Revista de Matemática: Teoría y Aplicaciones 2005 **12**(1 & 2) : 1–12
cimpa – ucr – ccss issn: 1409-2433

EL ANÁLISIS DE LA VARIACIÓN ESTACIONAL EN MEDICINA

Osvaldo Marrero*

Recibido/Received: 9 Feb 2004

Resumen

Se divulgan métodos recientes para el análisis de la variación estacional en medicina. Estos métodos son efectivos aun cuando los datos presentan pequeña amplitud y el tamaño de la muestra es bajo, dos características a menudo presentes cuando uno realiza este tipo de análisis. Los métodos tratan los dos casos: un solo grupo y múltiples grupos. Además, estos métodos sirven para analizar distintos tipos de variación estacional. La aplicación se ilustra por medio de ejemplos con datos reales.

Palabras clave: Medicina, Variación estacional.

Abstract

This is a survey of recent methods for the analysis of seasonal variation in medical research. These methods are effective even when the data show small amplitude and the sample size is small, two characteristics that are often present when one analyzes medical data for seasonal variation. The methods deal with both cases: one group and multiple groups. Also, these methods can be used to analyze different types of seasonal variation. Real-data examples are used to illustrate the application of these techniques.

Keywords: Medical research, Seasonal variation.

Mathematics Subject Classification: 62–02, 62F03, 62H15, 92C60.

*Department of Mathematical Sciences, Villanova University, Villanova, Pennsylvania 19085–1699, Estados Unidos; Fax: +1 (610) 519 6928; E-Mail: Osvaldo.Marrero@villanova.edu.

1 Introducción

El propósito de este artículo es divulgar en castellano unos métodos recientes (Marrero 1998, 1999) para el análisis estadístico de la variación estacional. El motivo para estas investigaciones proviene de la medicina. Como aspectos novedosos, aparte del idioma, este trabajo reune métodos para los dos casos, un solo grupo y múltiples grupos, y pone énfasis en la aplicación de estos métodos, incluso un nuevo ejemplo. En los ejemplos se usan datos reales.

La variación estacional es importante en las ciencias médicas por razones ecológicas. En efecto, si la incidencia de una enfermedad muestra variación estacional, entonces hay que considerar que un agente del ambiente pudiera formar parte de la etiología de esa enfermedad. Esto ayuda a los investigadores médicos a esclarecer el origen de ciertas enfermedades.

Como un ejemplo simple, es bien sabido que muchas personas padecen de coriza y estornudos dos veces al año debido a ciertos agentes que se manifiestan en el ambiente durante esas dos épocas del año. Pero hay otras enfermedades más graves, algunas de ellas congénitas, que también muestran variación estacional y cuyas etiologías no están completamente claras.

Los tipos de variación estacional más comunes en medicina son la sinusoidal anual, la sinusoidal semestral y la unimodal. Para el análisis, el estadígrafo dispone usualmente de frecuencias en forma de una serie de tiempo. Frecuencias mensuales a través de un mínimo de doce meses es un caso común. El análisis estadístico se complica a menudo por dos razones: la amplitud de la variación es pequeña y el tamaño de la muestra es bajo. Esto hace que tests aplicables más generalmente, como el chi cuadrado de Pearson, suelan tener poca potencia para detectar la presencia de la variación estacional; al parecer, lo que se gana en generalidad de aplicación es menos que lo se pierde en potencia en algunos casos. Por lo tanto estadígrafos han desarrollado tests más específicos para este problema; generalmente estos tests se han limitado a detectar un solo tipo de variación estacional. Para más detalles se puede consultar (Marrero 1979, 1983, 1984, 1988, 1992).

Los métodos presentados en este artículo sirven para analizar múltiples tipos de variación estacional. Todos estos tests se pueden explicar por medio de una misma idea sencilla; esta idea clave se presenta en la sección 2. Los métodos para un solo grupo y para múltiples grupos se presentan, respectivamente, en las secciones 3 y 4. La sección 5 está dedicada a los ejemplos.

2 La idea clave

Se supone que los datos son frecuencias de muestras que obedecen leyes multinomiales. Es crucial que las clases de la ley multinomial sean cíclicas, como un ciclo de estaciones, con orden fijo y sin ambigüedad. Como ejemplos uno tiene los meses del año y los días de la semana. Las clases de la ley multinomial son identificadas con arcos de la misma medida en la circunferencia de un círculo; esto se hace permitiendo la posibilidad que el mismo arco pudiera corresponder a dos clases distintas.

