Universidad de Oviedo

Asepelt España

Comunicaciones

XIV Reunión

UNA METODOLOGÍA PARA LA CONSTRUCCIÓN DE HISTOGRAMAS. APLICACIÓN A LOS INGRESOS DE LOS HOGARES ANDALUCES

Luis González Abril - luisgon@cica.es
José Manuel Gavilán Ruíz - jmgavir@cica.es
Universidad de Sevilla

Anales de Economía Aplicada

Oviedo 2 3 4
Junio 2000

Asepelt España

ISBN: 84-699-2357-9

Luis González Abril
Dpto. de Economía Aplicada I
Universidad de Sevilla
e-mail: luisgon@cica.es

José Manuel Gavilán Ruíz
Dpto. de Economía Aplicada I
Universidad de Sevilla
e-mail: jmgavir@cica.es

UNA METODOLOGÍA PARA LA CONSTRUCCIÓN DE HISTOGRAMAS. APLICACIÓN A LOS INGRESOS DE LOS HOGARES ANDALUCES.

Resumen

Es común agrupar los valores de una variable en una serie de intervalos cuando ésta toma una gran cantidad de valores diferentes y presentar el Histograma como un elemento que nos informa, entre otras cosas, sobre la forma de la variable (moda, asimetría, ...). Posteriormente es interesante suavizar la gráfica obtenida para conseguir una curva más parecida a una posible distribución teórica (modelo) de la variable.

En este contexto, el presente trabajo consta de dos partes: En primer lugar, a partir de un conjunto de datos y de un cierto número de intervalos dados resolvemos el problema de agrupar la variable en base a la minimización de la media de las desviaciones típicas dentro de cada intervalo formado[1]. En segundo lugar, nos proponemos suavizar el Histograma obtenido mediante una función splín parabólica. Para llevar a cabo ambas tareas implementamos sendos programas dentro del paquete MATHEMÁTICA.

Se finaliza el trabajo mostrando el resultado de la aplicación de la metodología desarrollada a la variable Ingresos Totales de los Hogares Andaluces con los 3665 datos obtenidos a partir de la EPF 90-91. Comparamos los resultados con los obtenidos al representar el Histograma usando el mismo número de intervalos (basado en la regla de Sturges), pero con amplitudes constantes, lo que se obtendría por defecto al usar cualquier software estadístico estándar. Se concluye que este último método es totalmente inadecuado ya que no refleja aspectos

[1] Este método aproximado pero fácil está inspirado en la lectura de la Sección 5A.7 del texto de Cochran (1976).

fundamentales de la forma de la variable, como en este caso es la bimodalidad, lo que si queda recogido con la metodología propuesta.

1. UNA CONSTRUCCIÓN DE HISTOGRAMAS: LA REGLA CUM

Vamos a dar aquí un método aproximado pero fácil para obtener los intervalos que conforman un histograma a partir de un número elevado de datos[2] observados de una variable *X* con función de densidad desconocida *f* y del número de intervalos deseados I.

Siendo $x_1,...,x_n$ las observaciones de la variable *X* ordenadas en forma creciente, las agrupamos en primer lugar usando una partición auxiliar en intervalos de amplitud constante a>0 del recorrido de la variable:

$$A_0 = x_0 < A_1 < ... < A_{m_I} = x_n,$$

aquí m_I es el número de intervalos de esta partición inicial, que se elegirá suficientemente grande como para que *f* sea aproximadamente constante en cada intervalo $[A_{i-1},A_i]$, llamaremos f_i^A al valor de dicha constante, i=1,...,m_I.

El objetivo es construir los I intervalos $[C_0,C_1]$, $[C_1,C_2]$,..., $[C_{I-1},C_I]$ que darán lugar al histograma con $C_0=x_0$ y $C_I=x_n$ y de manera que cada uno de ellos se obtenga uniendo un cierto número de subintervalos de la partición inicial; es decir, para cada h=1,...,I

$$[C_{h-1},C_h] = \bigcup_{i=m_{h-1}+1}^{m_h} [A_{i-1},A_i],$$

donde m_h es el número de intervalos de la partición inicial en $[C_0,C_h]$ (por definición se ha tomado $m_0=0$). De esta manera se tendrá

$$C_0 = A_{m_0}, C_1 = A_{m_1}, ..., C_I = A_{m_I}.$$

Así se elegirán los extremos C_h, h=1,...,I-1 de modo que la media de las desviaciones típicas en los intervalos finalmente construidos sea mínima; es decir que

[2] Necesitamos un elevado número de observaciones para que las aproximaciones de las probabilidades de ciertos intervalos que se considerarán por sus frecuencias relativas sean aceptables. Dando esta hipótesis por válida, nosotros escribiremos igualdades entre ambas cantidades.

$$\sum_{h=1}^{I} W_h S_h$$

sea mínima, donde W_h es la frecuencia relativa en el intervalo h-ésimo $[C_{h-1},C_h]$ y S_h la desviación típica en dicho intervalo, h=1,...,I.

