GEOS, Vol. 18, N0. 3, p. 174-179

HISTOGRAMAS REPARTIDOS PARA EL ANÁLISIS DE MUESTRAS ESTADÍSTICAS ESCASAS

F. Alejandro Nava
Depto. de Sismología, División de Ciencias de la Tierra, CICESE
A.P. 2732, Ensenada, 22800, B.C., México
E-mail: fnava@cicese.mx

RESUMEN

Se presenta un método de construcción de histogramas para el análisis de muestras escasas, que considera cada muestra repartida entre los intervalos de clase (celdas) según un criterio de contención o según las incertidumbres en la medición, y que reduce la dependencia de los histogramas en el tamaño y la localización de las celdas. Empíricamente parece que los histogramas producidos por este método pueden, en general, representar mejores aproximaciones a la distribución de la población madre de procesos naturales que los histogramas comunes. El método puede aplicarse al análisis de otros tipos de observaciones que pueden considerarse como histogramas ponderados.

INTRODUCCIÓN

En multitud de campos de la investigación científica, y de muchos otros tipos de investigaciones, es común el problema de tratar de identificar o estimar la distribución estadística $p(v)$ de una población V que tiene M elementos (M puede ser infinito), a partir de una muestra $\{v_j\}$, donde cada v_j es un dato (producto de la observación) identificado por el índice j=1,2,...,N, donde N es el número total de datos. Este problema es muy importante, pues conocer la distribución $p(v)$, que llamaremos *distribución madre*, permite hacer inferencias probabilísticas respecto a la ocurrencia de eventos de V. Por ejemplo, con base en las magnitudes observadas para sismos ocurridos en una falla dada, podemos tratar de determinar cuál es la distribución de magnitud para dicha falla y de allí calcular la probabilidad de ocurrencia de un sismo con una magnitud determinada.

Una herramienta muy utilizada para atacar este problema de estimación de distribuciones es la visualización de la información contenida en la muestra mediante histogramas. Un histograma es la representación del número n_i de elementos de la muestra cuyos valores están contenidos dentro de un intervalo que abarca de $v_1 + i\,\Delta v$ a $v_1 + (i + 1)\,\Delta v$, donde Δv es el tamaño de los intervalos de clase, o *celdas*, v_1 es un valor de referencia que indica el comienzo de la *i*ésima celda, e i es un entero que toma los valores necesarios para que las celdas cubran (por lo menos) toda la extensión de los N valores observados.

Es bien sabido que, conforme crece el tamaño de la muestra y la razón $N/M \rightarrow 1$ el histograma aproxima cada vez mejor a la distribución de la población. Una muestra abundante permite el uso de Δv pequeña y esto hace que el histograma pueda aproximarse con mayor detalle a la distribución.

Sin embargo, es muy común el caso de que no se cuente con una muestra suficientemente abundante; un ejemplo es el de las muestras de sismos grandes de una región sismogénica dada, pues éstos tienen usualmente tiempos de recurrencia de varias decenas a cientos de años, mientras que los registros instrumentales tienen apenas unas cuantas decenas de años, de manera que es imposible contar con muestras abundantes. Cuando la muestra es escasa la forma del histograma depende críticamente de la elección de Δv y v_1, y no se parecerá necesariamente a la distribución madre. Obviamente, el número de celdas debe ser menor que el número de muestras, lo que fija un límite inferior para Δv; por su parte, el efecto de la elección de v_1 es cíclico con período Δv y cambiar v_1 por α ($\alpha \leq \Delta v$) equivale a cambiarlo por Δv-α.

La discusión que sigue será ilustrada por un ejemplo sintético de muestras generadas a partir de la distribución $p(v)$ mostrada en la Figura 1, que es una distribución sencilla (como son usualmente las distribuciones para los procesos naturales) cuya forma se presta para una fácil identificación a partir de histogramas. La Figura 2 muestra el ajuste a la distribución de muestra de un histograma con N=50,000 muestras generadas aleatoriamente a partir de la distribución madre, Δv=0.5 y v_1=-5.0, normalizado para tener área unitaria. Las áreas sombreadas indican el error residual entre la distribución y el histograma y puede apreciarse que el histograma de una muestra abundante ajusta bastante bien a la distribución madre; el área total (valor absoluto) de error es 0.01521.

