# MODIFYING THE HISTOGRAM USING DECILES

**Juan Carlos Correa M.[1], Francisco Javier Castrillón M.[2]**

[1] Juan Carlos Correa Morales. PhD Matemático y Estadístico Profesor Universidad Nacional de Colombia Sede Medellín. Dirección: Calle 59A No 63-20 E-mail: jccorreamorales@unal.edu.co
[2] Francisco Javier Castrillón Meneses. MSc Matemático y Estadístico Profesor Universidad Nacional de Colombia Sede Medellín Dirección: Calle 59A No 63-20 E-mail: fjcastri@unal.edu.co

## ABSTRACT

We present some modifications that produce a histogram based on deciles which is visually more informative than the equal-width histogram and the *quartile* boxplot currently used to represent a dataset. We also present the asymptotic convergence of the deciles and their joint asymptotic convergence to conclude that the class limits actually are point estimations and, consequently, can be estimated through confidence intervals. The researcher has certain control of the information, since he or she knows the part of the dataset pertaining to each class; besides, the larger the sample size is the larger the number of classes can ever he or she choose knowing the amount of data included into each class. We also discuss some problems of the boxplot , and illustrate both the histogram and the boxplot using the Medellín 2009 half-marathon data.

**Keywords:** Histogram, Boxplot, Quantiles, Asymptotic Convergence, Joint Asymptotic Convergence.

Recibido 31 de Marzo de 2010. Aceptado 16 de Junio de 2010
*Received: March 31, 2010 Accepted: June 16, 2010*

## MODIFICACIÓN DEL HISTOGRAMA UTILIZANDO DECILES

### *RESUMEN*

*Presentamos algunas modificaciones que producen un histograma basado en los deciles, el cual es visualmente más informativo que el histograma de igual longitud de clases y el boxplot de cuartiles, más comúnmente utilizados para representar un conjunto de datos. Se muestra también la convergencia asintótica de los deciles lo mismo que su convergencia conjunta para llegar a la conclusión de que los límites de clase de las barras son realmente estimaciones puntuales y que consecuentemente pueden estimarse por intervalos de confianza. El investigador adquiere cierto dominio de la información en el sentido de que conoce el porcentaje de datos que cae dentro de cada barra; además, en la medida que aumente el tamaño poblacional, podrá extender el histograma al número de clases que desee, teniendo dominio siempre sobre el número de datos que cae en cada una de estas clases. Discutimos algunos problemas del boxplot e ilustramos ambas gráficas utilizando los datos de la media maratón de Medellín 2009.*

***Palabras claves:*** *Histograma, Boxplot, Percentiles, Convergencia Asintótica, Convergencia Asintótica Conjunta*

## 1. INTRODUCTION

The histogram is the most popular graphic that statisticians and the general public use to visualize data. In statistics, a histogram is a graphical display of tabulated frequencies shown as bars. It shows what proportion of cases fall into each of several categories: it is a form of data binning. The categories are usually specified as non-overlapping intervals of some variable. The categories (bars) must be adjacent. The intervals (or bands, or bins) are generally of the same size, but not necessarily so, since there is no "the best" number of bins, and different bin sizes can reveal different features of the data.

Nevertheless, some theoreticians have attempted to determine an optimal number of bins, but these methods generally make strong assumptions about the shape of the distribution. One should always experiment with bin widths before choosing one (or more) that illustrate the salient features in your data.

Although it is not the aim of this article, let us say that histograms also are used to approach to the density of the data, and often for density estimation: estimating the probability density function of the underlying variable. The total area of a histogram used for probability density is always normalized to 1. If the length of the intervals on the x-axis are all 1, then a histogram is identical to a relative frequency plot.

One of the typical rules of thumb that are given to the students in a statistical course is that they better use equal width intervals for drawing a histogram and, by the way, they receive the Sturges' Rule [1] for determining the number of bars that the histogram must have.

The number of bins, *k*, to draw a histogram can be calculated directly, or from a suggested bin width h:

$$k = \left\lceil \frac{Max(data) - Min(data)}{h} \right\rceil \quad (1)$$

where the braces indicate the ceiling function; i.e., the less integer great or equal than the argument.

