Métodos cualitativos de análisis gráfico

Importancia de la representación gráfica

La presentación y análisis de los resultados experimentales debe considerarse como parte integral de los experimentos. Es realmente útil que los datos obtenidos se presenten en un gráfico, donde quede concentrada la información para su apreciación y análisis. En la mayoría de los casos un gráfico es más útil que una tabla de valores, especialmente en los casos en que:[1]

- los experimentos se llevan a cabo midiendo una variable Y en función de otra X que se varía independientemente y se quiere interpretar o determinar la relación funcional entre ellas. Por ejemplo: medición del período de un péndulo en función de su longitud; medición de la caída de potencial en un alambre en función de la corriente aplicada; etc.

- interesa estudiar si dos variables mantienen una correlación (causal o no) y cuál es el grado de vinculación o grado de independencia. Por ejemplo: estudio de la relación entre el peso y la altura de personas; relación de consumo de gas natural y la temperatura; relación entre la velocidad máxima que alcanza un velero y su extensión desde proa a popa; etc.

Se trata de que la información que se quiere representar quede expuesta de una manera lo suficientemente clara y explícita como para que la representación gráfica "hable por sí sola". Lo importante es que un gráfico debe servir para un posterior tratamiento de los datos, que lleve a inferir las leyes subyacentes en ellos y ahondar así en las posibles implicaciones y generalizaciones de los resultados obtenidos en los experimentos.

Un gráfico debe construirse sobre la base de una elección adecuada tanto de las variables como de las escalas de los ejes. Dado que los experimentos propuestos en este curso están pensados para estudiar metodológicamente numerosos problemas de las ciencias, en este capítulo daremos las bases que nos ayuden a efectuar una adecuada representación gráfica de datos experimentales, ya sean recogidos por nosotros en un experimento o bien tomados de fuentes confiables. Comentaremos diversas opciones que se presentan y sobre algunos métodos numéricos de utilidad para el tratamiento general de los datos.

Elección de las variables

De una manera muy general, cuando estudiamos la fenomenología de un sistema cualquiera, tratamos de obtener las variaciones o respuestas del sistema ante ciertas perturbaciones que podemos aplicarle de manera controlada. La Fig. 1 representa esquemáticamente un sistema bajo estudio.

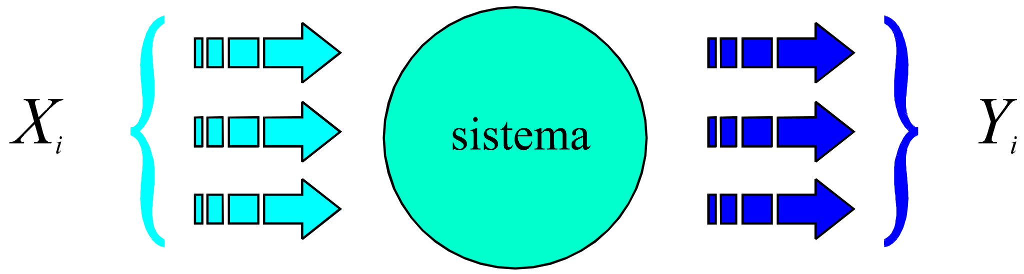


Figura 1 Representación de un sistema al que se estudia las respuestas Y_i cuando se varía el conjunto de variables X_i.

Hemos llamado X_i a las "*variables de entrada o independientes*", o sea aquellas que podemos controlar y variar. Ante los cambios de X_i, el sistema revela sus características o comportamientos a través de los cambios que sufren las variables Y_i, que pueden llamarse las "*variables de salida o dependientes*".[1,2, 3] Si deseamos estudiar un sistema, es conveniente, siempre que sea posible, aislar lo más posible las variables en estudio. Para ello es conveniente diseñar el experimento de modo tal que solo un parámetro relevante varíe por vez, manteniendo los restantes parámetros constantes. De este modo podremos concentrarnos en la respuesta de una de las variables de salida ante las variaciones de solamente una de las variables de entrada. Siempre que esto sea posible, esto es muy conveniente para simplificar el análisis. Afortunadamente esta es una situación muy común en los experimentos, aunque no siempre posible. En sistemas de mayor complejidad, en los que no podemos aislar los parámetros que varían de a uno por vez, el tipo de análisis que mostraremos puede generalizarse para tratar esos casos.[2] En lo que sigue nos apoyaremos en algunas relaciones funcionales simples con las que nos encontramos a menudo en el trabajo en el laboratorio y las usaremos para ejemplificar las ideas básicas.

Relación lineal

Una relación lineal entre las variables X e Y

$$Y = a \cdot X + b \tag{1}$$

es la más simple de todas. La representación gráfica de $Y(X)$ arrojaría una línea recta, de pendiente a y que corta al eje vertical en b (ordenada del origen o intersección con el eje y). Es importante notar que una recta es la forma geométrica más simple en dos dimensiones. Al mismo tiempo, una relación lineal entre dos variables cualesquiera es más fácil de ser identificada a simple vista, y no sería una exageración afirmar que éste es el único caso en que esta discriminación puede hacerse a simple vista. Entre una recta y una curva nuestro ojo siempre notará la diferencia, pero no discriminará a la función que define la curva.[4]

En la Fig. 2 están representadas dos series de datos. Para inferir cualitativamente cuál de las series puede aproximarse mejor por una relación lineal entre las variables X e Y, es útil la siguiente regla práctica: llevemos el papel hasta el nivel de nuestros ojos (podemos cerrar uno como cuando hacemos puntería) y veamos si los puntos se ven alineados. Este tipo de toma de decisión no debe desdeñarse en el momento de analizar datos experimentales. La decisión de aceptar o no una relación lineal entre las variables debe ser tomada por el experimentador, ya sea se espere o no una vinculación lineal entre las variables en juego. Una vez que decidimos que los datos “caen sobre una recta”, recién podremos estimar los parámetros (pendiente y ordenada al origen) de la *mejor recta* que aproxime la relación funcional: O bien podemos dibujar esa mejor recta y definirle los valores de la pendiente y la ordenada al origen, o usar métodos numéricos más generales para encontrarlos, como veremos más adelante.

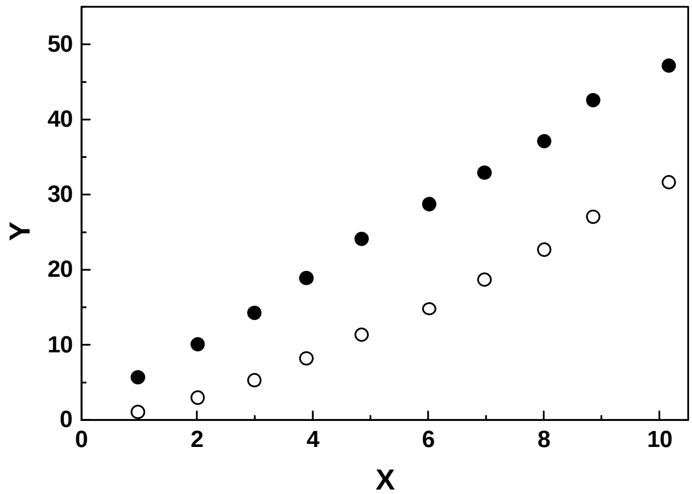


Figura 2 Representación de dos series de datos. ¿Cuál aproxima mejor una relación lineal entre las variables X e Y?

Función potencial

Supongamos que medimos pares de valores (X,Y) y tenemos conocimiento que la relación funcional que los vincula es del tipo

$$Y = aX^{b} \tag{2}$$

donde a y b son constantes. Esta forma funcional potencial es muy común en las ciencias puesto que sirve como aproximación del comportamiento en una gran variedad de casos.

La constante b suele llamarse "exponente de escala" y define la escala de variación de Y según varía X. Esto es, si X se multiplica por un factor f, Y cambiará consecuentemente f^{b} veces.

