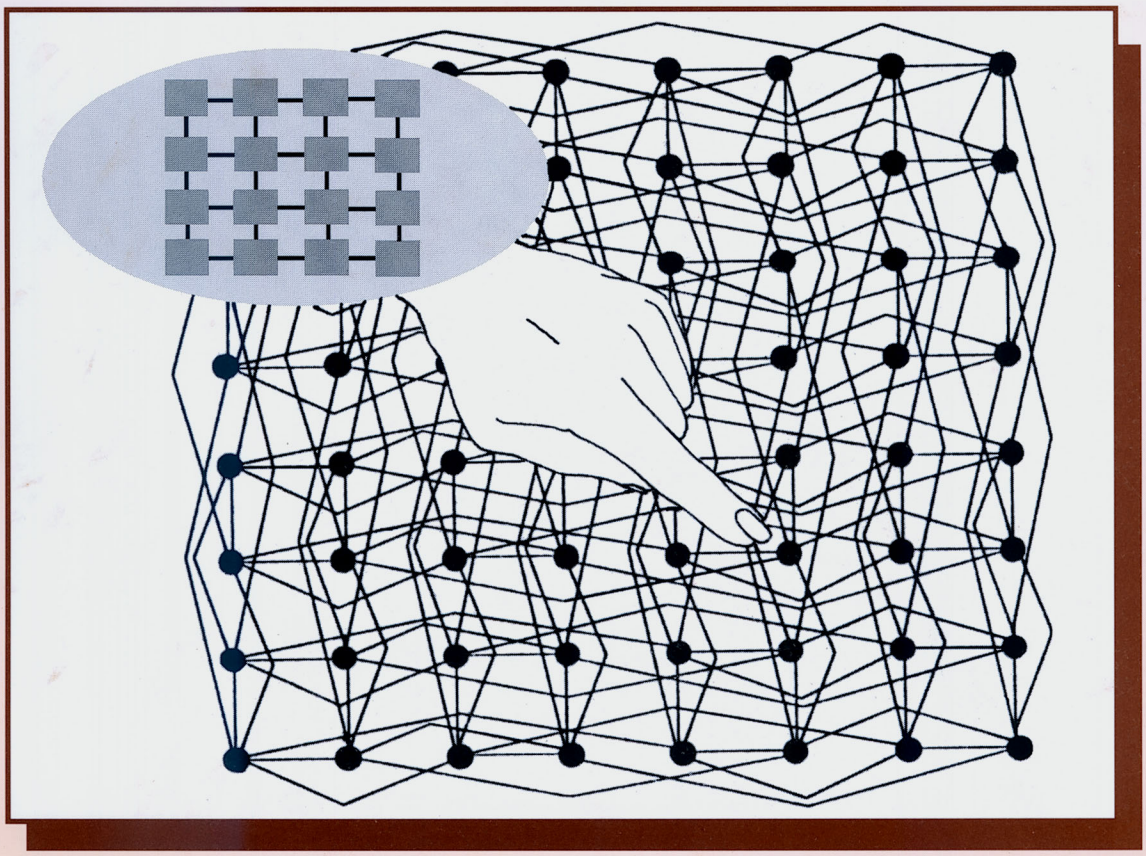
Ingeniería

Revista de la Universidad de Costa Rica
Julio/Diciembre 1996 VOLUMEN 6 Nº 2



INGENIERIA

Revista Semestral de la Universidad de Costa Rica
Volumen 6, Julio/Diciembre 1996 Número 2

CORRESPONDENCIA Y SUSCRIPCIONES

Editorial de la Universidad de Costa Rica
Apartado Postal 75
2060 Ciudad Universitaria Rodrigo Facio
San José, Costa Rica

CANJES

Universidad de Costa Rica
Sistema de Bibliotecas, Documentación e Información
Unidad de Selección y Aquisiciones-CANJE
Ciudad Universitaria Rodrigo Facio
San José, Costa Rica

Suscripción anual:
Costa Rica: ¢ 1 000,00
Otros países: US $ 25,00

Número suelto:
Costa Rica: ¢ 750,00
Otros países: $ 15,00

Edición aprobada por la Comisión Editorial de la Universidad de Costa Rica

Ciudad Universitaria Rodrigo Facio
San José, Costa Rica.

Revisión Filológica: *Lorena Rodríguez*

Diagramación:
José R. Argüello V.

Control de Calidad:
Unidad Diseño Revistas. Oficina de Publicaciones

Impreso en la Oficina de Publicaciones
de la Universidad de Costa Rica

Revista
620.005
I-46i Ingeniería / Universidad de Costa Rica. ——
Vol. I, no. 1 (ene./jun. 1991)— . —— San José, C. R. : Editorial de la Universidad de Costa Rica, 1991— (Oficina de Publicaciones de la Universidad de Costa Rica)
v. : il

Semestral.

1. Ingeniería - Publicaciones periódicas.

CCC/BUCR—250

Ingeniería 6 (2): 79-92 1996 San José, Costa Rica

LA ESTADÍSTICA Y LA COMPUTACIÓN EN APOYO A LA TOMA DE DECISIONES

Vladimir Lara V.[1]
Maureen Murillo R.[2]

Resumen

La mayoría de las decisiones a nivel táctico-estratégico de una organización necesitan un apoyo computacional que ofrezca más que simples datos descriptivos. Se esperaría que un sistema que apoye realmente este tipo de decisiones proponga al usuario diversas alternativas para que éste las conjugue, a su juicio, en la toma de decisiones. Estas decisiones, normalmente, involucran cierto grado de incertidumbre y de riesgo. Para crear un sistema de apoyo a estas decisiones en estas circunstancias, el profesional en el área de computación e informática no solo hace uso de las herramientas obtenidas en su formación académica, la cual está fuertemente basada en conceptos matemáticos, sino que debe recurrir a áreas del conocimiento ya desarrolladas que poseen técnicas que manejan la incertidumbre y que ayudan a pronosticar hechos con base en situaciones no claramente definidas, tal como la Estadística. Si se combina el conocimiento desarrollado en estadística, el poder de la computación y la experiencia del usuario, se podrán obtener verdaderos sistemas de soporte a la toma de decisiones.

