交通数据处理与分析技术

邵长桥

第一章 概述

我们的问题？

在日常科研活动中，我们是不是曾经有过这样的困惑？

1、数据与信息

1）数据与信息的区别和联系

- 数据定义
- 信息的定义
- 数据与信息的联系与区别

2）交通数据处理与交通信息

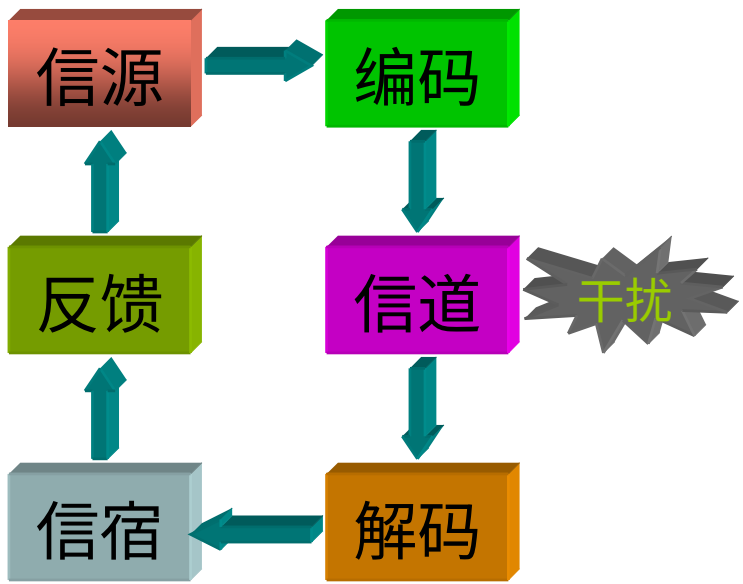
数据

- 数据
  - 对事实、概念或指令的一种特殊表达形式（约定的符号）
  - 能够被人工或自动化装置处理
  - 获取数据：测量、实验、调查

- 数据类型
  - 字符、图片、图像、动画、声音
  - 这是一个广义的定义

信息

- 信息
  - 有用的、能够影响人的行为的数据。
  - 信息要具备两个条件：
    - 可以传递
    - 能被通信双方所理解
  - 获取信息：消息、知识、加工数据
- 信息处理
  - 分类、比较、运算、推理等。



Shannon的信息传播理论

信息与数据区别和联系

- 信息是一种“数据”
  - 数据挖掘 ⟷ 信息挖掘
  - 数据融合 ⟷ 信息融合
- 信息是数据的一种抽象和概括
- 数据也可以看作一种信息，但只有加工处理后的数据，其中蕴含的信息才能容易被识别和揭示

- 获取信息的一种途径
  - 对数据进行加工处理和概括。

交通数据与信息

- 交通数据
  - 与交通行为有关的一切数据都可以纳入交通数据的范畴；
- 交通数据中往往含有一定的信息，这些信息可以是交通流自身运行的特征和规律，也可以是人们的交通行为或经济行为在交通活动中的反应。
- 因此，通过交通数据处理和分析可以挖掘出交通运行的信息、人们行为和交通系统的互动关系。

2、交通工程中数据分析的目的

- 直接目的
  - 1）交通信息（获取）
  - 2）交通预测
- 最终目的：
  - 为交通规划、交通管理和交通决策服务。

- １）交通信息分析（获取）
  - 交通信息分析一般是通过对交通因素或影响交通的其他因素进行分析，为道路使用者和交通管理部门提供交通、道路等方面的信息，为交通规划、管理和决策服务；
- 例如
  - 通过分析某路段上车辆（自由）行驶速度分布特征（如：85%位车速）确定限速值；
  - 分析交叉口事故发生的主要影响因素，通过改善措施或管理措施，来减少事故发生率；
  - 通过分析城市中出租车运行特性，发现其中存在问题，制定相应管理措施

- 信息获取手段（工具）
  - 描述性分析：通过简单的汇总和处理得到交通变量分布特征。如反映总体水平的均值、中位数；反映分散程度的方差、标准差。
  - 聚类分析：把具有相同（近）属性的交通因素归为一类，例如根据车辆在运行过程中对交通流的影响，进行聚类分析，提出了当量车概念
  - 相关性分析：分析多个交通因素之间相互影响关系，探寻其中是否存在因果关系

