ANALISIS EXPLORATORIO DE DATOS ECOLOGICOS Y BIOMETRICOS: GRAFICOS STEM-AND-LEAF (TALLO-Y-HOJA) Y BOXPLOT (CAJAS GRAFICAS)

CONDE, J.E.*, RULL, V. y VEGAS, T.****
* **Centro de Investigaciones Marinas, Universidad Nacional Experimental Francisco de Miranda, Venezuela.**
** **Centro de Ecología, Instituto Venezolano de Investigaciones Científicas, Caracas.**

INTRODUCCION

En los últimos años las ciencias biológicas han sufrido un proceso acelerado de cuantificación, pudiendo reconocerse en él cuatro vertientes. En primer lugar, un mayor uso de resultados numéricos para apoyar hipótesis de trabajo. En segundo lugar, un uso más frecuente de técnicas estadísticas básicas (por ejemplo, pruebas t y X^2). Finalmente, en tercer y cuarto lugar, también puede notarse un incremento notable en el número de publicaciones que utilizan técnicas estadísticas avanzadas, y otras donde el fin fundamental es el desarrollo de metodologías estadísticas y matemáticas específicas de la biología.

Esta tendencia hacia un mayor uso de la matemática y la estadística ha sido documentada por Sokal y Rohlf (1981), mediante la realización de una encuesta decenal de los artículos contenidos en la revista *The American Naturalist*, la cual, dada su política editorial multidisciplinaria, puede ser considerada como indicadora de tendencias en la investigación biológica.

Los resultados de esa encuesta, parcialmente reproducidos a continuación, expresados como porcentaje de artículos que contienen algún tipo de metodología estadística o matemática, para un año dado,

1940	11%
1950	31%
1960	52%
1970	61%
1980	83%

demuestran claramente que la matematización de la biología es un proceso sostenido, y que, además, en el presente, 4 de cada 5 artículos publicados en *The American Naturalist* incluyen algún tipo de análisis matemático o estadístico; bien como soporte, o bien, como metodología en la solución de problemas biológicos.

Es cierto, entonces, que la biología en general ha sufrido un proceso de cuantificación en los últimos años. La necesidad de analizar grandes cuerpos de datos, ha determinado que ese proceso haya sido más acelerado y de mayor alcance en Ecología.

En ésta, técnicas como superficies de respuesta (Wilbur, 1982), estadística multivariante (revisión de Green, 1979), Jackknife (Heltshe y Forrester, 1983) son ya de uso común, o comienzan a serlo. Sin embargo, existen metodologías, como por ejemplo la *Estimación Robusta* (Hoaglin et al., 1982; Launer y Wilkinson, 1979) y, en especial, el *Análisis Exploratorio de Datos* (AED) este último de reciente desarrollo (Tukey, 1970, 1972, 1977; Mosteller y Tukey, 1977; Hoaglin et al, 1982), las cuales han sido prácticamente ignoradas, a pesar de que presentan propiedades que hacen su uso deseable.

Dentro del ecléctico grupo de técnicas reunidas bajo la denominación de «Análisis Exploratorio de Datos», las representaciones gráficas juegan un papel preponderante. En este caso no sólo es importante la búsqueda de patrones a partir de grupos de datos, sino también el lograr una mayor y más eficiente transferencia de información al lector o al auditorio. Esta puede lograrse por medio de técnicas gráficas.

La preocupación por lograr un mayor rol de los métodos gráficos en el análisis estadístico de datos surge en la década de los setenta, alimentada fundamentalmente por la recién adquirida capacidad de las microcomputadoras y una mayor accesibilidad a su lenguaje y precio. Cox (1978) se hace eco de esa preocupación, al se-

ñalar la necesidad de una teoría de métodos gráficos que reúna de manera coherente todas las cosas que antes parecían no tener ninguna relación, al mismo tiempo que provea una base para tratar nuevas situaciones de manera sistemática».

Hoy en día existe un gran repertorio de métodos gráficos para analizar datos (para una completa visión y puesta al día, ver Chambers et al., 1983). De éstos hemos seleccionado dos: gráficos *Stem-and-leaf* (Tallo y Hoja) e intervalos *Boxplot* (Caja Gráfica), muy fáciles de usar y, dadas sus propiedades, potencialmente capaces de sustituir, respectivamente, al histograma y al intervalo de confianza, como representadores de datos.

Además de los ejemplos incluidos en algunas de las referencias citadas anteriormente, estos dos tipos de gráficos han mostrado su potencial en el análisis de datos en Ecología y Biometría (Conde et al., 1983), Educación (Hernández y Orellana, 1984) y Geoquímica (Torres y Conde, 1984). Igualmente, algunas de sus propiedades hacen su uso deseable en el proceso enseñanza-aprendizaje en cursos introductorios de bioestadística y estadística, tal como veremos más adelante.

Para ilustrar su uso, emplearemos datos obtenidos a partir de un estudio biométrico de granos de polen (Rull, 1983), un estudio limnológico (Paolini, datos no publicados) y mediciones de crustáceos (Díaz et al., 1979).

2. GRAFICOS «STEM-AND-LEAF» FRENTE A HISTOGRAMAS

A pesar de su popularidad y extendido uso, el histograma es quizás, junto con el análisis de regresión/correlación, el instrumento estadístico peor usado.

La construcción de un histograma cumple generalmente dos propósitos. En primer lugar, es utilizado para *representar datos,* con el fin de buscar patrones, tendencias y eventualmente, algún número —o números— que resuma los datos.

