统计软件期末复习

一、建立（逻辑）库

1. 建库的原因

A. **逻辑库**就是存放在同一**文件夹**中的一组 SAS 文件，每一个逻辑库对应了计算机中的一个文件夹；

B. SAS 数据文件均放在逻辑库中，使用 SAS 进行对数据操作分析前要指明它在哪一个逻辑库中；

C. SAS 的逻辑库分为临时库和永久库两种：临时库名为 Work，存放在 Work 中的 SAS 文件叫临时文件，这些临时文件当退出 SAS 系统时会被自动删除，永久库中的文件则不会被删除。

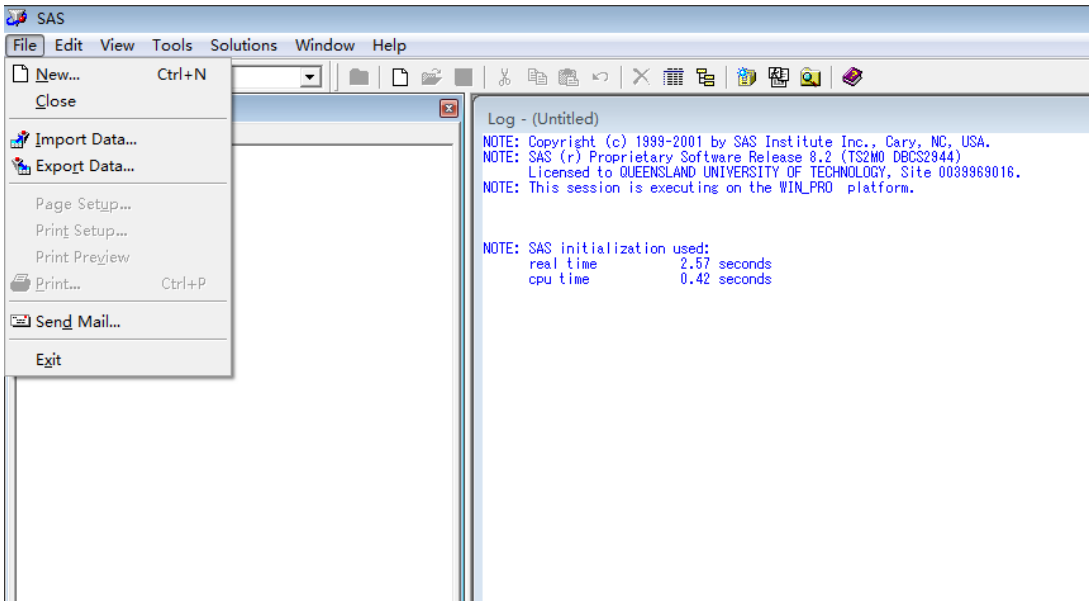
2. 建库操作

1) 编程操作

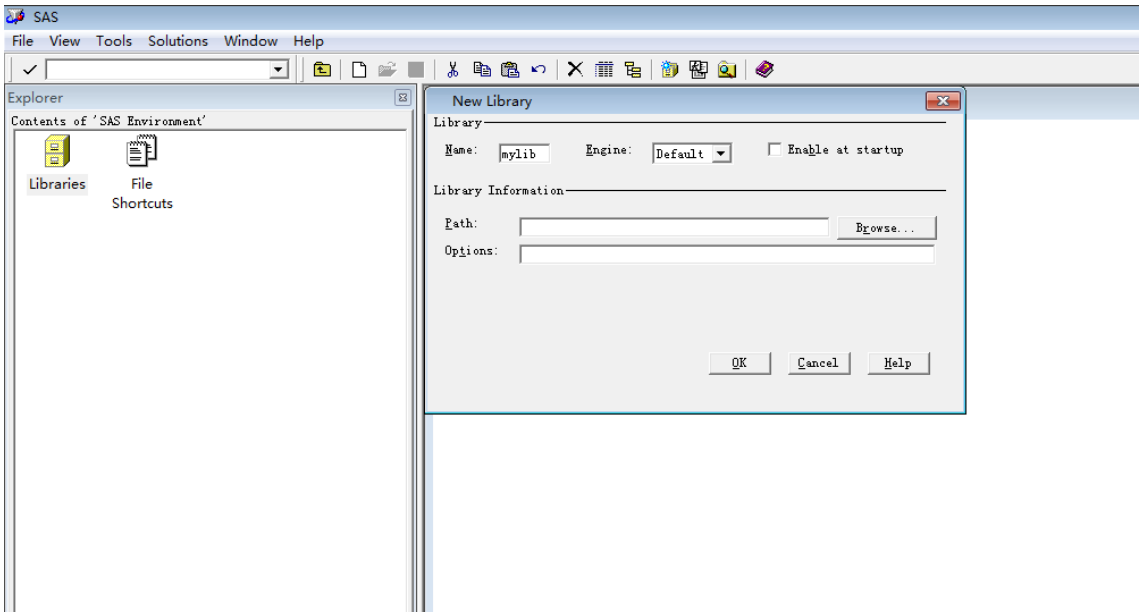
```
libname mysas 'E:\mysas';
```

2) 菜单操作

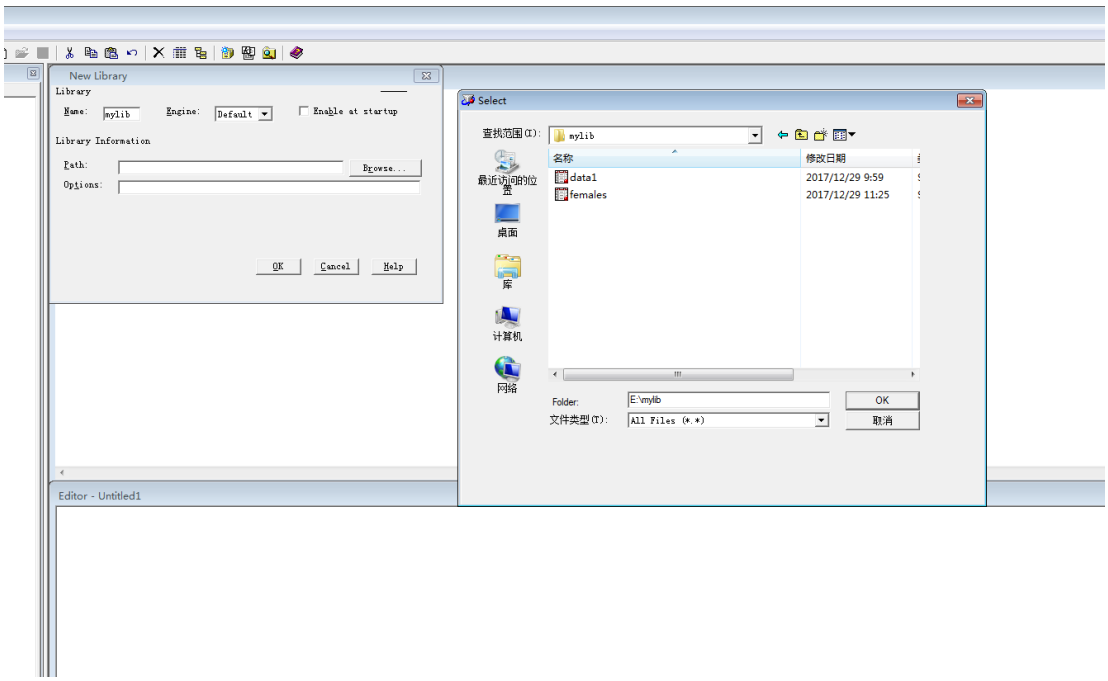
A. 选择 File 选项的 New 选项



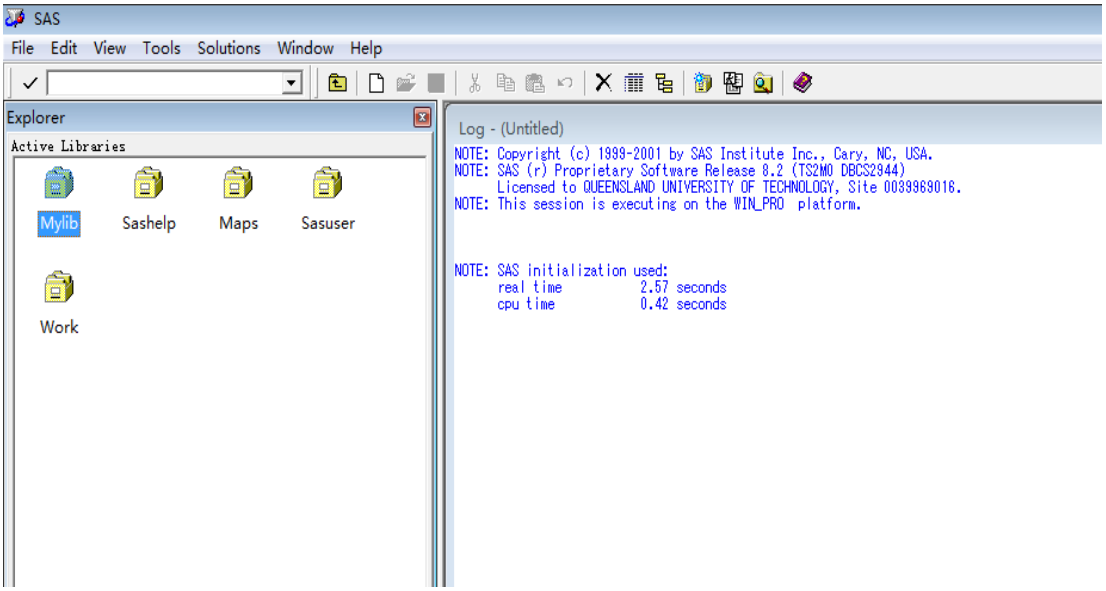
B. 输入逻辑库名，选择 brouse 键



C. 选择对应的文件夹

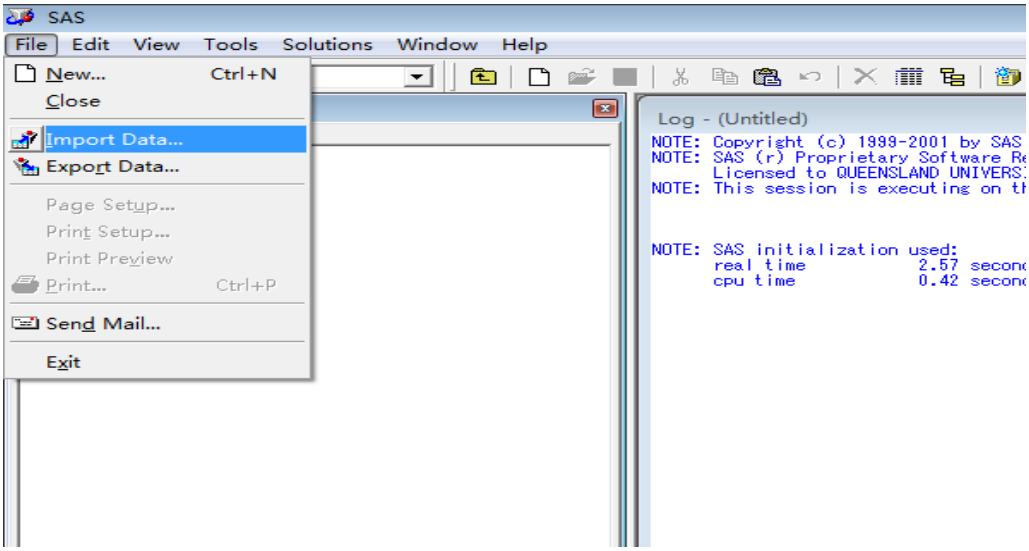


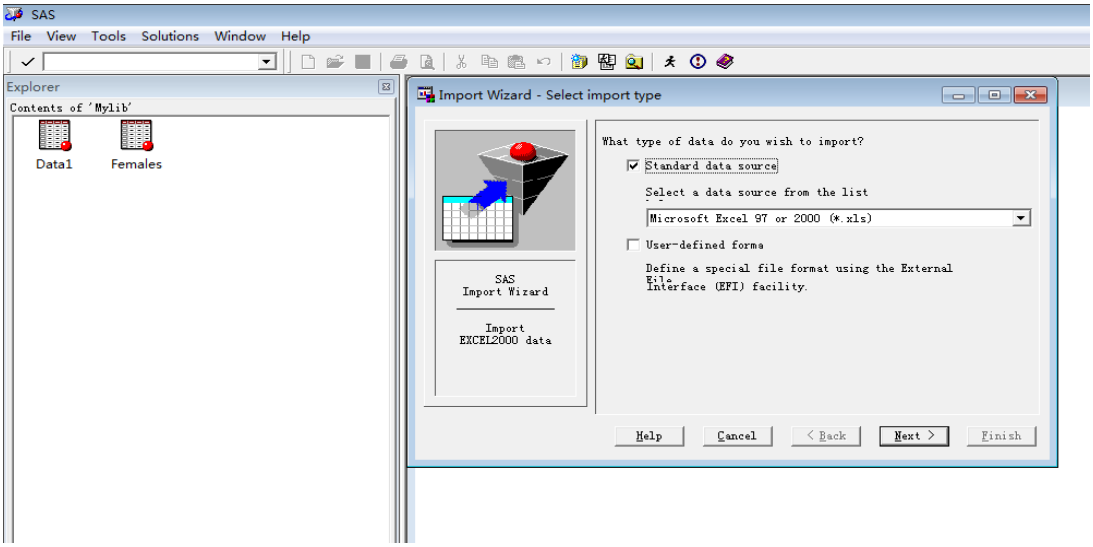
D. 点击 OK，完成建库

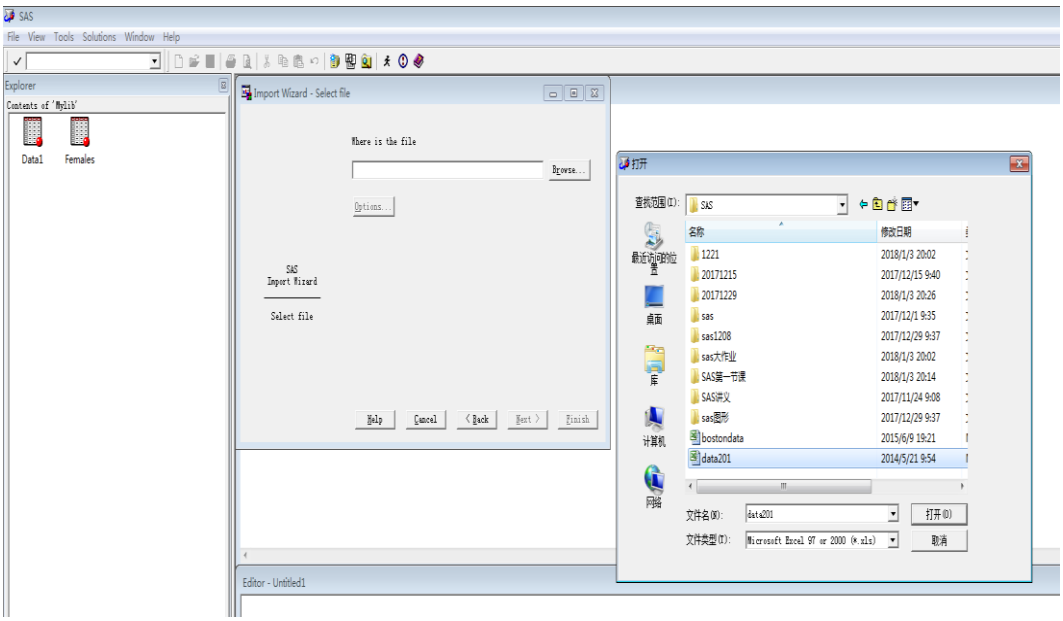


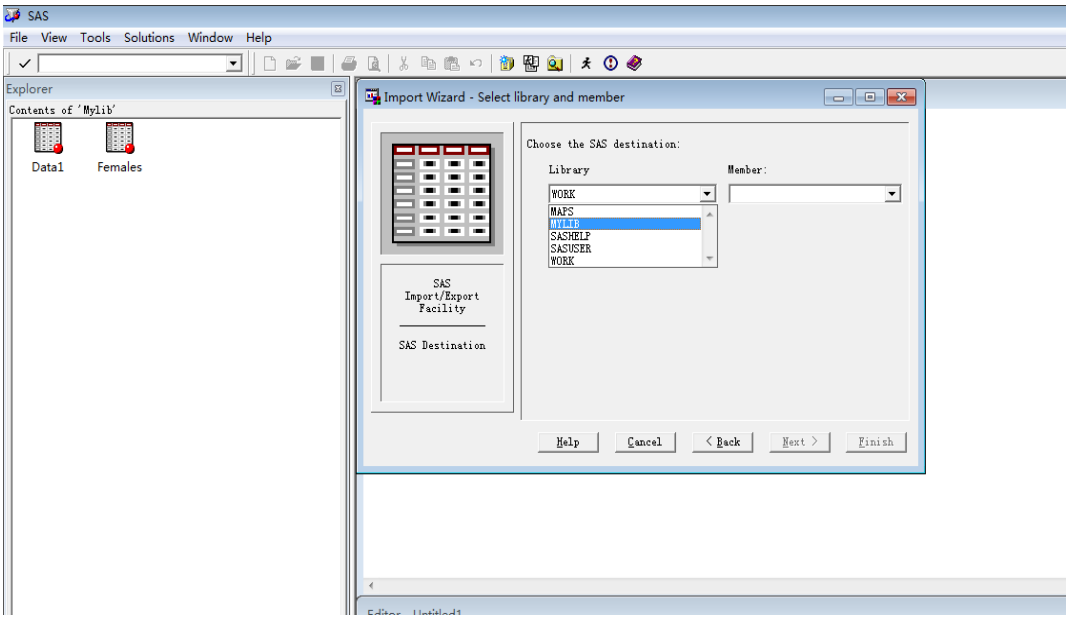
二、导入数据

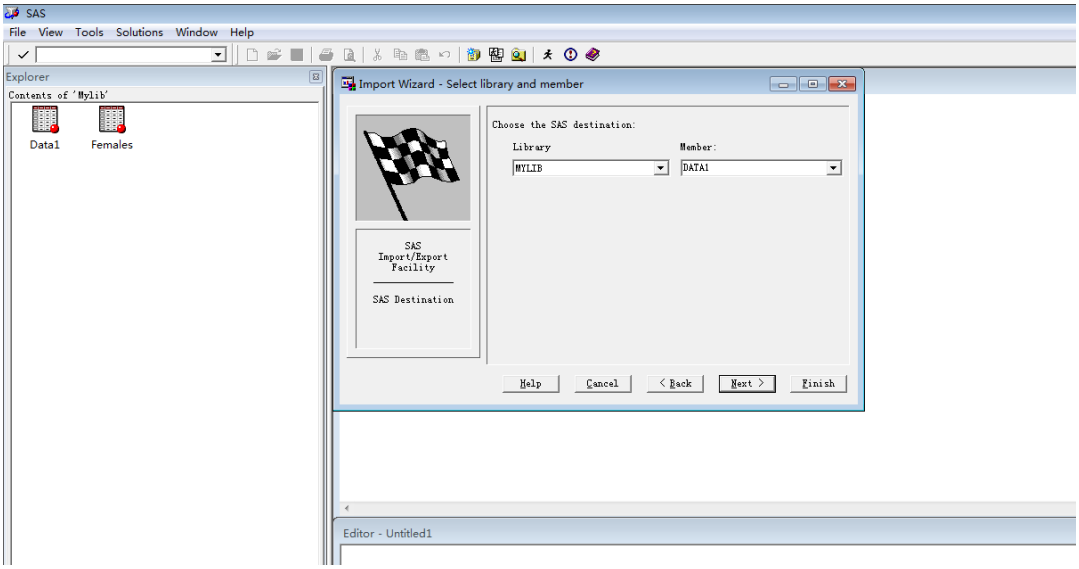
1. 菜单式操作




SAS
File View Tools Solutions Window Help
Explorer
Contents of 'Mylib'
Data1
Females
Import Wizard - Select import type
What type of data do you wish to import?
Standard data source
Select a data source from the list
Microsoft Excel 97 or 2000 (*.xls)
User-defined forma
Define a special file format using the External File Interface (EFI) facility.
SAS Import Wizard
Import EXCEL2000 data
Help
Cancel
< Back
Next >
Finish


SAS
Explorer
Contents of 'Mylib'
Data1
Females
Import Wizard - Select file
Where is the file
Browse...
Options...
SAS Import Wizard
Select file
Help
Cancel
< Back
Next >
Finish
打开
查找范围(I):
SAS
名称
修改日期
1221
20171215
20171229
sas
sas1208
sas大作业
SAS第一节课
SAS讲义
sas图形
bostondata
data201
文件名(N):
文件类型(T):
打开(O)
取消
Editor - Untitled1


SAS
File View Tools Solutions Window Help
Explorer
Contents of 'Mylib'
Data1
Females
Import Wizard - Select library and member
Choose the SAS destination:
Library
Member:
WORK
MAPS
MYLIB
SASHELP
SASUSER
WORK
SAS Import/Export Facility
SAS Destination
Help
Cancel
< Back
Next >
Finish



