Universidad Nacional del Litoral
Facultad de Ingeniería y Ciencias Hídricas

ESTADÍSTICA

Ingeniería Informática

TEORÍA

Mg.Ing. Susana Vanlesberg
Profesor Titular

UNIDAD 7
REGRESIÓN Y CORRELACIÓN

Muchos problemas en la ingeniería y la ciencia implican la exploración de las relaciones entre dos o más variables.

Se va a considerar la diferencia entre relación funcional y relación estadística:

Relación funcional entre dos o más variables se expresa por una fórmula matemática, por ejemplo en el cálculo de la velocidad de caída de un cuerpo, $v = (2gh)^{1/2}$, conociendo la altura de caída h y la gravedad del lugar, es posible obtener un valor exacto de velocidad; es por lo tanto una relación **determinística**.

Relación estadística, si se ajusta una curva a observaciones, existe variación de los puntos en torno de la curva que relaciona a las variables. Generalmente es posible encontrar una relación media con cierto grado de precisión.

El estudio de la asociación entre variables se hace a través de dos aspectos:

Análisis de regresión: es el que permite encontrar el modelo que vincula a las variables en cuestión, brindando así un mecanismo de pronóstico.

Análisis de correlación: determina la medida del grado de exactitud de la relación entre variables.

Se analizará el caso de asociación lineal simple entre dos variables, pero pueden darse casos más complejos como por ejemplo relación no lineal, relación múltiple, etc.

El modelo de regresión es una manera de expresar formalmente los aspectos esenciales de la relación estadística entre las variables:

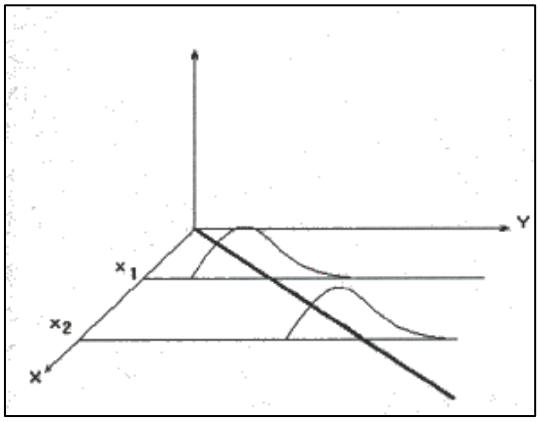
- La tendencia de la variable **y** (dependiente) a variar con la variable independiente de una manera sistemática.

- La dispersión de los puntos entorno a la curva que relaciona las variables.

Estos aspectos son contenidos en el modelo de regresión por enunciar lo siguiente:

- Existe una distribución de probabilidades de Y para cada valor de X: las variables X son fijas, es decir no aleatorias, mientras que las Y si lo son. Existen grupos de valores de Y para cada valor de la variable fija X, que se denominan **subpoblaciones**. Dependiendo del tipo de distribución de las variables aleatorias Y se clasifican en: -población tipo I es el caso en que la distribución de Y en cada subpoblación no está especificada; en el caso en que la distribución de Y en cada subpoblación es Normal se denomina población tipo II.

Las medias de estas distribuciones de probabilidad varían con la variación de x (variable fija).



Modelo de regresión bivariado

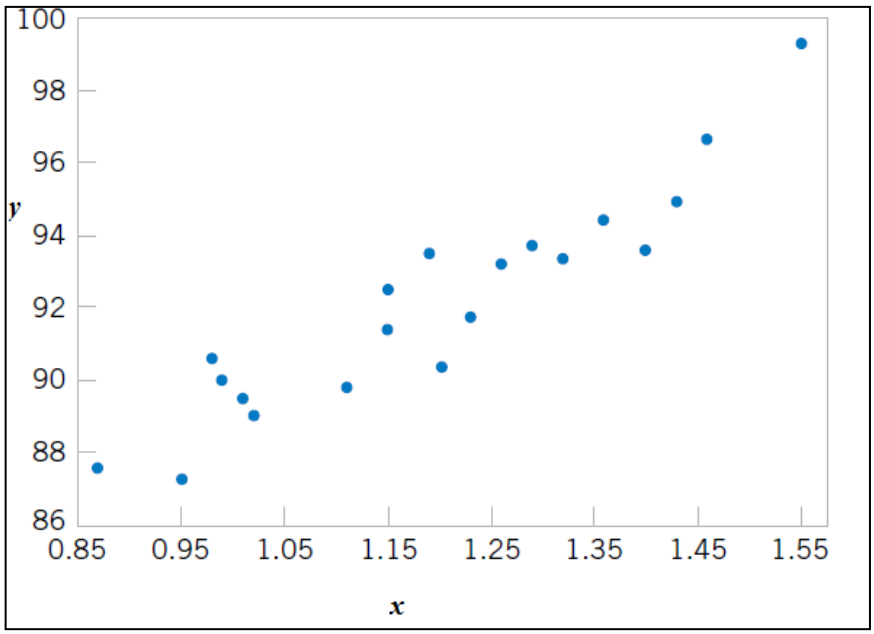


Figura N°1- Dispersiograma

La inspección del diagrama de dispersión indica que no hay una curva que pase exactamente por todos los puntos, pero es un fuerte indicio de que los puntos se encuentran dispersos al azar en torno a una línea recta. Por lo tanto, es razonable asumir que la media de la variable aleatoria Y está relacionada con x por la siguiente relación lineal:

$$E(Y/x) = \mu_{Y/x} = \alpha + \beta X_i \qquad (1)$$

α y **β** coeficientes de regresión.

Mientras que la media de Y es una función lineal de x, el valor observado y real no cae exactamente sobre una recta. La forma más adecuada para generalizar a un modelo lineal probabilístico es asumir que el valor esperado de Y es una función lineal de x, pero que para un valor fijo de x el real valor de Y está determinada por el valor medio de la función (el modelo lineal) más un término de error aleatorio, por ejemplo:

$$Y_i = \underbrace{\alpha + \beta X_i}_{I} + \underbrace{\varepsilon_i}_{II} \qquad (2)$$

donde ε_i es el término de error aleatorio.

siendo I la parte sistemática y II la parte estocástica, que hace que Y no pueda ser pronosticado exactamente como sucedería en un caso determinístico.

α y β son los parámetros del modelo.
Xi es la variable independiente, fija, conocida, variable explicativa.

Vamos a llamar a este modelo, **modelo de regresión lineal simple**, ya que cuenta con una sola variable independiente.

A veces, un modelo como este surgirá de una relación teórica. En otras ocasiones no vamos a tener ningún conocimiento teórico de la relación entre x e y, y la elección del modelo se basa en la inspección del diagrama de dispersión.

Para conocer más sobre este modelo, suponemos que podemos fijar el valor de x y observar el valor de la variable aleatoria Y. Ahora bien, si x es fijo, el componente aleatorio del lado derecho del modelo de la Ecuación (2) determina las propiedades de la variable dependiente Y.

Supongamos que la media y la varianza de ε son 0 y σ^2, respectivamente. Entonces:

$$E(Y/x) = E(\alpha + \beta x + \varepsilon) = \alpha + \beta x + E(\varepsilon) = \alpha + \beta x$$

Tener en cuenta que esta es la misma relación que en un principio escribimos empíricamente a partir de la inspección del diagrama de dispersión en la Figura Nº 1.

La varianza de Y dado x es:

$$Var(Y/x) = Var(\alpha + \beta x + \varepsilon) = Var(\alpha + \beta x) + Var(\varepsilon) = 0 + \sigma^2 = \sigma^2$$

Por lo tanto, el modelo de regresión es una línea de valores medios, es decir, el valor de la línea de regresión en cualquier valor de x es sólo el valor esperado de Y para que x.

La pendiente, puede ser interpretada como el cambio en la media de Y para un cambio unitario en x. Por otra parte, la variabilidad de Y para un valor particular de x se determina por la varianza del error $\in$, σ^2. Esto implica que hay una distribución de los valores Y para cada x, y que la varianza de esta distribución es la misma en cada x.

Significado de los parámetros α y β

α: intercepción de la línea de regresión con en eje Y.

β: pendiente de la recta, proporción de cambio en la media de la distribución de probabilidades de Y por unidad de incremento de X.

Es de destacar que el sentido con que son utilizados los términos **dependiente** e **independiente**, no es el mismo que el de dependencia e independencia de variables aleatorias.

Se dice que el modelo es de regresión simple cuando hay dos variables asociadas. Si esto no ocurre, el modelo es de regresión múltiple.

Se dice que el modelo es lineal si los parámetros y la variable independiente están elevados a la primera potencia. Si no es así, el modelo será no lineal.

¿Cómo puede asegurarse que el término de error $\in_i$ sea normalmente distribuido? Invocando el Teorema del Límite Central. Cuando existe una gran cantidad de causas independientes contribuyendo cada una con un pequeño efecto, la distribución de su suma es normal. En muchos casos en los que se aplica el análisis de regresión, las variables están influenciadas por un gran número de pequeños efectos independientes, por esto puede invocarse este teorema y justificar así la normalidad del término de error.

El modelo asume que la distribución de Y tiene igual varianza que el término de error, independientemente del valor de X. Esta propiedad se denomina *homoscedasticidad*.

Se asume independencia entre los términos de error. Esto significa que el resultado en alguna prueba no tiene efecto sobre el término de error de alguna otra prueba. ε no correlacionado con ε implica que Y_i no está correlacionado con Yj.

La completa especificación del modelo de regresión no solo incluye la forma del modelo (ecuación de regresión), sino una expresión de cómo son determinados los valores de la variable independiente y una especificación de la distribución de ε.