Un segundo propósito del histograma es servir como un *estimado no paramétrico* de la distribución subyacente. En ambos casos el, aparentemente, simple proceso de construir el histograma plantea, la mayoría de las veces, problemas de solución no inmediata.

En particular, las decisiones claves que se deben tomar se refieren al *ancho del intervalo* y al *número de intervalos.* Estos se toman generalmente en base a una combinación más o menos afortunada de intuición, experiencia y, principalmente, del deseo del analista para que aparezca en el papel la densidad de probabilidad específica en su mente (casi siempre la distribución normal).

En algunos casos, cuando el analista desea ser un poco más riguroso, dispone de algunos criterios para determinar el número de intervalos a usar en el histograma. De estos criterios, los más usados son los de:

$$\text{Sturges (1926)} \quad I = 1 + \log_2 n$$
$$\text{Dixon y Kronmal (1965)} \quad I = 10 \times \log_{10} n$$
$$\text{Velleman (1976)} \quad I = 2\sqrt{n}$$

donde

I = número de intervalos
n = número total de datos
$\log_2$ = logaritmo en base 2
$\log_{10}$ = logaritmo en base 10

Sin embargo, y a pesar de su relativa utilidad, todos los criterios anteriores son, en buena medida, arbitrarios, además de carecer de base teórica. Uno de los primeros intentos de seleccionar el número de intervalos y ancho de éstos en base a consideraciones teóricas, es el de Scott (1979), quien propone escoger el ancho del intervalo del histograma usando el criterio del Error Cuadrático Medio Integrado (ECMI), evaluando la integral de la expectativa de las diferencias entre la densidad de probabilidad observada y la postulada.

$$ECMI = \int E\,[fn\,(x) - f\,(x)]^{l2}\,dx$$

El ancho del intervalo h_n es seleccionado de tal manera, que el ECMI sea minimizado. Aunque, tal como puede desprenderse de la ecuación anterior, también

Tabla 1
Diámetro polar de granos de polen de *Poulsenia armata.*

Tabla 1. Diámetro polar de granos de polen de Poulsenia armata

17,0	16,1	14,1	14,2	17,0	16,5	15,1
15,8	16,0	14,4	14,0	17,3	14,3	15,7
15,5	17,5	14,5	16,4	16,2	16,7	15,0
16,6	13,6	17,8	17,4	18,3	18,2	18,5

Datos de Rull (1983).

en este caso se postula una densidad de probabilidad subyacente, f(x).

A diferencia del histograma, el gráfico «Stem-and-Leaf» no tiene como propósito fundamental el de ser usado como estimado de la densidad, sino el de optimizar la representación gráfica de datos, por lo que se soslaya uno de los problemas anteriores.

Para ilustrar la construcción del gráfico «Stem-and-Leaf» usaremos una serie de medidas del diámetro polar de los granos de polen de *Poulsenia armata* (Moraceae, Urticales), (Tabla 1).

Tal como podría inferirse del nombre, el gráfico «Stem-and-Leaf» tiene dos componentes. Uno básico, el tallo y otro apendicular que es la hoja. Para ilustrar su construcción usaremos los datos de la primera fila de la Tabla 1.

Estos son tomados uno a uno y sometidos al siguiente procedimiento: cada uno de los valores es cortado, asignando en este caso, la parte entera al «Stem» y la parte decimal al «Leaf», tal como puede verse en el diagrama siguiente:

Dato	Corte	"Stem"	"Leaf"
17,0	17|0	17	0
16,1	16|1	16	1
14,1	14|1	14	1
14,2	14|2	14	2
17,0	17|0	17	0
16,5	16|5	16	5
15,1	15|1	15	1

Al mismo tiempo, los datos, así tratados, se van colocando, en la medida en que son procesados, en lo que va a ser el gráfico «Stem-and-Leaf» (Figura 1).

fig. 1 Gráfico de «Stem-and-Leaf» para los valores 16,1; 14,1; 14,2; 17,0: 16,5 Y 15,1.

"Stem"	"Leaf"
14	1 2
15	1
16	1 5
17	0 0

Al incluir todos los valores de la Tabla 1 tendremos el siguiente gráfico de «Stem-and-Leaf», donde las hojas han sido anotadas en el orden en que van apareciendo en la misma tabla.

fig. 2 Gráfico de «Stem-and-Leaf» para todos los datos de la Tabla 1.

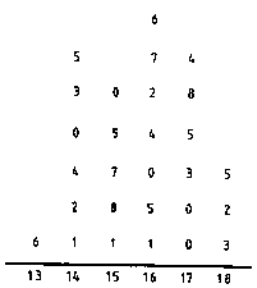
"Stem"	"Leaf"
13	6
14	1 2 4 0 3 5
15	1 8 7 5 0
16	1 5 0 4 2 7 6
17	0 0 3 5 8 4
18	3 2 5

En el proceso de construcción, la parte entera de los números se coloca en una posición fuerte, a la izquierda de la raya divisoria. Los dígitos restantes se colocan, según va apareciendo el número a partir de la Tabla 1, a la derecha de la división.

