**R**语言统计软件包接口说明文档

一．总体描述

1. 本次开发的R语言统计软件包包括：统计描述，t检验，方差分析等不少于18项常用的统计分析方法，每一个统计分析方法均保存在以.R结尾的脚本。
2. 每一个脚本均包含一个或多个函数，其中只有一个为主函数(名称同脚本名称一致)供python调用，其它函数皆被该主函数调用；脚本与脚本之间不存在跨脚本调用函数。
3. 由于统计分析经常返回多个结果，因此所有结果均以list形式返回。

二．数据结构与类型

1. 所有脚本主函数的第一个参数均为matrix或者dataframe格式的数据集（dataset）。如果该数据集只有一列，依然要求为dataframe格式。
2. 在python传入数据集实参时，对于数据集中的每一列，要事先设置好其数据类型，主要有整型（int），浮点型（double）, 字符串型（string），逻辑型（logic），时间日期型（date time）。其中整型和浮点型用一般常用的格式就行，不需区分长整型，和单精度浮点型等；字符串型设置比较重要，在R中根据需要会常常把字符串型转化为因子型（factor）。时间日期型数据稍微复杂一点，一般设置为“YYYY-MM-DD HH:MM:SS”或者“YYYY-MM-DD”。（这里我并不清楚rpy2中对日期时间数据的支持如何，一个替代的方法是用字符串格式代替）。
3. 主函数的结果均已list格式返回，list中只能包含数值，字符串和矩阵，其中用字符串分别记录矩阵的行名和列明。例如在ContingencyTableChisqTest 中，list(TableRowName = TableRowName, TablecolName = TablecolName, TableResults = TableResults,，ChiStatistic = ChiStatistic, PValue = PValue, Df = Df)
4. ErrorMsg代表返回错误。

三．具体函数说明

1. **DescriptiveAanlysis** **(dataset, rowname = NULL, colname = NULL)**

功能：对数据每一列做描述性统计

输入：

**dataset:** 数据集，类型为dataframe。

**rowname**：数据集每一行的名称,例如c(“1”, ”2”, “3”)，类型为字符串型，一般可缺少。

**colname**：数据集每一行的名称,例如c(“name”, ”age”, “sex”)，类型为字符串型。

如果**dataset**没有列名，则不可**colname**缺少。

输出：

**DespResultRowName:** 描述性统计结果矩阵的行名，类型为字符串向量。

**DespResultColName**：描述性统计结果矩阵的行名，类型为字符串向量。

**DespResult**：描述性统计结果矩阵，类型为字符串型矩阵。

1. **TTest**检验（文件夹下包含三个文件）
   1. **SingleTTest<-function(dataset, rowname = NULL, colname = NULL, side = "twotail", mu = 0, confidence = 0.95, varequal = FALSE)**

功能：

单总体t检验；

。

输入：

**dataset:** 一列，数值型。

**rowname**：同1。

**colname**：同1。

**side:** 指定双尾检验还是单尾检验，"twotail"=双尾检验，"lefttail" =左侧单尾检验，"rghttail " =左侧单尾，类型为字符串，缺省为"TwoTail"。

**mu:** 原假设均值, 类型为数值型，缺省为0。

**confidence:** 置信度，类型为数值型，缺省为0.95。

**varequal**：否存在异方差，如果存在，用Welch (or Satterthwaite)方法估计方差，类型未逻辑型，缺省为FALSE。

输出：

**TStatistic**：t检验统计量，类型为数值型。

**Pvalue**： p值，类型为数值型。

**LCI**：置信区间下界，类型为数值型。

**UCI**：置信区间上界，类型为数值型。

* 1. **DoubleUnPairTTest<-function(dataset, rowname = NULL, colname = NULL,** **numvar = NULL, chavar = NULL, side = "twotail", mu = 0, confidence = 0.95,**