There are several manners to determine *h;* we choose the following among the most utilized:

**Scott's choise**:

$$h = \frac{3{,}5\sigma}{n^{1/3}} \quad (2)$$

where s is the sample standard deviation [2].

Scott uses Scott's choice for a normal distribution based on the estimate of the standard error, unless that is zero where it returns 1.

**Freedman-Diaconi's choise[3]**:

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

which is based on the interquartile range, IQR[3].

**Sturges' Formula**:

$$k = \lceil log_2 n + 1 \rceil \quad (4)$$

which implicitly bases the bin sizes on the range of the data, and can perform poorly if $n < 30$.

## 2. THE MODIFIED HISTOGRAM

[4] presents a long chapter on histograms of unequal widths and their interpretations. They pointed out that, in the unequal case, the histogram represents numbers by area, but not height. They do not provide any recommendation on how to choose the limits of the classes that specify the histogram. If we choose the class limits in such a way that they correspond with some predefined percentiles, we also produce an unequal-class-width histogram, but each bar corresponds to a specific percentage of points in the sample.

We could use a modified histogram that is calculated from the estimated deciles, say $(\hat{\xi}_{0,1}, \hat{\xi}_{0,2}, \ldots, \hat{\xi}_{0,9})$, where $\hat{\xi}_\alpha$ is the estimator of $\xi_\alpha$, and $P(X \leq \xi_\alpha) = \alpha$. The advantage is that the number of classes is the same every time we draw a histogram and it is easy to be interpreted by the user.

One property of sample quantiles is that they are strongly consistent for the estimation of the

population quantiles[5]. This means that as the sample size increases, we are obtaining a histogram that will be closer to the histogram drawn using the true deciles. We could call this the *decile histogram* or *decil-gram*. It has also been proved that, under mild conditions, $(\hat{\xi}_{0,1}, \hat{\xi}_{0,2}, \ldots, \hat{\xi}_{0,9})$, converges asymptotically to a multinormal distribution[5]:

where $f$ is the probability density function of the data, *p=0.1, 0.2,…, 0.9,* and *n* is the sample size.

Now, for $0 < p_1 < p_2 < \cdots < p_k < 1$, assessing that $f$ is positive and continuous in a boundary of $\xi_{p_1}, \xi_{p_2}, \cdots, \xi_{p_k}$, then $\left(\hat{\xi}_{p_1}, \hat{\xi}_{p_2}, \cdots, \hat{\xi}_{p_k}\right)$ is asymptotically normal with mean vector $\left(\xi_{p_1}, \xi_{p_2}, \cdots, \xi_{p_k}\right)$ and covariance $\sigma_{ij} / n$ [5], where

$$\sigma_{ij} = \frac{p_i(1-p_j)}{f(\xi_{p_i}) f(\xi_{p_j})} \qquad (6)$$

for $i \leq j$, and $\sigma_{ij} = \sigma_{ji}$ for $i > j$.

One can see that with a finer partition of the interval [0,1], there is a better approximation to the real population density which the sample data come from.

## 3. EXAMPLE

In order to illustrate the above idea, we construct several histograms using the times that long-distance runners spent in the Medellín 2009 Half-Marathon-Race. One of the main characteristics of this kind of competition is that the runners conform well defined groups. Figure 1 presents the histogram that most of the statistical programs produce: equal-width bars where the number of bars follow the Scott's(2), Freedman and Diaconis'(3), and Sturge's(4) Rule. Scott and Freedman and Diaconis histograms have more classes than Sturges's. Also note that the number of classes can vary.