El significado físico de la constante a es el de representar el valor que toma Y cuando X vale la unidad. La dimensión de a es tal que da homogeneidad dimensional a la ecuación.

Lectura de ecuaciones: Algunas investigaciones muestran que la masa de los dinosaurios M estaba bien correlacionada con la longitud l de los animales medida desde la cabeza a la cola, según[4]

$$M = M_0 l^3$$

Leamos esta ecuación. El significado de M_0 es que representa la masa de un dinosaurio de "largo unidad". Por tanto, si la unidad elegida para la longitud es el metro y para la masa el kg, M_0 representa cuántos kilogramos pesaba un animal de largo igual a 1 m. La *unidad* de M_0 será tal que se igualen las unidades de los dos miembros de la ecuación. En este caso, M_0 tendrá la unidad kg/m^3. Sin embargo, M_0 no es la densidad de los animales, a pesar de su unidad, puesto que l^3 no es el volumen. Notemos que el valor de M_0 cambiará si se eligen otras unidades de medición. Por ejemplo, si el peso se midiera en gramos (g) y la longitud en centímetros (cm), M_0 adoptaría un nuevo valor, que sería M_0* (g / cm^3) =10^{-3}M$_0$(kg/m^3), a lo que se llega tras pasar los kilogramos a gramos y los metros a centímetros.

De manera más general, y sin recurrir a unidades particulares, podemos analizar cuál es la *dimensión* de M_0. Si usamos corchetes [...] para representar la dimensión de una cantidad, entonces $[M_0] = \frac{[M]}{[l]^3}$. Escribamos esta relación dimensional en términos de las dimensiones fundamentales masa, longitud y tiempo, a las que llamaremos M, L y T, respectivamente. Dado que $[M] = [m] = M$, resulta, luego de simplificar: $[M_0] = M / L^3$.

Este tipo de análisis puede usarse como prueba de consistencia de una fórmula complicada, o bien para determinar la dimensión de alguna variable introducida en un problema particular.

✍ La cantidad de potencia $\dot{Q}$ irradiada por unidad de área por un cuerpo negro que está a la temperatura absoluta T está dada por la ley de Stefan-Boltzmann

$$\dot{Q} = \sigma T^4$$

donde $\sigma = 5.67 \times 10^{-8}\ W/\ m^2\ K^4$ es la constante de Stefan-Boltzmann. a) Analice cuál es el significado físico de σ. b) Si el cuerpo negro estuviera a la temperatura de 2 K, ¿cuánto más irradiaría respecto a cuando se mantiene a 1 K? c) ¿A qué temperatura irradiará 25 veces más que a 2 K?

Si representamos datos medidos de Y en función de X , relacionados por una expresión como (2), lo que obtenemos en el caso $b \neq 1$ es una curva. De nuestro análisis cualitativo del gráfico observaremos una curva "cóncava hacia arriba" si $b > 1$, mientras que si $b < 1$, la curva se verá "cóncava hacia abajo". Lo que cualquiera de los casos precedentes significa es que una variación de la variable X a un dado ritmo, hace que la variable Y cambie a un ritmo distinto: más rápido si $b > 1$, más lento si $b < 1$. Esta observación cualitativa (en términos de "más rápido" o "más lento", bien puede ser buena en una gran variedad de casos de interés en el laboratorio cuando estemos interesados en la descripción general (tendencia) de algún fenómeno.

Transformación de variables

Si en la Ec. (2) transformamos las variables haciendo el cambio (suponiendo que conocemos el exponente b):

$$X^* = X^b \qquad\qquad Y^* = Y$$

y representamos las nuevas variables $(X^*, Y^*) = (X^b, Y)$, lo que obtenemos es una relación lineal entre las *variables transformadas* o *pseudovariables* y decimos que hemos *linealizado* la representación gráfica. En este caso hemos transformamos la variable X, pero bien podríamos haber optado por el cambio en la variable dependiente, o sea,

$$X^* = X \qquad\qquad Y^* = Y^{1/b}$$

y también habríamos obtenido una relación lineal entre las nuevas variables representadas $(X^*, Y^*) = (X, Y^{1/b})$.

Está claro que lo anterior es inmediato de realizar si conocemos el valor del exponente b. Además, observamos que un gráfico linealizado nos da el valor de la constante a [ver Ec. (2)] si evaluamos la pendiente de la recta que resulta.

📖 Se mide el período T de un péndulo simple para distintas longitudes L. En el caso de pequeñas amplitudes de oscilación, ambas variables están relacionadas por

$$T = 2\pi\sqrt{\frac{L}{g}}$$

donde g es la aceleración de la gravedad. La relación es del tipo

$$T = aL^b$$

con $a = \frac{2\pi}{\sqrt{g}}$ y $b = \frac{1}{2}$.

De un gráfico de T en función de L^b evaluamos la constante a, con lo que podemos obtener el valor de la aceleración gravitatoria g indirectamente. Es de esperar que el gráfico resulte en una recta que pase por el origen de coordenadas, dado que un péndulo de longitud nula tiene que tener un período de oscilación nulo.

En el caso más general, supongamos que no conocemos a a ni a b, y que ambas constantes deben encontrarse como resultado de la investigación experimental. Entonces, ¿cómo procedemos?

Para facilitar la tarea de encontrar tanto el exponente de escala b como la constante a, es conveniente representar $\log(Y)$ en función de $\log(X)$. Esto queda claro si transformamos nuestra ecuación original más general $Y = aX^b$, sacándole el logaritmo a ambos miembros

$$\log(Y) = \log(aX^b) \tag{3}$$

$$\log(Y) = \log(a) + \log(X^b) \qquad (4)$$

$$\log(Y) = \log(a) + b\log(X) \qquad (5)$$

Comparando esta última expresión con un gráfico de $\log(Y)$ en función de $\log(X)$ podremos ver que la ecuación representa una recta que tiene pendiente b y ordenada al origen $\log(a)$.

Este tipo de representación gráfica es extremadamente útil cuando se analizan ecuaciones algebraicas, se estudian correlaciones, leyes de crecimiento, etc.

Elección de las escalas

Hemos visto cómo elegir las nuevas variables con el fin de llevar la representación gráfica a una representación lineal. Lo que hemos propuesto es la transformación de las variables y la representación de las nuevas. Una manera alternativa de análisis es recurrir a gráficos en los que sus ejes tengan escalas logarítmicas.

Retomando el ejemplo del caso de variables *X*, *Y* relacionadas por la función potencial $Y = aX^b$, en vez de recurrir a un gráfico de variables transformadas [log (*X*), log(*Y*)], podemos representar directamente los pares de valores (*X*, *Y*) en un gráfico donde sus dos ejes tengan escalas logarítmicas. La Fig. 3 ejemplifica este método.