**varequal = FALSE)**

功能：

双总体独立样本t检验。

输入：

**dataset:** 两列，一列数值型向量，一列字符串型向量（2分类类别变量）

**rowname**：同1。

**colname**：同1。

**numvar**： **dataset**中数值型向量的列名，类型为字符串，不可缺省。

**chavar**： **dataset**中字符串型向量的列名，类型为字符串，不可缺省。

**side:** 指定双尾检验还是单尾检验，"twotail"=双尾检验，"lefttail" =左侧单尾检验，"rghttail " =左侧单尾，类型为字符串，缺省为"TwoTail"。

**mu:** 原假设均值, 类型为数值型，缺省为0。

**confidence:** 置信度，类型为数值型，缺省为0.95。

**varequal**：是否存在异方差，如果存在，用Welch (or Satterthwaite)方法估计方差，类型未逻辑型，缺省为FALSE。

输出：

**TStatistic**：t检验统计量，类型为数值型。

**Pvalue**： p值，类型为数值型。

**LCI**：置信区间下界，类型为数值型。

**UCI**：置信区间上界，类型为数值型。

* 1. **DoublePairTTest<-function(dataset, rowname = NULL, colname = NULL, numvar1 = NULL, numvar2 = NULL, side = "twotail", mu = 0, confidence = 0.95, varequal = FALSE))**

功能：

双总体成对样本t检验。

输入：

**dataset:** 两列，均为数值型向量。

**rowname**：同1。

**colname**：同1。

**numvar1**： **dataset**中第一列数值型向量的列名，类型为字符串，不可缺省。

**numvar2**： **dataset**中第二列数值型向量的列名，类型为字符串，不可缺省。

**side:** 指定双尾检验还是单尾检验，"twotail"=双尾检验，"lefttail" =左侧单尾检验，"rghttail " =左侧单尾，类型为字符串，缺省为"TwoTail"，

**mu:** 原假设均值, 类型为数值型，缺省为0。

**confidence:** 置信度，类型为数值型，缺省为0.95。

**varequal**：是否存在异方差，如果存在，用Welch (or Satterthwaite)方法估计方差，类型未逻辑型，缺省为FALSE。

输出：

**TStatistic**：t检验统计量，类型为数值型。

**Pvalue**： p值，类型为数值型。

**LCI**：置信区间下界，类型为数值型。

**UCI**：置信区间上界，类型为数值型。

**3. WilcoxonTest**（文件夹下包含三个文件）

3.1 **SingleWilcoxTest<-function(dataset, rowname = NULL, colname = NULL, side = "twotail", mu = 0, confidence = 0.95)**

功能：

单总体样本Wilcox 符号秩检验；

。

输入：

**dataset:** 一列，数值型。

**rowname**：同1。

**colname**：同1。

**side:** 指定双尾检验还是单尾检验，"twotail"=双尾检验，"lefttail" =左侧单尾检验，"rghttail " =左侧单尾，类型为字符串，缺省为"TwoTail"。

**mu:** 原假设中总体分布的中心位置, 类型为数值型，缺省为0。

**confidence:** 置信度，类型为数值型，缺省为0.95。

输出：

**W****Statistic**：W检验统计量，类型为数值型。

**Median**：估计的分布中位数

**Pvalue**： p值，类型为数值型。

**LCI**：估计的分布中位数的置信区间下界，类型为数值型。

**UCI**：估计的分布中位数的置信区间上界，类型为数值型。

3.2 **DoubleUnPairWilcoxTest<-function(dataset, rowname = NULL, colname = NULL, numvar = NULL, chavar = NULL, side = "twotail", mu = 0, confidence = 0.95)**

