Técnicas estadísticas de aplicación en las FAS

RAFAEL GARCIA MARTIN,
Capitán de Intendencia del Aire

TRATAR de condensar, incluso de forma somera en unas pocas líneas la totalidad de las técnicas estadísticas, es del todo tarea imposible. Reducir su número al ámbito de aplicación de las Fuerzas Armadas es inútil pues lógicamente todas tendrían cabida. Vamos pues a centrarnos en las de más uso como son aquellas que están dirigidas a la comparación, la predicción, y la clasificación.

Uno de los principales problemas con que se encuentra el poseedor de información homogénea de fuentes distintas, es el de la comparación entre estas fuentes. Cuestiones tales como ¿se avería más un equipo que otro?, ¿más en verano que en invierno?, o ¿más, los que más cerca están de la costa, y, menos los del interior? se plantean si existe un mínimo de curiosidad.

La resolución de este tipo de cuestiones está directamente relacionada con la **inferencia estadística.** A esta le incumbe el problema de sacar conclusiones de determinados sucesos, basándose en las observaciones registradas para una, generalmente reducida cantidad de ellos. Preguntas tales como: ¿representan lo mismo dos grupos de medidas de una característica?, ¿pertenece este elemento, cuyas características observadas tienen tal valor, a este conjunto, o a aquel otro?; se presentan a cada paso de la investigación, o en general en cualquier proceso encaminado a profundizar en el conocimiento de la realidad observada.

El esquema general del método para contestar tales preguntas es, en esencia, el mismo: planteamiento de la hipótesis, esto es aseveración a priori de una posición respecto a la pregunta formulada; construcción de un indicador (que recibe el nombre de estadístico), como función de los datos obtenidos; y por, respuesta a la pregunta formulada, mediante el rechazo, o aprobación, de nuestra postura, en base a la probabilidda de que bajo esta hipótesis de trabajo, el valor obtenido para el estadístico sea, o no, el que cabía esperar.

Es evidente la necesidad de que cualquier método desarrollado para resolver estas cuestiones ha de ser un método objetivo, y si además de objetivo tiene un substrato científico y racional, aún mejor, (porque el hecho de que todos estén de acuerdo en afirmar algo, no es garantía de que sea eso lo correcto).

Mediante el modelo general de **regresión,** se cuantifica cuánto, y de que forma, una variable dependiente está relacionada con una, o varias, variables dependientes, todas ellas numéricas.

La verdadera relación entre ambos grupos de variables es desconocida, y su estimación debe estar basada en la evidencia empírica de los datos de que disponemos, para llegar a un modelo de la forma general $\mathbf{Yi=f(Xi+\beta)+\varepsilon i}$ donde Xi es el conjunto de variables que afectadas de los parámetros β nos permiten inferir el valor de Yi.

La variedad de tipos que pueden ser considerados respecto a la forma de la relación funcional "f", y la facilidad de cálculo de las estimaciones de los parámetros β, hace de la regresión una técnica usada con frecuencia.

Sin embargo, las hipótesis que se hacen en la construcción del modelo son muy restrictivas, y un estudio de las diferencias entre el modelo construído y los datos reales, que nos permita asegurarnos de no haber olvidado relaciones entre las variables, de que estas no estén autocorrelacionadas, etc, es un paso definitivo a seguir, tras haber obtenido un valor del coeficiente de correlación que ponga de manifiesto una relación en el sentido de "f".

Algo, no por evidente, muchas veces olvidado es el hecho de que una relación causal fuerte entre las variables siempre implica una relación matemática también fuerte, pero lo contrario no tiene porque ser cierto, por ejemplo una fuerte correlación entre el tamaño de la población de cigüeñas, y los nacimientos de varones para iguales períodos de tiempo en un determinado país no implica, obviamente una relación causal entre ambas variables, sino a lo sumo una relación de estas con una o varias, no incluída en los datos.

La elección de la función no es siempre evidente, pues como se pone de manifiesto en la figura 1, un mismo conjunto de datos puede ser ajustado de diferentes maneras.

Cuando las variables dependientes no son numéricas, sino que existen categorías tales como "deficiente", "muy deficiente" o "sobresaliente", utilizaremos los modelos **logaritmos-lineales** desarrollados a raíz de los trabajos de Holland y otros desde 1975.

