Descriptive statistics

Descriptive statistics task can be invoked on matrix data node e.g. Gene Counts, Normalized Counts data node in bulk RNA seq analysis pipeline or Single Cell counts Data node etc. It calculates measures of central tendency and variability on observations or features of the matrix data.

Running Descriptive statistics

- Click on a counts data node
- Choose **Descriptive Statistics** in *Pre-analysis tools* section of the toolbox (Figure 1)

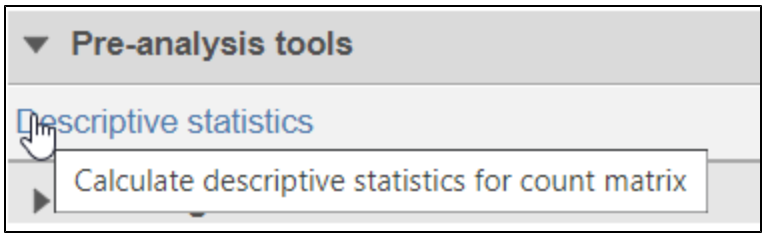

Figure 1. Descriptive statistics menu

This will invoke the dialog configuration dialog; use it to specify which calculation(s) will be performed on cells (or samples for a bulk analysis data node) or features (Figure 2).

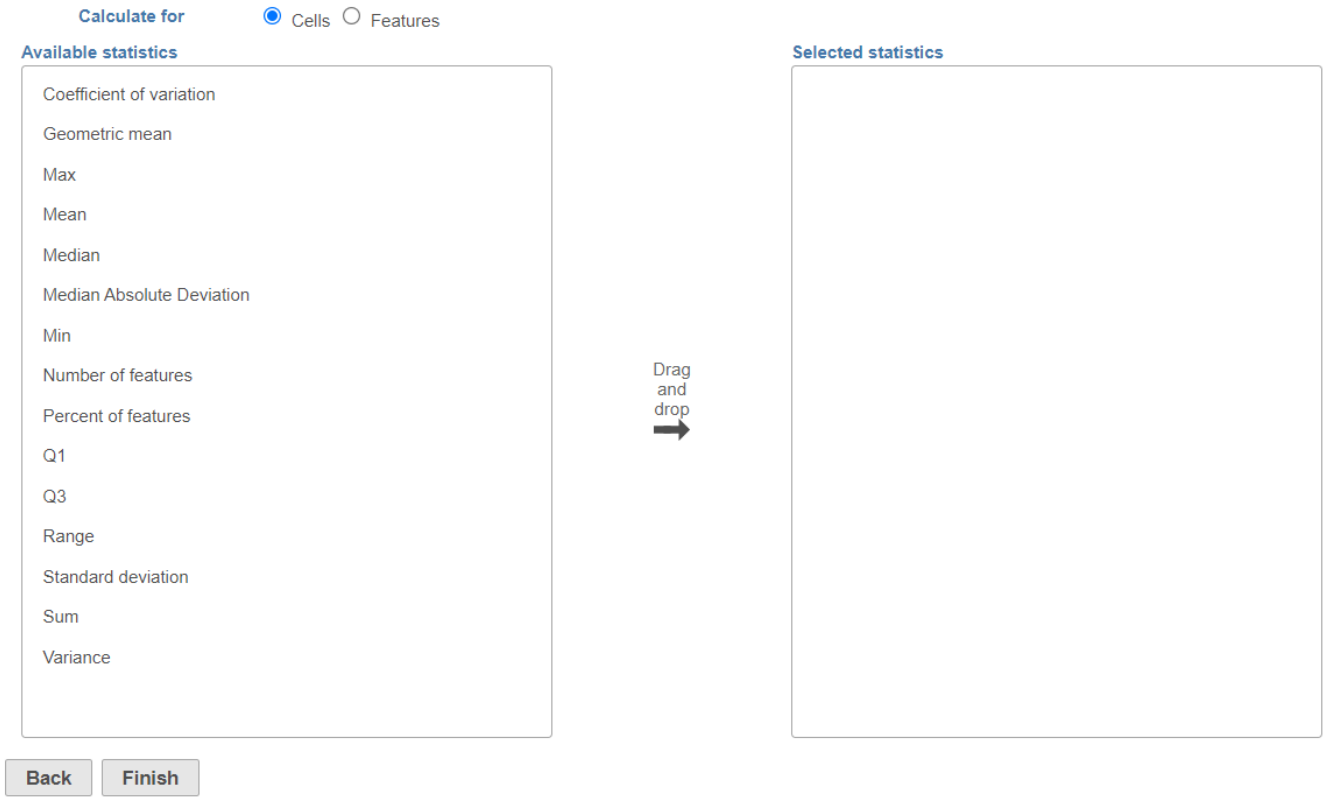

Figure 2. Select to calculate descriptive statistics on samples/cells or features

