La normalidad estadística y la Biología, una relación tortuosa

Pedro Cervantes Hernández*

Durante mis años dedicados al análisis de datos, he sido testigo de una gran variedad de investigaciones en el ámbito ecológico-biológico, en los cuales se evidencia una exagerada tendencia hacia la búsqueda de un "argumento" encaminado a probar la existencia de la normalidad en datos analizados.

Un argumento es un razonamiento con el que se pretende probar o desmentir una afirmación, convenciendo a alguien de su verdad o su falsedad. El argumento posee una doble función: a) probar, acreditar o demostrar *X* y b) desmentir, refutar o negar *X* (Bringas-Valdivia 2004).

En el terreno de la estadística paramétrica, uno de los "supuestos", que aparentemente se ha arraigado en el ámbito ecológico-biológico, es la normalidad, y de no cumplirse ésta, sobre todo en muestras menores a 30 datos, la conclusión acerca de la inferencia poblacional se considera errónea (Meneses 2005). A este respecto, en variadas ocasiones he escuchado a infinidad de estudiantes de licenciatura y postgrado exclamar: "¡mis datos no son normales! ¿qué debo hacer ahora?"; mientras que algunos otros, intentan lograr su objetivo al cobijo de la bienaventurada transformación de variables, sin comprender realmente la razón de dicha transformación.

El énfasis a "suponer" que los registros de cualquier característica (Y_n) en los individuos de una población, provenientes de *n* muestras, son normales, generalmente se realiza sin una conciencia plena de lo que realmente significa la variabilidad en el ámbito ecológico-biológico. Por esta razón, desde mi particular punto de vista, me permito señalar algunos aspectos que deberían de considerarse al tratar tan polémico tema, sobre todo cuando se hace referencia de que la normalidad es indispensable para el uso de la distribución (*F*), o en otras palabras el análisis de la varianza (ANDEVA).

Por otra parte, en el caso específico de la industria, y considerando una línea de producción, la estadística resulta ser una herramienta útil en el campo del control de calidad, en donde el argumento está dirigido a "demostrar" más no a "suponer", que entre un artículo y otro la variabilidad entre cualquier característica es mínima, considerando a la variabilidad como un sinónimo de error.

En el caso del análisis de la variabilidad ecológico-biológica, lo razonable es que no existan dos individuos iguales en ninguna de sus características, por lo que una muestra siempre será distinta a otra; sin embargo, lo que es obvio a simple vista en la naturaleza, no es considerado como tal por la mayoría de los analistas dedicados a esta línea de investigación (IMAS-INEGI 2003). A este respecto, lo recomendable es obtener de las *n* muestras un nivel de variabilidad tal, que permita mejorar la significación de la inferencia poblacional, interpretando a la variabilidad como la cantidad de información contenida en *n* para explicar cómo una variable *X* (ambiental o poblacional), incide sobre las características de los individuos de una población en estudio.

No necesariamente se requiere incrementar *n* para lograr esto último, ello depende de la característica que se desee analizar; esto es, si la característica es rara en los individuos de una población, entonces para poder detectarla, se

* *Universidad del Mar, campus Puerto Ángel. Apdo. Postal 47., C.P. 70902., Puerto Ángel, Oaxaca, MÉXICO*
Correo electrónico: pch@angel.umar.mx

requiere de un tamaño de muestra mayor, mientras que si ésta es común o no rara en la población, el tamaño de la muestra puede reducirse a tan sólo unos cuantos registros (IMAS-INEGI 2003).

A pesar de considerar la importancia de *n* en el caso ecológicobiológico, la dinámica poblacional y la variabilidad ambiental condicionan cambios en las características de los individuos de una población con respecto a la temporalidad (horas, días, semanas, meses y años). De acuerdo con Punt & Hilborn (1996), esta interacción genera respuestas con un alto grado de heterogeneidad tanto en ambientes estables como inestables, que frecuentemente son difíciles de registrar e interpretar dado la existencia de los errores de observación (v_n) y de proceso (w) asociados al muestreo.