Los métodos de **suavizado exponencial** (Exponential Smoothing), desarrollados a partir de los trabajos iniciales de Brown y Holt en 1950, gozan de una amplia popularidad en el sector industrial, que es donde mejor eco encuentran por lo general las técnicas estadísticas de previsión.

METODOS
DE REGRESION

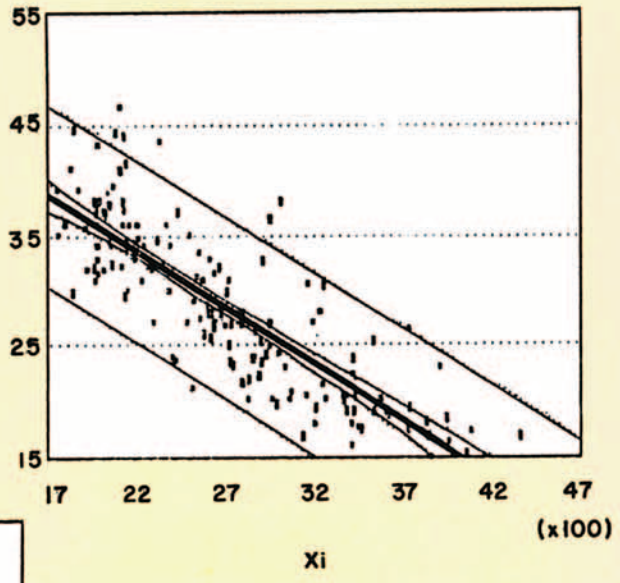
LA FORMA GENERAL DEL MODELO LINEAL ES:

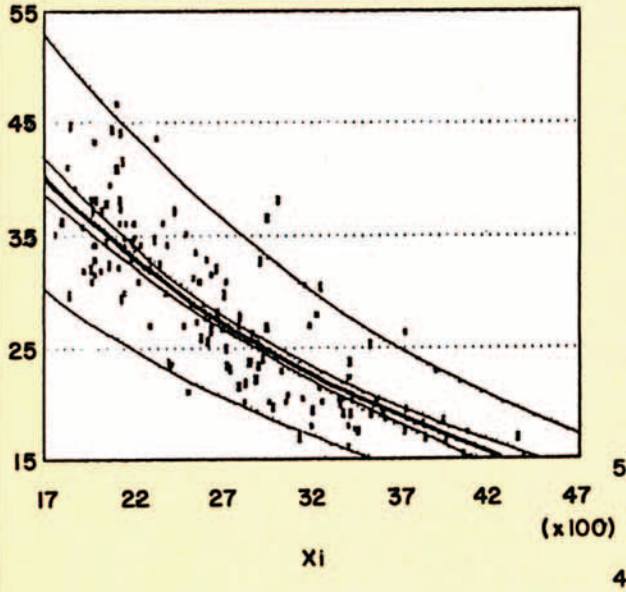
$$Y_i = a + bX_i$$

$a = 55.8971$
$b = -0.0101$

$Y_i = 55.8971 - 0.0101\ X_i$

Coeficiente de Correlación = —0.829
Coeficiente de Determinación = 68.74%





LA FORMA GENERAL DEL MODELO EXPONENCIAL ES:

$$Y_i = e^{(a + bX_i)}$$

$a = 4.333$
$b = -3.77\ E\text{-}4$

$Y_i = e^{(4.33 - 3.77E\text{-}4\ X_i)}$

Coeficiente de Correlación = —0.851
Coeficiente de Determinación = 72,44%

LA FORMA GENERAL DEL MODELO RECIPROCO ES:

