Fudan University Library 2020

Regular Training

SPSS数据分析基础

胡杰

2020-06-02

如何获取正版SPSS

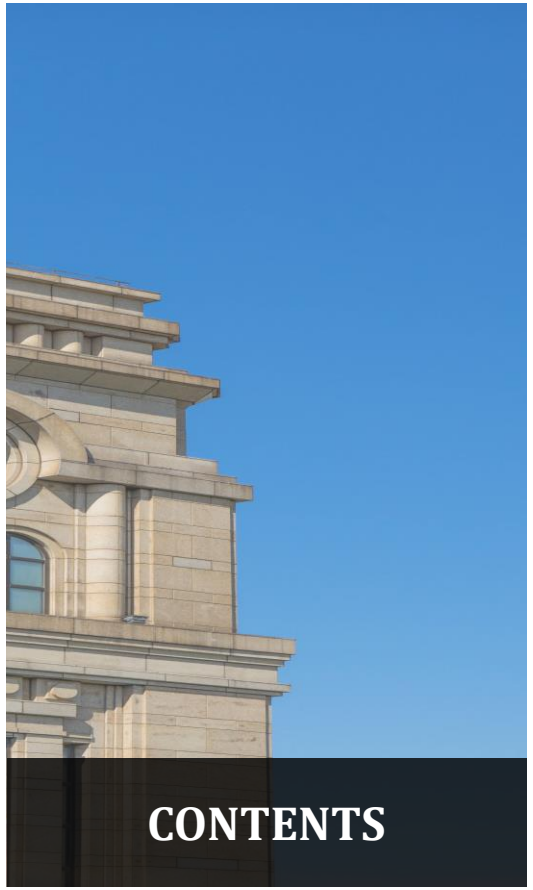
复旦信息办

服务对象：复旦大学在校师生

软件下载地址及安装说明请访问：

http://mvls.fudan.edu.cn

版本：SPSS 20.0



1 • SPSS快速入门

2 • 数据分析的一般步骤

3 • 问卷数据分析方法

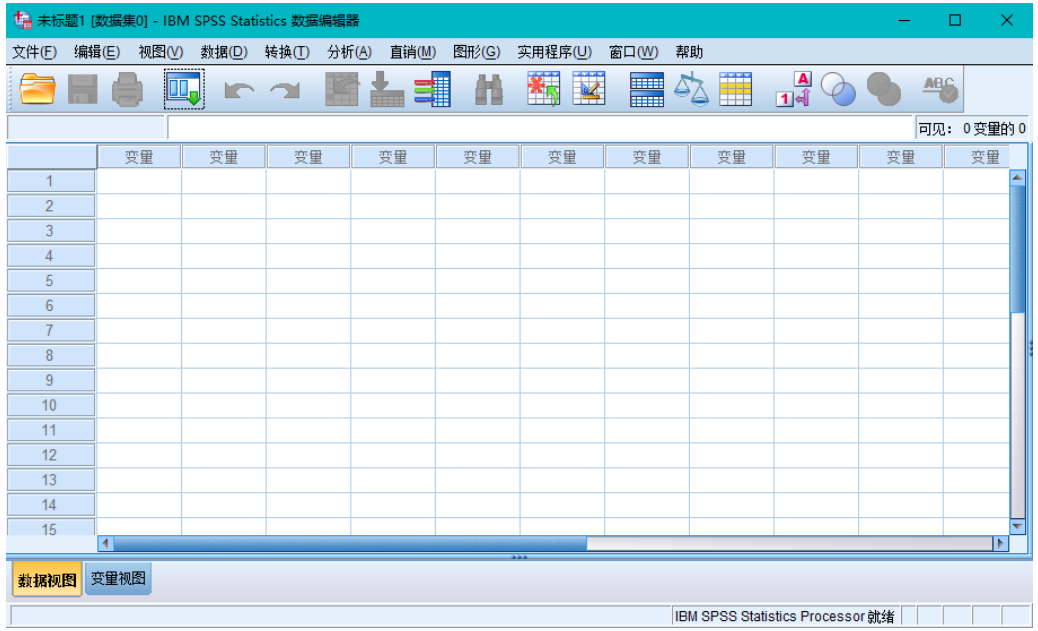
/01

SPSS快速入门

了解SPSS的基本窗口是学习使用SPSS的入门点

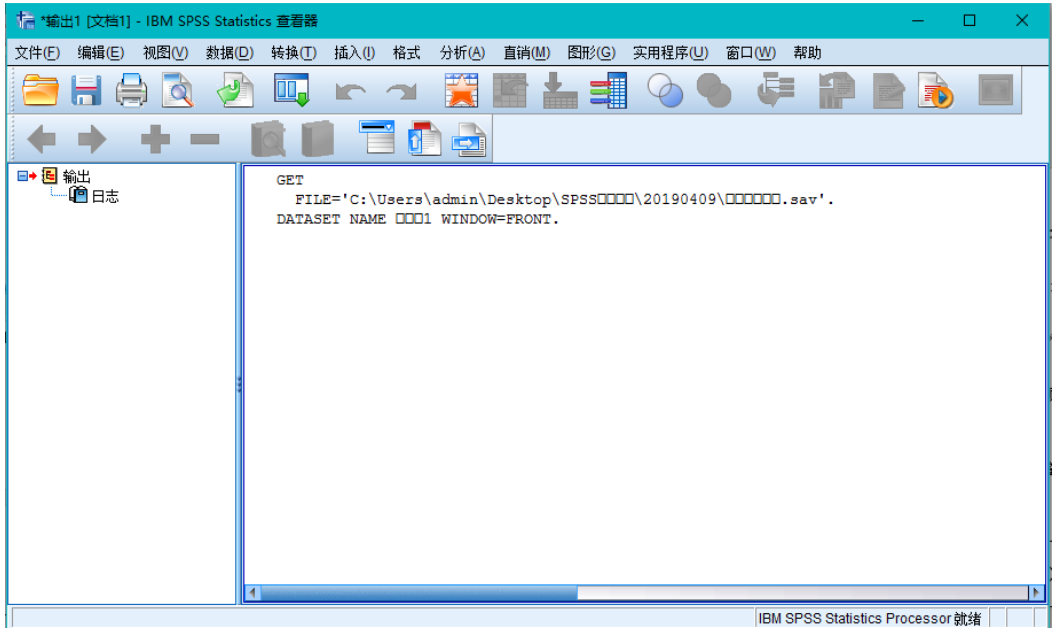
- SPSS有哪些基本操作窗口
- 各个窗口的功能特点是什么
- 窗口之间有怎样的关系