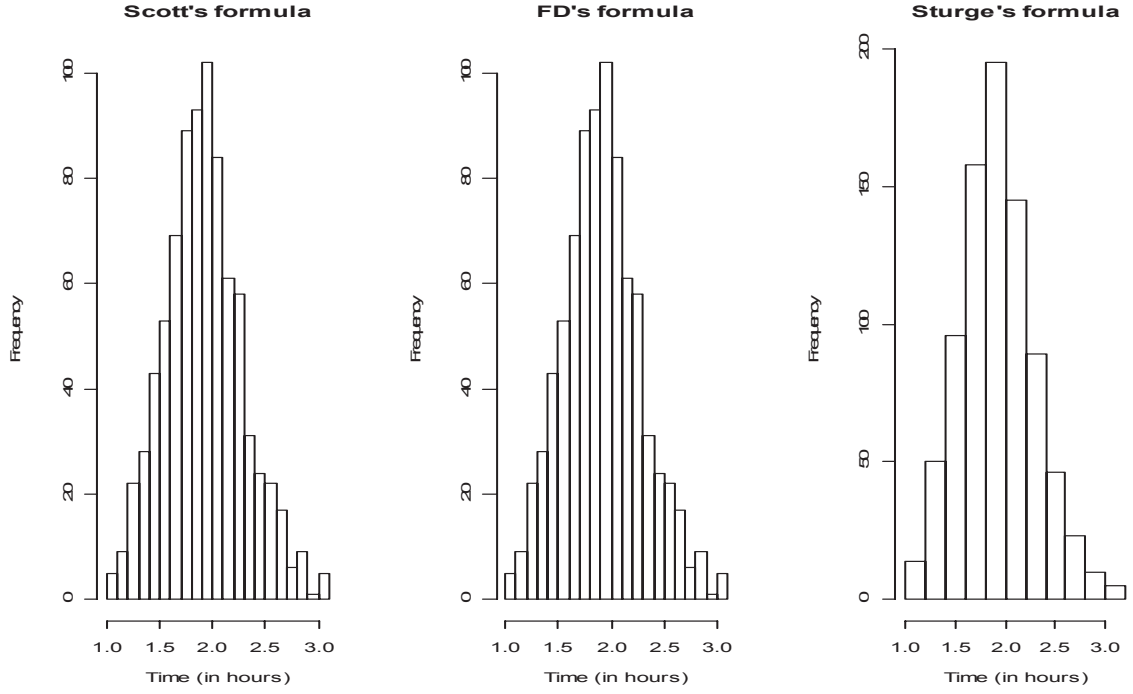


Figure 1. Typical histograms produced by a statistical package using Scott's, Freedman and Diaconis', and Sturges formula using the Medellín 2009 Half Marathon Race.

Note that if the Sturges' performs poorly for $n \leq 30$, then can be concluded that the performance of Scott's and FD's is worse.

Figure 2 shows two histograms: the left one is the histogram of the same dataset using deciles; that is, each bar contains 10% of the sample. This histogram shows a right long tail; histogram on the right is an eleven-equal-width-bar histogram, the details and clarity of the left histogram are lost here, the long right tail is a missed characteristic, for instance.