$$Y_i = \frac{1}{(a + bX)}$$

$a = 1.92\ E\text{-}3$
$b = 1.46\ E\text{-}5$

$$Y_i = \frac{1}{-1.92E\text{-}3 + 1.46E\text{-}5\ X_i}$$

Coeficiente de Correlación = —0.857
Coeficiente de Determinación = 73.43%

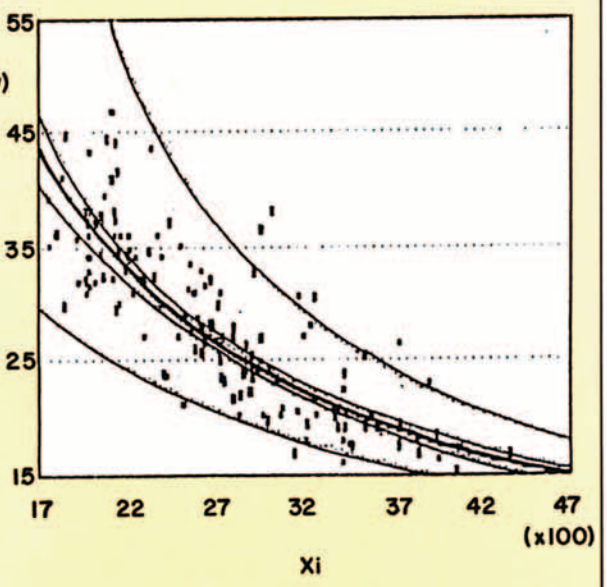


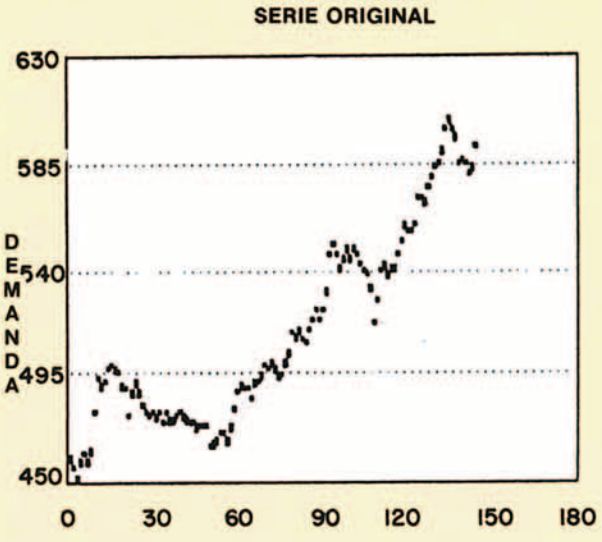
Figura 1

SUAVIZADO EXPONENCIAL

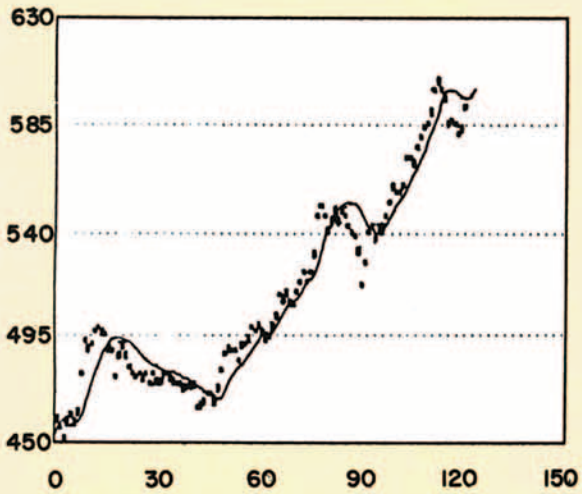
La forma general del modelo es:

$$Zt = f(t,B) + Et$$

donde:
- f(t) es una función del tiempo
- B es el conjunto de parámetros del modelo a estimar
- Et es el error cometido.



MODELO SIMPLE EXPONENCIAL



Un modelo lineal de un solo parámetro no tiene la suficiente flexibilidad para adaptarse a las posibles variaciones de los datos.

Al aumentar el número de parámetros aumenta la dificultad, pero la mayor flexibilidad del modelo mejora sus propiedades predictoras.