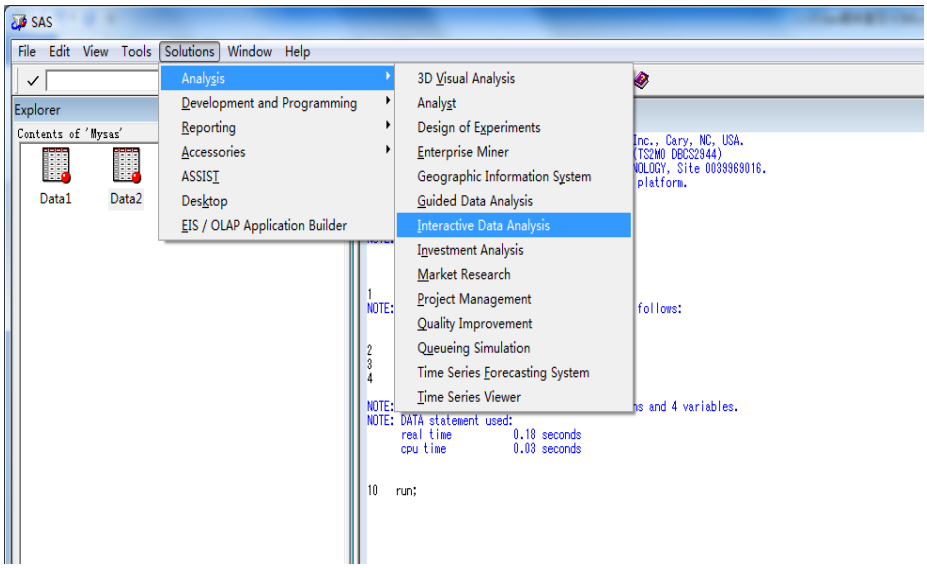
2. 编程式操作

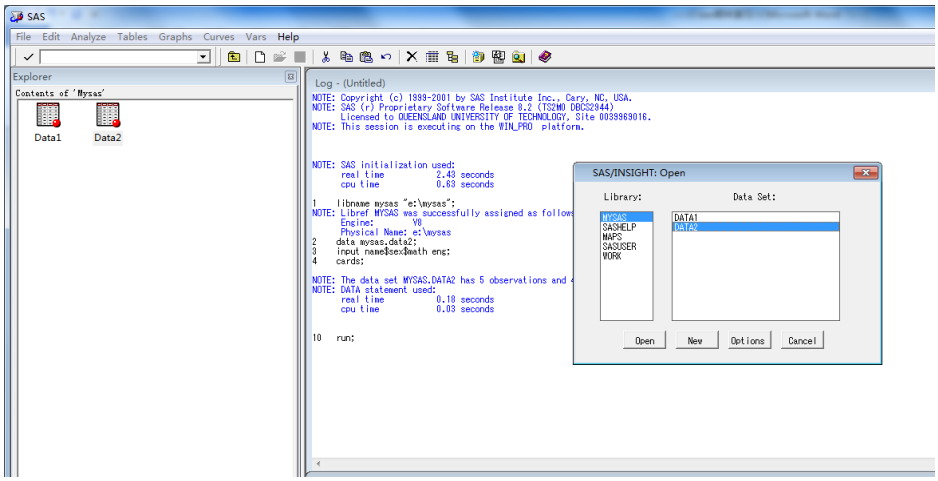
```
proc import datafile="E:\mylib\1.xls"  out=mylib.data;
run;/*第一行是变量名*/

libname mysas "e:\mysas";
data mysas.data2;
input name$sex$math eng;
cards;
李家富 男 92 98
张丽萍 女 89 106
王春雷 男 86 90
刘刚 男 98 109
张颖 女 80 110
run;
```

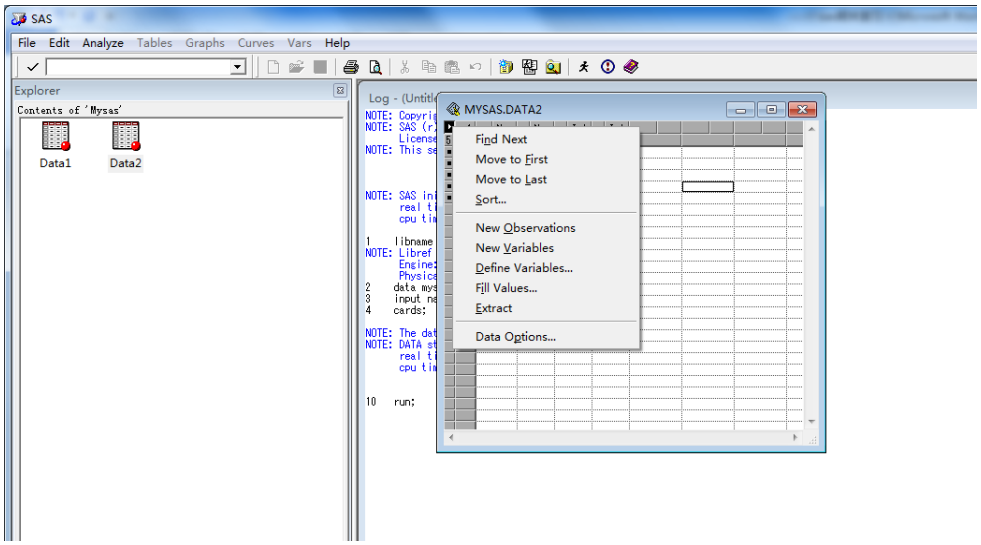
三、对数据的操作

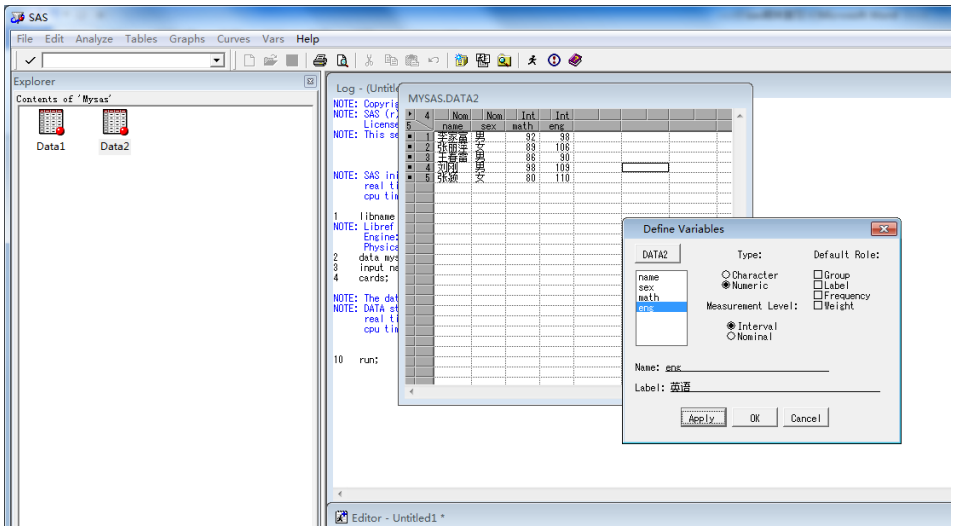
1. 数据的修改

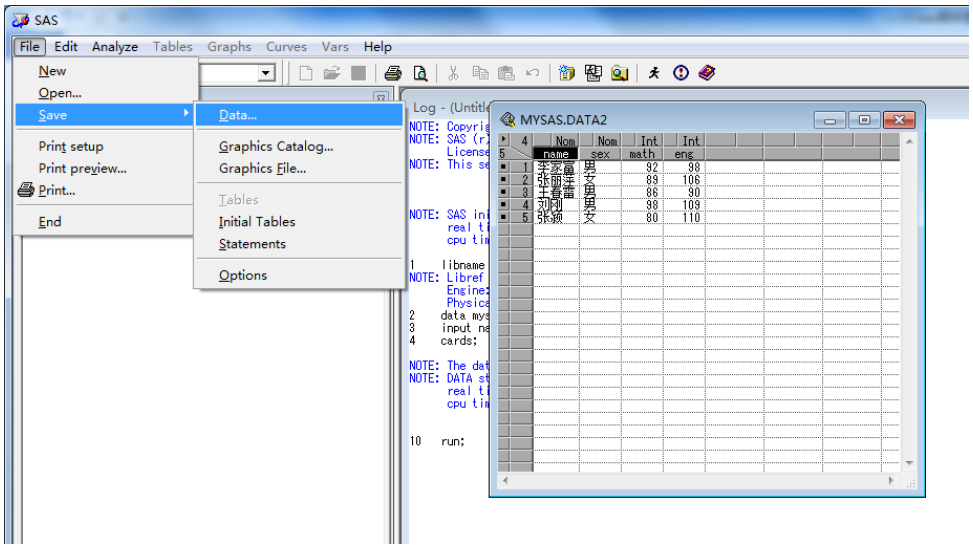
1) 对变量名，变量标签等的修改

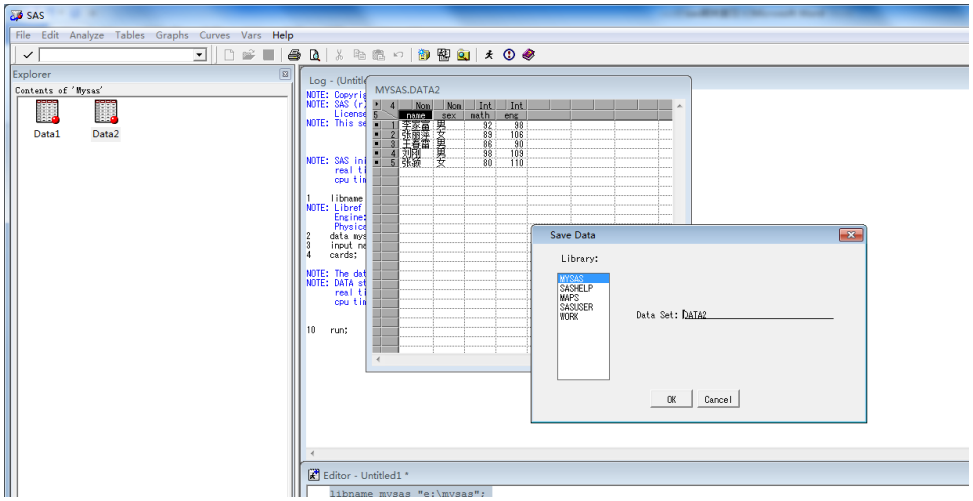

SAS
File Edit View Tools Solutions Window Help
Analysis
Development and Programming
Reporting
Accessories
ASSIST
Desktop
EIS / OLAP Application Builder
3D Visual Analysis
Analyst
Design of Experiments
Enterprise Miner
Geographic Information System
Guided Data Analysis
Interactive Data Analysis
Investment Analysis
Market Research
Project Management
Quality Improvement
Queueing Simulation
Time Series Forecasting System
Time Series Viewer
Explorer
Contents of 'Mysas'
Data1
Data2


SAS
File Edit Analyze Tables Graphs Curves Vars Help
Explorer
Contents of 'Mysas'
Data1
Data2
Log - (Untitled)
SAS/INSIGHT: Open
Library:
Data Set:
MYSAS
SASHELP
MAPS
SASUSER
WORK
DATA1
DATA2
Open
New
Options
Cancel


SAS
File Edit Analyze Tables Graphs Curves Vars Help
Explorer
Contents of 'Mysas'
Data1
Data2
MYSAS.DATA2
Find Next
Move to First
Move to Last
Sort...
New Observations
New Variables
Define Variables...
Fill Values...
Extract
Data Options...

