Introducción al análisis exploratorio de Datos Multidimensionales

JAIME CURTS G.[1 2]
LEONARDO ALCANTARA L.[2]
XAVIER CHIAPPA C.[3]

INTRODUCCION

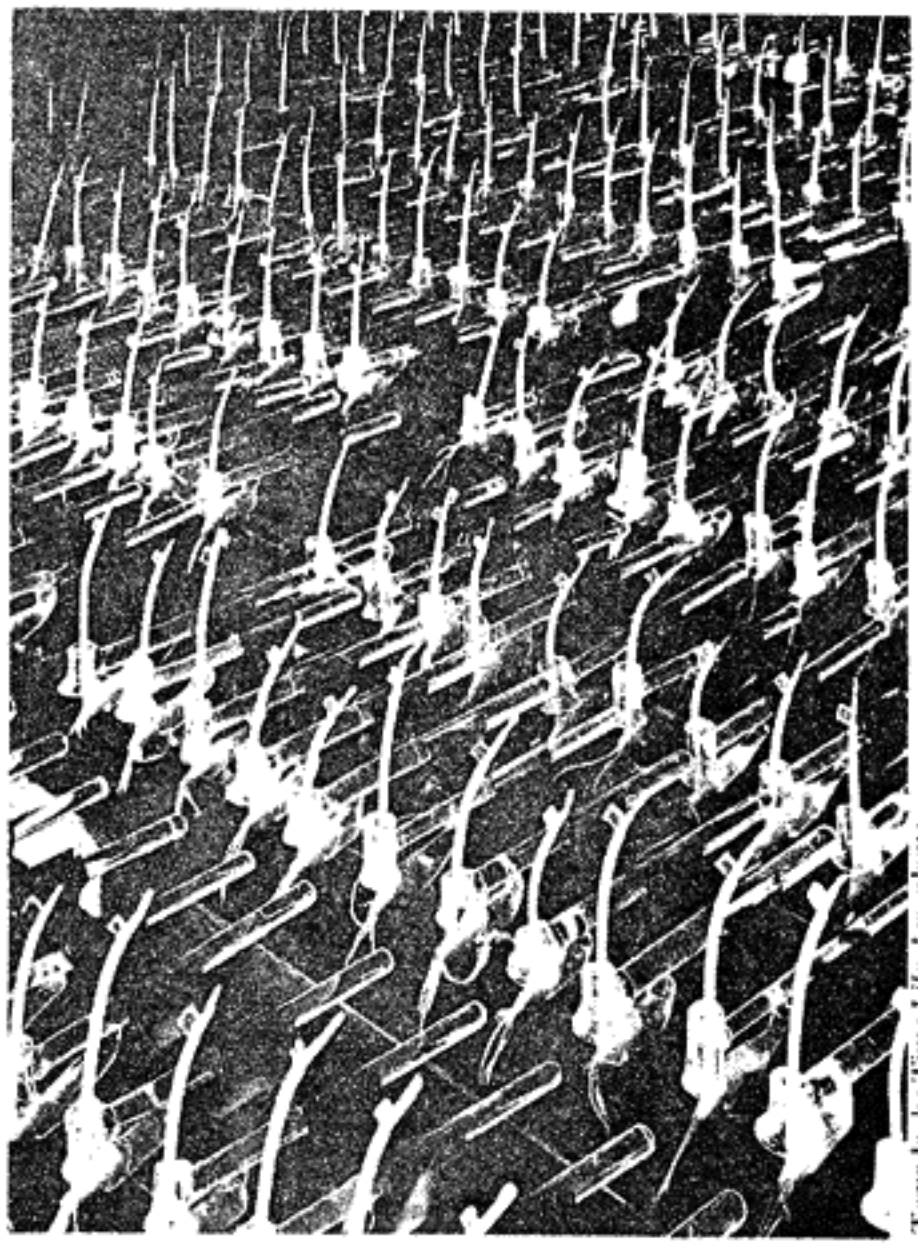
Un viejo proverbio chino dice sabiamente: "una imagen vale más que mil palabras" y pensamos que probablemente, también, más que cien estadísticos. Este artículo se desarrolla en ese sentido y, en efecto, no existe una herramienta estadística más poderosa que una gráfica bien seleccionada. Lo anterior tal vez cause asombro entre algunos de los lectores; sin embargo dicha actitud puede ser explicada al considerar erróneamente como verdad la siguiente ecuación: estadística = análisis de datos. Si bien "análisis de datos" significa la descomposición de los datos en sus componentes principales ($Y = \hat{Y} + e$; con Y = dato observado, $\hat{Y}$ = dato esperado y e = residuo) erróneamente éste análisis sólo se asocia al cómputo de resúmenes numéricos, omitiendo otros métodos de estudio o comparación.

La idea anterior implica degradar la importancia que tiene la visualización de datos, haciendo creer que el cómputo de un estadístico (como la media aritmética o un coeficiente de regresión) es mucho más "sólido" o "robusto" que la representación gráfica de los datos.

La importancia de graficar los datos antes de someterlos a un análisis estadístico ha sido ampliamente discutida en la literatura (Anscombe, 1973). Asimismo se ha subrayado la necesidad de evaluar por métodos gráficos la bondad de ajuste de los modelos estadísticos lineales (Curts, 1984).

En los últimos años se han desarrollado diversas técnicas gráficas para el análisis de datos unidimensionales, bidimensionales o multidimensionales. Cada una tiene sus bondades y limitaciones, pero en su conjunto constituyen herramientas muy poderosas para el analista de datos. Para el presente texto se escogió un problema sobre morfometría botánica a fin de divulgar algunas técnicas gráfico-exploratorias. Los datos del problema, colectados por el botánico Edgar Anderson, corres-

1. *Facultad de Ciencias. Depto. de Biología. Posgrado. UNAM*
2. *Escuela Nacional de Estudios Profesionales Iztacala. UNAM*
3. *Instituto de Ciencias del Mar y Limnología. UNAM*



Tomado de: Time Life, Las plantas.

Estudio del crecimiento en plantas. El análisis de los resultados de experimentos con sistemas vivos, puede ser muy complicado debido a las muchas variables involucradas. Esto hace que sea indispensable el conocimiento de técnicas estadísticas.

ponden a medidas de largo y ancho de sépalos y pétalos de tres especies del género *Iris* y fueron utilizados originalmente por Fisher (1936) en su ya clásico trabajo sobre la aplicación de distancias euclidianas en problemas taxonómicos. Vale la pena comentar que dicho trabajo, ampliamente citado en la literatura, generalizó el uso del denominado "análisis discriminante".

