

**CUADRO 5: ERRORES TIPO I Y II.**

| | ESTADO | VERDADERO |
|---|---|---|
| **DECISIÓN** | H0 cierta | H1 cierta |
| Se rechaza $H_0$ | Error tipo I (Probabilidad = α) | Decisión correcta (Probabilidad = Potencia) |
| Se rechaza $H_1$ | Decisión correcta | Error tipo II |

## 7. REGRESIÓN LINEAL

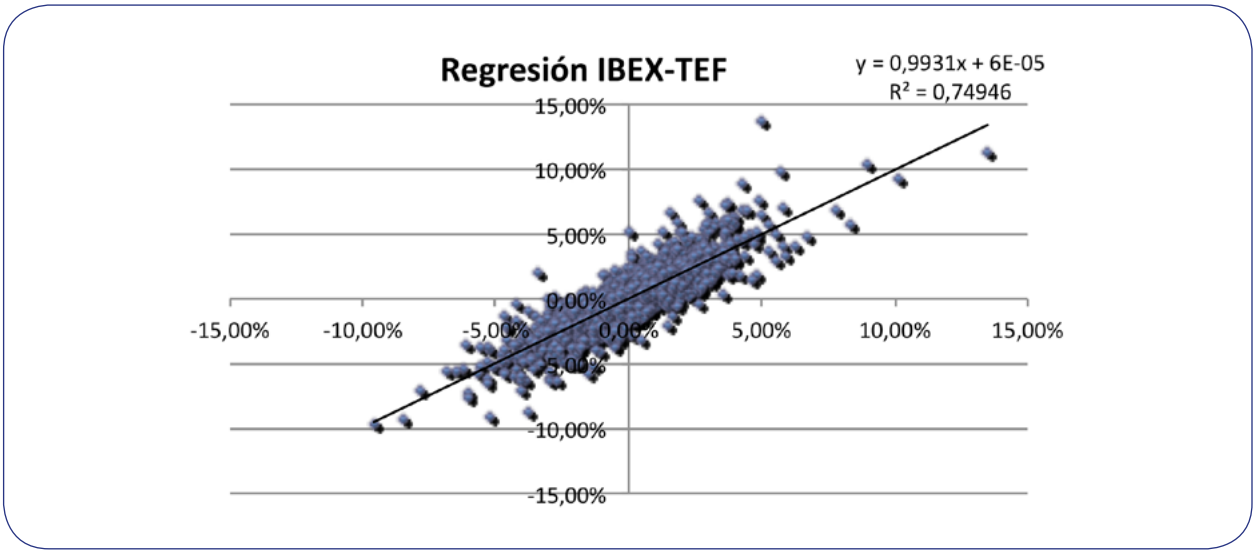
### RECTA DE REGRESIÓN: EL MÉTODO DE LOS MÍNIMOS CUADRADOS

Sea una muestra de dos variables $X$ e $Y$: $(x_1, y_1)$, $(x_2, y_2)$, …, $(x_n, y_n)$. Supongamos que:

- la variable $Y$ depende de la variable $X$. Expresado matemáticamente: $Y = f(X)$.
- la variable $Y$ está en función **lineal** de $X$. Expresado matemáticamente: $Y = aX + b$.

Se pretende, por tanto, encontrar la ecuación de una recta que aproxime los valores de la variable dependiente $Y$ en función de la variable independiente $X$. A tal recta la denominaremos **recta de regresión de Y sobre X**.

**GRÁFICO 7: REGRESIÓN DE IBEX Y TEF. MUESTRA UNA RELACIÓN ENTRE LAS VARIABLES.**
FUENTE: ELABORACIÓN PROPIA




Para determinar dicha recta emplearemos el método de los **mínimos cuadrados**. Se pretende minimizar la media de las distancias verticales de los puntos de la muestra a la recta de regre-

Desarrollando:

$$\text{E.C.M.}(a, b) = \frac{1}{n}\left(\sum_{j=1}^{n} y_j^2 + nb^2 + a^2\sum_{j=1}^{n} x_j^2 - 2b\sum_{j=1}^{n} y_j - 2a\sum_{j=1}^{n} x_j y_j + 2ab\sum_{j=1}^{n} x_j\right)$$

Derivando con respecto a *a* y a *b* llegamos a un sistema de ecuaciones el cual, tiene por solución:

$$a = \frac{\text{cov}_{x,y}}{v_x} \qquad\qquad b = \bar{y} - \frac{\text{cov}_{x,y}}{v_x}\bar{x}$$

Y sustituyendo en la recta de regresión:

$$y - \bar{y} = \frac{\text{cov}_{x,y}}{v_x}(x - \bar{x})$$

Se pretende ahora estimar si la recta representa de una manera eficaz a los puntos de la muestra, es decir, se quiere calcular el error que se comete al evaluar un dato mediante la recta de regresión. Para ello nos basamos en el error cuadrático medio, el cual, a partir de ahora lo denominaremos **varianza residual**:

Llamaremos **varianza residual** al error cuadrático medio cometido por la recta de regresión de *Y* sobre *X*.