Para ilustrar la idea clave, uno considera un círculo cuya circunferencia se ha dividido en arcos de la misma medida; conviene imaginarse el círculo suspendido por una cuerda que proviene del centro del círculo. En el medio de cada uno de estos arcos se pone una masa, y así se obtiene un círculo ponderado. Si todas estas masas son iguales, entonces el centro de gravedad del círculo ponderado se encuentra en el centro del círculo y el círculo ponderado se mantiene en equilibrio; esto corresponde a la hipótesis nula. Pero si todas las masas son iguales excepto una que es más pesada, entonces el centro de gravedad del círculo ponderado se aleja del origen del círculo, en dirección de la posición de la masa más pesada; además, el círculo se inclina en dirección del radio que pasa por la posición de la masa excepcional. En efecto, en este caso de una sola masa sobrepesada, el centro de gravedad del círculo ponderado se va alejando del centro del círculo a medida que el sobrepeso aumenta; al mismo tiempo, la inclinación del círculo ponderado también aumenta. Esto hace la hipótesis alternativa más plausible. Este caso de un solo sobrepeso corresponde a un ejemplo de la variación unimodal.

La idea clave consiste en imaginarse que las frecuencias de la muestra son masas cuyos pesos están determinados por los valores numéricos de las frecuencias, y entonces tales masas se colocan de forma adecuada a la hipótesis alternativa, como sigue.

Cuando las frecuencias son mensuales a través de un año y la hipótesis alternativa es la variación sinusoidal anual o la variación unimodal, entonces la circunferencia se divide en doce arcos distintos y de la misma medida. Enero corresponde al arco de $\pi/12$ a $\pi/4$ (15° a 45°), febrero corresponde al arco de $\pi/4$ a $5\pi/12$ (45° a 75°), etc.

Cuando las frecuencias son mensuales a través de un año y la hipótesis alternativa es la variación sinusoidal semestral, entonces la circunferencia se divide en seis arcos distintos y de la misma medida. Enero y julio corresponden al arco de $\pi/6$ a $\pi/2$ (30° a 90°), febrero y agosto corresponden al arco de $\pi/2$ a $5\pi/6$ (90° a 150°), etc.

Matemáticamente todo esto recuerda la trigonometría y el análisis armónico.

En las próximas dos secciones los métodos se presentan brevemente; para más detalles se pueden consultar los artículos de Marrero (1998, 1999).

3 Un solo grupo

Se supone que las frecuencias $N_1, \ldots, N_k$ de la muestra siguen una ley multinomial con parámetros n y $p_1, \ldots, p_k$. Además se supone que las k clases constituyen un ciclo. Se desea contrastar las hipótesis $H_0\colon p_1 = \cdots = p_k = 1/k$ y $H_1\colon p_i \neq p_j$ por lo menos para un par de distintos parámetros p_i y p_j. Esto significa, en particular, que la hipótesis alternativa no está limitada a un solo tipo de variación estacional. En la práctica, el investigador precisa un tipo de variación para la hipótesis alternativa y entonces el test se adapta a ese tipo de variación.

Se considera un círculo cuyo radio es de una unidad de largo y cuyo centro se encuentra en el origen de un sistema de coordenadas. Para construir el círculo ponderado de la muestra, sea $\theta_i := 2\pi ti/k$ el punto medio, en radianes, del arco de $\theta_i - \pi t/k$ a $\theta_i + \pi t/k$, donde t es un entero tal que $0 < t < k/2$. La masa N_i se pone en el punto $(\cos\theta_i, \operatorname{sen}\theta_i)$. El índice t determina la frecuencia cíclica t/k y el período correspondiente k/t de este

emplazamiento de las masas. En la práctica uno escoge el valor de t de acuerdo con la hipótesis alternativa; si $k := 12$, entonces $t := 1$ corresponde a la variación sinusoidal anual o la variación unimodal y $t := 2$ corresponde a la variación sinusoidal semestral.

En coordenadas rectangulares, el centro de gravedad del círculo ponderado de la muestra se encuentra en el punto

$$\left(\bar{X} := \left(\sum_{i=1}^{k} N_i \cos \theta_i \right) \Big/ n, \bar{Y} := \left(\sum_{i=1}^{k} N_i \operatorname{sen} \theta_i \right) \Big/ n \right).$$

Los estadísticos $\bar{X}$ y $\bar{Y}$ estiman sin sesgo, respectivamente, $\tau_{\bar{X}}$ y $\tau_{\bar{Y}}$, las coordenadas del centro de gravedad del círculo ponderado de la población. Si la hipótesis nula es cierta, $E_0(\bar{X}) = E_0(\bar{Y}) = 0$; además, $\operatorname{var}_0(\bar{X}) = \operatorname{var}_0(\bar{Y}) = 1/(2n)$ y $\operatorname{cov}_0(\bar{X}, \bar{Y}) = 0$.

Sean $\tau := (\tau_{\bar{X}}, \tau_{\bar{Y}})'$ y V la esperanza y la matriz de covarianzas del vector aleatorio $T := (\bar{X}, \bar{Y})'$. El estadístico del test es $T_t := (T - \tau)' V^{-1} (T - \tau)$, es decir,

$$T_t = \frac{2}{n} \left[\left\{ \sum_{i=1}^{k} N_i \cos(2\pi t i/k) \right\}^2 + \left\{ \sum_{i=1}^{k} N_i \operatorname{sen}(2\pi t i/k) \right\}^2 \right]$$

cuando la hipótesis nula es cierta. Suponiendo que la ley de probabilidad de las medias ponderadas $\bar{X}$ y $\bar{Y}$ sigue aproximadamente una ley binormal no singular, uno concluye que la forma cuadrática T_t obedece la ley del chi cuadrado con dos grados de libertad. Un estudio de simulación (Marrero 1999) demuestra que este estadístico funciona muy bien.