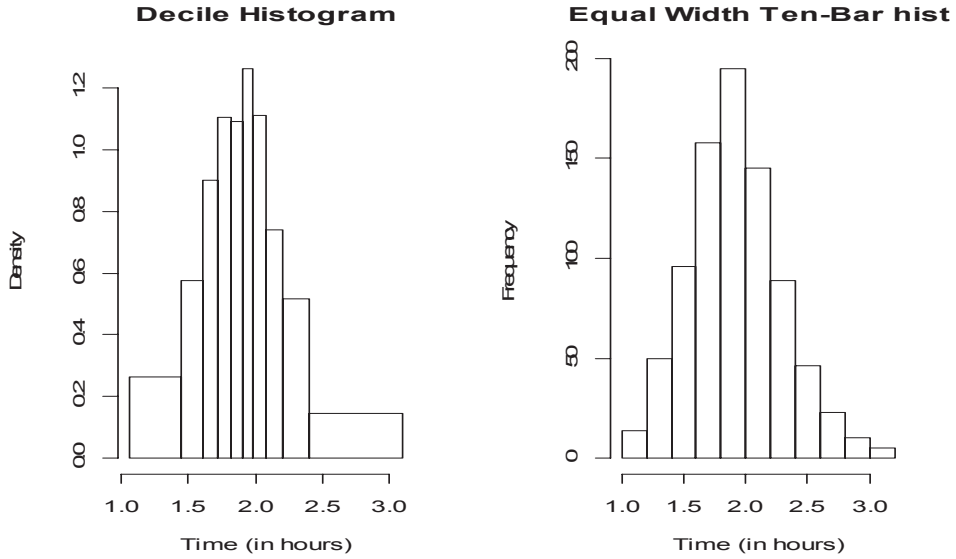


Figure 2. *Left:* Unequal width ten-bar histogram. Classes have limits defined at the deciles. Each bar contains 10% of the data. Special details of the distribution are easily visualized and interpreted.
*Right:* Equal width ten-bar histogram. We are not able to determine the percentage of the sample that falls in each bar

## 4. COMPARING THE QUARTILE HISTOGRAM WITH THE BOXPLOT

If we choose the limits of the classes of a histogram based on the 5-number summary that is used to construct the boxplot and add those points that do not fall in any of the classes explicitly, we obtain a histogram that has the same meaning as the boxplot. Figure 3 shows the boxplot [6] and the equivalent histogram for the marathon times. With this histogram we see the shape of the distribution better than with the boxplot. This histogram corresponds to a *quartile histogram*. Many people will not accept a histogram of this form because the number of classes is not enough to visualize the distribution of the data, but that people will accept the boxplot with no question at all; nevertheless, this hides multimodality when exists[7]. The figure shows the location of the 5-number summary histogram; it is not shown the place of atypical observations as it is usually shown on a boxplot, but they can be added.

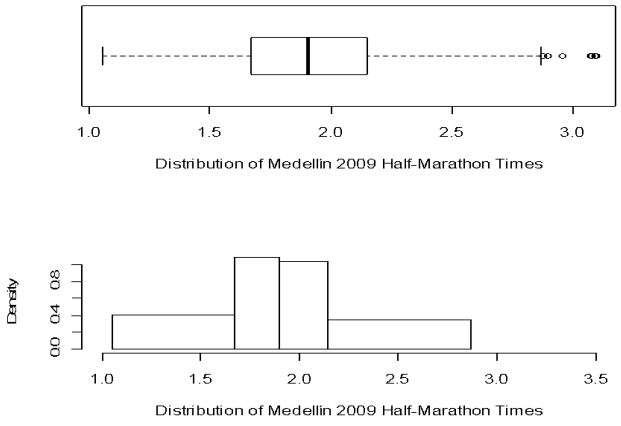


Figure 3. Constructing a histogram that is equivalent to a boxplot. The histogram and the boxplot were drawn using the same information.

## 5. CONCLUSIONS

We can modify the histogram to produce a graphic that conveys more information than the current histogram, the *decil-gram*. The limits of the classes are point estimations of the real limits, this consideration allows us to undertake the study of a new class of histogram in which the limits of bars are confidence intervals containing the actual ones. The histogram currently used is not quite accepted because we lose some characteristics of the distribution, such as the number of data per class, and needs the boxplot to analyze the whole data, this means that we cannot rely upon the Sturges based histogram (or Scott's or Freedman's and Diaconis' histogram) alone to analyze a dataset.

In regards to the boxplot, weather dataset presents multimodality, the graphic hides this characteristic; thus, the researcher must look for other kind of boxplot.

## 6. REFERENCES

[1] Sturges, H. A. The Choice of a Class Interval. *J. American Statistical Association*: 65–66. 1926.

[2] Scott, David W. On Optimal and Data-Based Histograms. *Biometrika* **66** (3): 605–610. 1979

[3] Freedman, David and Diaconis, P. On the Histogram as a Density Estimator: $L_2$ theory. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* **57** (4): 453–476. 1981

[4] Freedman, D., Pisani, R., and Purves, R. *Statistics.* W.W. Norton & Company: New York. 1978.

[5] Serfling, R.J. Approximation Theorems of Mathematical Statistics. John Wiley & Sons: New York 1980.

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

[7] Fridge, M., Hoaglin, D.C., and Iglewicz, B. Some Implementations of the Boxplot. *The American Statistician,* Vol. 43, No. 1, pp. 50-54. 1989.