El histograma con la TI-92: optimización de clases

Juan José González Henríquez

Resumen

El histograma es la herramienta más popular y antigua a la hora de representar gráficamente un conjunto de datos. Un aspecto crucial en su construcción es el número de clases. En este trabajo presentamos una actividad con calculadora gráfica para que los alumnos entiendan la importancia que tiene el número de clases en la apariencia del histograma. De esta manera se entiende que hay un número óptimo de clases que permite captar la estructura de los datos. Esto cuestiona una regla muy extendida en muchos libros que recomienda elegir entre 5 y 20 clases. En vez de esta regla, podría enseñarse una regla que tenga buenas propiedades teóricas como, por ejemplo, la regla de Scott.

Abstract

The histogram is the oldest and most popular tool for graphical display of a univariate set of data. An important parameter that needs to be specified when constructing a histogram is the number of bins. In this paper we present a graphing calculator activity helping pupils understand that the choice of number of bins has an enormous effect on the appearance of the resulting histogram. Acceptance of this viewpoint implies that the optimal number of bins should be chosen so that histogram displays the essential structure of the data. So, the rule widely recommended that select a number between 5 and 20 bins should be questioned. Instead of this rule, Scott's rule, which is theorecly well-stablished, could be teached.

Introducción

Una técnica habitual en el análisis exploratorio de datos numéricos de tipo continuo es la agrupación de los datos en clases o categorías, llamados intervalos de clase. Una vez establecidas las clases, se construye la distribución de frecuencias y su representación gráfica: el histograma. Esto permite vislumbrar ciertas características de la población de la cual proceden los datos, como unimodalidad, simetría, normalidad, etc. Desde luego estas conclusiones serían imposibles de deducir a partir de la simple observación del conjunto de datos; en cierto modo, los árboles no dejan ver el bosque.

Un aspecto fundamental en la enseñanza del histograma es el número k de clases que se escoge para agrupar los datos. La apariencia del histograma depende de este parámetro; si se toman muchas clases adopta una forma

muy recortada, con muchos picos, mientras que con pocas clases adquiere un aspecto muy suave, con pocos rectángulos. Una regla muy extendida en secundaria es tomar un número k de clases entre 5 y 20. Así, nos podemos encontrar que para un mismo conjunto de datos un alumno hace un histograma con 6 clases, mientras que su compañero de pupitre lo hace con 19 clases. De esta manera el alumno puede percibir cierta arbitrariedad en la elección de este parámetro. En este caso, la calculadora gráfica (por ejemplo, la TI-92) es una herramienta excepcional para hacer ver a los alumnos que, en realidad, la elección del número de clases no es arbitraria sino que obedece a cierto criterio de optimalidad. Para ello, vamos a generar 200 valores de una variable aleatoria X con distribución normal de media 4 y desviación típica 2. Desde luego el número de datos y los parámetros de la distribución normal no son relevantes por lo que puede escogerse otros números cualesquiera. A partir de estos datos, compararemos la verdadera curva de frecuencias (la campana de Gauss) con los distintos histogramas de frecuencias que surgen al variar el número k de clases.

Número óptimo de clases

El primer paso de esta actividad es generar el conjunto de datos. Si queremos que todos los alumnos tengan la misma muestra (así todos ven lo mismo) lo que haremos es introducir en la línea de entrada de la pantalla inicial de la calculadora la orden `RandSeed 1` (el número 1 es arbitrario, podemos poner cualquier número entero positivo). Ahora, generamos una muestra de tamaño 200 según lo establecido anteriormente y la guardamos en la variable d. Para ello, tecleamos la orden `seq(randNorm(4,2),n,1,200)→d`. Además, para tener una muestra más verosímil desde un punto de vista práctico redondearemos los datos a dos cifras decimales. En la figura 1 mostramos el área de historia de nuestra calculadora donde pueden observarse todas las órdenes introducidas hasta ahora.