En este artículo se expone cómo se complementó la aplicación de la Computación con técnicas del área de Estadística, enfatizando la investigación realizada en esta área, para construir un sistema de soporte a la toma de decisiones para un problema específico de la Escuela de Ciencias de la Computación e Informática (ECCI) de la Universidad de Costa Rica.

Summary

Most of the high level decisions in an organization need more than a simple and descriptive computarized data help to be made. A truly decision making support system should propose different alternatives that the user would combine during the decision making process. These decisions are usually made under uncertainty and risk. In order to create a decision making support system under these circunstances, the computer scientist not only uses the skills from his/her academic background, which is strongly based in mathematical concepts, but also uses developed knowledge areas, like statistics, that have tools to deal with uncertainty and help to build pronostics in situations not completely defined. The appropiate combination of statistical knowledge, the power of computer science and the decision maker's experience, would lead to the construction of truly decision making support systems.

This paper explains how the three areas mentioned above were complemented, enfasizing the research in statistics, in order to construct a decision making support system to solve a particular problem found in Computer Science Department of the University of Costa Rica.

1.- INTRODUCCION

Uno de los pasos que debe realizar la Dirección de la ECCI cuando tiene que decidir cuántos grupos abrir de cada curso en el semestre siguiente, es pronosticar la posible cantidad de estudiantes que matricularán cada uno de los cursos. En el momento de esta decisión, no se posee información certera que indique cuál será esta matrícula. De esta forma, el proceso de decisión debe basarse en los acontecimientos pasados y en la situación de los cursos en ese momento.

Un sistema automatizado que apoye este proceso deberá entonces tomar los datos pertinentes de semestres anteriores y utilizarlos de alguna manera para predecir el futuro. Ante esta situación, es conveniente recurrir a un área de estudio que provea métodos adecuados para el análisis de los datos, tal como la estadística.

La estadística, o los métodos estadísticos, pueden definirse como "la

[1] Ph.D, prof. Esc. Ciencias de la Computación e Informática, Fac. Ingeniería, UCR
[2] Licda., prof. Esc. Ciencias de la Computación e Informática, Fac. Ingeniería, UCR.

recopilación, presentación, análisis e interpretación de los datos numéricos. Los hechos que se estudian deben ser susceptibles de expresión numérica" [Croxton y Cowden, 1948, p.9]. La demanda de matrícula, que es el objeto de estudio de este sistema, puede expresarse totalmente en forma numérica, por lo que los métodos estadísticos son aplicables al problema de la estimación de la demanda.

El propósito del sistema realizado es analizar el comportamiento de la matrícula y brindar varios pronósticos de la misma. La rama de la estadística que se ocupa de este tipo de problemas es conocida como estadística inferencial. "El propósito de la estadística inferencial es obtener o formular inferencias (predicciones, decisiones) acerca de una población con base en información contenida en una muestra" [Mendenhall, 1990, p.5]. El sistema pone a disposición del usuario las herramientas de la estadística inferencial para que las complemente con su experiencia en la toma de decisiones.

2.- INVESTIGACION EN ESTADISTICA

La investigación realizada en el área de estadística se inicia con la especificación de los conceptos básicos del área que fundamentan los procesos, técnicas y métodos estadísticos relacionados con la problemática analizada. La acertada escogencia de los métodos que integran la solución depende fuertemente del conocimiento logrado en esta fase de la investigación.

La estadística trata de las técnicas para colectar, analizar y sacar conclusiones de datos [Snedecor y Cochran, 1981]. Estas técnicas se utilizan de acuerdo con un método estadístico definido que en general consta de cuatro pasos principales: recopilación de datos, presentación de los datos, análisis y conclusión acerca de los resultados obtenidos del análisis de los datos y de las técnicas aplicadas.

El problema de estimación de matrícula tiene un componente muy fuerte de análisis de datos, por lo que el resto de la investigación teórica se centra en las técnicas de análisis, con el fin de definir métodos que ayuden al pronóstico de la matrícula. Se investigaron los métodos de predicción que provee la Estadística; [Winston, 1994] propone clasificarlos en dos categorías:

Métodos de extrapolación: se usan para predecir valores futuros de una serie temporal a partir de valores en el pasado de otra serie temporal. En estos métodos se supone que los comportamientos y tendencias del pasado continuarán en los meses futuros, sin tomar en cuenta los factores o elementos que causaron el comportamiento de los datos.

Métodos de predicción causal: tratan de pronosticar valores futuros de una variable, la variable dependiente, mediante datos del pasado estimando la relación entre la variable dependiente y una o más variables independientes.

Esta clasificación corresponde a lo que, en estadística, se conoce comúnmente como análisis de series de tiempo y análisis de regresión, respectivamente. En las secciones siguientes se presentan estos dos tipos de métodos, las técnicas consideradas como adecuadas para la solución del problema y las herramientas computacionales que las soportan.

2.1.- Análisis de series de tiempo

Una serie de tiempo es "una colección de observaciones hechas secuencialmente en el tiempo" [Chatfield, 1980, p.1]. Uno de los objetivos de su análisis es predecir valores futuros de la serie. Tomando en cuenta que el objetivo del sistema es la predicción de la matrícula de los cursos y que los datos disponibles para obtener estos pronósticos son observaciones de cuál ha sido la matrícula a

través del tiempo, la teoría y las técnicas de series de tiempo son aplicables al problema.

