Société de Calcul Mathématique SA
Outils d'aide à la décision
depuis 1995

Présentation "canonique" d'un histogramme

par Bernard Beauzamy

Juillet 2022

Résumé

Il est commode de constituer un histogramme à partir d'un ensemble de données. Un histogramme est fait d'un certain nombre de classes consécutives, de même taille ; on compte combien de données tombent dans chacune des classes. A priori, la largeur des classes et le point de départ de la première sont à la disposition de l'utilisateur, mais certains choix sont plus pratiques que d'autres, et favorisent mieux la compréhension. En particulier, nous sommes habitués au système décimal, et nous préférerons les classes qui s'appuient sur ce système. En outre, le nombre de classes ne doit pas être trop élevé : 20 est un maximum.

Notons respectivement m, M la plus petite des valeurs présentes parmi les données ; on choisira des classes de largeur w, avec $w = 10^{\alpha}$, où :

$$\alpha = \text{int}\left(Log_{10}\left(\frac{M-m}{18} \right) \right) + 1$$

($\text{int}(x)$ désigne la partie entière de x, c'est-à-dire le plus grand entier inférieur ou égal à x)

La première classe commence à $k_0 10^{\alpha}$ avec $k_0 = \text{int}\left(\frac{m}{10^{\alpha}} \right)$.

La dernière classe finit à $k_1 10^{\alpha}$ avec $k_1 = \text{int}\left(\frac{M}{10^{\alpha}} \right) + 1$.

On note w la taille des classes, K le nombre de classes, A le début de la première classe, B la fin de la dernière classe, si bien que $B = A + Kw.$ Les bornes des classes sont $A, A+w, A+2w, ..., A+Kw.$ Par définition, chaque classe est fermée à gauche, ouverte à droite : $a \le x < b$.

A priori, w et K peuvent être choisis arbitrairement, pourvu que toutes les données recueillies tombent entre A et $B.$ Mais, en pratique, les utilisateurs préféreront les situations suivantes :

- Le nombre de classes ne doit pas être trop élevé, disons $K \le 20.$ Si on a des centaines de classes, on n'y comprend rien ;

- La taille des classes doit être une puissance de 10, puissance négative ou positive : $...,1/100,1/10,1,10,100,....$ Formellement, on a parfaitement le droit de prendre $w = 0{,}1234,$ mais les gens sont habitués au système décimal et il est préférable, pour présenter les résultats, de le faire à l'intérieur de ce système.

Nous avons donc les deux contraintes :

$K \le 20$
$w = 10^{\alpha}$, où α est un entier, positif, négatif ou nul.

Notons respectivement m, M la plus petite et la plus grande valeur dans la série de données que l'on doit ranger dans l'histogramme. On a, par construction, les inégalités :

$$A \le m < A + w,\ A + (K-1)w \le M < A + Kw$$

qui signifient que la première classe contient m et la dernière contient $M.$

On cherche donc à positionner une "grille" de pas 10^{α} (le "pas" est la distance entre deux éléments de la grille), de telle manière qu'elle contienne tous les éléments de la suite enregistrée, en utilisant au plus 20 intervalles.

Soit k_0 le plus grand entier tel que $k_0 10^{\alpha} \le m$ et k_1 le plus petit entier tel que $k_1 10^{\alpha} > M$; ou encore :

$$k_0 \le \frac{m}{10^{\alpha}},\ k_1 > \frac{M}{10^{\alpha}}$$

Par définition :

$$k_0 = \mathrm{int}\left(\frac{m}{10^{\alpha}}\right),\ k_1 = \mathrm{int}\left(\frac{M}{10^{\alpha}}\right) + 1,$$

où $\text{int}(x)$ désigne le plus grand entier inférieur ou égal à x (appelé "partie entière" de x).

L'entier α doit être choisi pour que $k_1 - k_0 \leq 20$, c'est-à-dire :

$$\text{int}\left(\frac{M}{10^\alpha}\right)+1-\text{int}\left(\frac{m}{10^\alpha}\right)\leq 20$$

$$\text{int}\left(\frac{M}{10^\alpha}\right)-\text{int}\left(\frac{m}{10^\alpha}\right)\leq 19$$

Ceci sera réalisé a fortiori dès que :

$$\frac{M}{10^\alpha}-\text{int}\left(\frac{m}{10^\alpha}\right)\leq 19 \quad (1)$$