功能：

双总体独立样本Wilcox 符号秩检验。

输入：

**dataset:** 两列，一列数值型向量，一列字符串型向量（2分类类别变量）

**rowname**：同1。

**colname**：同1。

**numvar**： **dataset**中数值型向量的列名，类型为字符串，不可缺省。

**chavar**： **dataset**中字符串型向量的列名，类型为字符串，不可缺省。

**side:** 指定双尾检验还是单尾检验，"twotail"=双尾检验，"lefttail" =左侧单尾检验，"rghttail " =左侧单尾，类型为字符串，缺省为"TwoTail"。

**mu:** 原假设中两总体分布差的中心位置, 类型为数值型，缺省为0。

**confidence:** 置信度，类型为数值型，缺省为0.95。

输出：

**W****Statistic**：W检验统计量，类型为数值型。

**Median**：估计的两总体分布差的中位数

**Pvalue**： p值，类型为数值型。

**LCI**：估计的分布中位数的置信区间下界，类型为数值型。

**UCI**：估计的分布中位数的置信区间上界，类型为数值型。

4. **ChisqTest**（文件夹下包含两个文件）

**4****.1 GoodnessOfFitChisqTest<-function(dataset, rowname = NULL, colname = NULL, theoryp = NULL, simulationp = FALSE)**

功能：

拟合优度检验，检验数据是否服从某种分布（拟合优度检验）。

输入：

**dataset:** 一列，整数型。

**rowname**：同1。

**colname**：同1。

**theoryp**：**dataset**对应事件的发生概率，长度和**dataset**一致，和为1，类型

为数值型，缺省为概率相等， 。

**simulationp**：是否使用 Monte Carlo simulation计算p值，类型为逻辑型，缺

省为FALSE。

输出：

**ChiStatistic**：卡方检验统计量，类型为数值型。

**PValue**：p值，类型为数值型。

**Df**：自由度，类型为数值型。

**4.2 ContingencyTableChisqTest<-function(dataset, rowname = NULL, colname = NULL, chavar1 = NULL, chavar2 = NULL, simulationp = FALSE)**

功能：

列联表检验。

输入：

**dataset:** 两列，均为字符串型向量。

**rowname**：同1。

**colname**：同1。

**chavar1**：指定**dataset**中的列名作为列联表的行，类型为字符串，不可缺省。

**chavar2**：指定**dataset**中的列名作为列联表的列，类型为字符串，不可缺省。

**simulationp**：是否使用 Monte Carlo simulation计算p值，类型为逻辑型，缺省为FALSE。

输出：

**TableRowName**：列联表矩阵的行名，类型为字符串向量。

**TablecolName**：列联表矩阵的咧名，类型为字符串向量。

**TableResults**：列联表矩阵，类型为字符串型矩阵

**ChiStatistic**：卡方检验统计量，类型为数值型。

**PValue**：p值，类型为数值型。

**Df**：自由度，类型为数值型。

**5. LinearRegression**（文件夹下包含一个文件）

**5.1 LinearRegression<-function(dataset, rowname = NULL, colname = NULL, yname = NULL, xname = NULL, formulastring = NULL, intercept = TRUE,plotstr = NULL)**