Existen muchas técnicas estadísticas de pronóstico para las series de tiempo. La efectividad de un método u otro para predecir un valor futuro depende de las propiedades que presente la serie cronológica. Se seleccionaron dos métodos de pronóstico de series de tiempo para implantarlos en el sistema: los promedios móviles y la atenuación exponencial simple, ya que son adecuados para el tipo y la cantidad de datos que se posee actualmente sobre la matrícula.

2.1.1.- Promedios móviles

Esta técnica proyecta valores en un período próximo, basándose en el valor promedio de una variable sobre un número específico de períodos previos. Sean x_1, x_2, ... , x_t, ... valores observados de una serie temporal, donde x_t es el valor de esa serie que se observa durante el período t. Se define $f_{t,1}$ como el pronóstico para el período t+1 que se hace después de observar x_t. Para el método de promedios móviles,

$f_{t,1}$ = media de las N observaciones últimas

$$f_{t,1} = (1/N) \sum_{j=1}^{N} x_{t-j+1}$$

donde N es un parámetro dado que indica la cantidad de observaciones tomadas en cuenta en el promedio móvil [Winston, 1994]. Para seleccionar el valor de N se utiliza la desviación absoluta media[3], que es una medida de exactitud de la predicción. Se selecciona un valor para N que reduzca esta desviación a un mínimo [Winston, 1994].

Las predicciones con promedios móviles trabajan bien para una serie de tiempo que varía alrededor de un nivel base constante. De manera formal, estos pronósticos trabajan bien si $x_t = b + \varepsilon_t$ en la cual b es el nivel base de la serie y ε_t es la fluctuación aleatoria en el período t con respecto al nivel base. Este método es adecuado para la estimación de la demanda de matrícula ya que al observar las series de tiempo, la mayoría de los cursos tienen un nivel base (b), aunque en algunos casos la fluctuación aleatoria (ε) es bastante grande. Un promedio móvil de los últimos semestres es un estimado bastante acertado si el curso ha mantenido su comportamiento en los últimos ciclos lectivos.

2.1.2.- Atenuación exponencial simple

El método de atenuación exponencial simple predice un valor basándose en el pronóstico del período anterior, ajustado con el error del pronóstico previo. Para representar esta técnica [Winston, 1994], sea A_t = promedio atenuado de una serie temporal después de observar a x_t. A_t es el pronóstico del valor de la serie durante cualquier período futuro después de observar x_t. La ecuación de la atenuación exponencial simple es

$$A_t = \alpha x_t + (1 - \alpha)A_{t-1}$$

En la ecuación, α es la constante de atenuación que satisface $0<\alpha< 1$. Si se trata de predecir un período más adelante, el error de predicción de x_t, representado por e_t, está dado por $e_t = x_t - A_{t-1}$. Al desarrollar la ecuación de la atenuación exponencial simple se obtiene

$$A_t = \alpha x_t + (1 - \alpha)A_{t-1}$$
$$A_t = \alpha x_t + A_{t-1} - \alpha A_{t-1}$$
$$A_t = A_{t-1} + \alpha(x_t - A_{t-1})$$
$$A_t = A_{t-1} + \alpha e_t$$

[3] Desviación absoluta media (DAM): medida de la exactitud de predicción. Consiste en el promedio de los valores absolutos de todos los errores de pronóstico, es decir, DAM = ($|e_1|$ + $|e_2|$ + ... + $|e_n|$) / n, donde e_t es el error de predicción para el período t dado por $e_t = x_t$ - (pronóstico de x_t) teniendo que x_t es la observación de la serie en el período t [Winston, 1994].

De esta forma se observa que la constante de atenuación α es la magnitud que determina qué tan fuerte responde el pronóstico a los errores de la predicción anterior. Esto significa que si se *sobrepronosticó* a x_t, se bajará el pronóstico, y que si se *subpronosticó* se elevará la predicción. En la práctica se escoge α en general como 0.10, 0.30 o 0.50 [Winston, 1994].

La atenuación exponencial simple es más efectiva como método de pronóstico cuando las influencias cíclica y regular representan los principales efectos sobre los valores de una serie de tiempo [Kazmier y Díaz, 1993]. También, si una serie de tiempo fluctúa con respecto a un nivel base, se puede utilizar el método de atenuación exponencial simple para obtener buenos pronósticos de valores futuros de la serie [Winston, 1994].

Las razones para utilizar este método para predecir la demanda de matrícula son similares a las observadas para los promedios móviles. La demanda, como se explicó anteriormente, varía en un nivel base por lo que el modelo es adecuado para los datos. Una estimación basada en la predicción anterior y ajustada con el error que hubo con respecto a la realidad, representa un buen pronóstico de la demanda, ya que se espera que la demanda de matrícula de un curso, bajo condiciones normales, sea similar a la que se dio en el semestre anterior o trasanterior. Las observaciones más antiguas van perdiendo relevancia para el pronóstico si se toma en cuenta que a través del tiempo las características de un curso cambian modificando su patrón de comportamiento.

El método de promedios móviles y el de atenuación exponencial simple le dan mayor peso a las últimas observaciones de la demanda, que es lo adecuado para los pronósticos de matrícula bajo la suposición de que los últimos semestres reflejan mejor el comportamiento de la demanda pues las características del curso ya se habrán estabilizado.

2.2.- Análisis de regresión

En muchos problemas existen dos o más variables que están inherentemente relacionadas y en donde es posible estudiar esa relación. El análisis de regresión es una técnica estadística para plantear el modelo e investigar esta relación [Hines y Montgomery, 1988]. El objetivo del análisis de regresión es deducir una fórmula de predicción que se base en una o más variables predictorias. Las técnicas regresivas son adecuadas para la predicción de la demanda de matrícula, ya que la demanda en un semestre puede definirse como el resultado de variables que la afectan directamente y que determinan su comportamiento.