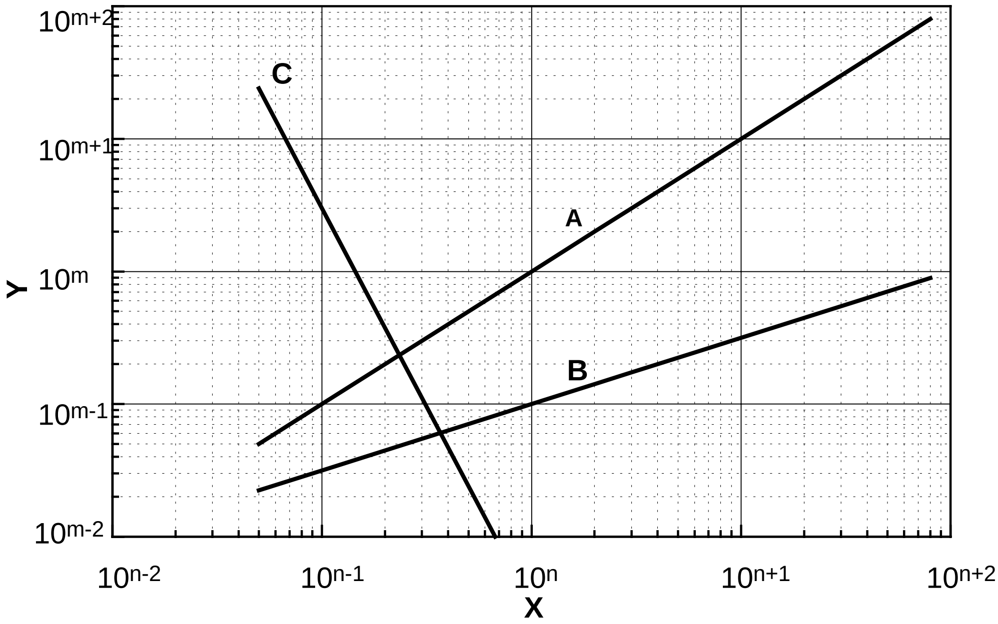


Figura 3 Ejemplo de un gráfico con escalas logarítmicas.

Un *gráfico doble-logarítmico* como el de la Fig. 3 también es llamado *gráfico log–log*. La posición de las grillas más gruesas identifica un valor igual a una potencia de 10. Por lo tanto, en cada eje, el espacio entre esas grillas representa una década de variación de las variables, es decir, entre 10^n y 10^{n+1}, cualquiera sea n. Las ocho grillas intermedias indexan los valores $k \times 10^n$, $k = 2, 3, 4, ..., 9$.

Esto hace muy simple la construcción de ejes en escalas logarítmicas. Esto requiere marcar intervalos fijos a distancias 1, 10, 100, 1000, ... (10^0, 10^1, 10^2, 10^3, ...). Si los datos a representar no cubren un rango tan amplio de valores, los intervalos pueden realizarse a distancias de 1, 2, 4, 8, 16, 32, ... (2^0, 2^1, 2^2, 2^3, 2^4, 2^5, ...).

Observando la Fig. 3 podemos darnos cuenta que las escalas logarítmicas son “más democráticas” que las lineales,[3] puesto que dejan ocupar el mismo espacio en el gráfico a los intervalos entre décadas entre valores “pequeños” que el espacio ocupado por los intervalos entre décadas entre valores “grandes”; podemos ver, por ejemplo, que el lugar reservado para los valores entre 10^{-5} y 10^{-4} es idéntico al reservado para el intervalo 10^8 y 10^9.

Si las variables X e Y se representan ambas en escalas logarítmicas, la función potencial de la Ec. (2) quedará representada por una recta cuya pendiente es b y cuya ordenada al origen $ord = \log(a)$, por lo que $a = 10^{ord}$.

A su vez, si los datos (X, Y) representados en un gráfico doble-logarítmico siguen una relación lineal, podemos inferir que $Y \propto X^b$, descubriendo en este caso la ley subyacente. Para calcular directamente del gráfico el valor de b, hay que contar cuántas décadas varía Y cuando X varía una. En el ejemplo de la Fig. 3, la línea A tiene pendiente $b = 1$, por tanto $Y \propto X$. Para la línea B, b = 1/2, lo que implica $Y \propto \sqrt{X}$. ¿Cuál es la pendiente de la recta C?

Esta representación puede hacerse sobre papeles especialmente diseñados (papel logarítmico) que se consigue en las librerías. Con las ventajas que ofrecen hoy en día los programas de computadora, este tipo de representación puede realizarse de manera inmediata para sacar mayor provecho al análisis de los datos experimentales. Muchos programas de análisis de datos o planillas de cálculo, tales como Excel, QuatroPro, Origin, etc., permiten realizar estos cambios muy fácilmente. Una vez realizado el gráfico en escala lineal, picando o activando con el *mouse* los ejes coordenados, se abre un sub-menú que permite variar la escala de los ejes (lineal, logarítmica, etc.).

Aplicaciones de gráficos log–log

Hay una gran variedad de casos donde es sumamente ventajosa la representación gráfica en escalas logarítmicas. En cierta manera, es a lo que recurre un experimentador de inmediato cuando quiere darse cuenta de "la forma de la ley que siguen sus datos". Efectivamente, es una manera rápida y eficiente de evaluar las tendencias de los resultados y dar un primer paso en el análisis. También es útil recurrir a estas escalas en los casos en los que el rango de valores es muy amplio y los datos que manejamos varían en varios órdenes de magnitud.

Amplio margen de valores: Como ejemplo podemos citar el problema de investigar si existe correlación entre el calor metabólico producido por mamíferos y su masa, analizando datos provenientes de experimentos que involucren desde ratones, con pesos de unas decenas de gramos (≈10 g), hasta elefantes que pesan varias toneladas (≈10^6 g), incluyendo especies de tamaños intermedios, como gatos (≈10^3 g), monos (≈10^4 g), etc. Es claro que si en un gráfico "calor – masa" usamos una escala lineal para el peso, la necesidad de incluir semejante margen de valores –entre 0 g y 10^6 g– tendrá como ingrata consecuencia el "amontonamiento", cerca del origen de coordenadas, de los datos de las especies más pequeñas. La elección de una escala logarítmica en el eje del peso eliminaría inmediatamente tal inconveniente.

Relación entre magnitudes: En una situación usual en el laboratorio, podríamos estar interesados en saber si una muestra conductora puede describirse como un conductor óhmico. Si hacemos circular una corriente eléctrica I por la muestra y medimos la diferencia de potencial V que se produce, y repetimos este procedimiento para varios valores de la intensidad de la corriente, tendremos los datos $V – I$ para llevar a un gráfico. En primera instancia, un gráfico con escalas lineales sirve para abrir el juego. Aquí caben varias posibilidades. Una de ellas es usar nuestro ojo, como ya comentamos, y decidir si la relación entre V e I puede considerarse lineal, lo que nos diría si el material conductor es óhmico (la relación entre V e I es lineal) o no. Pero supongamos que queremos ir más allá. Si verdaderamente el conductor es óhmico, un gráfico log–log en las variables V e I debería resultar lineal y, lo que es más importante, la pendiente de la recta debería ser igual a uno (el exponente de I es uno). De ser así, la conclusión sobre que el material es óhmico adquiere más valor tras este análisis.

Búsqueda de posibles correlaciones entre variables: Puede obtenerse mucha información cualitativa de un experimento si se conocen las proporcionalidades entre las variables involucradas. En este sentido podemos aprovechar un gráfico log–log para pronosticar tendencias. Podría ser deseable "anticiparnos" al resultado de un experimento –más aun si es caro o de largo aliento– estableciendo las leyes de escala entre las variables, para así saber cómo varía la variable de salida Y frente a un cambio de la variable de entrada *X*. Esto redundaría en mejoras sobre la marcha de nuestros diseños y estrategias experimentales.

✍ **Cuándo conviene usar escalas logarítmicas:** Se mide una propiedad Σ de 100 ml de un líquido puro. Luego se lo diluye en agua al 10% y se repite la medición. La solución se diluye otra vez al 10% y se mide de nuevo. La operación se repite cuatro veces más. Para cada muestra se obtiene Σ = 1 (líquido puro), 1.02, 1.04, 1.08, 1.24, 1.59, 1.95 (en unidades arbitrarias). Representar en un gráfico apropiado el resultado del experimento.