Para ilustrar el tratamiento de muestras escasas, consideraremos una muestra que consiste de los primeros 50 elementos de los 50,000 utilizados antes, indicados por líneas verticales discontinuas en la Figura 3. Esta muestra abarca una extensión de -3.9118 a 3.7891; el espaciamiento entre muestras (la distancia horizontal entre las líneas que indican las muestras) máximo es 1.0295, el siguiente menor al máximo es 0.5430, el mínimo es 0.0005 y el promedio es 0.1572

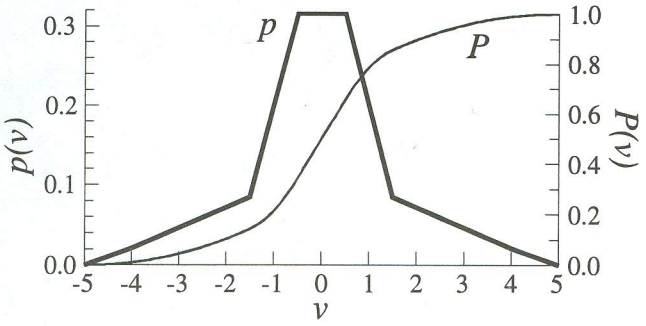

Figura 1. Distribución $p(v)$, referida al eje de la izquierda, y su cumulativa $P(v)$, referida al eje de la derecha. La variable v puede considerarse con unidades arbitrarias o adimensional.

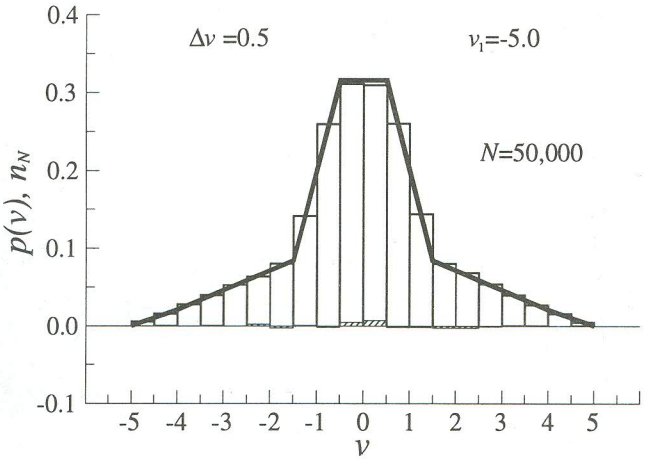

Figura 2. Ajuste de un histograma construido a partir de N = 50,000 muestras (barras con línea de grueso intermedio), a la distribución $p(v)$ (línea gruesa). n_N es el número de muestras en cada celda normalizado de forma que el área bajo el histograma sea unitaria. El área sombreada es la diferencia entre la distribución madre y el histograma.

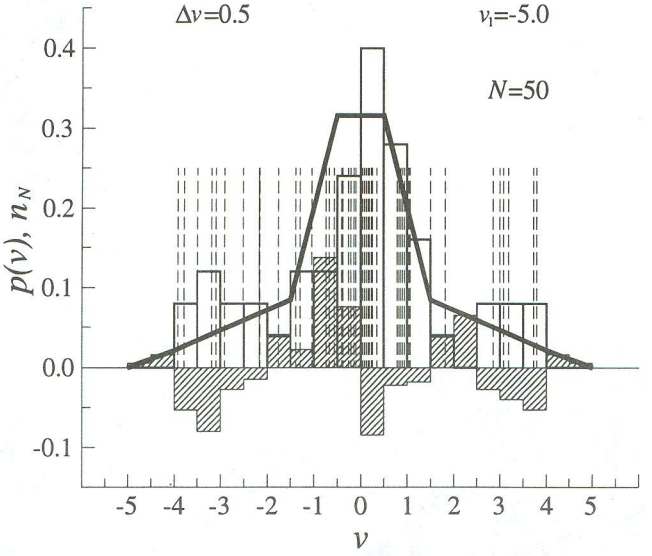

Figura 3. Ajuste de un histograma para N = 50 muestras normalizado. Convenciones como en la Figura 2. Las líneas discontinuas verticales indican la posición de los eventos.