En las secciones siguientes se presentan tres tipos de técnicas de regresión ampliamente utilizadas en Estadística: la regresión lineal simple que estudia la relación entre dos variables que se relacionan linealmente, el ajuste de relaciones no lineales que describe la relación no lineal entre dos variables, y la lineal múltiple que se aplica en los casos en que existe una relación lineal entre una variable dependiente y dos o más variables independientes.

2.2.1.- Regresión Lineal Simple

Frecuentemente se trata de predecir el valor de una variable, la variable dependiente (y), a partir del valor de otra variable, la variable independiente (x). Si la variable dependiente y la independiente se relacionan en forma lineal, se puede aplicar la regresión lineal simple para estimar la relación [Winston, 1994]. La finalidad de este método es modelar la relación funcional entre x_i y y_i mediante la siguiente fórmula:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

en donde β_0 es la intersección con el eje de las ordenadas, β_1 es la pendiente de la recta y ε_i es un término de error en el ajuste del modelo en el i-ésimo dato.

Cuando se aplica esta técnica de regresión lineal simple, deben seguirse tres pasos fundamentales. El primero es verificar que exista una relación lineal entre la variable dependiente y la variable independiente. Para llevar a cabo el primer paso se realiza un análisis llamado correlación. El análisis de correlación trata de medir el grado de relación entre dos variables por medio de un simple número denominado *coeficiente de correlación* (r_{xy}) [Walpole y Myers, 1986]. Los paquetes computacionales estadísticos brindan herramientas para obtener este coeficiente. El coeficiente de correlación varía entre los valores +1 y -1. Así, +1 indica una correlación perfecta positiva, es decir, que los valores de una variable aumentan a medida que aumentan los valores de la otra variable; -1 indica una correlación perfecta negativa, esto es, que los valores de una variable disminuyen al aumentar los valores de la otra variable, y 0 indica la falta de correlación [Croxton y Cowden, 1948].

Una vez que se ha verificado que existe una relación lineal entre la variable dependiente y la independiente, se localizan los puntos observados en una gráfica. Para graficar estos datos, se representa la variable dependiente en el eje de las ordenadas y la variable independiente en el eje de las abscisas.

El último paso es encontrar la recta que mejor se ajuste a los datos para así obtener la ecuación de regresión que servirá para predecir valores futuros. Para escoger esta recta se utiliza el criterio llamado *principio de mínimos cuadrados*, el cual puede establecerse de la siguiente manera:

> "Escoger como la recta de mejor ajuste la que minimice la suma de los cuadrados de las desviaciones de los valores observados de y respecto de los pronosticados" [Mendenhall, 1990, p.446].

Una vez encontrada la recta, es posible utilizarla para obtener los pronósticos. Luego de calcular el pronóstico, deben verificarse algunas medidas con respecto al análisis realizado para evaluar qué tan bueno es el ajuste y la predicción. Primero, para determinar qué tan bien se ajusta la recta de cuadrados mínimos a los puntos de los datos (bondad del ajuste), se define el coeficiente de determinación (R^2) que es el porcentaje de la variación de y que queda explicado por x [Winston, 1994][4].

Luego, para descartar la posibilidad de que la relación lineal entre las variables se deba a la casualidad se calcula con algún paquete computacional estadístico el nivel de significancia del valor F, es decir, la probabilidad de que la relación entre las dos variables se deba a la casualidad. Para ello se define un término o nivel α, al cual se le da generalmente un valor de 0.05 ó 0.10. Si el nivel de significancia de F es menor que el valor de α, entonces se concluye que la ecuación de regresión es útil para el pronóstico de valores futuros.

Por último, se define el error estándar de la estimación (s_e) que es una medida de la exactitud de las predicciones obtenidas con regresión. Se espera que un 68% de los valores de y queden dentro de una distancia s_e del valor predicho y que el 95% de los valores de y queden dentro de los márgenes $2s_e$ del valor predicho [Winston, 1994].

[4] En general, en el área de las ciencias exactas, un R^2 mayor a 0.7 se considera como un buen ajuste, mientras que en el área de las ciencias sociales, un R^2 mayor a 0.5 representa un ajuste aceptable (Información brindada por la licenciada en Estadística Sharon Kuhlmann).

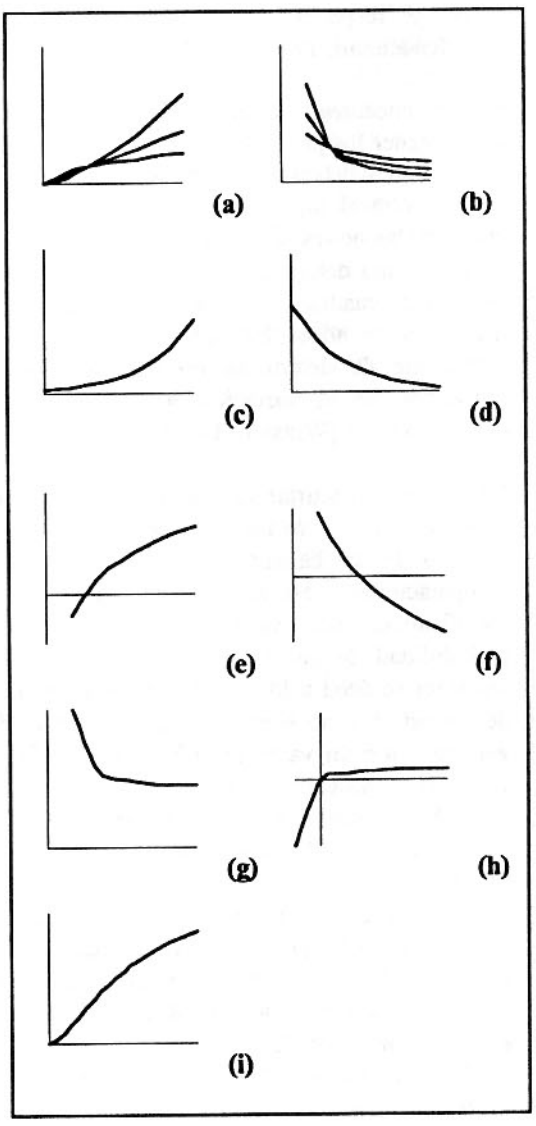


Figura N°1.Relaciones no lineales [Winston, 1994]