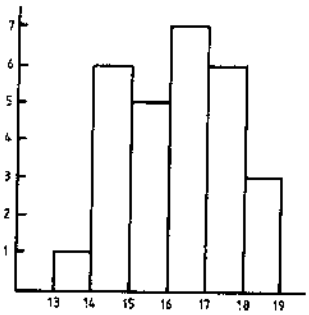
Para ver otros aspectos y para efectos de comparación, giremos el gráfico «Stem-and-Leaf» 90°, en sentido opuesto a las agujas del reloj, y construyamos, con los mismos datos, un histograma «clásico» con ancho de intervalo igual a una unidad, es decir, con la misma escala que el «Stem-and-Leaf» construido en el ejemplo en discusión.

fig. 3 Gráfico «Stem-and-Leaf» e Histograma construidos a partir de los datos contenidos en la Tabla 1.

a) Gráfico «Stem-and-Leaf»



b) Histograma



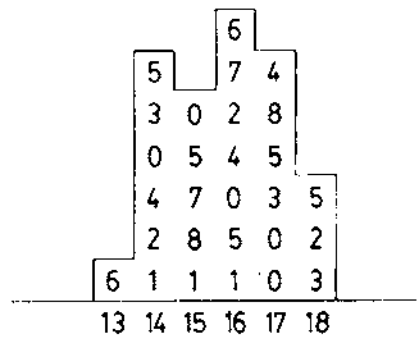
En la Figura 3 se pueden ver lado a lado, el histograma y el «Stem-and-Leaf» para los datos de la Tabla 1. Es posible ahora notar algunas virtudes del «Stem-and-Leaf».

Por una parte el «Stem-and-Leaf» conserva la información original, como fue presentada en la Tabla 1, a diferencia del histograma, donde se han perdido los datos básicos. Es decir, el histograma funciona básicamente como representador gráfico de la información procesada, mientras que el «Stem-and-Leaf», además de servir el mismo propósito, también llena el cometido de una tabla de trabajo, al conservar los datos crudos. De esta manera, es posible buscar rápidamente la mediana o cualquier otra característica de la distribución.

Otra ventaja del «Stem-and-Leaf sobre el histograma es que en el proceso de su construcción, el primero ha permitido que se produzca una virtual preordenación de los datos, lo cual puede ser una bonificación en el caso de que se desee aplicar posteriormente una prueba no paramétrica de rango.

Por último, puede observarse que el contorno o borde imaginario del «Stem-and-Leaf» reproduce la forma de histograma para los mismos datos, convirtiéndose así el primero de ellos en un potencial estimado no paramétrico de la densidad de probabilidad subyacente, tal como puede observarse en la Figura 4.

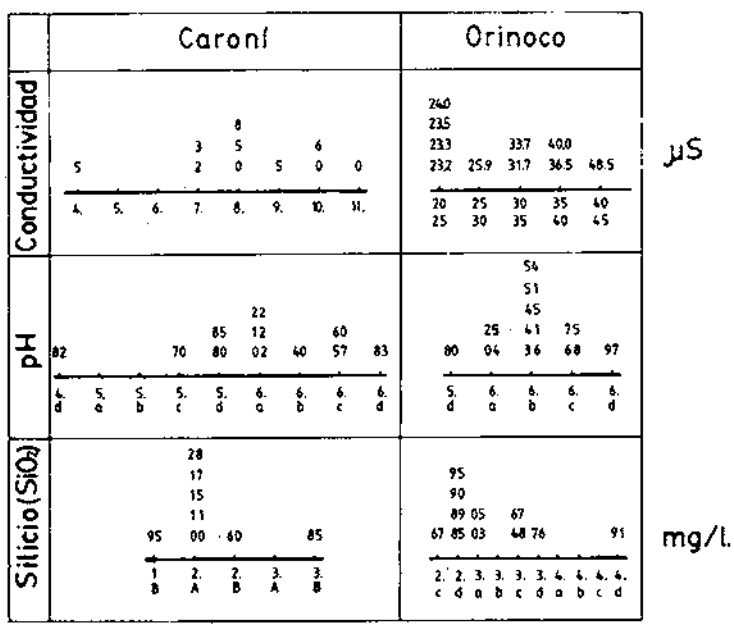
fig. 4 Gráfico «Stem-and-Leaf» y su borde, para los datos de la Tabla 1.



El gráfico Stem-and-Leaf», puede modificarse para contemplar situaciones diferentes a la básica presentada aquí.

Algunas de estas modificaciones han sido incluidas en la Figura 5, donde se han graficado datos fisicoquímicos de los ríos Caroní y Orinoco, tomados en 1981 y 1982.

figura 5



Es posible notar que la naturaleza básica de los datos y la importancia de dígitos a conservar determinan qué parte del número va a ser la rama y cual va a ser la hoja. Por ejemplo, para los datos, cada tallo incluye varios números enteros. Así, el primer tallo incluye valores que van entre 20 y 25, el segundo, aquellos entre 25 y 30 y así sucesivamente. Cada tallo está constituido por intervalos, debido a la dispersión de los datos y las escasez de los mismos. En otros casos, por ejemplo sílice en el Río Orinoco, dado el alto número de valores para cada entero, cada tallo corresponde a 25 unidades.

3. BOXPLOTS E INTERVALOS DE CONFIANZA

El «Boxplot» es una técnica gráfica que permite presentar los aspectos más importantes de la distribución empírica de cualquier grupo de datos, siendo especialmente útil para comparar varios grupos de datos. El «boxplot» podría sustituir al intervalo de confianza graficado, quedando este último como instrumento de contrastación de hipótesis y vehículo de proposiciones probabilísticas.