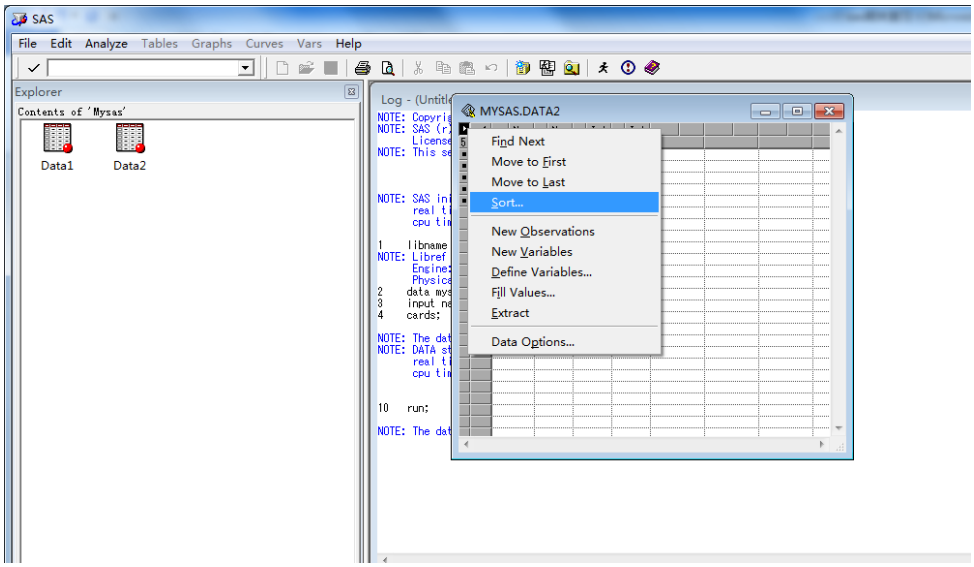






注意：修改变量名标签后一定要重新保存数据！（需把观看模式的数据页面关闭才能保存）

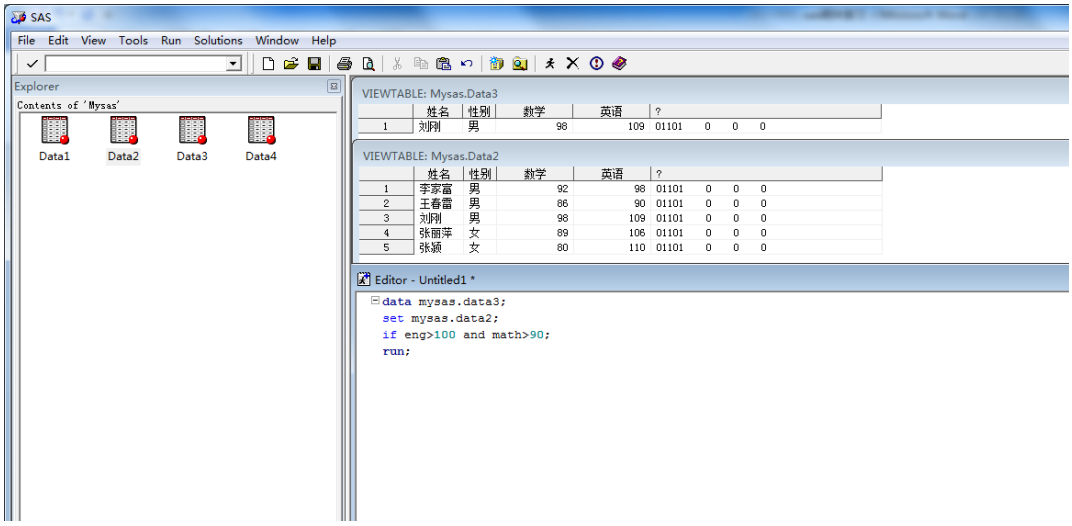
2) 对数据排序




```
proc sort data=mysas.data2;
by sex;
run;
```

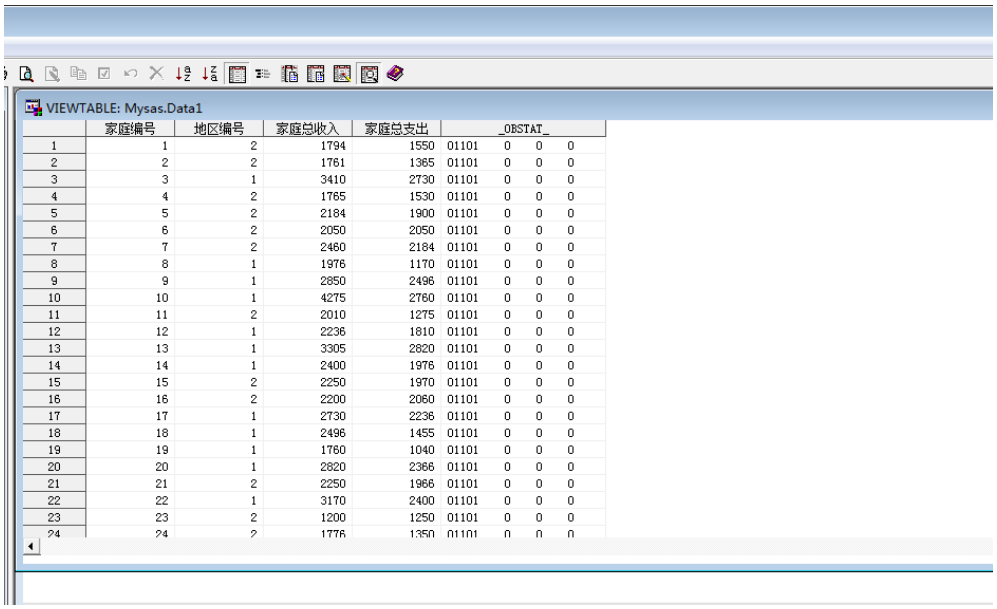
3) 筛选数据

```
data mysas.data3;
set mysas.data2;
if eng>100 and math>90;
run;
```

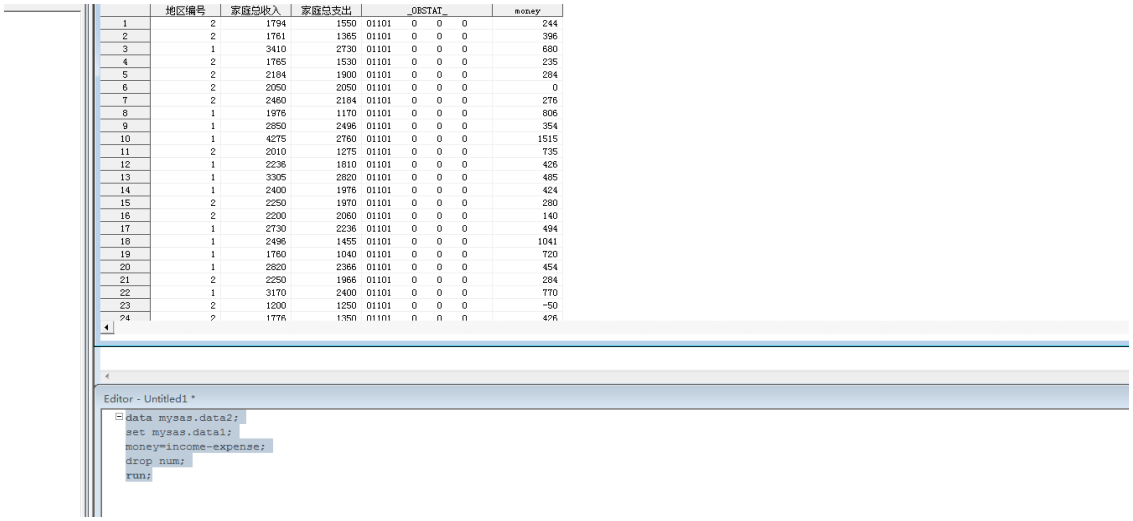


4) 数据的删除，创建新变量

```
data mysas.data2;
set mysas.data1;
money=income-expense;/*新变量*/
drop num;/*去掉变量*/
run;
```