En el caso industrial, ambos tipos de error son próximos a 0, por lo que la información extraída de las *n* muestras puede modelarse con base en cualquier distribución de probabilidad que se desee investigar (continua o discreta). Sin embargo, en el caso ecológico-biológico, v_n y $w \neq 0$, razón por la cual, la información extraída de las *n* muestras posee una distribución con parámetros $X_n(\overline{X},\ ^2)$; y dependido de n, es el error vn el que se aproxima a la distribución de probabilidad normal Zn (0,1), más no Yn o en su defecto w, pues no depende de n, sino de otras fuentes de error no contempladas en el procesamiento de los datos.

Los analistas de datos en el ámbito ecológico-biológico, en lugar de considerar a v_n (0,1), como en el caso de la "verosimilitud" en la estadística bayesiana, suponen a Y_n como normal, y para demostrarlo aplican una serie de transformaciones matemáticas sobre ésta, modificando tan sólo el valor observado de $Y_n(\log_{10}, \ln, \sqrt{Y}, arcsen(Y), 1/Y, etc.)$; sin embargo, dichas transformaciones no poseen un argumento firme que demuestre lo anterior. En términos prácticos, es preferible no suponer y ajustar directamente a la tan deseada distribución de probabilidad normal; sin embargo, el ajuste igualará a (1) la variabilidad contenida en Y_n, por lo que en vez de interpretarse como una fuente de información adicional, la variabilidad será analizada como en el caso del control de calidad.

Al realizar un muestreo, la información obtenida suele compararse con otros muestreos obtenidos en circunstancias diferentes o con métodos distintos, por lo que comúnmente se cometen errores de sobre-estimación o sub-estimación al confrontar las características de interés. En este caso, la transformación matemática de Y_n es útil para lograr una unidad de muestreo estandarizado ya sea por la unidad de área, volumen, tiempo, etc. (Sharon 1999). Lo anterior no implica que la palabra "estandarizar" se refiera a probar la normalidad en datos analizados, sino más bien a la homogenización de la varianza.

Es importante señalar que uno de los grandes problemas asociados al tema de la normalidad deriva del hecho de que un biólogo o un ecólogo, al estudiar estadística lo haga sobre una base de datos enfocados a las áreas de investigación industrial, administrativa, médica, social y contable, en donde la normalidad es de vital importancia. Romero-Cortés (2005) ha destacado el auge que actualmente ha tomado el criterio Geary como herramienta para analizar la normalidad en el campo del control de calidad. A este respecto, al buscar la frase "normalidad estadística" en la Web (dependiendo del buscador), el resultado es de aproximadamente 526 temas, de los cuales el 95% hacen referencia a éste en las áreas de investigación antes señaladas.

La identificación y el análisis correcto de la variabilidad resulta ser la clave que permite al argumento lograr una conclusión lógica referente a la base de datos. Lo anterior no aplica a las investigaciones de laboratorio, en donde las condiciones son estrictamente controladas. Sin embargo, en los casos dinámicos ecológico-biológico, los analistas deberán de considerar que éstos no son comparables con los casos de laboratorio; razón por la cual, se sugiere que al aplicar la ANDEVA, se considere que es ilógico analizar con una distribución sesgada (F) un grupo de datos que se suponen insesgados o con una distribución Z_n (0,1).

A este respecto, en algunas investigaciones en las que se ha ignorado dicho supuesto, los resultados y las conclusiones han sido estadísticamente confiables y ante todo lógicas

para el ámbito ecológico-biológico (Del Ángel-García 2002, Montaño-Juárez 2002, Cuesta-Castillo 2003, Frías-Velasco 2004, Vázquez-Gil *et al.* 2004, García-Ocampo 2005, Sánchez-Meraz 2005, Márquez-Reyes 2005, Flores-Gómez 2005, Gallardo-Berumen 2005).

El IMAS-INEGI (2003) señaló que al analizar e interpretar una base de datos, las conclusiones derivadas de éste consideran la probabilidad de equivocarnos. Lo anterior puede estimarse con base en el análisis de los errores de proceso y observación (Punt & Hilborn, 1996); sin embargo, esto último generalmente es pasado por alto, salvo en las investigaciones de tesis en las que se evidencia una actualización en sus procesos metodológicos.

Por último, si trabajamos en términos de un supuesto, la interpretación de los resultados podría considerarse mayormente errónea que el propio análisis de la normalidad; por lo que en los casos dinámicos ecológico-biológicos, se recomienda conocer o inferir el tipo de distribución de probabilidad sobre la cual se fundamentarán los resultados y las conclusiones de la investigación.