El «boxplot» (Figura 6) comunica información acerca de cinco características de la distribución de un grupo de datos: localización del valor central (media o mediana), dispersión central de los valores (distancia entre el cuarto superior e inferior), simetría de la distribución, longitud de las colas y valores extremos o fuera de ámbito, es decir aquellos valores tan grandes o tan pequeños que no pueden ser explicados por la distribución.

Para ilustrar la elaboración de un «boxplot», usaremos los datos de la Tabla 2. En particular, nos referiremos a aquellos números que reflejan las características de la distribución. El primero de ellos es la mediana, la cual estima el valor central de la distribución. Aunque la media es habitualmente usada para esto, nosotros preferimos usar la mediana, dadas sus propiedades robustas (Sokal y Rohlf, 1981). Esta puede computarse por medio de la fórmula

$M = [X_{(k)} + X_{(k+1)}]$ para un número par de datos.

donde: $k = \frac{n}{2}$

$M = X_{(k)}$ para un número impar de datos.

donde: $K = \frac{n + 1}{2}$

n en ambos casos es el número de datos y $X_{(k)}$ es el valor del dato en el lugar k, después de que aquellos han sido ordenados.

En el presente ejemplo $n = 10$ y $k = 5$; por lo tanto,

$$M = \frac{1}{2}[X_{(5)} + X_{(6)}] = \frac{1}{2}[17{,}4 + 18.6] = 18$$

Luego, para calcular cada cuarto es necesario usar el concepto de la profundidad, definida para cada uno de los valores de un grupo de datos como el menor de los rasgos ascendente y descendente, que corresponde a cada uno de ellos. Por ejemplo, el valor 23.1 de la Tabla 2, tiene rangos 9 y 2, por lo que su profundiad es 2.

La profundidad de los cuartos (PC) viene dada por la siguiente expresión:

$$PC = \frac{\text{Profundidad de la mediana} + 1}{2}$$

Tabla 2.
Ancho de caparazón de varios ejemplares del cangrejo *Aratus pisonii* (Brachyura, Grapsidae). Datos tomados de Díaz et al., (1979).

Valores X_i	Valores ordenados $X_{(i)}$	Rangos descendente	Rangos ascendente	Profundidad de los valores P
23.2	2.3	1	10	1
18.6	7.4	2	9	2
21.1	14.0	3	8	3
14.0	16.9	4	7	4
6.3	17.4	5	6	5
28.2	18.6	6	5	5
7.4	20.3	7	4	4
16.9	21.1	8	3	3
20.3	23.2	9	2	2
17.4	32.2	10	1	1

En el ejemplo que nos ocupa la profundidad de la mediana es PM = 5 y la profundidad de los cuartos es PC = 3. Por consiguiente, los dos cuartos son 14.0 y 21.1.

Para determinar los valores fuera del ámbito se hace necesario calcular los puntos de corte inferior y superior. Estos están dados por

$$PC_i = C_i - \frac{3}{2}\,dc \text{ y } PC_s = C_s + \frac{3}{2}\,dc$$

donde C_i : Cuarto inferior

C_s : Cuarto superior

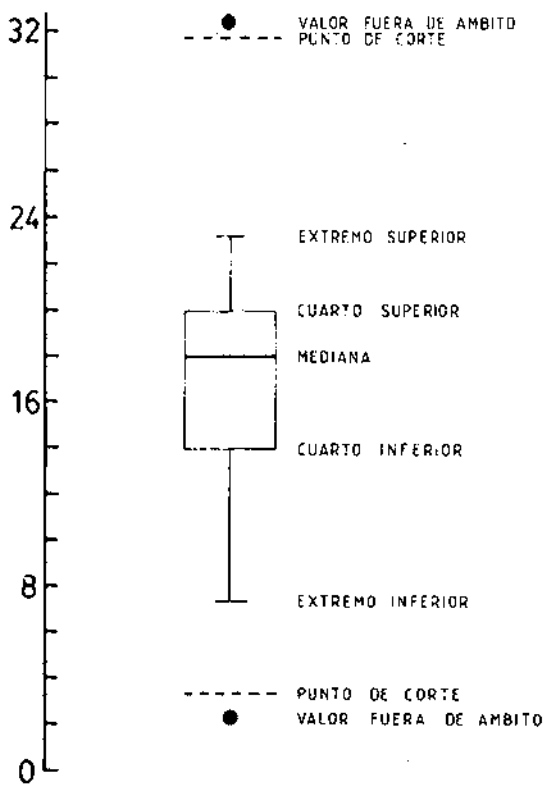
$dc = C_s - C_i$: Diferencia entre los cuartos.

Obteniéndose para el ejemplo en curso, como puntos de corte inferior y superior los valores 3.35 y 31.75, respectivamente. Es decir, que aquellos datos menores a 3.35, y mayores de 31.75 son valores fuera de ámbito.

Finalmente, se determinan los extremos, es decir aquellos valores máximo y mínimo del grupo de datos que no están fuera del ámbito o intervalo comprendido entre los puntos de corte. En este ejemplo los extremos son 7.4 y 23.2.

Una vez que todos estos valores han sido calculados es posible proceder a construir el boxplot respectivo (Figura 6).

fig. 6 Boxplot correspondiente a los datos de la Tabla 2. El eje vertical corresponde al ancho del caparazón en milímetros.