SPSS的基本窗口



数据编辑窗口

（输入和管理待分析数据）



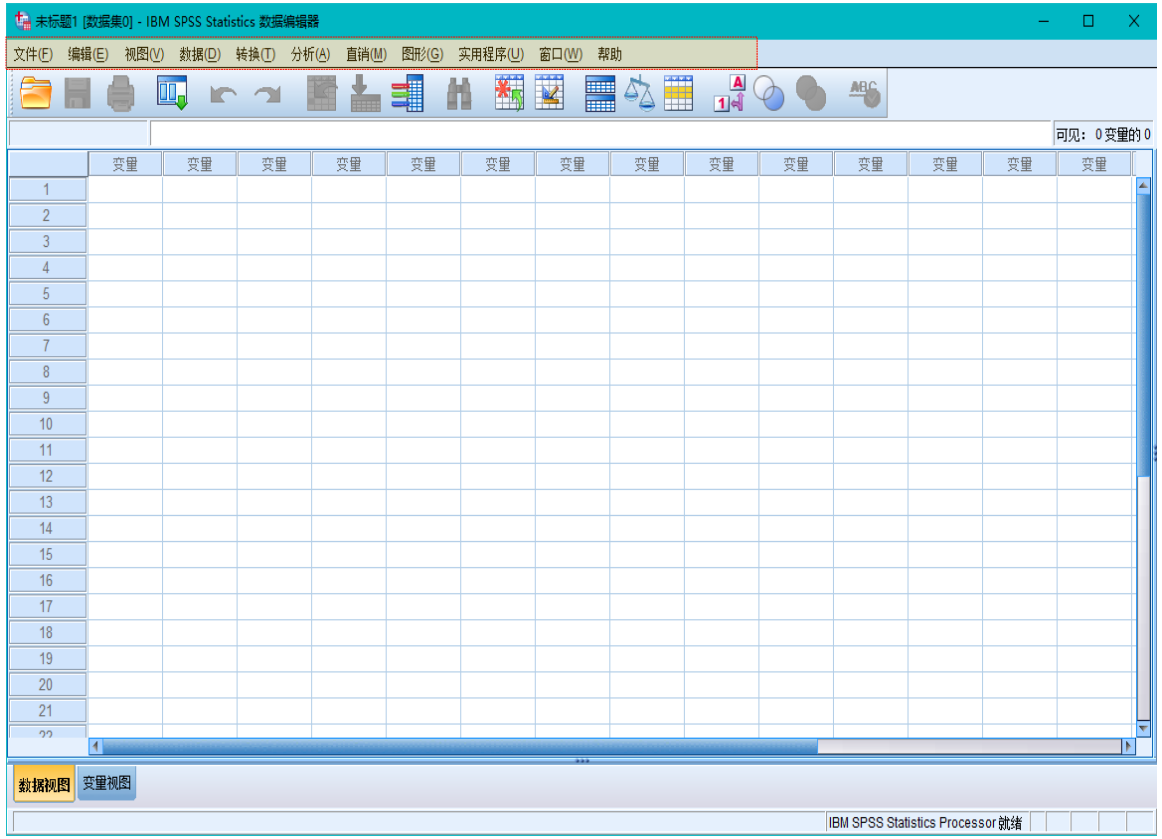
结果输出窗口

（接受和管理统计分析的结果）

数据编辑窗口

窗口主菜单

罗列了SPSS常用的数据编辑、加工和分析的功能



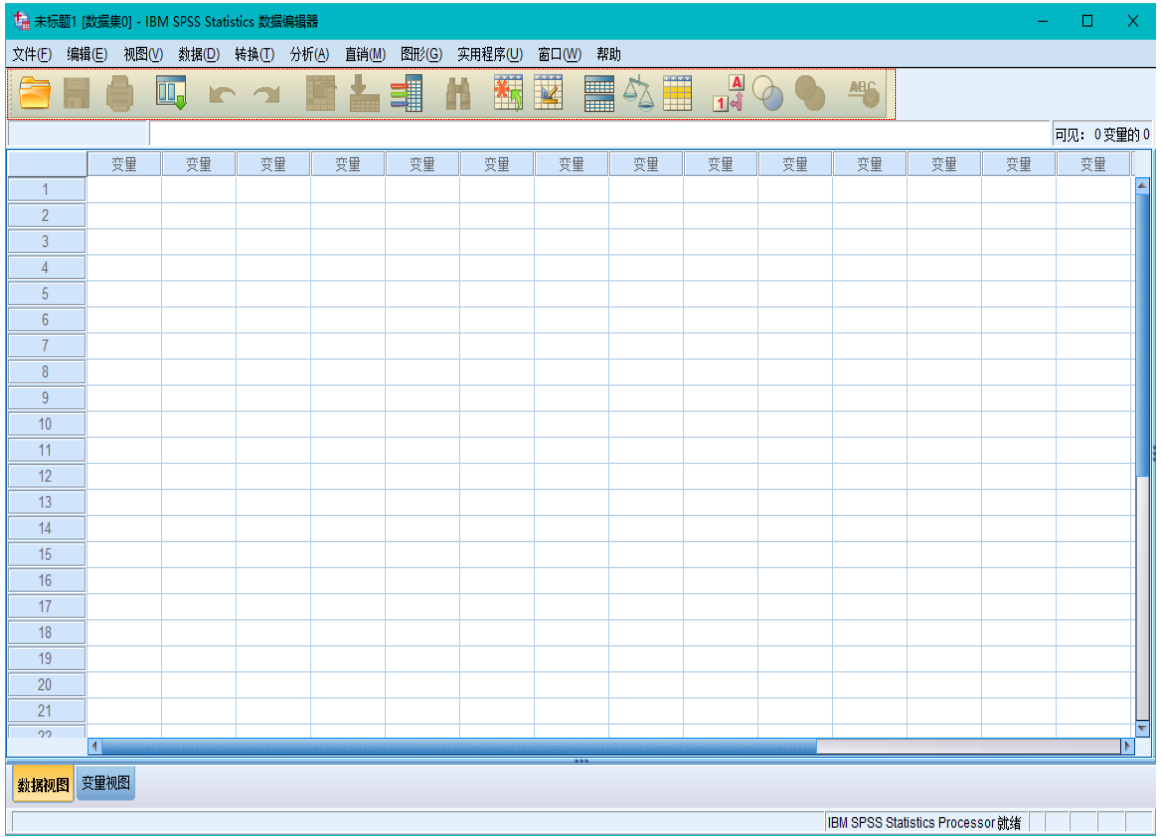
窗口主菜单

菜单名	功能	解释
文件（File）	文件操作	对SPSS相关文件进行基本管理：如新建、打开、保存、打印等
编辑（Edit）	数据编辑	对数据进行基本编辑，实现数据查找等功能：如撤销/恢复、剪切、复制、粘贴等
视图（View）	窗口外观状态管理	对SPSS窗口外观等进行设置：如状态栏、表格线等是否显示，字体设置等
数据（Data）	数据的操作和管理	对数据进行加工整理：如数据的排序、转置、抽样、分类汇总等
转换（Transform）	数据基本处理	对数据进行基本处理：如生成新变量、计数、分组等
分析（Analyze）	统计分析	对数据进行统计分析和建模：如基本统计分析、均值比较、回归分析等
图形（Graphs）	制作统计图形	对数据生成各种统计图形：如条形图、直方图、饼图、线图、散点图等
实用程序（Utilities）	实用程序	SPSS其他辅助管理：如显示变量信息、定义变量集、菜单编辑器等
窗口（Windows）	窗口管理	对SPSS的多个窗口进行管理：如窗口切换、最小化窗口等
帮助（Help）	帮助	实现SPSS的联机帮助：如语句检索、统计教程等