EL HISTOGRAMA: UN ENFOQUE CRITICO

El arte de exhibir datos es un tópico contemporáneo de considerable interés que ha ocupado la mente de algunos doctos desde tiempo atrás. Como en otras ramas de la Ciencia, la representación gráfica de datos es producto del avance en el conocimiento y la tecnología. Hoy día esta función es un hecho cotidiano. La disponibilidad de equipos de cómputo, periféricos auxiliares (impresoras, graficadoras, etc.), han permitido que el análisis gráfico y numérico sea llevado a cabo de manera rutinaria.

Sin embargo, lo rutinario no puede sustituir al sentido común o a la reflexión científica. Así, aunque parezca obvio e ineluso natural, el uso que desde hace más de doscientos años se le ha dado al histograma, ha sido presentar a la información numérica en forma resumida. De esta costumbre hay que ser lo suficientemente escépticos, pues aún incorporándole tecnología moderna (microcomputación con alta resolución gráfica), los resultados obtenidos dejan mucho que desear.

Con el objeto de dibujar histogramas rápida y automáticamente, se construyó un programa de alta resolución gráfica para una microcomputadora Apple. Dicho programa, adaptado de Korites (1982), construye el histograma de un lote de datos al especificarle el tamaño del intervalo de clase (TIC).

Para ilustrar la debilidad del histograma como instrumento de visualización, se escogieron los datos correspondientes a la longitud del sépalo de *Iris virginica* (Tabla 1) construyéndose, con ayuda del programa de cómputo anteriormente descrito, los cinco histogramas que se ilustran en la Figura 1. Cada histograma se elaboró incrementando el TIC desde 0.1 cm a 0.5 cm. El efecto de variar el TIC produce resultados bastante contrastantes. Se observa que el histograma con un TIC = 0.1 cm posee muchos "huecos" en su distribución (señalados con flechas en la figura), los cuales no aparecen en los otros histogramas. Asimismo se detecta claramente una barra muy aislada hacia la izquierda, que en los otros histogramas va incorporándose a medida que aumenta el intervalo de clase. Lo anterior es sumamente impactante pues se pasa de una distribución "rugosa" a una "lisa". En esta última se encubre un caso extremo potencial y su aspecto indicaría, de forma errónea, que los datos provienen de una distribución Gaussiana. Es evidente, a través de éste ejemplo, que el histograma es un "resumen visual de datos" poco confiable debido a la selección arbitraria del número y colocación del intervalo de clase. Del mismo modo la posición de las esquinas de las barras constituye un artificio en la construcción del histograma.

TABLA 1

Datos correspondientes a largo y ancho de sépalo y pétalo de tres especies del género *Iris*

Iris setosa				*Iris versicolor*				*Iris virginica*			
Sépalo		Pétalo		Sépalo		Pétalo		Sépalo		Pétalo	
l	a	l	a	l	a	l	a	l	a	l	a
5.1	3.5	1.4	0.2	7.0	3.2	4.7	1.4	6.3	3.3	6.0	2.5
4.9	3.0	1.4	0.2	6.4	3.2	4.5	1.5	5.8	2.7	5.1	1.9
4.7	3.2	1.3	0.2	6.9	2.3	4.9	1.5	7.1	3.0	5.9	2.1
4.6	3.1	1.5	0.2	5.5	2.8	4.0	1.3	6.3	2.9	6.6	1.8
5.0	3.6	1.4	0.2	6.5	2.8	4.6	1.5	6.5	3.0	5.8	2.2
5.4	3.9	1.7	0.4	5.7	2.8	4.5	1.3	7.6	3.0	6.6	2.1
4.6	3.4	1.4	0.3	6.3	3.3	4.7	1.6	4.9	2.5	4.5	1.7
5.0	3.4	1.5	0.2	4.9	2.4	3.3	1.0	7.3	2.9	6.3	1.8
4.4	2.9	1.4	0.2	6.6	2.9	4.6	1.3	6.7	2.5	5.8	1.8
4.9	3.1	1.5	0.1	5.2	2.7	3.9	1.4	7.2	3.6	6.1	2.5
5.4	3.7	1.5	0.2	5.0	2.0	3.5	1.0	6.5	3.2	5.1	2.0
4.8	3.4	1.6	0.2	5.9	3.0	4.2	1.5	6.4	2.7	5.3	1.9
4.8	3.0	1.4	0.1	6.0	2.2	4.0	1.0	6.8	3.0	5.5	2.1
4.3	3.0	1.1	0.1	6.1	2.9	4.7	1.4	5.7	2.5	5.0	2.0
5.8	4.0	1.2	0.2	5.6	2.9	3.6	1.3	5.8	2.8	5.1	2.4
5.7	4.4	1.5	0.4	6.7	3.1	4.4	1.4	6.4	3.2	5.3	2.3
5.4	3.9	1.3	0.4	5.6	3.0	4.5	1.5	6.5	3.0	5.5	1.8
5.1	3.5	1.4	0.3	5.8	2.7	4.1	1.0	7.7	3.8	6.7	2.2
5.7	3.8	1.7	0.3	6.2	2.2	4.5	1.5	7.7	2.6	6.9	2.3
5.1	3.8	1.5	0.3	5.6	2.5	3.9	1.1	6.0	2.2	5.0	1.5
5.4	3.4	1.7	0.2	5.9	3.2	4.8	1.8	6.9	3.2	5.7	2.3
5.1	3.7	1.5	0.4	6.1	2.8	4.0	1.3	5.6	2.8	4.9	2.0
4.6	3.6	1.0	0.2	6.3	2.5	4.9	1.5	7.7	2.8	6.7	2.0
5.1	3.3	1.7	0.5	6.1	2.8	4.7	1.2	6.3	2.7	4.9	1.8
4.8	3.4	1.9	0.2	6.4	2.9	4.3	1.3	6.7	3.3	5.7	2.1
5.0	3.0	1.6	0.2	6.6	3.0	4.4	1.4	7.2	3.2	6.0	1.8
5.0	3.4	1.6	0.4	6.8	2.8	4.8	1.4	6.2	2.8	4.8	1.8
5.2	3.5	1.5	0.2	6.7	3.0	5.0	1.7	6.1	3.0	4.9	1.8
5.2	3.4	1.4	0.2	6.0	2.9	4.5	1.5	6.4	2.8	5.6	2.1
4.7	3.2	1.6	0.2	5.7	2.6	3.5	1.0	7.2	3.0	5.8	1.6
4.8	3.1	1.6	0.2	5.5	2.4	3.8	1.1	7.4	2.8	6.1	1.9
5.4	3.4	1.5	0.4	5.5	2.4	3.7	1.0	7.9	3.8	6.4	2.0
5.2	4.1	1.5	0.1	5.8	2.7	3.9	1.2	6.4	2.8	5.6	2.2
5.5	4.2	1.4	0.2	6.0	2.7	5.1	1.6	6.3	2.8	5.1	1.5
4.9	3.1	1.5	0.2	5.4	3.0	4.5	1.5	6.1	2.6	5.6	1.4
5.0	3.2	1.2	0.2	6.0	3.4	4.5	1.6	7.7	3.0	6.1	2.3
5.5	3.5	1.3	0.2	6.7	3.1	4.7	1.5	6.3	3.4	5.6	2.4
4.9	3.6	1.4	0.1	6.3	2.3	4.4	1.3	6.4	3.1	5.5	1.8
4.4	3.0	1.3	0.2	5.6	3.0	4.1	1.3	6.0	3.0	4.8	1.8
5.1	3.4	1.5	0.2	5.5	2.5	4.0	1.3	6.9	3.1	5.4	2.1
5.0	3.5	1.3	0.3	5.5	2.6	4.4	1.2	6.7	3.1	5.6	2.4
4.5	2.3	1.3	0.3	6.1	3.0	4.6	1.4	6.9	3.1	5.1	2.3
4.4	3.2	1.3	0.2	5.8	2.6	4.0	1.2	5.8	2.7	5.1	1.9
5.0	3.5	1.6	0.6	5.0	2.3	3.3	1.0	6.8	3.2	5.9	2.3
5.1	3.8	1.9	0.4	5.6	2.7	4.2	1.3	6.7	3.3	5.7	2.5
4.8	3.0	1.4	0.3	5.7	3.0	4.2	1.2	6.7	3.0	5.2	2.3
5.1	3.8	1.6	0.2	5.7	2.9	4.2	1.3	6.3	2.5	5.0	1.9
4.6	3.2	1.4	0.2	6.2	2.9	4.3	1.3	6.5	3.0	5.2	2.0
5.3	3.7	1.5	0.2	5.1	2.5	3.0	1.1	6.2	3.4	5.4	2.3
5.0	3.3	1.4	0.2	5.7	2.8	4.1	1.3	5.9	3.0	5.1	1.8