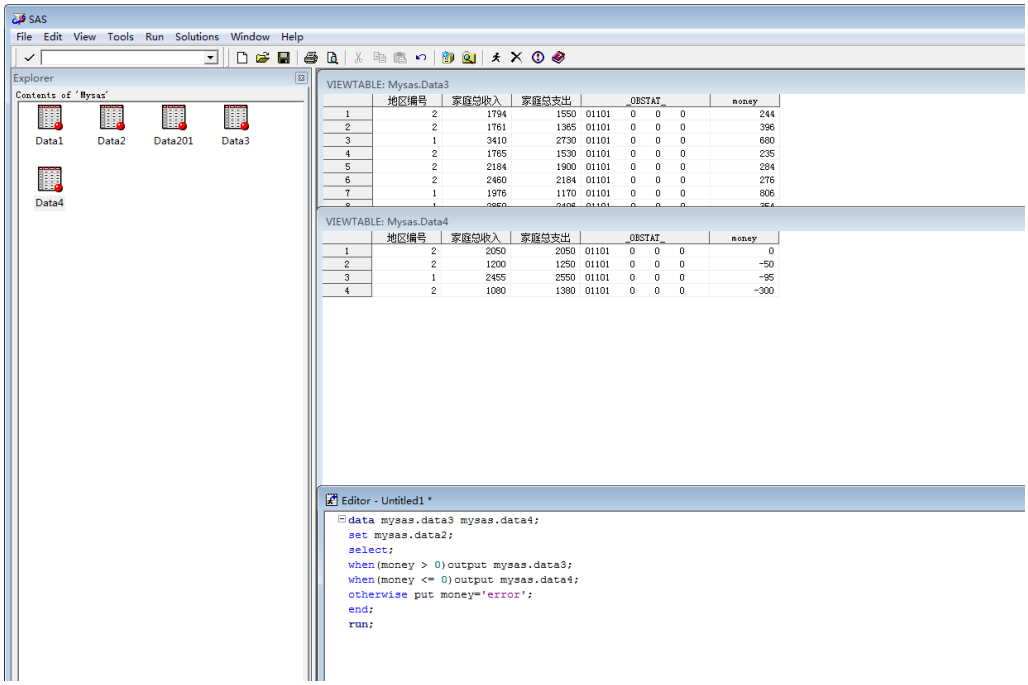
	家庭编号	地区编号	家庭总收入	家庭总支出	_OBSTAT_			
1	1	2	1794	1550	01101	0	0	0
2	2	2	1761	1365	01101	0	0	0
3	3	1	3410	2730	01101	0	0	0
4	4	2	1765	1530	01101	0	0	0
5	5	2	2184	1900	01101	0	0	0
6	6	2	2050	2050	01101	0	0	0
7	7	2	2460	2184	01101	0	0	0
8	8	1	1976	1170	01101	0	0	0
9	9	1	2850	2496	01101	0	0	0
10	10	1	4275	2760	01101	0	0	0
11	11	2	2010	1275	01101	0	0	0
12	12	1	2236	1810	01101	0	0	0
13	13	1	3305	2820	01101	0	0	0
14	14	1	2400	1976	01101	0	0	0
15	15	2	2250	1970	01101	0	0	0
16	16	2	2200	2060	01101	0	0	0
17	17	1	2730	2236	01101	0	0	0
18	18	1	2496	1455	01101	0	0	0
19	19	1	1760	1040	01101	0	0	0
20	20	1	2820	2366	01101	0	0	0
21	21	2	2250	1966	01101	0	0	0
22	22	1	3170	2400	01101	0	0	0
23	23	2	1200	1250	01101	0	0	0
24	24	2	1776	1350	01101	0	0	0



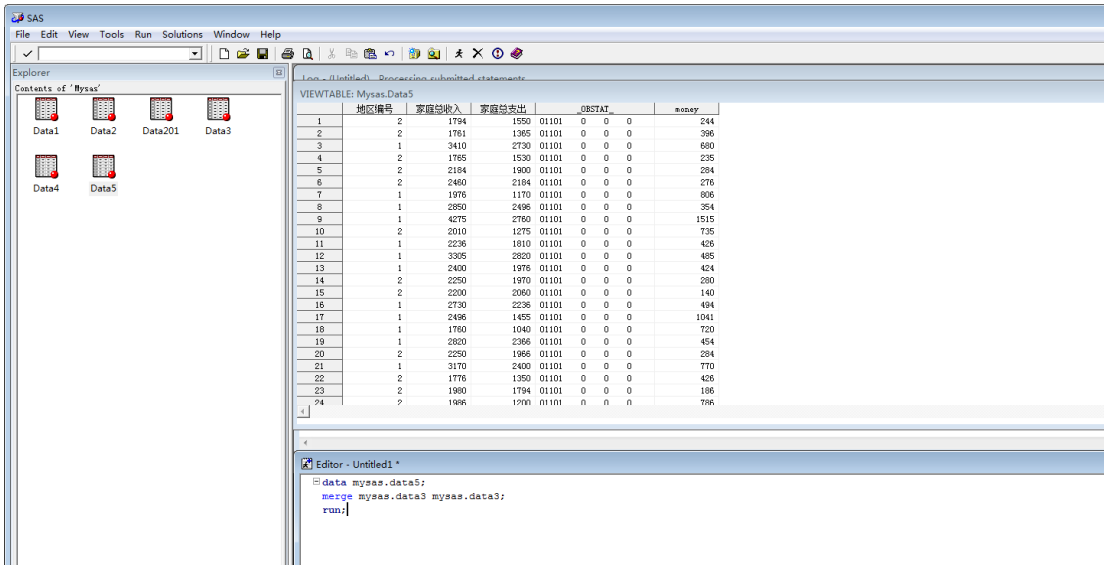
	地区编号	家庭总收入	家庭总支出	_OBSTAT_				money
1	2	1794	1550	01101	0	0	0	244
2	2	1761	1365	01101	0	0	0	396
3	1	3410	2730	01101	0	0	0	680
4	2	1765	1530	01101	0	0	0	235
5	2	2184	1900	01101	0	0	0	284
6	2	2050	2050	01101	0	0	0	0
7	2	2460	2184	01101	0	0	0	276
8	1	1976	1170	01101	0	0	0	806
9	1	2850	2496	01101	0	0	0	354
10	1	4275	2760	01101	0	0	0	1515
11	2	2010	1275	01101	0	0	0	735
12	1	2236	1810	01101	0	0	0	426
13	1	3305	2820	01101	0	0	0	485
14	1	2400	1976	01101	0	0	0	424
15	2	2250	1970	01101	0	0	0	280
16	2	2200	2060	01101	0	0	0	140
17	1	2730	2236	01101	0	0	0	494
18	1	2496	1455	01101	0	0	0	1041
19	1	1760	1040	01101	0	0	0	720
20	1	2820	2366	01101	0	0	0	454
21	2	2250	1966	01101	0	0	0	284
22	1	3170	2400	01101	0	0	0	770
23	2	1200	1250	01101	0	0	0	-50
24	2	1776	1350	01101	0	0	0	426

5) 数据的拆分，合并

```
data mysas.data3 mysas.data4;
set mysas.data2;
select;
when(money > 0)output mysas.data3;
when(money <= 0)output mysas.data4;
otherwise put money='error';
end;
run;
```




```
data mysas.data5;
merge mysas.data3 mysas.data3;
（by x/*根据实际情况，此处填相同的成员名*/）
run;
```

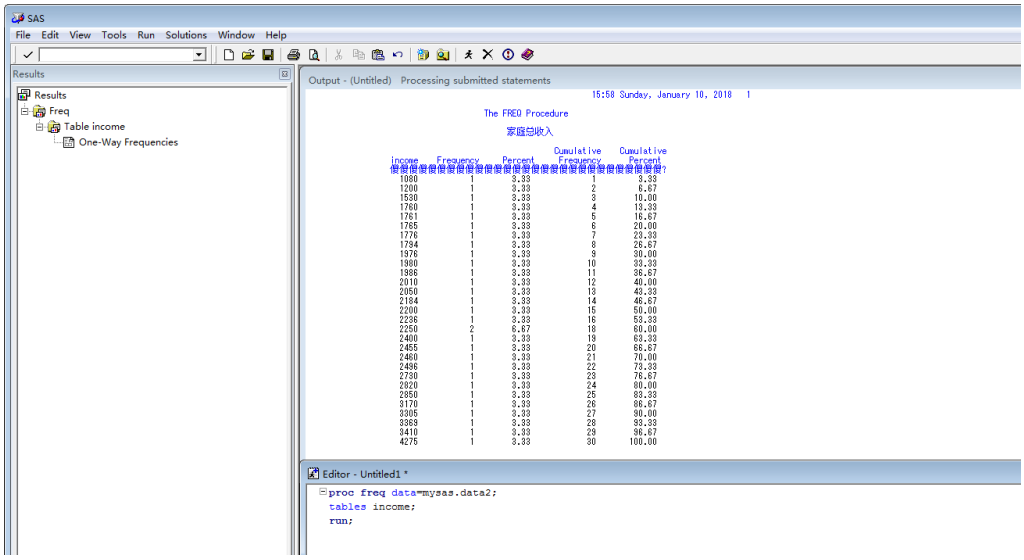


2. 数据的描述统计量

1) 输出描述统计量

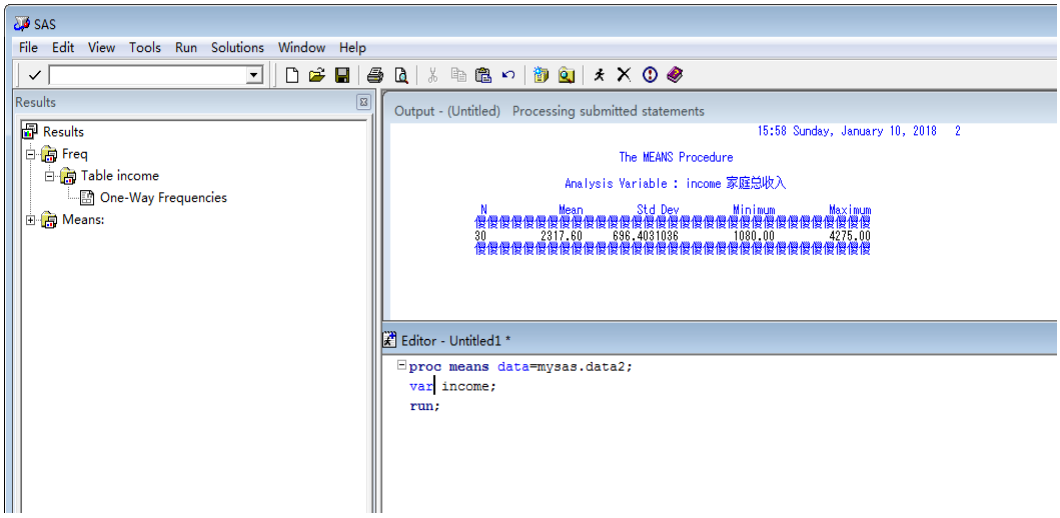
Freq 过程

```
proc freq data=mysas.data2;
tables income;
run;
```