Cambiando los supuestos referidos a $\in$ y a X se obtienen distintos modelos de regresión. Por ejemplo, decir:

a) **ε** es una variable aleatoria independiente.

b) **ε** es una variable aleatoria, pero no independiente.

c) La distribución de **ε** no está especificada.

d) La distribución de **ε** es normal.

e) X es un conjunto de números fijos.

f) X es una variable aleatoria, pero su distribución no está especificada.
g) X es una variable aleatoria con distribución normal.

El modelo que se desarrollará considerará los supuestos a), d) y e) y, por lo tanto, se denominará modelo de regresión que tiene una población de Tipo II.

El modelo correspondiente a población Tipo I se basa en los supuestos a), c) y e)

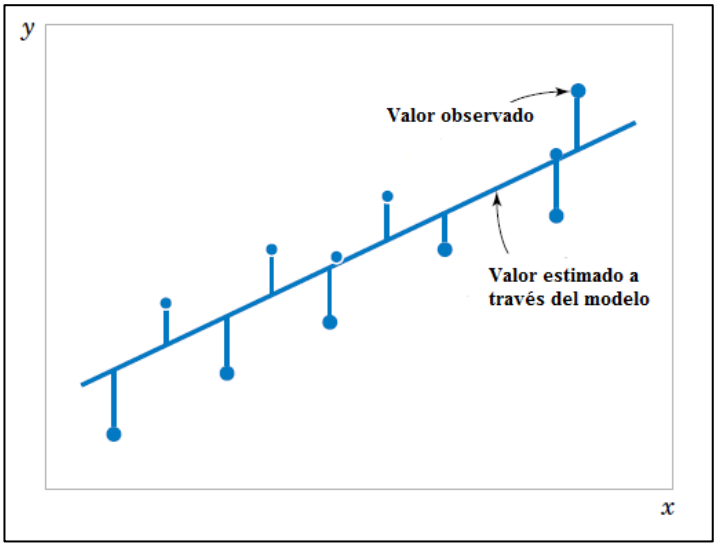
El modelo correspondiente a población Tipo III se basa en les supuestos a), c) y f).

El modelo correspondiente a población Tipo IV se basa en los supuestos a), d) y g).

La distribución de la variable Y, al ser una función lineal de $\in$, presenta su misma distribución.

Se trata de obtener el mejor estimador insesgado lineal del modelo planteado. El método empleado para esto es el de mínimos cuadrados. Una de las razones de su uso es la sencillez de su tratamiento matemático y, además, las estimaciones de α **y** β que produce son idénticas a las obtenidas por el método de máxima verosimilitud.

Estimación por el método de mínimos cuadrados



Se parte de considerar que la subpoblación de Y es normal, y que la suma de los cuadrados de las desviaciones de las observaciones respecto de la recta verdadera sea mínima:

$$S = \sum_{i=1}^{n} \left[Y_i - (\alpha + \beta X_i) \right]^2$$

$$S = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i \right)^2, con\_\hat{Y}_i = a + bX_i$$

$$luego \quad S = \sum_{i=1}^{n} \left[Y_i - (a + bX_i) \right]^2$$

Los estimadores de α **y** β serán aquellos que minimicen el valor de S:

$$\frac{\partial S}{\partial \alpha} = 0 \quad \text{y} \quad \frac{\partial S}{\partial \beta} = 0$$

$$\frac{\partial S}{\partial \alpha} = -2 \sum_{i=1}^{n} (Y_i - \alpha - \beta X_i)$$

$$\frac{\partial S}{\partial \beta} = -2 \sum_{i=1}^{n} X_i (Y_i - \alpha - \beta X_i)$$

Igualando a cero, aplicando sumatoria a todos los términos y reemplazando a α **y** β por a y b, se obtienen las siguientes ecuaciones normales, que conducen a obtener los estimadores a y b:

$$\begin{cases} \sum_{i=1}^{n} Y_i - na - b \sum_{i=1}^{n} X_i = 0 \\ \sum_{i=i}^{n} X_i Y_i - a \sum_{i=1}^{n} X_i - b \sum_{i=1}^{n} X_i^{\,2} = 0 \end{cases}$$

De la primera ecuación: $a = \overline{Y} - b\overline{X}$

Sustituyendo a en la segunda ecuación, se obtiene:

$$b = \frac{\sum_{i=1}^{n} X_i Y_i - \frac{\sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{n}}{\sum_{i=1}^{n} X_i^2 - \left(\frac{\sum_{i=1}^{n} X_i}{n} \right)^2} = \frac{\sum_{i=1}^{n} \left(X_i - \overline{X} \right)\left(Y_i - \overline{Y} \right)}{\sum_{i=1}^{n} \left(X_i - \overline{X} \right)^2}$$

$$b = \frac{\text{COV}}{S_x^2}$$

Luego la recta de regresión es $\hat{Y} = \hat{\alpha} + \hat{\beta}\, x$ (3)

Propiedades de los estimadores

Pueden considerarse a **a** y **b** como combinación lineal de las Y, que tienen distribución normal:

$$\hat{\beta} = b = \frac{\sum_{i=1}^{n} \left(X_i - \overline{X} \right)\left(Y_i - \overline{Y} \right)}{\sum_{i=1}^{n} \left(X_i - \overline{X} \right)^2}$$

Las propiedades estadísticas de los estimadores de mínimos cuadrados $\hat{\alpha}\,;\hat{\beta}$ se pueden describir fácilmente.

Recordemos que hemos supuesto que el término de error $\in$ en el modelo Y es una variable aleatoria con media cero y varianza σ^2. Dado que los valores de x son fijos, Y es una variable aleatoria con media $\mu_{Y/x} = \alpha + \beta x$ y varianza σ^2. Por lo tanto, los valores de $\hat{\alpha}$ y $\hat{\beta}$ dependen de los valores observados, por lo que los estimadores de mínimos cuadrados de los coeficientes de regresión pueden ser considerados como variables aleatorias. Vamos a investigar las propiedades de los estimadores.
Debido a que β es una combinación lineal de las observaciones Yi, podemos utilizar las propiedades de la esperanza (Ver Anexo Cap. VII) para demostrar que el valor esperado de β es:

$$E(b) = \beta \qquad (4)$$

Por lo tanto **b** es un estimador insesgado de β, y su distribución es Normal ya que se consideró a b como combinación lineal de variables normales independientes.

Para obtener las varianza el análisis se basa en el hecho de que las Y_i son variables independientes, cada una con varianza σ^2, y que las K_i utilizadas antes son constantes, pues dependen de X_i y $\overline{X}$:

$$\sigma^2(b) = \sigma^2\left(\sum_{i=1}^{n} K_i Y_i\right) = \sum_{i=1}^{n} K_i{}^2 \sigma^2(Y_i)$$

$$como$$

$$\sigma^2(Y_i) = \sigma^2(\in) = \sigma^2, entonces$$

$$\sigma^2(b) = \sum_{i=1}^{n} K_i{}^2 \sigma^2$$

$$\sigma^2(b) = \sigma^2 \frac{1}{\sum_{i=1}^{n}(X_i - \overline{X})^2} \qquad (5)$$

Se realizará algo similar para obtener la esperanza del estimador de α:

$$E(a) = E(\overline{Y} - b\overline{X}) = E(\overline{Y}) - E(b\overline{X}) = \alpha + \beta\overline{X} - \beta\overline{X} = \alpha$$

$$E(a) = \alpha \qquad (6)$$

Se comprueba así la insesgabilidad del estimador **a**, siendo su distribución Normal por ser también combinación lineal de variables normales independientes.

Se demuestra (Ver Anexo Cap VII) que la Varianza del estimador a es:

$$\sigma^2(a) = E(a - \alpha)^2$$

$$\sigma^2(a) = \sigma^2\left[\frac{1}{n} + \overline{X}^2 \frac{\sum_{i=1}^{n}(X_i - \overline{X})}{\left(\sum_{i=1}^{n}(X_i - \overline{X})^2\right)^2}\right]$$

$$\sigma^2(a) = \sigma^2 \left[\frac{1}{n} + \frac{\overline{X}^2}{\sum_{i=1}^{n} (X_i - \overline{X})^2} \right] \qquad (7)$$

Han sido así obtenidos los valores medios y varianzas de los estimadores puntuales de la pendiente y ordenada al origen del modelo de regresión lineal simple.

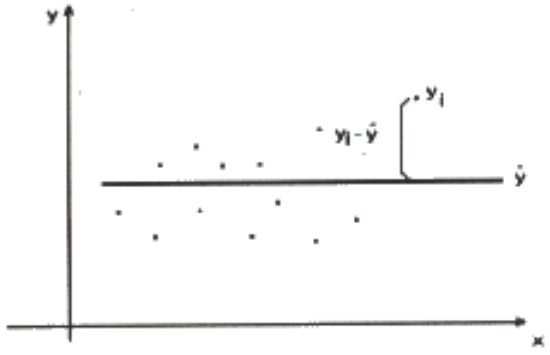
Varianza de la regresión

Se suele llamar ecuación predictiva a la ecuación de regresión muestral, ya que su principal objetivo es predecir valores medios de la variable dependiente asociados con un valor dado de la variable independiente. Pero para saber si es realmente conveniente utilizar esta ecuación para predicción, puede analizarse la variabilidad del valor pronosticado a través del modelo de regresión.

Una primera manera de analizar esta variabilidad puede ser a través de la inspección visual por trazar en el diagrama de puntos la recta obtenida. La medida numérica de la desviación de las observaciones respecto al modelo es el estimador de la varianza de la regresión de la población: $S^2_{y/x}$.