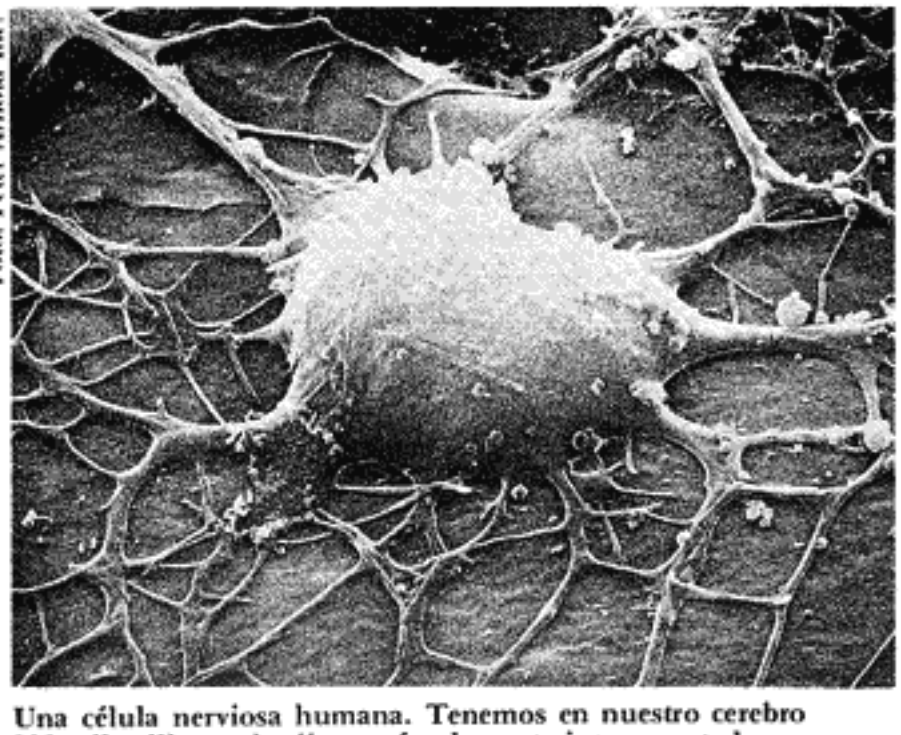
l = Largo (cm) a = Ancho (cm)

En síntesis la visualización de datos no tiene como objeto encubrir o crear falsas impresiones sobre su distribución, sino resaltar con claridad el verdadero patrón de los mismos. Debe quedar claro que el análisis gráfico de datos no está sujeto a "una cuestión de enfoques"; por el contrario, debe ser el retrato fiel de la información cuantitativa que éstos contienen.

EL DIAGRAMA DE TALLO-Y-HOJA

Con el objeto de comunicar simultáneamente los valores numé-

Foto: Peter Arnold Inc.



Una célula nerviosa humana. Tenemos en nuestro cerebro 100 mil millones de ellas profundamente interconectadas entre sí. Dentro de ciertos enfoques, el estudio de este sistema sin estadística es simplemente inadmisible.

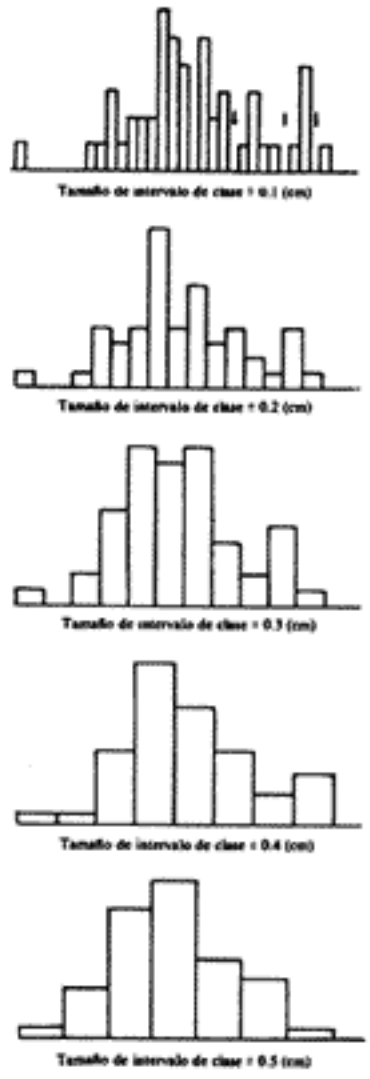


FIGURA 1

Histogramas de la longitud del sépalo de *Iris virginica* variando el tamaño de intervalo de clase. Nótese cómo a medida que se aumenta el intervalo de clase, la forma de la distribución va cambiando de manera drástica.