Si estos intervalos $[C_{h-1},C_h]$ son numerosos y estrechos se tendrá que *X* se distribuirá uniformemente (aproximadamente) en cada uno de estos intervalos y así podremos considerar que *f* es constante en cada uno de los mencionados intervalos $[C_{h-1},C_h]$ (llamaremos f_h^C a dicha constante) y

$$S_h = \frac{C_h - C_{h-1}}{\sqrt{12}}.$$

Por otra parte, y de forma también aproximada, tendremos

$$W_h = \int_{C_{h-1}}^{C_h} f(x)dx = \int_{C_{h-1}}^{C_h} f_h^C dx = f_h^C \left(C_h - C_{h-1}\right),$$

y por tanto el objetivo será elegir C_h, h=1,...,I-1 de manera que

$$\sqrt{12}\sum_h W_h S_h = \sum_h f_h^C \left(C_h - C_{h-1}\right)^2 = \sum_h \left(Z_h - Z_{h-1}\right)^2$$

sea mínima; donde por definición $Z(x) = \int_{x_0}^{x} \sqrt{f(x)}\,dx$ y $Z_h = Z(C_h)$. Siendo los extremos $Z_0=0$ y $Z_I = \int_{x_0}^{x_n} \sqrt{f(x)}\,dx = \sum_{i=1}^{m_I} \int_{A_{i-1}}^{A_i} \sqrt{f(x)}\,dx = a \cdot \sum_{i=1}^{m_I} \sqrt{f_i^A}$ constantes, es fácil verificar que el mínimo anterior se alcanza cuando

$$Z_h - Z_{h-1} = \int_{C_{h-1}}^{C_h} \sqrt{f(x)}dx = a \cdot \sum_{i=m_{h-1}+1}^{m_h} \sqrt{f_i^A}$$

es constante para cada h=1,...,I, o lo que es lo mismo de manera que $\sum_{i=m_{h-1}+1}^{m_h} \sqrt{f_i^A}$ sea constante. Como las f_i^A son las densidades de frecuencias relativas en intervalos de amplitud constante, lo anterior es equivalente a hacer $\sum_{i=m_{h-1}+1}^{m_h} \sqrt{n_i}$ constante para todo h=1,...,I, donde n_i es la frecuencia absoluta en el intervalo $[A_{i-1},A_i]$. Así la regla para construir los intervalos que darán lugar al histograma consiste en formar la cumulativa de las raíces de las frecuencias absolutas n_i, columna $\mathrm{CUM}\sqrt{n}$ en la siguiente tabla:

A_{i-1}-A_i	$CUM\sqrt{n}$
A_0-A_1	$\sqrt{n_1}$
A_1-A_2	$\sqrt{n_1}+\sqrt{n_2}$
...	...
A_{i-1}-A_i	$\sqrt{n_1}+\sqrt{n_2}+...+\sqrt{n_i}$
...	
$A_{m_{I-1}}-A_{m_I}$	$\sqrt{n_1}+\sqrt{n_2}+...+\sqrt{n_i}+...+\sqrt{n_{m_I}}=T$

y elegir C_h, h=1,...,I-1 de manera que estos formen intervalos iguales (lo más iguales posibles en realidad) en la escala $CUM\sqrt{n}$, para ello es suficiente calcular $\frac{T}{I}$ y elegir C_h como aquel de los A_i cuya $CUM\sqrt{n}$ correspondiente esté más cercana a $h\cdot\frac{T}{I}$.

Para llevar a cabo la tarea anterior hemos elaborado dos programas en el entorno MATHEMATICA cuyas líneas de código se presentan en el Anexo 1. Para usar dichos programas, el usuario debe introducir los valores observados $x_0,...,x_n$ y el número de intervalos m_I en la partición auxiliar (por defecto se toman 100), seguidamente el primero de los programas procede a hacer la partición auxiliar $[A_{i-1},A_i]$. A continuación, el usuario introduce la salida del programa anterior y el número de intervalos I deseados en el histograma y el programa CUM obtiene los intervalos $[C_{h-1},C_h]$ construidos con la regla anterior junto con sus correspondientes frecuencias (absolutas y $CUM\sqrt{n}$ éstas últimas serán aproximadamente constantes-).