(usaremos cuatro cifras decimales pues la aproximación necesaria para discutir las cantidades que encontraremos a lo largo de la discusión). Espaciamientos máximos tan superiores al promedio indican precisamente qué tan escaso es el muestreo.

La Figura 3 es un ejemplo de un histograma construido a partir de la muestra escasa, Δv=0.5 y $v_1 = 5.0$ como en el ejemplo anterior. Puede apreciarse que el error, con área total de 0.4128 es mucho mayor que para la muestra abundante y que la forma del histograma, si bien indica una mayor probabilidad cerca de los valores centrales de v, parece indicar un máximo desplazado a la derecha.

Un ejemplo de qué tanto cambian el histograma y su ajuste al cambiar el valor de referencia v_1 puede verse en la Figura 4, donde se muestra como referencia el mismo histograma de la figura anterior y, con línea gruesa, el histograma resultante de utilizar $v_1 = -4.8$, que tiene un área de error de 0.5550 (¡más del 50%!). Las líneas discontinuas verticales señalan los valores que toman los datos; es fácil ver como un corrimiento de las celdas hace que algunos datos cambien de celda, y para muestras escasas ésto puede cambiar bastante la forma del histograma, como sucede en este ejemplo.

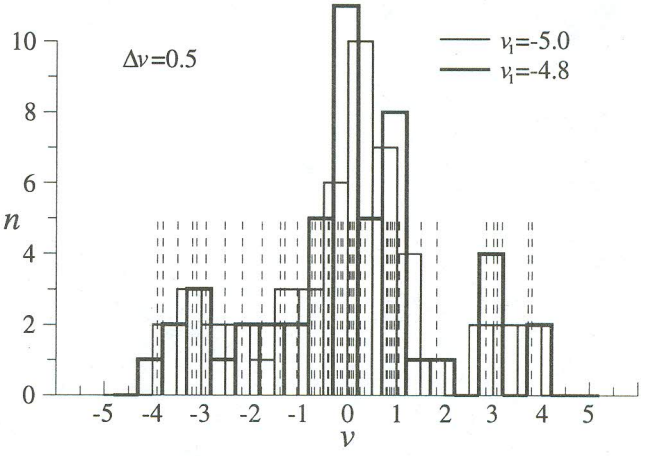

Figura 4. Comparación entre el histograma de la Figura 3 (línea delgada) y un histograma con el mismo tamaño de celdas pero inicio ligeramente distinto (línea gruesa). Las líneas discontinuas verticales indican la posición de los eventos.

El efecto de cambiar el tamaño de las celdas se ilustra con un ejemplo en la Figura 5, que muestra con línea delgada el histograma de referencia de la Figura 3 y con línea gruesa el histograma resultante de usar un tamaño de celda menor, que tiene un error de 0.5549.

Lo crítico de la elección de tamaño de celda y valor de referencia queda ilustrado por estos ejemplos.

MÉTODO

El método propuesto consiste en repartir cada muestra, considerada con área unitaria, sobre una extensión centrada en el valor medido y de acuerdo con alguna distribución de reparto

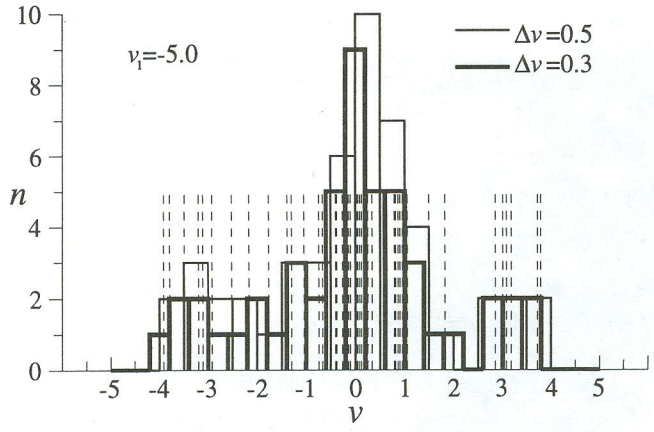

Figura 5. Comparación entre el histograma de la Figura 3 (línea delgada) y un histograma con el mismo valor de referencia pero celdas más pequeñas (línea gruesa). Las líneas discontinuas verticales indican la posición de los eventos.

