Calculation of the quantiles, mode(s), histogram and rank correlation

Scope

This document provides background information on and mathematical description of the algorithms used in the functions implemented in the module **ordered_functions** and responsible for the calculations of the statistical properties related to the shape of the sample data distirbution and the position of the individual data points within the sample.

Calculation of median value

The *median* of a distribution (or sample) **X** is a such value that exactly half of the data points in the distribution (or sample) have values less than or equal to that *median* value, therefore the other half of the points have values greater than or equal to that *median*.

In order to calculate the *median* the input data sequence **X** is sorted in the ascending order, resulting in a sorted sequence **S** of the length *N*, which elements have indexes from 0 to N-1. If *N* is odd, the *median* is the element exactly in the middle of the sequence, otherwise the *median* is calculated as the arithmetic mean of two elements around the mid-point. i.e.:

$$\texttt{Median} = \begin{cases} S[\lfloor \frac{N}{2} \rfloor]; \texttt{if}; N; \mathbf{mod}; 2 = 1 \\ \frac{S[\lfloor \frac{N}{2} \rfloor]+S[\lfloor \frac{N}{2} \rfloor+1]}{2}; \texttt{if}; N; \mathbf{mod}; 2 = 0 \end{cases}$$

Calculation of a generic quantile

The k-th of m-quantile, or Q_m^k of a distribution (or sample) **X** of length *N* is a such value that exactly $\frac{k}{m} \times N$ data points in the distribution (or sample) have values less than or equal to that value.

In order to calculate a quantile the input data sequence **X** is sorted in the ascending order, resulting in a sorted sequence **S** of the length *N*, which elements have indexes from 0 to N-1. Naturally, $S[0] \equiv \texttt{min}(X) = Q_m^0$ and $S[N-1] \equiv \texttt{max}(X) = Q_m^m$.

Then, for any $0 \leq k < m$ there is such index $0 \leq i < N-1$ that $\frac{i}{N-1} \leq \frac{k}{m} < \frac{i+1}{N-1}$. So, the quantile value is calculated using linear interpolation formula:

$$Q_m^k = S[i] \times (1-p) + S[i+1] \times p \; \texttt{where}; i = \lfloor \frac{k \times (N-1)}{m} \rfloor; \texttt{AND}; p = \frac{(k \times (N-1)); \mathbf{mod}; m}{m}$$

Within this definition the *first quartile* is $Q1 = Q_4^1$ and the *third quartile* is $Q3 = Q_4^3$.

Calculation of the mode(s) of a distribution

The *mode* of (*uni-modal*) distribution is a value, which occurs most frequently with continiuos independent sampling of the distibution. *Bi-modal* and *multi-modal* distributions have two or more values respectively, which occur with the same frequence, which is higher than the frequency of any other value.

In the case of a finite size sample **X** one, two or more values can occur same number of times, which is larger than the number of occurences of any other value, even if the sample is taken from a population

with uni-modal distribution. Thus, the result of the mode(s) calculation on a data sample is always a sequence (list) of values, which may include one or more values.

The data sample **X**, which can include *repetitions*, i.e. the same value occuring more than once, is transformed into a *set* of unique values $S_X = x_i;:;x_i \neq x_j;\forall;i \neq j$. Then, for each value $x_i \in S_x$ its frequency (number of occurences in **X**) is calculated as $F(x_i)$. The maximum frequency $\mathbf{max}(F)$ is determined, and all elements of the set are selected, for which the calculated frequency equals the maximal one; i.e.

$$\texttt{Mode(s)} = x_k;:;F(x_k) = \mathbf{max}(F);\forall;x_k \in S_x$$

Calculation of a histogram of a sample's distribution

Consider a finite size data sample **X**, which values span the range **min(X)** to **max(X)**. One can select a slightly larger interval covering the entire sample data range, i.e. $y_{min} \leq \mathbf{min}(X) \leq \mathbf{max}(X) \leq y_{max}$ and divide it into *N* equidistant bins, each of width *S* with the obvious relation $S \times N = y_{max} - y_{min}$. Each individual bin is characterized by its index *i* and the corresponding 'center' value y_i.

A value *x* from the sample **X** belongs to the i-th bin if $y_i - \frac{S}{2} \leq x < y_i + \frac{S}{2}$. Naturally, there is no sense to select too broad values range for the histogram that some left- and right-side bins are empty. The optimal combination of the number of bins and the bin width (size) is such that $y_0 - \frac{S}{2} \leq \mathbf{min}(x) < y_0 + \frac{S}{2}$ and $y_{N-1} - \frac{S}{2} < \mathbf{max}(x) \leq y_{N-1} + \frac{S}{2}$, where bins are indexed from zero.

With such arrangement, for any value *x* from the sample **X** the index *i* of the bin to which this value belongs can be calculated as $i = \lfloor \frac{x - y_o}{S} + \frac{1}{2} \rfloor$.

