Suavizando los histogramas

Sr. Director:

Hemos leído con interés la excelente revisión de Sánchez-Cantalejo y Ocaña-Riola[1]. Los autores presentan de una manera clara y comprensible los métodos de regresión no paramétrica y su utilización como métodos de diagnóstico en los modelos de regresión paramétrica. Tal y como indican los autores, debido a la amplia disponibilidad del software actual, las técnicas de alisamiento están siendo cada vez más utilizadas en la actualidad para estudiar los problemas relacionados con la salud.

Nuestro objetivo es ampliar, en la medida de lo posible, la revisión Sánchez-Cantalejo y Ocaña-Riola[1] presentando la técnica de alisamiento, o suavizado, para representaciones gráficas univariantes. Dentro de estos gráficos, el histograma el más utilizado para representar distribuciones univariantes con la finalidad de estimar la función de densidad de la variable a estudiar. Por ejemplo, consideremos una cierta variable aleatoria *X* con función de densidad *f* y supongamos que disponemos de una muestra aleatoria, *x*1,.....,*x*,, (de valores 1,5, 2,5, 3,2, 3,3, 3,7, 3,9, 5, 5,5, 6,5 y 7), extraída de la función de densidad desconocida f . La utilización del histograma como estimador de funciones de densidad presenta algunas limitaciones dependiendo del contexto en que se utiliza, por ejemplo en el análisis discriminante las discontinuidades que se dan en el histograma causan muchas dificultades si se necesitan calcular las derivadas de los estimadores[2]. Por otro lado, la representación gráfica del histograma tradicional depende en gran medida del punto x_0 que se considera como origen, así como del número y la amplitud de los intervalos de clase[3], tal y como presentamos con los datos del ejemplo (figuras 1a, 1b y 1c).

La alternativa al histograma es utilizar el estimador *naive.* Podemos deducir la expresión de este estimador a partir de la definición de densidad de probabilidad. Si la variable aleatoria *X* tiene por función de densidad *f,* entonces

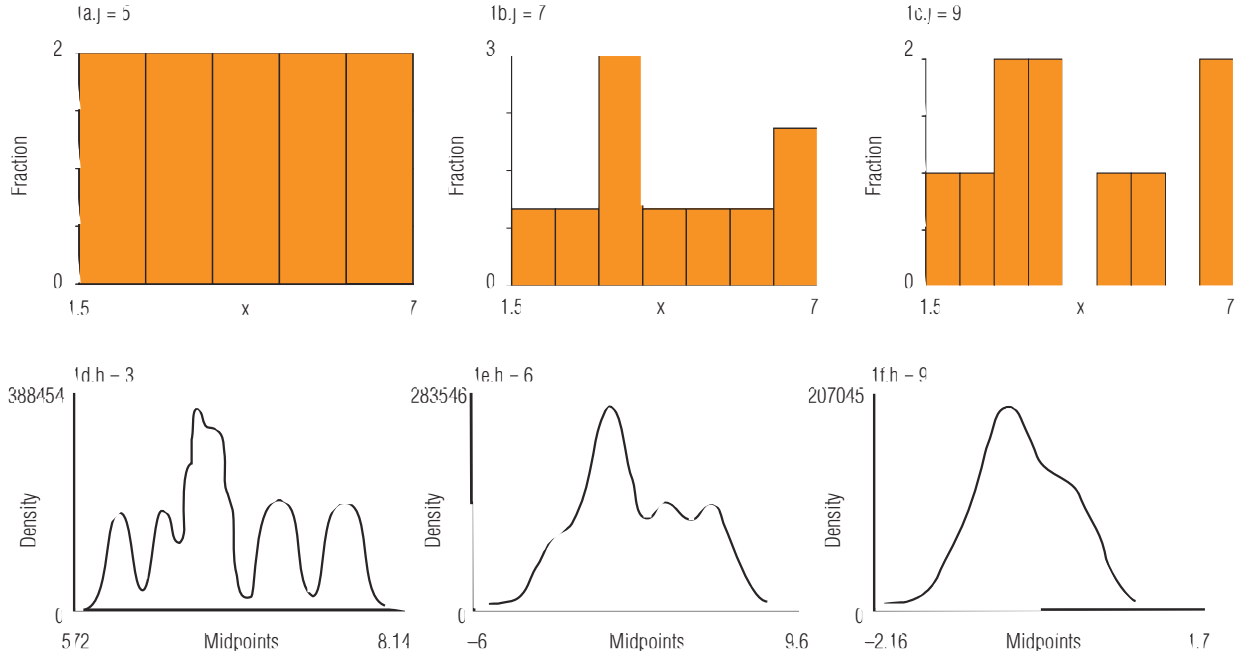
$$f(x) = \lim_{h \to 0} \frac{1}{2h} P(x - h < X < x + h)$$