Dicho valor vendrá dado por:

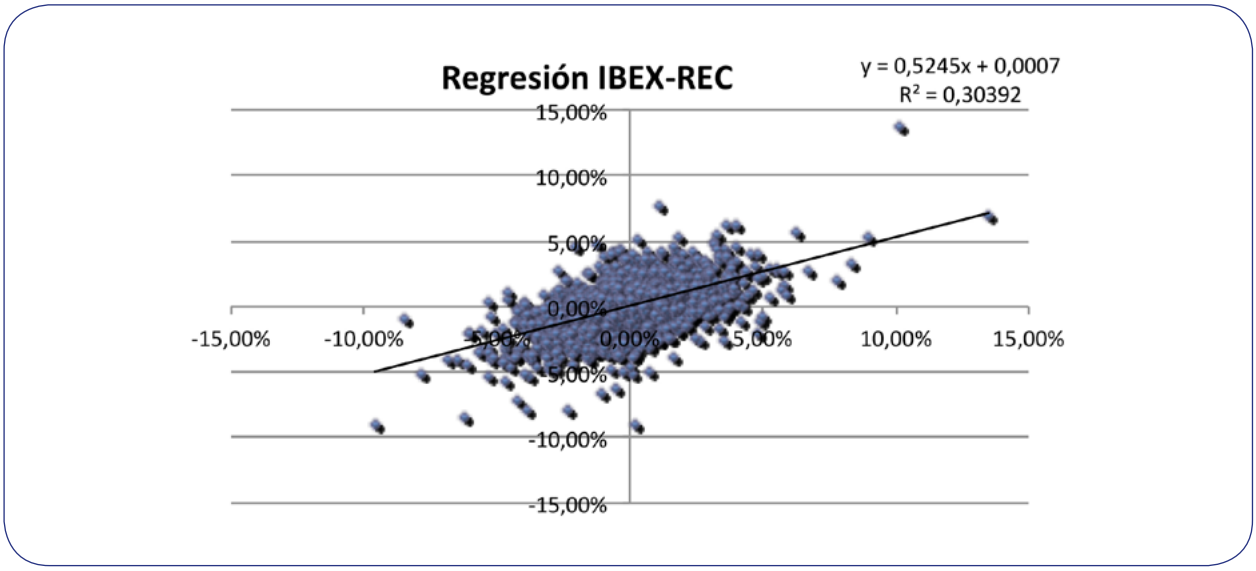
$$\text{Varianza residual} = \frac{1}{n}\sum_{j=1}^{n}(y_i - ax_i - b)^2 = \frac{1}{n}\sum_{j=1}^{n}\left(y_i - \bar{y} + \frac{\text{cov}_{x,y}}{v_x}(\bar{x} - x_i)\right)^2 = \ldots = v_y\left(1 - \frac{(\text{cov}_{x,y})^2}{v_x v_y}\right)$$

El cociente en la última expresión merece una mención especial:

Llamaremos **coeficiente de correlación muestral** entre *X* e *Y* (y se representará como *r*). El coeficiente de correlación muestral está comprendido entre –1 y 1.

$r^2$ es el coeficiente de bondad de ajuste e indica cómo de bien representa la recta de regresión al conjunto de datos. Es un valor entre 0 y 1, siendo 1 el ajuste perfecto (100%) y 0 el peor ajuste. Si la recta de regresión es representativa o no (fiabilidad de datos para estimaciones). Normalmente se considera que la recta de regresión es lo suficientemente representativa por encima de 0,75.

**GRÁFICO 8: REGRESIÓN ENTRE IBEX Y RED ELÉCTRICA MUESTRA UN R2 BASTANTE BAJO.**
FUENTE: ELABORACIÓN PROPIA




## REGRESIONES NO LINEALES

En la mayoría de los casos no es una recta la que mejor aproxima los datos de la muestra. Intentaremos aproximar la variable dependiente Y en función de la variable independiente *X* de la forma $Y = a \cdot g(X) + b$.

Procederemos de la siguiente manera:

a) Para ello definiremos una nueva variable aleatoria $T = g(X)$.

b) Hallaremos la recta de regresión de la variable Y en la variable T, obtendremos

## REGRESIONES NO LINEALES MÁS HABITUALES

Para saber el tipo de regresión que es más conveniente, conviene hacer la representación gráfica previamente.

**Regresión logarítmica:** Se ajustará mediante $Y = a \cdot \ln X + b$, (Aquí $T = \ln X$)

**Regresión exponencial:** Se ajustará mediante $Y = a \cdot e^{bx}$ (Tomando logaritmos se llega a que $\ln Y = \ln a + bX$, esto es, estamos en el caso anterior, llamando $T = \ln Y$ y haciendo la recta de regresión de $T$ sobre $X$)