SEP 2016: histplot() and histc() upgrade

Sommaire

A bit of history

The discussion in the bug report #6306 and the SEP #110 aimed

- to split the computation of an histogram on one hand, and its display on the other hand.
- to become able to get the histogram's data that -- up to then -- were not available from histplot()
- to introduce a "normalized" way to compute the histogram heights.

It introduced in Scilab 5.4 a new *histc()* macro having the same syntaxes as *histplot()* for input data, and returning histogram's heights and memberships of input data in defined bins. See https://codereview.scilab.org/#/c/13155/: new histc() + histplot() output added.
However,

- The *normalization* option has been badly designed, since several ways to normalize data can be defined, whereas a boolean can take only one active value.
- By the way, these ways to compute the histogram do not *normalize* heights or area, in such a way that the name of this option is misleading. It shall be changed.
- In *histplot()*,
    - *normalization* has been appended to the very long list of input parameters, instead of being inserted after **data** -- as for *histc()* -- and before graphical options.
    - the default *normalization* value has been set to *%T* and then breaks back-compatibility for nothing.

By chance, neither the *normalization* nor the *polygon* options implemented for histplot() were yet documented for it in Scilab 5.5.2.

Moreover, **other features are still missing**:

- Only one method to compute bins is available. Other methods could be implemented and called through their name as a string.
- A default method to compute bins is missing.
- Neither *histc()* nor *histplot()* return bin's edges when these ones are computed.
- With *histc()*, it is now possible to compute an histogram out of *histplot()*, but there is still no way to make *histplot()* just plotting it without recomputing it.
- As for the *normalization* option, the *histplot() polygon* option has been appended to the list of input parameters instead of being inserted before the *style* option.
- Only vertical-up histograms can be displayed. vertical-down and Horizontal histograms shall become available.
- Polar histograms shall be supported.
- There is no option to display heights values on the bars.
- Processing of *-%inf*, *%inf*, and *%nan* values is undefined. There should be ways to make then counted and become able to ignore them or to take them into account when required.
- *histc()* and *histplot()* do not yet support text data, despite *dsearch()* on witch *histc()* is based was extended to text since Scilab 5.5.0.

New syntaxes proposed

histc()

- **Existing syntaxes**
  - *[heights] = ..*
  - *[heights, memberships] = ..*
  - *.. = histc(nBins, data)*
  - *.. = histc(edges, data)*
- **Removed syntaxes**
  - *histc(.., data, normalization)*

    If backward-compatibility is of concern despite the short history of the *normalization* option, this one may be warn-obsoleted (for some while or forever) and automatically translated using the *histScale* option.
- **New syntaxes**
  - *histc(data)*
  - *histc(**binsMethod**, data)* with *binsMethod = "sqrt"(default) | "sturges" | "freediac"*:
    - *sqrt*: *nbins = sqrt(size(data,"*")*
    - *sturges*: Sturges criteria: *nbins = ceil(1 + log2(size(data,"*")))*
    - *freediac*: Freedman-Diaconis criteria: binWidth = 2*iqr(data)*size(data,"*")^(-1/3). This method can't be applied to text data.
  - *histc(.., data, **histScale**)* with *histScale = "counts"|"countsNorm"|"density"|"densityNorm"*:
    - *"counts"*: the height is the bin's number of members (default)
    - *"density"*: the bin's area is the bin's number of members
    - *"countsNorm"*: as *"counts"*, divided by the total number of data. Discussion: should *-%inf*, *%inf*, *%nan* values, or/and data out of defined bins be taken into account for the "normalization"?
    - *"densityNorm"*: as *"density"*, divided...
  - *[heights, memberships, **binsDef, outside**] = histc(..)* with
    - *binsDef* edges (continuous) or values (discrete)
    - *outside = [Nminf, Npinf, Nnan]* counts occurrences of *-%inf*, *%inf* and *%nan* values. In the *membership* array, *-%inf* will have the index *-%inf*, *%inf* will have the index *%inf*, and *%nan* will have the index *-1*

histplot()

- **Existing syntaxes**
  - *[heights, memberships] = **histplot(**Nbins|edges, data [, <graphical options>..]**)***
- **New syntaxes**
  - *[heights, memberships, **binsDef, outside**] = histplot(..)* : as for *histc(..)* (see here-above)
  - *histplot(data)*
  - *histplot(**binsMethod**, data)*
  - *histplot(.. data, **histScale** ..)*
  - *histplot(.. data, **histScale, dispOptions** ..)* with *dispOptions* being a vector of one to five strings *provided in any order*, among the following, specifying options to display the histogram:
    - positions: *"bottom"* or *""* (default) | *"up"* | *"left"* | *"right"* : draw the histogram with its base at the given position wrt its bars. **Are these values clear enough?**
    - *"cumulate"*: draw the cumulated "staired" histogram instead of the simple one.
    - *"polygon"* : draw as well the polygon of frequencies or densities etc.

        - *"polar"* : draw the histogram in polar mode, bins being rescaled over the full [0, 180°] fan. Values of bins edges are displayed on an external graduated half circle.
        - *"values"* : display values of heights on the bars

    - *histplot(**binsType, bins, heights, dispOptions** ..)* to display as is an histogram already computed,
        - with *binsType = "binsEdges" | "binsValues"* : must be explicit (no default value)
        - Since allowed values of the *binsType* string are all distinct from values of the *binsMethod* string, parsing input arguments can easily detect this specific syntax.

Questions an discussion

- Since now, since Scilab 5.4, *histc()* can compute and return histogram results, and that histplot() can be fed by them, **should *histplot()* still return the results?**
- In output, do we shift the *memberships* array (introduced recently, 5.5.0) in 3rd position to put computed *edges* in argout#2?
- How do we manage *%nan* and *%inf* ?

Author(s) : Samuel GOUGEON

CategorySep

public: SEP/2016 - histplot() and histc() upgrade (dernière édition le 2016-06-10 13:28:51 par sgougeon@free.fr)