En la práctica la matriz V se desconoce y es estimada por $\widehat{V}$; entonces el área elíptica $\left\{ \tau \colon (T - \tau)' \widehat{V}^{-1} (T - \tau) \leq \chi_2^2(\alpha) \right\}$ es una región de confianza aproximada de nivel $100(1 - \alpha)\%$ para el centro de gravedad del círculo ponderado de la población.

Las coordenadas polares del centro de gravedad del círculo ponderado de la muestra son (R, Θ), donde $R := (\bar{X}^2 + \bar{Y}^2)^{1/2}$ y

$$\Theta := \begin{cases} \operatorname{Arctan}(\bar{Y}/\bar{X}), & \text{si } \bar{X} > 0, \\ \operatorname{Arctan}(\bar{Y}/\bar{X}) - \pi, & \text{si } \bar{X} < 0 \text{ y } \bar{Y} < 0, \\ \operatorname{Arctan}(\bar{Y}/\bar{X}) + \pi, & \text{si } \bar{X} < 0 \text{ y } \bar{Y} \geq 0, \\ -\pi/2, & \text{si } \bar{X} = 0 \text{ y } \bar{Y} < 0, \\ \pi/2, & \text{si } \bar{X} = 0 \text{ y } \bar{Y} > 0. \end{cases}$$

Si el círculo ponderado de la muestra no se mantiene en equilibrio, Θ se puede utilizar para inferir donde se encuentran los valores extremos del círculo ponderado de la población. La varianza $\operatorname{var}_1(\Theta)$ es igual a zero cuando $\bar{X} = 0$, y en otro caso puede ser aproximada por una serie de Taylor limitada al primer término. En la práctica uno utiliza el estimador

$$\begin{aligned} \operatorname{estvar}_1(\Theta) := \Big[& \operatorname{estvar}_1(\bar{X}) \left\{ \widehat{E}_1(\bar{Y}) \right\}^2 + \operatorname{estvar}_1(\bar{Y}) \left\{ \widehat{E}_1(\bar{X}) \right\}^2 \\ & - 2 \operatorname{estcov}_1(\bar{X}, \bar{Y}) \left\{ \widehat{E}_1(\bar{X}) \right\} \left\{ \widehat{E}_1(\bar{Y}) \right\} \Big] \Big/ \left[\left\{ \widehat{E}_1(\bar{X}) \right\}^2 + \left\{ \widehat{E}_1(\bar{Y}) \right\}^2 \right]^2, \end{aligned}$$

donde todas las estimaciones se obtienen reemplazando p_i por $\widehat{p}_i := N_i/n$.

4 Múltiples grupos

Se consideran dos tests. Uno de los tests es para detectar variación estacional simultáneamente en los grupos. El otro es un test de homogeneidad para múltiples distribuciones multinomiales con clases cíclicas. Los dos tests están basados en círculos ponderados y siguen ideas de la estadística multivariada.

4.1 Variación estacional en múltiples grupos

El estadígrafo dispone de gk datos: k frecuencias de cada uno de g grupos mutuamente independientes. Para cada $j = 1, \ldots, g$, se supone que las frecuencias $N_{j1}, \ldots, N_{jk}$ del grupo j siguen una ley multinomial con parámetros n_j y $p_{j1}, \ldots, p_{jk}$. También se supone que las clases en cada una de estas distribuciones multinomiales están ordenadas cíclicamente y sin ambigüedad.

La hipótesis nula es que en cada grupo las probabilidades son iguales; es decir, para cada $j = 1, \ldots, g$, uno tiene $p_{j1} = \cdots = p_{jk} = 1/k$. La hipótesis alternativa es que la hipótesis nula es falsa. Por lo tanto la hipótesis alternativa no está limitada a un solo tipo de variación estacional y, además, se permite que distintos grupos tengan distintas variaciones estacionales.

Como en el caso de un solo grupo, uno construye un círculo ponderado de la muestra para cada uno de los grupos. Para cada $j = 1, \ldots, g$, sea $\theta_{ji} := (2\pi t_j i)/k$; este es el punto medio, en radianes, del arco de $\theta_{ji} - (\pi t_j)/k$ a $\theta_{ji} + (\pi t_j)/k$ que corresponde a la $i$ésima clase de la distribución multinomial del grupo j. Entonces, para cada $i = 1, \ldots, k$, la masa N_{ji} se pone en el punto medio θ_{ji} del arco. El índice t_j es un entero tal que $0 < t_j < k/2$.