Mais puisque

$$\text{int}\left(\frac{m}{10^\alpha}\right)\leq \frac{m}{10^\alpha} < \text{int}\left(\frac{m}{10^\alpha}\right)+1$$

$$\frac{m}{10^\alpha}-1 < \text{int}\left(\frac{m}{10^\alpha}\right)\leq \frac{m}{10^\alpha}$$

$$\frac{M}{10^\alpha}-\text{int}\left(\frac{m}{10^\alpha}\right)\leq \frac{M}{10^\alpha}-\frac{m}{10^\alpha}+1$$

et par conséquent (1) sera réalisé dès que :

$$\frac{M}{10^\alpha}-\frac{m}{10^\alpha}+1\leq 19$$

ou encore

$$\frac{M-m}{10^\alpha}\leq 18 \quad (2)$$

L'entier α doit être le plus petit possible : la grille doit être la plus fine possible, tout en respectant la contrainte de ne pas dépasser 20 intervalles. Il vient :

$$\frac{M-m}{18}\leq 10^\alpha\ ,\ \alpha \geq Log_{10}\left(\frac{M-m}{18}\right)$$

et finalement :

$$\alpha = \mathrm{int}\left(Log_{10}\left(\frac{M-m}{18} \right) \right) + 1$$

Par exemple, si $m = 0$ et $M = 95,$ on trouve $\alpha = 1,$ ce qui est conforme à l'intuition : toutes les classes auront pour largeur 10 ; ce seront $0-10, 10-20, ..., 90-100.$

La première classe commence à $k_0 10^\alpha$ avec $k_0 10^\alpha \le m$ et donc :

$$k_0 = \mathrm{int}\left(\frac{m}{10^\alpha} \right)$$

La dernière classe finit à $k_1 10^\alpha$ avec $k_1 10^\alpha > M$ et donc :

$$k_1 = \mathrm{int}\left(\frac{M}{10^\alpha} \right) + 1$$

Implémentation en VBA sous Excel ; voici le code :

```
Sub macro1()
Sheets(2).Cells(1, 1) = "données à introduire"
Sheets(2).Cells(1, 2) = "données triées"
Ntot = Sheets(2).Range("A2").End(xlDown).Row
Sheets(1).Cells(3, 2) = Ntot - 1
Sheets(1).Cells(3, 1) = "nombre de données"

val_min = Application.WorksheetFunction.Min(Sheets(2).Range("A2:A" & Ntot))
Sheets(1).Cells(4, 2) = val_min

val_max = Application.WorksheetFunction.Max(Sheets(2).Range("A2:A" & Ntot))
Sheets(1).Cells(5, 2) = val_max

Dim DATA As Variant 'données mises en mémoire
DATA = Sheets(2).Cells(1, 1).Resize(Ntot, 1)

Dim w As Double 'taille de la classe (width)
Dim alpha As Integer
alpha = Int(Log((val_max - val_min) / 18) / Log(10)) + 1
w = 10 ^ (alpha)

Sheets(1).Cells(6, 2) = w

Dim k_min As Long ' première classe
k_min = Int(val_min / w)
Sheets(1).Cells(7, 2) = k_min
```

```
Dim A_min As Double
A_min = w * k_min 'début de la première classe
Sheets(1).Cells(7, 2) = A_min
Sheets(1).Cells(7, 1) = "début 1ere classe"

Dim k_max As Long 'numéro de la dernière classe
k_max = Int(val_max / w) + 1

Sheets(1).Cells(8, 2) = k_max

Dim A_max As Double
A_max = w * k_max 'début de la dernière classe
Sheets(1).Cells(8, 2) = A_max
Sheets(1).Cells(8, 1) = "début dernière classe"

Dim Ktot As Long 'nb de classes
Ktot = k_max - k_min + 1
Sheets(1).Cells(9, 2) = Ktot
Sheets(1).Cells(9, 1) = "nb classes"

Dim histo() As Long
ReDim histo(k_min - 1 To k_max + 1) As Long

Dim m1 As Double
Dim i As Long
Dim j As Long
Dim dd As Double

For i = 2 To Ntot
If IsNumeric(DATA(i, 1)) = True Then
j = Int(DATA(i, 1) / w)
m1 = m1 + DATA(i, 1)
histo(j) = histo(j) + 1
End If
Next i

Sheets(1).Cells(10, 2) = Round(m1 / Ntot, 2)
Sheets(1).Cells(10, 1) = "moyenne"

Sheets(3).Cells(1, 1) = "intervalle"
Sheets(3).Cells(1, 2) = "nb données"
Sheets(3).Cells(1, 3) = "histo représentatif"
Sheets(3).Range("A2:B1000000").ClearContents

For j = k_min To k_max
Sheets(3).Cells(j - k_min + 3, 2) = histo(j)
Sheets(3).Cells(j - k_min + 3, 3) = histo(j) / Ntot
```

```
        Sheets(3).Cells(j - k_min + 3, 1) = w * j
        Next j

        Sheets(1).Cells(13, 1) = "Statut"
        Sheets(1).Cells(13, 2) = "Terminé"
End Sub
```