ricos de un lote de datos con la forma de su distribución, John Tukey (1977) ideó el diagrama de tallo-y-hoja, como un híbrido donde se combinan los aspectos visuales de un histograma con la información numérica que proporciona una tabla de distribución de frecuencias. Su uso y construcción están ampliamente explicados en Curts (1986). De acuerdo a los propósitos de este artículo, los datos referentes a la longitud del sépalo de *Iris virginica* se utilizarán nuevamente para ilustrar la construcción del diagrama de tallo-y-hoja y tener un marco de comparación con los histogramas obtenidos en la sección anterior.

Para obtener un diagrama de tallo-y-hoja de un lote de datos, se ordena ese último en mágnitud creciente y se localiza la mediana. En el ejemplo la mediana (M) del lote de datos ordenados es M = 6.4 cm y el rango está definido por el intervalo (4.9 cm, 7.9 cm). Para formar el tallo y sus hojas se escogen dígitos que permitan fraccionar en dos partes el lote de datos. Por ejemplo, el valor más pequeño del lote corresponde al valor de 4.9 cm y puede ser fraccionado de la siguiente forma:

Valor del dato	*Fracción*	*Tallo*	*Hoja*
4.9	4/9	4	9

A partir del ejemplo anterior se construye el tallo escribiendo verticalmente los dígitos enteros entre 4 y 7, asociando a cada uno su hoja respectiva. Para los primeros tres valores tenemos el siguiente diagrama:

Valor del dato	*Fracción*	*Tallo*	*Hoja*
4.9	4|9	4	9
5.6, 5.7	5/6 , 5/7	5	6 7

El diagrama completo se muestra en la Figura 2a. Nótese que el diagrama contiene la siguiente información adicional. Primero tiene un recordatorio que las unidades del diagrama son 0.1 cm —es decir, que un 4 y un 9 deben leerse como 4.9—; posteriormente la figura incluye una columna de números hacia el extremo izquierdo y claramente se aprecia al número 31 entre paréntesis (31). Este número indica que la mediana del lote de datos, M = 6.4 cm, se localiza en esa línea. Por lo tanto el número entre paréntesis, (31), indica que en esa línea se encuentra el centro de la distribución y que, además, contiene 31 hojas. A partir del número entre paréntesis, el resto de ellos indica las frecuencias acumuladas por línea. Estas frecuencias se cuentan del extremo superior hacia la última línea que no contiene la mediana y viceversa. Por ejemplo, en la Figura 2A se observa que el primer valor del tallo (4) tiene una sola hoja (9) y el segundo (5) posee seis hojas (6,7,8,8,8,9); por tanto la frecuencia acumulada hasta ese punto es de siete.

Como se puede apreciar, la Figura 2a es demasiado tupida, mostrando muchas hojas por línea. Una forma efectiva de romper el amontonamiento consiste en "alargar el tallo", duplicando los dígitos de la siguiente forma:

```
4*
4.
5*
5.
6*
6.
7*
7.
```

```
UNIDAD = 0.1

  1       4  9
  7       5  678889
(31)      6  00112233333344444555577777788999
 12       7  122234677779
```

Figura 2b

```
UNIDAD = 0.1

          4*
  1        .  9
  1       5*
  7        .  678889
(17)      6*  00112233333344444
 26        .  55557777788999
 12       7*  122234
  6        .  677779
```

Figura 2c

```
UNIDAD = 0.1

  1       4.  9
  1       5*
  1        t
  1        f
  3        s  67
  7       5.  8889
 11       6*  0011
 19        t  22333333
 (9)       f  444445555
 22        s  77777
 17       6.  88999
 12       7*  1
 11        t  2223
  7        f  4
  6        s  67777
  1       7.  9
```

FIGURA 2

Diagramas de tallo y hoja para la longitud de sépalos de *Iris virginica*
a) Notación normal
b) Notación estrella-punto
c) Notación estrella, t,f,s, punto
Nótese que el cambio de una notación a otra equivale a pasar de una "lente de aumento de menor resolución" a "una de mayor resolución".

TABLA II

Ecuaciones de regresión lineal resistente para cada una de las relaciones estudiadas para el género *Iris*

Iris setosa	*Iris versicolor*	*Iris virginica*
Largo de sépalo contra ancho de sépalo		
$\hat{y} = -0.325 + 0.75\ x$	$\hat{y} = 0.847 + 0.32\ x$	$\hat{y} = 1.258 + 0.26\ x$
Largo de sépalo contra largo de pétalo		
$\hat{y} = 0.545 + 0.18\ x$	$\hat{y} = 0.745 + 0.59\ x$	$\hat{y} = 1.312 + 0.65\ x$
Largo de sépalo contra ancho de pétalo		
$\hat{y} = 0.2$	$\hat{y} = 0.486 + 0.14\ x$	$\hat{y} = 0.375 + 0.25\ x$
Ancho de sépalo contra largo de pétalo		
$\hat{y} = 1.17 + 0.08\ x$	$\hat{y} = 1.45 + 1.0\ x$	$\hat{y} = 2.875 + 0.87\ x$
Ancho de sépalo contra ancho de pétalo		
$\hat{y} = 0.2$	$\hat{y} = -0.1 + 0.5\ x$	$\hat{y} = 0.1 + 0.67\ x$
Largo de pétalo contra ancho de pétalo		
$\hat{y} = 0.2$	$\hat{y} = -0.196 + 0.36\ x$	$\hat{y} = 0.778 + 0.22\ x$

Posteriormente, las hojas cuyos valores estén dentro del intervalo (0,4) se colocan en la línea que contenga la estrella (*). Las hojas cuyos valores estén en el intervalo (5,9) se colocan en la línea que contenga el punto (.).

