Société de Calcul Mathématique SA
*Outils d'aide à la décision*
*Tools for decision help*

Probabilistic Studies:

# Normalizing the Histograms

Bernard Beauzamy

December, 2012

## I. General construction of the histogram

Any probabilistic study usually starts with the construction of an histogram: one defines some classes and counts how many points fall into each class. The most common situation is as follows : we have a sample of real values $x_i$, $i = 1,...,Itot$ ; let $m = \min x_i$ and $M = \max x_i$. We want to build an histogram with $K$ classes, from this sample.

What people do in general is to divide the interval $[m, M]$ into $K$ classes, of width $\frac{M-m}{K}$.

But this approach has several drawbacks, and people are not often conscious of them:

- The boundaries of the classes are strongly dependent of the values of $m$ and $M$, and would be modified if these values were changed, for instance if the sample grew bigger;

- These boundaries do not take into account the uncertainties which certainly exist upon the values of $m$ and $M$ ;

- All classes are of the form $a \leq x < b$, except the last one, which is of the form $a \leq x \leq b$, since the value $M$ is necessarily met.

An histogram should be viewed as a measurement device, just like a thermometer. It gives an information, with some accurracy. Therefore, the measurement device should be as independent as possible from the sample. Of course, it cannot be totally independent.

The number of classes itself, namely $K$, is not arbitrary but should reflect the performances of the measurement device. Indeed, it is linked with the precision we expect. When we build an histogram, two points in the same class are considered as the same point. When we make a measurement, we consider that, if the value $x$ is read, it might as well be anything between $x-\varepsilon$ and $x+\varepsilon$, $\varepsilon$ being considered as the precision of the measurement device. So we have some rough link between both concepts :

$$\frac{M-m}{K}=2\varepsilon, \tag{1}$$

since $\frac{M-m}{K}$ is the width of each class, from the histogram point of view, and $2\varepsilon$ is the width of each class, from the precision point of view.

In order to answer the difficulties mentioned above, we will build classes such that the first one is centered at $m$ and the last one centered at $M$.

Therefore, the centers of the classes will be:

$$c_k=m+\frac{k}{K-1}(M-m)\ ,\ k=0,...,K-1 \tag{2}$$

The half-width of a class is:

$$l=\frac{M-m}{2(K-1)} \tag{3}$$

A point $x$ belongs to the class $C_k$, with center $c_k$, if:

$$m+\frac{k}{K-1}(M-m)-\frac{M-m}{2(K-1)}<x\le m+\frac{k}{K-1}(M-m)+\frac{M-m}{2(K-1)} \tag{4}$$

So, all our classes will be here of the form $a<x\le b$.

Condition (4) may be written:

$$k<(x-m)\frac{K-1}{M-m}+\frac{1}{2} \tag{4a}$$

and:

$$(x-m)\frac{K-1}{M-m}-\frac{1}{2}\le k \tag{4b}$$

which means that $k$ is defined by:

$$k = \left[ (x-m)\frac{K-1}{M-m} + \frac{1}{2} \right], \tag{5}$$

where $[x]$ is the integer part of $x$, that is the largest integer smaller than $x$.
So, a VBA code may be written as follows; $Itot$ is the total number of lines is the table, $\min\_values$ and $\max\_values$ are respectively the min and the max, and $Ktot$ is the number $K$ above:

```
for i = 1 to Itot
k= int( (x(i)-min_values)*(Ktot-1)/(max_values-min_values) +1/2)
histo(k)=histo(k)+1
next i
```

As it stands now, the method has a drawback: the extremities of each class are rational numbers, usually with many decimal digits, which looks unnatural, with respect to the requirement for a given precision. For instance, a class might appear as 443.556-464.444. Its width is almost 1, which means that we do not want to distinguish between numbers with a difference say of 0.5. But still, we give 3 digits after the decimal point, which looks absurd. So, we have to study how to round up the values.

## II. Rounding up the values

If we accept the idea that all values in our sample are subject to some measurement error, the simplest way of taking it into account is to round up each value. Let $\varepsilon$ be the precision we accept, and let $\varepsilon = 10^{-\nu}$ for some integer $\nu > 0$. Then each value $x_i$ is replaced by

$$rx_i = round(x_i, \nu)$$

which is the number with $\nu$ decimal places closest to $x_i$.