Ejemplo: Consideramos una muestra aleatoria simple de tamaño 5000, de una variable normal centrada de desviación típica 2, generada por simulación. Las instrucciones necesarias para ello son las siguientes:

```
nd=5000;
datos=Table[Random[NormalDistribution[0,2]]//N,{nd}];
```

A continuación se genera con el programa **ordenar** una partición auxiliar formada por 100 intervalos de clases. Las salida de este programa junto con el número de intervalos por nosotros deseados (en este caso 9) se dan como entrada en el programa CUM:

```
ordenacion=ordenar[datos];
nc=9;
clases=CUM[ordenacion,nc]
```

proporcionando éste la siguiente salida:

	Extr . inferior	Extr . superior	Frec . absoluta	Frec . Abso . Cum
1	-7.5984	-3.6091	193	57.2561
2	-3.6091	-2.1316	530	71.5959
3	-2.1316	-1.2451	612	60.5626
4	-1.2451	-0.3586	787	68.5109
5	-0.3586	0.3802	727	60.1560
6	0.3802	1.1189	706	59.2992
7	1.1189	2.1532	742	71.9177
8	2.1532	3.3352	484	61.6977
9	3.3352	7.1767	219	60.8536

Cabe destacar en esta salida como los intervalos más pequeños se encuentran en la parte central de la distribución, que es donde hay una mayor densidad de frecuencias. Esto responde al hecho deseable de que al deducir la regla CUM las desviaciones típicas de cada intervalo aparecen multiplicadas por las frecuencias de los mismos.

2. UNA FORMA DE SUAVIZAR HISTOGRAMAS: SPLINES PARABÓLICOS

Partimos ahora de un histograma de una variable X, construido por el método anterior o algún otro. Es común señalar la conveniencia de suavizar el histograma de manera que sea más parecido a la posible distribución teórica de la variable. En este punto es donde este trabajo proporciona un planteamiento teórico de como puede conseguirse una suavización donde se tengan en cuenta algunas propiedades fundamentales de los histogramas.

Supongamos dado un histograma que deseamos suavizar. Esto significa que tenemos los puntos $C_0 < C_1 < ..... < C_{I-1} < C_I$, y los números $h_1 , h_2 , ... , h_n$, los cuales suelen ser generalmente no negativos, con h_i el peso (frecuencia absoluta, frecuencia relativa, densidad de frecuencia,..) sobre el intervalo $[C_{i-1},C_i]$ para i=1,...,I. Como la regla CUM para construir histogramas conduce generalmente a intervalos de amplitudes diferentes, en ese caso trabajaremos con las densidades de frecuencia.

El histograma se obtiene a partir de diferentes rectángulos, los cuales tienen de área $h_i \cdot \Delta C_i = h_i \cdot (C_i - C_{i-1})$. Una interpretación usual es que dichas áreas coinciden aproximadamente con la integral de la distribución desconocida *f* sobre $[C_{i-1} , C_i]$, con i=1,2,...,I. Por ello, éstas serán condiciones que impondremos a nuestra función *g* suavizadora del histograma

$$\int_{C_{i-1}}^{C_i} g(x)\, dx = \int_{C_{i-1}}^{C_i} p_i(x)\, dx = h_i \bullet \Delta C_i \quad ; i = 1,2,...,I. \qquad \textbf{(1), I ecuaciones.}$$

La función que vamos a elegir *g* va a ser una función a trozos de tal forma que en cada intervalo $[C_{i-1} , C_i]$ coincida con un polinomio de 2º grado $p_i(x)$, e imponemos la condición de ser derivable con continuidad en los puntos de enlace[3] C_i, para obtener una curva suave (nótese que no sería posible conseguir la condición de diferenciabilidad construyendo un polígono de frecuencias, es decir tomando *g* como una función poligonal).

Especificamos más la función suavizadora: *g* es un polinomio de 2º grado en $[C_{i-1} , C_i] \Rightarrow g(x) = p_i(x) = a_i + b_i x + c_i x^2 \Rightarrow g(x) = p_i(x) = a_{1,i} + a_{2,i} (x - C_{i-1}) + a_{3,i} (x - C_{i-1})^2/2$ en $[C_{i-1} , C_i]$. Escribimos el polinomio de 2º grado como un desarrollo en serie en el punto C_{i-1}. (se puede obtener esta expresión sin más que aplicar un desarrollo en serie de orden dos en el punto C_{i-1}).

Tenemos I polinomios ($p_1(x) , p_2(x) , ... , p_I(x)$) cada uno de ellos con 3 coeficientes por determinar ($a_{1,i} , a_{2,i} , a_{3,i}$), por tanto para poder hallar unívocamente todos los polinomios deseados debemos de plantear un sistema con al menos 3I incógnitas. En este momento tenemos ya I ecuaciones planteadas en (1), y vamos a ver las que se deducen de las condiciones de continuidad y derivabilidad impuestas a g.

Continuidad de *g* en $C_i \rightarrow \quad p_i(C_i) = p_{i+1}(C_i)$; i=1,2,...., I-1. **(I-1) ecuaciones**

[3] f es una pp-función de orden 3 con continuidad en la 1ª derivada; es decir $f \in P_{3,\varepsilon} \cap C^1$.

Derivada continua de g en $C_i \rightarrow p_i'(C_i) = p_{i+1}'(C_i)$; i=1, 2, ... , I-1. **(I-1) ecuaciones**

En total tenemos **I+(I-1)+(I-1) = 3I-2 ,** aún necesitamos al menos dos ecuaciones más. Para obtenerlas imponemos condiciones en los extremos de nuestra función suavizadora

$g(C_0)=p_1(C_0)$=a; $g(C_I)=p_I(C_I)$=b, **2 ecuaciones.**

En distribuciones en las que las densidades de frecuencia "parecen" decrecer a cero conforme nos acercamos a los intervalos extremos (como sucede en el ejemplo de la Sección anterior y en la variable analizada en la sección siguiente) son coherentes condiciones del tipo

$g(C_0)=p_1(C_0)=0$; $g(C_I)=p_I(C_I)=0$.