El cuerpo de la caja (box) ha sido construido usando como bordes los cuartos superior e inferior. La línea horizontal dentro de la caja corresponde a la mediana. Luego, se unen los bordes del cuerpo de la caja con los valores extremos que no sobrepasen los puntos de corte, los cuales están indicados en la Figura 6 por líneas punteadas horizontales. Finalmente se indican mediante equis o puntos grandes, como en este caso, los valores fuera de ámbito.

Toda la información anterior cumple una función. La mediana es un estimador de localización. La longitud de la caja, es decir, la diferencia entre los dos cuartos, da una idea de la dispersión central de los datos. La posición relativa de la mediana respecto a los dos cuartos permite detectar posibles asimetrías de la masa central de los datos. En el caso que nos ocupa puede verse que la distribución es bastante asimétrica.

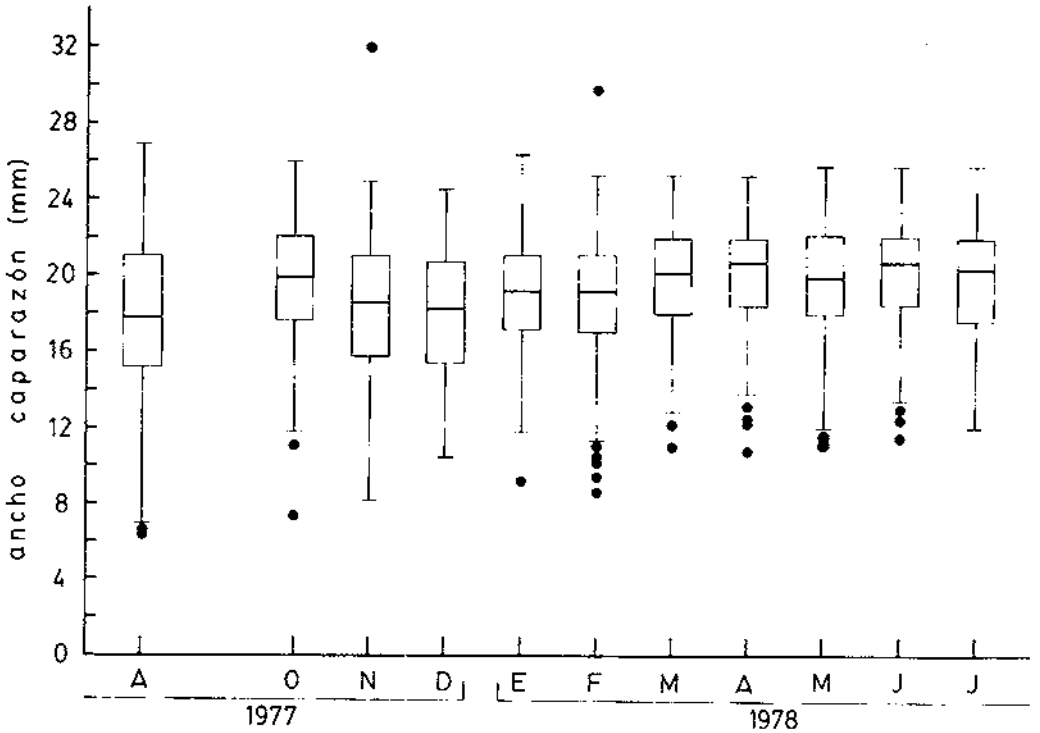
Finalmente, el «boxplot» también incluye información acerca de la longitud de las colas de la distribución y los valores fuera de ámbito. En la Figura 6 se han incluido, además, leyendas correspondientes a los diferentes aspectos destacados por el boxplot. Puede notarse la gran cantidad de información ofrecida en un solo gráfico. No obstante, el boxplot alcanza su mayor utilidad, cuando varios de ellos son presentados en forma de serie de tiempo (Figura 7). En este caso puede verse que, exceptuando Agosto, el cual fue un mes de puesta a punto de los métodos de muestreo y establecimiento de estaciones, las distribuciones de tamaños de caparazón del cangrejo *Aratus pisonii* a lo largo del tiempo, son bastante similares. Igualmente puede notarse que las distribuciones son relativamente simétricas y que la cola de los valores pequeños de cada una de las distribuciones es más larga, la mayoría de las veces, que la cola de los valores grandes. Añadiendo a esto el hecho de que la mayoría de los valores fuera de ámbito, se encuentran en la primera de las colas mencionadas. Si se desease describir los datos anteriores por medio de alguna distribución probabilística, el análisis exploratorio por medio de boxplots, sugiere que debería seleccionarse una que fuese bastante simétrica en torno a la mediana, pero que al mismo tiempo tuviese una larga cola hacia la izquierda; por ejemplo, la distribución canónica de valores extremos (Johnson y Kotz, 1970) o una distribución log gamma (Lawless, 1982).

Para efectos de ilustración, en el presente caso es interesante construir un intervalo de confianza al 95% para la media de los datos de la Tabla 2. Los límites de dicho intervalo vienen dados por la expresión

$$L = \overline{X} \pm 2.064\frac{s}{\sqrt{n}}$$

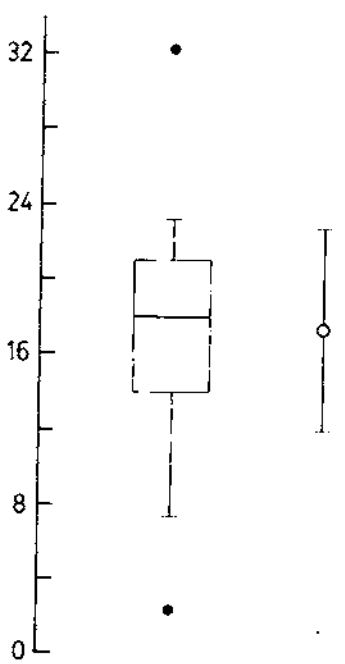
donde $\overline{X}$, s y n son, respectivamente la media muestral, la desviación típica y el número de datos. Los límites superior es inferior son entonces $L_s = 22.74$ y $L_i = 11.94$.

fig. 7 Serie de gráficos boxplot para anchos de caparazón del cangrejo *Aratus pisonii*.



