

1. SAS INSIGHT 启动:

方法1: Solution→Analysis→Interactive Date Analysis

方法2: 在命令栏内输入insight

方法3: 程序编辑窗口输入以下代码, 然后单击 Submit按钮:

Proc insight;

Run;

1.1 一维数据分析

用 sas insight做直方图、盒形图、马赛克图。

直方图: Analysis→Histogram/Bar Chart

盒形图: Analysis→Box plot

马赛克图: Analysis→Box plot/Mosaic plot (Y)

1.2 二维数据分析

散点图: Analysis→Scatter plot(Y X)

曲线图: Analysis→Line plot (Y X)

1.3 三维数据分析

旋转图: Analysis→Rotating Plot

曲面图: Analysis→Rotating Plot 设置 Fit Surface

等高线图: Analysis→Contour plot

1.4 分布分析

包括: 直方图、盒形图、各阶矩、分位数表, 直方图拟合密度曲线, 对特定分布进行检验。

1.4.1 Analysis→Distribution (Y)

第一部分为盒形图, 第二部分为直方图, 第三部分为各阶矩, 第四部分为分位数表。

1.4.2 添加密度估计

A: 参数估计: 给出各种已知分布(正态, 指数等), 只需要对其中参数进行估计;

Curves→Parametric Density

B: 核估计: 对密度函数没有做假设, 曲线性状完全依赖于数据;

Curves→Kernel Density

1.4.3 分布检验

Curves→CDF confidence band

Curves→Test for Distribution

1.5 曲线拟合

Analysis→Fit (Y X):分析两个变量之间的关系

1.6 多变量回归

Analysis→Fit (Y X)

1.7 方差分析

Analysis→Fit (Y X)

1.8 相关系数计算

Analysis→Multivariate

1.9 主成分分析

Analysis→Multivariate

2.SAS ANALYST启动:

方法1: Solution→Analysis→Analyst

方法2: 在命令栏内输入analyst

2.1 分类计算统计量: Data→Summarize by group

2.2 随机抽样: Data→Random Sample

2.3 生成报表: Report→Tables

2.4 变量计算: Date→Transform

2.5 绘制统计图

2.5.1 条形图: Graph→Bar Chart→Horizontal

2.5.2 饼图: Graph→Pie Chart

2.5.3 直方图: Graph→Histogram

2.5.4 概率图: Graph→Probability plot

2.5.5 散点图: Graph→Scatter plot

2.6 统计分析与计算

2.6.1 计算描述性统计量

Statistics →De s c r i p t i v e→Summart Statistics 只计算简单统计量

Statistics →De s c r i p t i v e→Distribution 可计算一个变量的分布信息

Statistics →De s c r i p t i v e→Correlations可计算变量之间的相关关系

Statistics →De s c r i p t i v e→Frequency counts 可计算频数

2.6.2 列联表分析

Statistics →Table Analysis

2.7假设检验

2.7.1单样本均值Z检验： 检验单样本均值与某个给定的数值之间的关系

Statistics →Hypothesis tests →One-Sample Z-test for a mean

2.7.2单样本均值t检验： 适用于不了解变量的方差情形

推断该样本来自的总体均数 μ 与已知的某一总体均数 μ_0 是否相等

Statistics →Hypothesis tests → One-Sample t-test for a mean

2.7.3单样本比例检验： 检验取离散值的变量取某个值的比例

Statistics →Hypothesis tests →One-Sample test for a proportion

2.7.4单样本方差检验： 检验样本方差是否等于给定的值。零假设方差等于某个给定的。

Statistics →Hypothesis tests→One-Sample test for a variance

2.7.5两样本均值t检验： 独立的两个总体的均值是否相等或者是否相差给定的值

Statistics →Hypothesis tests →Two-Sample t-test for means

2.7.6成对样本均值t检验： 成对样本检验中总体是相关的。

Statistics →Hypothesis tests →Two-Sample paired t-test for means

2.7.7两样本比例检验： 检验两个总体中某个比例的值是否相等。

Statistics →Hypothesis tests →Two-Sample test for proportions

2.7.8两样本方差检验

Statistics →Hypothesis tests→Two Sample test for variance

2.8ANOVA过程

2.8.1单因素ANOVA过程

Statistics →ANOVA→One-Way Anova

2.8.2非参数的单因素方差分析： 适用于正态分布假定或方差相等假设不能满足的单因素问题

Statistics →ANOVA→nonparameter one-way Anova test

Wilcoxon法、Median法、Van der Waerden法、Savage法。

2.8.2因素方差分析： 实验结果是连续数值而分类变量是两个以上的离散型数值。

Statistics →ANOVA→Factorial Anova

2.8.3线性模型： 用最小二乘法拟合一般线性模型

Statistics →ANOVA→Linear Model

2.9回归分析： Statistics →Regression

2.9.1simple回归： 简单一类回归分析，单一的自变量，单一的因变量，模型可以是一次、二次、三次。

Statistics →Regression→simple

2.9.2linear回归：线性回归，回归模型可以有多个因变量，多个自变量，但是对因变量分别进行回归

Statistics →Regression→linear

2.9.3logistic回归：用于解决因变量是一个二元变量

Statistics →Regression→logistic

3.报表以及图形输出

3.1 print过程

```
Proc print data = sasuser.score;    //数据库.数据集
```

```
Run;
```

```
Proc print data = sasuser.score;
```

```
Var name math Chinese;           //变量
```

```
Run;
```

```
Proc print data = sasuser.score noobs;    //去掉第一列（观测序号）
```

```
Var name math Chinese;
```

```
Run;
```

```
Proc print data= sasuser.score;
```

```
Where sex in ('f');                //通过where语句
```

```
Run;
```

```
Proc print data = sasuser.score noobs label;
```

```
Title '女生成绩单';
```

```
Label name = '姓名'
```

```
Sex = '性别'
```

```
Math = '数学'
```

```
Chinese = '语文'
```