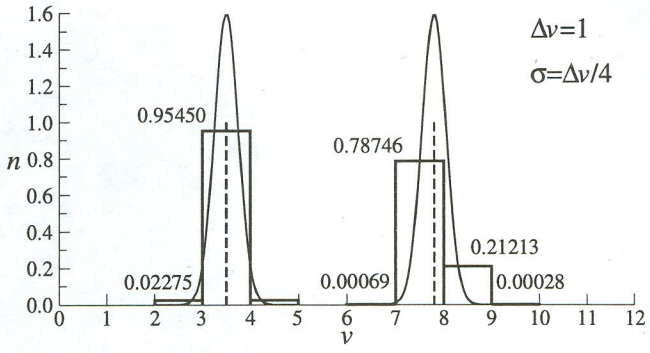

Figura 6. Ejemplo de repartición de dos datos (líneas discontinuas) cuyos valores se encuentran a diferentes distancias dentro de sus celdas respectivas, mediante el uso de la distribución Gaussiana con (línea delgada). El número sobre cada celda del histograma (línea gruesa) indica el área de ésta.

g dada (ésto es, convolucionar la muestra con esta distribución que debe tener también área unitaria), de manera que la contribución de la muestra a la celda en que cae su valor y a las celdas cercanas a ésta ya no es todo/nada (ó 1/0), respectivamente, sino que depende del área de la distribución de reparto que cae dentro de cada celda.

Este proceso se ilustra en la Figura 6, que muestra el uso de una distribución normal (o Gaussiana)

$$g(v-v_i) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(v-v_i)^2}{2\sigma^2}}, \qquad (1)$$

con desviación estándar s y centrada sobre el valor v_i. El evento *i* contribuye a la celda *j*, que se extiende de h_j a $h_{j+1}=h_j+\Delta v$, el área de la Gaussiana a_{ij} contenida entre estos límites:

$$a_{ij} = \int_{hj}^{h_{j+1}} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(v-v_i)^2}{2\sigma^2}} dv; \qquad (2)$$

expresión que puede evaluarse fácilmente haciendo $x_j \equiv (h_j - v_i)/\sigma$, $x_{j+1} \equiv (h_{j+1} - v_i)/\sigma$ y

$$a_{ij} = \text{sgn}(x_j)\tfrac{1}{2}\text{erf}\left(\frac{x_j}{\sqrt{2}}\right) - \text{sgn}(x_{j+1})\tfrac{1}{2}\text{erf}\left(\frac{x_{j+1}}{\sqrt{2}}\right), \qquad (3)$$

donde sgn(x) es la conocida función *signum*:

$$\text{sgn}(x) \equiv \begin{cases} -1; x < 0 \\ 1; x \geq 0 \end{cases}, \qquad (4)$$

y la función de error es representable como:

$$\text{erf}\left(\frac{x}{\sqrt{2}}\right) = \sqrt{\frac{2}{\pi}}\, x\left[1 - \frac{x^2}{2\cdot 1!\cdot 3} + \frac{x^4}{2^2\cdot 2!\cdot 5} - \frac{x^6}{2^3\cdot 3!\cdot 7} + \ldots\right]; \quad [x^2 < \infty] \qquad (5)$$

(Dwight,1961), función ésta que, para fines prácticos, sólo requiere evaluarse para $x < 6$ y puede considerarse igual a 1 para $x \geq 6$.

En la Figura 6, las celdas de ancho unitario abarcan la distancia entre cada dos enteros del eje *v*; la posición de cada dato de este ejemplo, uno en $v_1 = 3.5$ y otro en $v_2 = 7.8$, se indica por una línea discontinua; la distribución normal centrada sobre cada dato se grafica con línea delgada; el valor de cada celda se ilustra con línea gruesa.

Para la gaussiana con $\sigma = \Delta v/4$, un dato que cae exactamente a la mitad de una celda, como el de la izquierda, contribuye a esta celda con el área de la gaussiana entre $\pm 2\sigma$ que es 0.95450, y a las dos celdas adyacentes con sólo 0.02275 a cada una, de manera que está casi completamente contenida en la celda a la que pertenece.