El análisis de la varianza de regresión se basa en la partición de la suma de cuadrados.

La variación de la variable dependiente Y_i generalmente se mide en términos de las desviaciones respecto al valor medio:

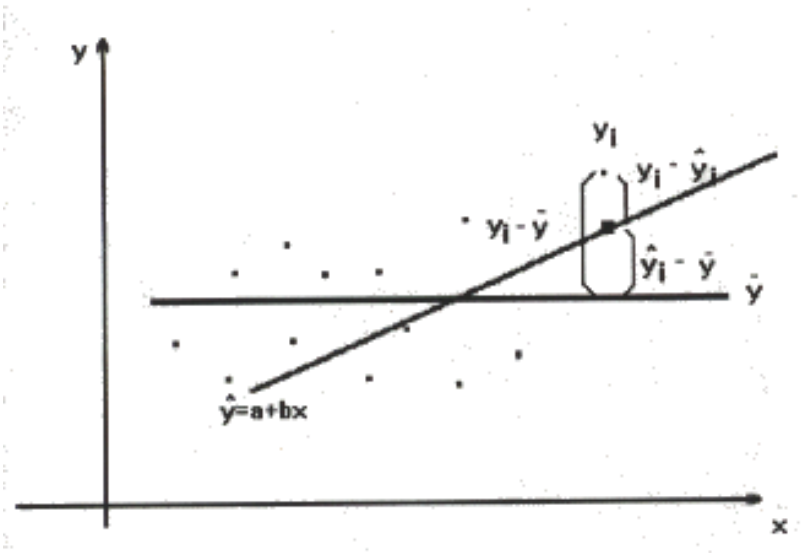


La medida de variación total es para todos los puntos:

$$\sum_{i=1}^{n}\left(Y_i-\bar{Y}\right)^2$$

Cuanto mayor es este valor, mayor es la variación de la curva ajustada respecto a las observaciones.

Utilizando el modelo ajustado, *la variación se da por la diferencia de los valores observados con los valores ajustados o estimados:*



$$Y_i-\hat{Y}_i$$

Por lo tanto la variación total será: **SSE** suma de desvíos cuadrados o suma de errores cuadrados:

$$\sum_{i=1}^{n}\left(Y_i-\hat{Y}\right)^2$$

Si este valor es igual a cero todos los puntos caen sobre el modelo ajustado; cuanto mayor sea, mayor será la variación o dispersión alrededor de la recta.

Particionando la suma total, o sea la dispersión respecto al valor medio, se obtiene:

$$\left(Y_i-\bar{Y}\right)=\left(\hat{Y}_i-\bar{Y}\right)+\left(Y_i-\hat{Y}_i\right)$$

siendo:

$Y_i-\bar{Y}$ *: desviación total*

$\hat{Y}_i - \overline{Y}$ *: desviación de la recta respecto al valor medio*

$Y_i - \hat{Y}_i$*: desviación respecto a la línea ajustada*

$$\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2 = \left(\sum_{i=1}^{n}\left(\hat{Y}_i - \overline{Y}\right)^2 + \sum_{i=1}^{n}\left(Y_i - \hat{Y}i\right)^2\right)$$

desarrollando el último término de la suma:

$$\left(Y_i - \hat{Y}_i\right) = \left(Y_i - \overline{Y}\right) - \left(\hat{Y}_i - \overline{Y}\right)$$

$$\left(Y_i - \hat{Y}_i\right) = \left(Y_i - \left(a + b\overline{X}\right)\right) - \left(a + bX_i - \overline{Y}\right)$$

siendo $a = \overline{Y} - b\overline{X}$, $\overline{Y} = a + b\overline{X}$, $\hat{Y}_i = a + bX_i$

$$\left(Y_i - \hat{Y}_i\right) = \left(Y_i - \overline{Y}\right) - \left(\left(a + bX_i\right) - \left(a + b\overline{X}\right)\right)$$

$$\left(Y_i - \hat{Y}_i\right) = \left(Y_i - \overline{Y}\right) - a - bX_i + a + b\overline{X}$$

$$\left(Y_i - \hat{Y}_i\right) = \left(Y_i - \overline{Y}\right) - b\left(X_i - \overline{X}\right)$$

Como se quiere encontrar una expresión para la varianza de la estimación, a la expresión anterior se le deberá aplicar el operador esperanza y elevarla al cuadrado, y esto para todos los puntos, es decir sumatoria. Luego de estos pasos matemáticos, se obtiene la expresión de la varianza de la regresión:

$$S^2_{y/x} = \frac{\sum (Y_i - \hat{Y}_i)^2}{n-2} \qquad (8)$$

La distribución en el muestreo de este estimador, recuérdese que la varianza estaba relacionada con la variable χ^2:

$$\chi^2_{n-2} = \frac{(n-2)S^2_{y/x}}{\sigma^2}$$

Esto sirve para realizar inferencias (construir intervalos de confianza o realizar test de hipótesis) respecto a los parámetros de la ecuación de regresión.

Predicción y pronóstico

Como ya se dijo, uno de los objetivos principales del análisis de regresión es la predicción. Pero es importante hacer notar claramente la diferencia entre predicción y pronóstico:

***Predicción*:** es la estimación del valor medio de Y dado un valor particular de X:

$$\hat{Y}_h = a + bX_h$$

a + bX es el estimador insesgado de α + βX, y su distribución es Normal, ya que es una combinación lineal de variables aleatorias normales. Por lo tanto, para poder encontrar un intervalo para cualquier punto de la recta de regresión poblacional, faltaría encontrar la varianza o error de la regresión. Para este caso, la variación depende de la variación en ambos estimadores, a y b:

$$\sigma^2\left(\hat{Y}_h\right) = \sigma^2\left(a + bX_h\right) = \sigma^2\left(\overline{Y} - b\overline{X} + bX_h\right) = \sigma^2\left(\overline{Y} + b(X_h - \overline{X}\right)$$

Como $\overline{Y}$ y b son variables independientes, y X_h y $\overline{X}$ son constantes, es posible hallar la varianza por términos:

$$\sigma^2\left(\hat{Y}_h\right) = \frac{\sigma^2}{n} + \sigma^2\left(b\left(X_h - \overline{X}\right)\right) =$$

$$= \frac{\sigma^2}{n} + \left(X_h - \overline{X}\right)^2 \sigma^2\left(b\right) =$$

$$= \frac{\sigma^2}{n} + \left(X_h - \overline{X}\right)^2 \sigma^2 \frac{1}{\sum_i \left(X_i - \overline{X}\right)^2} = \qquad (9)$$

$$\sigma^2\left(\hat{Y}_h\right) = \sigma^2\left(\frac{1}{n} + \frac{\left(X_h - \overline{X}\right)^2}{\sum_i \left(X_i - \overline{X}\right)^2}\right)$$

Siendo σ^2 la varianza de la regresión.

***Pronóstico*:** es la proyección de un solo valor de Y correspondiente a un valor de X particular:

$$\tilde{Y} = a + bX_h \qquad (10)$$

Se ve que las expresiones son las mismas, pero existen diferencias:

σ^2 *(error de pronóstico): consta de dos partes:*

1- σ^2 *del error de predicción (ya analizada)*

2- σ^2 *debida a errores casuales, tenida en cuenta por* σ^2

$$\sigma^2\left(Y_i - \hat{Y}_h\right) = \sigma_p^2 + \sigma^2 = \sigma^2\left(\frac{1}{n} + \frac{\left(X_h - \overline{X}\right)^2}{\sum_i \left(X_i - \overline{X}\right)^2}\right) + \sigma^2$$

$$\sigma^2\left(Y_i - \hat{Y}_h\right) = \sigma^2\left(1 + \frac{1}{n} + \frac{\left(X_h - \overline{X}\right)^2}{\sum_i \left(X_i - \overline{X}\right)^2}\right) \qquad (11)$$

Regresión no lineal

La regresión lineal no siempre da buenos resultados ya que veces la relación entre *Y* e *X* no es lineal. La estimación directa de los parámetros de funciones no-lineales es un proceso bastante complicado. No obstante, a veces se pueden aplicar las técnicas de regresión lineal por medio de transformaciones de las variables originales.

Por ejemplo los resultados de este análisis pueden proporcionar una buena indicación sobre el comportamiento de los costos para un banco "típico", aunque la naturaleza misma de un estudio de este tipo no puede arrojar resultados estrictamente aplicables a cada uno de los bancos considerados individualmente. No obstante, a pesar de esto, un estudio de este tipo de todas maneras puede ser muy útil, porque los resultados pueden proporcionar una "norma" o "estándar" contra el cual se pueden comparar los costos administrativos en un banco particular. En ausencia de un estudio de este tipo, un banco no tiene realmente un criterio para determinar si sus costos son "muy elevados," "aceptables," o "normales," ya que los bancos difieren enormemente en cuanto a cantidad de activos, número de sucursales, etc., de modo que el único criterio objetivo sería el de compararse con un banco de similar tamaño y características. Sin embargo, si se pudiera obtener una fórmula empírica que permita calcular un valor "normal" o "promedio" para los costos administrativos en función de unas pocas variables que permitan una medición numérica, entonces se podría fácilmente determinar si el banco en cuestión está "mejor" o "peor" que el banco "típico" a ese respecto.

Una función no-lineal que tiene muchas aplicaciones es la *función potencial*:

$$Y = A * X^b$$

donde *A* y *b* son constantes desconocidas. Si se aplica logaritmos, esta función tambiénpuede ser expresada como:

log(Y) = log(A) + *b* .log(X)

Considerando ahora la siguiente regresión lineal:

$\log(Y) = b_0 + b_1 \log(X)$

En esta regresión (denominada *regresión doble-log*), en lugar de calcular la regresión de Y en X, calculamos la regresión del *logaritmo* de *Y* vs el *logaritmo* de X.