Finalmente, el análisis de regresión lineal simple requiere la validez de los siguientes supuestos: [Winston, 1994]

1. La variancia del término de error no debe depender del valor de la variable independiente x. Para probar esta hipótesis, se grafican los errores en dirección de las ordenadas y el valor de x en el eje de las abscisas. Si en el gráfico se observa que no hay tendencia a que la magnitud de los errores dependa de x entonces se satisface esta hipótesis.
2. Los errores se distribuyen normalmente. Para probar este supuesto, se grafican los errores contra la frecuencia relativa acumulada en una escala de probabilidad. Si el resultado se parece a una línea recta, puede afirmarse que se cumple este supuesto.
3. Los errores deben ser independientes. La independencia de los errores significa que el conocimiento del valor de un error no dice nada acerca del valor del error siguiente, o de cualquier otro de los que sigan. Se puede comprobar esta hipótesis graficando los errores en una sucesión a través del tiempo.

2.2.2.- Ajuste de relaciones no lineales

Con frecuencia la gráfica de los puntos (x_i, y_i) indica que y no es función lineal de x. Sin embargo, en esos casos la gráfica puede indicar que hay una relación no lineal entre x y y.

Para aplicar la regresión a este tipo de relaciones, es necesario transformar cada punto de datos siguiendo algunas reglas según sea el tipo de relación existente. Por ejemplo, la gráfica de los puntos $(x_{i'i})$ podría ser similar a alguna de las presentadas en la figura N°1.

Para estimar una relación no lineal se aplica el siguiente procedimiento [Winston, 1994]:

Paso 1. Graficar los puntos y ver cuál de las gráficas de la figura N°1 se ajusta mejor a los datos.

Paso 2. Se determina la relación funcional entre x y y de acuerdo con la segunda columna del cuadro N°1.

Paso 3. Transformar cada punto de datos siguiendo las reglas de la tercera columna de la tabla anterior. Los datos transformados deben, si se grafican, seguir una relación lineal.

Paso 4. Estimar la recta de regresión de cuadrados mínimos para los datos transformados. Si β_0 es la ordenada al origen de esa línea, para los datos transformados, y β_1 es la pendiente de esa línea para los datos transformados, entonces la relación estimada es la que se presenta en la columna 4 del cuadro N°1.

Si la gráfica se parece a la figura	Relación funcional entre *x* y *y*	Transformar (x_i, y_i) en	Estimado de la relación funcional
(a) o (b)	$y = \beta_0 x^{\beta 1}$	($\ln x_i$, $\ln y_i$)	$y = \exp(\beta_0)\, x^{\beta 1}$
(c) o (d)	$y = \beta_0 \exp(\beta_1 x)$	(x_i, $\ln y_i$)	$y = \exp(\beta_0 + \beta_1 x)$
(e) o (f)	$y = \beta_0 + \beta_1 (\ln x)$	($\ln x_i$, y_i)	$y = \beta_0 + \beta_1 (\ln x)$
(g) o (h)	$y = x / (\beta_0 x + \beta_1)$	($1/x_i$, $1/y_i$)	$y = x / (\beta_0 x + \beta_1)$
(i)	$y = \exp(\beta_0 + \beta_1/x)$	($1/x_i$, $\ln y_i$)	$y = \exp(\beta_0 + \beta_1/x)$

Cuadro N°1. Relaciones no lineales [Winston, 1994]

2.2.3.- Regresión Lineal Múltiple

Muchos problemas de regresión utilizan más de una variable independiente para explicar la variación de la variable dependiente. En estos casos se aplica el método de la regresión lineal múltiple y se plantea el modelo lineal de la siguiente forma:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} + \varepsilon_i$$

donde k es el número de variables independientes, i es el i-ésimo punto de datos y ε_i es un término de error con promedio 0 que representa el hecho de que el valor real de y_i puede no ser igual a $\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki}$. En general, β_j (coeficiente de regresión) es el aumento de *y* si aumenta en 1 el valor de la j-ésima variable independiente, quedando constantes las demás variables independientes.

Igual que en la regresión lineal simple, para estimar los coeficientes de regresión de la ecuación de mejor ajuste se emplea el método de mínimos cuadrados. Estas estimaciones se pueden obtener con un programa computacional estadístico.

Para este método también se define la bondad del ajuste (R^2), que significa el porcentaje de variación de *y* explicado por las *k* variables independientes. Además, se obtiene el error estándar de la estimación (s_e), que significa que un 68% de los valores de *y* quedarán a una distancia del pronóstico menor que s_e y que un 95% del los valores de *y* quedarán a una distancia del pronóstico menor que $2s_e$.

También debe determinarse si en realidad existe una relación lineal entre la variable dependiente y el conjunto de variables independientes. Si la significancia de F, que es la probabilidad de concluir erróneamente que hay una relación lineal entre la variable dependiente y las independientes, es menor que el α definido (generalmente 0.05 o 0.10), entonces la ecuación de regresión es útil para predecir futuros valores y al menos una variable independiente contribuye significativamente al modelo [Hines y Montgomery, 1988].