数据编辑窗口

窗口主菜单

罗列了SPSS常用的数据编辑、加工和分析的功能



工具栏

一些常用功能以图形按钮的形式组织在工具栏中，操作更加快捷和方便。

数据编辑窗口

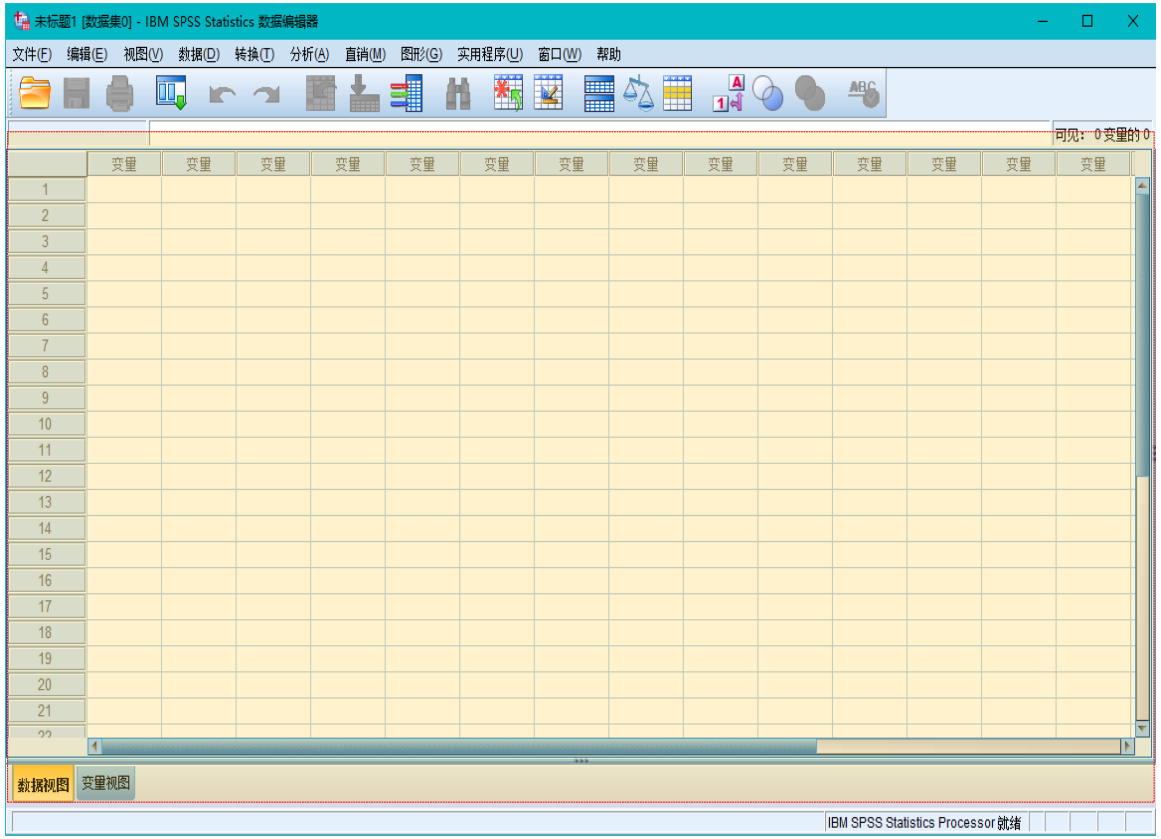
窗口主菜单

罗列了SPSS常用的数据编辑、加工和分析的功能

数据编辑区

数据视图：录入、编辑和管理数据内容

变量视图：定义和修改数据结构



工具栏

一些常用功能以图形按钮的形式组织在工具栏中，操作更加快捷和方便。

数据编辑窗口

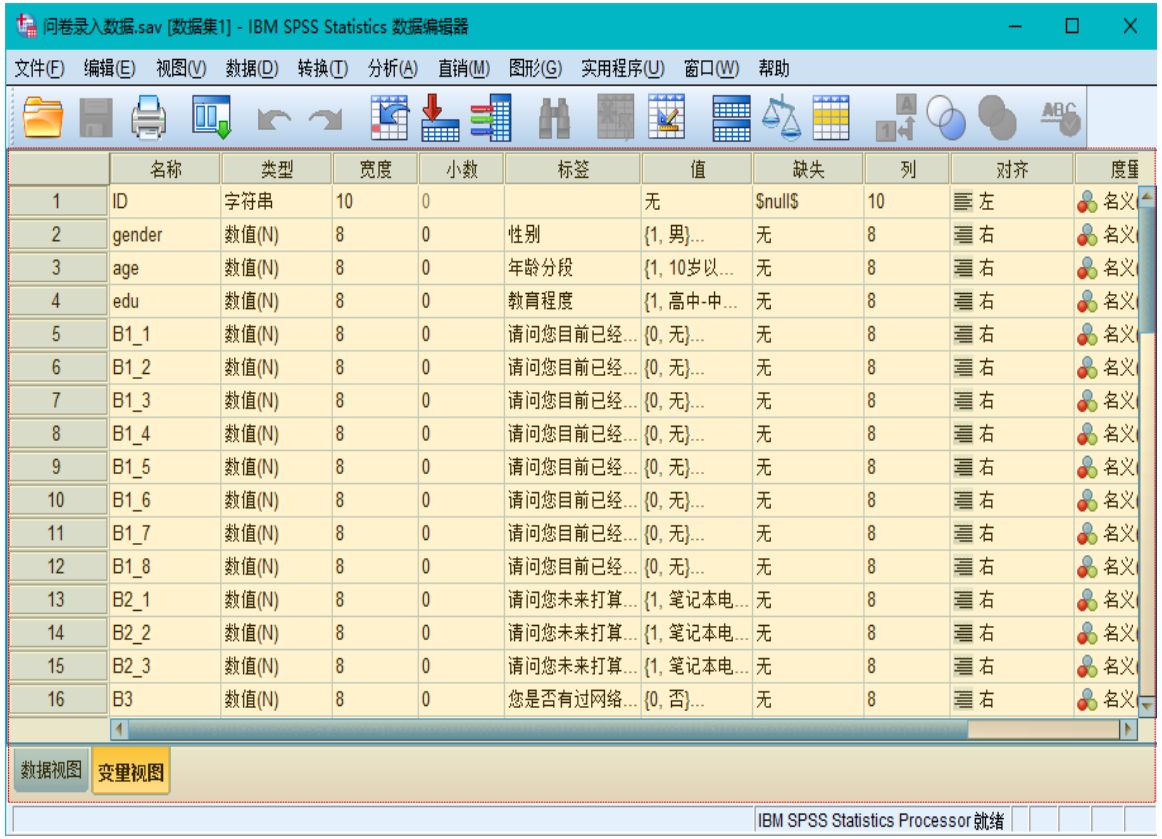
窗口主菜单

罗列了SPSS常用的数据编辑、加工和分析的功能

数据编辑区

数据视图：录入、编辑和管理数据内容

变量视图：定义和修改数据结构



	名称	类型	宽度	小数	标签	值	缺失	列	对齐	度量
1	ID	字符串	10	0		无	$null$	10	左	名义
2	gender	数值(N)	8	0	性别	{1, 男}...	无	8	右	名义
3	age	数值(N)	8	0	年龄分段	{1, 10岁以...	无	8	右	名义
4	edu	数值(N)	8	0	教育程度	{1, 高中-中...	无	8	右	名义
5	B1_1	数值(N)	8	0	请问您目前已经...	{0, 无}...	无	8	右	名义
6	B1_2	数值(N)	8	0	请问您目前已经...	{0, 无}...	无	8	右	名义
7	B1_3	数值(N)	8	0	请问您目前已经...	{0, 无}...	无	8	右	名义
8	B1_4	数值(N)	8	0	请问您目前已经...	{0, 无}...	无	8	右	名义
9	B1_5	数值(N)	8	0	请问您目前已经...	{0, 无}...	无	8	右	名义
10	B1_6	数值(N)	8	0	请问您目前已经...	{0, 无}...	无	8	右	名义
11	B1_7	数值(N)	8	0	请问您目前已经...	{0, 无}...	无	8	右	名义
12	B1_8	数值(N)	8	0	请问您目前已经...	{0, 无}...	无	8	右	名义
13	B2_1	数值(N)	8	0	请问您未来打算...	{1, 笔记本电...	无	8	右	名义
14	B2_2	数值(N)	8	0	请问您未来打算...	{1, 笔记本电...	无	8	右	名义
15	B2_3	数值(N)	8	0	请问您未来打算...	{1, 笔记本电...	无	8	右	名义
16	B3	数值(N)	8	0	您是否有过网络...	{0, 否}...	无	8	右	名义