Comparando estas dos ecuaciones, podemos apreciar que el coeficiente b_0 es un estimador de log(A), mientras que b_1 es un estimador de *b* (el exponente de la función potencial). Este modelo es particularmente interesante en aplicaciones econométricas, porque el exponente *b* mide la *elasticidad* de Y respecto de X.

CASO APLICADO

Desempleo y Crecimiento Económico

En 1963 el economista norteamericano Arthur M. Okun planteó un modelo macroeconómico para explicar la relación entre el crecimiento económico y las variaciones en la tasa de desempleo. Según este modelo, que se conoce hoy en día como la "ley de Okun," existe una relación lineal entre el *cambio* en la tasa de desempleo y la tasa de crecimiento del Producto Interno Bruto (PIB) real. En el cuadro adjunto se muestran datos anuales para la tasa de desempleo y el cambio porcentual en el PIB real en Alemania Occidental durante el período 1960-1981. Usar estos datos para estimar el modelo de Okun, y explicar el significado de los resultados obtenidos.

Año	Crecimiento PIB Real (%)	Desempleo (%)
1960	4.6	1.2
1961	5.1	0.9
1962	4.4	0.7
1963	3.1	0.9
1964	6.7	0.8
1965	5.5	0.7
1966	2.6	0.7
1967	-0.1	2.1
1968	5.9	1.5
1969	7.5	0.8
1970	5.1	0.7
1971	3.1	0.8
1972	4.2	1.1
1973	4.6	1.2
1974	0.5	2.6
1975	-1.7	4.8
1976	5.5	4.7
1977	3.1	4.6
1978	3.1	4.4
1979	4.2	3.8
1980	1.8	3.8

1981	0.1	5.5

Fuente: Frank Wolter, "From Economic Miracle to Stagnation: On the German Disease," en A. C. Harberger, ed., *World Economic Growth* (San Francisco: ICS Press, 1984), Table A-3, p. 119.

Medida del grado de asociación entre las variables

La medida del grado de relación entre dos variables se denomina **Coeficiente de Correlación ρ**.

Consideraciones a tener en cuenta en este análisis:

1 - Las variables X e Y son variables aleatorias, esto significa que no es fijo decir variable dependiente o independiente, cualquiera de las dos puede ser la variable independiente o a la inversa.

2 - Las variables proceden de una población Normal bivariada, o sea X e Y están distribuidas conjuntamente como Normal.

3 - X e Y tienen cada una distribución Normal:

$$X \approx N(\mu_x;\sigma_x) \qquad Y \approx N(\mu_y;\sigma_y)$$

4 - La relación entre X e Y es lineal; este supuesto implica decir que las medias de y para valores de X caen sobre la recta $Y_i = \alpha + \beta X_i$ de la misma manera que para $X_i = \alpha + \beta Y_i$.

5 - Si las dos rectas de regresión (con X dependiente o con Y dependiente) son iguales, quiere decir que la relación es perfecta.

El coeficiente de correlación poblacional se define como:

$$\rho = \frac{E(X-\mu_x)E(Y-\mu_y)}{\sqrt{E(X-\mu_x)^2 E(Y-\mu_y)^2}} = \frac{\text{cov}arianza}{\sigma_x \sigma_y} \qquad (12)$$

Siendo μ_x , μ_y ,σ_x , σ_y , ρ, los parámetros de la distribución Normal bidimensional.

De la expresión de ρ puede decirse:

- se aprecia que un cambio en el orden de las variables no afecta su valor, luego es un número adimensional.

- la covarianza y de aquí ρ, serán positivos si grandes valores medios de X se asocian con grandes valores medios de Y (y pequeños valores medios con pequeños valores medios). Por el contrario si grandes valores medios se asocian con pequeños valores medios (o viceversa), la covarianza y por lo tanto ρ, serán negativos. En ambos casos puede decirse que existe al menos alguna vinculación o dependencia estocástica entre X e Y.

Específicamente el **coeficiente de correlación** es una medida de la dependencia lineal entre dos variables aleatorias. Dado solamente el valor de ρ, puede decirse que un alto valor implica dependencia estocástica alta y de esta manera se puede decir que existe entre X e Y una tendencia lineal conjunta. Lo cual no significa necesariamente relación de causa y efecto, mientras que un bajo valor implica que las variables no tienen un comportamiento lineal conjunto y esto no asegura que falte dependencia estocástica. Es por esto, que debe tenerse **CUIDADO EN SU INTERPRETACIÓN**.

Correlación espuria: suele aparecer cuando se busca normalizar los datos, dividiendo por algún factor, el cual es en sí mismo una variable aleatoria, las variables originales pueden ser independientes pero los pares formados por los cocientes pueden presentar alta correlación, cuando en realidad no existe.

Valores posibles de ρ

$$Cov(x,y) = E(x - \mu_x)E(y - \mu_y)$$

$$x \text{ e } y \text{ son } N(\mu, \sigma)$$

$Cov(x^*,y^*) = E(x^*)E(y^*)$, *las esperanzas son* 0 *y los desvìos* 1. *Luego:*

$$Cov(x,y) = E(xy) - E(x)E(y) = E(x^*y^*)$$

$$\rho = \frac{Cov(xy)}{\sigma_x \sigma_y}. \text{ Siendo } \sigma_x \text{ y } \sigma_y \text{ iguales a } 1, \text{ luego:}$$

$$Var(x^* \pm y^*) = Var(x^*) + Var(y^*) \pm 2Cov \geq 0$$

$$1 + 1 \pm 2\rho \geq 0$$

$$2 \pm 2\rho \geq 0$$

$$-1 \leq \rho \leq 1$$

Coeficiente de correlación muestral

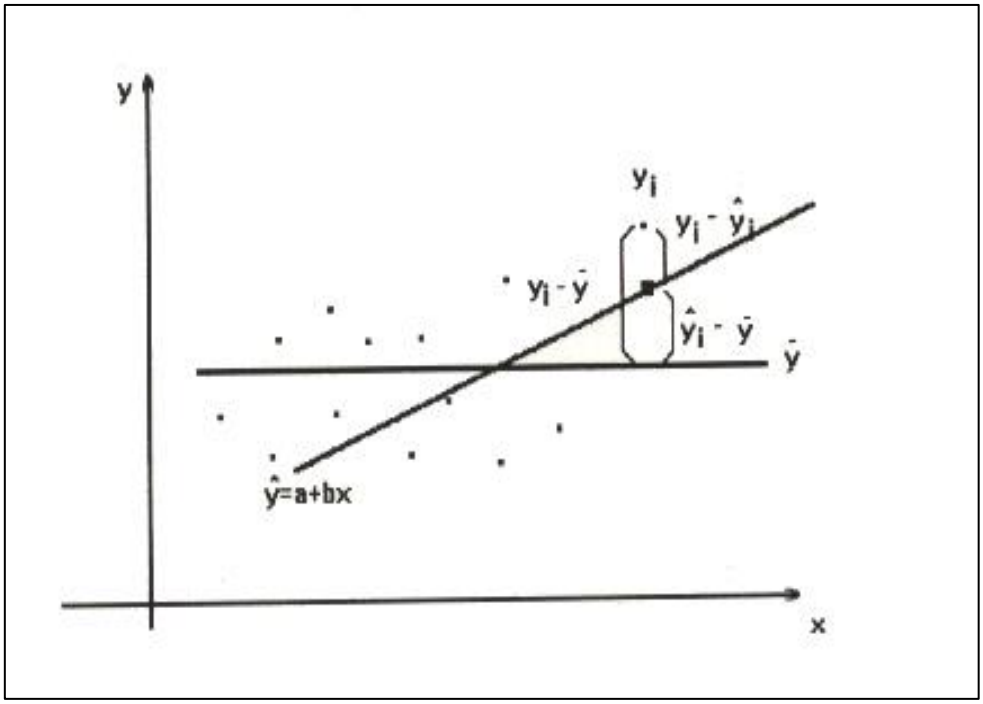
El estimador de ρ se obtiene considerando los momentos muestrales:

$$r = \hat{\rho} \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2} \sqrt{\sum_i (y_i - \bar{y})^2}} = \frac{m_{1,1}}{S_x S_y} \qquad (13)$$

Su variación es la misma que la del coeficiente de correlación poblacional.

Otro coeficiente usado en este análisis es el de determinación, que está relacionado con el de correlación.