Sin embargo en otras situaciones (como ocurriría al generar datos a partir de una distribución exponencial donde a medida que nos acercamos al intervalo inferior las frecuencias aumentan) no son en absoluto convenientes ni coherentes.

Llegados a este punto tenemos un sistema de 3I ecuaciones con 3I incógnitas que queda planteado de la siguiente forma:

a_{11}	$= a$ (valor estimado de g en C_0)
$a_{11} + a_{21}\,\Delta C_1/2 + a_{31}\,\Delta C_1^2/6$	$= h_1$ (área en $[C_0,C_1]$)
$a_{11} + a_{21}\,\Delta C_1 + a_{31}\,\Delta C_1^2/2 - a_{12}$	$= 0$ (Continuidad de *g* en C_1)
$a_{21} + a_{31}\,\Delta C_1 - a_{22}$	$= 0$ (Continuidad de *g'* en C_1)
$a_{12} + a_{22}\,\Delta C_2/2 + a_{32}\,\Delta C_2^2/6$	$= h_2$ (Área en $[C_1,C_2]$)
.....	
.....	

Que resulta ser un sistema lineal compatible determinado (ver Carl De Boor, 1976), el cual a pesar de tener muchas ecuaciones homogéneas resulta muy complicado de resolver sin la ayuda de un programa informático. Aquí nosotros trabajamos con el paquete MATHEMATICA**,** implementando un nuevo programa (cuyas líneas de código aparecen en el Anexo 2) el cual nos resuelve el sistema de ecuaciones a partir de una distribución y además nos representa la función suavizadora *g* sobre el histograma.

Podríamos pensar en buscar otro tipo de suavizamiento, por ejemplo con polinomios de grado 3, 4,.... En estos casos las mejoras no son significativas respecto a los polinomios de segundo grado, y sin embargo el tiempo de ejecución del módulo es mucho mayor ya que el número de ecuaciones aumenta mucho[4].

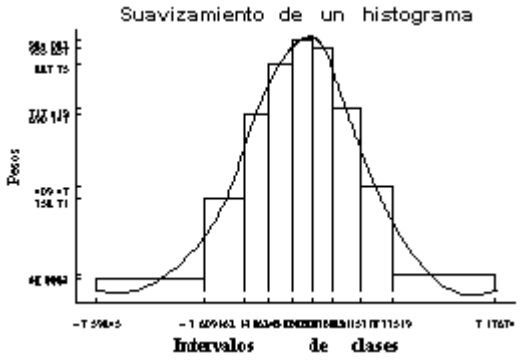
Ejemplo: Si consideramos la distribución de frecuencias construida en la Sección anterior y planteamos la suavización por splines parabólicos, se tiene un sistema lineal formado por 27 (3*9) ecuaciones con 27 incógnitas. En este ejemplo concreto sería coherente asignar el valor cero en los extremos de la función suavizadora. La implementación es la siguiente:

En primer lugar, modificamos ligeramente las salidas del programa CUM:

```
a=Transpose[clases[[1]]];
intervalos=AppendTo[a[[1]],a[[2,nc]]]
alturas=a[[3]]/(Drop[intervalos,1]-Drop[intervalos,-1])
```

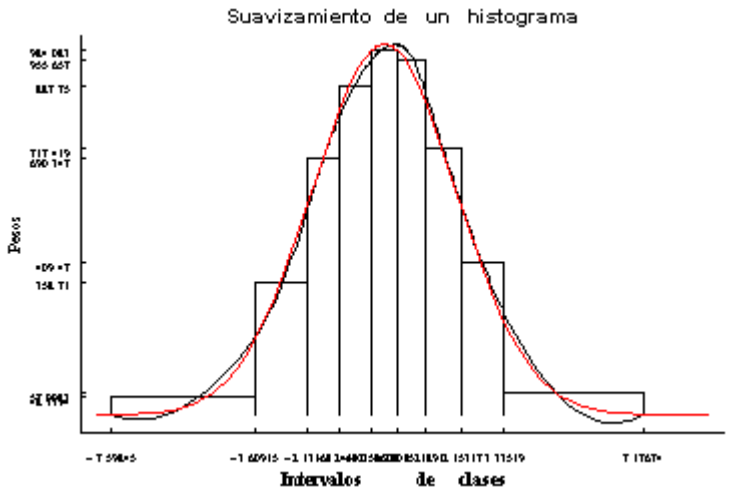
Conseguimos, de esta forma, las entradas del programa SUAHI01. Ejecutando este programa obtenemos lo siguiente:

```
Suahi01[intervalos,alturas,0,0]
```



[4] De grado 2→3I ecuaciones, de grado 3→ 4I ecuaciones, de grado 4→ 5I ecuaciones,....