Estudio de Caso

En un estudio oceanográfico se estimó la concentración de clorofila *a* (mg m^{-3}) en la capa superficial del mar para la temporada "NortesEl Niño" 97-98 en las áreas de surgencia de los golfos de Tehuantepec, Papagayo y Panamá. Los registros fueron obtenidos a través de imágenes de color del mar nivel L_3 GAC (Global Area Coverage), provenientes de los sensores OCTS (Ocean Color and Temperature Sensor) de noviembre de 1996 a diciembre de 1997 (excepto entre mayo-agosto de 1997) y SeaWIFS (Sea-Viewing Wide Field-of-View Sensor) de enero a mayo de 1998 (Fig. 1).

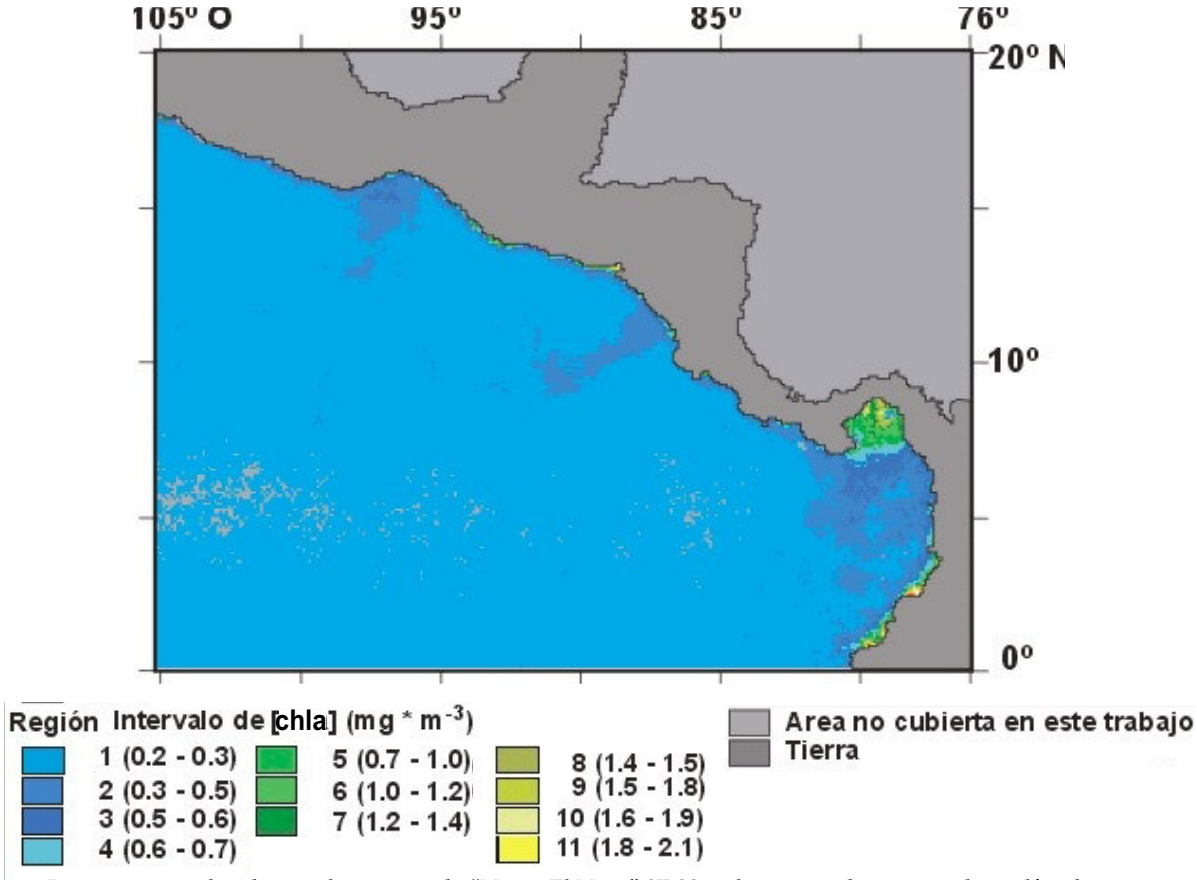

Figura 1. Imagen regionalizada para la temporada "NortesEl Niño" 97-98 en las áreas adyacentes a los golfos de Tehuantepec, Papagayo y Panamá. Tomado de Frías-Velasco (2004). [chla] Concentración de Clorofila a (mg * m-3)

Tabla I. Resultados para los cuatro casos de análisis del ANDEVA

Caso de análisis	Valor de F	Valor de p
Datos puros	34.64	0.000059
Transformación $\sqrt{chla}$	35.99	0.000051
Transformación $\log_{10}(chla)$	27.01	0.000157
Transformación Z(0,1)	0	1.0

chla es la concentración de Clorofila *a*.