Por otro lado, el dato con v=7.8 de la Figura 6 se encuentra cerca del extremo derecho de la celda, y los valores en la figura indican cómo se distribuye su área entre las celdas próximas; la mayor parte corresponde a la celda que lo contiene, pero una parte significativa corresponde a la celda de la derecha. Un dato situado exactamente entre dos celdas contribuirá a ambas por igual.

Antes de considerar los resultados de este filtrado conviene discutir cuáles son los posibles significados de y/o justificaciones para repartir los datos entre las celdas de un histograma. Una justificación es la ya mencionada e ilustrada que tiene que ver con la manera en que consideramos que un dato debe estar contenido en una celda de ancho finito; vimos que $\sigma=\Delta v/4$ hace que un dato centrado en una celda quede mayormente contenido en ésta; $\sigma=\Delta v/6$ hace que dicho dato quede (para cualquier fin práctico) totalmente contenido, pero

si el dato se acerca al límite de la celda parte de su valor se asignará a la celda vecina, y si se encuentra entre dos celdas, se repartirá por igual en éstas; cuando $\sigma \rightarrow 0$ volvemos a un histograma usual, excepto por el caso de un dato en la frontera entre dos celdas, que en un histograma usual se asigna completamente a una celda u otra según algún criterio (usualmente arbitrario).

Por otro lado, aumentar σ hace que al dato se reparta más entre las celdas y veremos más adelante que ésto puede contribuir a eliminar efectos de baja longitud de onda introducidos por la falta de datos en una muestra escasa.

Otra justificación importante es la de la posible incertidumbre y/o error en la determinación de los datos; de hecho es difícil pensar en algún proceso de medición de datos de interés científico (y, en particular, geofísico) que esté exento de incertidumbre y/o error. Desde este punto de vista, repartir el dato según la distribución más apropiada para describir el posible error es un proceso enteramente razonable y, de hecho, encomiable, para que el histograma esté menos influido por los posibles errores.

Cabe mencionar que para ilustrar el método se ha escogido como $g(v)$ una distribución normal porque es una distribución apropiada para una gran cantidad de procesos físicos, pero el método puede utilizar cualquier distribución que se considere apropiada. Por ejemplo, el error en la determinación de tiempos de arribo de fases sísmicas no está distribuido normalmente, sino que tiene un sesgo hacia residuales negativos causado por la presencia de ruido en los sismogramas (Lomnitz y Nava, 1980).

La Figura 7 muestra el histograma repartido para la muestra escasa de 50 elementos, Δv=0.5, v_1 = -5.0 y $\sigma = \Delta v/4$, normalizado a área unitaria. Muestra también la distribución madre y la diferencia entre ésta y el histograma, que ha disminuido a 0.33624, de manera que el histograma repartido ajusta mejor a la distribución madre.

Tomando en cuenta que la distribución madre, como la gran mayoría de las distribuciones encontradas en la naturaleza, no tiene detalles de longitud de onda pequeña, se puede considerar el uso de una mayor desviación estándar. La Figura 8 muestra el efecto de utilizar $\sigma = \Delta v/2$, el área de error disminuye a 0.27971. Naturalmente, mientras mayor es σ, menos sensitivo es el histograma repartido a cambios en v_1 y en Δv.

Recordando que el efecto de la elección de v_1 es cíclico con período Δv, el siguiente paso en histogramas repartidos consiste en, una vez determinado el valor de σ, considerar celdas con $\Delta v < \sigma$. La Figura 9 muestra el efecto de repartir los dos datos considerados en la Figura 3, con σ=0.25 y $\Delta v = s/5$; por claridad se ha dibujado sólamente la parte superior de cada barra del histogramas (sin regresar al cero entre barras) y las pequeñas líneas verticales en el eje v indican las fronteras de las celdas. Se puede ver que los datos que antes se habían