功能：

多元线性回归。

输入：

**dataset:**。两列及以上，**yname**指定的列必须为数值型向量；**xname**指定的列可以为数值型或者字符串型向量。

**rowname**：同1。

**colname**：同1。

**yname**：因变量名称，类型为字符串，不可缺省。

**xname**：自变量名称，类型为字符串向量，不可缺省。

**formulastring:** 回归模型方程，类型为字符串，缺省为“**yname~xname**”。

**intercept**：是否有截距，类型为逻辑型，缺省为**TRUE**，有截距。

**plotstr**：图片输出文件夹目录，类型为字符串，不可缺省。

**resifitname** ：残差-拟合值（Residuals vs Fitted）图的输出名称，类型为字符串，缺省不输出。

**normalqqname** ：QQPlot图的输出名称，类型为字符串，缺省不输出。

**scallocname** ：扩展定位图（Spread-Location plot）的输出名称，类型为字符串，缺省不输出。

**resilevname**：残差杠杆值图（Residuals vs Leverage）的输出名称，类型为字符串，缺省不输出。

输出：

**RegResultRowName**：回归结果矩阵的行名，类型为字符串向量。

**RegResultColName**：回归结果矩阵的列名，类型为字符串向量。

**RegResult**：回归结果矩阵，包括估计系数(Estimate)，估计系数的标准差(Std. Error)， 估计系数检验的t统计量(t value)，估计系数检验的P值(Pr(>|t|)和显著性水平(significance)。

**Rsquare**：回归决定系数，类型为数值型。

**AdjRsquare**：调整后的回归决定系数，类型为数值型。

**MSE:** 残差均方误差，类型为数值型。

**MSEDf:** 残差均方误差的自由度，类型为数值型。

**FStatistic**：回归F检验的统计量，类型为数值型。

**FDf1**：回归F检验的第一个自由度，类型为数值型。

**FDf2**：回归F检验的第二个自由度，类型为数值型。

**FPValue**：回归F检验的p值，类型为数值型。

**ROC.png**：**R**OC曲线图。

**6. LogisticRegression**（文件夹下包含三个文件）

**6.1 BinaryLogistic<-function(dataset, rowname = NULL, colname = NULL, yname=NULL, xname=NULL, formulastring=NULL, plotstr = NULL**，**plotname = NULL)**

功能：

二元反应变量Logistic回归，需要加载ROCR包。

输入：

**dataset:**。两列及以上，**yname**指定的列必须为符串型向量；**xname**指定的列可以为数值型或者字符串型向量。

**rowname**：同1。

**colname**：同1。

**yname**：因变量名称，类型为字符串，不可缺省。

**xname**：自变量名称，类型为字符串向量，不可缺省。

**formulastring:** 回归模型方程，类型为字符串，缺省为“**yname~xname**”。

**plotstr**：图片输出文件夹目录，类型为字符串，缺省不输出。

**plotname**：图片输出名称，类型为字符串，缺省不输出。

输出：

**RegResultRowName**：Logitstic回归结果矩阵的行名，类型为字符串向量。

**RegResultColName**：Logitstic回归结果矩阵的列名，类型为字符串向量。

**RegResult**：Logitstic回归结果矩阵，包括估计系数(Estimate)，估计系数的标准差(Std. Error)， 估计系数检验的Z统计量(z value)，估计系数检验的P值(Pr(>|z|)和显著性水平(significance)。

**AIC**：AIC得分，类型为数值型。

**ROC.png**：**R**OC曲线图。

**6.2 MultiNormLogistic<-function(dataset, rowname = NULL, colname = NULL, yname=NULL, xname=NULL, formulastring=NULL)**

功能：

多元无序元反应变量Logistic回归，需要加载mlogit包。

输入：

**dataset:**。两列及以上，**yname**指定的列必须为符串型向量；**xname**指定的列可以为数值型或者字符串型向量。

**rowname**：同1。

**colname**：同1。

**yname**：因变量名称，类型为字符串，不可缺省。

**xname**：自变量名称，类型为字符串向量，不可缺省。

**formulastring:** 回归模型方程，类型为字符串，缺省为“**yname~xname**”。

输出：

**RegResultRowName**：Logitstic回归结果矩阵的行名，类型为字符串向量。

**RegResultColName**：Logitstic回归结果矩阵的列名，类型为字符串向量。

**RegResult**：Logitstic回归结果矩阵，包括估计系数(Estimate)，估计系数的标准差(Std. Error)， 估计系数检验的t统计量(t value)，估计系数检验的P值(Pr(>|t|)和显著性水平(significance)。

**LogLikelyhood**：估计模型得到的对数极大似然估计值ç

**McFaddenR2**： 伪R方，McFaddenR2。

**LRatioRowName**：似然比检验结果矩阵的行名，类型为字符串向量。

**LRatioColName**：似然比检验结果矩阵的列名，类型为字符串向量。

**LRatioTest**：似然比检验结果矩阵，H0 = 模型为”yname~1”，H1 = 模型为输

入的模型，检验结果包括卡方统计量(ChiStatistic)，自由度 (Df)