Means 过程

```
proc means data=mysas.data2;
var income;/*输出简单统计量*/
run;
```



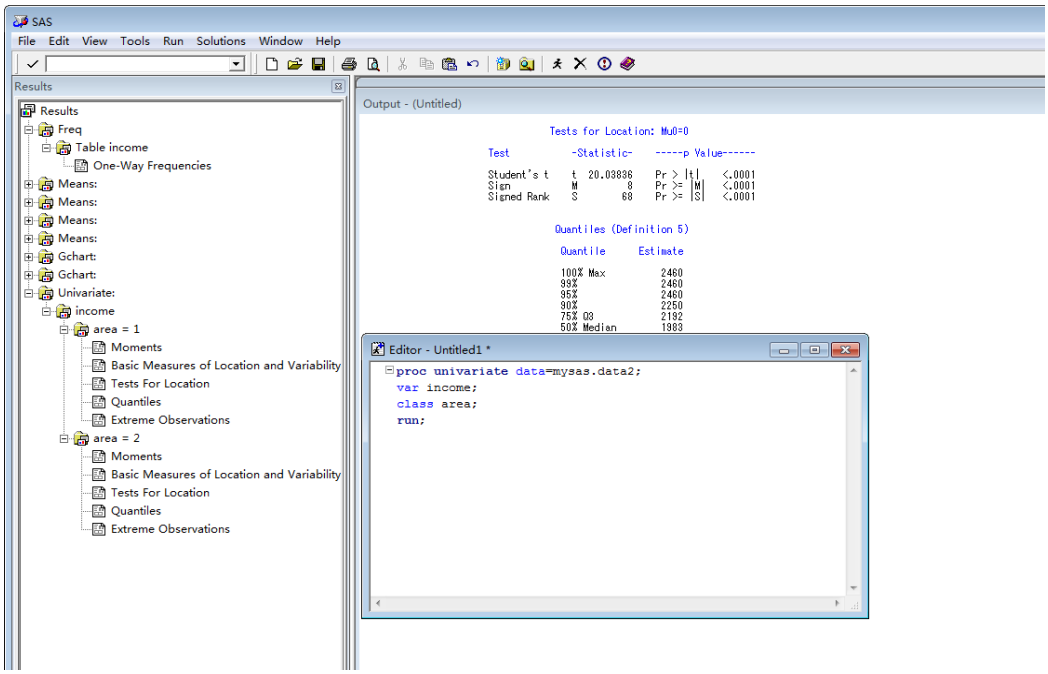

```
proc means data=mysas.data2  n mean median p1 p5 p95 p99 q1 q3 max min;/*指定输出统计量，数目，均值，中位数，第一位百分数，第五位百分数，第九十五位百分数，四分之一分位数，四分之三分位数*/
var income;
run;
```

统计参数的关键词	含　义	统计参数的关键词	含　义
N	样本数	Cv	变异系数
Mean	平均数	Var	方差
Std	标准差	Stderr	均值的标准误
Min	最小值	Skewness	偏度
Max	较大值	Kurtosis	峰度
Nmiss	缺失值个数	Q1|P25	四分之一分位数
Mode	众数	Q3|P75	四分之三分位数

Median	中位数	P1	第 1 百分位数
Range	极差	P5	第 5 百分位数
Uss	加权平方和	P10	第 10 百分位数
Css	均值偏差的加权平方和	P90	第 90 百分位数
Uclm	置信度上限	P95	第 95 百分位数
Lclm	置信度下限	P99	第 99 百分位数
Clm	置信度上限和下限	QRANGE	百分位数极差
Sum	累加和	PROBT|PRT	T 分布的双尾 p 值
Sumwgt	权数和	T	总体均值为 0 的 t 统计量

Univariate 过程

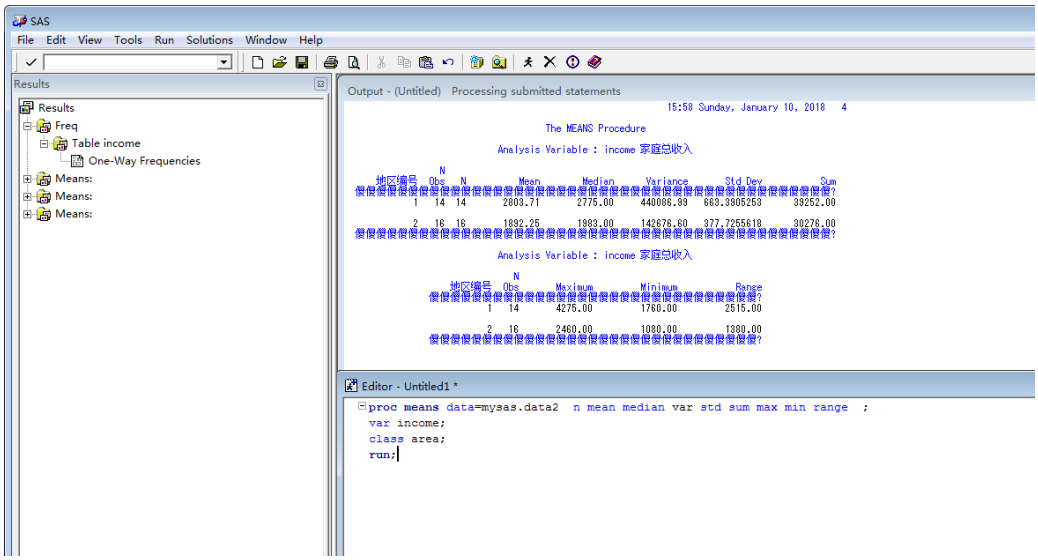
```
proc univariate data=mysas.data2;
var income;
class area;
run;
```



2) 分组输出描述统计量

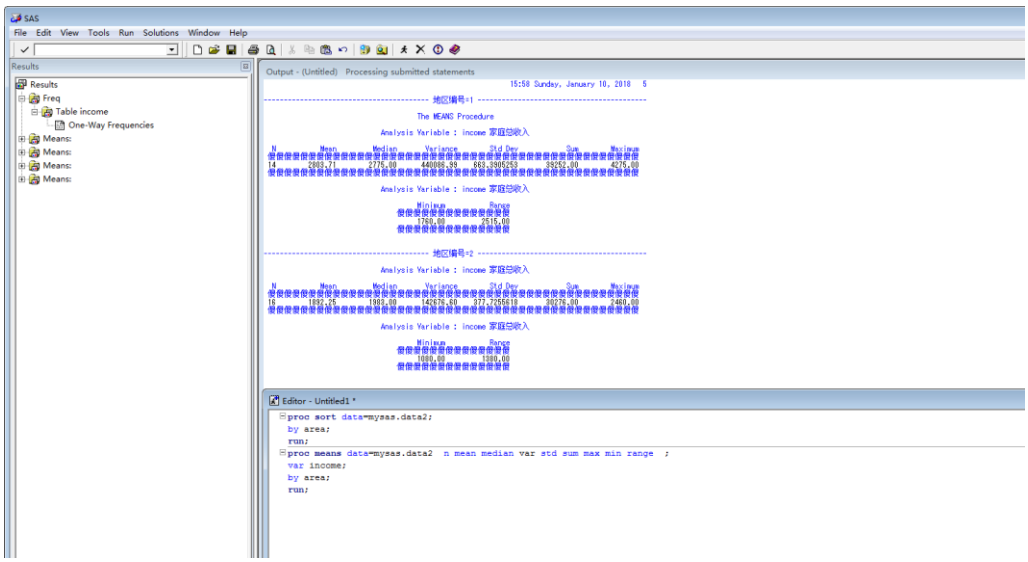
Class 法

```
proc means data=mysas.data2  n mean median var std sum max min
range  ;
var income;
class area;
run;
```

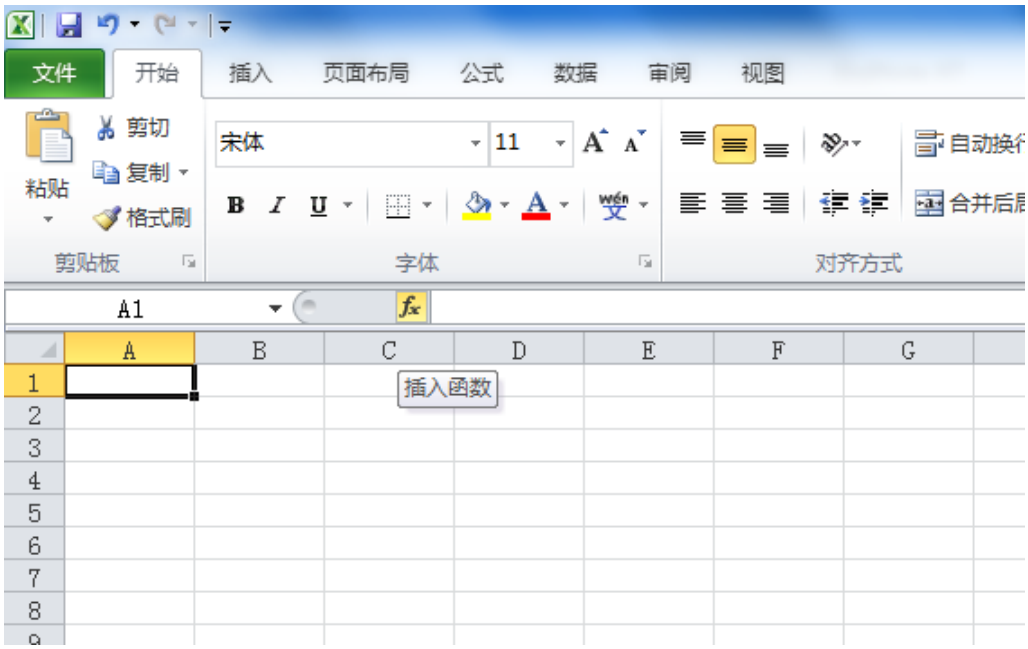


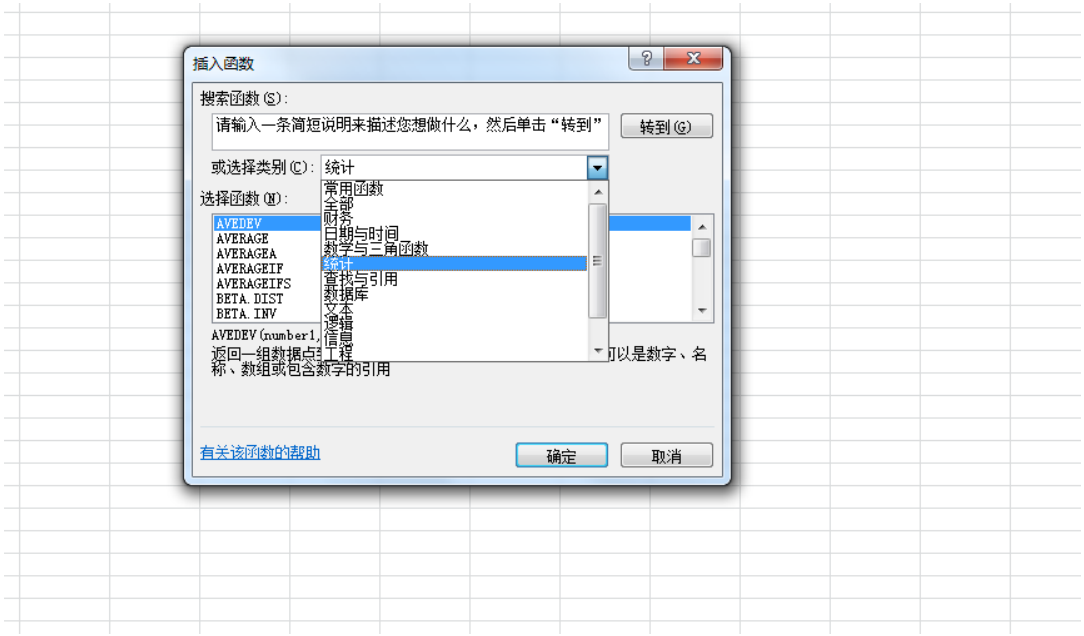
By 法（注意要先排序）

```
proc sort data=mysas.data2;
by area;
run;
proc means data=mysas.data2  n mean median var std sum max min range  ;
var income;
by area;
run;
```