El intervalo de confianza al 95% al lado del boxplot correspondiente, ofrece mucha menos información que este último (Figura 8).

fig. 8 «Boxplot» e intervalo de confianza al 95% para la media de los datos de la Tabla 2. El eje vertical corresponde al ancho en milímetros del caparazón de un cangrejo.



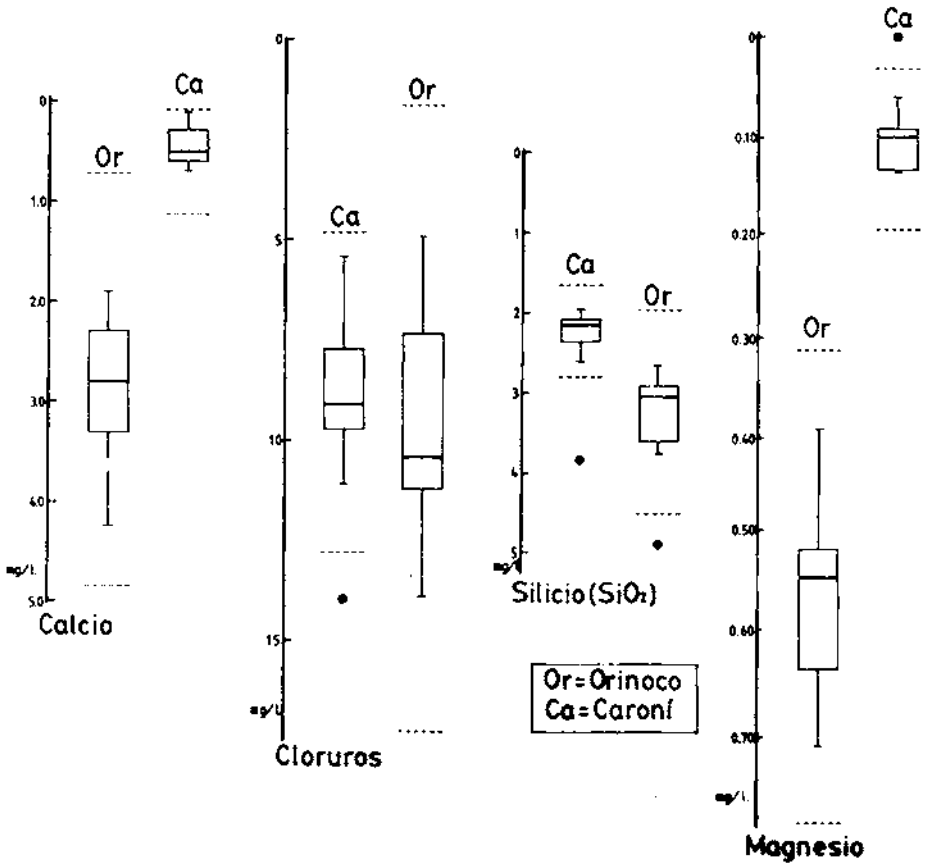
Un último ejemplo del uso del boxplot puede observarse en la Figura 9. En ella se ha graficado las distribuciones de calcio, cloruros, silicio y magnesio en los ríos Orinoco y Caroní. Un rápido vistazo a los gráficos nos permite aprender un cúmulo de información relevante acerca de dichos ríos.

De manera muy general, podría decirse que un buen representador gráfico es aquel en el que el impacto visual de sus componentes es proporcional a la importancia de los diferentes aspectos del cuerpo de datos. El boxplot cumple muy bien con esta recomendación. El énfasis mayor de esta figura lo da la caja, la cual a su vez representa el 50% central de los valores de la distribución con la localización central indicada por una raya que representa la mediana. Luego, con un menor énfasis visual, vienen representadas las colas de la distrubución y por último, como puntos aislados, pero con un cierto énfasis visual, dada su importancia, aquellos valores que caen fuera de ámbito.

4. OBSERVACIONES FINALES

El uso de métodos gráficos en el análisis de datos estadísticos alcanzó la excelencia en el siglo pasado. Minard, Playfair y Walker son los autores más reconocidos de una escuela que popularizó el uso de las series de tiempo para efectos analíticos, y, al mismo tiempo, se caracterizó por alto grado de refinación y estética de sus productos. En particular, el famoso gráfico de

fig. 9 Gráficos «boxplot» de la distribución de varios elementos en los ríos Caroní y Orinoco.



Minard realizado en 1861, donde se representa la campaña rusa de Napoleón, durante los años de 1812 y 1813, ha sido considerado por Tufte (1983) y Wainer (1984), como el mejor gráfico estadístico que se haya dibujado.