Recordando el análisis de varianza ya realizado, se partirá de estas expresiones para obtener el **coeficiente de determinación**.



$$\sum_i (y_i - \bar{y})^2 = \sum_i (\hat{y}_i - \bar{y})^2 + \sum_i (y_i - \hat{y}_i)^2$$

$$\sum_i (y_i - \bar{y})^2 = \ (SCT)$$

$$\sum_i (\hat{y}_i - \bar{y})^2 = \ (SCR)$$

$$\sum_i (y_i - \hat{y}_i)^2 = \ (SCE)$$

$$SCR = \sum_i (\hat{y}_i - \bar{y})^2 = \sum_i (a + bx_i - \bar{y})^2 = \sum_i (\bar{y} - b\bar{x} + bx_i - \bar{y})^2 =$$

$$= \sum_i \left[b(x_i - \bar{x})\right]^2 = b^2 \sum_i (x_i - \bar{x})^2$$

$$SCT = \sum_i (y_i - \bar{y})^2$$

$$SCE = SCT - SCR$$

$$Dividiendo por\ SCT:$$

$$\frac{SCE}{SCT} = \frac{SCT}{SCT} - \frac{SCR}{SCT}$$

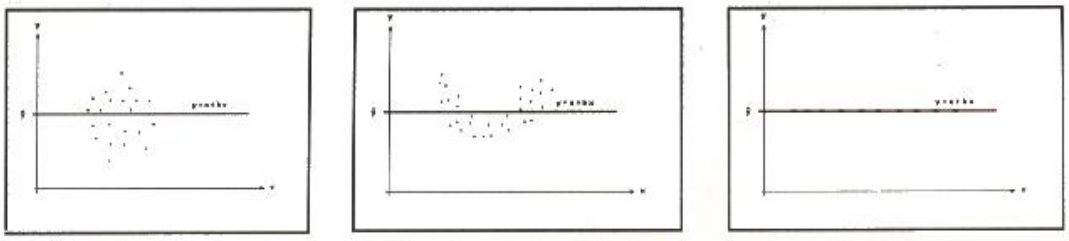
$$1 = \frac{SCR}{SCT} + \frac{SCE}{SCT}$$

$$r^2 = 1 - \frac{SCE}{SCT} = \frac{SCR}{SCT} \qquad (14)$$

$$r^2 = \frac{suma\ de\ cuadrados\ \text{explicados}}{suma\ de\ cuadrados\ total}$$

r^2 varía entre 0 y 1, ya que SCR es menor o igual que SCT.
Pueden hacerse algunos comentarios respecto de su valor:
Si SCE = 0, esto implica que SCR = SCT, lo que lleva a decir que r^2 es igual a 1. Esto significa que todos los puntos están sobre la recta estimada.
Si SCR = 0, implica que SCE = SCT, con lo cual $r^2 = 0$. Esto significa que la pendiente de la recta es igual a cero. Esto puede deberse a que la línea de regresión sea horizontal, y esto ser debido a distintas causas:

a) las observaciones se dispersan alrededor del valor medio en forma aleatoria.
b) las observaciones se dispersan alrededor de una curva tal que la línea mejor ajustada es una línea recta horizontal.
c) todas las observaciones tienen el mismo valor, cualquiera sea el valor de x.



Este coeficiente es también denominado **índice de correlación**, y se utiliza para medir el grado de asociación entre las variables cuando la regresión es lineal y no lineal.

Universidad Nacional del Litoral
Facultad de Ingeniería y Ciencias Hídricas

ESTADÍSTICA

Ingeniería Informática

TEORÍA

Mg.Ing. Susana Vanlesberg
Profesor Titular

UNIDAD 8
SERIES CRONOLÓGICAS

SERIES CRONOLÓGICAS

Son aquellas que están formadas por valores de una variable observada a <u>intervalos regulares de tiempo</u>.

Pueden ser

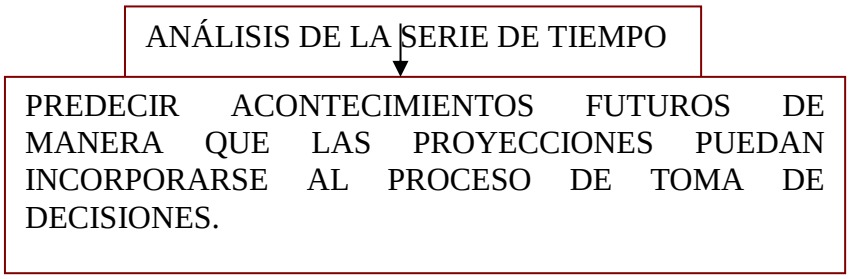
- ANUALES
- MENSUALES
- BIMESTRALES
- OTROS

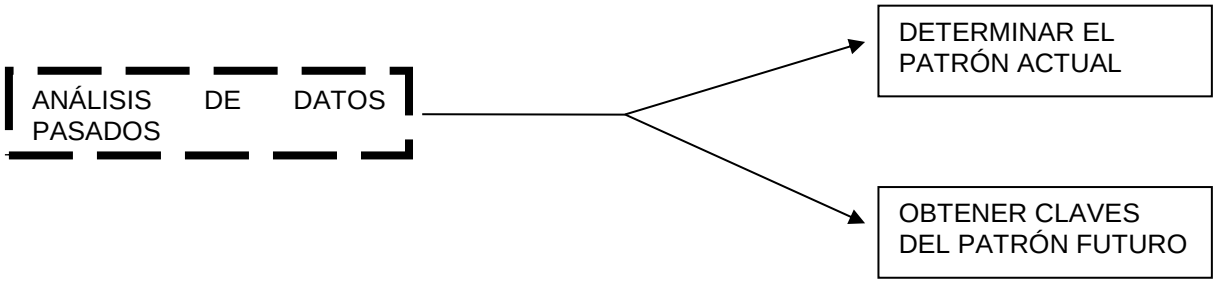
EJEMPLOS:

- Ventas semanales de un supermercado.
- Producción de una fábrica textil durante un mes.
- Informes anuales de un Municipio en cuanto a la recaudación de impuestos.
- Informes mensuales de un determinado Banco en cuanto al ingreso de depósitos.

EL ANÁLISIS DE SERIES DE TIEMPO

PERMITE → CONTROLAR LAS OPERACIONES PRESENTES Y LLEVAR ADELANTE EL PLANEAMIENTO FUTURO A TRAVÉS DE LA APLICACIÓN DE TÉCNICAS ESTADÍSTICAS





COMPONENTES DE UNA SERIE DE TIEMPO

El análisis de una serie de tiempo es un tema complicado; hay una variedad de opiniones en cuanto a como se tendrían que realizar los análisis.

Uno de los enfoques de mayor aceptación es considerar una serie de tiempo como una combinación de 4 elementos, los cuales superpuestos y actuando en forma conjunta contribuyen a los cambios que se observan en un período de tiempo. Estos elementos son:

a)- Tendencia a largo plazo

b)- Variación estacional

c)- Variación cíclica

d)- Variación aleatoria e irregular, impredecible

Estos componentes se aíslan y se ajustan utilizando algunos métodos y son:

a)- Tendencia a largo plazo: Es el movimiento de una serie de tiempo (creciente o decreciente) gradual en el tiempo de acuerdo a una curva .

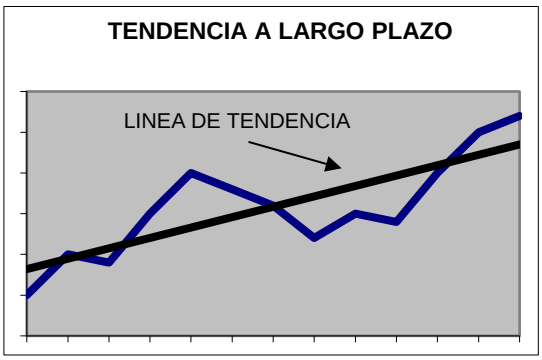
b)- Variación estacional: representa la tendencia de la serie de tiempo a variar hacia arriba y hacia abajo durante épocas especificas del año y más o menos con igual intensidad, pueden

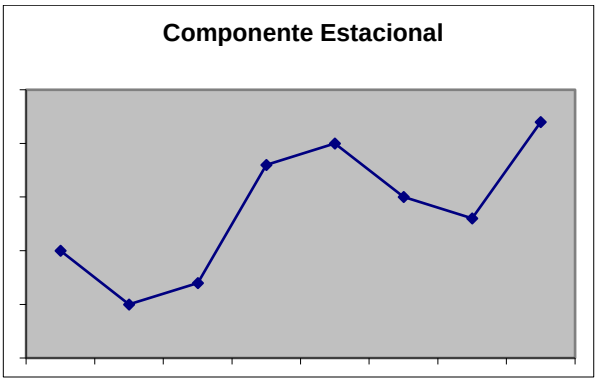
ser meses, bimestres o trimestres, además puede presentarse con movimientos periódicos por naturaleza. Es decir que esta componente se presenta cuando se trabaja con datos mensuales.

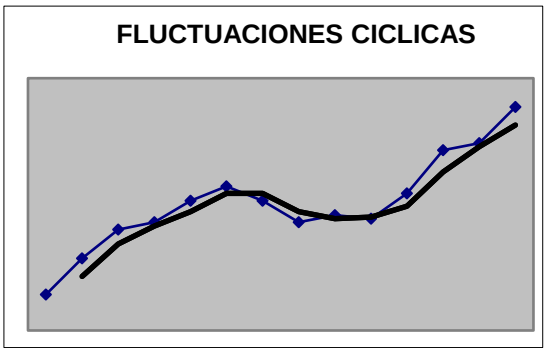
c)- Variación Cíclica: Los componentes cíclicos de una serie de tiempo tienden a subir o bajar según un patrón cíclico alrededor de la curva de la tendencia. Difieren de la componente anterior en que se extienden por períodos de tiempo más largos derivándose de causas diferentes.

d)- Variación aleatoria e irregular: Esta variación se añade a las anteriores. Se presentan de manera casual debido a efectos inexplicados como por ejemplo:

- Guerras
- Inundaciones
- Huelgas
- Etc.