Es necesario entonces, determinar si cada coeficiente de crecimiento de las variables independientes es útil en forma individual en el modelo de regresión, ya que el uso de variables

sin importancia puede reducir la efectividad de las ecuaciones de predicción [Walpole y Myers, 1986]. Para llevar a cabo esta prueba, los programas estadísticos calculan el nivel de significancia del valor *t* (valor de probabilidad *p*) para cada una de las variables independientes. Si este nivel de significancia es menor que el α definido, entonces la variable sí tiene efecto significativo sobre *y* cuando las demás variables independientes entran en la ecuación de regresión; en caso contrario esta variable debería eliminarse del modelo y volver a calcular la regresión [Winston, 1994].

Como resultado de este proceso se obtienen varias ecuaciones de regresión con distintos conjuntos de variables independientes. Generalmente se escoge la ecuación con el menor valor de s_e (error estándar), porque será la que dará pronósticos más exactos. Pero a la vez se desea que las medidas estadísticas *t* para todas las variables de la ecuación sean significativas y estos dos objetivos pueden contraponerse, en cuyo caso será difícil determinar la mejor ecuación [Winston, 1994].

Debe verificarse que las variables independientes no presenten una relación lineal muy fuerte entre ellas, ya que podrían tener efectos graves sobre los estimadores de los coeficientes de regresión y sobre la generalidad del modelo [Hines y Montgomery, 1988]. Además, debe comprobarse que los errores de la regresión sean independientes y que no dependan de los valores pronosticados de *y*.

La regresión lineal múltiple es adecuada para la estimación de la matrícula ya que su demanda depende de otros factores que la afectan directamente. Este método toma en cuenta dos tipos de elementos que son importantes en la demanda de la matrícula. El primero es de tipo histórico que se refleja en los coeficientes de la ecuación de regresión y cuya estimación se basa en los datos históricos de la variable dependiente y de las variables independientes. El otro tipo de elementos representan la actualidad por medio de las variables independientes, a las cuales el usuario les asigna valores recientes.

3.- VALIDACION DE LOS DATOS

La validación de los datos consiste en el análisis de las características pertinentes de los datos para determinar su conveniencia al ser utilizados por los modelos. Los datos relacionados con la demanda de matrícula de los cursos se analizaron desde dos enfoques diferentes: como series de tiempo y como datos para análisis de regresión.

3.1.- Características de los datos de matrícula como series de tiempo

El análisis de los datos como series de tiempo debe tomar en cuenta la fiabilidad de los datos, ya que de ellos depende la confiabilidad de los resultados obtenidos con los métodos de pronóstico. Los datos históricos de la matrícula de los cursos se tomaron de la Oficina de Registro, que es la entidad que mantiene la información oficial de la Universidad, lo que garantiza su fiabilidad.

La cantidad de datos disponibles es fundamental para determinar un comportamiento claro de las series. Para un análisis cronológico 30 datos son apenas los suficientes. En promedio cada curso de la ECCI tiene 20 datos disponibles. A pesar de que en este momento la cantidad de información disponible no es la adecuada, el sistema incluye los métodos de series de tiempo considerando que con el tiempo la cantidad de información aumentará. Además, los métodos seleccionados le dan mayor peso a las últimas observaciones de la matrícula, que posiblemente son las que representan mejor la situación actual de los cursos.

Por otro lado, la mayoría de los cursos se ubican en alguno de los siguientes grupos: los que luego de una matrícula irregular estabilizaron su comportamiento, los que se mantienen en un nivel base constante, los que muestran una tendencia creciente o decreciente y los que aparentemente tienen un comportamiento totalmente irregular. A pesar de estas caracterizaciones, no puede asegurarse que los cursos muestran un patrón de comportamiento definido, razón por la cual aparentes datos "anormales" en un curso no se eliminan para la utilización de los modelos, ya que podrían constituir el cambio a otra fase de la serie de tiempo aún desconocida.

3.2.- Características de los datos utilizados en regresión

Para un análisis de regresión son necesarios dos tipos de datos: los datos históricos de la variable que se quiere predecir (variable dependiente) y los datos de las variables que determinan su comportamiento (variables independientes). En el sistema la variable dependiente es la matrícula de cada curso y las variables independientes están definidas de la siguiente forma: para los cursos del primer nivel del plan de estudio que no tienen requisitos las variables independientes son el cupo de admisión a carrera y el número de estudiantes reprobados en el curso; para el resto de los cursos las variables son el número de reprobados en el curso y la cantidad de estudiantes aprobados en cada uno de sus requisitos.

Al igual que para los métodos de series de tiempo, se garantiza la fiabilidad de los datos ya que fueron obtenidos de la Oficina de Registro. En cuanto a la cantidad de datos históricos, el análisis de regresión es menos exigente, por lo que se posee un número adecuado de datos para los cálculos en la mayoría de los cursos.

Otra característica de los datos que hay que observar desde el enfoque del análisis de regresión, es que las variables independientes identificadas influyan realmente en el comportamiento de la matrícula de los cursos. Esto se verificó utilizando correlaciones. Además, los métodos de cálculo de las regresiones utilizados dan como resultado el mejor modelo de regresión combinando variables que no estén correlacionadas entre sí.

Por último, en el análisis previo a la confección de los modelos de regresión de los datos deben identificarse datos "anormales" de los cursos que se deban a situaciones no comunes y descartarlos como datos válidos para el análisis, ya que no reflejan las condiciones normales de estos cursos, sin embargo, deben mantenerse en el sistema como datos históricos.

4.- ESPECIFICACION DE LOS MODELOS DE PRONOSTICO

La especificación de los modelos consiste en tomar los conceptos teóricos de los métodos y a partir de ellos definir su funcionamiento y calcular algunos parámetros que los modelos necesitan para su operación. El sistema incluye cuatro métodos de pronóstico: promedios móviles, atenuación exponencial simple, regresión lineal múltiple y el de análisis por requisitos que es resultado del análisis del problema de la estimación de matrícula en la ECCI que se realizó para este sistema. Para cada uno de los métodos se presentan las especificaciones realizadas antes de ser incorporados en el sistema.