MODELO DOBLE EXPONENCIAL

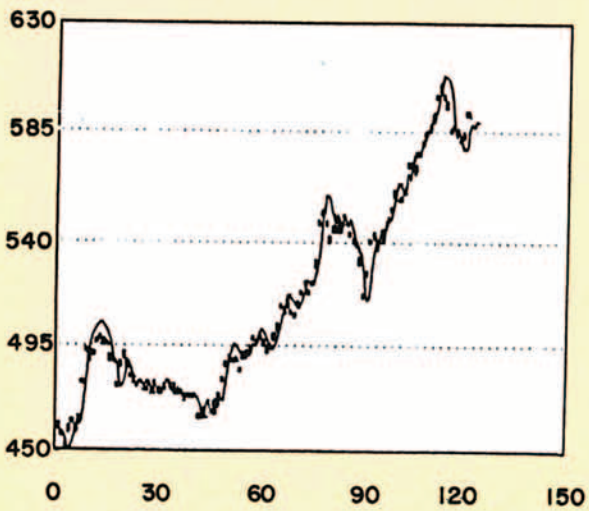


FIGURA 2

Esta popularidad es debida a consideraciones eminentemente prácticas: la formulación de los modelos existentes es relativamente simple; tanto los parámetros como los componentes del modelo tienen un significado intuitivo para el usuario; y por último las necesidades de cálculo y almacenamiento de datos no son demasiado gravosas.

Pero quizá la razón más importante de esta popular aceptación es la sorprendente relación entre la calidad de predicción y el esfuerzo necesario.

Al contrario que el modelo general de regresión, en esta técnica, los datos usados para confeccionar el modelo que prediga el comportamiento futuro de una variable, no son los de otras variables usadas como predictoras, sino los valores anteriores de esa misma variable, a través de un modelo genérico que hace exhaustivo uso de la información contenida en las realizaciones anteriores.

La forma del modelo general es $\mathbf{Zt=f(t,\beta)+\varepsilon t}$ donde $f(t,\beta)$ es una función del tiempo y los coeficientes β, (que será necesario estimar). Este modelo general incluirá dos componentes: el de la tendencia con sus clases: constante, lineal, exponencial, etc.; y el de la estacionalidad con sus clases: no estacionalidad, aditiva, multiplicativa, etc.

Como metodología, el suavizado exponencial, tiene un defecto: la ausencia de procedimientos objetivos para la identificación del modelo, así como para su exacto diagnóstico; sin embargo ofrece, aparte de las ya mencionadas evidentes ventajas, su precisión en horizontes de predicción limitados al futuro más próximo, y de que su facilidad se deriva el hecho de que sea el único método válido para grandes entornos de predicción como puedan serlo los grandes sistemas logísticos de las Fuerzas Armadas.

Sin duda alguna, la induscutible estrella de los métodos de predicción, es la metodología **ARIMA** (Auto Regressive Integrated Moving Average), desarrollada por Box y Jenkins en 1970. El impacto causado por el trabajo de estos dos autores ha sido tal que, en la actualidad los "modelos de Box Jenkins", y las "series temporales", son para la mayoría sinónimos.

Las técnicas mencionadas, tanto la de regresión como el suavizado exponencial, asumen en líneas generales la independencia de las observaciones a lo largo del tiempo, esto es las sucesivas realizaciones de la variable a predecir, no se ven influenciadas por las realizaciones anteriores. Por ejemplo, la demanda de un artículo dentro de un almacén, tendrá un valor el día de hoy, en el que para nada influyen los valores de las demandas en días anteriores. Esta suposición, se ve con frecuencia descartada, y lo más probable es que nos encontremos con variables autocorrelacionadas, en las que los valores anteriores tienen una mayor o menor influencia, que nunca debe ser considerada despreciable.

Ni la regresión, ni el suavizado exponencial, explotan adecuadamente esta frecuente propiedad de las series temporales. El hecho de que el modelo desarrollado por Box y Jenkins saque máximo provecho de esta relación es, sin duda, la causa de su excelente comportamiento en la predicción de valores futuros. Los pasos a seguir en la construcción de un modelo ARIMA son tres:

a) Identificación, fase en la que partiendo del conocimiento de como se relaciona la serie con ella misma, se identifica el grado de cada uno de los parámetros básicos del modelo general, hasta conseguir mediante las transformaciones que sean necesarias una serie estacionaria.

b) Estimación, de cada uno de los parámetros del modelo ya identificado, (generalmente a través del método de mínimos cuadrados, análogo al empleado en la regresión).

c) Validación de cada uno de los parámetros del modelo ya identificado y parametrizado que se somete a prueba, obteniéndose valoraciones objetivas de la confianza que nos merece respecto a otros tipos posibles de modelos.

