Société de Calcul Mathématique SA
Outils d'aide à la décision

Building Histograms:

An attempt to normalize the construction

by Bernard Beauzamy
February 2014

Summary

Building a histogram is usually the first step towards any probabilistic or statistical interpretation; this concerns risks analysis, reliability of equipments, epidemiology, and so on. Therefore, there is an obvious need for normalization of the construction: all people with the same data should get the same result, which is not the case presently.

The present paper is an attempt to define "good practices" in this respect.

We show how to define the lower bin (not necessarily from the lowest value) and the upper bin, how to compute the number of bins, and how to give a consistent definition for all bins. Finally, we give a VBA code for the implementation.

*****

I. Introduction

The construction of a histogram is the preliminary step towards the exploitation of any real-life experiment: one defines "bins" and counts how many data fall into each bin. Then the representation is made with bars, the height of which is proportional to the number of values. This is quite easy and classical. Usually, for further use, only this histogram is kept; the original values are lost.

Here are some examples:

- The bins are classes of age, from instance 0 to 10 years, 10 -20, and so on;
- The bins are months of operation for the components of some plant;
- The bins are classes of concentration for some pollutant, for instance 0-1 mg/l, 1-2, and so on.

In such examples, the definition of the bin comes from some "rounding up" procedure: people aged 11 are considered in the bin 10-20, and so on.

But the result of the tests which are performed later are often dependent upon the way the histogram has been produced. For instance, if we are interested in the number of occurrences of children leukemia around nuclear plants, we may consider bins made with circles of radius 1 km, that is, 0-1, 1-2, and so on, or we can do the same with radii in miles. Since the numbers of occurrences of the disease are extremely small (usually 0), the way we define the bins may have some influence. For instance, if we know that some occurrence was found at 4.5 km from the plant, and if we define our bins with 0-5 km, 5-10, and so on, then of course we will conclude that the closest bin to the plant has some occurrence of the disease; the conclusion would be quite different if we had kept 0-1, 1-2, and so on.

Such "tricks" are often used by people willing to obtain a desired conclusion, either by ignorance or by intention.

The obvious conclusion of our remark is that the outcome of the study should be more or less independent of the width of the bins. In order to show that, the statisticians have to demonstrate that they tried at least two different widths, and obtained the same results (a precaution which is often ignored).

So, there is a clear need for standardization of the histograms, meaning that from the same set of data different people should produce the same histograms. We present here a way to obtain this standardization.

II. Defining the scale

The definition of the bins usually involves some scale, as we saw earlier: km, miles, g/l, or whatever is appropriate. This scale should be linked with the objective of the study (what do we want ?) and not really with the data themselves. It may happen, and happens quite often, that the data are too poor to allow the precision one expected; this means that more data will have to be collected, and one should not try to produce an artificial precision from the existing data.

Let us take an example: we measure temperatures and are interested in a precision of 0.2°C. Let $t_{\min}$ be the lowest temperature we want to consider. It is usually the smallest in our set of data, but not necessarily: perhaps, we met $-25.34°C$, but we might be interested in the value $-25.41°C$ which has been met somewhere else. Similarly, let $t_{\max}$ be the highest value we want to consider.

III. The width

Let w be the width of the bins; in our example, $w = 0.2°C$. We want to define K bins, all of them with width w, containing all our numbers, from $t_{\min}$ to $t_{\max}$. Since we want to use some "rounding up" procedure, all the extremities of the bins will be of the form kw, where k is an integer (positive or negative). In our case, all bins will be multiples of $0.2°C$.

Usually, there are disputes about the type of bin: should they be of the form $c_k \le x < c_{k+1}$, or $c_k < x \le c_{k+1}$, and how should we handle consistently the first and the last ?

In order to answer these small difficulties, we introduce an empty bin, left of the first non-empty one, and an empty bin, right of the last non-empty one. So all our bins are of the same model, namely $c_k \le x < c_{k+1}$ and our histogram will always start and finish with empty bins. This is a good precaution, because some new data might come later and, this way, we have fewer chances to need to redefine our histogram.

So we define:

$$k_{\min} = \text{int}\left(\frac{t_{\min}}{w}\right) - 1$$

where $\text{int}(\ )$ defines the integral part of a real number, that this the largest integer smaller than this number (sometimes written as $[.]$), and we set:

$$c_{\min} = w\, k_{\min}$$

This will be the lower bound of the first bin.

Similarly, we define:

$$k_{\max} = \text{int}\left(\frac{t_{\max}}{w}\right) + 1$$

and:

$$c_{\max} = w\, k_{\max}$$

This will be the lower bound of the last bin.

The number of bins will be:

$$K = k_{\max} - k_{\min} + 1$$

IV. VBA implementation

So, if we want to implement the construction, for instance in VBA, we set:

dim histo(1 to Ktot) as integer (or as "long" if we expect many values)

Let N be the total number of observations, or points of interest. We construct the histogram the following way:

```
dim t as double 'to contain the data
for i = 1 to N
t= sheets(1).cells(i+1,1) 'or whatever column contains the data
j = int(t/w)
histo(j)=histo(j)+1
next i
```

Note that with this definition of the histogram, the classes are of the form $c_k \le x < c_{k+1}$; here again we recommend a normalization of such a choice.

When this construction is done, the representation of the histogram is very simple:

```
for k = 1 to Ktot
sheets(2).cells(k+1,1) = k × w 'the values on the x axis
sheets(2).cells(k+1,2) = histo(k)
next k
```