Sin embargo, la irrupción de la obra *Statistical Methods for Research Workers* de Sir Ronald A. Fisher en 1925, poniendo a la disposición de los usuarios un instrumento de análisis de datos, que combinaba la sencillez matemática con la claridad conceptual, fue relegando los métodos gráficos, al punto que a comienzo de los años setenta la mayoría de los libros de estadística incluía únicamente gráficos de frecuencia y de dispersión y series de tiempo.

Esta situación, como ya dijimos en la introducción, ha cambiado rápidamente en los últimos diez años. La preocupación de Cox (1978) ha sido recogida por muchos autores (entre otros, Andrews et al., 1980; Chambers et al., 1983; Cleveland et al., 1982a, 1982b; Cleveland y McGill, 1983, 1985; Kleiner y Hartigan, 1981; Tufte, 1983; Tukey, 1970, 1972, 1977; Wainer, 1984), quienes han contribuido a dilatar el cuerpo de conocimientos de los métodos gráficos, tanto en sus aspectos empíricos, como en el desarrollo de una incipiente teoría.

Los métodos gráficos de análisis y manejo de datos, a pesar de su sencillez, constituyen poderosos instrumentos de comunicación de resultados y, por supuesto, de análisis, principalmente en la fase exploratoria. No obstante, poco de este potencial está siendo usado en biología (de hecho prácticamente en ninguna ciencia). Es por ello que se requiere una mayor difusión de los métodos para vencer la resistencia inicial de los usuarios. Esta ha sido precisada por Boen y Zahn (1982) al señalar que:

The simple graph, when appropiate, is a major consulting virtue. Often it's all you need. Consultants who worry that their analysis isn't complicated enough to impress mathematical statisticians have a problem. This brings upon an essential point about consulting: Effective consultants very seldom in their consulting do things that impress mathematical statisticians. Mathematical statisticians are on the lookout for new mathematical statistical concepts or the use of previously unused techniques.

La utilidad de los gráficos discutidos anteriormente no sólo se limita al análisis y presentación de resultados de investigación, sino que también se extiende al campo de la enseñanza estadística y bioestadística.

La forma que debe asumir la enseñanza de estas disciplinas ha sido bastante controvertida (para una visión completa de este problema, consultar Rustagi y Wolfe, 1982) y tiende a caer en extremos. Por una parte se encuentran los matemáticos y estadísticos matemáticos para quienes el análisis de datos y las aplicaciones de la estadística en situaciones de «mundo-real» no son más que recetas de cocina y carpintería. En el otro extremo se encuentran los prácticos. Para ellos los fundamentos matemáticos de la teoría estadística no son más que reflejos de la conspiración matemática para hacer a la estadística una materia compleja y difícil de digerir; por lo tanto, esos principios deben ser apartados para pasar inmediatamente a lo que es «sustancial»: el ordeño de los datos.

Obviamente, cualquiera de los dos extremos es aborrecible. Al igual que en cualquier otra disciplina, el equilibrio es recomendable y digno de ser alcanzado. En estadística, el área de equilibrio podría estar en el Análisis Exploratorio de Datos, donde se conjugan los elementos inmediatos del análisis de datos con los más formales de la fundamentación matemática subyacente.

Nuestra experiencia personal (no necesariamente generalizable) en la enseñanza del análisis exploratorio de datos es llamativa: por ejemplo, los «boxplots», además de proveer rápidamente al estudiante con un instrumento para resolver sus problemas de manejo de datos, también lo motivan fuertemente hacia la comprensión de aspectos teóricos relacionados con las formas de las distribuciones, valores fuera de ámbito («outliers»), percentiles e intervalos de confianza. Además de ofrecer al Profesor el ambiente para introducir conceptos relativos a estadística robusta y no paramétrica.

Para finalizar este artículo nos referiremos brevemente a las posibles traducciones de los duros términos anglosajones, «boxplot» y «stem-and-leaf». Torres (en Torres y Conde, 1984) ha sugerido que el primero de ellos debería traducirse como «caja gráfica». Esto da una buena idea de la forma del instrumento, aunque no de su propósito y utilidad. El mismo autor ha sugerido que una traducción adecuada para «stem-and-leaf» (tallo-y-hoja, literalmente) sería «tabligraf», poniendo en este caso el énfasis en las funciones del gráfico y no en su forma.

Agradecimientos

Nuestro agradecimiento muy especial al Dr. Jorge Rabinovich (Centro de Ecología, Instituto Venezolano de Investigaciones Científicas), quien leyó detenidamente el artículo, haciendo valiosas sugerencias.

Deseamos dar las gracias al Lic. José Javier Alió (Fondo Nacional de Investigaciones Agrícolas y Pecuarias, Cumaná Venezuela) y al Profesor Ricardo Bitter (Universidad Nacional Experimental Francisco de Miranda, Coro Venezuela) por sus constructivas observaciones. Igualmente vaya nuestro agradecimiento al Dr. Ernesto Medina (Centro de Ecología, Instituto Venezolano de Investigaciones Científicas) por su interés en el artículo y sus palabras positivas. Al Dr. Ramón Margalef (Departamento de Ecología, Universidad de Barcelona, España) por su interés en que el artículo fuese publicado.

Finalmente nuestro reconocimiento a Noris Chirino Brett y a Berta Sánchez de García por su paciencia y por su impecable trabajo mecanográfico.

REFERENCIAS BIBLIOGRAFICAS

ANDREWS, H.P., SNEE, R.D. y SARNER, M.H., 1980, Graphical display of means. *The American Statistician* 34: 195-199.

