NAME

AnalyzeSDFilesData.pl - Analyze numerical data field values in SDFile(s)

SYNOPSIS

AnalyzeSDFilesData.pl SDFile(s)...

AnalyzeSDFilesData.pl [--datafields "fieldlabel,[fieldlabel,...]" | All] [--datafieldpairs "fieldlabel,fieldlabel,[fieldlabel,fieldlabel,...]" | AllPairs] [-d, --detail infolevel] [-f, --fast] [--frequencybins number | "number,number,[number,...]"] [-h, --help] [--klargest number] [--ksmallest number] [-m, --mode DescriptiveStatisticsBasic | DescriptiveStatisticsAll | All | "function1, [function2,...]"] [--trimfraction number] [-w, --workingdir dirname] SDFiles(s)...

DESCRIPTION

Analyze numerical data field values in *SDFile(s)* using a combination of various statistical functions; Non-numerical values are simply ignored. For *Correlation, RSquare, and Covariance* analysis, the count of valid values in specified data field pairs must be same; otherwise, column data field pair is ignored. The file names are separated by space.The valid file extensions are *.sdf* and *.sd*. All other file names are ignored. All the SD files in a current directory can be specified either by **.sdf* or the current directory name.

OPTIONS

--datafields *"fieldlabel,[fieldlabel,...]"* | *Common* | *All*

Data fields to use for analysis. Possible values: list of comma separated data field labels, data fields common to all records, or all data fields. Default value: *Common*. Examples:

```
ALogP,MolWeight,EC50
"MolWeight,PSA"
```

--datafieldpairs *"fieldlabel,fieldlabel,[fieldlabel,fieldlabel,...]"* | *CommonPairs* | *AllPairs*

This value is mode specific and is only used for *Correlation, PearsonCorrelation, or Covariance* value of -m, --mode option. It specifies data field label pairs to use for data analysis during *Correlation* and *Covariance* calculations. Possible values: comma delimited list of data field label pairs, data field label pairs common to all records, or all data field pairs. Default value:*CommonPairs*. Example:

```
MolWeight,EC50,NumN+O,PSA
```

For *AllPairs* value of --datafieldpairs option, all data field label pairs are used for *Correlation* and *Covariance* calculations.

-d, --detail *infolevel*

Level of information to print about column values being ignored. Default: *0*. Possible values: 0, 1, 2, 3, or 4.

-f, --fast

In this mode, all the data field values specified for analysis are assumed to contain numerical data and no checking is performed before analysis. By default, only numerical data is used for analysis; other types of column data is ignored.

--frequencybins *number* | *"number,number,[number,...]"*

Specify number of bins or bin range to use for frequency analysis. Default value: *10*

Number of bins value along with the smallest and largest value for a column is used to group the column values into different groups.

The bin range list is used to group values for a column into different groups; It must contain values in ascending order. Examples:

```
10,20,30
0.1,0.2,0.3,0.4,0.5
```

The frequency value calculated for a specific bin corresponds to all the column values which are greater than the previous bin value and less than or equal to the current bin value.

-h, --help

Print this help message.

--klargest *number*

Kth largest value to find by *KLargest* function. Default value: *2*. Valid values: positive integers.

--ksmallest *number*

Kth smallest value to find by *KSmallest* function. Default values: *2*. Valid values: positive integers.

-m, --mode *DescriptiveStatisticsBasic* | *DescriptiveStatisticsAll* | *All* | *"function1, [function2,...]"*

Specify how to analyze data in SDFile(s): calculate basic or all descriptive statistics; or use a comma delimited list of supported statistical functions. Possible values: *DescriptiveStatisticsBasic* | *DescriptiveStatisticsAll* | *"function1,[function2]..."*. Default value: *DescriptiveStatisticsBasic*

DescriptiveStatisticsBasic includes these functions: *Count, Maximum, Minimum, Mean, Median, Sum, StandardDeviation, StandardError, Variance*.

DescriptiveStatisticsAll, in addition to *DescriptiveStatisticsBasic* functions, includes: *GeometricMean, Frequency, HarmonicMean, KLargest, KSmallest, Kurtosis, Mode, RSquare, Skewness, TrimMean*.

All uses complete list of supported functions: *Average, AverageDeviation, Correlation, Count, Covariance, GeometricMean, Frequency, HarmonicMean, KLargest, KSmallest, Kurtosis, Maximum, Minimum, Mean, Median, Mode, RSquare, Skewness, Sum, SumOfSquares, StandardDeviation, StandardDeviationN, StandardError, StandardScores, StandardScoresN, TrimMean, Variance, VarianceN*. The function names ending with N calculate corresponding values assuming an entire population instead of a population sample. Here are the formulas for these functions:

Average: See Mean

AverageDeviation: SUM(ABS(x[i] - Xmean)) / n

Correlation: See Pearson Correlation

Covariance: SUM((x[i] - Xmean)(y[i] - Ymean)) / n

GeometricMean: NthROOT(PRODUCT(x[i]))

HarmonicMean: 1 / (SUM(1/x[i]) / n)

Mean: SUM(x[i]) / n

Median: Xsorted[(n - 1)/2 + 1] for even values of n; (Xsorted[n/2] + Xsorted[n/2 + 1])/2 for odd values of n.