Características de las Componentes

Componente	***Definición***	Razón	***Duración***
Tendencia	Patrón de movimientos ascendentes o descendentes, persistente a largo plazo.	Debido a cambios en la tecnología, riqueza, población.	Varios años.
Estacional	Fluctuaciones periódicas regulares que ocurren dentro de cada período de 12 meses, año a año.	Debido a condiciones de costumbres, tiempo, etc.	Dentro de los 12 meses.
Cíclica	Movimientos repetitivos ascendentes y descendentes mediante cuatro fases: -punto más alto: Prosperidad -contracción: Recesión -sima: Depresión -expansión: Recuperación.	Interactúan una serie de combinaciones de factores que influyen en la economía.	Generalmente de 2 a 10 años con diferente intensidad para cada ciclo completo.
Irregular	Fluctuaciones que existen en una serie luego de tomar en cuenta los efectos sistemáticos anteriores.	Se relacionan con acontecimientos imprevistos como huelgas, inundacio-nes, etc,.	Breves y no repetitivas.

Se supone además que estas cuatro componentes están ligadas por una relación, vamos a nombrar dos relaciones, si bien, hay que aclarar que no son los únicos esquemas de análisis.

Modelo o hipótesis *Aditiva*: supone que los cuatro componentes son independientes unos de otros:

$$Yi = Ti + Si + Ci + Ii$$

Modelo o hipótesis *multiplicativa:* supone que los cuatro componentes se deben a diferentes causas, y que se relacionan entre sí por un efecto multiplicador:

$$Yi = Ti * Si * Ci * Ii$$

El modelo clásico multiplicativo que se analizará considera que cualquier valor observado en una serie de tiempo es el producto de los factores componentes:

$$Y_i = T_i * S_i * C_i * I_i$$

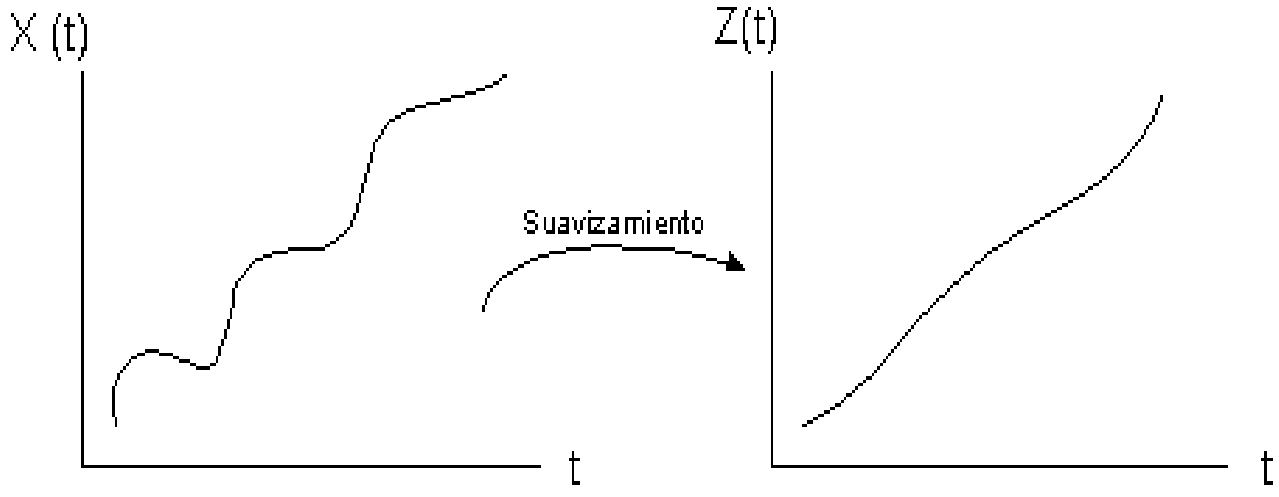
Donde: i es el año

T_i = Valor de la componente de la tendencia
S_i = Valor del componente estacional
C_i = Valor del componente cíclico
I_i = Valor del componente irregular o aleatorio

SUAVIZACIÓN DE SERIES DE TIEMPO

Muchas veces resulta difícil al examinar la información, decir si la **tendencia** es descendente o ascendente debido a que existen amplias fluctuaciones en sus componentes cíclicos e irregular. Entonces antes de tratar de modelar una serie de tiempo es útil graficarla para determinar la naturaleza de los componentes secular, cíclica y estacional si es que existen. Pueden utilizarse métodos para **suavizar o alisar la serie** y poder así distinguir los distintos movimientos libre de los efectos de la variación aleatoria.

La idea central es definir a partir de la serie observada una nueva serie en la que suavizan los efectos ajenos a la tendencia (estacionalidad, efectos aleatorios), de manera que se pueda determinar claramente la tendencia.



Lo que se hace es usar una expresión lineal que transforma la serie *X(t)* en una serie *suavizada Z(t): Z(t) = F(X(t)), t* = 1,...,n

de tal modo que F(*X(t)) = T(t).* La función F se denomina Filtro Lineal. El filtro lineal más usado es el ***promedio móvil***.

PROMEDIOS MÓVILES

Dado un conjunto de números

Y1, Y2, Y3 ...

se define un movimiento medio de orden N el que viene dado por la sucesión de medias aritméticas.

$$\frac{Y_1+Y_2+...+Y_N}{N}, \quad \frac{Y_2+Y_3+...+Y_N+_1}{N}, \quad \frac{Y_3+Y_4+...+Y_N+_1}{N}, \ ...$$

Las sumas de los numeradores se llaman movimientos totales de orden N.

Si los datos son anuales o mensuales, se llama movimiento medio de N años o un movimiento medio de N meses, respectivamente. Así, se habla de movimientos medios de cinco años, movimientos medios de doce meses, etc. Naturalmente que cualquier otra unidad de tiempo puede igualmente utilizarse.

Los movimientos medios tienen la propiedad de tender a reducir la cantidad de variación presente en un conjunto de datos. En el caso de series de tiempo, esta propiedad se utiliza a menudo para eliminar las fluctuaciones no deseadas y el proceso se llama **suavización de series de tiempo.**

Para aplicar el método del promedio móvil a una serie de tiempo, los datos deben tener una tendencia bastante lineal y un esquema de fluctuaciones rítmico definido (que se repite, por ejemplo cada tres años). Cuando no hay componente estacional o sea para datos anuales lo que se hace en el método de promedios móviles es en realidad promediar C e I. El residuo es la tendencia.

Si la duración de los ciclos es constante y si las amplitudes de tales ciclos son iguales, las fluctuaciones cíclica e irregular pueden eliminarse por completo usando el método del promedio móvil.

El problema principal en los promedios móviles es la elección apropiada de período para el promedio, esto depende de la naturaleza de los datos y el propósito que se persigue. Generalmente el **objetivo** de aplicar un promedio móvil es eliminar, tanto como sea posible, las variaciones indeseables de los datos, tratando de darle a la serie un aspecto más uniforme.

Si por ejemplo, a una serie temporal de 7 observaciones se le aplica este método, tomando medias aritméticas de tres observaciones (promedio móvil de orden tres), las operaciones a seguir son:

Ti	**Yi**	$\overline{Y}$
T1	Y1	
T2	Y2	$\overline{Y}2$
T3	Y3	$\overline{Y}3$
T4	Y4	$\overline{Y}4$
T5	Y5	$\overline{Y}5$
T6	Y6	$\overline{Y}6$
T7	Y7	

$$\text{donde } \overline{Y}2 = \frac{Y1 + Y2 + Y3}{3}$$

$$\text{donde } \overline{Y}3 = \frac{Y2 + Y3 + Y4}{3}$$

$$\text{donde } \overline{Y}4 = \frac{Y3 + Y4 + Y5}{3}$$

$$\text{donde } \overline{Y}5 = \frac{Y4 + Y5 + Y6}{3}$$

$$\text{donde } \overline{Y}6 = \frac{Y5 + Y6 + Y7}{3}$$

Como se observa cada media aritmética se obtiene de la anterior con solo eliminar el primer valor Yi y añadir el siguiente; de ahí su denominación de medias móviles. Si se tomase un número par de valores para obtener las medias móviles, la nueva serie de promedios quedaría descentrada en el sentido de que sus valores no corresponderían ya a los momentos originales de tiempo, sino a momentos intermedios; luego deberá calcularse una nueva serie de medias, promediando los valores medios obtenidos; o sea:

Ti	**Yi**	$\overline{Y}$	$\overline{\overline{Y}}$
T1	Y1	$\overline{Y}2$ →	$\overline{\overline{Y}}2$
T2	Y2	$\overline{Y}3$ →	$\overline{\overline{Y}}3$
T3	Y3	$\overline{Y}4$ →	$\overline{\overline{Y}}4$
T4	Y4	$\overline{Y}5$ →	$\overline{\overline{Y}}5$
T5	Y5	$\overline{Y}6$ →	$\overline{\overline{Y}}6$
T6	Y6	$\overline{Y}7$	
T7	Y7		

$$\text{donde } \overline{Y}2 = \frac{Y1 + Y2}{2}$$

$$\text{donde } \overline{Y}3 = \frac{Y2 + Y3}{2}$$

$$\text{donde } \overline{Y}4 = \frac{Y3 + Y4}{2}$$

$$\text{donde } \overline{Y}5 = \frac{Y4 + Y5}{2}$$

$$\text{donde } \overline{Y}6 = \frac{Y5 + Y6}{2}$$

$$\text{donde } \overline{Y}7 = \frac{Y6 + Y7}{2}$$

$$\text{donde } \overline{\overline{Y}}2 = \frac{\overline{Y}2 + \overline{Y}3}{2}$$

$$\text{donde } \overline{\overline{Y}}3 = \frac{\overline{Y}3 + \overline{Y}4}{2}$$

$$\text{donde } \overline{\overline{Y}}6 = \frac{\overline{Y}6 + \overline{Y}7}{2}$$

Los inconvenientes de este método son los siguientes:

1. Se pierden los períodos al comienzo y a la finalización de la serie.
2. Se pueden generar componentes que los datos originales no tenían
3. Estos promedios móviles están fuertemente afectados por los valores extremos

Sin embargo permite suavizar la serie original y es muy utilizado como base para el futuro análisis de la componente estacional.