- 2） 预测
- 因果关系分析
  - 通过分析交通变量之间的因果关系，通过控制“因”，来预测“果”；常见的方法有：回归分析法、趋势预测等
- 时间序列分析
  - 交通是动态的，通过时间序列分析，可以发现交通因素随时间变化特性，来预测未来某天某个时段交通状况
- 交通预测
  - 顾名思义就是预测将来交通状态。这也可以看作是数据分析在交通工程中的应用，
  - 交通预测技术有许多种，除了回归预测方法和时间序列方法外，还有离散选择模型等方法

- 以上几种分析目的是泛泛而谈的。
- 但对一个具体的问题，进行数据分析是应该有着明确的目的的。

3、数据分析的任务

- 目的驱使下的数据分析
  - 很多情况下的数据分析是在明确了研究目的的情况下进行的（例如，在进行数据调查之前，就已经明确了研究目的，从而调查的数据是围绕着研究目的（是否支持你的结论）进行的，数据分析也是在此目的驱使下进行的）。
  - 数据调查之前要想到数据如何分析，数据分析要考虑到数据是如何调查的。

- 目的“不明确”的数据分析：
  - 探索性数据分析，该类分析就是“用数据说话”。在分析之前没有明确的结论。但探索性数据分析，可以为进一步用数学模型分析奠定基础（指明方向）。

- 分析任务可以分为以下几种

- 1） 分析变量的分布和变化特征（探索性分析）
  - 如对北京居民出行OD的分析
  - 如分析北京市机动车发展趋势

- 2）分析变量之间的关系
  - 如交通三参数之间的关系
  - 旅行时间和交通量、道路几何条件之间的关系
  - 交通需求与社会经济等之间的关系

- 3） 以恰当的方式体现信息
  - 解释或表达出数据中含有的信息

- 4）应用信息
- 制定规划方案、制定管理措施等
  - 制定规范

4、数据分析中的几个问题

- 1）、数据类型
  - 定性变量（属性变量）
  - 数值性变量
- 2）、数据质量
  - 噪音
  - 污染
  - 错误
  - 缺失

- 数据类型(data type)
  - 根据数据属性
  - 名义型的(nominal)：如设备ID，车道位置。
    - 没有实际上的数据大小含义
    - 可用于表示一种偏好或选择
  - 数值型的(numerical：速度，流量等
  - 顺序型的(order)：大，中，小等
    - 表示序列、等级。

- 数据定义：对事实、概念或指令的一种特殊表达形式

设备ID号	地点	车道位置	开始时间	结束时间	流量	速度	路面
2048	卫昆桥	2	07:00	07:05	43	042 mph	Dry
2048	卫昆桥	2	07:01	07:10	29	039 mph	Dry
2048	卫昆桥	2	07:02	07:11	48	040 mph	Dry
2048	卫昆桥	2	07:03	07:12	37	044 mph	Dry
2048	卫昆桥	2	07:04	07:13	47	041 mph	Dry
2048	卫昆桥	2	07:05	07:14	40	042 mph	Dry