工具栏

一些常用功能以图形按钮的形式组织在工具栏中，操作更加快捷和方便。

数据编辑窗口

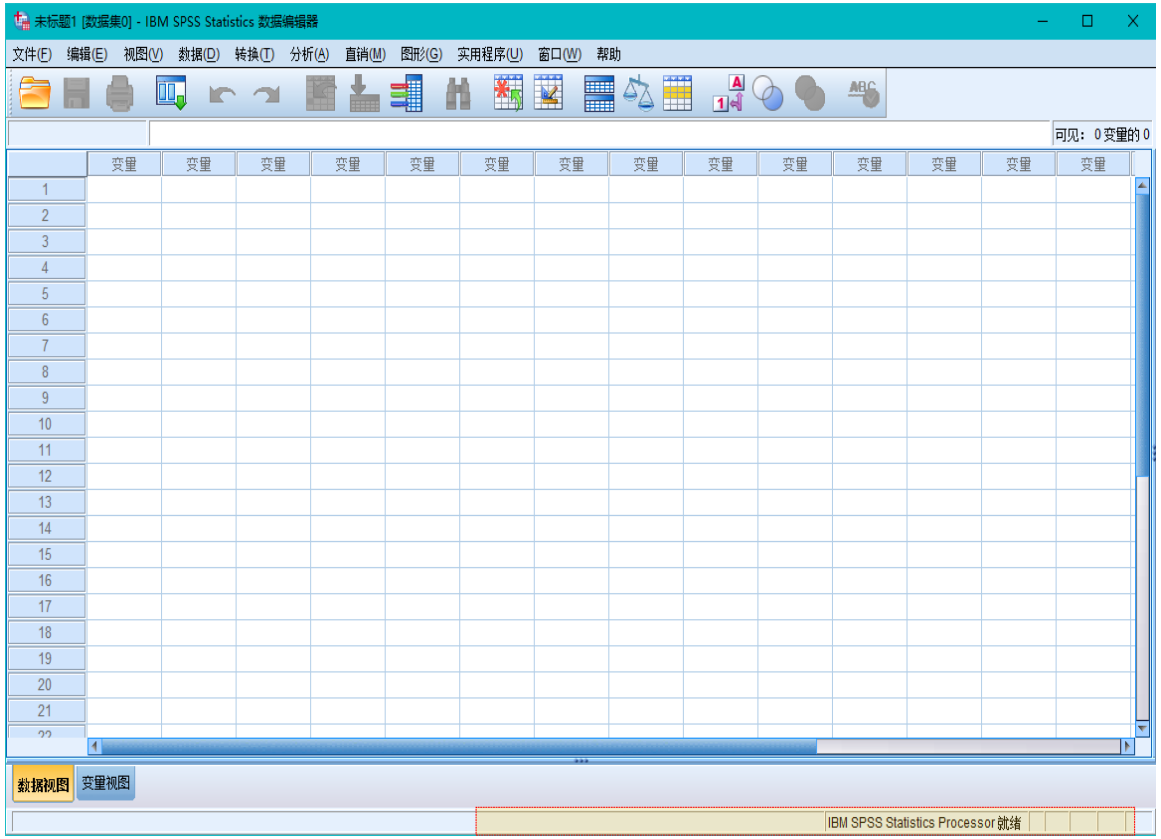
窗口主菜单

罗列了SPSS常用的数据编辑、加工和分析的功能

数据编辑区

数据视图：录入、编辑和管理数据内容

变量视图：定义和修改数据结构



工具栏

一些常用功能以图形按钮的形式组织在工具栏中，操作更加快捷和方便。

系统状态显示区

显示系统的当前运行状态

数据编辑窗口

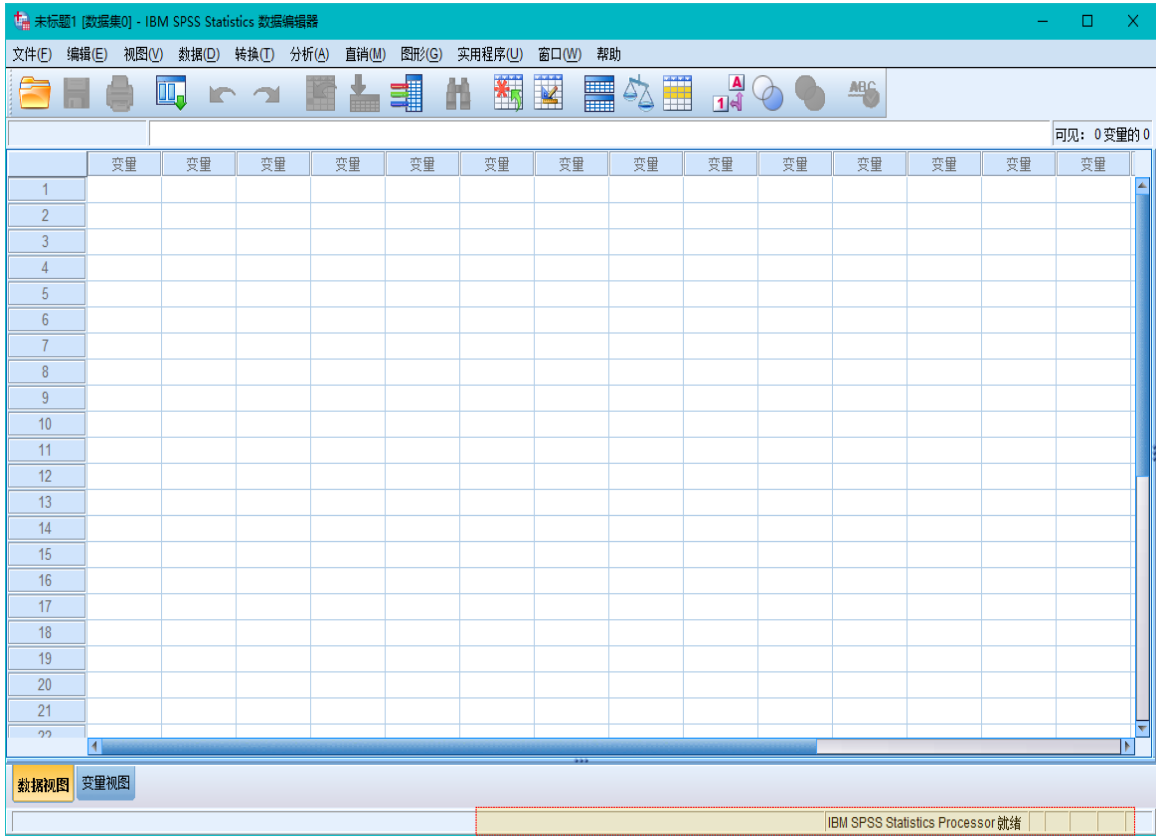
窗口主菜单

罗列了SPSS常用的数据编辑、加工和分析的功能

数据编辑区

数据视图：录入、编辑和管理数据内容

变量视图：定义和修改数据结构



工具栏

一些常用功能以图形按钮的形式组织在工具栏中，操作更加快捷和方便。

系统状态显示区

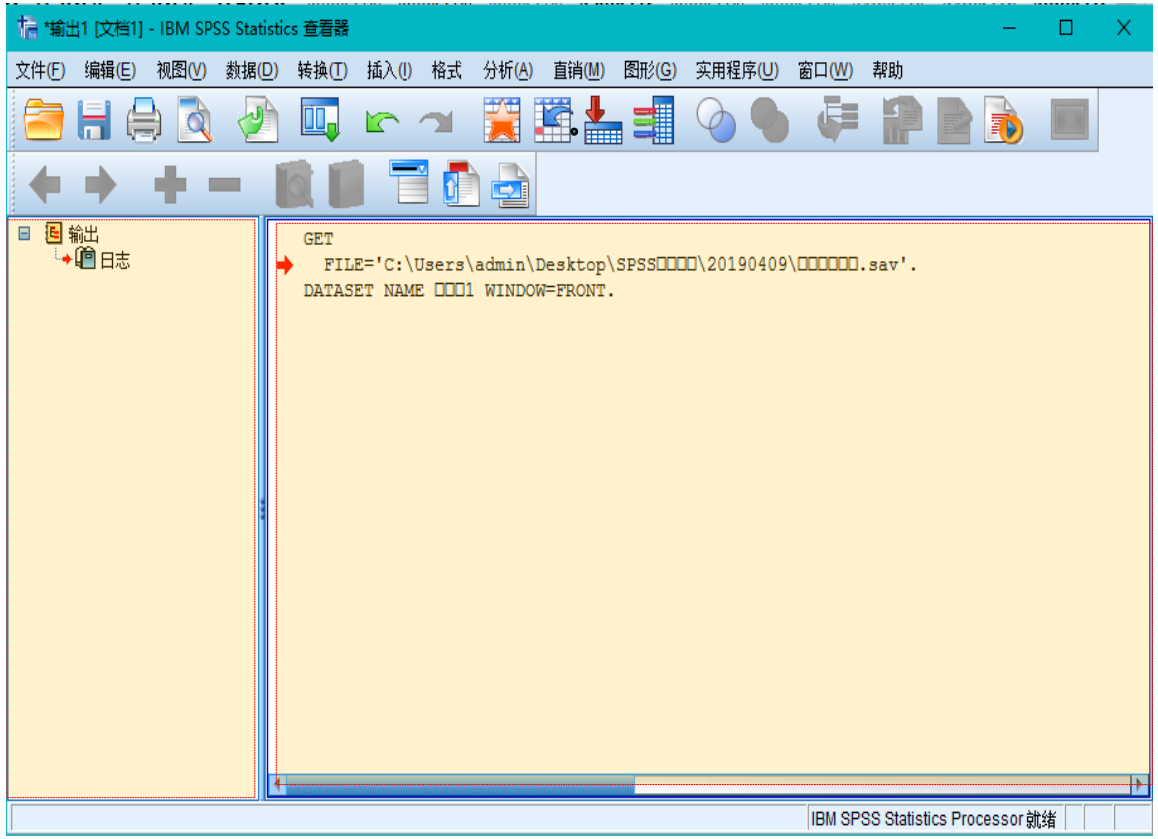
显示系统的当前运行状态

- ✓ 数据编辑窗口是SPSS的主程序窗口，可以同时打开多个数据编辑窗口。
- ✓ 这些数据通常以SPSS数据文件的形式保存，文件扩展名为.sav。
- ✓ 数据编辑窗口的主要功能是：定义SPSS数据的结构、录入编辑和管理待分析的数据。

结果输出窗口

窗口主菜单

与编辑窗口主菜单选项大致相同
有两项独有的菜单，分别是插入和格式



工具栏

保留数据编辑窗口中的某些功能按钮
还增添了一些自己特有的功能按钮

分析结果显示区

目录区：分析结果的目录
内容区：分析结果的详细报告

系统状态显示区

显示系统的当前运行状态

✓ 结果输出窗口是SPSS的主要窗口，允许同时创建或打开多个输出窗口。
✓ 这些数据通常以SPSS输出文件的形式保存，文件扩展名为.spv。
✓ 结果窗口的主要功能是：
显示、管理SPSS统计分析结果、报表及图形。

SPSS基本运行方式

SPSS为用户提供了以下三种基本运行方式

分别适合于不同的用户和不同的统计分析要求

今天的培训内容都基于完全窗口菜单方式

完全窗口菜单方式

- 所有分析操作过程都通过菜单和按钮及对话框方式进行
- 适用于一般分析和SPSS初学者

程序运行方式

- 手工编写SPSS命令程序，一次性提交计算机运行
- 适用于大规模的分析工作和熟练的SPSS程序员

菜单程序混合运行方式

- 使用菜单的时候同时编辑SPSS程序
- 适用于熟练的SPSS程序员

/02

数据分析的一般步骤

数据分析的一般步骤

明确数据分析目标是数据分析的出发点！

1

收集数据

- 从分析目标出发，排除干扰项，正确收集数据。准确的数据对于实现数据分析目标将起到关键作用。

2

加工和整理数据

- 大致掌握数据的分布特征
- 数据预处理：缺失值处理、数据分组、数据转换等

3

数据分析

- 选择恰当的统计分析方法进行探索分析
- 明确统计方法的含义和适用范围，切忌滥用和误用统计分析方法

4

读懂结果

- 正确理解统计量的含义
- 将统计结果与实际问题相结合

利用SPSS进行数据分析的一般步骤

1

数据准备

- 定义数据结构
- 录入和修改数据

2

加工和整理数据

- 大致掌握数据的分布特征
- 数据预处理：缺失值处理、数据分组、数据转换等

3

数据分析

- 选择恰当的统计分析方法进行分析
- 数据建模.....

4

读懂结果

- 正确理解统计量的含义
- 将统计结果与实际问题相结合

/03

问卷数据分析案例

以2016年中国流动人口动态监测调查数据为例

数据样例：

http://hdl.handle.net/20.500.12291/10215

案例介绍

中国流动人口动态监测调查数据

国家卫生健康委自2009年起一年一度大规模全国性流动人口抽样调查数据，覆盖全国31个省（区、市）和新疆生产建设兵团中流动人口较为集中的流入地，内容涉及流动人口基本信息、流动范围和趋向、就业和社会保障、收支和居住、基本公共卫生服务、婚育和计划生育服务管理、子女流动和教育等。

1
- 数据结构
- 问卷录入（多选题、矩阵题）

2
- 数据预处理（变量计算、数据选取、分类汇总、数据分组）
- 基本统计分析（频数分析、基本描述统计量、多选项分析）

3
- t 检验
- 方差分析

数据结构

对每列变量及其相关属性的描述

变量名（Name）

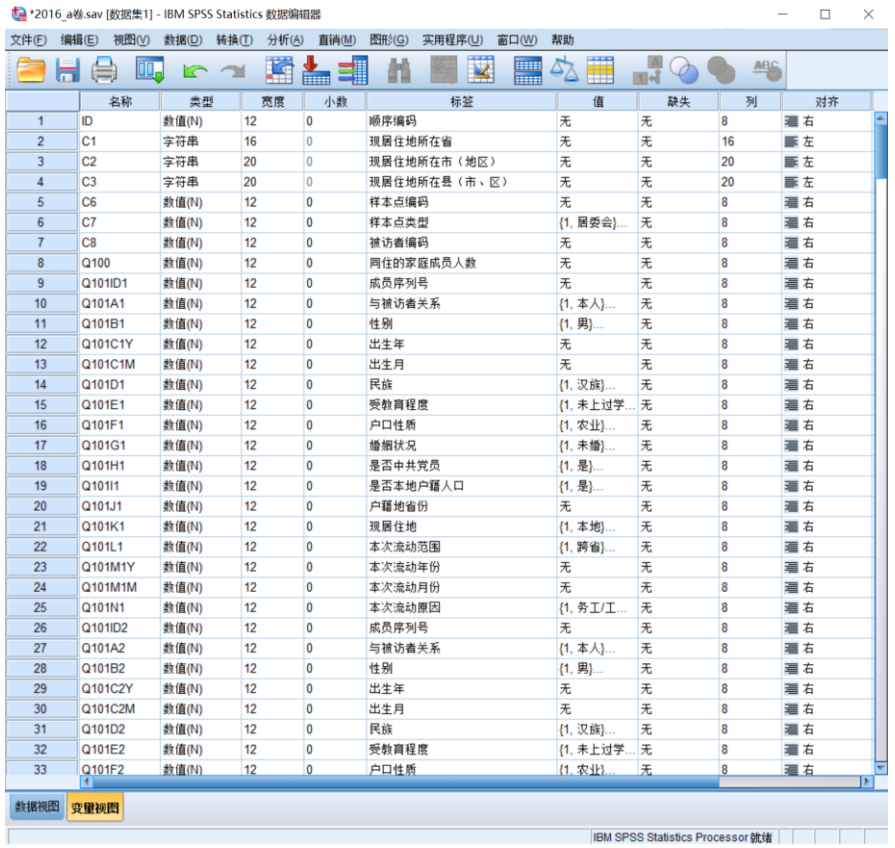
变量存取的唯一标志

变量名标签（Label）

对变量名的一些解释说明，增强分析结果的可视性和分析结果的可读性

变量值标签（Values）

对变量取值含义的一些解释说明，对定类型和定距型数据尤为重要



	名称	类型	宽度	小数	标签	值	缺失	列	对齐
1	ID	数值(N)	12	0	顺序编码	无	无	8	右
2	C1	字符串	16	0	现居住地所在省	无	无	16	左
3	C2	字符串	20	0	现居住地所在市（地区）	无	无	20	左
4	C3	字符串	20	0	现居住地所在县（市、区）	无	无	20	左
5	C6	数值(N)	12	0	样本点编码	无	无	8	右
6	C7	数值(N)	12	0	样本点类型	{1, 居委会}...	无	8	右
7	C8	数值(N)	12	0	被访者编码	无	无	8	右
8	Q100	数值(N)	12	0	同住的家庭成员人数	无	无	8	右
9	Q101ID1	数值(N)	12	0	成员序列号	无	无	8	右
10	Q101A1	数值(N)	12	0	与被访者关系	{1, 本人}...	无	8	右
11	Q101B1	数值(N)	12	0	性别	{1, 男}...	无	8	右
12	Q101C1Y	数值(N)	12	0	出生年	无	无	8	右
13	Q101C1M	数值(N)	12	0	出生月	无	无	8	右
14	Q101D1	数值(N)	12	0	民族	{1, 汉族}...	无	8	右
15	Q101E1	数值(N)	12	0	受教育程度	{1, 未上过学...	无	8	右
16	Q101F1	数值(N)	12	0	户口性质	{1, 农业}...	无	8	右
17	Q101G1	数值(N)	12	0	婚姻状况	{1, 未婚}...	无	8	右
18	Q101H1	数值(N)	12	0	是否中共党员	{1, 是}...	无	8	右
19	Q101I1	数值(N)	12	0	是否本地户籍人口	{1, 是}...	无	8	右
20	Q101J1	数值(N)	12	0	户籍地省份	无	无	8	右
21	Q101K1	数值(N)	12	0	现居住地	{1, 本地}...	无	8	右
22	Q101L1	数值(N)	12	0	本次流动范围	{1, 跨省}...	无	8	右
23	Q101M1Y	数值(N)	12	0	本次流动年份	无	无	8	右
24	Q101M1M	数值(N)	12	0	本次流动月份	无	无	8	右
25	Q101N1	数值(N)	12	0	本次流动原因	{1, 务工/工...	无	8	右
26	Q101ID2	数值(N)	12	0	成员序列号	无	无	8	右
27	Q101A2	数值(N)	12	0	与被访者关系	{1, 本人}...	无	8	右
28	Q101B2	数值(N)	12	0	性别	{1, 男}...	无	8	右
29	Q101C2Y	数值(N)	12	0	出生年	无	无	8	右
30	Q101C2M	数值(N)	12	0	出生月	无	无	8	右
31	Q101D2	数值(N)	12	0	民族	{1, 汉族}...	无	8	右
32	Q101E2	数值(N)	12	0	受教育程度	{1, 未上过学...	无	8	右
33	Q101F2	数值(N)	12	0	户口性质	{1, 农业}...	无	8	右

