VII Coloquio Regional de Estadística
XII Seminario de Estadística Aplicada IASI
III Escuela de Verano CEAES
Universidad Nacional de Colombia, Sede Medellín
Medellín, 20-23 de Julio de 2010
Histograma Basado en Deciles
Convergencia Asintótica

Juan Carlos Correa M., PhD.
E-Mail:jccorreamorales@unal.edu.co
Francisco J. Castrillón M. M.Sc.
E-Mail:fjcastri@unal.edu.co

June 15, 2010

Abstract

In this paper we present some modifications to the current histogram in which the length of classes depends on Sturges formula, although Scott's and Freedman-Diaconis',are also utilized. The propossed histogram is drawn from the deciles of the dataset used. It has analytic, intuitive, and visual advantages over the formers. We illustrate this histogram using the Medellín 2009-Male-Half-Marathon data.

Key words: Histogram, Quantiles, Asymptotic Convergence

Resumen

Presentamos algunas modificaciones al histograma común cuyo número de intervalos de clase se halla de acuerdo con la fórmula de Sturges, aunque también son muy utilizadas las de Scott y Freedman-Diaconis, las cuales también se muestran. El histograma propuesto está basado en los deciles de la muestra de datos con que se cuenta, éste tiene ventajas analíticas, intuitivas y visuales que lo hacen más potente para representar el conjunto de datos. Para hacer las comparaciones, se hace una ilustración utilizando los tiempos de la Media Maratón de Medellín 2009, medidos con un chip.

Palabras claves: Histograma, Cuantiles, Convergencia Asintótica

1. Introducción

El histograma es el gráfico más popular que utilizan no sólo los estadísticos sino también el público en general para visualizar un conjunto de datos. Un histograma es un gráfico de frencuencias tabuladas representadas en barras, muestra la proporción de datos que caen dentro de cada una de varias categorías, es una forma de acumulación en barras. Las clases usualmente son especificadas en forma de intervalos adyacentes que no se intersectan. Las categorías o barras generalmente son del mismo ancho, aunque no necesariamente.
Esta clase de gráfico es realmente una aproximación visual de la densidad de los datos, realmente es una estimación de la densidad de la población de la cual provienen los datos. El área total del histograma se normaliza en 1. Si la longitud de los intervalos sobre el eje X es igual a 1, el histograma es idéntico a una gráfica de frecuencia relativa. Una regla sencilla que se presenta a los estudiantes en un curso de estadística es que utilicen intervalos de ancho igual y que utilicen la regla de Sturges [12] para determinar el número de barras. Pero realmente no hay un "mejor"número de clases y muy probablemente diferente logitud de intervalos pueden revelar detalles ocultos de los datos.
Problemas como el número de barras, límites de clases y demás, han impulsado la búsqueda de alternativas más simples. Algunos teóricos han intentado determinar un número óptimo de clases, pero estos métodos generalmente hacen suposiciones muy fuertes acerca de la forma de la distribución. Normalmente se experimenta tanto con el número de clases como con el ancho de éstas para detectar algún comportamiento oculto en los datos.

El número de clases k puede calcularse directamente o de un ancho de clase h sugerido:

$$k = \left\lceil \frac{Max(data) - Min(data)}{h} \right\rceil$$

donde los corchetes indican la función *ceiling*; es decir, el menor entero mayor o igual que el argumento.

Fórmula de Scott:

$$h = \frac{3{,}5s}{n^{1/3}}$$

donde s es la desviación estándar muestral. Los histogramas de Excel están todos basados en $k = \sqrt{n}$.

Se utiliza la fórmula de Scott para distribuciones normales basadas en el estimador del error estándar.

Fórmula de Freedman-Diaconi:

$$h = 2\frac{IQR(X)}{n^{1/3}}$$

Basada en el rango intercuartil, IQR, a menos que éste sea igual a 0, en cuyo caso devuelve la desviación mediana absoluta.

Fórmula de Sturges:

$$k = \lceil log_2 n + 1 \rceil$$

La cual basa los tamaños de las clases en el rango de los datos, pero se comporta muy pobre si $n < 30$.