和P值(PValue)。

**7. Cluster**（文件夹下包含四个文件）

**7.1 KMeansCluster<-function(dataset, rowname = NULL, colname = NULL,**

**culstervar = NULL, centers = 3, algorithm = "Hartign-Wong")**

功能：

K均值聚类。

输入：

**dataset:**。两列及以上，全部为数值型向量；字符串型向量会被自动转为数值型（1，2，3…）。

**rowname**：同1。

**colname**：同1。

**culstervar**：聚类变量名称，类型为字符串，缺省为dataset中所有列。

**centers:** 聚类的类别数目，类型为数值型，缺省为3。

**algorithm**：聚类算法名称，字符串型，取值范围为（"Hartigan-Wong",

"Lloyd", "Forgy", “MacQueen"），缺省为"Hartigan-Wong"。

输出：

**ClusterCenterRowName**：聚类中心矩阵的行名，类型为字符串向量

**ClusterCenterColName**：聚类中心矩阵的列名，类型为字符串向量。

**ClusterCenter**：聚类中心矩阵。

**ClusterLabelRowName**：聚类结果矩阵的行名，类型为字符串向量。

**ClusterLabelColName**：聚类结果矩阵的列名，类型为字符串向量。

**ClusterLabel**：聚类结果矩阵。

**7.2 KMedoidsCluster<-function(dataset, rowname = NULL, colname = NULL,**

**culstervar = NULL, centers=NULL, metric = "euclidean", stand = FALSE)**

功能：

K均值聚类。

输入：

**dataset:**。两列及以上，全部为数值型向量；字符串型向量会被自动转为数值型（1，2，3…）。

**rowname**：同1。

**colname**：同1。

**culstervar**：聚类变量名称，类型为字符串，缺省为dataset中所有列。

**centers:** 聚类的类别数目，类型为数值型，缺省为3。

**metric**：样本相似度的距离（测度），类型为字符串，取值范围为

（"euclidean", “manhattan”）缺省为"euclidean"。

**stand**：是否标准化数据，类型为逻辑型，缺省为Fasle。

输出：

**ClusterCenterRowName**：聚类中心矩阵的行名，类型为字符串向量

**ClusterCenterColName**：聚类中心矩阵的列名，类型为字符串向量。

**ClusterCenter**：聚类中心矩阵。

**ClusterLabelRowName**：聚类结果矩阵的行名，类型为字符串向量。

**ClusterLabelColName**：聚类结果矩阵的列名，类型为字符串向量。

**ClusterLabel**：聚类结果矩阵。

**7.3 AgglomerativeHierarchicalCluster<-function(dataset, rowname = NULL,**

**colname = NULL, culstervar = NULL, metric = "euclidean", method = "average",**

**stand = F**asle**, plotstr = NULL, bannername = NULL, treename = NULL)**

功能：

凝聚层次聚类。

输入：

**dataset:**。两列及以上，全部为数值型向量；字符串型向量会被自动转为数值型（1，2，3…）。

**rowname**：同1。

**colname**：同1。

**culstervar**：聚类变量名称，类型为字符串，缺省为dataset中所有列。

**metric**：样本相似度的距离（测度），类型为字符串，取值范围为

（"euclidean", “manhattan"）缺省为"euclidean"。

**method**：聚类算法，类型为字符串，取值范围为( "average","single",

"complete", "ward", "weighted", “”gaverage””)，缺省为 “average"。

**stand**：是否标准化数据，类型为逻辑型，缺省为Fasle。

**plotstr**：**图片输出文件夹目录，类型为字符串，不可缺省。**

**bannername**： **Banner**图输出名称，类型为字符串，缺省不输出。

**treename**： **Dendrogram**图输出名称，类型为字符串，缺省不输出。

输出：