Como en este caso los datos han sido generados a partir de una distribución conocida (normal centrada de desviación típica 2) podemos presentar la función de densidad teórica y la suavizada a partir del histograma, obteniendo el siguiente resultado:



3. APLICACIÓN A LOS INGRESOS DE LOS HOGARES ANDALUCES.

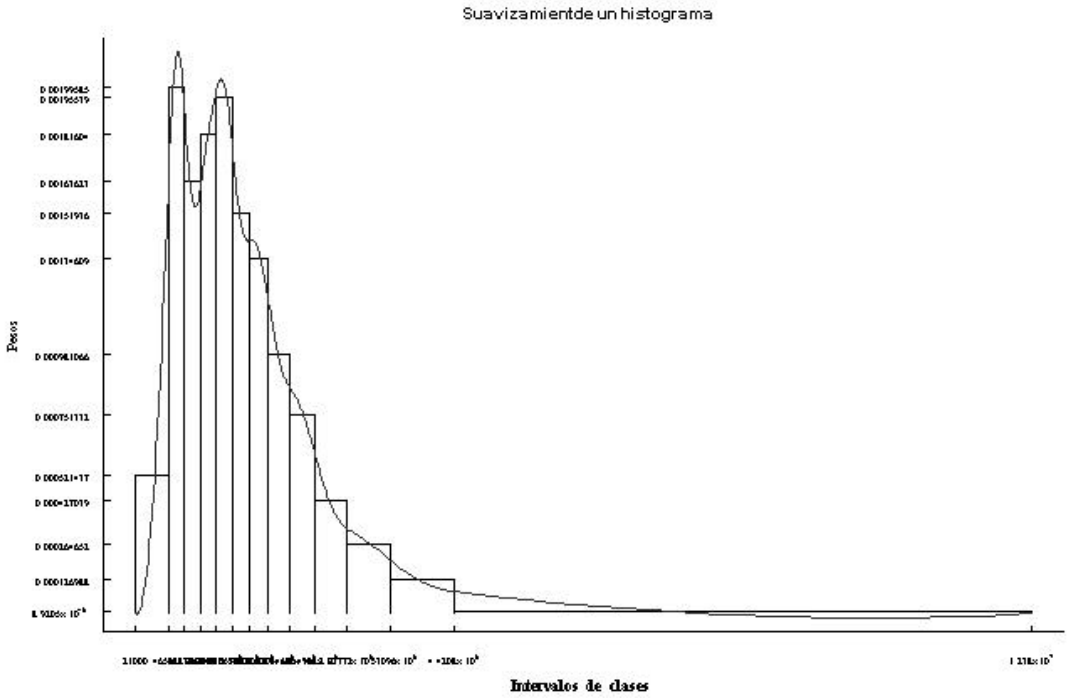
Para la implementación práctica de la metodología a la variable ingresos de los hogares andaluces hemos utilizado la Encuesta de Presupuestos Familiares 1990-91, donde el hogar es la unidad de análisis, por lo que el ingreso monetario del hogar se obtendrá sumando el ingreso de todos los miembros perceptores de dicho hogar. Para la interpretación de los datos se ha utilizado la metodología propuesta por la comunicación "La pobreza y las prestaciones sociales en Andalucía" (Domínguez A., González R. y Martín D.- 1999 ASEPELT), donde se define el ingreso monetario del hogar como la suma de los "ingresos procedentes del trabajo", "del capital", las "prestaciones sociales de carácter monetario" y los "ingresos monetarios por transferencias regulares". Se ha considerado oportuno, no incluir los "ingresos de carácter extraordinarios", puesto que dado su carácter ocasional podría distorsionar el comportamiento de la variable. Es conveniente además considerar Escalas de Equivalencias que tengan en cuenta el tamaño y composición del hogar, con el fin de que los ingresos de los hogares sean comparables, se toma aquella escala que solo tiene en cuenta el tamaño del hogar, dando lugar al ingreso per cápita.

Se han considerado los 3665 hogares andaluces muestrales obtenidos al eliminar previamente 9 hogares a los cuales no se les había imputado ningún tipo de ingresos monetario, al considerar que ello podría ser debido a un error de imputación de datos.