数据结构

对每列变量及其相关属性的描述

数据类型（Type）、列宽（Width）、小数（Decimals）

三种基本数据类型：

数值型（Numeric）、字符型（String）、日期型（Date）

相应的类型有默认的列宽或小数位宽

缺失值（Missing）

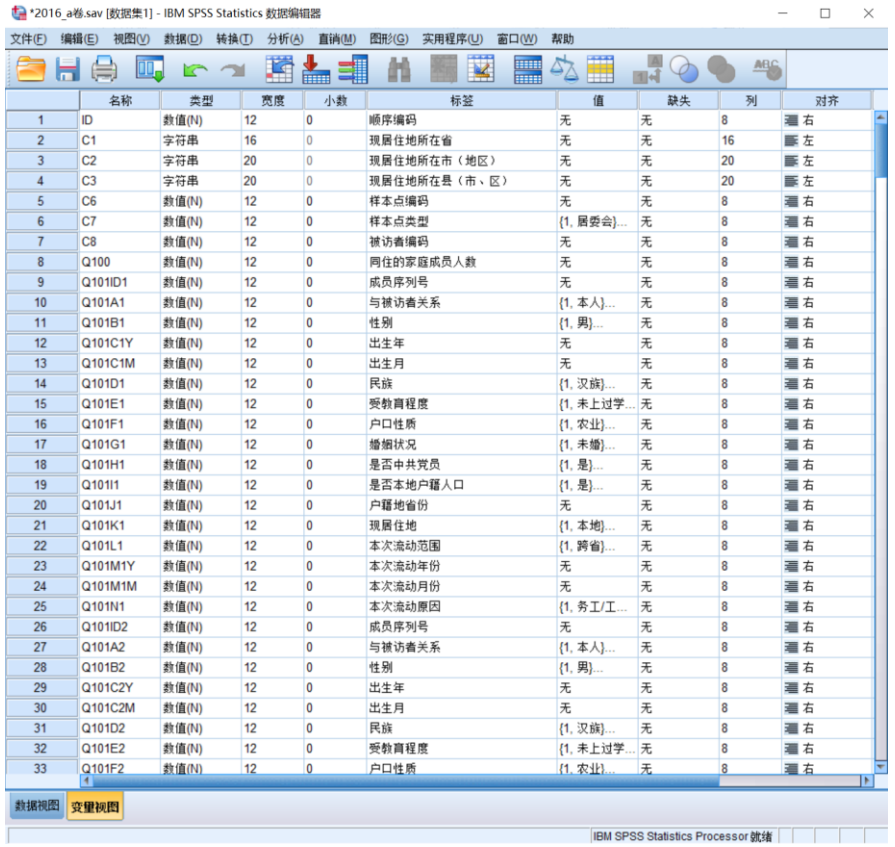
漏填数据、明显错误或不合理的数据等

用户缺失值：指定某个特定值为缺失值

系统缺失值：数值型：点（·）；字符型：空

度量标准（Measure）

Scale：定距型数据；Ordinal：定序型数据；Nominal：定类型数据（无顺序）



	名称	类型	宽度	小数	标签	值	缺失	列	对齐
1	ID	数值(N)	12	0	顺序编码	无	无	8	右
2	C1	字符串	16	0	现居住地所在省	无	无	16	左
3	C2	字符串	20	0	现居住地所在市（地区）	无	无	20	左
4	C3	字符串	20	0	现居住地所在县（市、区）	无	无	20	左
5	C6	数值(N)	12	0	样本点编码	无	无	8	右
6	C7	数值(N)	12	0	样本点类型	{1, 居委会}...	无	8	右
7	C8	数值(N)	12	0	被访者编码	无	无	8	右
8	Q100	数值(N)	12	0	同住的家庭成员人数	无	无	8	右
9	Q101ID1	数值(N)	12	0	成员序列号	无	无	8	右
10	Q101A1	数值(N)	12	0	与被访者关系	{1, 本人}...	无	8	右
11	Q101B1	数值(N)	12	0	性别	{1, 男}...	无	8	右
12	Q101C1Y	数值(N)	12	0	出生年	无	无	8	右
13	Q101C1M	数值(N)	12	0	出生月	无	无	8	右
14	Q101D1	数值(N)	12	0	民族	{1, 汉族}...	无	8	右
15	Q101E1	数值(N)	12	0	受教育程度	{1, 未上过学...	无	8	右
16	Q101F1	数值(N)	12	0	户口性质	{1, 农业}...	无	8	右
17	Q101G1	数值(N)	12	0	婚姻状况	{1, 未婚}...	无	8	右
18	Q101H1	数值(N)	12	0	是否中共党员	{1, 是}...	无	8	右
19	Q101I1	数值(N)	12	0	是否本地户籍人口	{1, 是}...	无	8	右
20	Q101J1	数值(N)	12	0	户籍地省份	无	无	8	右
21	Q101K1	数值(N)	12	0	现居住地	{1, 本地}...	无	8	右
22	Q101L1	数值(N)	12	0	本次流动范围	{1, 跨省}...	无	8	右
23	Q101M1Y	数值(N)	12	0	本次流动年份	无	无	8	右
24	Q101M1M	数值(N)	12	0	本次流动月份	无	无	8	右
25	Q101N1	数值(N)	12	0	本次流动原因	{1, 务工/工...	无	8	右
26	Q101ID2	数值(N)	12	0	成员序列号	无	无	8	右
27	Q101A2	数值(N)	12	0	与被访者关系	{1, 本人}...	无	8	右
28	Q101B2	数值(N)	12	0	性别	{1, 男}...	无	8	右
29	Q101C2Y	数值(N)	12	0	出生年	无	无	8	右
30	Q101C2M	数值(N)	12	0	出生月	无	无	8	右
31	Q101D2	数值(N)	12	0	民族	{1, 汉族}...	无	8	右
32	Q101E2	数值(N)	12	0	受教育程度	{1, 未上过学...	无	8	右
33	Q101F2	数值(N)	12	0	户口性质	{1, 农业}...	无	8	右

多选题

二分法

将多选题的每个答案设为一个变量

每个变量具有两个可能值：

是/否、有/无、选中/未选中

多重分类法

用多个变量来记录问题的答案

变量值标签一致，均包含所有选项

通常用于“最多选几项”或者“多选答案存在顺序”时使用

名称	类型	宽度	小数	标签	值
Q303A	数值(N)	12	0	您是否在本地购买住房	{1, 是}...
Q303B	数值(N)	12	0	您是否在户籍地-区县政府所在地购买住房	{1, 是}...
Q303C	数值(N)	12	0	您是否在户籍地-乡政府所在地购买住房	{1, 是}...
Q303D	数值(N)	12	0	您是否在户籍地-村购买住房	{1, 是}...
Q303E	数值(N)	12	0	您是否在其他地方购买住房	{1, 是}...

多选题

多重分类法

用多个变量来记录问题的答案

变量值标签一致，均包含所有选项

通常用于“最多选几项”或者“多选答案存在顺序”时使用

名称	类型	宽度	小数	标签	值
B2_1	数值(N)	8	0	请问您未来打算优先购买的数码产品有哪些？（最多三项）	{1, 笔记本电脑}...
B2_2	数值(N)	8	0	请问您未来打算优先购买的数码产品有哪些？（最多三项）	{1, 笔记本电脑}...
B2_3	数值(N)	8	0	请问您未来打算优先购买的数码产品有哪些？（最多三项）	{1, 笔记本电脑}...

矩阵题

219 您目前参加下列何种社会保障？

社会保障	1. 是否参保 1 是 2 否（跳问下一行） 3 不清楚（跳问下一行）	2. 在何处参保 1 本地 2 户籍地 3 其他地方
A 养老保险(含新农保、养老金等)	|__|	|__|
B 失业保险	|__|	|__|
C 工伤保险	|__|	|__|
D 生育保险	|__|	|__|
E 住房公积金	|__|	|__|

名称	类型	宽度	小数	标签	值
Q219A1	数值(N)	12	0	您是否参加养老保险(含新农保、养老金等)	{1, 是}...
Q219A2	数值(N)	12	0	您在何地参加养老保险(含新农保、养老金等)	{1, 本地}...
Q219B1	数值(N)	12	0	您是否参加失业保险	{1, 是}...
Q219B2	数值(N)	12	0	您在何地参加失业保险	{1, 本地}...
Q219C1	数值(N)	12	0	您是否参加工伤保险	{1, 是}...
Q219C2	数值(N)	12	0	您在何地参加工伤保险	{1, 本地}...
Q219D1	数值(N)	12	0	您是否参加生育保险	{1, 是}...
Q219D2	数值(N)	12	0	您在何地参加生育保险	{1, 本地}...