4.1.- Promedios móviles

La definición teórica de este método no requiere de ningún trabajo adicional previo a su incorporación en el sistema. El único parámetro que necesita el método para su operación es el número de datos que se tomarán en cuenta para el pronóstico, pero este número es

proporcionado por el usuario en el momento de la estimación de la demanda de matrícula.

4.2.- Atenuación exponencial simple

El método de atenuación exponencial simple tampoco requiere un trabajo adicional a partir de su definición teórica para ser incorporado en el sistema. Para su funcionamiento lo único que requiere es de un valor α (constante de atenuación) que sirve para determinar en cuánto se ajustará el pronóstico al error de la predicción anterior, pero éste es especificado por el usuario en el momento del cálculo de los pronósticos.

Para calcular los pronósticos mediante el método de atenuación exponencial simple, el sistema utiliza el programa *Excel*. El sistema le transfiere los datos necesarios a *Excel* para que éste efectúe los cálculos y le devuelva el resultado final.

4.3.- Regresión lineal múltiple

Para cada uno de los cursos de la ECCI, se seleccionó un modelo de regresión lineal múltiple particular. El cálculo de estos modelos se puede realizar según los métodos siguientes:

- *Stepwise*: elabora iterativamente una secuencia de modelos de regresión agregando o quitando variables en cada paso [Hines y Montgomery, 1988].
- *Forward*: agrega una variable a la vez, hasta que no queden variables candidato restantes que produzcan un incremento significativo en la suma de cuadrados de la regresión [Hines y Montgomery, 1988].
- *Backward*: comienza con todas la variables del modelo y luego va eliminando la variable con el menor estadístico F parcial si este estadístico F es insignificante, hasta que no se puedan eliminar más variables [Hines y Montgomery, 1988].

Para cada curso se compararon los modelos de regresión resultantes de estos métodos y se seleccionó el mejor de acuerdo con las medidas que el mismo modelo proporciona. Además, se obtuvieron otras regresiones variando algunos parámetros de los métodos para determinar cuál combinación de parámetros y métodos de cálculo proporcionaban el mejor modelo.

En todos los cursos se obtuvieron mejores resultados cuando la recta de regresión pasaba por el origen. Esto se interpretó como el hecho de que la matrícula está determinada prácticamente en su totalidad por las variables incluidas en el modelo, es decir, si todas las variables asumieran un valor 0 el cálculo de la matrícula daría también 0 (pasa por el origen). Para todos los modelos obtenidos se verificó el cumplimiento de los supuestos de la regresión lineal múltiple acerca de la independencia y de la distribución normal de los errores.

4.4.- Análisis por requisitos

Este método se creó específicamente para el pronóstico de matrícula en la ECCI. No está basado en conceptos teóricos predefinidos, sino que es el resultado del análisis del problema y se basa en las características de cada curso y en los factores que aportan o contribuyen a determinar la matrícula. A continuación se describe la especificación final del método.

Para un curso del primer nivel del plan de estudio, el método permite tomar en cuenta el cupo de admisión a la carrera y la cantidad de estudiantes reprobados en el mismo curso en el semestre anterior al que se está estimando. Para el resto de los cursos del plan, el método da la posibilidad de tomar en cuenta el número de estudiantes reprobados en el curso en el semestre anterior al que se estima y la cantidad de estudiantes que aprueban los requisitos del curso

en el período previo al que se pronostica. El usuario tiene la posibilidad de seleccionar cuáles de los requisitos del curso serán tomados en cuenta para el pronóstico.

Para dar un estimado de la posible matrícula de un curso, este método utiliza una combinación de datos históricos y de datos actuales. Los datos históricos los utiliza para obtener posibles estimados de algunos datos que requiere el método para sus cálculos pero que en el momento de la estimación no se tiene certeza de su valor, por ejemplo, el porcentaje de aprobación de alguno de los requisitos. Los datos actuales son útiles para describir la situación del curso en el momento de la estimación, por ejemplo, la matrícula actual del curso.

Un elemento importante en este método es la utilización de un *factor de atenuación.* Este factor indica qué porcentaje de todos los posibles estudiantes a matricularse en un curso, realmente se matriculan. Este factor trata de tomar en cuenta el hecho de que posiblemente cierta cantidad de estudiantes matriculados en un requisito también están matriculados en los otros requisitos, por lo que no deben ser tomados en cuenta en forma repetida.

El factor de atenuación de un curso se obtiene a través de sus datos históricos mediante el siguiente procedimiento general: para cada período en que se haya impartido el curso se calculan todos los posibles estudiantes a matricularse, en donde "todos los posibles estudiantes a matricularse" se define en forma general como la suma de los reprobados en el curso y los aprobados de todos los requisitos en el período anterior; luego este número se divide entre el número real de estudiantes que se matricularon obteniéndose el factor del período que se analiza. Finalmente, se promedian los factores de todos los períodos en que se impartió el curso obteniéndose su factor final. En el caso de que el curso no tenga datos históricos su factor no se calcula, asumiéndose un factor igual a 1 en el cálculo de los pronósticos.

A continuación se presentan los pasos generales que sigue el método de análisis por requisitos para obtener el pronóstico de la matrícula de un curso particular para un semestre específico.