Siguiendo la notación de "estrella y punto", el diagrama completo del ejemplo anterior se muestra en la Figura 2b. Otra variante al diagrama consiste en la notación "estrella, t, f, s, punto", en la cual los intervalos son, respectivamente, (0,1), (2,3), (4,5), (6,7), (8,9). El diagrama con esta notación se muestra en la Figura 2c. Nótese que en esta figura se pueden apreciar detalles que el histograma hubiera encubierto. Claramente se señala la presencia y frecuencia del caso extremo, la forma de la distribución y el tipo de simetría. Otra variante del diagrama de tallo-y-hoja está próxima a publicarse (Curts & Romberg, 1987).

Cabe señalar que las distintas versiones del diagrama (Figuras 2a, b y c) proveen la flexibilidad requerida para determinar tanto el número de líneas en el diagrama como un método para manejar los casos que marcadamente se salen de la distribución. Estas versiones no son análogas a los incrementos de tamaño de intervalo de clase del histograma que se discutió, pues el elemento de construcción y ordenamiento utilizado en todas las variantes del diagrama son los propios dígitos que contiene el lote de datos.

En forma global, todos los diagramas correspondientes a los largos y anchos de sépalos y pétalos para las tres especies del género *Iris* se contemplan en la Figura 3. En cada uno de ellos es posible señalar su unidad de lectura, determinar dónde se concentra la mayoría de los datos, describir la simetría del lote, identificar la presencia de "huecos" y casos extremos. En relación a estos últimos resulta evidente que, tanto en el diagrama de longitud de sépalo de *I. virginica* como en el ancho de sépalo de *I. setosa* se detecta la presencia de casos extremos. En ambos casos ha de investigarse la causa de su existencia (error de medición, error de transcripción, etc.). Sin embargo, en el primero (*I. virginica*) puede sospecharse que no se trata de un caso extremo, sólo la falta de valores entre el rango (5.1 cm, 5.4 cm), los cuales producen un gran "hueco". En el segundo caso (*I. setosa*), también se avizora una situación parecida a la anterior, probablemente debida a un error de transcripción. Asimismo no se debe descartar la posibilidad de errores de medición en ambos casos.

Nótese que la mayoría de los diagramas mostrados en la Figura 3 son asimétricos, sospechándose que en muchos de ellos se llevaron a cabo muestreos defectuosos. Una posible excepción está constituida por la longitud del sépalo de *Iris setosa*. En este caso puede advertirse que existe proporcionalidad entre las frecuencias acumuladas que giran alrededor de la mediana.

Observar cómo se comportan las frecuencias acumuladas alrededor de la mediana resulta un buen indicador exploratorio en torno a qué tan cerca o lejos se está de una distribución simétrica. La distribución Normal o Gaussiana representa un caso particular de la simetría, la cual es asumida como condición en la mayoría de los análisis estadísticos. Algunos de estos métodos toleran o son "robustos" con respecto a ligeras desviaciones de la normalidad, pero difícilmente este criterio tendría aplicación al problema aquí estudiado. Lo anterior no implica que no sea posible seguir adelante tan sólo porque "no se pueda aplicar la estadística tradicional". Todo lo contrario, se debe asumir una actitud abierta y explorar la factibilidad de que existan otras relaciones interesantes entre las variables. La sección que continúa presenta una forma novedosa de explorar relaciones multidimensionales.

EL DIAGRAMA DE ESCALERA

El invento del sistema de coordenadas, atribuido a René Descartes (1595-1650), generalizó el uso del espacio bidimensional y tridimensional para estudiar relaciones de naturaleza cuantitativa. Las palabras espacio-tiempo y la cuarta dimensión son hoy familiares, no obstante es útil mencionar que mucho antes de Einstein los matemáticos y los físicos estrecharon su imaginación para trabajar con cualquier número de dimensiones. En el campo de la biología el plano multidimensional generalizó el marco de trabajo del análisis morfométrico y el de la taxonomía numérica.

Cabe señalar que los estudios de los cuales se derivan datos para ser estudiados en planos multidimensionales, se generan en nuestro mundo real de tres dimensiones. Y aunque mucho tiempo atrás los espacios multidimensionales eran cuestiones de teoría abstracta, debemos comprender que, por ejemplo, el estudio de las cuatro dimensiones lineales del problema aquí tratado (longitud y ancho de sépalos y pétalos del género *Iris*) no provienen de un espacio imaginario o abstracto, sino de nuestro propio mundo tridimensional.

Con lo anterior en mente es factible elaborar bosquejos bidimensionales de planos multidimensionales para estudios que contemplan un sinnúmero de variables dimensionales. La idea anterior se ejemplifica suponiendo la colocación de las seis tapas que componen un cubo en un solo plano. Equivale a estudiar, como en la novela de Abbott (1976), "espacilandia en planilandia".

El bosquejo dimensional, atendiendo las cuatro variables de este estudio (longitud y ancho de pétalos y sépalos del género *Iris*) se ilustra en la Figura 4. Esta figura, tomada de Chambers, Cleveland, Keiner y Tukey (1983), muestra el arreglo que sigue cada una de las gráficas bivariadas. El arreglo en forma de "escalera" está condicionado a que cualquier par de gráficas adyacentes debe compartir un eje en común. Nótese cómo en la Figura 4 el ancho del pétalo (última hilera) comparte su eje vertical con los siguientes ejes horizontales: largo del sépalo, ancho del sépalo y largo del pétalo. De este modo se pueden "barrer" simultáneamente combinaciones de variables y comparar los distintos patrones que se llegan a formar. De igual manera el bosquejo puede generalizarse para cualesquiera de las otras gráficas que compartan un eje común.

Lo más sobresaliente de la Figura 4, que por simplicidad denominaremos "Diagrama de Escalera", son los dos cúmulos de puntos distintivos que se forman en cada gráfica bivariada. Incluso es válido afirmar que en cada una de las gráficas el cúmulo "grande" lo comparten las características de *Iris versicolor* e *Iris virginica*, y el cúmulo pequeño es exclusivo de *Iris setosa*.