数据类型对数据分析的影响

1）不同的数据类型，其分析方法是不一样的

属性变量分析方法：离散数据分析

如可以求分布，但其均值、方差等统计量没有实际意义。

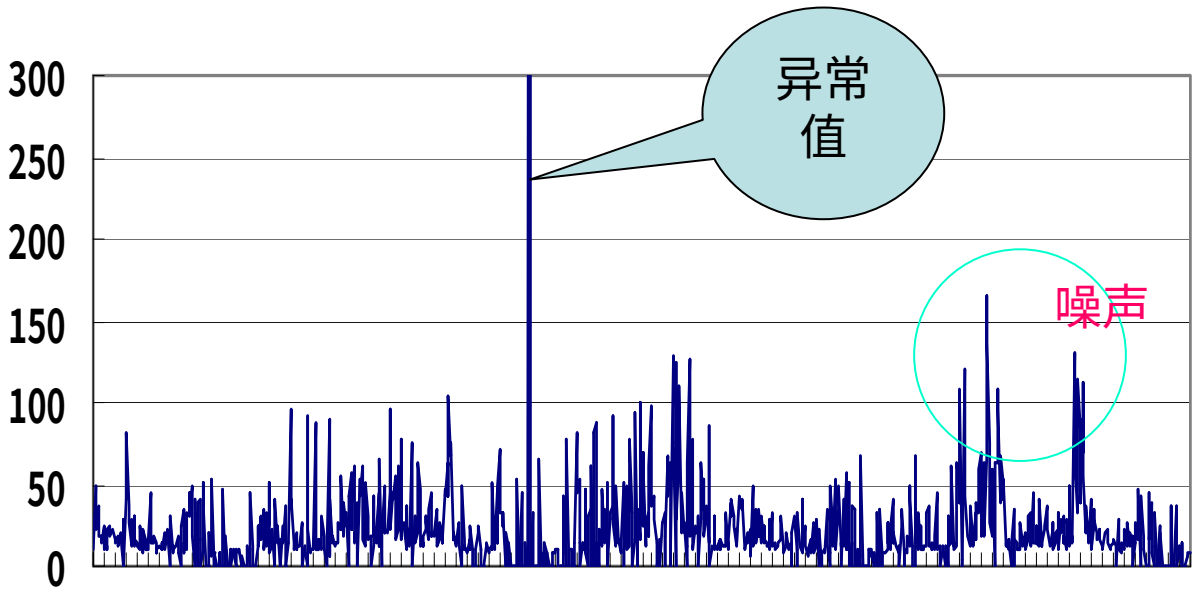
数值性变量分析方法：一般数据分析

可求均值、方差等

2）属性变量一般采用重新定义的方法

3）数据变量可以根据需要采用数据变换的方法

- 数据质量(data quality)
  - 噪音（noise):各种干扰等引起的数据和真实值有偏差
  - 错误（人为或仪器故障）
  - 污染（数据异常outlier)）
  - 缺失（missing value)


300
250
200
150
100
50
0
异常
值
噪声

5、数据予处理

- 数据予处理
  - 去污（data cleaning)：
    - 缺失值
    - 异常值
    - 噪音
  - 数据整合：数据可能来自不同的统计资料
  - 冗余处理：剔除重复或多余信息
  - 数据变换

数据清理（data cleaning)
—缺失值

- 缺失值（missing value)
  - 缺失数据模式
    - 单一变量缺失
    - 联合观测变量中，某个或某些变量观测值缺失
  - 数据缺失机制
    - 完全随机缺失：缺失部分不依赖于观测部分
    - 不完全随机缺失：

- 缺失值的处理
  - 忽略缺失记录
    - 完全随机缺失情况下，忽略缺失记录
    - 不完全随机缺失情况下，如果观测样本足够大，对于有少量缺失数据的情况下，可以忽略缺失记录。

– 填充方法

- 填充的方法就是对确实部分用一定的数值补充上。对于填充方法很多。但一般常用估计值（或预测值）代替。
- 对于时间序列数据，常用其邻近几个观测值的平均值填充，例如用前后邻近平均值:

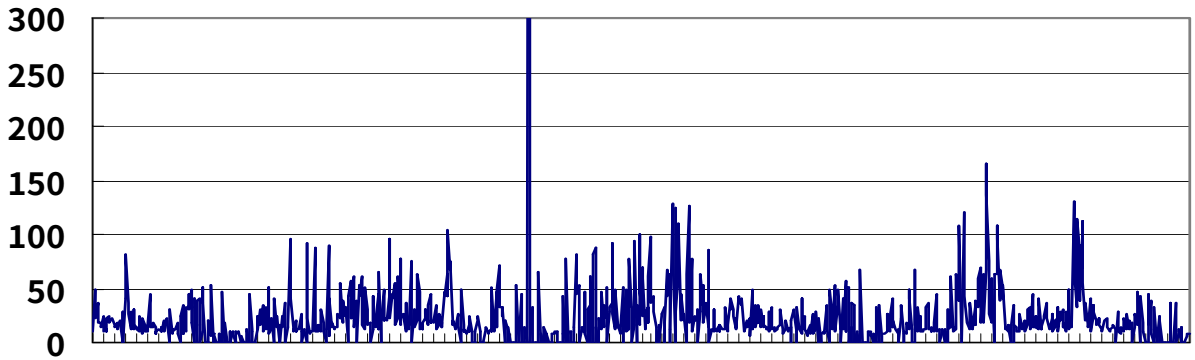
$$x_i' = \frac{x_{i+1} + x_{i-1}}{2}$$

- 对于多元变量的情况，常用回归预测值填充。

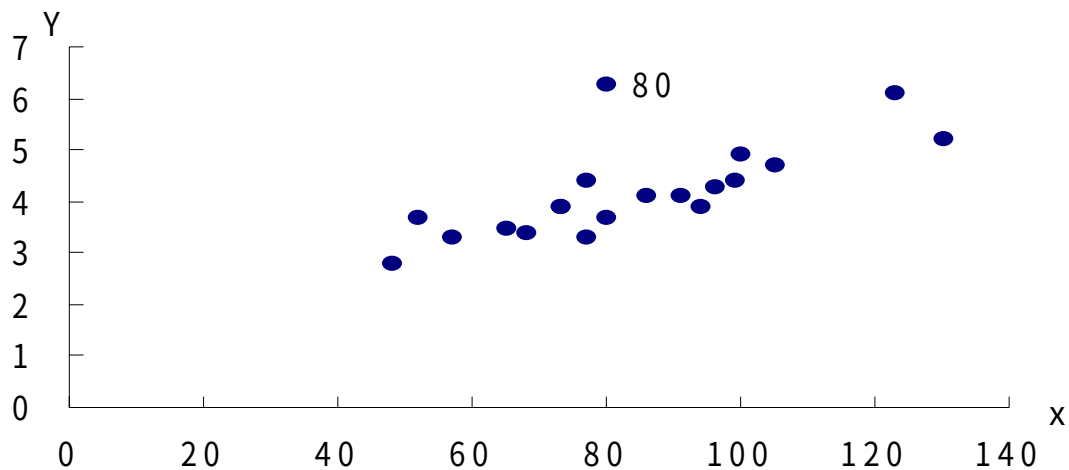
数据清理（data cleaning)
—异常值

- 异常值（outlier)
  - 定义：
    - “不一致(inconsistence）”“不和谐”，“离群”
  - 异常值的诊断
    - 图形法：趋势图、散点图、残差图等
    - 概率判别法

图形法——直接观测异常点的情况



时间序列



散点图法

概率方法

- 假设 $X_1, X_2, \cdots, X_n$ 是来自正态分布的一组观测值

- 2σ 或 3σ 检测法

$$P(|x-u|>2S_0)=0.05$$

$$P(|x-u|>3S_0)=0.003$$

- T 检验法

$$K=\frac{|x_{n+1}-\bar{x}|}{S_x}$$

- 其他方法
  - 箱形图法：

- 异常值的处理
  - 方法和缺失值处理相同

6、数据变换

- 为什么要数据变换？
  - 数据不能直接满足分析的需要
    - 数据不具有对称性
    - 对数据拟合的简单模型含有大而偶然的方差
    - 数据不能满足分析方法：（如方差分析中的正态性假设）

- 为具有稳定的方差
  - 多元回归分析为例

- 为直线性变换
  - 1）解释更容易
  - 2）容易检测出对于拟合的偏离
  - 3）容易内插和外推
- 例如

- 增强可解释性
  - 例如：多元线性回归分析中的标准化变化
  - 通过标准化变化，可以分析出每个自变量对因变量的影响的大小（消除自变量量纲的影响）。

- 例如，影响一个家庭出行次数的变量有收入、年龄和家庭人口等，显然这些变量的测量单位是不一样的。为了比较不同的变量对出行次数的影响程度，则需要消除测量单位的影响。一个常用方法就是标准化处理。

- 常用数据变换
  - 标准化 $x_i' = \frac{x_i - \bar{x}}{s_x}$
  - 对数变化 $x_i' = \ln x_i$
  - 其他变换
    - 幂变换
    - 指数变换