Then of course $m$ and $M$ will also be rounded to $\nu$ decimal places. But even so, the centers of the other classes will not be rounded to the same number of decimal places, because they are multiples of $\frac{M-m}{K-1}$.

It is important to keep the fact that all classes should have the same width: for instance, when we generate random numbers, the percentage of points in each class depends on the width of the class.

So, what we do is as follows: we do not try to replace all centers by approximate values; we keep the rational value. But still, in the Excel cells, we may present the result with a given number of digits. We will write for instance :

```
For k = 0 To Ktot
Sheets(3).Cells(k + 2, 1) = Round(min_values + k / (Ktot - 1) * (max_values - min_values) - (max_values - min_values) / (2 * (Ktot - 1)), 2) & "-" & Round(min_values + k / (Ktot - 1) * (max_values - min_values) + (max_values - min_values) / (2 * (Ktot - 1)), 2)
Sheets(3).Cells(k + 2, 2) = histo(k)
Next k
```

This way, we will have an histogram of the following sort:

| interval | number of occurrences |
|---|---|
| 0-0,01 | 43 |
| 0,01-0,02 | 90 |
| 0,02-0,03 | 115 |
| 0,03-0,04 | 98 |

The endpoints look simple, but still the classes have the same width. If this example, the value of $l$ was 0,00504934908163668, the value of the min was 0.000189483165740967, the value of the max 0.999960601329803.

## III. Avantages of the method

This method answers the difficulties mentioned previously:

- If the sample grows bigger, the classes are not necessarily modified, as long as no value becomes smaller than $m-l$ or larger than $M+l$. Of course, if more points appear below $m$ or above $M$, the values $m, M$ will not be centers of classes anymore, but the definition of the classes will not be modified.

- The construction incorporates the uncertainties upon the values.

- All classes have the same form, namely $a < x \leq b$.

- The method may be fully automatized. All we need is $m, M, K$.

An interesting application, which we recently met, is that this method allows us to show that some variables have identical laws. Assume for instance that we have one random variable $X$ and another one $Y$ which turns to be $Y = 100X$. If we build the histograms the usual way, by hand, we might not notice this. Assume for instance that the minimum value for $X$ is 0.04 and we want $K = 100$ classes. We would take for first interval for $X$ the interval $0 - 0.1$.

For $Y$, the smallest value is 4, and we would probably take as first interval $0 - 5$. We would not see that this is the same variable, up to a multiplication by a constant factor 100.

If the classes are defined in an automatic manner, as was previously explained, the link between $X$ and $Y$ is obvious. All endpoints are multiplied by 100, and the number of points in each class is the same.

## IV. Loss of information

Quite clearly, when we perform an histogram, some information is lost: all points belonging to the same class are identified together, and identified to the center of the class. Simply consider the extreme values $m$ and $M$ : it is therefore better to have them as centers of classes, and no information will be lost upon them. So the indicator "total loss of information when performing the histogram" is one more reason for the choice we indicate.

## V. Refining the definition of the grid

In the work above, we decided that the extreme classes would be centered at the extreme values of the sample. We may wonder if there are better choices. We now investigate this question.

The set of classes will be called a "grid". As before, there are $K$ classes, denoted by $C_K$, and the number $K$ is fixed. The width of the classes is denoted by $2l$ ; it is the same for all classes, and it is fixed, since it results from the precision which is required. We denote by $c_k$ , $k = 1,...,K$ , the center of the class $C_k$. Our question now is how to choose the position of the center $c_1$, since all other centers will follow.

If some points $x_i$ fall into the class $C_k$, they are identified to its center $c_k$ ; so there is a loss of information equal to $|x_i - c_k|$ for each of them. We are looking for the position of the grid, that is the position of $c_1$, which will minimize this loss of information.