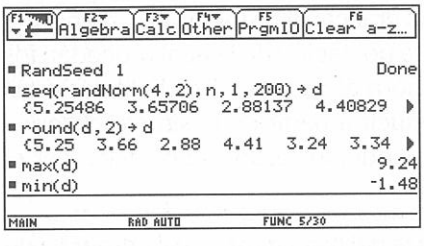

Figura 1

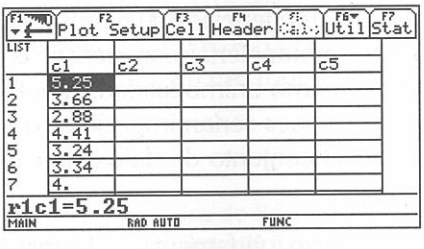

Figura 2

A continuación vamos a visualizar, por ejemplo, el histograma con 5 clases. Para ello, debemos pulsar consecutivamente las teclas [APPS] [6] [2] y, en el menú open, elegimos dentro de la carpeta principal la variable de tipo lista `d (Type List)`. Al final obtenemos la pantalla que aparece en la figura 2, donde puede observarse los datos en la primera columna. A partir de aquí pulsamos [F2] [F1] y elegimos la opción histograma; después, nos situamos en el recuadro de la `x` y ponemos `d`; finalmente, en el recuadro `Hist. Bucket Width` ponemos (9.24-(-1.48))/5 y pulsamos [ENTER] dos veces. Obviamente, el cociente anterior nos da la amplitud de los intervalos de clases para el caso que estamos considerando, a saber, 5 clases. Ahora pulsando [♦] [GRAPH] veremos el histograma (figura 3). Antes de hacerlo, debemos pulsar [♦] [WINDOW] y poner como parámetros de la ventana de visualización aquellos que aparecen en la figura 4. Además, para que no aparezcan los ejes de coordenadas en la figura 3, debemos pulsar [♦] [F] y cambiar `Axes a OFF`.

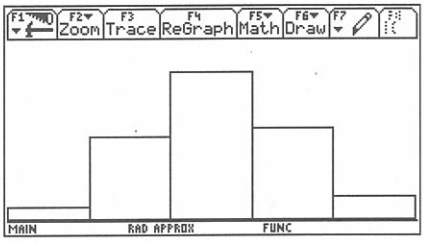

Figura 3

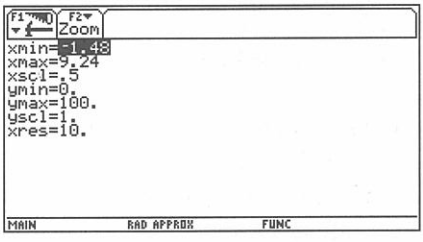

Figura 4

Para decidir si el histograma de la figura 3 estima correctamente a la verdadera curva de frecuencias debemos superponerle dicha curva. Para definir nuestra curva, nos vamos al editor de funciones ([♦] [Y=]) e introducimos la curva en el primer lugar de la lista (figura 5). Después pulsando [♦] [GRAPH] veremos la superposición de la verdadera curva de frecuencias con el histograma anterior (figura 6). Desde luego, el histograma no estima muy bien (sobreestima) a la verdadera curva de frecuencias. Es necesario probar con otros valores de k (el número de clases) para buscar el «óptimo».

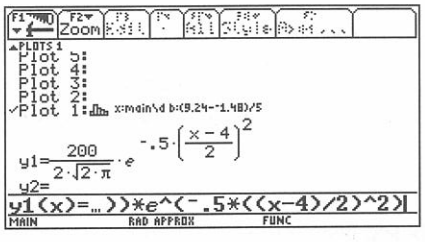

Figura 5

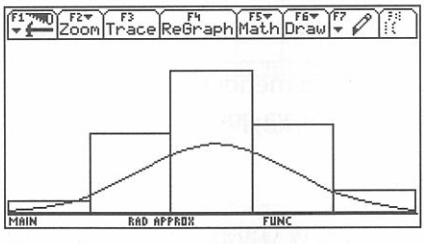

Figura 6

Para probar con 10 clases, por ejemplo, nos vamos a la gráfica de la figura 6 en nuestra calculadora, pulsamos [♦] [Y=] y nos situamos con el cursor en «plot 1». Posteriormente, pulsamos [F3], cambiamos el 5 que aparece en el recuadro Hist. Bucket Width por un 10, pulsamos [ENTER] dos veces y finalmente [♦] [GRAPH] (figura 7). Para esta última figura es conveniente cambiar el parámetro ymax de la ventana de visualización a 50 (para ello, pulsar [♦] [WINDOW] y cambiar ymax).