Kurtosis: [{n(n + 1)/(n - 1)(n - 2)(n - 3)} SUM{ ((x[i] - Xmean)/STDDEV)^4 }] - {3((n - 1)^2)}/{(n - 2)(n-3)}

PearsonCorrelation: SUM((x[i] - Xmean)(y[i] - Ymean)) / SQRT(SUM((x[i] - Xmean)^2) (SUM((y[i] - Ymean)^2)))

RSquare: PearsonCorrelation^2

Skewness: {n/(n - 1)(n - 2)} SUM{ ((x[i] - Xmean)/STDDEV)^3 }

StandardDeviation: SQRT (SUM((x[i] - Mean)^2) / (n - 1))

StandardDeviationN: SQRT (SUM((x[i] - Mean)^2) / n)

StandardError: StandardDeviation / SQRT(n)

StandardScore: (x[i] - Mean) / (n - 1)

StandardScoreN: (x[i] - Mean) / n

Variance: SUM((x[i] - Xmean)^2 / (n - 1))

VarianceN: SUM((x[i] - Xmean)^2 / n)

-o, --overwrite

Overwrite existing files.

--outdelim *comma* | *tab* | *semicolon*

Output text file delimiter. Possible values: *comma, tab, or semicolon* Default value: *comma*.

-p, --precision *number*

Precision of calculated values in the output file. Default: up to *2* decimal places. Valid values: positive integers.

-q, --quote *yes* | *no*

Put quotes around column values in output text file. Possible values: *yes or no*. Default value: *yes*.

-r, --root *rootname*

New text file name is generated using the root: <Root>.<Ext>. Default new file name: <InitialSDFileName><Mode>.<Ext>. Based on the specified analysis, <Mode> corresponds to one of these values: DescriptiveStatisticsBasic, DescriptiveStatisticsAll, AllStatistics, SpecifiedStatistics, Covariance, Correlation, Frequency, or StandardScores. The csv, and tsv <Ext> values are used for

comma/semicolon, and tab delimited text files respectively. This option is ignored for multiple input files.

--trimfraction *number*

Fraction of data to exclude from the top and bottom of the data set during *TrimMean* calculation. Default value: *0.1* Valid values: > 0 and < 1.

-w --workingdir *text*

Location of working directory. Default: current directory.

EXAMPLES

To calculate basic statistics for data in all common data fields and generate a NewSample1DescriptiveStatisticsBasic.csv file, type:

```
% AnalyzeSDFilesData.pl -o -r NewSample1 Sample1.sdf
```

To calculate basic statistics for MolWeight data field and generate a NewSample1DescriptiveStatisticsBasic.csv file, type:

```
% AnalyzeSDFilesData.pl --datafields MolWeight -o -r NewSample1
Sample1.sdf
```

To calculate all available statistics for MolWeight data field and all data field pairs, and generate NewSample1DescriptiveStatisticsAll.csv, NewSample1CorrelationMatrix.csv, NewSample1CorrelationMatrix.csv, and NewSample1MolWeightFrequencyAnalysis.csv files, type:

```
% AnalyzeSDFilesData.pl -m DescriptiveStatisticsAll --datafields
MolWeight -o --datafieldpairs AllPairs -r NewSample1 Sample1.sdf
```

To compute frequency distribution of MolWeight data field into five bins and generate NewSample1MolWeightFrequencyAnalysis.csv, type:

```
% AnalyzeSDFilesData.pl -m Frequency --frequencybins 5 --datafields
MolWeight -o -r NewSample1 Sample1.sdf
```

To compute frequency distribution of data in MolWeight data field into specified bin range values, and generate NewSample1MolWeightFrequencyAnalysis.csv, type:

```
% AnalyzeSDFilesData.pl -m Frequency --frequencybins "100,200,400"
--datafields MolWeight -o -r NewSample1 Sample1.sdf
```

To calculate all available statistics for data in all data fields and pairs, type:

```
% AnalyzeSDFilesData.pl -m All --datafields  All --datafieldpairs
AllPairs -o -r NewSample1 Sample1.sdf
```

AUTHOR

Manish Sud <msud@san.rr.com>

SEE ALSO

FilterSDFiles.pl, InfoSDFiles.pl, SplitSDFiles.pl, MergeTextFilesWithSD.pl

COPYRIGHT