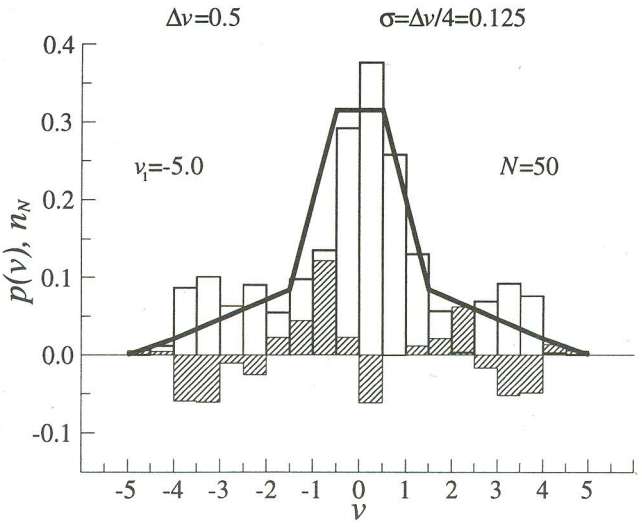

Figura 7. Ajuste del histograma repartido con Δv = 0.5 y σ = $\Delta v/4$ para N = 50 muestras y normalizado. Convenciones como en la Figura 2.

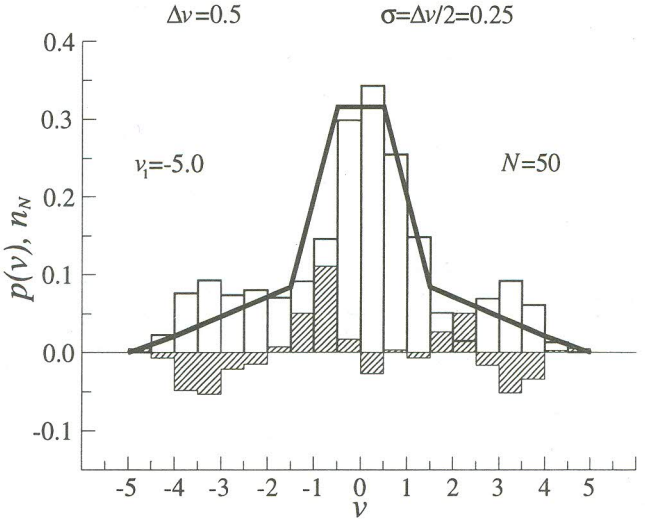

Figura 8. Ajuste del histograma repartido con Δv = 0.5 y σ = $\Delta v/2$ para N = 50 muestras y normalizado. Convenciones como en la Figura 2.

repartido de forma distinta entre celdas, ahora se reparten en forma idéntica. Naturalmente, si cambiamos v_1 por menos de Δv cambiará, para ambos datos de la misma manera, la forma en que se reparten, pero este cambio de forma ahora será sólamente de detalle como se muestra en la Figura 10, y no cambia significativamente la forma del histograma. Llevar la disminución de las celdas al límite $\Delta v \rightarrow 0$, equivale a substituir la muestra por la convolución de los datos con la distribución g:

$$g(v) * \sum_{j=1}^{N} \delta(v_j - v), \qquad (6)$$

La Figura 11 muestra la aplicación de este método a la muestra escasa considerada antes, con σ=0.25 y $\Delta v = \sigma/5$=0.05.

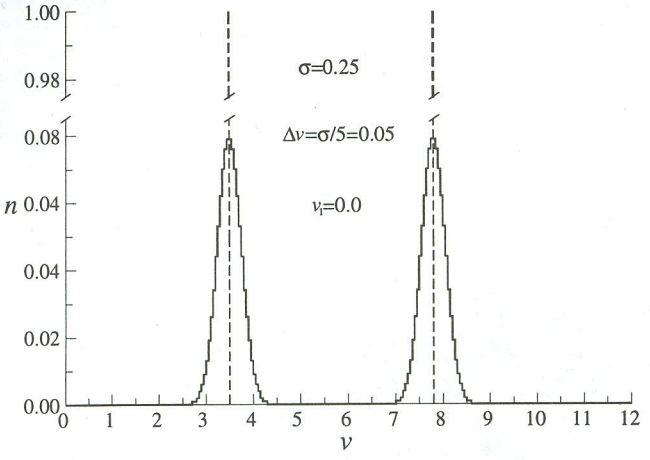

Figura 9. Ejemplo de repartición de los datos mostrados en la Figura 6, para σ = 0.25 y Δν = σ/5. La suma de los elementos contenidos en cada celda es igual al número de valores (2 en este caso).