De la imagen anterior y con ayuda del programa WIM 6.11, se obtuvieron 20 muestras al azar de la concentración de clorofila *a* en cada una de las áreas de surgencia, de modo que se pretende establecer un argumento que demuestre si entre éstas existe diferencia significativa en cuanto a la variable de interés. Para ello, se consideraron cuatro casos para el ANDEVA (Tabla I, Fig. 2), la interpretación de los resultados es la siguiente:

En primera instancia, de acuerdo con la escala de colores, la figura 1 sugiere a simple vista, que entre las plumas de clorofila *a* adyacentes a los golfos de Tehuantepec y Papagayo, existe una aparente similitud con respecto al golfo de Panamá, por lo que se espera diferencias significativas entre éstas. El ANDEVA a partir de los datos crudos y considerando las transformaciones raíz cuadrada ($\sqrt{chla}$) y logaritmo base 10 ($\log_{10}(chla)$), culmina-

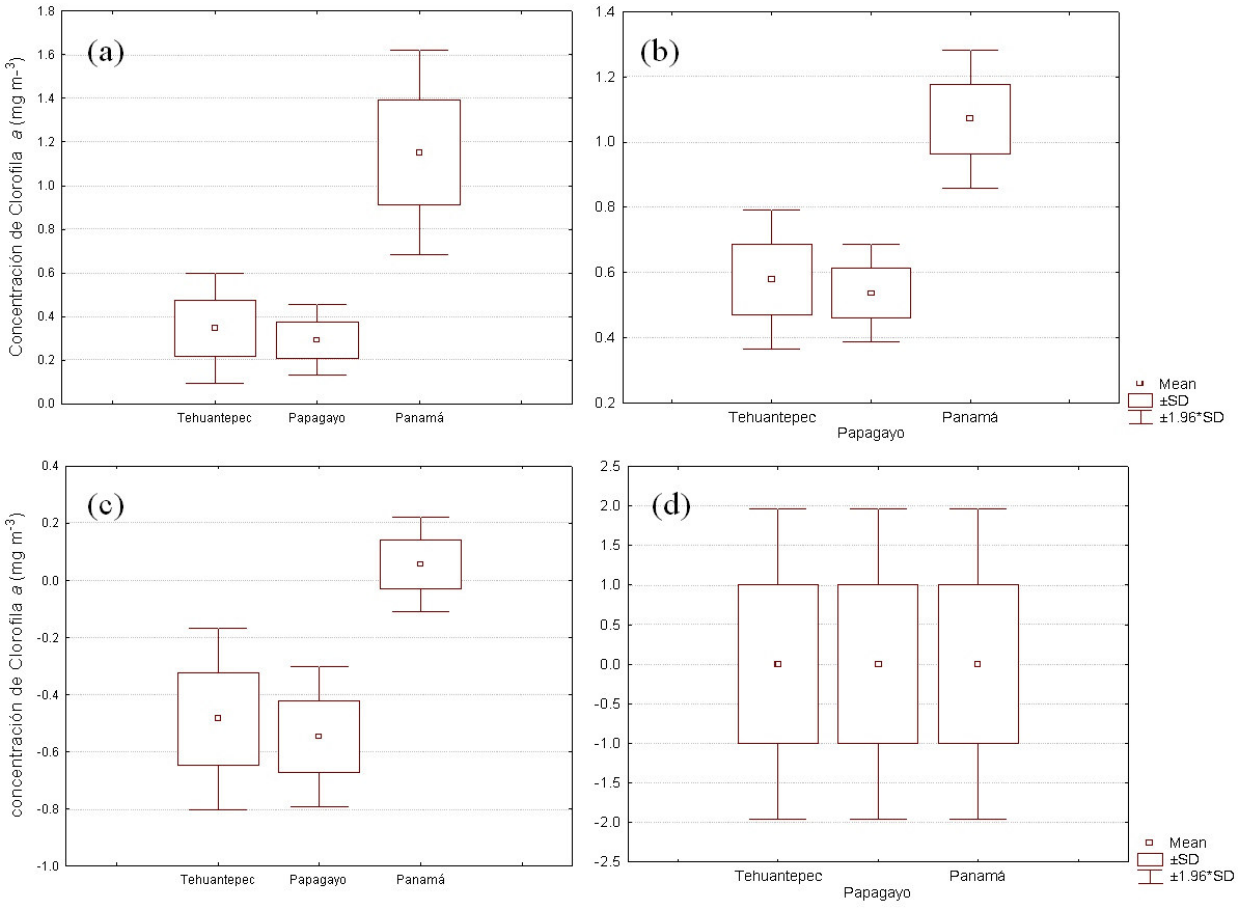

Figura 2. Diagramas de caja para los cuatro análisis del ANDEVA. (a) datos crudos, (b) transformación $\sqrt{chla}$, (c) transformación $\log_{10}(chla)$ y (d) transformación Z(0,1). Chla es la concentración de Clorofila a.

ron en la aceptación de la hipótesis alternativa (Ha). Por lo que se comprueba, que existe diferencia significativa en la concentración de clorofila *a* entre los golfos Tehuantepec, Papagayo y Panamá durante el periodo "Nortes-El Niño" de 1997-98 (Tabla I, Fig. 2a-c).

La transformación matemática de la clorofila *a* se llevó a cabo con el objeto de estandarizar las estimaciones superficiales entre ambos sensores; sin embargo, la diferencia entre los algoritmos para estimar la concentración de clorofila *a* en los sensores OCTS y SeaWIFS es mínima, por lo que el ANDEVA con los datos crudos resultó en la aceptación de Ha.

Suponiendo que las transformaciones sobre la variable de interés fueron las correctas y una vez realizado el ANDEVA, es importante destacar que los datos transformados podrían ser problemáticos si se pretende realizar inferencia estadística, ya que las nuevas unidades podrían ser no compatibles con respecto a la media y a la varianza original de la población en estudio.