English = ‘英语’;

Where sex in (‘f’) ;

Run;

Title “the sas system”; //恢复系统标题

Proc print data = sasuser.score;

Footnote = ‘分数列表’; //加分数列表的脚注

Run;

Proc sort data = sasuser.score;

By sex;

Run;

Proc print data = sasuser.score; //使用by分组输出前用sort排序

By sex;

Run;

Proc print data = sasuser.score;

Sum math;

Run;

3.2 tabulate过程

Proc tabulate data =数据集名称;

Class 分类变量;

Var 分析变量;

Table 页面说明 行维说明 列维说明/选项;

Run;

3.3 sort过程

Proc sort data = 数据集名称; //默认升序排列

By 变量名;

Run;

Proc sort data = 数据集名称;

By descending 变量名; //降序排列

Run;

3.4 means过程: 数量(N)、均值(Mean)、标准差(Std Dev)、最大值(Maximum)、最小值(Minimum)

Proc means data = sasuser.stock;

Var price;

Run;

3.5 univariate过程

Proc univariate data =数据集;

Var 分析变量;

Run;

结果:

Moments: 统计量的各阶矩, 例如一阶矩就是均值, 二阶矩就是方差等;

Basic Statistical Measures: 基本统计量;

Tests for location: 检验均值是否为零;

Quantiles: 分位数表;

Extreme Observations: 极端观测值。

3.6 freq过程: 离散变量的分布情况

Proc freq data =数据集名;

Tables 变量名;

Run;

结果:

变量取值、频数、百分比、累计频数、; 累计百分比

3.7 corr过程: 相关系数

Proc corr data =数据集;

Var 变量名 变量名;

Run;

结果:

简单统计量

相关系数及p值

3.8 gplot过程: 绘制散点图和曲线图,绘制回归曲线。

Proc gplot data = 数据集名称;

Symbol 曲线类型;

Plot 竖轴变量*横轴变量;

Run;

Proc gplot data = sasuser.score;

Symbol I = none v=star;

Plot English*Chinese;

Run;

3.9 gchart过程: 绘制直方图、饼图、三维直方图等。

Proc gchart data = 数据集名称;

Vbar/pie/block =变量;

Run;

3.10 G3D过程绘制三维曲面

Proc g3d data =数据集;

Plot 变量x*变量y=变量z;

Run;

gcontour过程: 画出曲面的等高线

Proc gcontour data =数据集名;

Plot x*y=z;

Run;

4.基本统计分析

4.1正态性检验: univariate过程

Proc univariate data= sasuser.stock normal;

Var eps;

Run;

Proc univariate data= sasuser.stock normal;

Var eps;

Histogram eps; //画出直方图

Probplot eps; //画出概率分布图

Run;

4.2单变量均值检验

4.2.1如果一个变量服从正态分布, 那么可以用t检验来对变量进行均值检验

Proc ttest data =数据集 ho = 均值;

Var 检验变量;

Run;

4.2.2t检验还可以检验方差相同的两个独立样本均值是否相等

Proc ttest data =数据集;

Class 分类变量;

Var 检验变量;

Run;

结果

第一部分简单统计量

第二部分t检验结果

第三部分两者方差是否相等检验

T检验要求两个独立样本都必须服从正态分布，如果不服从正态分布，则无法进行t检验。这时可用非参数的方法，常用的非参数方法是NPAR1WAY过程，它是 `noparameter 1 way` 缩写。

4.3成对总体均值检验

4.4回归分析：reg（回归）过程、rsreg（二次响应面回归）过程、orthoreg（病态数据回归）过程、nlin（非线性回归）过程、transreg（变换回归）过程、calis（线性结构方程和路径分析）过程、glm（一般线性回归）过程、genmod（广义线性回归）过程

4.4.1 REG过程

Proc reg data = 输入数据集 选项;

Var 变量列表;

Model 因变量 = 自变量列表;

Print 输出结果;

Plot 诊断图形;

Run;

4.4.2 nlin过程

指明模型的表达式并给定系数初值

4.4.3glm过程：使用最小二乘法回归线性模型，还可以进行回归，分差，协方差，多变量方差、偏相关系数分析

4.5方差分析

4.5.1单因素方差分析

Proc anova data =数据集名称;

Class 因素;

Model 实验结果 =因素;

Run;

Proc anova data =数据集名称;

Class 因素;

Model 实验结果 =因素;

Means brand;

Run;

Proc anova data =数据集名称;

Class 因素;

Model 实验结果 =因素;

Means brand/t; //t检验

Run;

Proc anova data =数据集名称;

Class 因素;

Model 实验结果 =因素;

Means brand/bon; //bonferroni t检验 控制第一类错误的概率，但是具有较大第二类错误概率

Run;

Proc anova data =数据集名称;

Class 因素;

Model 实验结果 =因素;

Means brand/regwq; //regwq检验 控制第一类错误的概率

Run;

Proc anova data =数据集名称;

Class 因素;

Model 实验结果 =因素;

Means brand/tukey; //tukey检验 控制第一类错误的概率，但是第二类错误概率通常高于regwq检验

Run;

4.5.2多因素方差分析

4.5.3列联表检验

Proc freq data = 数据集;

Tables 因素a*因素b / chisq;

Weight 实验结果;

Run;

5相关知识

因变量—Depender (Y)

自变量—Independent (X1 X2...)