数据预处理

1

变量计算

- SPSS算术表达式
- SPSS条件表达式
- SPSS函数

2

数据选取

- 按指定条件选取
- 随机选取

3

分类汇总

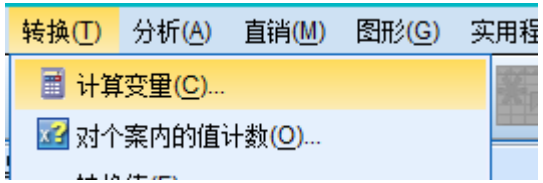
- 按照某分类变量进行分类计算
- 根据多个分类变量汇总计算时称为多重分类汇总

4

数据分组

- 根据需要将数据按照某种标准重新划分为不同的组别

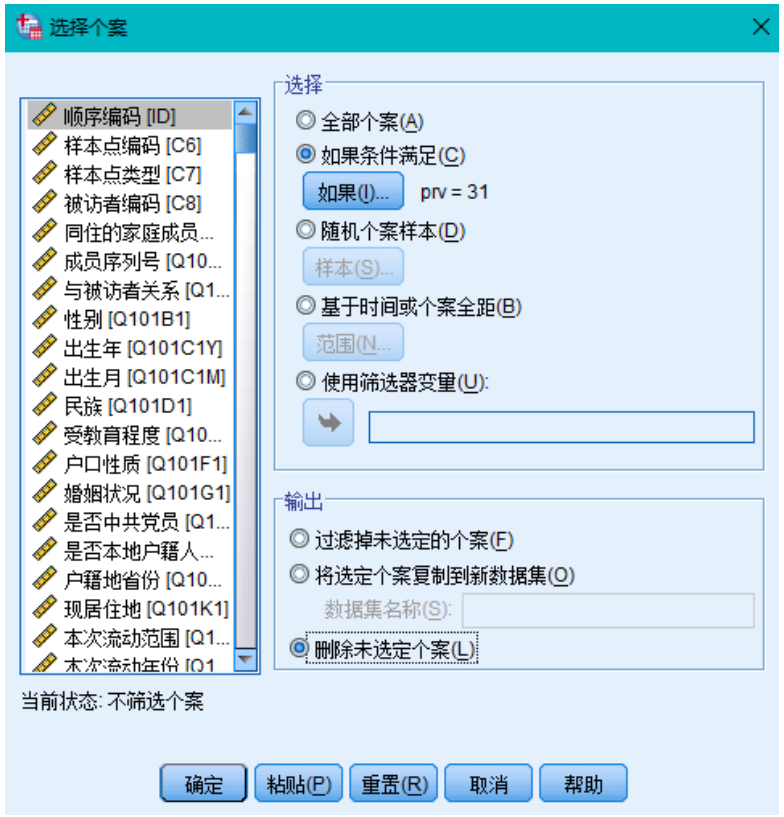
变量计算

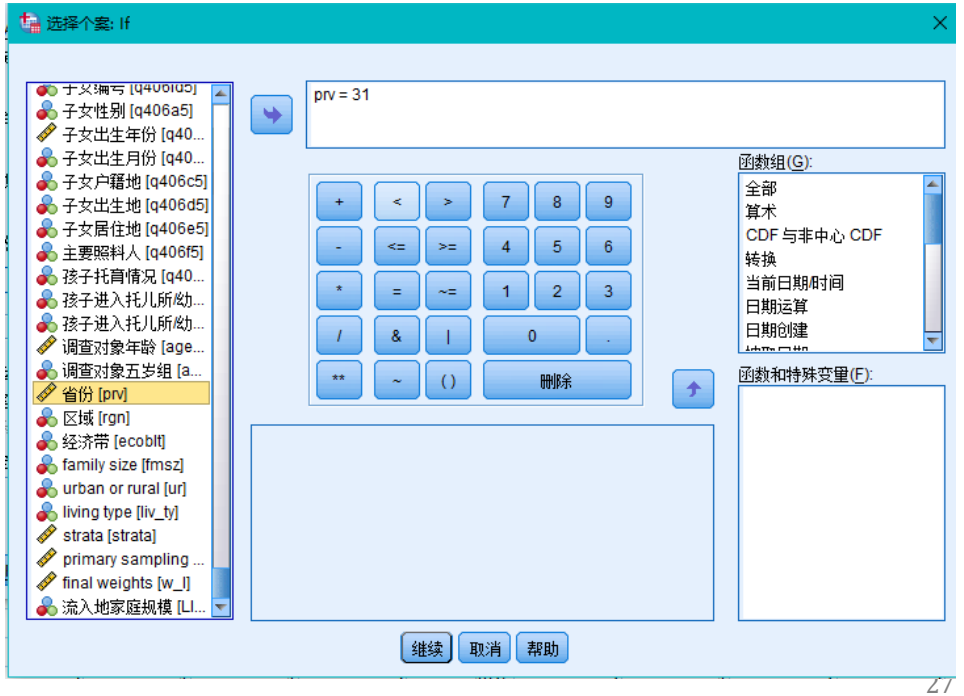


根据出生年、月计算年龄

数据选取

选取上海地区的样本



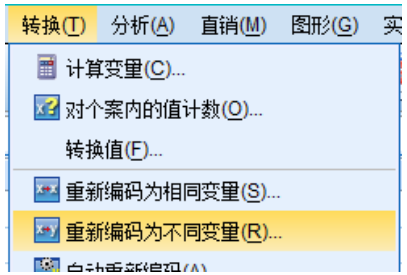


分类汇总

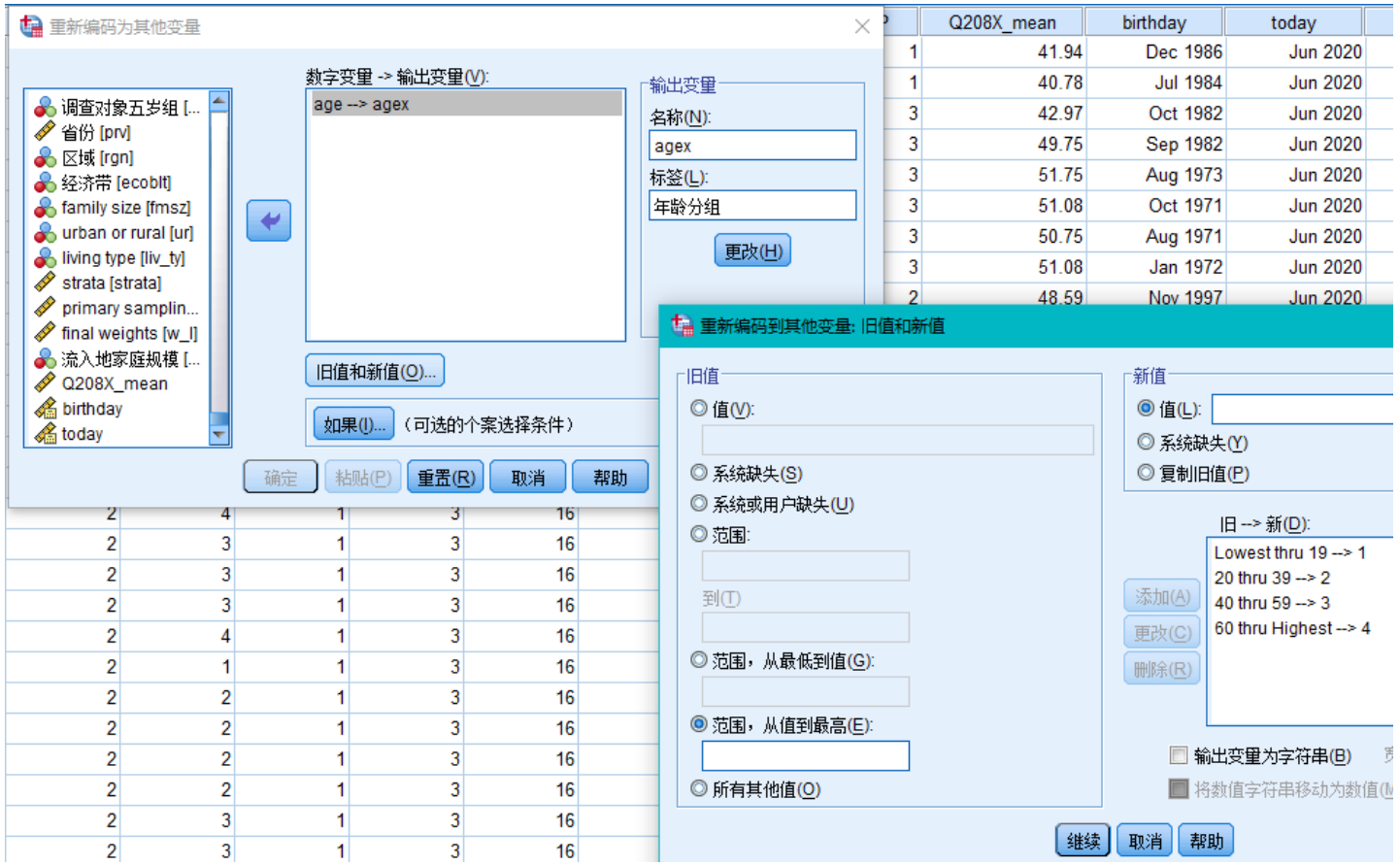
按受教育程度进行分类，计算流动人口的上月平均收入

Q101E1	平均收入	样本量
1	3324.58	117
2	3765.08	686
3	4776.13	2724
4	5423.06	1434
5	7906.62	875
6	9567.44	1019
7	11826.76	145

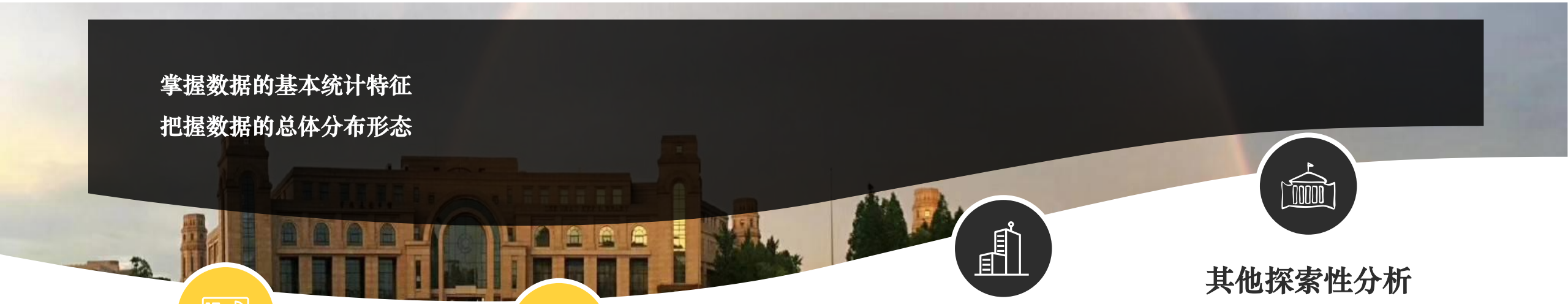
数据分组


转换(T)
分析(A)
直销(M)
图形(G)
计算变量(C)...
对个案内的值计数(O)...
转换值(F)...
重新编码为相同变量(S)...
重新编码为不同变量(R)...