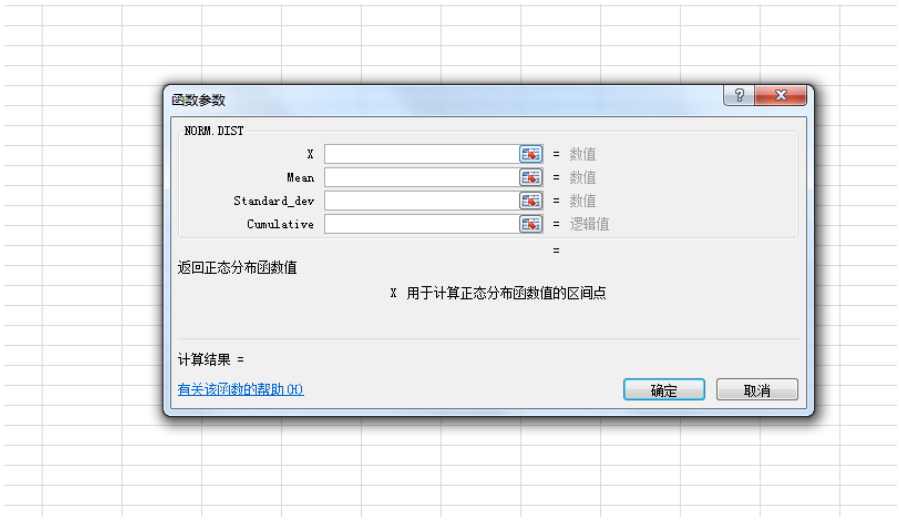


3) 三个典型分布（Excel）

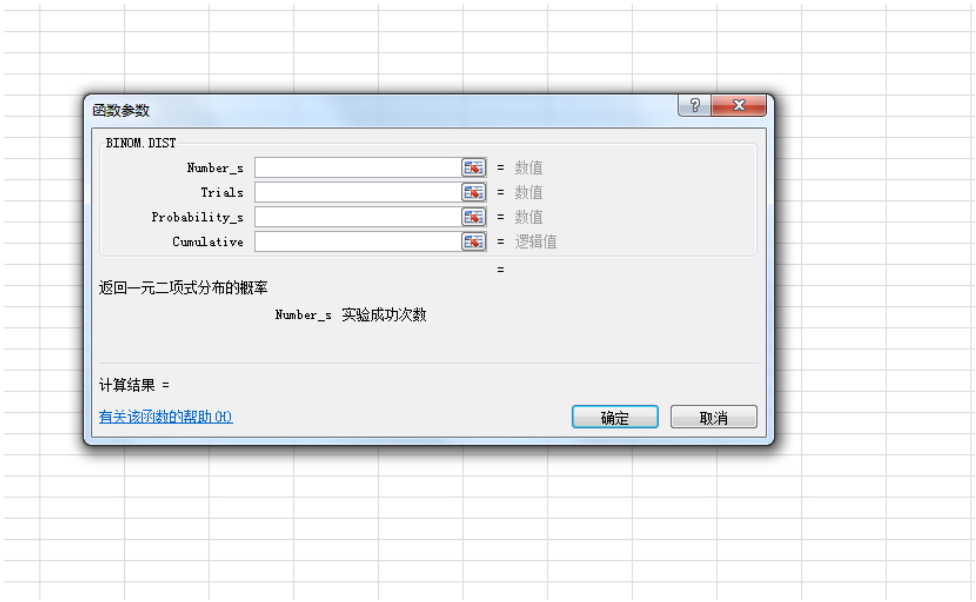

文件
开始
插入
页面布局
公式
数据
审阅
视图
剪切
复制
格式刷
粘贴
宋体
11
剪贴板
字体
对齐方式
A1
插入函数


插入函数
搜索函数(S):
请输入一条简短说明来描述您想做什么，然后单击“转到”
转到(G)
或选择类别(C): 统计
选择函数(N):
常用函数
全部
财务
日期与时间
数学与三角函数
统计
查找与引用
数据库
文本
逻辑
信息
工程
AVEDEV
AVERAGE
AVERAGEA
AVERAGEIF
AVERAGEIFS
BETA.DIST
BETA.INV
AVEDEV(number1,
返回一组数据点
称、数组或包含数字的引用
有关该函数的帮助
确定
取消

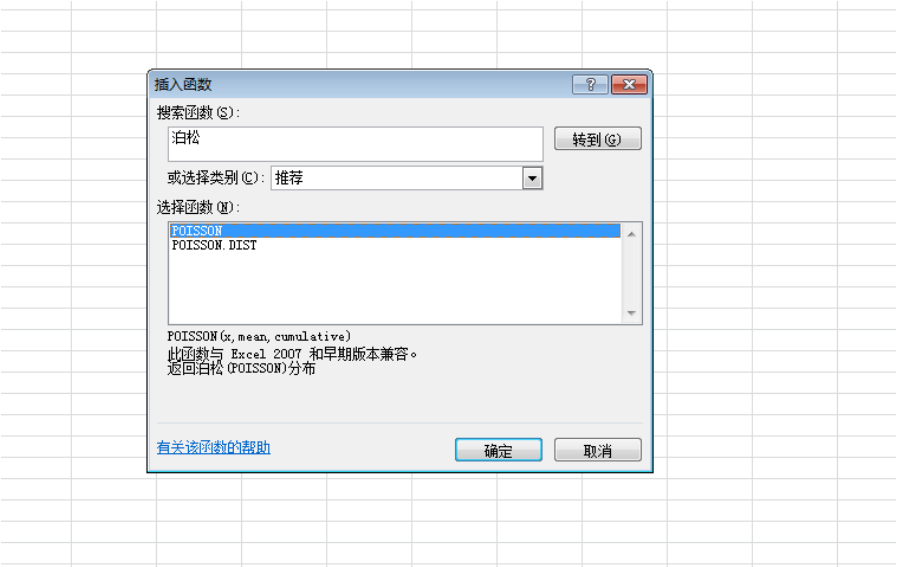
A. 正态分布


函数参数
NORM.DIST
X
Mean
Standard_dev
Cumulative
= 数值
= 数值
= 数值
= 逻辑值
返回正态分布函数值
X 用于计算正态分布函数值的区间点
计算结果 =
有关该函数的帮助(H)
确定
取消

B. 二项分布


函数参数
BINOM.DIST
Number_s
Trials
Probability_s
Cumulative
= 数值
= 数值
= 数值
= 逻辑值
返回一元二项式分布的概率
Number_s 实验成功次数
计算结果 =
有关该函数的帮助(H)
确定
取消

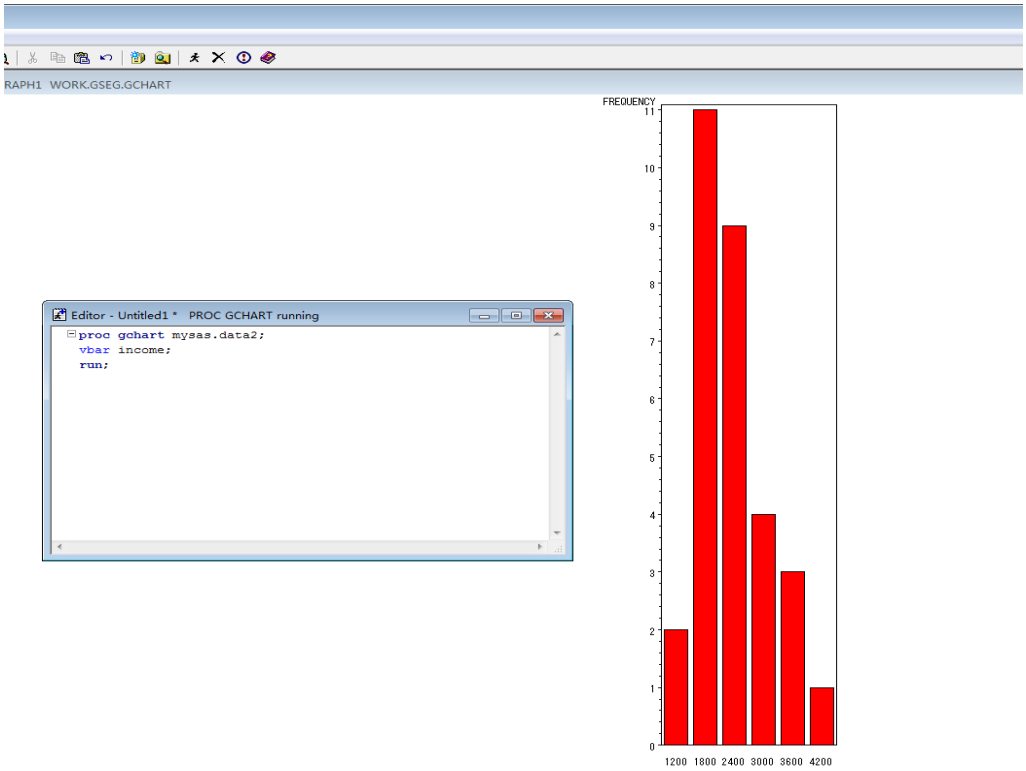
C. 泊松分布



3. 数据的图形描述

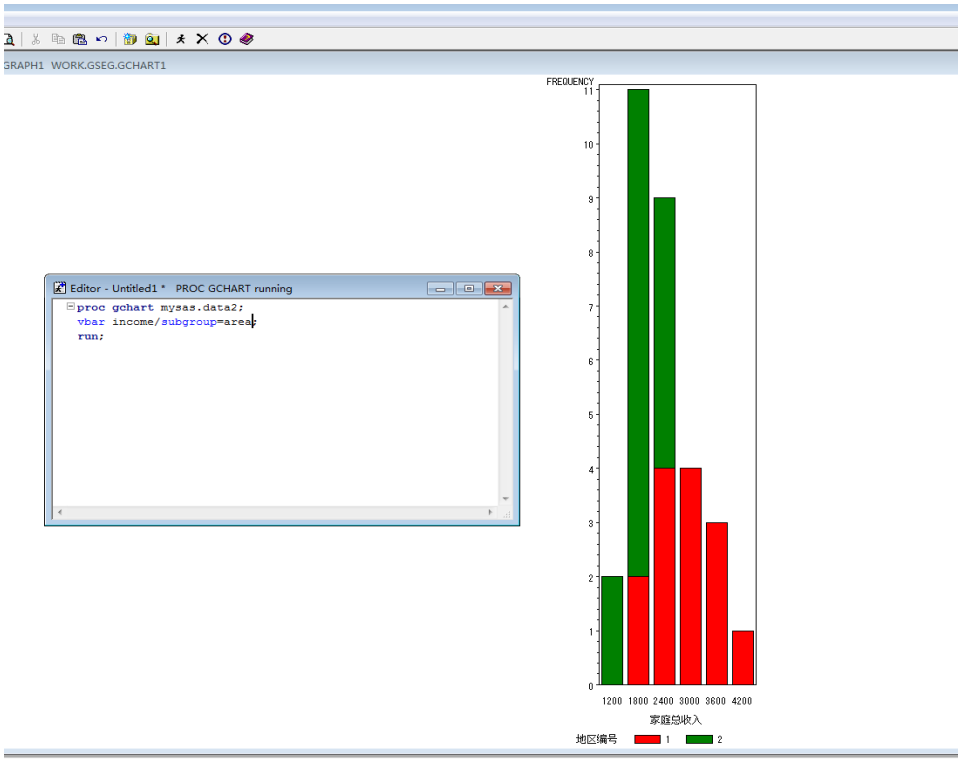
1) 条形图

```
proc gchart mysas.data2;
vbar income;
run;
```