**ClusterOrderRowName**：聚类中心矩阵的行名，类型为字符串向量

**ClusterOrderColName**：聚类中心矩阵的列名，类型为字符串向量。

**ClusterOrder**：聚类中心矩阵。

**ClusterHeightRowName**：聚类结果矩阵的行名，类型为字符串向量。

**ClusterHeightColName**：聚类结果矩阵的列名，类型为字符串向量。

**ClusterHeight**：聚类结果矩阵。

8. **Anova**（文件夹下包含两个文件）

**8.1 UinFactorAnova<-function(dataset, rowname = NULL, colname = NULL, yname**

**= NULL, xname = NULL, formulastring = NULL, plotstr = NULL, plotname =**

**NULL)**

功能：

单因素方差分析。需加载ggplot2包

输入：

**dataset:**。两列及以上，**yname**指定的列必须为数值型向量；**xname**指定的列必须为字符串型向量。

**rowname**：同1。

**colname**：同1。

**yname**：因变量名称，类型为字符串，不可缺省。

**xname**：自变量名称，类型为字符串，不可缺省。

**formulastring:** 回归模型方程，类型为字符串，缺省为“**yname~xname**”。

**plotstr**：图片输出文件夹目录，类型为字符串，不可缺省。

**plotname**：箱线图（box plot）的输出名称，类型为字符串，缺省不输出。

输出：

**AnoResultRowName**：单因素分析结果矩阵的行名，类型为字符串向量。

**AnoResultColName**：单因素分析结果矩阵的列名，类型为字符串向量。

**AnoResult**：单因素分析结果矩阵，包括估计系数(Estimate)。

**FStatistic**：单因素分析F检验的统计量，类型为数值型。

**FDf1**：单因素分析F检验的第一个自由度，类型为数值型。

**FDf2**：单因素分析F检验的第二个自由度，类型为数值型。

**FPValue**：单因素分析F检验的p值，类型为数值型。

**Box.png**：箱线图图。

**9. FactorAnalysis**（文件夹下包含一个文件）

**9.1 FactorAnalysis<-function(dataset, rowname = NULL, colname = NULL)**

功能：

只能因素分析，根据数据自动判断适合的因素分析工具。

输入：

**dataset:**。一列及以上。0-1的数值型列向量会自动转为类别变量。

**rowname**：同1。

**colname**：同1。

当**dataset**只有一列时：

如果为数值型，则使用单总体t检验。

如果为字符型，则使用拟合优度卡方检验。

当**dataset**只有两列时：

如果全部为数值型，使用双总体成对比较t检验。

如果全部为字符型，使用列联表卡方检验。

如果一列为数值型，一列为字符型，则使用双总体t检验

当**dataset**大于两列时，自动把最后一列作为因变量，其余为自变量：

如果因变量是数值型：

如果自变量全部为字符型，则使用多因素方差分析；

如果自变量不全部为字符型，则使用多元线性回归分析；

如果因变量是字符型：

如果因变量只有一个取值，则报错，；

如果自变量只有两个取值，则使用二元响应变量logistic回归分析；

如果自变量大于两个取值，使用多元无需响应变量logistic回归分析；

输出：

根据选择的模型返回相应的结果。

Method: 自动选择的模型名称。

四．附录

1. **Formula**

**formulastring** 是指R中用字符串表示的回归公式，如（“y ~ x1 + x2 + x1:x2”）。~左边的是因变量，右边的X1， x2 是自变量，其中x1:x2表示两者的交互效应，作为第三个自变量出现。自变量之间用+连接。

1. **Kernal Function in ‘’kernlab’’ Package**

括号内为参数名称，如sigma, 其取值需要由外部指定：

1）rbfdot(sigma = 1)： Gaussian RBF kerne

2）polydot(degree = 1, scale = 1, offset = 1)：Polynomial kernel

3）tanhdot(scale = 1, offset = 1)：Hyperbolic tangent kernel

4）vanilladot()：Linear kernel

5）laplacedot(sigma = 1)：Laplacian kernel

6）besseldot(sigma = 1, order = 1, degree = 1)：Bessel kernel

7）anovadot(sigma = 1, degree = 1)： ANOVA RBF kernel

8）splinedot()：Spline kernel