Lo anterior puede observase al considerar el cambio de escala asociado con la trasformación $\log_{10}(chla)$, la cual asignó una escala de intervalo a una variable que, biológicamente, debería ser analizada con escala de razón (Figs. 2a, c). Lo anterior advierte que es relevante considerar las bases teóricas para la correcta aplicación de las transformaciones matemáticas y de los procesos de estandarización contenidos en la teoría de muestreo.

Por otro lado, si en vez de suponer que los datos son normales a través de un tipo de transformación u otro proceso, y verdaderamente ajustamos éstos a una distribución normal Z (0,1); el resultado de dicho proceso provocará que en todos los tratamientos analizados, las variables involucradas tengan media 0 y varianza 1 (Fig. 2d).

A este respecto, la conclusión del ANDEVA para el estudio de caso que nos compete, resultó en la aceptación de la hipótesis nula (Ho) (Tabla I, Fig. 2d); por lo que con datos normales no existe diferencia significativa en la concentración de clorofila *a* entre los golfos Tehuantepec, Papagayo y Panamá durante el periodo "Nortes-El Niño" de 1997-98.

Este último ajuste ha cambiado totalmente la lógica que se percibía a simple vista en la figura 1: los datos son normales; sin embargo, el análisis de la ANDEVA no tiene sentido alguno para el objetivo que se planteó, o ¿lo tiene?

La distribución normal es una herramienta relevante en el análisis multivariado, en la conformación de la matriz de correlación, en el análisis de la probabilidad, en las pruebas de hipótesis y en el campo de la estadística bayesiana, entre otros; sin embargo, es necesario conocer bajo qué criterios teóricos ésta distribución puede ayudarnos en el planteamiento, análisis e interpretación en los casos dinámicos ecológico-biológico.

Finalmente, dejo a discusión los argumentos expuestos en este trabajo, emitiendo una recomendación acerca de la existencia de nuevos textos y artículos del 2000 al presente, en los cuales la estadística clásica de los años 60 y 70, es aplicada con nuevos enfoques capaces de sorprender hasta al más experimentado analista de datos, o ¡quizá no¡

Agradecimientos

Para Alfredo Frías Velasco (UMAR) por permitir la utilización de los compuestos regionalizados. A Mario Alejandro Gómez Ponce (UMAR), Blanca Sánchez-Meraz y Antonio López Serrano (UMAR) por las sugerencias y comentarios. Se agradecen los comentarios de los revisores internos de la revista.