☺ La imposibilidad de alcanzar físicamente el cero absoluto de temperatura siempre ha cautivado la atención de los hombres de ciencia. Una manera de hacer una analogía de esta imposibilidad la ofrecen los físicos que estudian propiedades de la materia a bajas temperaturas, combinando temperatura con dinero.

a) Imaginemos un cuerpo que está inicialmente a la temperatura de 100 K y que cuesta $1 reducirle la temperatura 10 K. Cuando está a 90 K nos cuesta $1 llevarlo a 80 K, y $1 más para llevarlo a 70 K, y así sucesivamente. Con este procedimiento, ¿cuánto cuesta enfriarlo hasta el cero absoluto?
b) Ahora consideremos el mismo cuerpo a la temperatura inicial de 100 K. Pero el procedimiento de enfriamiento consiste en pagar $1 para llevarlo a 10 K. Cuando está a 10 K nos cuesta $1 para llevarlo a 1 K, y otro peso para que alcance 0.1 K, y así sucesivamente. Con este procedimiento, ¿cuál es el costo de enfriarlo hasta el cero absoluto?

La ley exponencial

Un caso particular de mucho interés por su aplicación en muchos problemas físicos, biológicos, de ingeniería y hasta financieros y económicos, es el de una relación exponencial entre dos variables. Para fijar ideas supongamos que estamos considerando dos variables, Y_1 e Y_2, como función del tiempo *t*. Si las relaciones entre estas variables son:

$$Y_1(t) = Ae^{-\lambda_1 t} \qquad (6)$$

y

$$Y_2(t) = A(1 - e^{-\lambda_2 t}) \qquad (7)$$

sus representaciones gráficas lucirán en escala semilogarítmica como se muestra en la Fig. 4. Es fácil notar que, si bien la primera de estas relaciones (Y_1) se "linealiza" en este tipo de gráfico, la segunda (Y_2) no lo hace. En este caso, es conveniente recordar que la derivada de ambas expresiones sí tienen una relación funcional simple, a saber:

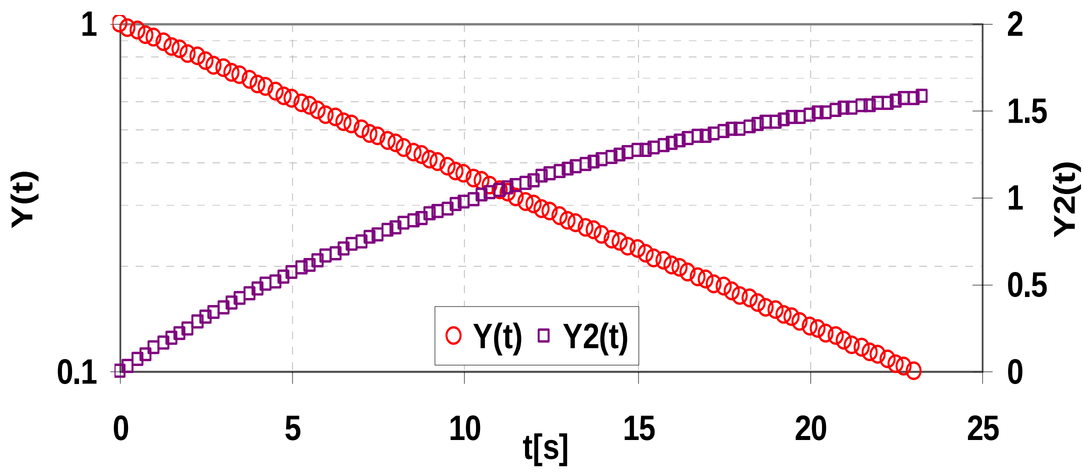


Figura 4 Representación en escala semilogarítmica de las funciones (xx.6) y (xx.7)

$$\frac{dY_1(t)}{dt} = -A\lambda_1 e^{-\lambda_1 t} = -\lambda_1 Y_1(t) \qquad (8)$$

y

$$\frac{dY_2(t)}{dt} = A\lambda_2 e^{-\lambda_2 t} = \lambda_2 (A - Y_2(t)) \qquad (9)$$

Por lo tanto, usando una representación de la *derivada en función de la variable dependiente* (Y_1 o Y_2) es cuando obtenemos una recta. De los valores de la pendiente y la ordenada al origen de estas rectas (8) y (9), tenemos información sobre este tipo de relación, puesto que de ellos extraemos los parámetros λ_1 y λ_2. En la Fig. 5 se muestran las mismas funciones que en la Fig. 4 usando la representación propuesta. Es claro que esta alternativa es muy útil para estos problemas.

Una dificultad de esta representación es que requiere conocer la derivada de la función en cuestión y que para hacerlo debemos usar un procedimiento numérico. Si disponemos de mediciones de Y_1 e Y_2 en función de t lo que hacemos es aproximar la derivada calculando las diferencias finitas usando pares de datos consecutivos:

$$\frac{dY(t)}{dt} \approx \frac{Y_{i+1} - Y_i}{t_{i+1} - t_i} \tag{9}$$

Sin embargo, como los datos tienen errores, la diferencia (Y_{i+1} – Y_i) puede ser en algunos casos menor que el error de medición, y en tal caso el valor obtenido con (9) carece de significado. Una manera de mejorar la estimación de la derivada de datos experimentales consiste en usar un grupo de datos que estén en un intervalo donde *a priori* no se espere mucha variación en la derivada. Usando este grupo de valores elegidos (Y_i, t_i) aproximamos una recta que pase por ellos, cuya pendiente m tomamos como una estimación de la pendiente de la curva en el entorno de esos datos, o sea, hacemos una estimación local de la derivada dY/dt usando un grupo de valores en vez de usar pares consecutivos. El gráfico que hacemos finalmente es uno de m en función de Y. La mayoría de la hojas de cálculo usan este procedimiento para el cálculo de la derivada de una función representada por un conjunto de datos.

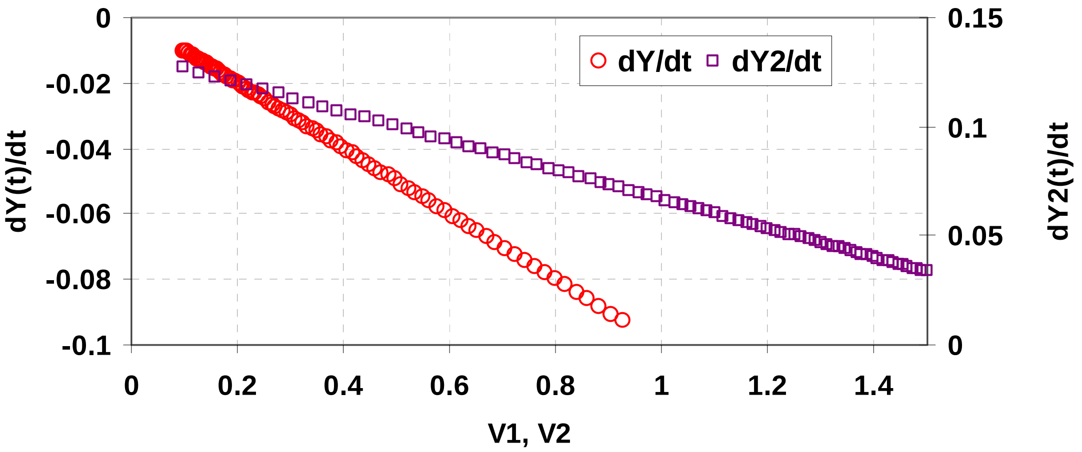


Figura 5 Representación en escala lineal de las derivadas en función de las variables dependientes. En este ejemplo, λ_2 es 50% menor que λ_1.