Para cada $j = 1, \ldots, g$, en coordenadas rectangulares, el centro de gravedad del círculo ponderado de la muestra para el grupo j se encuentra en el punto $(\bar{X}_j, \bar{Y}_j)$, donde

$$\bar{X}_j := \left(\sum_{i=1}^{k} N_{ji} \cos\theta_{ji}\right) \Bigg/ \left(\sum_{i=1}^{k} N_{ji}\right) = \frac{1}{n_j}\sum_{i=1}^{k} N_{ji}\cos\theta_{ji}$$

$$\bar{Y}_j := \left(\sum_{i=1}^{k} N_{ji} \operatorname{sen}\theta_{ji}\right) \Bigg/ \left(\sum_{i=1}^{k} N_{ji}\right) = \frac{1}{n_j}\sum_{i=1}^{k} N_{ji}\operatorname{sen}\theta_{ji}.$$

Para cada $j = 1, \ldots, g$, sea V_j la matriz de covarianzas del vector aleatorio $(\bar{X}_j, \bar{Y}_j)'$, y sea

$$T_{t_j} := \left(\begin{bmatrix}\bar{X}_j \\ \bar{Y}_j\end{bmatrix} - \begin{bmatrix}E(\bar{X}_j) \\ E(\bar{Y}_j)\end{bmatrix}\right)' V_j^{-1} \left(\begin{bmatrix}\bar{X}_j \\ \bar{Y}_j\end{bmatrix} - \begin{bmatrix}E(\bar{X}_j) \\ E(\bar{Y}_j)\end{bmatrix}\right).$$

El estadístico del test se define por $T := \sum_{j=1}^{g} T_{t_j}$. Se supone que para cada $j = 1, \ldots, g$, las medias ponderadas $\bar{X}_j$ y $\bar{Y}_j$ obedecen una ley de probabilidad que puede ser aproximada bien por una distribución normal bivariante no singular; entonces el estadístico T sigue la ley del chi cuadrado con $2g$ grados de libertad. Además, cuando H_0 es cierta,

$$T = \sum_{j=1}^{g} \frac{2}{n_j} \left\{ \left(\sum_{i=1}^{k} N_{ji} \cos \frac{2\pi t_j}{k} i \right)^2 + \left(\sum_{i=1}^{k} N_{ji} \operatorname{sen} \frac{2\pi t_j}{k} i \right)^2 \right\}.$$

La magnitud de T_{t_j} determina la contribución del grupo j al valor de T así como la intensidad de la inferencia para ese grupo.

4.2 Test de homogeneidad

El propósito es evaluar la homogeneidad de las probabilidades multinomiales para f de los g grupos, $f \leq g$. Los datos son como en la sección 4.1; además, $m := \sum_{j=1}^{f} n_j$. Suponiendo que los grupos tienen el mismo tipo de variación (e.g., sinusoidal anual), uno desea determinar si hay diferencia en esa variación a través de los grupos (e.g., diferentes amplitudes en diferentes grupos). La hipótesis nula es que para cada clase multinomial, la probabilidad es la misma en todos los grupos; es decir,

$$\begin{array}{ccccccc} p_{11} & = & \cdots & = & p_{f1} & =: & p_1 \\ & & \vdots & & & & \\ p_{1k} & = & \cdots & = & p_{fk} & =: & p_k. \end{array}$$

La hipótesis alternativa es que H_0 es falsa.

El estadístico del test se basa en f círculos ponderados de la muestra que se construyen como en la sección 4.1, excepto que hay diferencias en los índices de adaptación y en las masas. Como se supone que los grupos muestran el mismo tipo de variación, los índices de adaptación son iguales para todos los grupos; es decir, $t_1 = \cdots = t_f =: t$. Por lo tanto, los puntos medios en los arcos son también iguales para todos los grupos; es decir, $\theta_{ji} := (2\pi t_j i)/k = (2\pi t i)/k =: \theta_i$. Para cada $j = 1, \ldots, f$, uno usa las masas $N_{j1}/n_j, \ldots, N_{jk}/n_j$ para construir el círculo ponderado de la muestra para el grupo j.

Una vez más, la idea que conduce al estadístico del test es sencilla: si la hipótesis nula es cierta, entonces el centro de gravedad del círculo ponderado de la muestra se encuentra en el mismo punto—no necesariamente el centro—para todos los grupos.