As before, we set $m = \min(x_i)$ et $M = \max(x_i)$ ; we admit the fact that the grid is larger enough to cover all the sample, with half a class on each side. This gives the inequality:

$$2(K-1)l \geq M - m \tag{1}$$

It is useless to have empty classes; either before $m$, or after $M$. So the first class will contain $m$ and the last one will contain $M$ , and we get the conditions:

$|m - c_1| \leq l$

$|M - c_K| \leq l$

Since $c_K - c_1 = (2K-2)l$ , this is compatible with condition (1), since:

$|M - m| \leq |M - c_K| + |c_K - c_1| + |c_1 - m| \leq 2l + (2K-2)l = 2Kl$

The total number of classes, taking (1) into account, is:

$$K = \text{int}\left( \frac{M-m}{2l} \right) + 1 \tag{2}$$

For instance, if $m = 0,\ M = 1$ and $l = 1/20$ (classes of width 1/10), we find $K = 11$.

So, the difference with the paragraphs above is that now $c_1$ may not be exactly in $m$. More precisely, we want to position $c_1$, under the constraint:

$$m - l \le c_1 \le m + l \tag{3}$$

and we want to minimize the quantity:

$$Q = \sum_k \sum_{i \in C_k} |c_k - x_i| \tag{4}$$

In the definition of this quantity, we consider that the total loss of information is simply the sum of all individual losses of information; we do not see any reason to take, for instance, a quadratic sum.

Since $c_k = c_1 + 2(k-1)l$, $k = 1,...,K$, the quantity $Q$ may be written:

$$Q = \sum_k \sum_{i \in C_k} |c_1 + 2(k-1)l - x_i| \tag{5}$$

If we move $c_1$, but still keeping each $x_i$ in the same class, then obviously $Q$ is a linear function of $c_1$: the absolute value becomes a quantity $a - c_1$ or $c_1 - a$ and their sum is linear.

The function $Q(c_1)$ is continuous and piecewise linear. The discontinuities of the derivative appear for the values of $c_1$ such that, for some $i$ and some $k$:

$$|c_1 + 2(k-1)l - x_i| = l$$

that is:

$$c_1 + 2(k-1)l - x_i = \pm l$$

So, these are points $c_1$ of the form:

$$c_1 = x_i - (2k-3)l$$

or of the form:

$$c_1 = x_i - (2k-1)l \tag{6}$$

for $i = 1,...,N$ and $k = 1,...,K$. Both forms are equivalent, if we replace $k$ by $k-1$.

So we have $NK$ points of discontinuity for the derivative, and all we have to do is to compute the values of the function at these points. The minimum value of $Q$ may be reached only at such points, since the function is linear in between.

A given point $x_i$ belongs to the class $k = k(c_1)$ defined by:

$$k = \text{int}\left( \frac{x_i - c_1}{2l} + \frac{1}{2} \right) \tag{7}$$

So our program goes as follows. Here, we generated a random sample x(i) between 0 and 1, of size Itot=10 000. We take lc = 1 / 20 (half width of a class). We have Ktot = 11. Let c(k) be the centers of the classes.

```
Dim c1 As Double 'position of the first center | Dim c0 As Double | Dim dist As Double
Dim d_min As Double 'shortest distance | d_min = 10000 'initialization with high value
Dim i1 As Integer | Dim k1 As Integer
For i1 = 1 To Itot
For k1 = 1 To Ktot
c1 = x(i1) - (2 * k1 - 1) * lc 'enumeration of all possible first centers
If c1 > -lc And c1 < lc Then
For k = 1 To Ktot
c(k) = c1 + 2 * (k - 1) * lc 'enumeration of all centers, the first one being given
Next k

For i = 1 To Itot
kk = Int((x(i) - c1 + lc) / (2 * lc)) + 1 'the index of the center closest to x(i)
dist = dist + Abs(c(kk) - x(i))
Next i
If dist < d_min Then
d_min = dist
c0 = c1
End If 'If dist < d_min Then
End If 'If c1 > -1 / 20 And c1 < 1 / 20 Then
dist = 0
Next k1
Next i1
```

The result is the value of $c_1$ . In the present case, we find $c_1 = 0.026$, which means that the value $c_1 = 0$ was not best : a slight shift of the grid to the right minimizes the loss of information.

The values of $c_1$ to be searched are of the form $x_i - l, x_i - 3l, x_i - 5l, \ldots$ so, quite obviously, for a given $x_i$, only one of them may be in the interval $[m - l, m + l]$.