Diseño de gráficos

Los programas de representación gráfica disponibles en las computadoras incluyen entre sus opciones el diseño de gráficos usando los distintos tipos de escalas descriptas en este capítulo. Pero, ya sea que el gráfico vaya a realizarse usando estos programas o a mano, es conveniente considerar algunos "trucos del buen dibujante" para que la información contenida en el dibujo adquiera la relevancia que le corresponde. Es así que, además de la correcta elección de las variables y de las escalas, un gráfico adquirirá una mejor presentación si se cuidan algunos detalles, entre ellos:

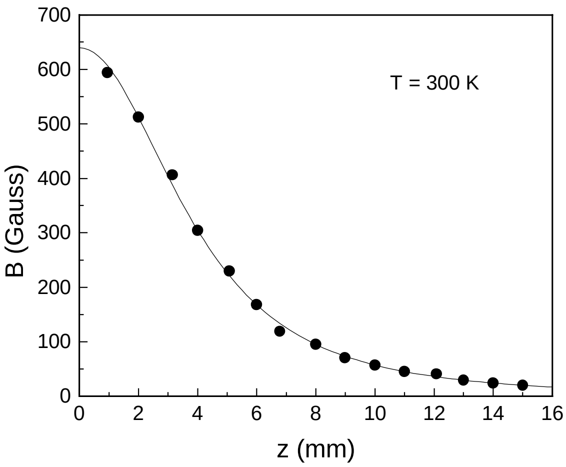
- identificación de los ejes con rótulos bien ubicados que digan qué variables se representan y en qué unidades se miden,

- uso de símbolos que ubiquen los datos (cuadrados, círculos, rombos, etc.), en lo posible con sus incertidumbres (en la forma de barras que indiquen el intervalo de incertidumbre); que haya una diferenciación de distintas series de datos cuando se presenten varios resultados, para lo que es recomendable el uso de diferentes símbolos,

- inclusión de un epígrafe, que es un texto descriptivo de lo que está representado en el gráfico y que además puede aportar alguna información adicional,

- carteles interiores al gráfico, con información complementaria relevante para entender en qué contexto se muestran los datos o sobre las condiciones experimentales particulares bajo las que se los han obtenido,

- una clara diferenciación entre lo que es propio del resultado experimental del trabajo y lo que corresponde a una comparación con una teoría o modelo propuesto (por ejemplo, usando líneas continuas) o a resultados extraídos de otras fuentes.



Campo magnético axial de un imán de Ne-Fe-B a temperatura ambiente medido con una sonda de efecto Hall. La línea es un ajuste de los datos.

Figura 6 Ejemplo de gráfico y epígrafe.

Referencias

1. S. Gil y E. Rodríguez, *Física re-Creativa*, Prentice Hall, Buenos Aires, 2001.; http://www.fisicarecreativa.com
2. D. C. Baird, *Experimentación*, 2ª ed., Prentice-Hall Hispanoamericana S.A., México, 1991.
3. Christopher Deacon, "The importance of graphs in undergraduate physics," *Phys. Teach.* **37**, 270, 1999.
4. E. Martínez, *Logarithmic Park*, Instituto Balseiro, Bariloche, 1997; http://cabbat1.cnea.gov.ar/apfa/apfa.htm

Métodos cuantitativos de análisis gráfico

Método de cuadrados mínimos – Regresión lineal

Hemos enfatizado sobre la importancia de las representaciones gráficas y hemos visto la utilidad de las versiones linealizadas de los gráficos (X, Y) junto a las distintas maneras de llevar a cabo la linealización. A menudo nos confrontamos con situaciones en las que encontramos o suponemos que existe una relación lineal entre las variables X e Y, surge de modo natural la pregunta: ¿cuál es la relación analítica que mejor se ajusta a nuestros datos? El *método de cuadrados mínimos* es un procedimiento general que nos permite responder esta pregunta. Cuando la relación entre las variables X e Y es lineal, el método de ajuste por cuadrados mínimos se denomina también *método de regresión lineal*. En este capítulo discutiremos este último caso. El lector puede consultar en el Apéndice F de la Ref. [1] una discusión del caso general de cuadrados mínimos cuando el modelo es no lineal y los datos están afectados de errores.

La Fig. 1 ilustra el caso lineal. La dispersión de los valores está asociada a la fluctuación de los valores de cada variable. Observamos o suponemos una tendencia lineal entre las variables y nos preguntamos sobre cuál es la *mejor recta*:

$$y(x) = a\ x + b \tag{1}$$

que representa este caso de interés.

Es útil definir la función χ^2 (Chi-cuadrado)[1-3]:

$$\chi^2 = \sum_i \left(y_i - (a \cdot x_i + b) \right)^2 \tag{2}$$

que es una medida de la desviación total de los valores observados y_i respecto de los predichos por el modelo lineal $a\ x + b$. Los mejores valores de la pendiente a y la ordenada al origen b son aquellos que minimizan esta desviación total, o sea, son los valores que remplazados en la Ec.(1) minimizan la función χ^2 de la Ec.(2). Los parámetros a y b pueden obtenerse usando técnicas matemáticas que hacen uso del cálculo diferencial. Aplicando estas técnicas, el problema de minimización se reduce al de resolver el par de ecuaciones:

$$\frac{d\chi^2}{da} = 0 \quad \text{y} \quad \frac{d\chi^2}{db} = 0 \tag{3}$$

de donde resulta.[1-4]

$$a = \frac{N\sum x_i y_i - \sum x_i \sum y_i}{N\sum x_i^2 - \left(\sum x_i\right)^2} \tag{4}$$

$$b = \frac{N\sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i}{N\sum x_i^2 - \left(\sum x_i\right)^2} \tag{5}$$

Actualmente, la mayoría de los programas de análisis de datos y planillas de cálculo, realizan el proceso de minimización en forma automática y dan los resultados de los mejores valores de *a* y *b*, o sea los valores indicado por la ecuaciones (4) y (5).

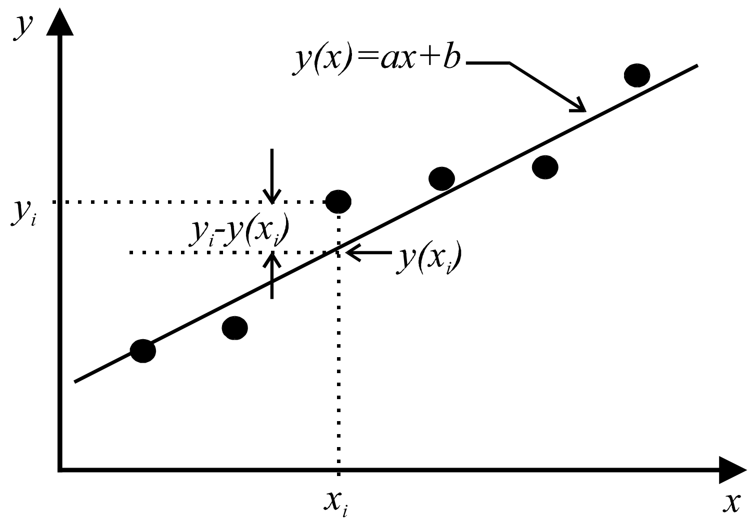


Figura 1 Gráfico de datos asociados a un modelo lineal. La cantidad y_i - $y(x_i)$ representa la desviación de cada observación de y_i respecto del valor predicho por el modelo $y(x)$.

El criterio de mínimos cuadrados reemplaza el juicio personal de quien mire los gráficos y defina cuál es la mejor recta. En los programas como Excel, Origin, etc., este cálculo se realiza usando la herramienta “regresión lineal” o “ajuste lineal”. Los resultados (4) y (5) se aplican en el caso lineal cuando todos los datos de la variable dependiente tienen la misma incertidumbre absoluta y la incertidumbre de la variable independiente se considera despreciable.