年龄分组


重新编码为其他变量
调查对象五岁组 [...
省份 [prv]
区域 [rgn]
经济带 [ecoblt]
family size [fmsz]
urban or rural [ur]
living type [liv_ty]
strata [strata]
primary samplin...
final weights [w_l]
流入地家庭规模 [...
Q208X_mean
birthday
today
数字变量 -> 输出变量(V):
age --> agex
输出变量
名称(N):
agex
标签(L):
年龄分组
更改(H)
旧值和新值(O)...
如果(I)...
（可选的个案选择条件）
确定
粘贴(P)
重置(R)
取消
帮助
Q208X_mean
birthday
today
41.94 Dec 1986 Jun 2020
40.78 Jul 1984 Jun 2020
42.97 Oct 1982 Jun 2020
49.75 Sep 1982 Jun 2020
51.75 Aug 1973 Jun 2020
51.08 Oct 1971 Jun 2020
50.75 Aug 1971 Jun 2020
51.08 Jan 1972 Jun 2020
重新编码到其他变量: 旧值和新值
旧值
值(V):
系统缺失(S)
系统或用户缺失(U)
范围:
到(T)
范围，从最低到值(G):
范围，从值到最高(E):
所有其他值(O)
新值
值(L):
系统缺失(Y)
复制旧值(P)
旧 --> 新(D):
Lowest thru 19 --> 1
20 thru 39 --> 2
40 thru 59 --> 3
60 thru Highest --> 4
添加(A)
更改(C)
删除(R)
输出变量为字符串(B)
将数值字符串移动为数值(M)
继续
取消
帮助

基本统计分析



掌握数据的基本统计特征

把握数据的总体分布形态

频数分析

- 编制频数分布表
- 绘制统计图
- 交叉列联表

计算描述统计量

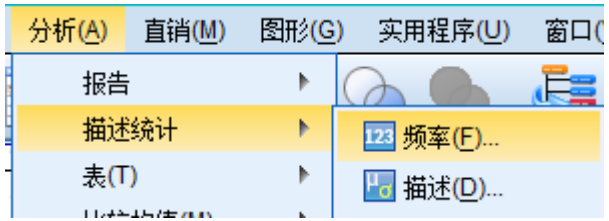
- 集中趋势：均值、中位数、众数
- 离散程度：样本标准差、样本方差
- 分布形态：偏度系数、峰度系数

多选项分析

- 按二分法或多重分类法设置变量
- 多选项频数分析
- 多选项交叉分组下的频数分析

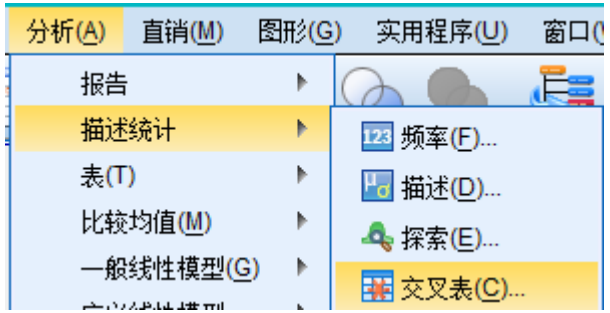
其他探索性分析

频数分析



查看就业单位性质的频数分布

交叉列联表



性别、婚姻状况交叉列联表

卡方检验

	值	df	渐进 Sig. (双侧)
Pearson 卡方	28.286[a]	5	.000
似然比	29.559	5	.000
线性和线性组合	12.218	1	.000
有效案例中的 N	7000		

a. 0 单元格(0.0%) 的期望计数少于 5。最小期望计数为 17.77。

性别* 婚姻状况 交叉制表

计数

		婚姻状况						
		未婚	初婚	再婚	离婚	丧偶	同居	合计
性别	男	365	2869	51	37	7	32	3361
	女	405	3005	89	57	30	53	3639
合计		770	5874	140	94	37	85	7000

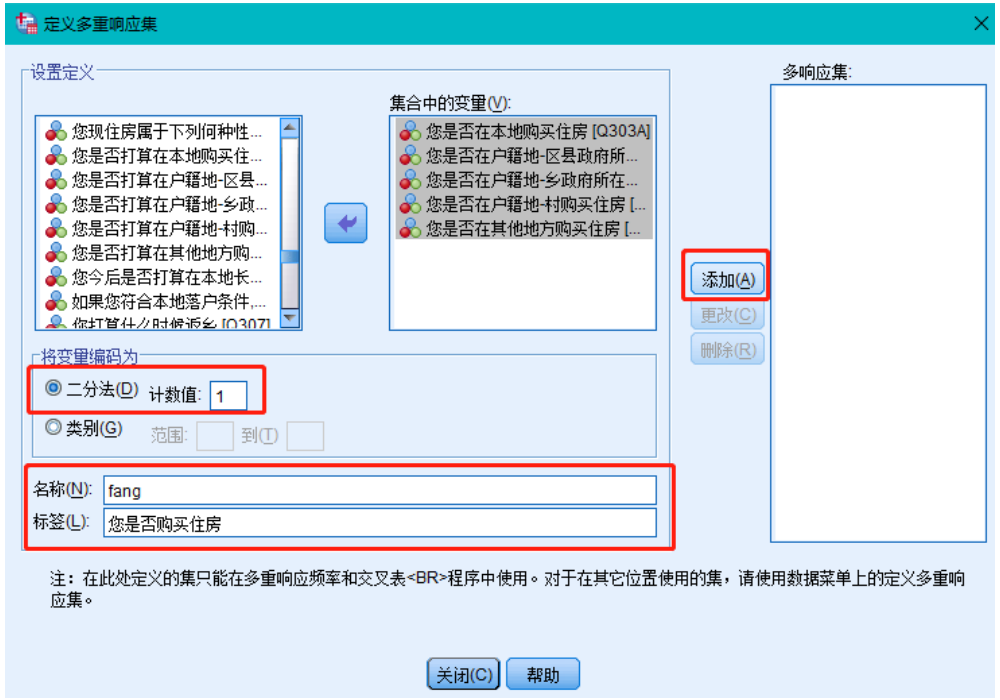
计算描述统计量


分析(A)
直销(M)
图形(G)
实用程序(U)
窗口(
报告
描述统计
表(T)
频率(F)...
描述(D)...


变量(V):
过去一年,全家在本地每月总支出多少...
选项(O)...
Bootstrap(B)...
重置(R)
取消
帮助
描述: 选项
均值(M)
合计
离散
标准差(T)
最小值
方差
最大值
范围
均值的标准误(E)
分布
峰度
偏度
显示顺序
变量列表(B)
字母顺序(A)
按均值的升序排序(C)
按均值的降序排序(D)
继续
取消
帮助

多选项分析


分析(A)
直销(M)
图形(G)
实用程序(U)
窗口(W
报告
描述统计
表(T)
比较均值(M)
一般线性模型(G)
广义线性模型
混合模型(X)
相关(C)
回归(R)
对数线性模型(O)
神经网络
分类(F)
降维
度量(S)
非参数检验(N)
预测(T)
生存函数(S)
多重响应(U)
小数
定义变量集(D)...


定义多重响应集
设置定义
集合中的变量(V):
您现住房属于下列何种性...
您是否打算在本地购买住...
您是否打算在户籍地-区县...
您是否打算在户籍地-乡政...
您是否打算在户籍地-村购...
您是否打算在其他地方购...
您今后是否打算在本地长...
如果您符合本地落户条件,...
您是否在本地购买住房 [Q303A]
您是否在户籍地-区县政府所...
您是否在户籍地-乡政府所在...
您是否在户籍地-村购买住房 [...
您是否在其他地方购买住房 [...
多响应集:
添加(A)
更改(C)
删除(R)
将变量编码为
二分法(D) 计数值: 1
类别(G) 范围: 到(T)
名称(N): fang
标签(L): 您是否购买住房
注：在此处定义的集只能在多重响应频率和交叉表
程序中使用。对于在其它位置使用的集，请使用数据菜单上的定义多重响应集。
关闭(C)
帮助


分析(A)
直销(M)
图形(G)
实用程序(U)
窗口(W
报告
描述统计
表(T)
比较均值(M)
一般线性模型(G)
广义线性模型
混合模型(X)
相关(C)
回归(R)
对数线性模型(O)
神经网络
分类(F)
降维
度量(S)
非参数检验(N)
预测(T)
生存函数(S)
多重响应(U)
缺失值分析(Y)...
多重归因(T)
小数
定义变量集(D)...
频率(F)...
交叉表(C)...