PARA RESUMIR: la técnica de promedios móviles auxilia en la identificación de la tendencia a largo plazo en una serie de tiempo ya que amortigua las fluctuaciones a corto plazo. Sirve para revelar cualquiera de las fluctuaciones cíclicas y estacionales.

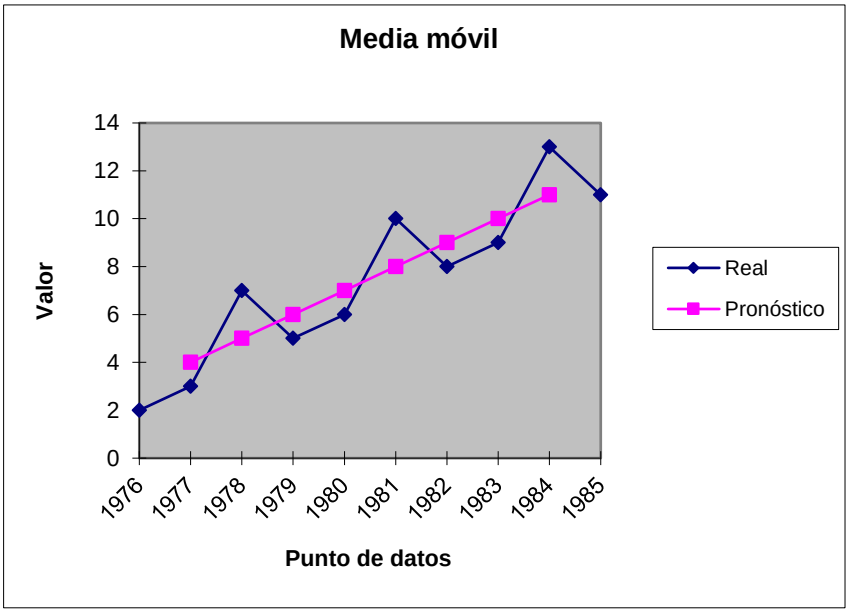
Ejemplo:

Sea la serie ventas de una determinada empresa comercial:

Año	**Ventas en miles**	**Prom. Móvil de Orden 3**	**Prom. Móvil de Orden 2 (1ª Media)**	**Prom. Móvil de Orden 2 (2ª Media)**
1976	**2**		**2,5**	
1977	**3**	**4**	**5**	**3,75**
1978	**7**	**5**	**6**	**5,50**
1979	**5**	**6**	**5,5**	**5,75**
1980	**6**	**7**	**8**	**6,75**
1981	**10**	**8**	**9**	**8,50**
1982	**8**	**9**	**8,50**	**8,75**

1983	**9**	**10**	**11**	**9,75**
1984	**13**	**11**	**12**	**11,50**
1985	**11**			

Promedio Móvil de orden 3



Como se puede ver en el gráfico anterior, los valores faltantes al comienzo y al final de la serie artificial son una característica de este tipo de suavizado: se pierde un valor en cada extremo de un promedio móvil de tres años, dos en un promedio de cinco, tres en uno de siete, etc. Esto puede o no tener consecuencias, pero puede ocasionar problemas cuando la serie es muy corta y se necesite realizar un pronóstico a largo plazo.

SUAVIZACIÓN EXPONENCIAL

Es otro método para suavizar los movimientos generales a largo plazo en la información. Puede ser utilizado también para obtener pronósticos a corto plazo para series de tiempo en las que resulta dudoso el tipo de efecto de tendencia a largo plazo.

La suavización exponencial es un tipo especial de promedio móvil, pero su naturaleza es muy diferente del cálculo del Promedio Móvil.

La *suavización exponencial*, una clase especial de *promedio móvil ponderado*, es útil en pronóstico a corto plazo proporciona una impresión de los movimientos globales a largo plazo de los datos.

Esta técnica posee una ventaja con respecto al promedio móvil porque proporciona un promedio móvil ***exponencialmente ponderado*** a través de la serie de tiempo, ya que cada valor suavizado depende de todos los valores anteriormente observados (en los promedios móviles no se toman en cuenta todos los valores observados).

Si se desea pronosticar el valor de una serie de tiempo para el período *t + 1* sobre la base de la información obtenida inmediatamente después del período t, el pronóstico se considera mejor como una función de dos componentes: el valor real de una serie para el período t, y el valor pronosticado para el mismo período hecho en el período anterior *t – 1*. El uso de valores observado y estimado disponibles ahora para predecir valores futuros es mejor que el uso de cualquiera de ellos sólo, porque el valor real en el periodo t podría haber sido indebidamente influido por factores aleatorios o porque puede que las condiciones que condujeron al pronóstico para el período t no se cumplan ya, o por todos estos factores juntos.

Un valor suavizado exponencialmente para una serie de tiempo está dado por:

$$S_i = W * Y_i + (1 - W) * S_i - 1$$

Donde: **S1 = Y1**

Siendo:

S_i: valor de la serie exponencialmente suavizada calculada en el período **i**
$S_i - 1$: valor de la serie exponencialmente suavizada ya calculada en el período **i – 1**
Y_i: valor observado de la serie de tiempo en el período **i**
W: ponderación o coeficiente de suavización asignado en forma subjetiva ($0 < W < 1$)

Como elegir W:

Si el valor de *W* es demasiado grande, se le dará un valor muy grande a los datos actuales y no se suavizan adecuadamente las variaciones irregulares; pero si *W* es demasiado chico, se le dará un valor muy pequeño a los datos actuales de la serie y el promedio móvil será insensible a las variaciones que se pueden dar en realidad. Esto podría considerarse como una desventaja en la utilización de este método, ya que el rango de valores que se le pueden asignar a W es muy grande, se debe tener mucho cuidado dependiendo de lo que se quiera reflejar con los resultados.

Si sólo se quiere suavizar una serie mediante la eliminación de variaciones cíclicas e irregulares que no se desean, debe seleccionarse un valor pequeño de W (cercano a 0); pero si se quieren hacer pronósticos se elegirá un valor grande de W (cercano a 1). En el primer caso las tendencias generales de la serie a largo plazo serán aparentes, y en el segundo caso tal vez se pronostiquen en forma adecuada las direcciones a corto plazo.

Importante:

Si se quieren utilizar los valores suavizados exponencialmente para realizar pronósticos, sólo se toma el valor suavizado en el período **i** como el valor pronosticado en el período **i + 1** por lo tanto:

$$Y_i + 1 = S_i$$

ANÁLISIS DE TENDENCIA

Tendencia

Es la componente de una serie que más se estudia, se puede hacer con fines de pronóstico a largo y mediano plazo.

Tendencia Lineal:

El método de mínimos cuadrados permite ajustar una línea recta de la forma:

$$Y = a + bx$$

a: Ordenada al origen, es el valor de Y cuando X=0
b: Pendiente de la Tendencia, indica el aumento o disminución medio en Y por cambio unitario en X.

Recordamos el análisis de regresión lineal que permite calcular los estimadores de los parámetros de la regresión.

Cuando utilizamos el método de mínimos cuadrados para ajustar tendencias en series de tiempo, nuestros esfuerzos de cálculo pueden minimizarse si se codifican apropiadamente los valores de X (con software estos pasos no se realizan)

Se debe distinguir además números de años par e impar:

- Si la serie tiene numero impar de años se elige como inicio de la sucesión al año que esta en el centro, asignándole valores crecientes a los posteriores 1,2,3,..... y enteros consecutivos decrecientes a los anteriores -1,-2,-3,.....
- Si la serie tiene un número par se toma como origen el primer año X = 0 y se le asigna números consecutivos crecientes a los años siguientes.

Una vez calculada la ecuación de la tendencia puede ser utilizada para calcular la tendencia en un año dado sustituyendo el valor X que corresponde a ese año en la ecuación obtenida. Esta ecuación puede utilizarse para pronóstico.

Tendencias no Lineales:

Si el modelo a ajustar a la tendencia no es una recta sino una curva, puede pensarse en un "*polinomio de segundo grado*"

$$\hat{y}_i = B_0 + B_1 X_i + B_2 X_i^2$$

B_0 = intersección con el eje Y
B_1 = efecto lineal estimado
B_2 = efecto curvilíneo estimado

Los coeficientes se estimarán a través del método de mínimos cuadrados resolviendo el sistema de ecuaciones correspondiente.

Los cálculos de los coeficientes B_0, B_1, B_2 pueden simplificarse usando la condición de codificación mencionada anteriormente.

Tendencia Potencial:

Cuando una serie parece estar incrementando con rapidez cada vez mayor tal que la diferencia porcentual de una observación a otra es constante se puede considerar una ecuación de tendencia potencial de la forma

$$Y_i = B_0 * B_1^{Xi}$$

En la que B_0 = intersección estimada con el eje Y

$(B_1 - 1)*100\%$ = Tasa de crecimiento compuesta estimada anual (en porcentaje)

Si se toma logaritmo (Base 10) de ambos lados de la ecuación se tiene:

$$\log \hat{y}i = \log B_0 + X_i \log B_1$$

Puesto que la ecuación tiene forma lineal, se puede utilizar el método de mínimos cuadrados, si se trabaja con el logaritmo de los valores de Y_i, en lugar de hacerlo con los valores de Y_i, y obtener la pendiente ($\log B_1$) y la intersección ($\log B_2$)

Aislamiento y eliminación de la tendencia en datos anuales y mensuales

Si los pronósticos que se desean elaborar son a *corto plazo*, debería ser eliminado el efecto del componente tendencia ya que esta es la componente a largo plazo.