Una medida de la calidad o *bondad del ajuste* realizado viene dado por el *coeficiente de correlación* R^2 entre las variables X e Y, definido como:

$$R^2 = \frac{Cov(x,y)^2}{Var(x) \cdot Var(y)} \tag{6}$$

donde

$$Cov(x,y) = \frac{N\sum_{i=1}^{N} x_i \cdot y_i - \sum_{i=1}^{N} x_i \cdot \sum_{i=1}^{N} y_i}{N^2} \tag{7}$$

$$Var(x) = \frac{\sum_{i=1}^{N} x_i^2}{N} - \left(\frac{\sum_{i=1}^{N} x_i}{N} \right)^2 = <x^2> - <x>^2 \tag{8}$$

y

$$Var(y) = \frac{\sum_{i=1}^{N} y_i^2}{N} - \left(\frac{\sum_{i=1}^{N} y_i}{N} \right)^2 = <y^2> - <y>^2 \tag{9}$$

El valor de ***R*** varía entre –1 y 1. Si ***R*** es ±1 o próximo a estos valores, decimos que el modelo lineal es adecuado para describir los datos experimentales. Cuando ***R*** se aparta de estos extremos decimos que una expresión lineal no es una buena descripción de los datos. En este caso, conviene analizar el gráfico y buscar una relación no–lineal que aproxime mejor la dependencia. Dado que ***R*** mide el grado de correlación lineal entre los datos, si, por ejemplo, los pares de puntos (X, Y) tienen una relación tal que caen sobre un círculo, aunque ellos están correlacionados, tendríamos $\boldsymbol{R} \approx 0$. Desde luego, si los pares (X, Y) no tienen correlación alguna entre ellos, también tendríamos $\boldsymbol{R} \approx 0$. Ver la Figura 2.

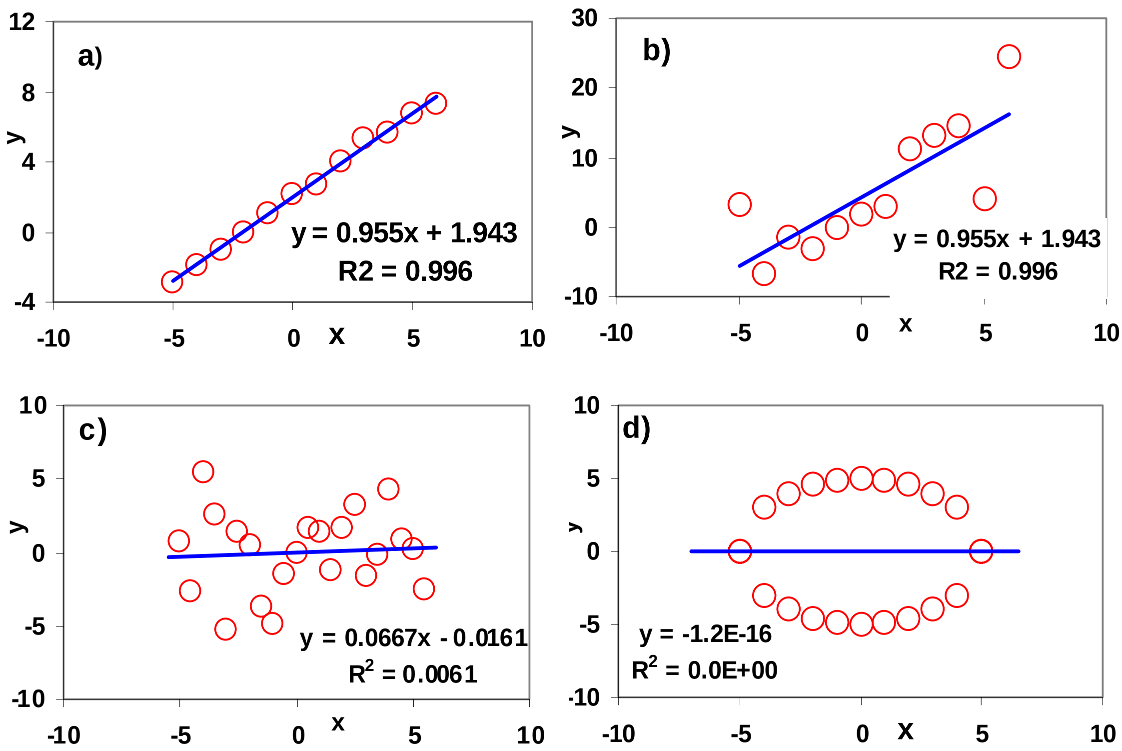


Figura 2 Ajuste de datos experimentales por un modelo lineal. a) Caso de una buena correlación lineal, b) aceptable, c) ex un caso en el prácticamente no ha correlación entre X e Y. , d) tiene un buen correlación pero el modelo lineal es inadecuado.

Frecuentemente el resultado que deseamos determinar de nuestro experimento es alguno de los parámetros de la Ec. (1). Por ejemplo, si deseamos determinar la constante elástica k de un resorte a partir de mediciones de fuerzas aplicadas F_i y estiramientos x_i que le producen al resorte, k será precisamente la pendiente de la recta que mejor se ajusta a los datos. Otro ejemplo es la obtención de la resistencia eléctrica R de un conductor, que deseamos determinar a partir de mediciones de tensión V_i y la corriente que lo atraviesa I_i. Por consiguiente, es útil disponer de un modo de estimar las incertidumbres asociadas a la determinación de los parámetros a y b de la Ec. (1). La importancia del método de cuadrados mínimos reside en el hecho que nos permite obtener valores de la desviación estándar o sea los errores asociados a los parámetros a

y *b* de la Ec. (1)[4], que denotaremos con los símbolos σ_a y σ_b**.** En esta sección sólo presentamos los resultados de utilidad más frecuente en el laboratorio; el lector interesado podrá encontrar un tratamiento más exhaustivo en las Ref.[1-4]. Las incertidumbres de los parámetros del ajuste vienen dadas por las expresiones:

$$\sigma_a = \sqrt{\frac{\chi_N^2}{N \cdot Var(x)}} \tag{10}$$

$$\sigma_b = \sqrt{\frac{\chi_N^2 \cdot \sum_{i=1}^{N} x_i^2}{N \cdot Var(x)}} \tag{11}$$

donde χ_N^2, conocido como el valor de Chi-cuadrado por grado de libertad, viene dada por:

$$\chi_N^2 = \frac{1}{N-2} \cdot \chi^2 \tag{12}$$

Las incertidumbres de los parámetros *a* y *b* también pueden escribirse en términos del coeficiente de correlación ***R*** del siguiente modo[5]:

$$\sigma_a = \sqrt{\frac{a^2}{(N-2)} \cdot \left(\frac{1}{R^2} - 1\right)} \tag{13}$$

$$\sigma_b = \sigma_a \sqrt{< x^2 >} \tag{14}$$

donde

$$< x^2 >= \frac{\sum_{i=1} x_i^2}{N} \tag{15}$$

Estas expresiones son de mucha utilidad para estimar σ_a y σ_b, ya que la mayoría de las planillas de cálculo y programas de ajuste, por los regular indican los valores de los parámetros *a* y *b* que mejor ajustan los datos y el valor de ***R***.

Precauciones en el análisis

No siempre es suficiente admitir que dos variables siguen una relación lineal guiándonos por lo que muestra un gráfico de los datos en escalas lineales. Menos aun si *sólo* evaluamos el coeficiente de correlación del ajuste lineal que propondríamos a partir de este gráfico. Un gráfico de $Y = X^{1.1}$ (variables sin correlación lineal) puede ajustarse por una recta y obtenerse a la vez un coeficiente de correlación lineal (inexistente) de, por ejemplo, 0.998. Un gráfico de datos experimentales de $Y = X$ con algo de dispersión fortuita de los puntos, podría devenir en un coeficiente de, por ejemplo, 0.995, menor que el anterior. Entre los coeficientes hay una diferencia, apenas, del 3 por mil. Pero en un gráfico log-log, la diferencia de pendientes será la que hay entre 1.1 y 1.0, lo que representa un 10% de discrepancia entre los exponentes de la variable *X*. Estos métodos de análisis nos enseñan que los efectos de correlación pueden estar enmascarados por el efecto del “ruido” de los datos. En ocasiones lo difícil es establecer si existe correlación entre las variables, aun cuando los datos provengan de fuentes “limpias”, que hayan producido datos con relativamente poca dispersión. Muchas veces la decisión entre dos alternativas debe hacerse usando otros criterios. Por ejemplo, la consistencia con otros conjuntos de datos o sobre la base de consideraciones de simetría o concordancia con teorías bien establecidas.