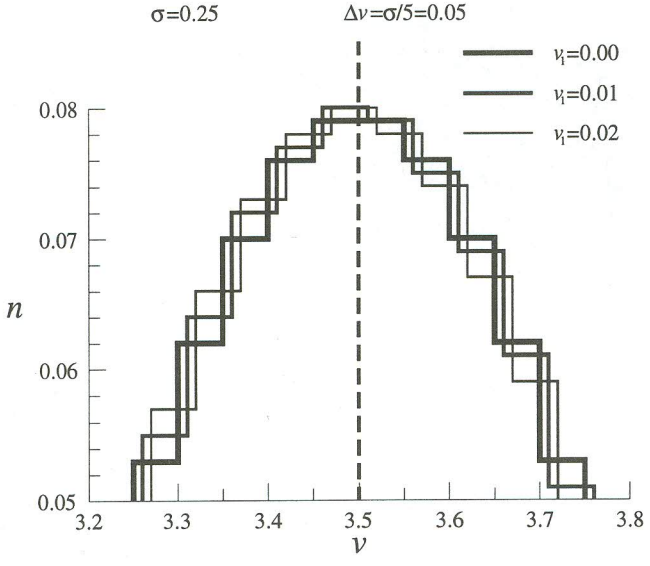

Figura 10. Ejemplo de las diferencias en la repartición de un evento para diferentes valores de referencia v_1.

La forma del histograma es muy robusta pues depende ya muy poco de la elección de v_1 y casi no cambia ante variaciones de Dν. El área de error para este histograma es de 0.33894, lo que significa que no ajusta tan bien a la distribución madre como el histograma de la Figura 8 que usaba la misma σ, pero hay que tener en cuenta el filtrado pasabajos implícito en el uso de celdas, que es más fuerte conforme más ancha es la celda, de manera que para lograr una distribución de datos equivalente con celdas más pequeñas es necesario aumentar σ. Efectivamente, como puede verse en la Figura 12, una mayor σ=0.5 (del orden de la mitad del máximo espaciamiento entre muestras) para el mismo tamaño de celda (Δν=0.05=σ/10) se parece mucho más a la distribución madre y disminuye el área de error a 0.23588.

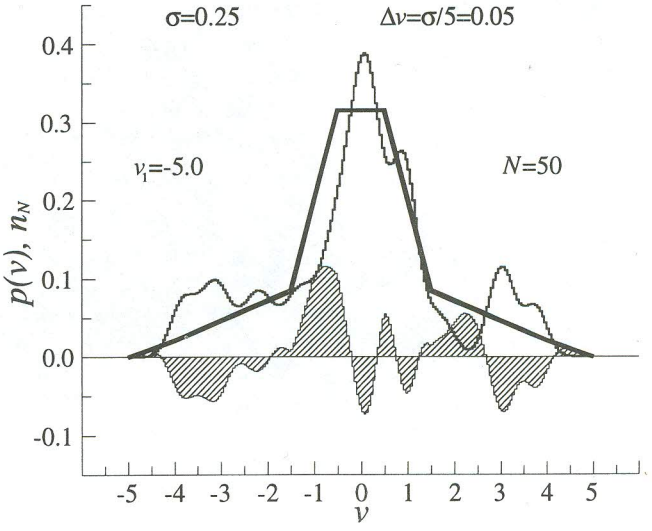

Figura 11. Ajuste del histograma repartido con σ = 0.25 y Δν = σ/5 para N = 50 muestras y normalizado. Convenciones como en la Figura 2.