En cuanto a la estimación de la densidad, varios autores han dado estimadores importantes con base en desarrollos teóricos, entre ellos, Van Ryzin [16], Kim and Van Ryzin [6], Scott [9], Freedman and Diaconis [**?**]alias3), Kogure [7], Taylor [13]. Scott [9] mostró el *histograma promediado* como un estimador de densidad.

2. Histograma basado en deciles

2.1. Cuantiles muestrales [10]

Si definimos para una función de distribución univariada F, y para $0 < p < 1$, la cantidad

$$F^{-1}(p) = inf\{x : F(x) \geq p\}$$

el p-ésimo *cuantil* de F, denotado ξ_p, (es decir, $\xi_p = F^{-1}(p)$), entonces F y la función no decreciente continua por la izquierda $F^{-1}(t), 0 < t < 1$, (llamada la función *inversa* de F), cumplen las siguientes propiedades:

- i) $F^{-1}(F(x)) \leq x, -\infty < x < \infty$, y
- $(ii) F(F^{-1}(t)) \geq t, 0 < t < 1$. Luego
- $(iii) F(x) \geq t$ si y solamente si $x \geq F^{-1}(t)$

Nótese que ξ_p satisface

$$F(\xi_p -) \leq p \leq F(\xi_p)$$

Definicion 1. El p-ésimo *cuantil muestral* está definido como el p-ésimo cuantil de la función de distribución muestral F_n, es decir, como $F_n^{-1}(p)$. considerando que el p-ésimo cuantil muestral es un estimador de ξ_p, denotamos éste por $\hat{\xi}_{pn}$, o ξ_p cuando no hay lugar a ambiguedades.
La probabilidad

$$P\left(sup_{\{m \geq n\}} |\hat{\xi}_{pm} - \xi_p| > \epsilon\right)$$

converge a 0 a una tasa *exponencial*, más específicamente,

$$P\left(|\hat{\xi}_{pn} - \xi_p| > \epsilon\right) \leq 2e^{-2n\delta_\epsilon^2}$$

para todo n y $\delta_\epsilon = min\{F(\xi_p + \epsilon) - p, p - F(\xi_p - \epsilon)\}$.

Teorema 1. Sea $0 < p < 1$. Supóngase que F es continua en ξ_p.
(i) si existe la derivada por la izquierda de F en ξ_p, y es positiva, es decir, $F^{'}(\xi_p-) > 0$, entonces para $t < 0$,

$$\lim_{n\to\infty} P\left(\frac{n^{1/2}(\hat{\xi}_{pn} - \xi_p)}{[p(1-p)]^{1/2}/F'(\xi_p-)} \leq t\right) = \Phi(t)$$

(ii) Si existe la derivada de F por la derecha en ξ_p, y si ésta es positiva, es decir, $F^{'}(\xi_p+) > 0$, entonces para $t > 0$,

$$\lim_{n\to\infty} P\left(\frac{n^{1/2}(\hat{\xi}_{pn} - \xi_p)}{[p(1-p)]^{1/2}/F'(\xi_p+)} \leq t\right) = \Phi(t).$$

(iii) En todo caso,

$$\lim_{n\to\infty} P\left(n^{1/2}(\hat{\xi}_{pn} - \xi_p) \leq 0\right) = \Phi(0) = \frac{1}{2}$$

Corolario 1. Sea $0 < p < 1$. Si F es diferenciable en ξ_p y $F^{'}(\xi_p) > 0$, entonces,

$$\hat{\xi}_{pn} \approx AN\left(\xi_p, \frac{p(1-p)}{[F'(\xi_p)]^2 n}\right)$$

Corolario 2. Sea $0 < p < 1$. Si F tiene como densidad a f en una vecindad de ξ_p, y f es positiva y continua en ξ_p, entonces

$$\hat{\xi}_{pn} \approx AN\left(\xi_p, \frac{p(1-p)}{f^2(\xi_p) n}\right)$$

Los corolarios 1 y 2 muestran que $\hat{\xi}_p$ es asimptóticamente normal, es decir, que $\hat{\xi}_p$ converge en distribución a $N(0,1)$:

$$f(\xi_p)\sqrt{\frac{n}{p(1-p)}}\left(\hat{\xi}_{pn} - \xi_p\right) \approx AN(0,1)$$

En realidad, se cumple

$$sup_{-\infty<t<\infty}\left|P\left[f(\xi_p)\sqrt{\frac{n}{p(1-p)}}\left(\hat{\xi}_{pn} - \xi_p\right) \leq t\right] - \Phi(t)\right| = O(n^{-1/2})$$

cuando $n \to \infty$.