Como metodología, pocas pegas pueden ponerse a esta de los modelos ARIMA, quizá el hecho de que si bien las herramientas informáticas actuales, ofrecen al estadístico facilidades que hace apenas diez años eran impensables, nunca se conseguirá la "regla automática de parada", esto es el desarrollo de un algoritmo, como conjunto finito de reglas objetivas, que solucione cualquier problema de predicción dentro de la generalidad de los métodos ARIMA, lo que convierte al modelado de series temporales en un proceso altamente interactivo, que algunos no dudan en calificar como auténtico arte.

De su precisión no cabe hacer ningún comentario, pues el modelo genérico puede ser ampliado tanto como se desee, a costa naturalmente del aumento de necesidades de cálculo.

Con frecuencia nuestro interés se centra en el conocimiento de la probabilidad de ocurrencia de un determinado suceso, la estadística ha desarrollado a partir de una axiomática simple, procedimientos analíticos para el cálculo de probabilidades; el análisis matemático y el cálculo combinatorio, forman con frecuencia la base de estos procedimientos, desgraciadamente casi con la misma frecuencia la complejidad del problema a resolver es tal, que estos métodos analíticos son inaplicables.

El conocido como **Método de Montecarlo,** es entonces la única posibilidad factible de obtener una aproximación de suficiente calidad, a través de una serie de acciones para cuyo desarrollo no son necesarios los elevados conocimientos de análisis que pudiera requerir la resolución analítica.

Supongamos por ejemplo una pista de aterrizaje sometida a un bombardeo cuyo objeto es conseguir su interdicción, y también, que deseamos conocer, el número medio de impactos que contendrá la "pista mínima posible", entendiendo como tal, aquella sección rectangular de la pista atacada, cuyas dimensiones son suficientes para permitir el despegue de un avión.

Podemos estar más o menos seguros de qué variables intervienen en el problema, los parámetros del lanzamiento, altura, velocidad, número de atacantes; podemos tener toda la información referente al armamento usado y su comportamiento balístico, podemos saberlo todo desde la probabilidad de que se