推断统计与参数检验

推断统计

- 根据样本数据推断总体特征
  总体数据无法全部收集
  采集数据需要大量投入

参数检验

- 参数检验是推断统计的重要组成部分
- 当总体分布已知（如总体为正态分布），根据样本数据对总体分布的统计参数进行推断
- 也可对两个或多个总体的参数进行比较

t 检验

假设检验的一般步骤：

- 第一步：提出原假设
- 第二步：选择检验统计量
- 第三步：计算检验统计量观测值发生的概率
- 第四步：给定显著性水平α，并作出统计决策

利用SPSS进行假设检验时，首先需明确原假设，第二步和第三步是SPSS自动完成的，最终决策需要人工判定。

单样本t检验

- 对总体均值的假设检验
- 前提：样本来自的总体应服从或近似服从正态分布
- 检验统计量为t统计量

两独立样本t检验

- 推断两个总体的均值是否存在显著差异
- 前提1：样本来自的总体应服从或近似服从正态分布
- 前提2：两样本相互独立，样本量可以不等

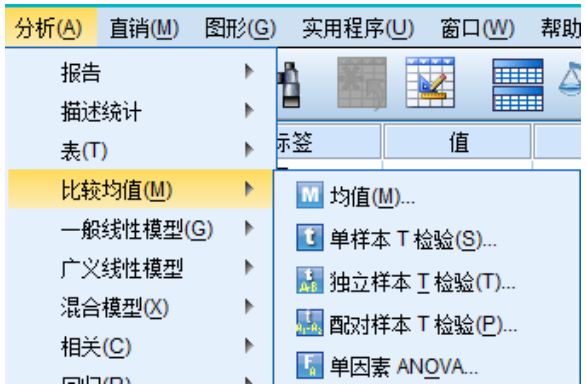
两配对样本t检验

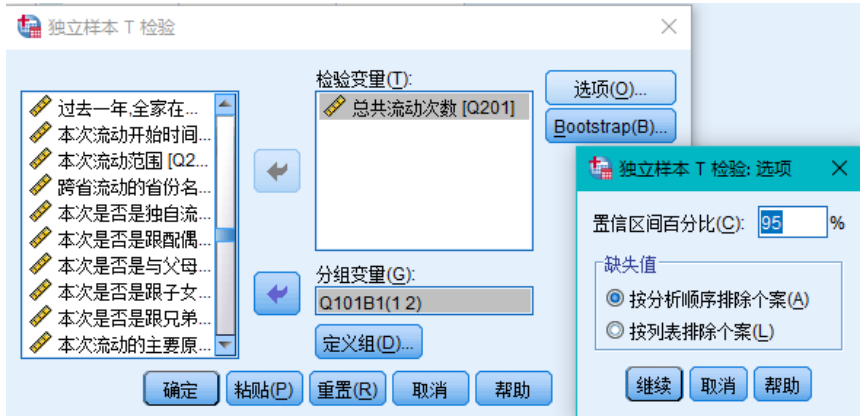
- 配对样本通常具有两个特征：

  一、两组样本量相同

  二、两组样本观测值先后顺序一一对应

t 检验





组统计量

	性别	N	均值	标准差	均值的标准误
总共流动次数	男	88087	1.40	1.193	.004
	女	80911	1.27	.811	.003

男性和女性流动人口的流动次数存在显著差异

独立样本检验

		方差方程的 Levene 检验		均值方程的 t 检验						
									差分的 95% 置信区间	
		F	Sig.	t	df	Sig.(双侧)	均值差值	标准误差值	下限	上限
总共流动次数	假设方差相等	1862.863	.000	25.833	168996	.000	.129	.005	.120	.139
	假设方差不相等			26.240	156011.381	.000	.129	.005	.120	.139

方差分析

- 通过推断控制变量各水平下各观测变量的总体均值是否存在显著差异，分析控制变量是否给观测变量带来了显著影响。
- 基本假设前提：
  - 一、观测变量各总体服从正态分布
  - 二、观测变量各总体方差相同

单因素方差分析

- 仅研究单个因素对观测变量的影响
- 检验统计量为F统计量
- 方差齐性检验
- 多重比较检验

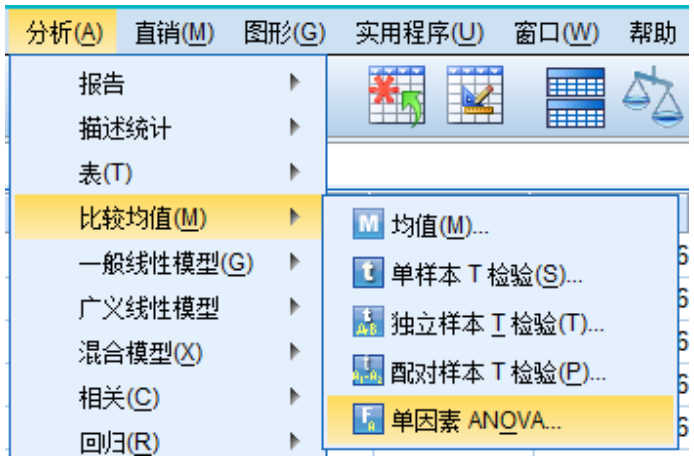
多因素方差分析

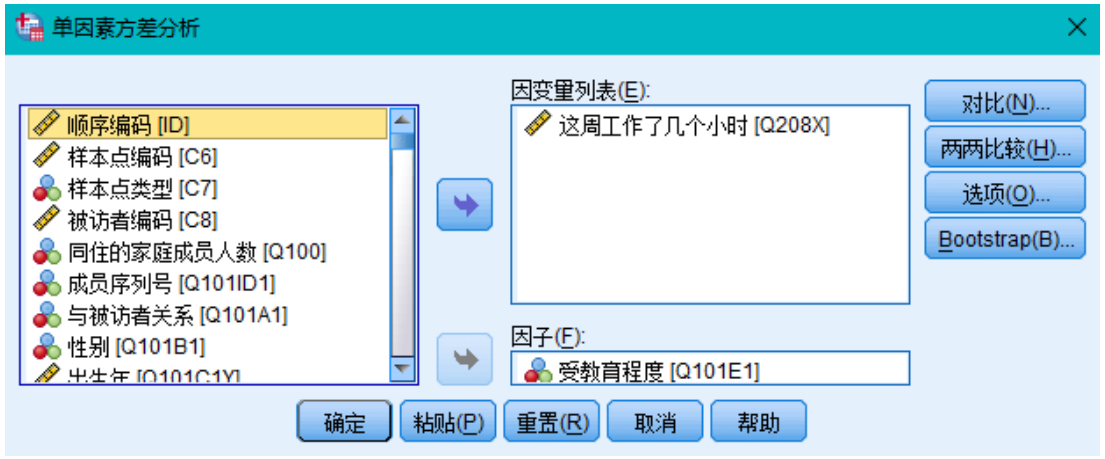
- 研究多个因素对观测变量的影响
- 控制变量独立作用的影响
- 控制变量交互作用的影响
- 随机因素的影响

协方差分析

- 协变量：人为因素很难控制的因素

方差分析


分析(A)
直销(M)
图形(G)
实用程序(U)
窗口(W)
帮助
报告
描述统计
表(T)
比较均值(M)
一般线性模型(G)
广义线性模型
混合模型(X)
相关(C)
回归(R)
均值(M)...
单样本 T 检验(S)...
独立样本 T 检验(T)...
配对样本 T 检验(P)...
单因素 ANOVA...


单因素方差分析
顺序编码 [ID]
样本点编码 [C6]
样本点类型 [C7]
被访者编码 [C8]
同住的家庭成员人数 [Q100]
成员序列号 [Q101ID1]
与被访者关系 [Q101A1]
性别 [Q101B1]
因变量列表(E):
这周工作了几个小时 [Q208X]
因子(F):
受教育程度 [Q101E1]
对比(N)...
两两比较(H)...
选项(O)...
Bootstrap(B)...
确定
粘贴(P)
重置(R)
取消
帮助


单因素 ANOVA: 选项
统计量
描述性(D)
固定和随机效果(F)
方差同质性检验(H)
Brown-Forsythe(B)
Welch(W)
均值图(M)
缺失值
按分析顺序排除个案(A)
按列表排除个案(L)
继续
取消
帮助


单因素 ANOVA: 两两比较
假定方差齐性
LSD(L)
Bonferroni(B)
Sidak
Scheffe(C)
R-E-G-W F(R)
R-E-G-W Q(Q)
S-N-K(S)
Tukey
Tukey s-b(K)
Duncan(D)
Hochberg's GT2(H)
Gabriel(G)
Waller-Duncan(W)
类型 I/类型 II 误差比率(/):
100
Dunnett(E)
控制类别:
最后一个(L)
检验
双侧(2)
< 控制(O)
> 控制(N)
未假定方差齐性
Tamhane's T2(M)
Dunnett's T3(3)
Games-Howell(A)
Dunnett's C(U)
显著性水平(F):
0.05
继续
取消
帮助

方差分析

方差齐性检验

这周工作了几个小时

Levene 统计量	df1	df2	显著性
189.919	6	5899	.000

单因素方差分析

这周工作了几个小时

	平方和	df	均方	F	显著性
组间	88044.778	6	14674.130	88.946	.000
组内	973205.502	5899	164.978		
总数	1061250.280	5905			

多重比较

因变量: 这周工作了几个小时

	(I) 受教育程度	(J) 受教育程度	均值差 (I-J)	标准误	显著性	95% 置信区间 下限	95% 置信区间 上限
LSD	未上过学	小学	1.032	1.609	.521	-2.12	4.18
		初中	2.126	1.537	.167	-.89	5.14
		高中/中专	5.760*	1.559	.000	2.70	8.82
		大学专科	9.970*	1.585	.000	6.86	13.08
		大学本科	11.202*	1.572	.000	8.12	14.28
		研究生	10.927*	1.860	.000	7.28	14.57
	小学	未上过学	-1.032	1.609	.521	-4.18	2.12
		初中	1.094	.606	.071	-.09	2.28
		高中/中专	4.729*	.660	.000	3.43	6.02
		大学专科	8.938*	.720	.000	7.53	10.35
		大学本科	10.170*	.691	.000	8.82	11.52
		研究生	9.895*	1.211	.000	7.52	12.27
	初中	未上过学	-2.126	1.537	.167	-5.14	.89
		小学	-1.094	.606	.071	-2.28	.09
		高中/中专	3.635*	.460	.000	2.73	4.54
		大学专科	7.844*	.542	.000	6.78	8.91
		大学本科	9.076*	.503	.000	8.09	10.06
		研究生	8.801*	1.114	.000	6.62	10.99

受教育程度对工作时长是否产生显著影响

Thanks

復旦大學 大数据研究院
人文社会科学数据研究所
Institute for Humanities and Social Science Data