Note que si F tiene una densidad f, entonces, la distribución G_n de $\hat{\xi}_{pn}$ también tiene una densidad $g_n(t) = G_n^{'}(t)$ que cumple:

$$G_n(t) = P(\hat{\xi}_{pn} \leq t) = P(F_n(t) \geq p) = P(nF_n(t) \geq np)$$

$$= \sum_{i=m}^{n} \binom{n}{i} [F(t)]^i [1 - F(t)]^{n-i}$$

donde

$$m = \begin{cases} np & \textbf{si } np \textbf{ es un entero}, \\ [np] + 1 & \textbf{si } np \textbf{ no es un entero} \end{cases}$$

Así

$$g_n(t) = n\binom{n-1}{m-1} [F(t)]^{m-1} [1 - F(t)]^{n-m} f(t)$$

la cual es otra manera de ver la normalidad asintótica de $\hat{\xi}_{pn}$. La densidad de la variable $n^{1/2}(\hat{\xi}_{pn} - \xi_p)$ es

$$h_n(t) = n^{-1/2} g_n(\xi_p + tn^{-1/2})$$

que cumple:

$$\lim_{n\to\infty} h_n(t) = \phi(tf(\xi_p)) [p(1-p)]^{-1/2}, \textbf{para cada t}$$

esto es, $h_n(t)$ converge a la densidad $N(0, p(1-p)/f^2(\xi_p))$

2.2. Caso multivariado [10]

Bajo condiciones de suavidad de F en las vecindades de los puntos $\xi_{p_1}, \xi_{p_2}, \cdots, \xi_{p_k}$, el vector de cuantiles muestrales $(\hat{\xi}_{p_1}, \hat{\xi}_{p_2}, \cdots, \hat{\xi}_{p_k})$, es *asintóticamente normal***:**

Teorema 2. Sea $0 < p_1 < p_2 < \cdots < p_k < 1$. Supóngase que F tiene una densidad f en una vecindad de $\xi_{p_1}, \xi_{p_2}, \cdots, \xi_{p_k}$ y que f es positiva y continua en $\xi_{p_1}, \xi_{p_2}, \cdots, \xi_{p_k}$. Entonces $(\hat{\xi}_{p_1}, \hat{\xi}_{p_2}, \cdots, \hat{\xi}_{p_k})$ es asintóticamente normal con vector de medias $(\xi_{p_1}, \xi_{p_2}, \cdots, \xi_{p_k})$ y covarianzas σ_{ij}/n, donde

$$\sigma_{ij} = \frac{p_i(1-p_j)}{f(\xi_{pi}) f(\xi_{pj})}$$

para $i \leq j$, y $\sigma_{ij} = \sigma_{ji}$, para $i > j$.

Nótese que, aunque conocemos la convergencia asintótica de $(\hat{\xi}_{p_1}, \hat{\xi}_{p_2}, \cdots, \hat{\xi}_{p_k})$ (para $n \to \infty$) para k finita sin importar su magnitud, no se puede decir lo mismo acerca de la convergencia de $(\hat{\xi}_{p_1}, \hat{\xi}_{p_2}, \cdots, \hat{\xi}_{p_k})$ para $k \to \infty$, en cambio sí podemos afirmar que si $\{p_k\}_{k\in Z}$ es una sucesión no decreciente con $0 < p_k < 1$ para todo k y:

$$\lim_{k\to\infty} p_k = 1,$$

la distribución de $\hat{\xi}_{pk}$ es sintóticamente normal.