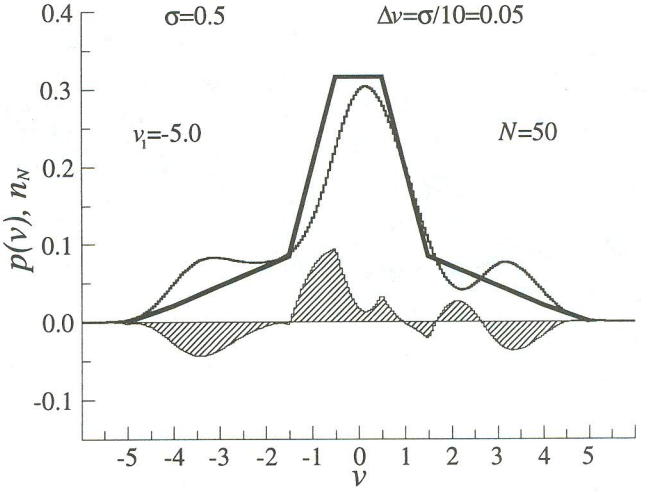

Figura 12. Ajuste del histograma repartido con σ = 0.5 y Δν = σ/10 para N = 50 muestras y normalizado. Convenciones como en la Figura 2.

Se ha aplicado este método a diferentes muestras escasas generadas a partir de varias distribuciones madre (distribuciones con formas razonables para representar procesos de interés geofísico) y en cada caso han resultado histogramas que, como en el ejemplo mostrado, se ajustan a éstas con menor error que los histogramas comunes.

CONCLUSIONES Y DISCUSIÓN

Para muestras escasas de distribuciones madres conocidas, se encuentra empíricamente que los histogramas repartidos se ajustan mejor a las distribuciones madre que los histogramas comunes. Esto significa que, cuando se trata de estimar o identificar una distribución madre desconocida a partir de una

muestra escasa, el uso de histogramas repartidos puede hacer más fácil esta tarea.

Además, los histogramas repartidos son menos dependientes del tamaño y la localización de las celdas; de hecho, para celdas mucho menores que la desviación estándar de la distribución de reparto, los histogramas repartidos son casi independientes de estos parámetros.

Aparentemente, distribuir los valores observados contrarresta parcialmente la falta de información implícita en una muestra escasa; para una distribución de reparto Gaussiana, una desviación estándar del orden de la mitad del máximo espaciamiento da buenos resultados.

Distribuir las muestras según una distribución de reparto cuya forma y anchura (o desviación estándar) correspondan a la posible incertidumbre en el valor de las muestras, hace el histograma más confiable, puesto que será menos dependiente de los errores en el proceso de muestreo.

Si bien conviene suponer que toda la muestra comparte la misma distribución de error, es posible utilizar distintas desviaciones estándar para cada valor, que reflejen la confianza que se tiene en la determinación de éste. Por ejemplo, si estamos observando la distribución de magnitudes sísmicas para sismos ocurridos en una región dada, la confianza en la determinación de cada magnitud depende, entre otros factores, del tamaño del sismo, del número de estaciones que lo hayan registrado (sin saturarse) y de la localización hipocentral, de manera que con base en estos factores puede asignarse a cada magnitud un número que indique qué tan confiable es y, por ello, qué tanto debe distribuirse.

Por último, mencionaremos la aplicación del método propuesto al análisis de información mediante lo que podemos llamar *histogramas ponderados*. Por ejemplo, el daño que causa un sismo depende de su magnitud, pero también depende de la presencia de poblaciones o construcciones u otros sistemas vulnerables, y de qué tan vulnerables sean; así, si queremos estudiar los daños causados por sismos como función de la magnitud, podemos hacerlo agrupando los daños causados por sismos cuyas magnitudes caen dentro de un intervalo de clase, ésto es, en vez de la serie considerada en (6) que asigna a cada evento un peso unitario, consideramos la serie

$$\sum_{j=1}^{N} A_j \delta \quad (v_j - v) \qquad (7)$$

que asigna a cada valor v_j un peso A_j, que para el ejemplo serían la magnitud de cada sismo y el daño causado por éste.

AGRADECIMIENTOS

Agradezco a Luis Delgado y a Juan García Abdeslem su interés y entusiasmo, a José Frez sus comentarios y a dos árbitros anónimos sus útiles recomendaciones.

REFERENCIAS

Dwight, H. (1961) Tables of integrals and other mathematical data. The Macmillan Co., New York, 336pp.

Lomnitz, C. y Nava,F. (1980) A method for epicenter determination. Geofísica Internacional, 19, 45-62.