El cociente entre dos dimensiones lineales da idea del tamaño y forma de cada especie. Así tenemos que, de las tres especies, *Iris setosa* se caracteriza por ser una flor pequeña con un ancho de sépalo relativamente grande. Por otra parte, cuando se

DIAGRAMAS DE TALLO Y HOJA

Iris setosa
LONGITUD DE SEPALO
UNIDAD = 0.01

1	43	0
4	44	000
5	45	0
9	46	0000
11	47	00
16	48	00000
20	49	0000
(8)	50	00000000
22	51	00000000
14	52	000
11	53	0
10	54	00000
5	55	00
3	56	
3	57	00
1	58	0

Iris versicolor
LONGITUD DE SEPALO
UNIDAD = 0.1

1	4.	9
4	5*	001
5	t	2
11	f	455555
21	s	6666677777
(5)	5.	88899
24	6*	00001111
16	t	22333
11	f	445
8	s	66777
3	6.	89
1	7*	0

Iris virginica
LONGITUD DE SEPALO
UNIDAD = 0.1

1	4.	9
1	5*	
1	t	
1	f	
3	s	67
7	5.	8889
11	6*	0011
19	t	22333333
(9)	f	444445555
22	s	77777
17	6.	88999
12	7*	1
11	t	2223
7	f	4
6	s	67777
1	7.	9

ANCHO DE SEPALO
UNIDAD = 0.1

1	t	3
1	f	
1	s	
2	2.	9
12	3*	0000001111
19	t	2222233
(15)	f	4444444445555555
16	s	666777
10	3.	888899
4	4*	01
2	t	2
1	f	4

ANCHO DE SEPALO
UNIDAD = 0.01

1	20	0
1	21	
3	22	00
6	23	000
9	24	000
13	25	0000
16	26	000
21	27	00000
(6)	28	000000
23	29	0000000
16	30	00000000
8	31	000
5	32	000
2	33	0
1	34	0

ANCHO DE SEPALO
UNIDAD = 0.1

1	t	2
5	f	5555
11	s	667777
21	2.	8888888899
(16)	3*	00000000000001111
13	t	22222333
5	f	44
3	s	6
2	3.	88

LONGITUD DE PETALO
UNIDAD = 0.01

1	10	0
2	11	0
4	12	00
11	13	0000000
24	14	00000000000000
(13)	15	0000000000000
13	16	0000000
6	17	0000
2	18	
2	19	00

LONGITUD DE PETALO
UNIDAD = 0.1

1	3*	0
3	t	33
5	f	55
7	s	67
11	3.	8999
19	4*	00001111
25	t	2222333
25	f	44445555555
14	s	66677777
6	4.	8899
2	5*	01

LONGITUD DE PETALO
UNIDAD = 0.1

1	f	5
1	s	
6	4.	88999
16	5*	0001111111
20	t	2233
25	f	44555
25	s	666777
16	5.	88899
11	6*	00111
6	t	3
5	f	4
4	s	677
1	6.	9

ANCHO DE PETALO
UNIDAD = 0.01

5	1.	00000
(29)	2.	00000000000000000+
16	3.	0000000
9	4.	0000000
2	5.	0
1	6.	0

ANCHO DE PETALO
UNIDAD = 0.01

7	10	0000000
10	11	000
15	12	00000
(13)	13	0000000000000
22	14	0000000
15	15	0000000000
5	16	000
2	17	0
1	18	0

ANCHO DE PETALO
UNIDAD = 0.01

1	14	0
3	15	00
4	16	0
5	17	0
16	18	0000000000
21	19	00000
(6)	20	000000
23	21	000000
17	22	000
14	23	00000000
6	24	000
3	25	000

FIGURA 3

Diagramas de tallo y hoja para cada una de las dimensiones lineales de este estudio (largo y ancho de sépalo y pétalo)

compara la relación entre el ancho y largo de pétalo, se concluye —en una primera aproximación— que tanto para flores pequeñas como para grandes, la forma del pétalo no varía significativamente en su ontogenia.

Es útil señalar que en la gráfica donde se muestra la relación entre el ancho y el largo del pétalo, existe un "hueco" al cual han de atribuírsele varias explicaciones. A nuestro juicio, primero debe procederse a examinar la calidad de la muestra antes de aventurar conclusiones biológicas. En este sentido, si volvemos nuestra atención a la Figura 3 —diagramas de tallo-y-hoja correspondientes al ancho y largo de pétalo de cada especie— notamos el grado de asimetría que poseen. En efecto, en ellos se observan defectos de muestreo, sobre todo en los extremos de las distribuciones. Posiblemente parte del "hueco" de la relación entre ancho y largo del pétalo sea por errores ocurridos durante el muestreo.

En relación al punto señalado con el símbolo "*" en la Figura 4, éste corresponde a un caso extremo (2.3 cm) detectado en el ancho de sépalo de *Iris setosa*. Es posible que se trate de un error de transcripción, porque al invertirse el dígito dejaría de ser caso extremo.

Con la finalidad de conocer el comportamiento individual de cada una de las especies del género *Iris* en el diagrama de escalera, la Figura 5 identifica con tres símbolos distintos a cada una de las especies. Con punto se identifica a *Iris setosa*; con "x" a *Iris versicolor* y con "o" a *Iris virginica*. De este diagrama nuevamente se observa que *Iris setosa* está marcadamente aislada de las otras dos especies; pero también claramente se muestra que entre *Iris versicolor* e *Iris virginica* existe una tendencia a la separación, excepto cuando es considerada la relación entre el largo y ancho del sépalo.

Para describir el crecimiento relativo de las tres especies del género Iris aquí tratadas, se utilizó un algoritmo robusto (Velleman & Hoaglin, 1981) para calcular las ecuaciones de regresión entre las posibles combinaciones de largo y ancho de pétalo y sépalo, definidas por el diagrama de escalera.

El diagrama de escalera que contiene las ecuaciones de regresión están ilustradas en la Figura 6 y describen resumidamente el comportamiento entre las dimensiones consideradas, al eliminar el "ruido" que se observa cuando se contemplan los datos en crudo. La eliminación del "ruido" llega a ser determinante en la interpretación de los datos ya que, por ejemplo, se esperaba que las ecuaciones de regresión para ancho-largo de sépalos, de *Iris versicolor* e *Iris virginica*, fueran similares; sin embargo los resultados de la Tabla 2 muestran todo lo contrario.

Asimismo, las ecuaciones de regresión para las relaciones ancho-largo de pétalo indican que el crecimiento diferencial entre estas dos dimensiones no es igual para las tres especies, como se había sospechado en la Figura 5. Incluso de manera inesperada se ve que esta relación en *Iris setosa* tiene una pendiente de cero. Lo anterior equivaldría a afirmar erróneamente que "el ancho del pétalo en esta especie se mantiene constante durante su ontogenia". De ser así se llegaría a un punto en el que la estructura del pétalo se debilitaría. ¿Cabe señalar entonces que el sesgo de los resultados son producto de la baja calidad del muestreo que se tiene para el ancho del pétalo de *Iris setosa*. Lo anterior es evidente si se observa en la Figura 3 el diagrama de tallo y hoja correspondiente a esta dimensión lineal.