Ejemplo: Imaginemos un experimento donde se mide la distancia que recorre un móvil sobre una línea recta mientras una fuerza constante actúa sobre él. Esperamos que el movimiento sea uniformemente acelerado. Supongamos que el cuerpo parte del reposo, que medimos *x(t)* a tiempos largos y que los datos colectados son los representados en la Fig. 3.a.

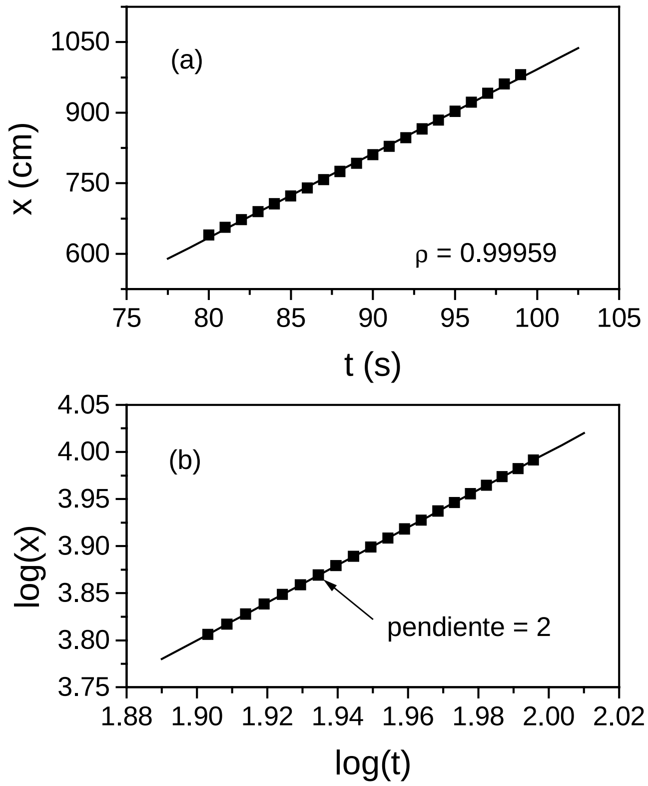


Figura 3 Representación de *x(t)* para un cuerpo que se mueve con movimiento uniformemente acelerado. (a) A tiempos largos no se aprecia bien la curvatura de la curva y, dado que el coeficiente de correlación lineal es muy cercano a la unidad, podría suponerse que la correlación es lineal. (b) *log(x)* en función de *log(t)*, de donde se deduce que la relación no es lineal sino cuadrática.

Si los datos experimentales se analizan sobre este gráfico de escalas lineales, el ajuste por un modelo lineal es más que tentador. Hecho esto, se obtiene la ecuación de la mejor recta y un coeficiente de correlación muy alto, R = 0.99959. Sin embargo, un modelo basado en las ecuaciones de la dinámica dice que

$$x = \frac{1}{2}at^2$$

donde a es la aceleración. En la Fig. 3.b están los logaritmos de los mismos datos, de donde se ve claramente la proporcionalidad $x \propto t^2$ que predice el modelo, difícilmente demostrable a partir del gráfico de la Fig. 3.a. Evidentemente, el uso de una aproximación lineal no es buena en este problema y el mero juicio del valor del coeficiente de correlación no es suficiente.

Referencias

1. S. Gil y E.Rodríguez, *Física re-Creativa*, Prentice Hall, Buenos Aires 2001.
2. P. Bevington and D. K. Robinson, *Data reduction and error analysis for the physical sciences*, 2nd ed., McGraw Hill, New York, 1993.
3. Stuardt L. Meyer, *Data analysis for scientists and engineers*, John Willey & Sons, Inc., New York, 1975.
4. D. C. Baird, *Experimentación*, 2ª ed., Prentice Hall Hispanoamericana S.A., México, 1991.

Aplicaciones de los métodos de análisis gráfico

Elija dos de los siguientes problemas para analizar y resolver. En cada caso confeccione un gráfico utilizando los datos de las tablas y proponga la mejor función que represente la relación entre las variables. Elabore sus conclusiones y preséntelas por escrito acompañadas de los gráficos que haya utilizado para el análisis.

I) Imagine que una empresa está interesada en invertir en el sector de Gas Natural Comprimido (GNC) en la Argentina. Para ello decide estudiar como ha evolucionado la construcción de bocas de expendio en los últimos años (ver tabla).

Año	Cantidad de estaciones
1984	1
1985	3
1986	8
1987	26
1988	49
1989	70
1990	93
1991	165
1992	341
1993	423
1994	463
1995	485
1996	528
1997	580

- Encuentre la mejor función que represente el crecimiento del número de bocas de expendio en función del tiempo.
- ¿Qué representan cada uno de los parámetros de la función de ajuste elegida?
- ¿Puede predecir cuántas estaciones habrán sido instaladas en el año 2005? ¿Cómo haría esa predicción? ¿Es ese dato absolutamente confiable? ¿Por qué?

II) El organismo que regula el uso de Gas Natural Comprimido (GNC) estudia la cantidad de usuarios que están regularizados, es decir, tienen la oblea habilitante. Suponga que una empresa distribuidora de GNC está interesada en conocer el número de usuarios de GNC para analizar consumos e inversiones en el sector, a partir de los datos suministrados por el organismo.

Año	**Cantidad de obleas vendidas**
1996	305655
1997	352095
1998	436155
1999	494099
2000	559335
2001	661125
2002	824070

- Encuentre la mejor función que represente el crecimiento del número de usuarios en función del tiempo
- ¿Qué representan cada uno de los parámetros de la función de ajuste elegida?
- ¿Puede predecir cuántos usuarios habrá en el año 2003? ¿Cómo haría esa predicción? Analice la confiabilidad de la predicción.

III) Un portal de Internet desea vender espacios de publicidad en sus páginas en España. Por ello hace un estudio de la cantidad de computadoras conectadas a Internet en ese país. Los datos obtenidos se vuelcan en la siguiente tabla:

Año	**Número de computadoras conectadas a Internet (España)**
1981	213
1982	235
1983	562
1984	1024
1985	2308
1986	5089
1987	28174
1988	80000
1989	159000
1990	376000
1991	727000
1992	1313000
1993	2217000
1994	5846000
1995	14352000
1996	21819000
1997	29670000

- Encuentre la mejor función que represente el número de computadoras conectadas a Internet en función del tiempo.
- ¿Qué representan cada uno de los parámetros de la función de ajuste elegida?
- ¿Puede predecir cuántas computadoras habrán sido conectadas hacia el año 2005? ¿Cómo haría esa predicción? Analice la validez de su predicción.

IV) Un grupo de investigadores analiza la evolución de la población mundial de los últimos cincuenta años para evaluar planes de acción para distintos organismos, tanto económicos (FMI, Banco Mundial y OMC, entre otros) como de desarrollo social y humano (UNESCO, UNICEF, OIT y OMS, entre otros). A partir de los datos suministrados por un organismo oficial se confeccionó la siguiente tabla:

Año	**Población x 10^9**		**Año**	**Población x 10^9**
1950	2.56		1973	3.94
1951	2.59		1974	4.01
1952	2.64		1975	4.09
1953	2.68		1976	4.16
1954	2.73		1977	4.23
1955	2.78		1978	4.31
1956	2.83		1979	4.38
1957	2.89		1980	4.46
1958	2.94		1981	4.53
1959	3.00		1982	4.61
1960	3.04		1983	4.69
1961	3.08		1984	4.77
1962	3.14		1985	4.85
1963	3.20		1986	4.94
1964	3.28		1987	5.02
1965	3.34		1988	5.11
1966	3.41		1989	5.19
1967	3.48		1990	5.28
1968	3.56		1991	5.37
1969	3.63		1992	5.45
1970	3.71		1993	5.53
1971	3.78		1994	5.61
1972	3.86		1995	5.69

- Encuentre la mejor función que represente el número de habitantes como función del tiempo.
- ¿Qué representan cada uno de los parámetros de la función de ajuste elegida?
- ¿Puede predecir cuántos habitantes habrá en el año 2005? ¿Cómo haría esa predicción? ¿Es ese dato absolutamente confiable? ¿Por qué?