Para cada $j = 1, \ldots, f$, en coordenadas rectangulares, el centro de gravedad del círculo ponderado de la muestra para el grupo j se encuentra en el punto $(\bar{X}_j, \bar{Y}_j)$, donde

$$\bar{X}_j := \left(\sum_{i=1}^{k} (N_{ji}/n_j) \cos \theta_{ji} \right) \Bigg/ \left(\sum_{i=1}^{k} (N_{ji}/n_j) \right) = \frac{1}{n_j} \sum_{i=1}^{k} N_{ji} \cos \theta_{ji}$$

$$\bar{Y}_j := \left(\sum_{i=1}^{k} (N_{ji}/n_j) \operatorname{sen} \theta_{ji} \right) \Bigg/ \left(\sum_{i=1}^{k} (N_{ji}/n_j) \right) = \frac{1}{n_j} \sum_{i=1}^{k} N_{ji} \operatorname{sen} \theta_{ji}.$$

Si la hipótesis nula es cierta, la esperanza del centro de gravedad es la misma para todos los círculos ponderados de la muestra; es decir,

$$\left(E_0 \left(\bar{X}_1, \right), E_0 \left(\bar{Y}_1 \right) \right) = \cdots = \left(E_0 \left(\bar{X}_f, \right), E_0 \left(\bar{Y}_f \right) \right) = \left(\sum_{i=1}^{k} p_i \cos \theta_i, \sum_{i=1}^{k} p_i \operatorname{sen} \theta_i \right).$$

Sea V_Z la matriz de covarianzas del vector aleatorio $Z := (\bar{X}_1, \ldots, \bar{X}_f, \bar{Y}_1, \ldots, \bar{Y}_f)'$. Generalmente los elementos de V_Z dependen de probabilidades multinomiales que son desconocidas. Por lo tanto en la práctica uno usa la matriz $\widehat{V}_Z$ que se obtiene de V_Z reemplazando las probabilidades multinomiales desconocidas por sus estimaciones.

Cuando H_0 es cierta las probabilidades multinomiales son $p_1, \ldots, p_k$; para cada $i = 1, \ldots, k$, p_i se estima por $\widehat{p}_i := c_i/m$, donde $c_i := \sum_{j=1}^{f} N_{ji}$. También bajo la hipótesis nula, los elementos de $\widehat{V}_Z$ se calculan por medio de las expresiones siguientes:

$$\operatorname{estvar}_0(\bar{X}_j) = \frac{1}{mn_j}\left\{\sum_{i=1}^{k} c_i \cos^2\theta_i - \frac{1}{m}\left(\sum_{i=1}^{k} c_i \cos\theta_i\right)^2\right\} \text{ para cada } j = 1, \ldots, f;$$

$$\operatorname{estvar}_0(\bar{Y}_j) = \frac{1}{mn_j}\left\{\sum_{i=1}^{k} c_i \operatorname{sen}^2\theta_i - \frac{1}{m}\left(\sum_{i=1}^{k} c_i \operatorname{sen}\theta_i\right)^2\right\} \text{ para cada } j = 1, \ldots, f;$$

$$\begin{aligned}\operatorname{estcov}_0(\bar{X}_j, \bar{Y}_j) = \frac{1}{mn_j}\Bigg\{&\sum_{i=1}^{k} c_i \cos\theta_i \operatorname{sen}\theta_i \\ &- \frac{1}{m}\left(\sum_{i=1}^{k} c_i \cos\theta_i\right)\left(\sum_{i=1}^{k} c_i \operatorname{sen}\theta_i\right)\Bigg\} \text{ para cada } j = 1, \ldots, f;\end{aligned}$$

$$\operatorname{estcov}_0(\bar{X}_r, \bar{X}_s) = \operatorname{cov}(\bar{X}_r, \bar{X}_s) = 0 \text{ para cada } r, s \in \{1, \ldots, f\} \text{ con } r \neq s;$$

$$\operatorname{estcov}_0(\bar{X}_r, \bar{Y}_s) = \operatorname{cov}(\bar{X}_r, \bar{Y}_s) = 0 \text{ para cada } r, s \in \{1, \ldots, f\} \text{ con } r \neq s;$$

$$\operatorname{estcov}_0(\bar{Y}_r, \bar{Y}_s) = \operatorname{cov}(\bar{Y}_r, \bar{Y}_s) = 0 \text{ para cada } r, s \in \{1, \ldots, f\} \text{ con } r \neq s.$$

Sean $X := (\bar{X}_1, \ldots, \bar{X}_f)'$ y $Y := (\bar{Y}_1, \ldots, \bar{Y}_f)'$. Además, sean C_1 y C_2 matrices de contraste, de dimensiones $(f-1) \times f$ y tales que $E_0(C_1 X) = 0$ y $E_0(C_2 Y) = 0$. Entonces la matriz C de dimensiones $(2f-2) \times 2f$ se define por

$$C := \begin{bmatrix} C_1 & 0 \\ 0 & C_2 \end{bmatrix}.$$

Uno supone que el vector aleatorio Z obedece una ley de probabilidad que puede ser aproximada bien por una distribución normal multivariante no singular. Entonces el vector aleatorio CZ obedece una ley normal multivariante no singular con esperanza $CE(Z)$ y matriz de covarianzas $CV_Z C'$.