Al contemplar los resultados que producen las Figuras 5 y 6 simultáneamente, refuerzan la idea de que un buen analista de datos siempre asumirá una posición escéptica pero con criterio abierto, para la interpretación de sus datos. Por otra parte, este caso nos ejemplifica que "cantidad no implica calidad".

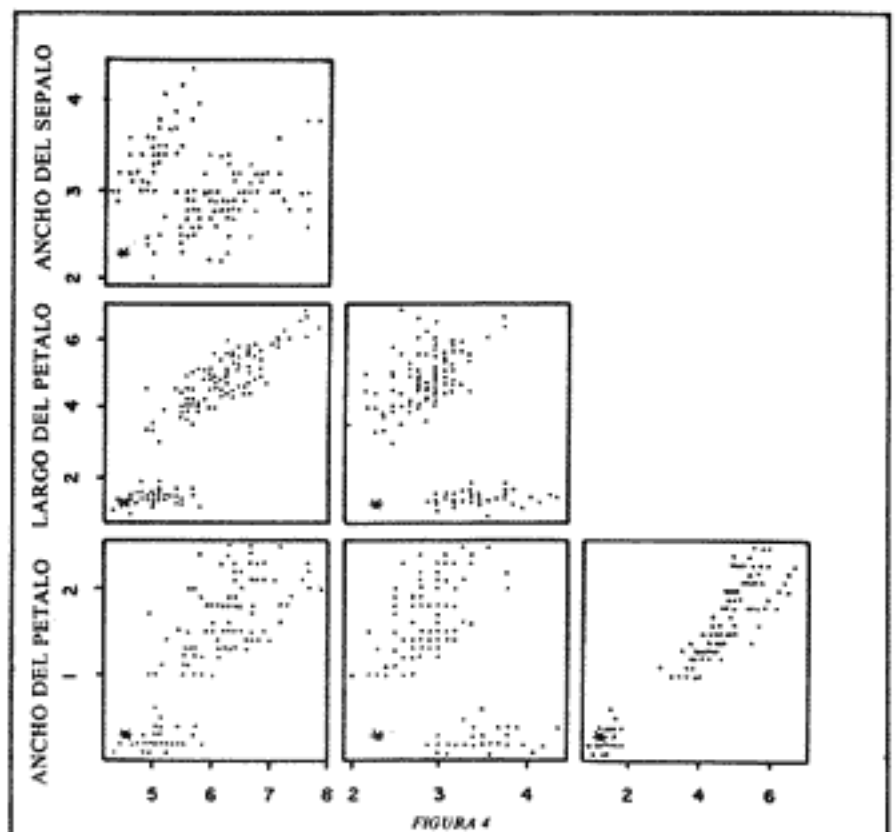
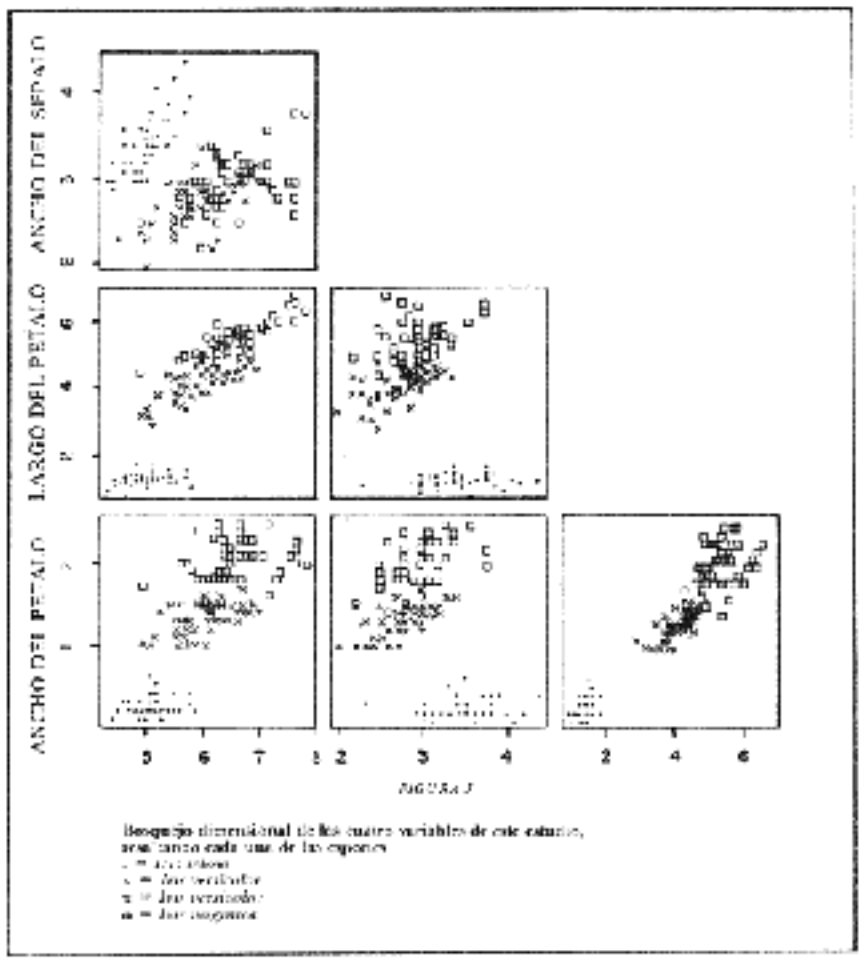


FIGURA 4

Diagrama de escalera del bosquejo dimensional de las cuatro variables de este estudio (longitud y ancho de pétalos y sépalos del género *Iris*).



Bosquejo dimensional de las cuatro variables de este estudio, mostrando cada una de las especies
s = Iris setosa
v = Iris versicolor
g = Iris virginica

COMENTARIOS Y CONCLUSIONES

John Tukey, progenitor moderno de métodos gráficos para el análisis de datos, ha señalado que aunque "la ciencia pueda estar sujeta a múltiples hipótesis", la pictografía de datos nos permite ser sensibles "no sólo a las múltiples hipótesis que sostenemos, sino a muchas otras de las que no hemos pensado, consideradas como improbables o imposibles de pensar" (Tukey, 1974).

Es verdad que parte del análisis de datos está estrechamente vinculada con el problema de la medición, es decir, con el empleo de números para representar propiedades. Sin embargo, con frecuencia nos encontramos que el modo ideal de expresar alguna propiedad de los datos a través de resúmenes numéricos no es la más idónea; tal vez porque dicho resumen numérico no produce la evidencia que se requiere, o si lo es, quizá se necesiten procedimientos prohibitivos, caros o impracticables por un motivo u otro.