- 常用正态性变换方法
- （1）观察值为计数数据，其可能服从泊松分布，可考虑平方根变换：

$$Z = \sqrt{X}$$

- 或

$$Z = \sqrt{X + 1}$$

- 或

$$Z = \sqrt{X + 0.5}$$

- （2）观察值为比值时，其分子部分可能服从二项分布，此时可以考虑反正弦变换：

$$Z = \arcsin\sqrt{X}$$

- 3）观察值服从对数正态分布，可用对数变换：

$$Z = \ln X$$

补充-数据精确要求（保留小数）

名称	单位	保留小数位数
交通量	辆/h	0
	辆/m	1
	辆/s	2
速度	km/h	0
	m/min	2
密度	辆/km	1
车头时距	s	2

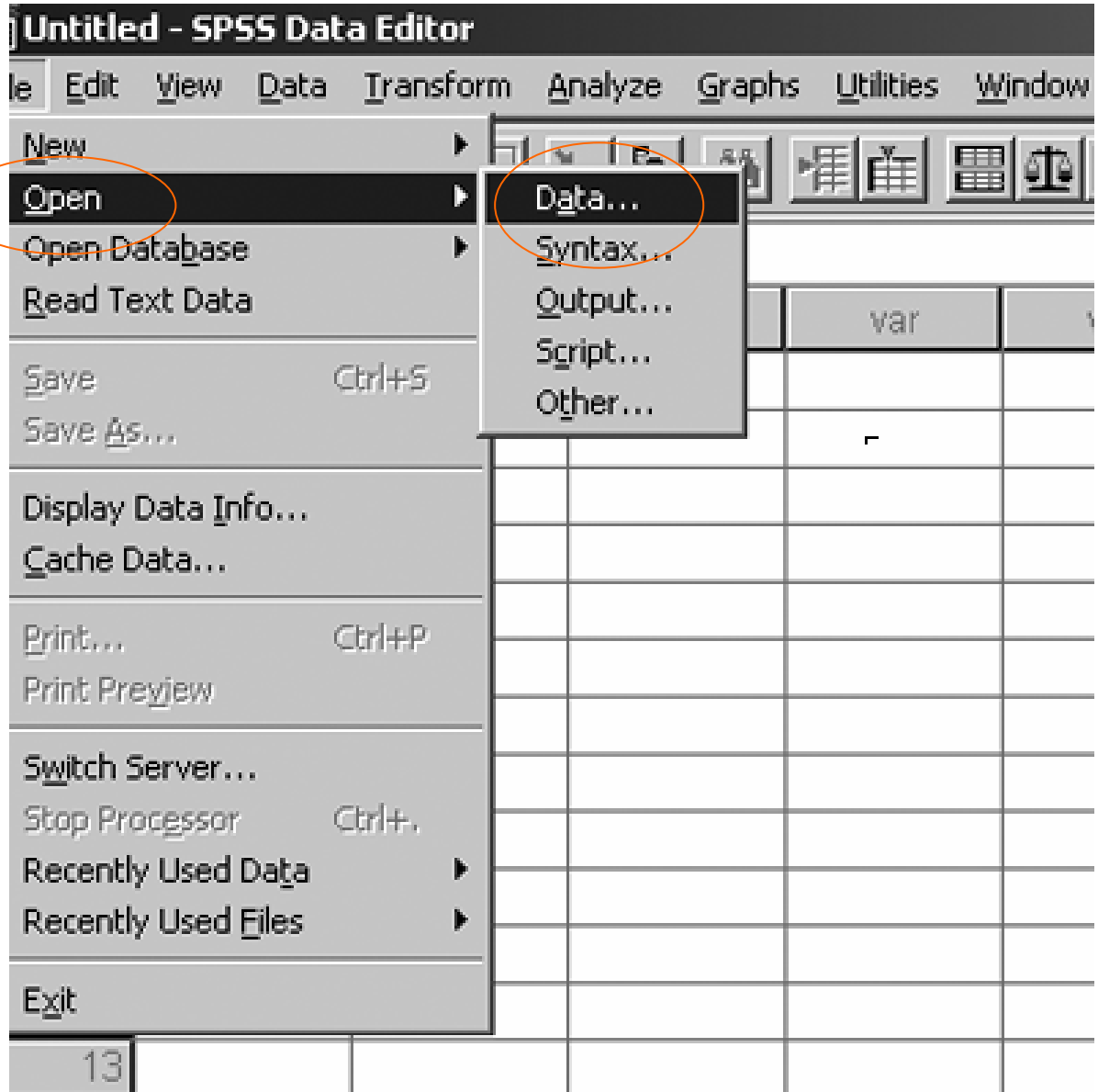
小结

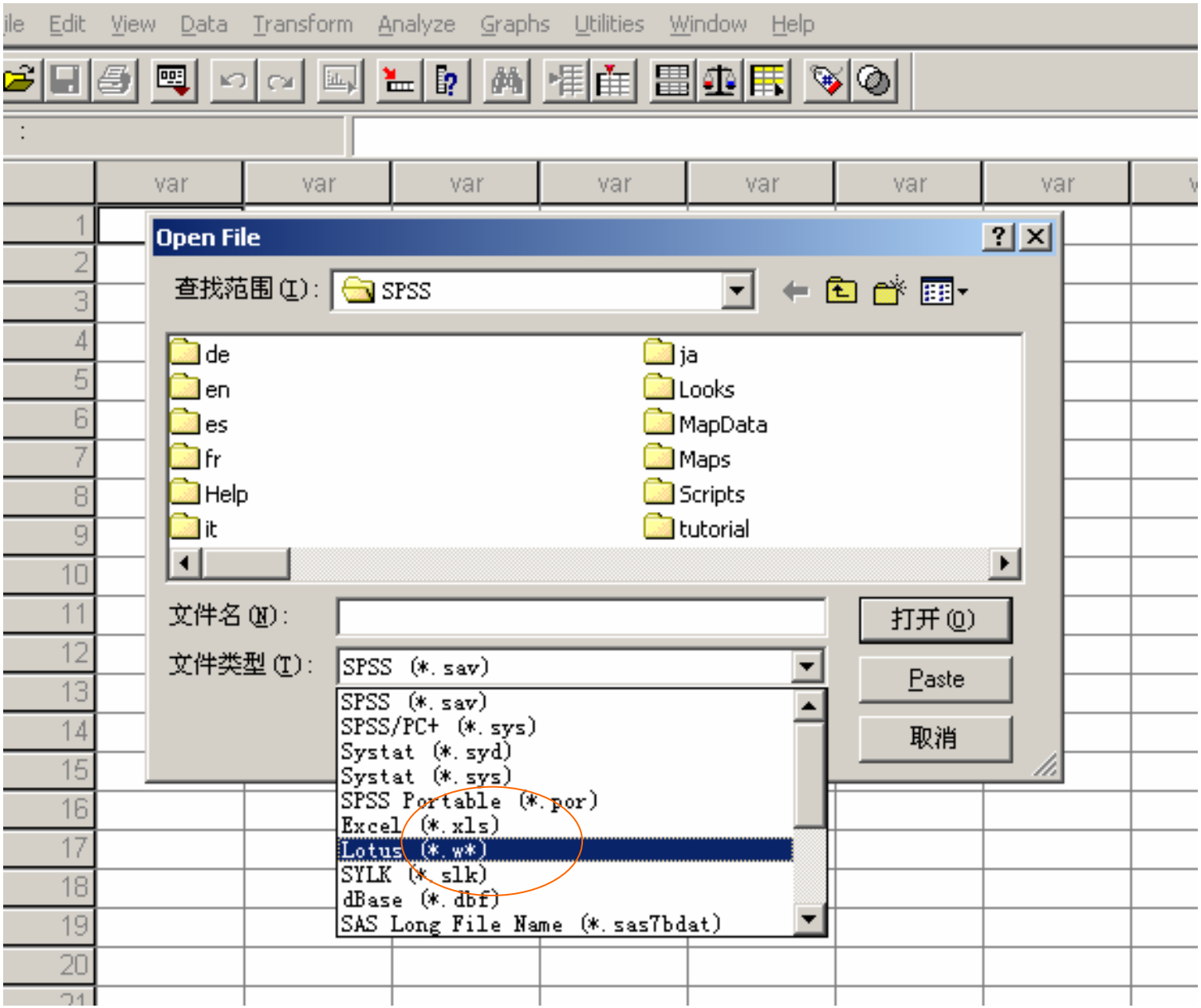
- 主要介绍了以下几个内容：
  - 在数据分析之前，要明白分析的目的：是描述性分析，还是预测性分析
  - 明确分析类型（数据）：是单个变量？还是多个变量？是时间序列变量还是其他
  - 数据予处理（不要忽略背景分析）
  - 选择分析方法