Algoritmo:

- Obtener cupo actual de admisión a la carrera (cupo)
- Obtener factor de atenuación (factor)
- Si el factor no existe
  - factor=1
- Si el usuario desea tomar en cuenta los reprobados en el curso
  - suma = # de estudiantes matriculados en el curso en el semestre anterior * % de reprobación del curso
- Si el usuario desea tomar en cuenta el cupo de la carrera
  - suma = suma + cupo
- Si el usuario desea tomar en cuenta los requisitos del curso
  - Si desea tomar en cuenta todos los requisitos
    - Para todos los requisitos
      - suma = suma + matriculados en requisito * % de aprobación del requisito
    - Si el curso tiene un único requisito
      - factor = 1
  - Si no
    - factor = 1
  - Si desea tomar en cuenta el requisito con el menor número de aprobados
    - Obtener el requisito con el menor número de estudiantes aprobados
    - suma = suma + matriculados en requis. menor * % de aprob. del requis. menor
  - Si desea tomar en cuenta el requisito con el mayor número de aprobados
    - Obtener el requisito con el mayor número de estudiantes aprobados
    - suma = suma + matriculados en requis. mayor * % de aprob. del requis. mayor

- •Si desea tomar en cuenta un requisito específico
  - •suma = suma + matriculados en requis. seleccionado * %de aprob. del requis. seleccionado

•Pronóstico = suma * factor

Con base en los datos que se analizaron y en los métodos especificados en esta sección, se implementó el sistema de soporte a la toma de decisiones para la estimación de matrícula de la ECCI.

5.- IMPLANTACION DE LA SOLUCION

Parte de la investigación en el área de estadística consiste en la búsqueda y selección de herramientas computacionales que soporten los métodos estadísticos estudiados. Esto con el fin de llevar a cabo el análisis preliminar de los datos y a la vez que la herramienta sirva de apoyo al sistema desarrollado. Se debe tratar de utilizar para la puesta en funcionamiento de los métodos estadísticos del sistema, aplicaciones ya desarrolladas que los provean, a menos que se decida desarrollarlos en el propio sistema debido a causas tales como eficiencia y tiempo de respuesta.

Para el análisis estadístico de los datos y de los métodos de análisis del sistema estudiado se utilizaron los programas *SPSS* y *Excel* el cual posee un módulo para el análisis estadístico. Estas herramientas aportan todas las técnicas necesarias para el análisis, pero con la característica de ser sencillas de utilizar para un usuario poco experimentado en el área de la estadística.

En esta aplicación el usuario es parte integral del sistema, ya que es el que toma la decisión final después de evaluar las diferentes alternativas propuestas por el sistema; además interviene directamente en el cálculo de los pronósticos que realiza la aplicación. Los sistemas de soporte a la toma de decisiones (DSS) satisfacen eficazmente esta relación del usuario con el sistema y es la orientación técnica de la aplicación de estimación de matrícula sobre la que se montaron los métodos estadísticos.

El sistema está formado por tres módulos: pronósticos, edición de modelos y edición de datos. En el primer módulo se obtienen los resultados que se buscan, en el segundo se modifican los mecanismos de cálculo de estos resultados y en el tercero se actualizan los datos que permiten efectuar estos cálculos.

Para probar la efectividad de la incorporación de los métodos estadísticos en el sistema computacional de soporte a la toma de decisiones, se presentan en el cuadro Nº2 los resultados de los porcentajes de acierto de los pronósticos, obtenidos luego de aplicar los cuatro métodos a situaciones reales durante el año 1996. También se incluye para efectos de comparación los porcentajes de acierto de los pronósticos obtenidos en forma manual por la Dirección de la ECCI en el mismo año.

Como puede observarse en el cuadro Nº2, los métodos del sistema mejoraron significativamente los pronósticos de la demanda de matrícula que normalmente se realizan manualmente. Esto indica que la combinación de estadística con computación fue efectiva en el trato de un problema poco estructurado como el que se abordó en este artículo.

	Dirección ECCI	Promedios Móviles	Atenuación Exponencia l	Regresión Lineal Múltiple	Análisis por Requisitos
Pronósticos buenos (76-100% de acierto)	52%	75%	72%	86%	84%
Pronósticos regulares (51-75% de acierto)	23%	19%	22%	7%	12%
Pronósticos malos (0-50% de acierto)	25%	6%	6%	7%	4%

Cuadro N°2. Resultado de pronósticos

6.- CONCLUSION

De lo expuesto en este artículo se concluye que:

- Se debe fomentar la incorporación, por parte de los profesionales en Computación e Informática, de conocimientos de otras disciplinas en los sistemas computacionales, de tal forma que las soluciones que se propongan estén sustentadas sobre bases teóricas más sólidas con respecto al problema que se enfrenta.

- Debe dársele un apoyo computacional más directo a los niveles táctico y estratégico para la toma de decisiones, las cuales son el motor que mueve a toda organización.

- La Estadística es un área que los profesionales en Computación e Informática pueden explotar para proveer soluciones efectivas en un sistema de soporte a la toma de decisiones a situaciones no estructuradas.

7.- BIBLIOGRAFIA

Chatfield, C. The Analysis of Time Series: An Introduction. Segunda edición. Chapman and Hall. New York. 1980.

Croxton, Frederick E y Cowden, Dudley J. Estadística General Aplicada. Fondo de Cultura Económica. México. 1948.

Hines, William W. y Montgomery, Douglas C. Probabilidad y Estadística para Ingeniería y Administración. Compañía Editorial Continental, S.A. México. 1988.

Kazmier, Leonard y Díaz Mata, Alfredo. Estadística aplicada a la Administración y a la Economía. Segunda edición. McGraw-Hill. México. 1993.

Mendenhall, William. Estadísitica para Administradores. Segunda Edición. Grupo Editorial Iberoamérica. México. 1990.

Snedecor, George W. y Cochran, William G. Métodos Estadísticos. Compañía Editorial Continental, S.A. México. 1981.

Walpole, Ronald E. y Myers, Raymond H. Probabilidad y Estadística para Ingenieros. Tercera edición. Nueva Editorial Interamericana. México. 1986.

Winston, Wayne L. Investigación de Operaciones. Grupo Editorial Iberoamérica. México. 1994.