3. El histograma modificado

Freedman, Pizani y Purves [2], presentan un extenso capítulo relativo a histogramas con longitudes diferentes y sus interpretaciones. No dan recomendaciones acerca de la forma como escogen los límites de las clases.
Si tomamos los límites de clase de tal manera que éstos se correspondan con algunos percentiles predeterminados, producimos un histograma con intervalos de clase desiguales y las barras corresponden a un porcentaje específico de datos de la muestra. El histograma modificado es realizado a partir de los deciles calculados, $(\hat{\xi}_{0,1}, \hat{\xi}_{0,2}, \ldots, \hat{\xi}_{0,9})$, ya vimos $\hat{\xi}_p$ es el estimador de ξ_p, y $P(X \leq \xi_p) = p$. La gran ventaja de este nuevo histograma es que el número de clases es el mismo siempre y su interpretación es bastante fácil para el usuario.

Ya vimos que los cuantiles muestrales son fuertemente consistentes para la estimación de los cuantiles poblacionales [10]; esto quiere decir que a medida que la muestra aumenta de tamaño, obtenemos un histograma que será más cercano al histograma obtenido de los verdaderos deciles; el nombre de este histograma es *histograma decil* **o** *decilgrama***. Recuérdese también la multinormalidad asintótica de $(\hat{\xi}_{0,1}, \hat{\xi}_{0,2}, \ldots, \hat{\xi}_{0,9})$, bajo ciertas condiciones de suavidad.**

4. Ejemplo

Para ilustrar el uso y las bondades del histograma basado en deciles, hemos realizado los histogramas basados en las fórmulas de Scott, Freedman-Diaconis y Sturges tomando como datos los de la Media Maratón de Medellín 2009. Una de las características más importantes de los datos que arroja esta competencia es que los participantes forman grupos bien definidos. Esto produce distribuciones en su mayoría multimodales.

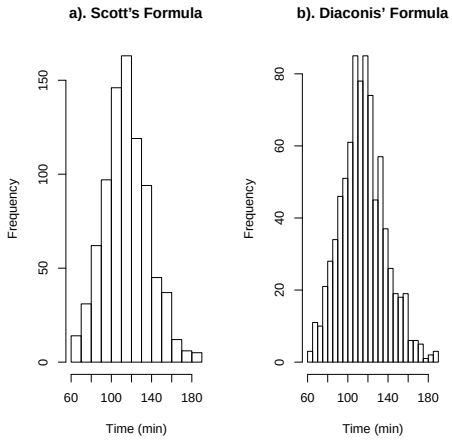

Fig. 1. Histograma de los datos de la Media Meratón de Medellín 2009, Fórmula de Scott y Fórmula FD

En la Figura 1 se observa la diferencia entre el número de clases. El histograma de Freedman-Diaconis tiene mayor número de clases; en ambos el número de datos debe ser relativamente grande y todavía más grande si se considera la convergencia multivariada de los estimadores conjuntamente tomados.

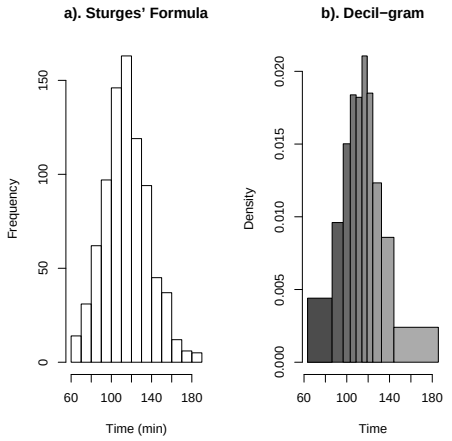

Fig. 2. Histograma de los datos de la Media Meratón de Medellín 2009, Fórmula de Sturges vs Deciles

En la figura 2 se pueden notar varias características. Por ejemplo, el histograma basado en la fórmula de Sturges es muy similar al del histograma basado en la fórmula de Scott. No obstante, en éste y en

los otros no se tiene conocimiento del número de datos que caen en cada una de las barras o clases; esta información sí se conoce en el histograma basado en deciles, pues, se sabe que el 10 % de los datos están en la barra correspondiente. Este histograma puede extenderse a un histograma percentil o incluso más, haciendo una partición más fina del intervalo $(0, 1)$. En todos estos casos, se sabe cuántos elementos de los datos caen en cada barra. Por supuesto, entre más divisiones se le hagan al intervalo $(0, 1)$, más datos serán necesarios.