Referencias

Bringas-Valdivia, J.M. 2004. Los argumentos falaces en la investigación jurídica, un comentario. Accedido en 2005. *In* www.juridicas.unam.mx/inst/becarios/eureka/1/art4.htm

Cuesta-Castillo, L.B. 2003. Abundancia y distribución de los foraminíferos planctónicos de la Bahía de la Paz, México y su relación con la dinámica oceánica. Tesis de licenciatura, Facultad de Ciencias, Universidad Nacional Autónoma de México.

Del Ángel-García, G. 2002. Variación estacional de la migración vertical de *Litopenaeus setiferus* en la Bahía de Campeche, México. Tesis de licenciatura, Facultad de Ciencias, Universidad Nacional Autónoma de México.

Flores-Gómez, A. 2005. Modelo dinámico de biomasa para el

camarón café *Farfantepenaeus californiensis* (Holmes, 1900) en el Golfo de Tehuantepec, Oax., México. Tesis de licenciatura, Universidad del Mar, Puerto Ángel, Oaxaca.

Frías-Velasco, A. 2004. Regionalización de los Golfos de Tehuantepec , Papagayo, Panamá y áreas adyacentes mediante la biomasa fitoplanctónica estimada a partir de imágenes satelitales. Tesis de licenciatura, Universidad del Mar, Puerto Ángel, Oaxaca.

García-Ocampo, M.R. 2005. Patrones de reclutamiento de las colonias de coral del género *Pocillopora* Lamarck 1816 (Anthozoa: Scleratinia), en cinco localidades de Ixtapa-Zihuatanejo, Guerreo, México. Tesis de licenciatura, Universidad del Mar, Puerto Ángel, Oaxaca.

Gallardo-Berumen, M. I. 2005. Análisis del sistema de vedas sobre la explotación del recurso camarón en el Golfo de Tehuantepec. Tesis de licenciatura, Universidad del Mar, Puerto Ángel, Oaxaca.

IMAS-INEGI. 2003. Variabilidad y muestreo (vídeo). Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, UNAM e Instituto Nacional de Estadística Geográfica e Informática. Directora cinematográfica: Vanesa Gil Tejada.

Márquez-Reyes, L.A. 2005. Efecto de la temperatura y la hormona fluoximesterona en la reversión sexual, sobrevivencia y crecimiento de la tilapia del nilo *Oreochromis niloticus* (Linnaeus, 1758). Tesis de licenciatura, Universidad del Mar, Puerto Ángel, Oaxaca.

Meneses, B. 2005. La aplicación de la estadística no paramétrica en la administración. Enlace accedido en 2005: www.uv.mx/iiesca/revista2/bety1.html

Montaño-Juárez, K. 2002. Variación estacional del patrón de migración vertical de larvas de camarón *Sicyonia* spp., en la Bahía de Campeche, México. Tesis de licenciatura, Facultad de Ciencias, Universidad Nacional Autónoma de México.

Punt, A.E. & R. Hilborn. 1996. Biomass dynamic models. Computerized Information Series. Food and Agriculture Organization of the United Nations (FAO), Biodyn, Roma. 62 pp.

Romero Cortés, J.C. 2005. Prueba de normalidad de Geary. Accedido en 2005. Accedido en 2005: www.azc.uam.mx/publicaciones/enlinea2/num1/1-4.htm

Sánchez Meraz, B. 2005. Respuestas del reclutamiento del camarón café (*Farfantepenaeus californiensis*, Holmes 1900) a la variación interanual de la temperatura superficial del mar en el Golfo de Tehuantepec, Oaxaca. Tesis de maestría. Universidad del Mar Puerto Ángel, Oaxaca, México.

Sharon, L. 1999. Muestreo, diseño y análisis. Thomsom, México. 480 pp.

Vázquez-Gil, C.A., P. Cervantes-Hernández, S. Serrano-Gúzman, R.P. Cid-Rodriguez & M.E. Fuente-Carrasco. 2004. Análisis de la mortalidad en la población del caracol púrpura *Plicopurpura pansa* (Gould, 1853) en bahías de Huatulco, Oaxaca. Ciencia y Mar 8(24): 21-29.