Si $C\widehat{V}_Z C'$ es no singular, el estadístico del test U_f se define por

$$U_f := \{CZ - E(CZ)\}' \left(C\widehat{V}_Z C'\right)^{-1} \{CZ - E(CZ)\}.$$

La ley de probabilidad de U_f es aproximadamente el chi cuadrado con $2f-2$ grados de libertad. El valor del estadístico U_f es invariante bajo cambios de las matrices de contraste C_1 y C_2. Si H_0 es cierta, entonces $E_0(CZ)=0$, y $U_f=(CZ)'(C\widehat{V}_ZC')^{-1}(CZ)$. El estadístico del test evalúa simultáneamente para todos los f grupos las diferencias que hay entre las posiciones de los centros de gravedad de los círculos ponderados de la muestra.

Los datos de los grupos que son juzgados homogéneos se juntan para formar un solo grupo. Entonces uno puede calcular regiones de confianza y hacer otros análisis usando los métodos de la sección 3.

5 Ejemplos

El objeto de estos ejemplos es ilustrar la aplicación de los métodos presentados en este artículo. No hay intención de disputar ningún otro análisis de los mismos datos. Las hipótesis alternativas son de acuerdo con los artículos de donde provienen los datos.

En el ejemplo 1 los datos son descritos por simples modelos de un armónico. El fin de estos modelos no es más que ayudar a visualizar el comportamiento de los datos a través del tiempo y, además, ayudar a interpretar los resultados de los análisis estadísticos. Por lo tanto no hay intención de obtener modelos óptimos. Es claro que en general armónicos adicionales pudieran servir para obtener una mejor modelación, pero en la práctica uno encuentra que modelos con uno o dos armónicos son adecuados.

La modelación también sirve para corroborar los resultados del círculo ponderado de la muestra. Por ejemplo, si un modelo sinusoidal anual muestra un máximo en la primera mitad de enero, entonces el valor de Θ debe ser alrededor de 22.5°, que en este caso indica el punto medio de la primera mitad del arco correspondiente a enero.

El ejemplo 2 demuestra la potencia de estos métodos aun cuando la muestra es de solamente 30 efectivos.

En los ejemplos, el *valor de probabilidad* p es la probabilidad que el estadístico del test tome un valor igual o más extremo al valor actualmente observado cuando la hipótesis nula es cierta; "más extremo" quiere decir un resultado que favorezca a la hipótesis alternativa. Usualmente uno rechaza la hipótesis nula si $p \leq 0.05$.

5.1 Ejemplo 1

Separadamente para cada sexo y reunidos en doce totalidades mensuales, los datos son el número de casos de diabetes mellitus en jóvenes de 10 a 17 años, tal como fueron registrados en Colorado durante 1978–83 por mes de diagnosis (Jones *et al.* 1988). De enero a diciembre las listas de datos para los varones y para las hembras son, respectivamente, (37, 23, 16, 10, 14, 14, 8, 17, 17, 13, 21, 27) y (27, 22, 18, 12, 11, 12, 13, 14, 9, 14, 17, 17). Jones *et al.* comentan que desde los años 1860, investigadores sospechan que este tipo de diabetes tiene una etiología infecciosa o viral. Por lo tanto tiene sentido investigar si estos datos muestran variación estacional.

En el test de variación estacional para múltiples grupos, la hipótesis alternativa es la variación sinusoidal anual para cada grupo. Por lo tanto uno escoge $t_1 := 1$ para los

varones y $t_2 := 1$ para las hembras. El valor del estadístico del test es $T := T_{t_1} + T_{t_2} = 23.30 + 11.59 = 34.89$, con valor de probabilidad de $p = 0.0000005$. De esta manera uno concluye que al parecer ambos grupos muestran variación sinusoidal anual, y esta inferencia es más intensa para los varones ya que $T_{t_1} > T_{t_2}$.

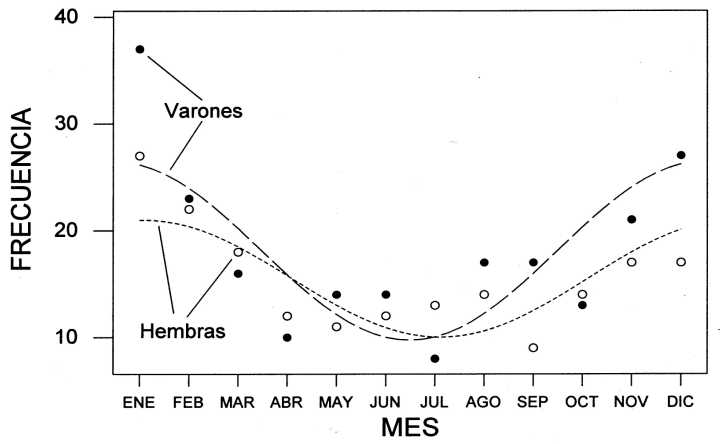

Figura 1: Frecuencia mensual de diagnosis de diabetes mellitus en jóvenes de 10 a 17 años, Colorado, 1978–1983: datos y modelo sinusoidal anual para cada sexo.