5. Conclusiones

El histograma decil tiene ciertas ventajas sobre los histogramas más comúnmente utilizados; se destaca que el observador tiene algún control sobre los datos en la medida que sabe cuántos puntos están en cada barra; la longitud de las clases no necesariamente es la misma y conoce la distribución(asintótica) de los deciles, percentiles y fracciones de éstos; conoce, además, su distribución conjunta.

La convergencia asintótica puede asegurarse para un número k fijo de percentiles y su distribución conjunta; sin embargo, ésto mismo no puede afirmarse para vectores de fracciones de percentiles con dimensión infinita, incluso bajo el supuesto de que el tamaño muestral tiende a infinito a una tasa mayor que la longitud del vector.

Referencias

[1] **Benjamini, Y. (1988). Opening the Box of a Boxplot. *The American Statistician* , Vol. 42, No. 4, pp. 257-262**

[2] **Freedman, D., Pisani, R., and Purves, R. (1978). *Statistics*. W. W. Norton Company: New York**

[3] **Freedman, D. and Diaconis, P. (1981a) On the Maximum Deviation Between the Histogram and the Underlying Density. *Z. Wahrschein-lichkeitstheorie*, Vol. 58, pp. 139-167**

[4] **Freedman, D. and Diaconis, P. (1981b) On the Histogram as a Density Estimator: L2 Theory. *Z. Wahrscheinlichkeitstheorie*, Vol. 57, pp. 453-476**

[5] **Frigge, M., Hoaglin, D.C. y Iglewicz, B. (1989) Some Implementations of the Boxplot. *The American Statistician*, Vol. 43, No. 1, pp. 50-54**

[6] **Kim, B. K. and Van Ryzin, J. (1975) Uniform Consistency of a His- togram Density Estimator and Modal Estimation. *Communication in Statistics*, Vol. 4, No. 4, pp. 303-315**

[7] **Kogure, A. (1987) Asymptotically Optimal Cells for a Histogram.** ***The Annals of Statistics,*** **Vol. 15, No. 3, 1023-1030**

[8] **Scott, D. W. (1979) On Optimal and Data-Based Histograms.** ***Biometrika*****, Vol. 66, No. 3, pp. 605**

[9] **Scott, D. W. (1985) Averaged Shifted Histograms: Effective Nonpara- metric Density Estimators in Several Dimensions.** ***The Annals of Statis- tics,*** **Vol. 13, No. 3, 1024-1040**

[10] **Serfling, R. J. (1980).** ***Approximation Theorems of Mathematical Statis- tics.*** **John Wiley Sons: New York**

[11] **Scott, David W. (1979). On Optimal and Data-Based Histograms.** ***Biometrika*****, Vol. 66, No. 3, pp. 605-610**

[12] **Sturges, H.A. (1926). The Choise of a class Interval.** ***Journal of The American Statistical Association*****, pp 65-66**

[13] **Taylor, C. C. (1987) Akaike's Information Criterion and the Histogram.** ***Biometrika,*** **Vol. 74, No. 3, pp. 636-639**

[14] **Tufte, E. (1983) The Visual Display of Quantitative Information.** ***Graph- ics Press: Cheshire***

[15] **Tukey, J.W. (1977) Exploratory Data Analysis.** ***Addison-Wesley Pub- lishing Company: Reading, Massachusetts***

[16] **Van Ryzin, J. (1973) A Histogram Method of Density Estimation.** ***Communications in Statistics,*** **Vol. 2, No. 6, pp. 493-506**

[17] **Wainer, H. (1981). Graphical Data Analysis.** ***Ann. Rev. Psychol.*** **Vol. 32, pp. 191-204**