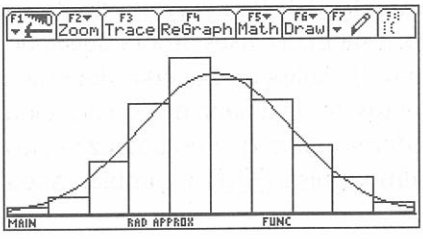

Figura 7

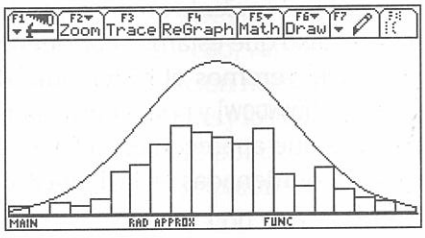

Figura 8

Desde luego la figura 7 «habla por sí misma» y no cabe la menor duda de que es mejor coger 10 clases que 5 clases. A partir de aquí podemos probar con el número de clases *k* que queramos; basta repetir los pasos dados anteriormente y cambiar 10 por el *k* que queramos. Si probamos con 20 clases obtenemos el histograma de la figura 8, el cual subestima a la verdadera curva. Después de probar con diferentes valores de *k* se llega a la conclusión de que para este conjunto de datos el número óptimo de clases es 10. Una pregunta que podemos hacernos ahora es: ¿existe alguna relación entre el número de datos y el número óptimo de clases? En el caso que estamos considerando (distribución normal) ¿dependerá, a su vez, de la dispersión de los datos? Investigaciones, relativamente recientes, han puesto de manifiesto que para poblaciones normales el número óptimo de clases que minimiza asintóticamente el error cuadrático medio integrado depende del número de datos y de la dispersión de los datos[1]. En la próxima sección presentaremos dicha regla (conocida como regla de Scott).

La elección del número de clases en la práctica

No cabe la menor duda de que la situación del párrafo anterior no se da en la práctica, ya que es extremadamente raro conocer de antemano la distri-

[1] Scott, D. W. (1979). On Optimal and data-based histograms. *Biometrika*, 66, 3, 605-610

bución de la cual proceden los datos. No obstante, queda claro que existe un número óptimo de clases para realizar una distribución de frecuencias, por lo que no debiera elegirse de manera arbitraria un número de clases en un rango de números naturales, como 5-20. Es preferible dar reglas precisas para la elección del número de clases que tengan algún sustento teórico. Al respecto han surgido una serie de reglas prácticas; las más citadas son:

i) $k = \sqrt{n}$, siendo n el número de datos,

ii) $k = 1 + \log_2 n$, (regla de Sturges[2]) y

iii) Tomar el entero superior más próximo a $\frac{\max - \min}{3.5 \cdot s} \cdot \sqrt[3]{n}$ (regla de Scott),

siendo, *max*, *min*, y s, el máximo, el mínimo y la desviación típica muestral, respectivamente. De las reglas anteriores, la primera de ellas aparece recomendada en algunos libros de estadística, pero carece de la base teórica que tienen las otras dos. Sin duda, de las tres, la que mejor consideración tiene dentro de la literatura es la regla de Scott. A pesar de ser una regla basada en normalidad, es muy robusta, como lo demuestra el autor en su artículo, frente distribuciones no Gaussianas. La regla de Sturges es muy común en paquetes estadísticos como R, S-plus y SPSS ya que para valores de n (tamaño de la muestra) moderados (n menores que 200) da resultados parecidos a la regla de Scout. Sin embargo para valores grandes de n esta regla no funciona bien ya que sobreestima a la verdadera función de densidad. A continuación ejemplificamos el cálculo de las reglas anteriores para el conjunto de datos que venimos manejando en nuestra calculadora. En la figura 9 mostramos el área de historia de nuestra calculadora donde puede verse los cálculos realizados para las tres reglas anteriores.

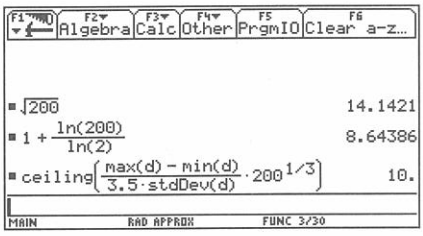

Figura 9

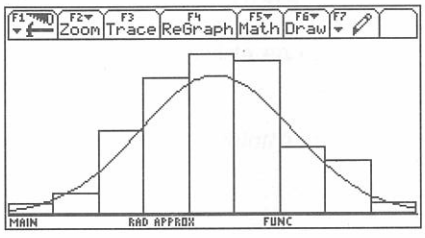

Figura 10

[2] Sturges, H. A. (1926). The choice of a class interval. *Journal of American Satatistical Association* , 21, 65-66

Al observar la figura 9 puede concluirse que regla de Scott y nuestro criterio visual coinciden en número de clases. Desde luego esto da cierta credibilidad a dicha regla. No obstante, la regla de Sturges sugiere un histograma con 9 clases, lo cual no está nada mal, como puede observarse en la figura 10. En esta figura puede apreciarse como, coincidiendo con otros estudios realizados al respecto, la regla de Sturges sobreestima a la verdadera función de densidad.

Juan José González Henríquez. Universidad de las Palmas de Gran Canaria
Correo electrónico: jjglez@dma.ulpgc.es