La visualización gráfica de datos constituye la base fundamental del análisis estadístico de éstos, ya que permiten al investigador penetrar en la estructura de sus datos sin imponer *a priori* su posiciones probabilísticas sobre el comportamiento de tales datos. Asimismo, son herramientas de "baja sofisticación" que iluminan el camino hacia procedimientos más formales.

Las técnicas de visualización deben ser seleccionadas de tal manera que reflejen adecuadamente a los datos; y aunque éstas significan "lentes de aumento" indispensables para el investigador, no son todos los elementos de juicio que constituyen un análisis estadístico de datos. Toda lente de aumento tiene un límite de resolución. Lo anterior se hizo evidente cuando se compararon las ecuaciones de regresión para ancho de sépalo contra largo de sépalo, de *Iris virginica* e *Iris versicolor* (Fig. 6 y Tabla 2), y los datos crudos para esa misma relación (Fig. 5). Por lo tanto, la interpretación de los datos no debe estar sujeta a una sola forma de análisis. Tal aseveración nos recuerda una de las cartas anónimas publicadas por Jean Rostand (1972), que dice:

> No porque sepa de ranas tiene Usted la capacidad para opinar sobre todo. Las ranas no contestan a todo...

En este artículo se resalta la importancia de seleccionar apropiadamente una técnica de visualización, al comparar las bondades y limitaciones entre el histograma y el diagrama de tallo-y-hoja. El histograma reveló ser una herramienta poco confiable para estudiar la forma de distribución de un lote de datos, ya que con facilidad encubre información o genera falsas espectativas. En cambio se demostró la conveniencia de utilizar al diagrama de tallo-y-hoja como una forma robusta de visualizar la distribución de un lote de datos.

Por otra parte, los bosquejos multidimensionales resumidos en los diagramas de escalera mostraron ser una buena alternativa exploratoria para datos multidimensionales.

BIBLIOGRAFIA

1. Abbot, E. A. (1976) **Planilandia.** Madrid, España. Ediciones Guadarrama.
2. Anscombe, F. J. (1973) Graphs in statistical analysis. **The American Statistician 27.** 17-21.
3. Chamber, J. M., Cleveland, W. S., Kleiner, B. & Tukey, P. A. (1983). **Graphical methods for data analysis**. Belmont, CA: Wadsworth International Group.
4. Curts, J. B. (1984). Introducción al análisis de residuos en biología. **Biótica, 9 (3),** 271-278.
5. Curts, J. B. (1986). El diagrama de tallo y hoja. **Biología, 15** (1-4). 7-12.
6. Curts, J. B. & Romberg, T. A. (1987). **A note on stem and leaf diagrams.** (Sujeto a publicación).
7. Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. **Ann. Eugen. London, 7,** 179-188.
8. Korites, B. J. (1982). **Data plotting software for micros.** Duxbury, MA: Kern Publications.
9. Rostand, J. (1972). **El correo de un biólogo.** Madrid: Editorial Alhambra.
10. Tukey, J. W. (1974). **Mathematics and the picturing of data.** Proc. Int. Congr. Math. Vancouver.
11. Tukey, J. W. (1977). **Exploratory data analysis.** Reading, MA: Addison Wesley Publishing Company.
12. Velleman, P. F. & Hoaglin, D.C. (1981). **Applications, basics and computing exploratory data analysis.** Boston, MA: Duxbury Press.

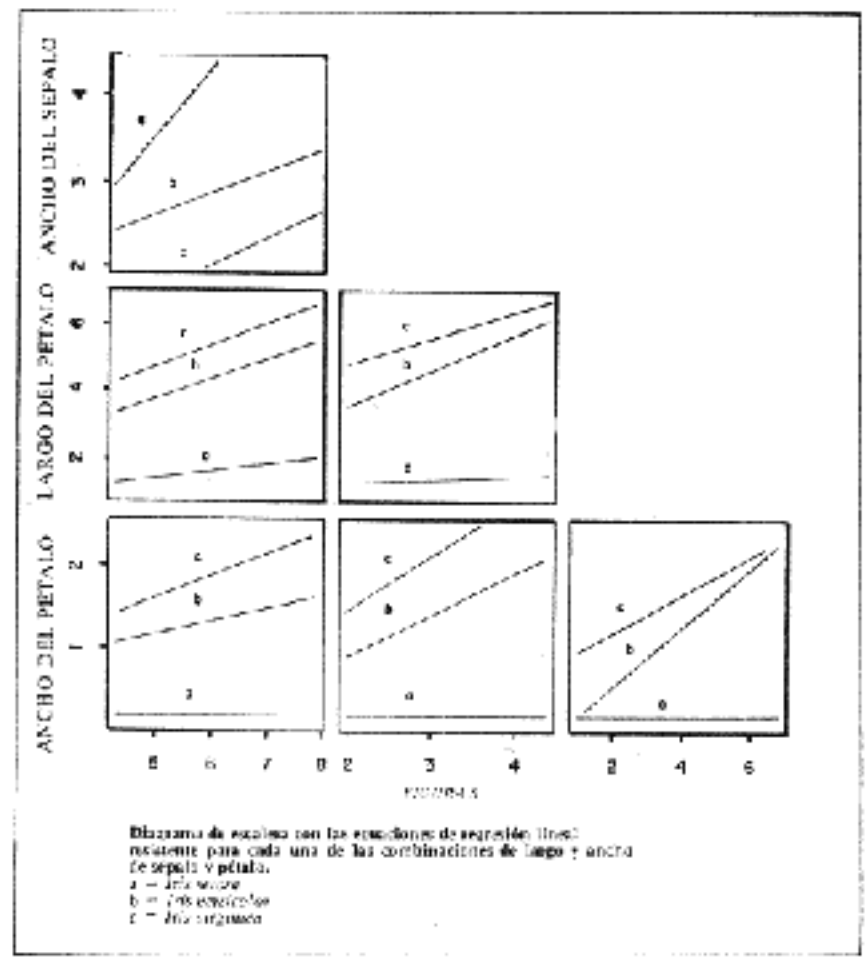


Diagrama de escalera con las ecuaciones de regresión lineal resultante para cada una de las combinaciones de largo y ancho de sépalo y pétalo.
a = Iris setosa
b = Iris versicolor
c = Iris virginica