If the specific number of bins *N* is required in the histogram, the bin width is calculated as $S = \frac{\mathbf{max}(X) - \mathbf{min}(X)}{N-1}$ and the min and max value are centered to the mid-points of the first and the last bins respectively, i.e. $y_0 = \mathbf{min}(X)$ and $y_{N-1} = \mathbf{max}(X)$.

If the specific bin width *S* is required instead, the arithmetic mean of the distribution *<X>* is centered to the mid-point of its respective bin, and the total number of bins is calculated as $N = \lceil \frac{\mathbf{max}(X) - \langle X \rangle}{S} - \frac{1}{2} \rceil + \lceil \frac{\langle X \rangle - \mathbf{min}(X)}{S} - \frac{1}{2} \rceil + 1$. The position of the mid-point of the first bin is calculated as $y_0 = \langle X \rangle - \lceil \frac{\langle X \rangle - \mathbf{min}(X)}{S} - \frac{1}{2} \rceil \times S$.

Note that with such design the left-most and right-most bins include 'dead-space', thus, effectively, introduce correction factor 0 < *q* < 1 to the far tails of the sample's distribution, which can influence the visual perception of shape of the distribution, especially, if it is close to uniform. On the other hand, it reduces the distribution shape distortion in the case of small number of bins and 'long-tailed' data distribution.

This choice of the first bin position also minimizes variation of the shape of the histogram with slight variation of the bin width, e.g. when requesting a specific bin width vs specific number of bins resulting in a different bin size. In short, wih the fixed bin width the 'dead-space' in one or both left- / right-most bins is expected in the majority of the cases, which can vary from a tiny fraction of to almost entire width of a bin. With fixed number of bins the bin size could be defined as $S = \frac{\mathbf{max}(X) - \mathbf{min}(X)}{N}$, when the histrogram values range equals to the sample values range. However, due to rounding errors few extreme values close

to the **max**(X) may be lost in some cases. Thus, bin width selection as $S = \frac{\mathbf{max}(X) - \mathbf{min}(X)}{N-1}$ eliminates the possible data points loss and results in the 'average' expected size of the 'dead-space' with a fixed arbirary bin width.

Calculation of fractional ranks and Spearman rank correlation

Consider a sample **X** of length *N*. The same data sorted in ascending order is sequence **S** such that $x_i \to s_j ; : ; s_m \leq s_k ; \forall ; m < k$. Each element of the sorted sequence **S** has index (or *natural rank*) from 1 to N. If the original sample sequence **X** contains *repeatitions* (i.e. the same value apears more the once), all appearances of repetitive value from the original sequence will be placed one after another in the sorted sequence **S**.

Consider such a group of the repeating value in the sorted sequence, which starts at the position *K* + 1 (i.e. $K \geq 0$ elements are to the left from this group in the sorted sequence) and has length of *M* > 1. In other words, it includes elements with the *natural ranks* from K + 1 to K + M. The *fractional rank* is calculated as the arithmetic mean of the *natural ranks*, i.e.

$$r = \frac{(K+1) + (K+2) + \cdots + (K+M)}{M} = K + \frac{M+1}{2}$$

Each element in this group receives the *fractional rank r*, whereas all unique elements in the sorted sequence retain their *natural ranks*. Using this process $x_i \to s_j \to r_j$ each element of the initial data sequence receives its own rank. Arranging the calculated ranks in the same order as the corresponding elements of the original sequence **X** ones produces the *ranks* of the distribution **R**(**X**).

The Spearman rank correlation coefficient ρ is calculated as the Pearson's correlation coefficient *r* of the *ranks* of the initial **X** and **Y** data sets. In short:

$$\rho(X, Y) = r(R(X), R(Y))$$

Calculation of Kendall rank correlation

The Kendall rank correlation coefficient τ-*b* is calculated from the analisis of the relation between pairs of 2-D data poinst. Considering the 2-D data sample (**X**, **Y**) length *N* the pairs are constructed as $(x_i, y_i) | (x_j, y_j) ; \forall ; i \neq j$. In total there are $\frac{N \times (N-1)}{2} = N_C + N_D + N_X + N_Y + N_{XY}$ pairs, where N_C is the number of *conciding* pairs, N_D is the number of *discording* pairs, N_X is the number of X-tied pairs, N_Y is the number of Y-tied pairs, and N_{XY} is the number of X-Y tied pairs.

A pair is:

- Concording if
  - $x_i > x_j$ AND $y_i > y_j$; OR
  - $x_i < x_j$ AND $y_i < y_j$
- Discording if
  - $x_i > x_j$ AND $y_i < y_j$; OR
  - $x_i < x_j$ AND $y_i > y_j$
- X-tied if $x_i = x_j$ AND $y_i \neq y_j$
- Y-tied if $x_i \neq x_j$ AND $y_i = y_j$
- X-Y tied if $x_i = x_j$ AND $y_i = y_j$

Then, the rank correlation coefficient is calculated as

$$-1 \leq \tau_b = \frac{N_C - N_D}{\sqrt{(N_C + N_D + N_X) \times (N_C + N_D + N_Y)}} \leq 1$$