Visto que los dos grupos aparentan tener el mismo tipo de variación estacional, uno aplica el test de homogeneidad. El resultado del estadístico es $U_2 = 1.34$, con valor de probabilidad de $p = 0.5117$. Por lo tanto uno concluye que no hay diferencia significativa entre las variaciones estacionales de los dos grupos.

Los resultados de estos tests están de acuerdo con lo que se ve en la figura 1. En cada grupo los datos parecen tener una variación sinusoidal anual. La amplitud de los varones parece más grande, de acuerdo con $T_{t_1} > T_{t_2}$. Pero los dos grupos muestran más o menos el mismo tipo de variación, de acuerdo con el resultado del test de homogeneidad.

Para análisis adicionales los datos se juntan en un solo grupo. Para los datos juntos, la coordenada polar $\Theta = 22.1°$. Por lo tanto uno infiere que la población parece tener frecuencia máxima en la primera mitad de enero y frecuencia mínima en la primera mitad de julio. Esto está de acuerdo con lo que se ve en la figura 2, donde los datos son descritos por el modelo sinusoidal anual

$$n_i = \alpha_0 + \alpha_1 \cos \frac{\pi}{6} i + \beta_1 \operatorname{sen} \frac{\pi}{6} i + e_i,$$

y las estimaciones de los parámetros (con los respectivos errores estándares) por el método de mínimos cuadrados son $\widehat{\alpha}_0 = 33.6$ (2.3), $\widehat{\alpha}_1 = 12.7$ (3.3) y $\widehat{\beta}_1 = 5.1$ (3.3).

De acuerdo con este análisis uno concluye que este tipo de diabetes pudiera tener una etiología infecciosa o viral que aparenta ser igual para ambos sexos.

5.2 Ejemplo 2

Reunidos en doce totalidades mensuales, los datos son el número mensual de casos de atresia biliar extrahepática, tal como fueron registrados en la región del norte de Texas durante 1972–80. De enero a diciembre la lista de datos es (0, 0, 3, 2, 1, 1, 3, 5, 5, 4, 2, 4).

La hipótesis alternativa es la variación unimodal con un intervalo modal de tres meses. Usando el índice $t := 1$ en el test de variación estacional, uno obtiene el resultado del estadístico $T_1 = 6.75$, con valor de probabilidad de $p = 0.0342$. Esto lleva a la conclusión que la distribución de las probabilidades multinomiales no parece ser uniforme a través del año. En comparación, una aplicación del chi cuadrado de Pearson con 11 grados de libertad resulta en un valor de probabilidad de $p = 0.2330$. Este es un caso concreto de la superioridad en potencia de estos métodos sobre el chi cuadrado cuando el tamaño de la muestra es bajo.

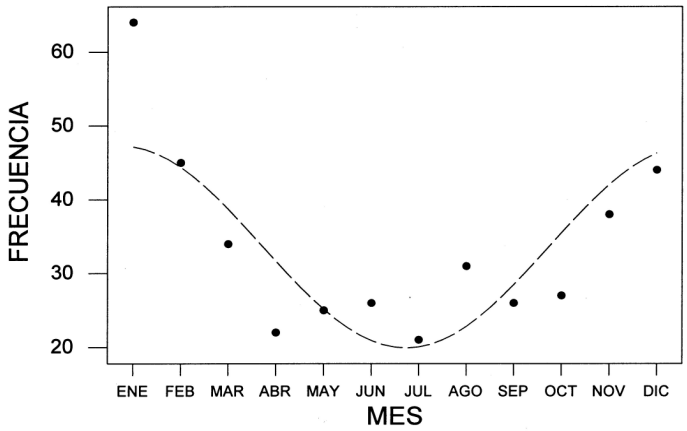

Figura 2: Frecuencia mensual de diagnosis de diabetes mellitus en jóvenes de 10 a 17 años, Colorado, 1978–1983: datos y modelo sinusoidal anual.

La modelación se hace usando un modelo unimodal rectangular (Marrero 1999). En este modelo las probabilidades multinomiales son iguales a $1/(12+\beta x)$, excepto durante los x meses del intervalo modal cuando dichas probabilidades aumentan a $(1+\beta)/(12+\beta x)$. La *altura* δ de este modelo es $\delta := \beta/(12+\beta x)$. La estimación del parámetro β por el método de máxima verosimilitud es $\widehat{\beta} := (12M - nx)/\{x(n-M)\}$, donde M es la totalidad de las frecuencias observadas durante los x meses del intervalo modal. Esta estimación tiene una varianza aproximada de $\text{estvar}_1(\widehat{\beta}) = \{Mn(12-x)^2\}/\{x^2(n-M)^3\}$. La altura se estima sin sesgo por $\widehat{\delta} := (12M - nx)/\{nx(12-x)\}$, con $\text{estvar}_1(\widehat{\delta}) := \{144M(n-M)\}/\{n^3x^2(12-x)^2\}$.