Así, un estimador $\hat{f}$ para nuestra función desconocida *f,* será

$$\hat{f}(x) = \frac{1}{2hn} * k$$

(donde *k* es el número de observaciones generadas en el intervalo *(x – h, x + h),* eligiendo un valor pequeño para *h* (que normalmente varía entre 0 y 1). Este estimador es lo que se

Figura 1. Histograma tradicional (según el número de intervalos de clase, *j*), e histograma suavizado con una función *kernel Gaussiana* (según el parámetro de alisamiento, *h)*



conoce como estimador *naive.* La expresión anterior puede ser escrita de la siguiente forma

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}\,w\left(\frac{x - x_i}{h}\right) \text{ siendo } w(x) = \begin{cases} \frac{1}{2} & |x| < 1 \\ 0 & c.c \end{cases}$$

Podemos pensar que este estimador es un intento de construir un histograma donde cada punto es el centro de su respectiva clase, logrando así que no dependa tanto de la elección del origen. La elección de la amplitud de cada clase depende del parámetro *h,* que controla el valor mediante el cual los datos son suavizados. La generalización del estimador *naive* es el estimador *kernel.* Sustituyendo la función *w* por la función kernel *K,* el estimador *kernel* se define como

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n}\left(\frac{x - x_i}{h}\right)$$

donde *h* es el parámetro de alisamiento. La función *K* determina la forma de estas funciones mientras que el parámetro *h* determina su amplitud. Pero así como la elección de la función K (rectangular, triangular, Gaussiana o Epanechinkov) no influye de manera sustancial[3], la adecuada elección del parámetro *h* es muy importante. Si elegimos un valor muy pequeño de *h,* introducimos ruido inherente al sistema (figura 1d) pero si lo elegimos demasiado grande, entonces el estimador *kernel* será demasiado suavizado y características importantes, como la multimodalidad de los datos, podrían quedar ocultadas (figura 1f). Es por ello que la elección del parámetro h siempre involucra un balance comparativo entre estas dos consideraciones[3,4] (figura 1e).

Como conclusión, el histograma suavizado evita algunos de los problemas más frecuentes que se encuentran con los histogramas tradicionales, representando con mayor claridad la naturaleza continua de los datos. Su construcción con el software actual resulta tan sencilla como la del histograma tradicional, en especial con el paquete estadístico STATA[5,6]. Sin embargo su forma final todavía se ve afectada, aunque en menor medida respecto al histograma tradicional, por dos decisiones relativamente arbitrarias, la elección de la función *kernel* y del parámetro de alisamiento.

Bibliografía

1. Sánchez-Cantalejo E, Ocaña Riola R. Actualizaciones en regresión: suavizando las relaciones. Gac Sanit 1997;11: 24-32.

2. Härdle W. Smoothing Techniques, with Implementation in S. New York: Springer-Verlag; 1991.

3. Jacoby WG. Statistical Graphics for Univariate and Bivariate Data. CA: Sage; 1997.

4. Salgado-Ugarte IH, Shimizu M, Taniuchi T. snp6.2: Practical rules for bandwidth selection in univariate density estimation. Stat Technical Bulletin 1995;25:5-19.

5. Salgado-Ugarte IH, Shimizu M, Taniuchi T. snp6: Exploring the shape of univariate kernel density estimators. Stat Technical Bulletin 1993;16:8-19.

6. Salgado-Ugarte IH, Shimizu M, Taniuchi T. snp6.1: WARPing and kernel density estimation for univariate data. Stat Technical Bulletin 1995;26:23-31.