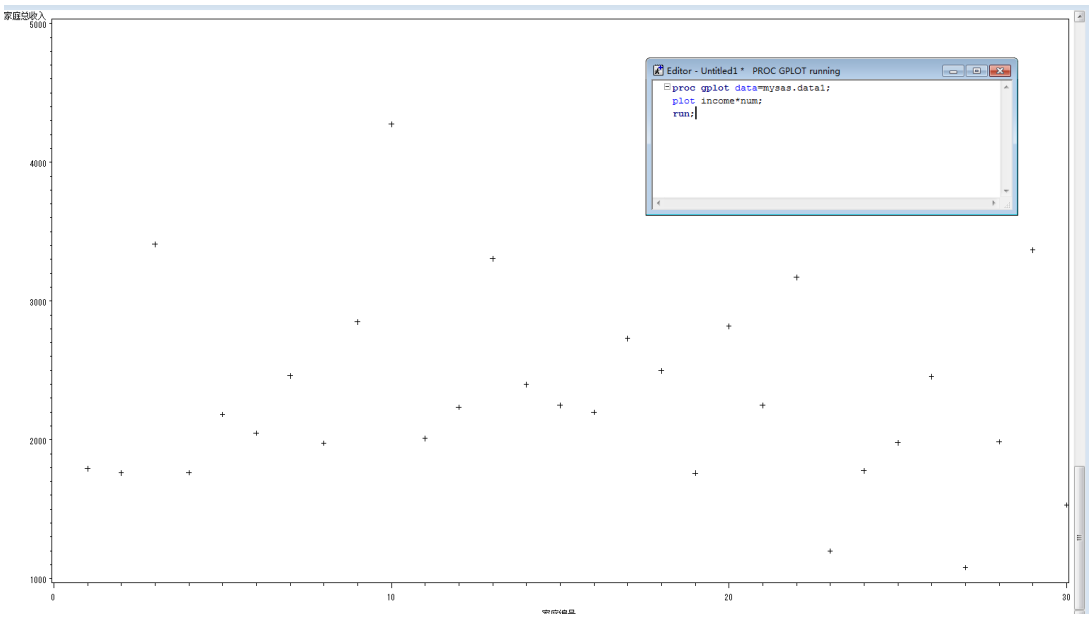
分组条形图

```
proc gchart mysas.data2;
vbar income/subgroup=area;
run;
```

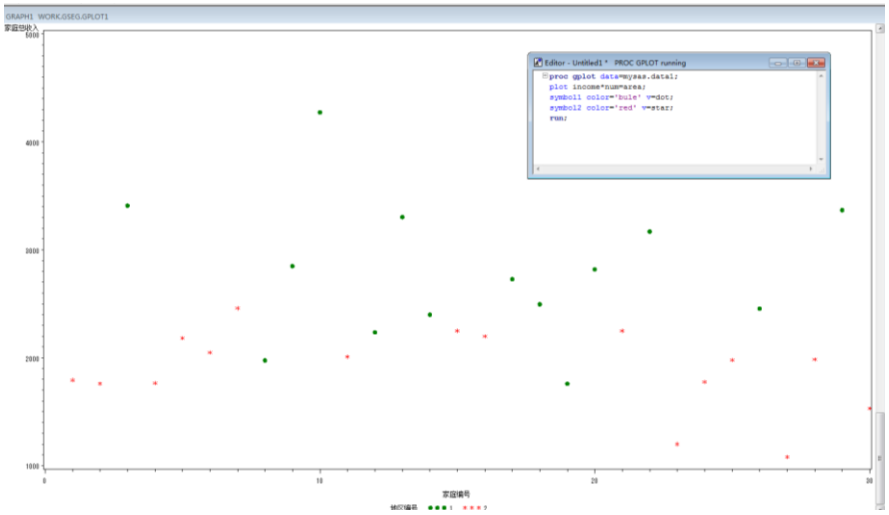


2) 散点图

```
proc gplot data=mysas.data1;
plot income*num;（先y后x）
run;
```

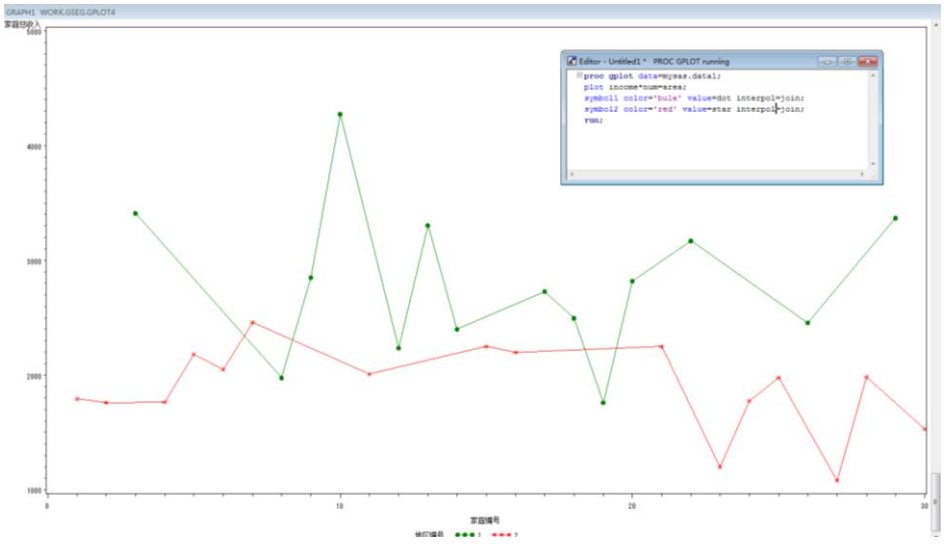



```
proc gplot data=mysas.data1;
plot income*num=area;
symbol1 color='bule' value=dot;
symbol2 color='red' value=star;
run;
```

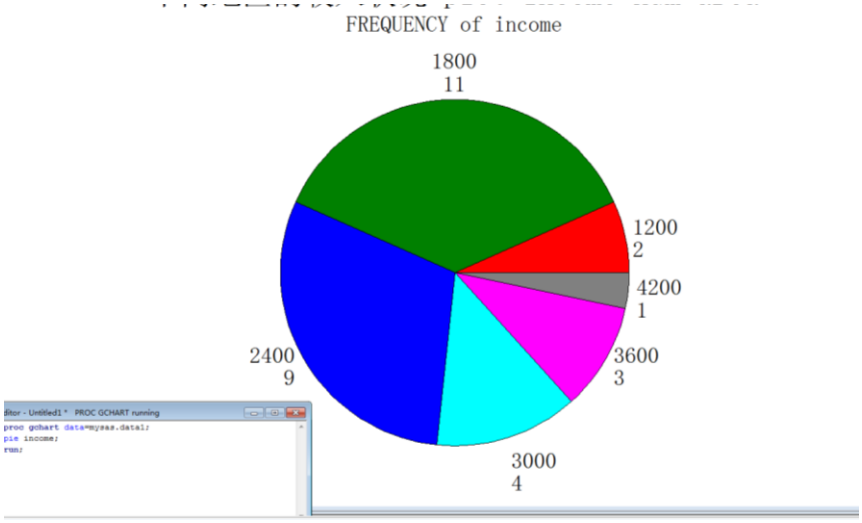


3) 折线图

```
proc gplot data=mysas.data1;
plot income*num=area;
symbol1 color='bule' value=dot interpol=join;/*color可用cv代替value可用v代替 Interpol可用i代替*/
symbol2 color='red' value=star interpol=join;
run;
```

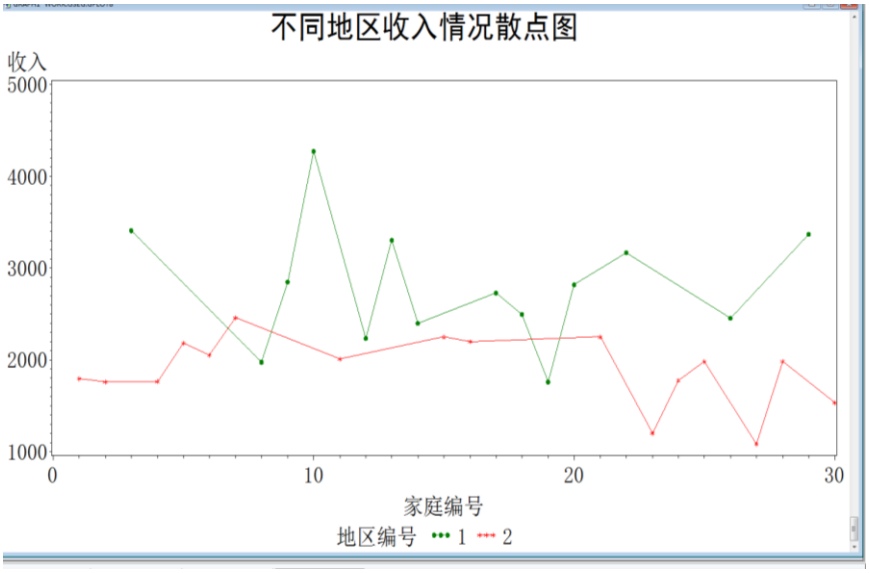


4) 饼图



附：添加标题，修改图形标签等

```
title f='黑体' '不同地区收入情况散点图';
proc gplot data=mysas.data1;
plot income*num=area;
symbol1 color='bule' value=dot;
symbol2 color='red' value=star;
label income='收入' ;
run;
```



PS:

整理较匆忙，如果有错误或者遗漏的重要知识点，欢迎补充.

By 凌宛莹

2018.1