La coordenada polar $\Theta = 268.7°$. Por lo tanto uno estima que la mitad del intervalo modal se encuentra alrededor del medio de septiembre; por consiguiente, uno concluye que el intervalo modal va desde el principio de agosto al final de octubre. La estimación del parámetro del modelo es $\widehat{\beta} = 1.6$, con $\text{estvar}_1(\widehat{\beta}) = 0.9$. Además, la estimación de la altura es $\widehat{\delta} = 0.1$, con $\text{estvar}_1(\widehat{\delta}) = 0.002$. La figura 3 muestra este modelo y los datos.

Habitualmente uno interpreta los resultados de un estudio estadístico con prudencia. En este ejemplo la interpretación de los resultados pide precauciones especiales porque el tamaño de la muestra es muy bajo. De todas maneras, uno puede proponer una explicación. De acuerdo con el test de variación estacional, la repartición de estos datos a través del año no es uniforme; hay algo que explicar. En la figura 3 uno puede ver que los datos están generalmente de acuerdo con el modelo, excepto los datos de enero, febrero y diciembre. Además, los valores de los datos son generalmente más grandes durante la segunda mitad del año. Lo que se ve en la figura 3 está suficientemente bien de acuerdo con la conclusión que resulta del test de variación estacional. Entonces tiene sentido pensar que un agente ecológico está implicado en la etiología de esta enfermedad.

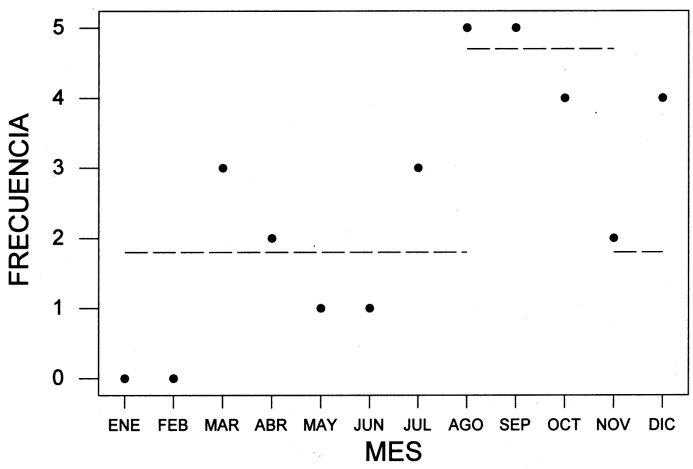

Figura 3: Frecuencia mensual de casos de atresia biliar extrahepática, norte de Texas, 1972–1980: datos y modelo unimodal rectangular.

Después de reflexionar sobre varias posibilidades, Strickland & Shannon llegaron a la conclusión siguiente: lo más plausible parece ser que una substancia tóxica sea el agente ecológico que está envuelto en la etiología de la enfermedad. Entonces Strickland & Shannon consideraron la relación entre el intervalo modal de los datos y el período de actividad del supuesto agente ecológico. Estos autores dieron algunas razones para pensar que el período de mayor intensidad de tal agente ecológico duraría de la mitad de marzo a la mitad de junio. Ya que este período de marzo a junio coincide con una época de actividad agrícola intensa en el norte de Texas, Strickland & Shannon han sugerido que este agente ecológico está relacionado quizás a toxinas que salen de la tierra cuando uno la trabaja, a insecticidas o a abonos.

Agradecimientos

El autor agradece la ayuda del editor de la *Revista*, Sr. Prof. Javier Trejos, con los comandos de LaTeX en la preparación de este artículo.

Referencias

[1] Jones, R. H.; Ford, P. M.; Hamman, R. F. (1988) "Seasonality comparisons among groups using incidence data", *Biometrics* **44**: 1131–1144.

[2] Marrero, O. (1979) *Statistical Tests for Seasonality in Epidemiological Data*. M.P.H. essay, Department of Epidemiology and Public Health, Yale University, New Haven, Connecticut.

[3] Marrero, O. (1983) "The performance of several statistical tests for seasonality in monthly data", *Journal of Statistical Computation and Simulation* **17**: 275–296.

[4] Marrero, O. (1984) "Statistical testing for seasonality in data with multiple peaks and troughs", *Biometrical Journal* **26**: 591–608.

[5] Marrero, O. (1988) "The power of a nonparametric test for seasonality", *Biometrical Journal* **30**: 495–502.

[6] Marrero, O. (1992) "A maximum rank-sum test for one-pulse variation in monthly data", *Biometrical Journal* **34**: 485–500.

[7] Marrero, O. (1998) "Multigroup analysis of seasonal variation: assessing the homogeneity of multiple cyclically ordered multinomial distributions", *Environmetrics* **9**: 151–163.

[8] Marrero, O. (1999) "L'analyse de la variation saisonnière quand l'amplitude et la taille sont faibles", *Revue Canadienne de Statistique (Canadian Journal of Statistics)* **27**: 875–882.

[9] Strickland, A. D.; Shannon, K. (1982) "Studies in the etiology of extrahepatic biliary atresia: time-space clustering", *Journal of Pediatrics* **100**: 749–753.