BOEN, J.R. y ZAHN, D.A., 1982 *The Human Side of Statistical Consulting*. Belmont, California: Lifetime learning Publications.

CHAMBERS, J.M., CLEVELAND, W.S., KLEINER, B. y TUKEY, P.A., 1983, *Graphical Methods for Data Analysis*. Belmont, California: Wadsworth.

CLEVELAND, W.S., Mc GILL, R., 1983, A color-caused illusion on a statistical graph. *The American Statistician*. 37: 101-105.

CLEVELAND, W.S. y Mc GILL, R., 1985, Graphical perception and graphical methods for analyzing scientific data. *Science* 229: 828-833.

CLEVELAND, W.S., DIACONIS, P. y Mc GILL, R., 1982a, Variables on Scaterplots look more highly correlated when the scales are increased. *Science* 216: 1138-1141.

CLEVELAND, W.S., HARRIS, C.S. y Mc GILL, R., 1982b, Judgements of circle size on statistical maps. *Journal of the American Statistical Association*, 77: 541-547.

CONDE, J.E., RULL, V. y VEGAS, T., 1983, Gráficos Stem-and-Leaf y Boxplots en Ecología. *Acta Científica Venezolana* 34 (Suplemento 1): 138 [abstract].

COX, D.R., 1978, Some remarks on the role in statistics of graphical methods. *Journal of the Royal Statistical Society, Series C* 27: 4-9.

DIAZ, H., CONDE, J.E., y BEVILACQUA, M., 1979, Patrón de historia vital y dinámica poblacional de cangrejos de ambientes marino y estuarino. Un estudio comparativo. Proyecto S1-0766. CONICIT. Informe de Avance.

DIXON, W.J. y KRONMAL, R.A., 1965, The choice of origin and scale for graphs. *Journal of the Association for Computing Machinery* 12: 259-261.

GREEN, R.D., 1979, *Sampling Design and Statistical Methods for Environmental Biologists*. New York: Wiley.

HELTSHE, J.F. y FORRESTER, N.E., 1983, Estimating species richness using the Jacknife Procedure. *Biometrics* 39: 1-14.

HERNANDEZ, O.E. y ORELLANA, R., 1984, Análisis exploratorio de los resultados del certamen preliminar de la Octava Olimpíada Matemática Venezolana. *Acta Científica Venezolana* 35 (Suplemento 1): 156 [Abstract].

HOAGLIN, D.C., MOSTELLER, F. y TUKEY, J.W., (Eds) 1983, *Understanding Robust and Exploratory Data Analysis*. New York: Wiley.

JOHNSON, N.L. y KOTZ, S., 1970, *Continuous Univariate Distributions-1* New York: Wiley.

KLEINER, B. y HARTIGAN, J.A., 1981, Representing points in many dimensions by trees and castles. *Journal of the American Statistical Association* 76: 260-276.

LAUNER, R.L. y WILKINSON, G.N., (Eds), 1979, *Robustness in Statistics*. New York: Academic Press.

LAWLESS, J.F., 1982, *Statistical Models and Methods for Lifetime Data*. New York: Wiley.

MOSTELLER, F. y TUKEY, J.W., 1977, *Data Analysis and Regression*. Reading, Massachusetts: Addison-Wesley.

RULL, V., 1983, El polen de las Urticales de los Andes Venezolanos. *Acta Científica Venezolano* 34 (Suplemento 1): 54 [Abstract].

RUSTAGI, J.S. y WOLFE, D.A., (Eds), 1982, *Teaching of Statistics and Statistical Consulting*. New York: Academic Press.

SCHMID, C.F., 1983, *Statistical Graphics. Design Principles and Practices*. New York: Wiley-Interscience.

SCOTT, D.W., 1979, On optimal and data-based histograms. *Biometrika* 66: 605-610.

SOKAL, R.R. y ROHLF, F.J., 1981, *Biometry. The Principles and Practice of Statistics in Biological Research*. Segunda Edición. San Francisco: W.H. Freeman.

STURGES, H.A., 1926, The choice of a class interval. *Journal of the American Statistical Association* 21: 65-66.

TORRES, J. y CONDE, J.E., 1984, Análisis exploratorio de datos geoquímicos: abundancia y distribución de resistatos pesados en la cuenca del Río Paragua, Edo. Bolívar. *Acta Científica Venezolana* 35 (Suplemento 1): 76 [Abstract].

TUFTE, E.R., 1983, *The Visual Display of Quantitative Information*. Cheshire, Connecticut: Graphic Press.

TUKEY, J.W., 1970, *Exploratory Data Analysis*. Edición Preliminar Limitada Vol. 1. Reading, Massachusetts: Addison-Wesley.

TUKEY, J.W., 1972, Some graphic and semigraphic displays. En: T.A. Bancrof (Ed.), *Statistical Papers in Honor of George W. Snedecor*. Ames, Iowa: Iowa State University Press.

TUKEY, J.W., 1977, *Exploratory Data Analysis*. Reading, Massachusetts: Addison-Wesley.

VELLEMAN, P.F., 1976, Interactive computing for exploratory data analysis. I: display algorithms. *1975 Proceedings of the Statistical Computing Section*, American Statistical Association, Washington, D.C.

WAINER, W., 1984, How to display data badly. *The American Statistician*, 38:137-147.

WILBUR, H.M., 1982, Competition between tadpoles of *Hyla femoralis* and *Hyla gratiosa* in laboratory experiments. *Ecology* 63:278-282.