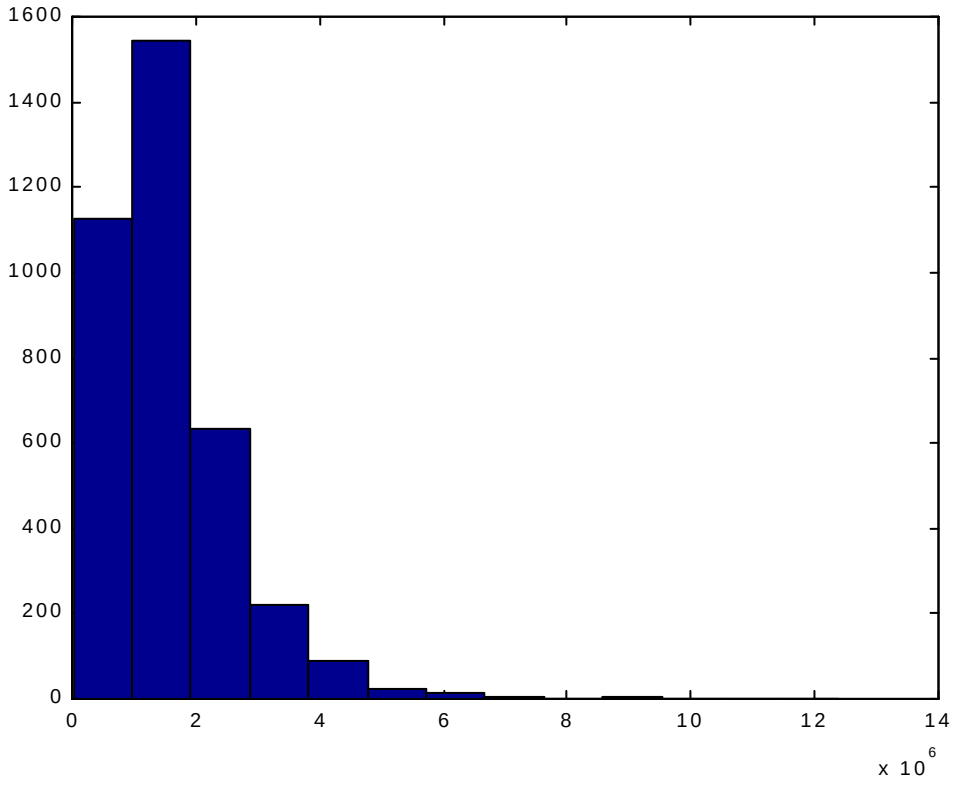
A continuación se muestran los resultados obtenidos aplicando la metodología descrita en las secciones anteriores a la variable en cuestión, tomando como número de intervalos, 13, el número más próximo a la fórmula propuesta por Sturges, $\log_2 n+1$, donde n es el tamaño de la muestra; y la cual esta basada en una distribución normal (Scott, 1992, página 48).

	Extr. inferior	Extr. superior	Frec. absoluta	Frec. Abso. Cum
1	21000	465924.	232	56.4726
2	465924.	688386.	444	63.1249
3	688386.	910848.	364	56.9299
4	910848.	1.13331×10^6	404	59.9964
5	1.13331×10^6	1.35577×10^6	435	61.8907
6	1.35577×10^6	1.57823×10^6	338	54.9272
7	1.57823×10^6	1.85013×10^6	366	63.1631
8	1.85013×10^6	2.14675×10^6	291	58.5681
9	2.14675×10^6	2.4928×10^6	260	59.9932
10	2.4928×10^6	2.93772×10^6	190	57.3082
11	2.93772×10^6	3.53096×10^6	157	60.1134
12	3.53096×10^6	4.4208×10^6	113	59.3232
13	4.4208×10^6	1.238×10^7	71	58.8024

A la representación gráfica de esta distribución de frecuencias (histograma) le realizamos una suavización como se plantea en las secciones iniciales, obtenido el siguiente gráfico:

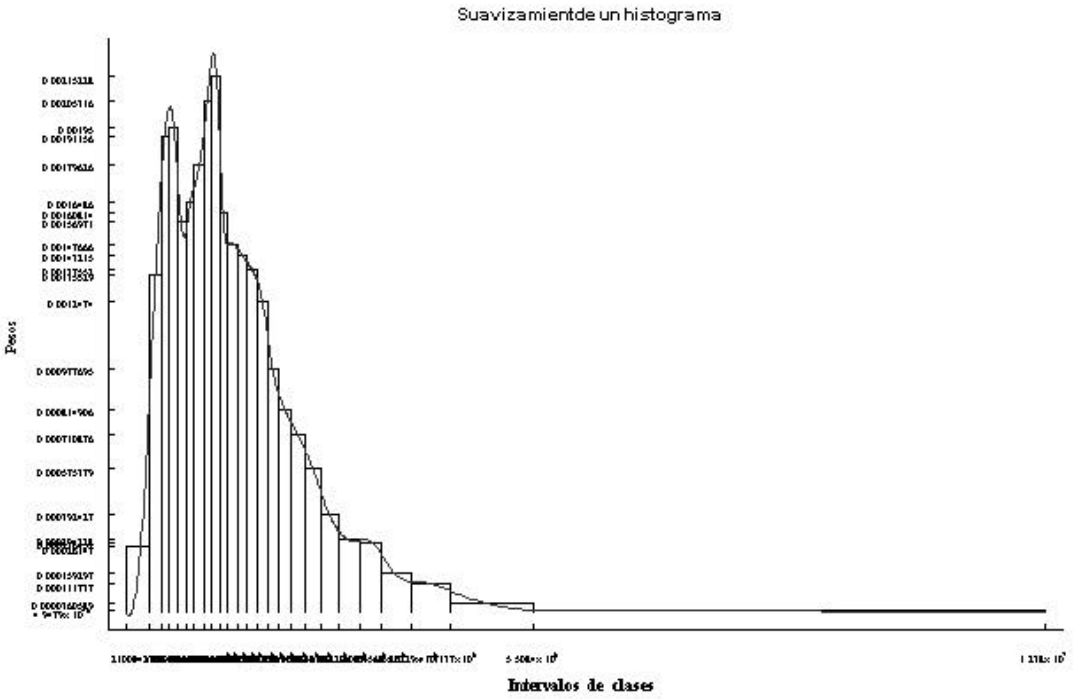


Usando el mismo número de intervalos, mostramos el histograma que se obtendría por defecto utilizando cualquier software estadístico estándar, el cual esta construido a partir de intervalos de la misma amplitud. (Se ha utilizado el programa Matlab versión 5.3)



Concluimos que la metodología propuesta es manifiestamente mejor que la obtenida por defecto con el software estándar a la hora de reflejar el comportamiento de la variable. En particular observamos que al haber una mayor densidad de frecuencia en la zona central de la distribución es conveniente elegir aquí una menor amplitud en los intervalos para recoger el comportamiento de la variable en esta zona, lo cual es recogido por el método propuesto y no por el estándar. En particular, el método propuesto recoge la bimodalidad de la variable y sugiere que esta podría comportarse como una mixtura de dos variables unimodales, comportamiento importante que no es en absoluto recogido en la metodología estándar.

Finalizamos esta sección mostrando como sería el correspondiente histograma cuando se toman 25 intervalos de clase y posteriormente se realiza su suavizamiento con la metodología propuesta en este trabajo:



ANEXO 1: IMPLEMENTACIÓN DE LA REGLA CUM

Este programa lo configuran dos funciones:

La función **ordenar** que a partir de unos datos y un número de particiones nos proporciona la entrada para la siguiente función.

Los parámetros que hay que suministrar al programa **ordenar** son:

datos:= Corresponde a los datos iniciales.

Particion->m_I = Donde m_I es el numero de intervalos de la partición auxiliar (si no se indica se toma por defecto 100).

La función CUM la cual a partir de una muestra y un número prefijado de intervalos I genera los intervalos de clase siguiendo el algoritmo CUM.

Los parámetros que hay que suministrar al programa CUM son:

datos:= Corresponde a una lista que contiene los extremos de los intervalos de clases, las frecuencias absolutas de cada intervalo y su correspondiente frecuencia absoluta en la escala CUM.

I= Número de intervalos.

Las ordenes que configuran el programa son:

```
Off[General::spell1]

ordenar::usage =
"ordenar[datos,opciones] ordenar los datos en intervalos. En
opciones indicando particion->100, divide los datos en 100
intervalos (opción por defecto)";

CUM::usage =
"Cum[datos,I] construye “I” intervalos de clases a partir de los
datos siguiendo el algoritmo CUM "

Options[ordenar] = {particion->100};
ordenar[datos_List, opciones___Rule]:=
     Module[{amplitud,intervalo,Contar,k,datosff},
          partic = particion /. {opciones} /. Options[ordenar];
```

```
          amplitud=(Max[datos]-Min[datos])/partic//N;
          intervalo=Table[{Min[datos]+k amplitud,
            Min[datos]+(k+1) amplitud},{k,0,partic-1}];
          Contar=BinCounts[datos,{Min[datos],Max[datos],amplitud
     }];
          Contar[[1]]=Contar[[1]]+Count[datos,Min[datos]];
          datosff=Transpose[Append[Append[Transpose[intervalo],
            CumulativeSums[Contar]],
          CumulativeSums[Sqrt[Contar]//N]]]
          ]

CUM[datos_List,estratos_Integer]:=
     Module[{longitud,a1,a2,k},
          longitud=datos[[Length[datos],4]]/estratos;
          Array[final,estratos];
          a2=Length[datos]-Length[Select[Column[datos,4],
             (#> longitud)&]];
          If[a2==0,a2=1];

     final[1]={Column[datos,1][[1]],Column[datos,2][[a2]],
              Column[datos,3][[a2]],Column[datos,4][[a2]]};
            a1=a2;
          Do[
            a2=Length[datos]-Length[Select[Column[datos,4],
            (#>k longitud)&]];
            If[a2<=a1,a2=a1+1];
            If[Column[datos,4][[a2+1]]-k longitud >
              k longitud - Column[datos,4][[a2]] ,

     final[k]={Column[datos,2][[a1]],Column[datos,2][[a2]],
              Column[datos,3][[a2]],Column[datos,4][[a2]]};a1=a2,
            final[k]={Column[datos,2][[a1]],
                  Column[datos,2][[a2+1]],
```

```
Column[datos,3][[a2+1]],Column[datos,4][[a2+1]]}
            ;a1=a2+1],{k,2,estratos-1}];

     final[estratos]={Column[datos,2][[a1]],
          Column[datos,2][[Length[datos]]],
          Column[datos,3][[Length[datos]]],
          Column[datos,4][[Length[datos]]]};
     final[0]={0,0,0,0};
     TableForm[Table[{final[k][[1]],final[k][[2]],
          final[k][[3]]-final[k-1][[3]],
          final[k][[4]]-final[k-1][[4]]},
       {k,1,estratos}], TableAlignments->Right,
       TableHeadings->{Automatic,{Extr. inferior,
          Extr. superior,
         Frec. absoluta, Frec. Abso. Cum}} ]
     ]
```