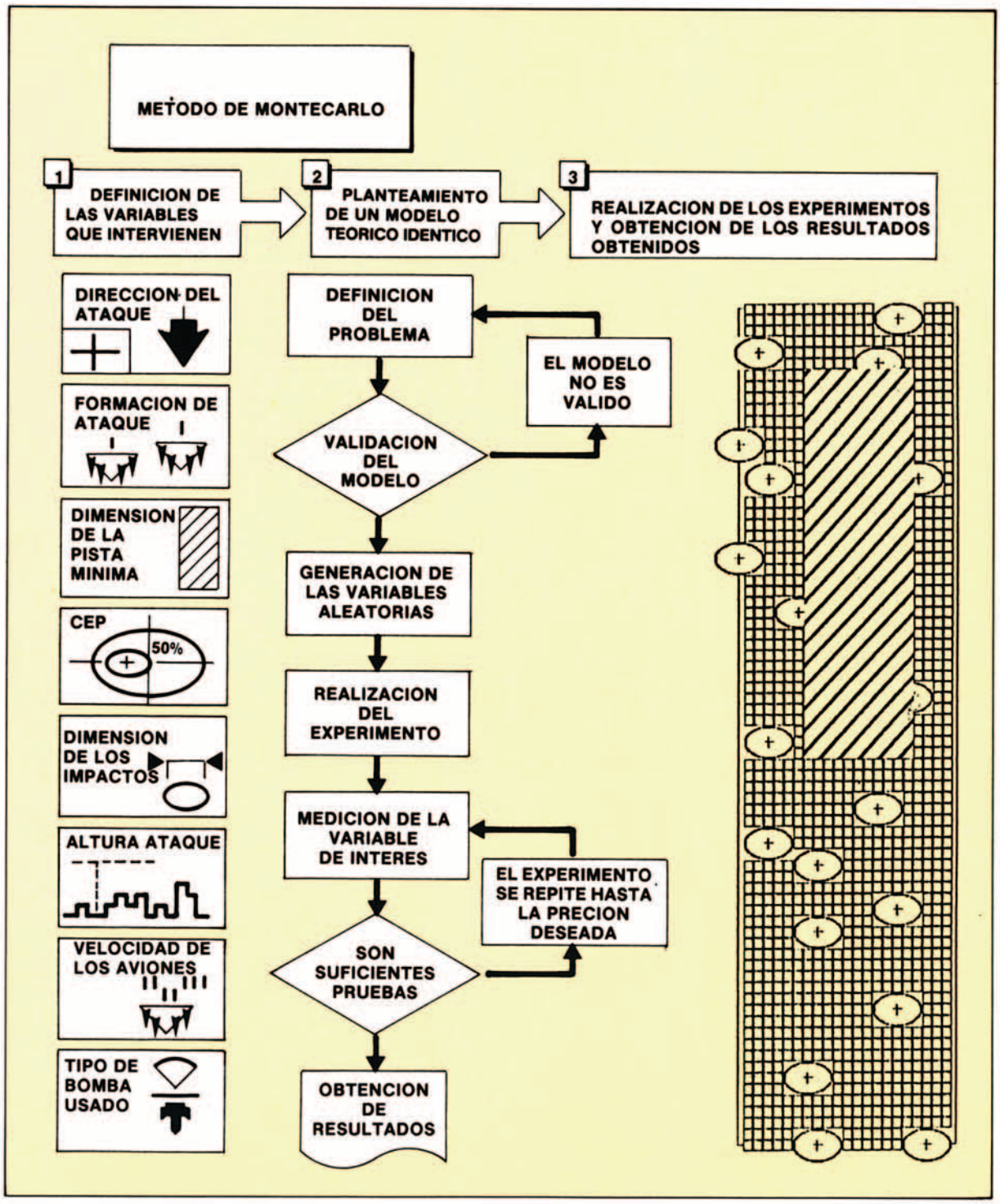


Figura 3

abra el paracaídas de la bomba, hasta el CEP (Circular Error Probable), del piloto que probablemente realizará el bombardeo.

Sin embargo conjugar todas estas variables, sus posibles combinaciones, sus complejas interacciones, y finalmente derivar una fórmula kilométrica que conteste nuestra pregunta requerirá también otros dos elementos: una fe ciega en las matemáticas por parte del usuario final, y el concurso de un genio matemático eficazmente vacunado contra el desánimo, sin contar el gasto extra de papel que posiblemente fuera necesario para escribirla.

Cuando en 1945, dos eminentes científicos que empeñaban sus esfuerzos en el desarrollo de la primera

bomba atómica, encontraron un problema de quizá no menos complejidad que el anterior, no fueron capaces de encontrar ni lo uno ni lo otro. Fue entonces cuando desarrollaron los principios básicos de lo que más tarde sería conocido como el método de Montecarlo.

Dos peculiaridades distinguen la resolución de este tipo de problemas: la primera consiste en que su algoritmo tiene una estructura muy sencilla, como regla se elabora primero un programa de ordenador, que simule la realización de un fenómeno cuya base probabilística subyacente sea idéntica al del problema bajo estudio, en nuestro ejemplo bastaría, una vez el sistema estuviera correctamente definido, la generación de las variables aleatorias pertinentes cuya distribución ha de ser escrupulosamente identificada, estas serían tal vez las correspondientes a la caída de las bombas, las variaciones aleatorias de rumbo, altura, etc., después un sencillo cálculo permitiría encontrar la pista mínima y los impactos que ha recibido, bastaría entonces repetir el experimento un número suficientemente grande de veces, generando en cada ocasión valores distintos de cada una de las variables.