Se estima la tendencia y para eliminarla es suficiente con dividir por ella en la expresión del modelo multiplicativo *sin considerar el efecto estacional*, se lo logra ya que los datos son anuales:

$$Y_i = T_i * C_i * I_i$$

$$\frac{Y_i}{\hat{y}_i} \text{ como } \hat{y}_i = T_i \quad \text{luego} \quad \frac{Y_i}{\hat{y}_i} = \frac{T_i * C_i * I_i}{T_i}$$

$$\frac{Y_i}{\hat{y}_i} = C_i * I_i$$

Este cociente se denomina Relativas cíclicas irregulares ya que muestran el desarrollo tanto cíclico como irregular de la serie de datos una vez eliminada la tendencia en los datos. Gráficamente estos valores fluctuarán alrededor -1 y 1.

Aislamiento y eliminación de la tendencia en datos mensuales

En la serie de datos mensuales existe la componente estacional, además de la tendencia, cíclica e irregular.

Al igual que para datos anuales, interesa pronosticar algunos movimientos futuros de los movimientos mensuales, sólo que, se agregarán los datos mensuales para llevarlos a anuales y luego de obtenida la ecuación de tendencia deberá tenerse en cuenta lo que significa cada parámetro de la regresión para su transformación.

a es la ordenada al origen, por lo tanto dará el valor de la variable serie de tiempo expresada en utilidades por año con lo cual será necesario dividir su valor por doce (12), para expresarlo en unidades por meses.

b es la pendiente que da el cambio medio de Y por cambio en X, entonces el valor b se dividirá por 144 (12^2) para expresarlo en unidades por mes.

VARIACIÓN CÍCLICA:

A la mayoría de las empresas les interesan los pronósticos a corto plazo, por ejemplo el próximo mes, bimestre o año, estas predicciones se usan para planificar y controlar los negocios día a día en un futuro próximo.

En las predicciones a corto plazo, el objetivo principal al analizar la componente cíclica es identificar la posición actual y poder predecir el comportamiento futuro.

Es muy difícil encontrar un modelo regular, en promedio, que permita la proyección mecánica hacia el futuro en este tipo de casos, esto se debe a que existe mucha variabilidad de un ciclo a otro en la mayor parte de las actividades económicas; por esto, es que no se puede obtener predicciones confiables de los movimientos cíclicos en la actividad industrial o de empresas para poder proyectarlos mecánicamente hacia el futuro.

Además la identificación del estado actual de los movimientos cíclicos se ve obstaculizada por la presencia de movimientos irregulares.

Por lo tanto, con datos anuales, con la descomposición se llega hasta la obtención de las relativas cíclicas-irregulares.

VARIACIÓN ESTACIONAL

Dada una serie cronológica con valores mensuales, nos proponemos aislar su componente estacional; es decir, medir la influencia que depende del calendario y, por lo tanto, de las estaciones del año. Las variaciones estacionales generalmente provienen, en economía, de causas climáticas que determinan los ciclos vegetativos, lo que a su vez, influyen en la producción, el consumo, la ocupación y otros factores económicos, de manera que esta variación se repite año tras año de manera parecida.

Existen también modelos estacionales que se repiten en periodos inferiores a un año como por ejemplo los modelos diarios de rendimiento de productividad por hora en una planta.

Un conjunto de números mostrando los valores relativos de una variable durante los meses del año se llama **índice estacional** de la variable. Aunque dichos índices suelen determinarse en forma mensual, se pueden elaborar para otras subdivisiones de un año, por ejemplo bimestre, trimestrales, semanales, etc.

Como se dijo anteriormente a dicha variación estacional se la denomina índice estacional por ser mensuales los datos, este índice consta de 12 valores, uno por cada mes y cada uno expresa la actividad de ese mes en particular como porcentaje de la actividad del mes promedio.

Cada índice es un porcentaje con el promedio del año igual a 100 o sea cada índice mensual indica el nivel de ventas de producción, o de otra variable, en relación con el promedio anual de 100. Por ejemplo decir que el índice para octubre es de 107 significa que la variable en forma característica está 7% por arriba del promedio anual.

El método más usado el llamado método de razón a promedio móvil que elimina las componentes de tendencia, cíclica e irregular de los datos originales. Los números que resultan se denominan *índice estacional.*

Por ejemplo:

Se sabe que la venta durante enero, febrero, marzo, etc., son de 50, 120, 90,... por ciento de la venta media mensual del año completo, los números 50, 120, 90,... suministran el índice estacional del año y a veces se conocen el número del índice estacional. El promedio (media) del índice estacional deberá ser 100% es decir, la suma de los números índice deberá ser 12.

Al elaborar un índice estacional, todos los esfuerzos se encaminan a la eliminación de las variaciones de tendencia, cíclicas e irregulares de la serie para que se lo quede el estacional. La manera de lógralo, en el método básico de la razón del promedio móvil es relativamente simple.

INDICE ESTACIONAL

Obtención:

1) Se comienza obteniendo una serie de promedio móviles de 12 meses para eliminar los movimientos estacionales de la serie. Como un promedio móvil de n periodos elimina por completo cualquier movimiento recurrente pero absolutamente uniforme en los n

períodos, el promedio móvil de 12 meses suprimirá todos los movimientos estacionales de la serie.

Dichos patrones estacionales varían año tras año con absoluta regularidad; de manera que no se puede eliminar por completo las variaciones estacionales; sin embargo eliminará la mayor parte de esta variación, por lo tanto el promedio móvil de 12 meses será una estimación de la componente de la tendencia y de la variación cíclica.

2) El resultado se centra entre los dos meses centrales que forman cada total móvil, por ejemplo el primer total móvil que consta de los meses de enero a diciembre del primer año se coloca entre junio y julio de ese año, el segundo total móvil que consta de los meses de febrero del primer año a enero del segundo año se coloca entre julio y agosto del primer año y sucesivamente.

3) Para centrar este resultado dentro de un mes en particular se obtiene totales móviles de dos meses de los totales de 12 meses. El primer resultado que consiste en el total registrado entre junio y julio más el de julio y agosto se centra en julio del primer año. Al dividir esto totales por 24 se obtiene promedio móviles centrados. Se dice que estos promedios móviles centrados constan de las componentes cíclicas y tendencia de la serie.

4) Los datos originales se dividen por este promedio y esto hace que únicamente se tengan los factores de la variación estacional e irregular ya que en el modelo que se está analizando se obtendría:

Yi / (promedio móvil centrado)= SI =TSCI / TC

Para elaborar el índice estacional los datos de estas razones se reordenan de acuerdo a los valores mensuales para cada año. Se obtienen así para cada mes un valor de índice que se repetirán año año.

Entonces ahora resta eliminar hasta donde sean posibles las variaciones irregulares para obtener solo la parte estacional. Por ejemplo una manera de reducir estas fluctuaciones es a través del uso de la mediana de los valores dados de cada mes. Estos valores de la mediana se ajustan de manera tal que el valor total de los índices estacionales durante el año sea 12 y el promedio de cada índice estacional (mensual) sea 1.Este factor de corrección será 12/total de las doce medias.

RESUMIENDO: los datos originales contienen las cuatro componentes T C S I. El objetivo es eliminar S de los datos originales. Al obtener los promedios móviles se han eliminado las fluctuaciones estacionales e irregulares solo queda T y C. A continuación al dividir los datos originales por los promedios móviles se obtiene los valores de estacionalidad específicos SI que se expresan en forma de índice multiplicándolos por 100. Por último se toma la media o la mediana de todos índices mensuales ordenados para eliminar la mayor parte de las fluctuaciones irregulares y los valores resultantes indican el patrón de la variación estacional.

DESESTACIONALIZACIÓN DE DATOS

Un conjunto de índices estacionales es muy útil para ajustar las series respecto a fluctuaciones estacionales. La serie resultante se denomina serie desestacionalizada. La razón para desestacionalizar las series es eliminar las fluctuaciones estacionales a fin de estudiar la tendencia y el ciclo. Se consigue dividiendo a cada dato original por e l índice obtenido para cada período, así los datos contienen las componentes de T C e I.

RESUMEN DE LOS PASOS EN EL ANÁLISIS DE SERIES DE TIEMPO

1 Coleccionar los datos de la serie de tiempo, procurando asegurarse de que estos datos sean dignos de confianza. En la colección de datos se debe siempre tener en cuenta el propósito que se persigue en cada caso con el análisis de la serie de tiempo.

2 Representar la serie de tiempo, anotando cualitativamente la presencia de la tendencia de larga duración, variaciones cíclicas y variaciones estacionales.

3 Construir la curva o recta de tendencia de larga duración y obtener los valores de tendencia apropiados mediante cualquiera de los métodos, de mínimos cuadrados, libre, movimientos medios o semimedias.

4 Si están presentes variaciones estacionales, obtener un índice estacional y ajustar los datos a estas variaciones estacionales, es decir, desestacionalizar los datos.

5 Ajustar los datos desestacionalizados a la tendencia. Los datos resultantes contienen solamente las variaciones cíclicas e irregulares. Un movimiento medio de 3, 5 o 7 meses sirve para eliminar las variaciones irregulares y poner de manifiesto las variaciones cíclicas.

6 Representar las variaciones cíclicas obtenidas anteriormente, anotando cualquier periodicidad que pueda aparecer.

7 Combinando los resultados con cualquier otro tipo de información útil, hacer una predicción (si se desea) y si es posible discutir las fuentes de error y su magnitud.