ANEXO 2: IMPLEMENTACIÓN DEL SUAVIZAMIENTO DE HISTOGRAMAS

A partir de un conjunto de valores referente a los extremos de los intervalos de clases y sus correspondientes pesos, construye una función polinómica de segundo grado a trozos en cada intervalo de clase que suaviza el histograma. La salida nos proporciona la representación gráfica del histograma y la función suavizadora.

Los parámetros que hay que suministrarle al programa son:

datos:= Corresponde a una lista que contiene los extremos de los intervalos de clases, ordenados de forma creciente.

peso:= Corresponde a una lista que contiene el peso de cada uno de los intervalos.

Las ordenes que configuran el programa son:

Suahist[datos_,pesos_] :=

```
Module[{n,m,p,listaec1,listaec2,listaec3,listaec4,listaf,dd
1,dd,ww,i,j,c},

m=Length[datos]; n=Length[pesos];

If[n!=m-1,
     Show[Graphics[{RGBColor[1.000,0.000,0.000],
          Text[FontForm["              ERROR \n El número
          de datos ha de ser\n uno más que el de pesos.",
          {"Arial",16}],{0,0}]}]];,

listaf=Flatten[Table[c[i,j],{i,3},{j,n}]];
Do[p[i_,x_]:=c[1,i]+(x-datos[[i]])*(c[2,i]+(x-datos[[i]])
     *c[3,i]/2),{i,1,n}];

(* Condiciones del área *)
```

```
listaec1=Table[Integrate[p[i,x],{x,datos[[i]],datos[[i+1]]}
]
     -pesos[[i]]*(datos[[i+1]]-datos[[i]])==0,{i,1,n}];

(* Condiciones de continuidad *)
listaec2=Table[p[i,datos[[i+1]]]-
     p[i+1,datos[[i+1]]]==0,{i,1,n-1}];

(* Condiciones de derivada continua *)
listaec3=Table[(D[p[i,x]-p[i+1,x],x]/.x->datos[[i+1]])==0,
     {i,1,n-1}];

listaec4=Flatten[{listaec1,listaec2,listaec3,
     p[1,datos[[1]]]==0,p[n,datos[[n+1]]]==0}];

listaec2=Solve[listaec4,listaf];

$DefaultFont={"Times",6};
dd1=Graphics[Table[Line[{{datos[[i]],0},{datos[[i]],pesos[[
i]
     ]},{datos[[i+1]],pesos[[i]]},{datos[[i+1]],0}}],{i,1,n
}]
     ];
dd=Table[Plot[Evaluate[p[i,x]/.listaec2,{x,datos[[i]],
     datos[[i+1]]},DisplayFunction->Identity,
     PlotStyle->GrayLevel[0.2]]],{i,1,n}];

ww=((datos[[n+1]]-datos[[1]])/(2*m));
Show[{dd1,dd},DisplayFunction->$DisplayFunction
     ,PlotRange->{{datos[[1]]-ww,datos[[n+1]]+ww},All}
     ,Frame->{True,True,False,False}
     ,FrameLabel->
     {FontForm["Intervalos de clases",{"Times-Bold",10}]
```

```
      ,FontForm["Pesos",{"Times-Bold",8}]}
      ,PlotLabel->FontForm["Suavizamiento de un histograma"
      ,{"Arrus-Bold",10}],FrameTicks->{datos,pesos}];
]];
```

BIBLIOGRAFÍA.

COCHRAN, W. G. (1976) *Técnicas de muestreo*. Compañía Editorial Continental, México. 6ª edición.

CASTILLO, E. (1993). *Introducción a la Estadística Aplicada con Mathematica*. Gráficas Calima.

DE BOOR, C. (1978). *A practical guide to splines.* Springer-Velag.

DOMINGUEZ A., GONZÁLEZ R., y MARTÍN D. (1999). *La Pobreza y las Prestaciones Sociales en la Comunidad Autónoma de Andalucía.* Comunicación ASEPELT 1999

SCOTT, D.W. (1992). *Multivariante Density Estimation. Theory, Practice, and Visualization.* New York. John Wiley and Sons.

WOLFRAM. S. (1992). *MATHEMATICA. A System for Doing Mathematics by Computer*. Addison-Wesley Publishing Company.