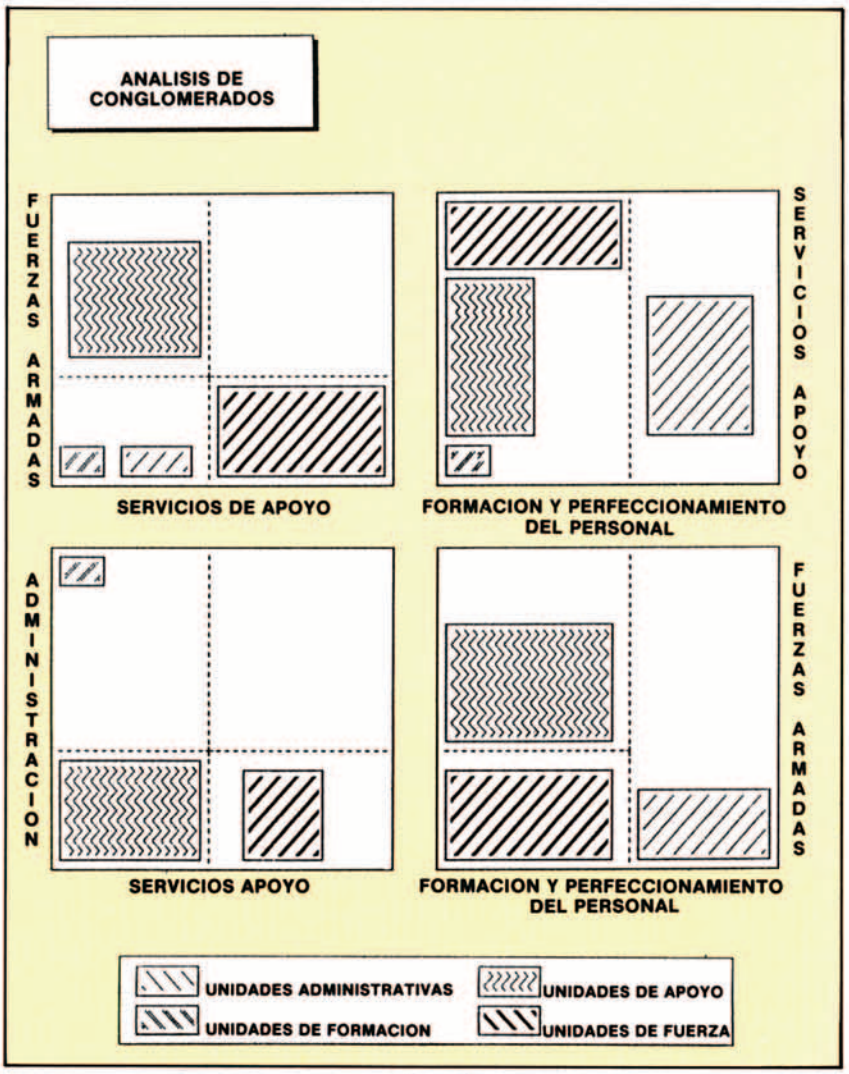


Figura 4

La segunda particularidad consiste en que el error es inversamente proporcional al número de realizaciones del experimento, en una proporción de forma cuadrática, la precisión obtenida puede ser cualquiera, pero el número de pruebas necesarias, y por tanto el tiempo de resolución crece de manera a veces insufrible.

Los modernos y cada vez más rápidos ordenadores han permitido el afianzamiento del método; sólo un potente ordenador es capaz de repetir experimentos de tipo complejo, en un tiempo razonable, aunque la precisión deseada exija repeticiones de millones de veces.

La última técnica que citaremos está relacionada con las tareas que implican una taxonomía numérica esto es la clasificación de individuos en grupos homogéneos.

Este se conoce por **análisis de conglomerados** o en inglés "Cluster Analysis".

Supongamos, por ejemplo que queremos obtener una clasificación de las unidades de Ejército del Aire, que esté basada en los importes de los créditos gastados por cada unidad, para el desarrollo de cada una de las funciones propias de las Fuerzas Armadas, según la clasificación presupuestaria vigente (Administración, Fuerza, Potenciación, Apoyo y Formación).

El análisis de conglomerados nos establece, en función de la similaridad-disimilaridad entre los gastos de las unidades, una clasificación que las agrupa en grupos homogéneos respecto a los importes gastados.

En la figura 4, se observa que a grupos distintos le corresponden, como es natural grupos distintos, obteniéndose una clasificación en cuatro grupos correspondientes a las variables más significativas: Fuerza, Apoyo, Administración, Formación.

La importancia, cada día más afianzada, de la estadística en todas las ramas de la ciencia, no hace sino confirmarla como eficaz instrumento de decisión e investigación, superando el tradicional papel de técnica meramente descriptiva, para convertirse poco a poco en una técnica plenamente operativa en la que poder confiar plenamente como soporte para la toma de, cada vez, más complejas decisiones. ■