V) Suponga que trabaja para una empresa y debe hacer un informe a la casa matriz de las ganancias de dicha empresa desde su instalación en el país. A partir de los datos suministrados por la gerencia contable, se confeccionó la siguiente tabla:

Mes de actividad	**Ganancias (en miles de pesos)**
1	159
4	252
7	287
10	329
13	352
16	381
19	395
22	434
25	434
28	456
31	475
34	495
37	524
40	512
43	529
46	534
49	538
52	536
55	588
58	599

- Encuentre la mejor función que represente los ingresos de la empresa en función del tiempo.
- ¿Qué representan cada uno de los parámetros de la función de ajuste elegida?
- ¿Puede predecir cuáles serán los ingresos de la empresa al cabo de 84 meses de su instalación? ¿Cómo haría esa predicción? ¿Es confiable ese dato? ¿Por qué?

VI) Suponga que el gerente de una PyME debe hacer un informe al directorio de las ganancias de dicha empresa desde 1980. Para ello recurre a los datos suministrados por la gerencia contable, a partir de los cuales confecciona la siguiente tabla:

Año	**Ganancias (en miles de pesos)**
1980	190
1981	197
1982	209
1983	239
1984	250
1985	311
1986	317
1987	363
1988	382
1989	441
1990	514
1991	502
1992	557
1993	635
1994	715
1995	811
1996	894
1997	941
1998	1149
1999	1121
2000	1241

- Encuentre la mejor función que represente los ingresos de la empresa en función del tiempo.
- ¿Qué representan cada uno de los parámetros de la función de ajuste elegida?
- ¿Puede predecir cuales serán los ingresos de la empresa en el año 2002? ¿Cómo haría esa predicción? ¿Es ese dato absolutamente confiable? ¿Por qué?

VII) Suponga que el gerente de una PyME debe hacer un informe al directorio de las ganancias de dicha empresa desde el año 1962. Para ello recurre a los datos suministrados por la gerencia contable, a partir de los cuales confecciona la siguiente tabla:

Año	**Ganancias (en miles de dólares)**
1962	1048
1964	870
1966	767
1968	772
1970	705
1972	708
1974	664
1976	647
1978	659
1980	607
1982	598
1984	618
1986	598
1988	590
1990	597
1992	591
1994	593
1996	556
1998	570
2000	579

- Encuentre la mejor función que represente las ganancias de la empresa en función del tiempo.
- ¿Qué representan cada uno de los parámetros de la función de ajuste elegida?
- ¿Puede predecir cuales serán los ingresos de la empresa en el año 2010? ¿Cómo haría esa predicción? ¿Es ese dato absolutamente confiable? ¿Por qué?

VIII) Los médicos pediatras utilizan tablas de crecimiento promedio de los niños como función de la edad para evaluar si sus pacientes evolucionan de acuerdo a lo esperado. Entidades pediátricas se encargan de confeccionar dichas tablas, distinguiendo entre el crecimiento de niños y niñas. A continuación se muestra una tabla para la altura media de los niños entre el mes de vida y los seis años:

Edad (meses)	**Estatura media para el varón (cm)**
1	54
2	57.09
3	60.4
4	62.25
5	65
6	66.74
7	68.01
8	69.6
9	71.11
10	72.3
11	73.65
12	75.01
15	78.2
18	81.3
21	84
24	86.7
30	91.1
36	95.2
42	95.2
48	102.5
54	105.7
60	108.7
66	111.8
72	114.1

- Encuentre la mejor función que represente la altura de los niños en función del tiempo.
- ¿Qué representan cada uno de los parámetros de la función de ajuste elegida?
- ¿Puede predecir qué altura tendrá un niño promedio a la edad de 10 años? ¿Cómo haría esa predicción? Analice el grado de validez de su predicción.

IX) Los médicos pediatras utilizan tablas de crecimiento medio de los niños como función de la edad para evaluar si sus pacientes evolucionan de acuerdo a lo esperado. Entidades pediátricas se encargan de confeccionar dichas tablas, distinguiendo entre el crecimiento de niños y niñas. A continuación se muestra una tabla para la altura media de las niñas entre el mes de vida y los seis años:

Edad (meses)	**Estatura media para las niñas (cm)**
1	53.1
2	56.5
3	58.9
4	62
5	63.9
6	65.3
7	67
8	68.1
9	69.4
10	71
11	72.1
12	73.3
15	77
18	79.8
21	83
24	85.4
30	89.7
36	94.1
42	97.8
48	101.5
54	105
60	108.9
66	111
72	114

- Encuentre la mejor función que represente la altura de las niñas en función del tiempo.
- ¿Qué representan cada uno de los parámetros de la función de ajuste elegida?
- ¿Puede predecir qué altura tendrá una niña promedio a la edad de 10 años? ¿Cómo haría esa predicción? Analice el grado de validez de la predicción.

X) Los médicos pediatras utilizan tablas de la ganancia de peso medio de los niños como función de la edad para evaluar si sus pacientes evolucionan de acuerdo a lo esperado. Entidades pediátricas se encargan de confeccionar dichas tablas, distinguiendo entre la ganancia de los niños y niñas. A continuación se muestra una tabla para el peso medio de los niños entre el mes de vida y los seis años:

Edad (meses)	**Peso medio para el varón (kg)**
1	4.4
2	5.38
3	6.2
4	6.88
5	7.6
6	7.99
7	8.45
8	8.83
9	9.24
10	9.58
11	9.78
12	10.15
15	10.9
18	11.5
21	12.6
24	12.7
30	13.8
36	14.8
42	15.8
48	17
54	17.9
60	18.7
66	19.9
72	20.9

- Encuentre la mejor función que represente el peso de los varones en función del tiempo.
- ¿Qué representan cada uno de los parámetros de la función de ajuste elegida?
- ¿Puede predecir cuánto pesará un niño promedio a la edad de 8 años? ¿Cómo haría esa predicción? Analice el grado de validez de la predicción.

XI) Los médicos pediatras utilizan tablas de la ganancia de peso medio de los niños como función de la edad para evaluar si sus pacientes evolucionan de acuerdo a lo esperado. Entidades pediátricas se encargan de confeccionar dichas tablas, distinguiendo entre la ganancia de las niñas y los niños. A continuación se muestra una tabla para el peso medio de las niñas entre el mes de vida y los seis años:

Edad (meses)	**Peso medio para las niñas (kg)**
1	4.3
2	5
3	5.7
4	6.2
5	7
6	7.4
7	8
8	8.2
9	8.6
10	8.9
11	9.1
12	9.6
15	10.2
18	10.9
21	11.3
24	12.1
30	13.3
36	14.5
42	15.5
48	16.7
54	17.7
60	18.4
66	19.6
72	20.7

- Encuentre la mejor función que represente el peso de las niñas en función del tiempo.
- ¿Qué representan cada uno de los parámetros de la función de ajuste elegida?
- ¿Puede predecir cuánto pesará una niña promedio a la edad de 9 años? ¿Cómo haría esa predicción? Analice el grado de validez de la predicción.