The available statistics are listed on the left panel, suppose "x_1, x_2, ..., x_n"represent an array of numbers

- Coefficient of variation (CV): $cv = \frac{s}{x}$ s represent the standard deviation

- Geometric mean: g=$\sqrt[n]{x_1 \times \cdots \times x_n}$
- Max: $x_{\min} = \max_i (x_i)$
- Mean: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
- Median: when n is odd, median is $x_{\frac{n+1}{2}}$, when n is even, median is $\frac{1}{2}x_{\frac{n}{2}} + \frac{1}{2}x_{\frac{n}{2}+1}$
- Median absolute deviation: $\mathrm{MAD} = \mathrm{median}(|X_i - \tilde{X}|)$, where $\tilde{X} = \mathrm{median}(X)$
- Min: $x_{\min} = \min_i (x_i)$
- Number of cells: Available when *Calculate for* is set to *Features*. Reports the number of cells with the value [<, <=, =, !=, > >=] (select one from the drop down list) than the cut off value entered in the text box. The cut off will be applied to the values present in the input data node, i.e. if invoked on non-normalised data node, the values are raw counts. For instance, use this option if you want to know the number of cells in which each feature was detected; possible filter: *Number of cells whose value > 0.0*
- Percent of cells: Available when *Calculate for* is set to *Features*. Reports the number of cells with the value [<, <=, =, !=, > >=] (select one from the drop down list) than the cut off value entered in the text box.
- Number of features: Available when *Calculate for* is set to *Cells*. Reports the number of features with the value [<, <=, =, !=, > >=] (select one from the drop down list) than the cut off value entered in the text box. The cut off will be applied to the values present in the input data node, i.e. if invoked on non-normalised data node, the values are raw counts. For example, use this option if you want to know the number of detected genes per each cell; filter: *Number of features whose value > 0.0*
- Percent of features: Available when *Calculate for* is set to *Cells*. Reports the fraction of features with the value [<, <=, =, !=, > >=] (select one from the drop down list) than the cut off value entered in the text box.
- Q1: 25th percentile
- Q3: 75th percentile
- Range: x_{max} - x_{min}
- Standard deviation: $s = \sqrt{s^2}$ where $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$
- Sum: $\sum_{i=1}^{n} x_i$
- Variance: $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$

Left click to select measurement and drag to move to the right panel one at a time, or when you mouse over on a measurement, click on the **green plus** button to move to the right panel. When *Sample (Cell)* is select, the calculation will be performed on all the features in the input matrix for each sample (or cell). When *Feature* is selected, the calculation will be performed across all the samples (cells) in the input matrix for each feature.

In addition, when *Feature* is selected, there is an extra *Group by* option (Figure 3)

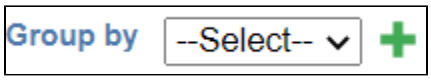

Figure 3. Choose a categorical attribute to calculate the statistics on each subgroup

From the drop-down list, choose a categorical attribute to calculate the descriptive statistics on all the subgroups for each feature.

The output of the task is a matrix: *Cell stats* (result of *Calculate for Cells*) or *Feature stats* (result of *Calculate for Features*) (Figure 4). The results can be visualized in the *Data Viewer*.

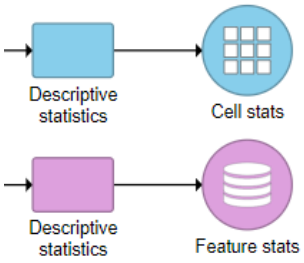

Figure 4. Descriptive statistics task produces either a Cell stats (calculation per cell) or Feature stats (calculation per feature) data node

Additional Assistance

If you need additional assistance, please visit our support page to submit a help ticket or find phone numbers for regional support.