应用SPSS分析数据

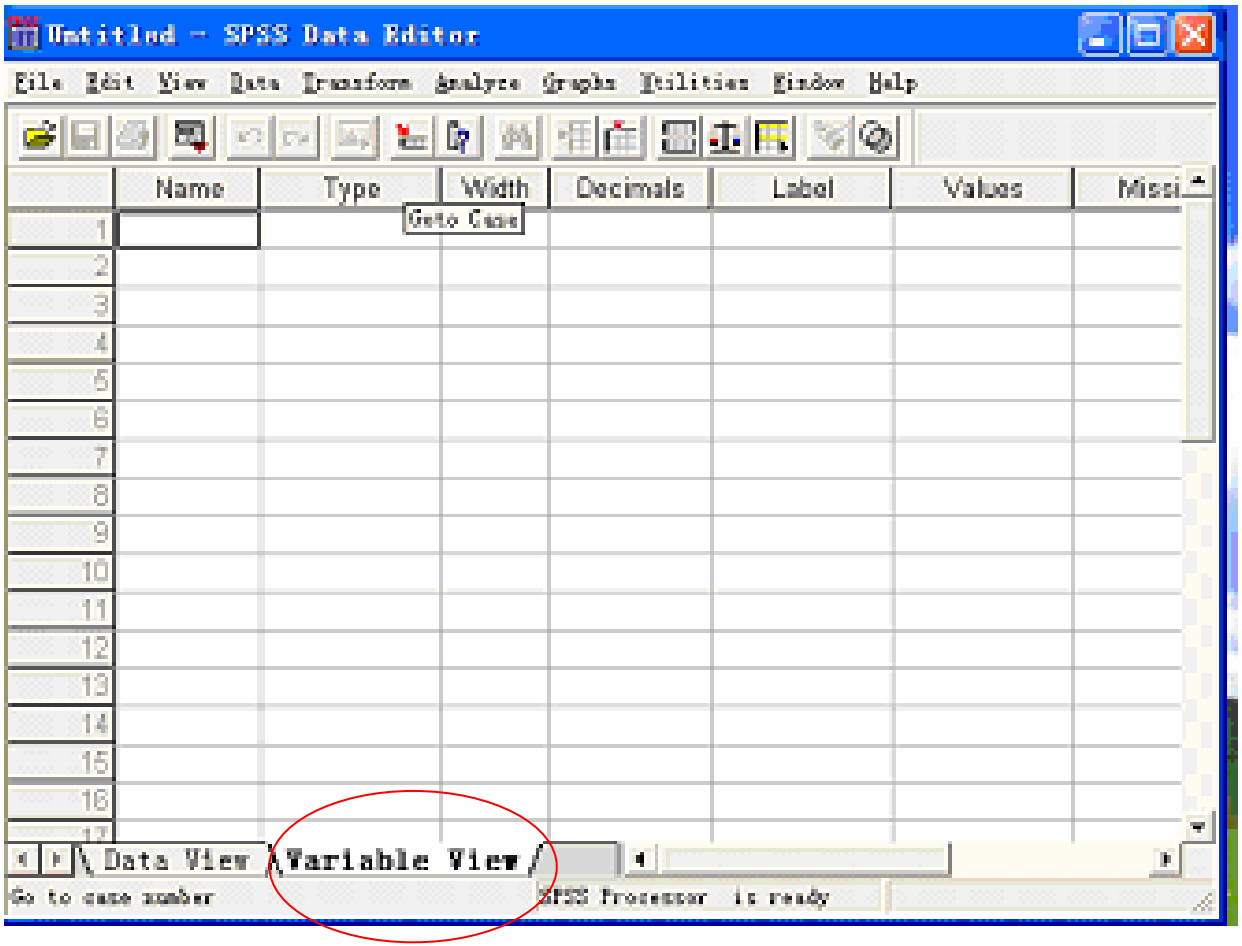
- 目的：
- 结合所讲课程，简单介绍部分应用，希望抛砖引玉，大家自己学习应用SPSS软件
- 只讲解部分内容，结合了SPSS，因此，介绍的是片面的，相关内容大家自己课下自己尝试。

打开已有数据文件

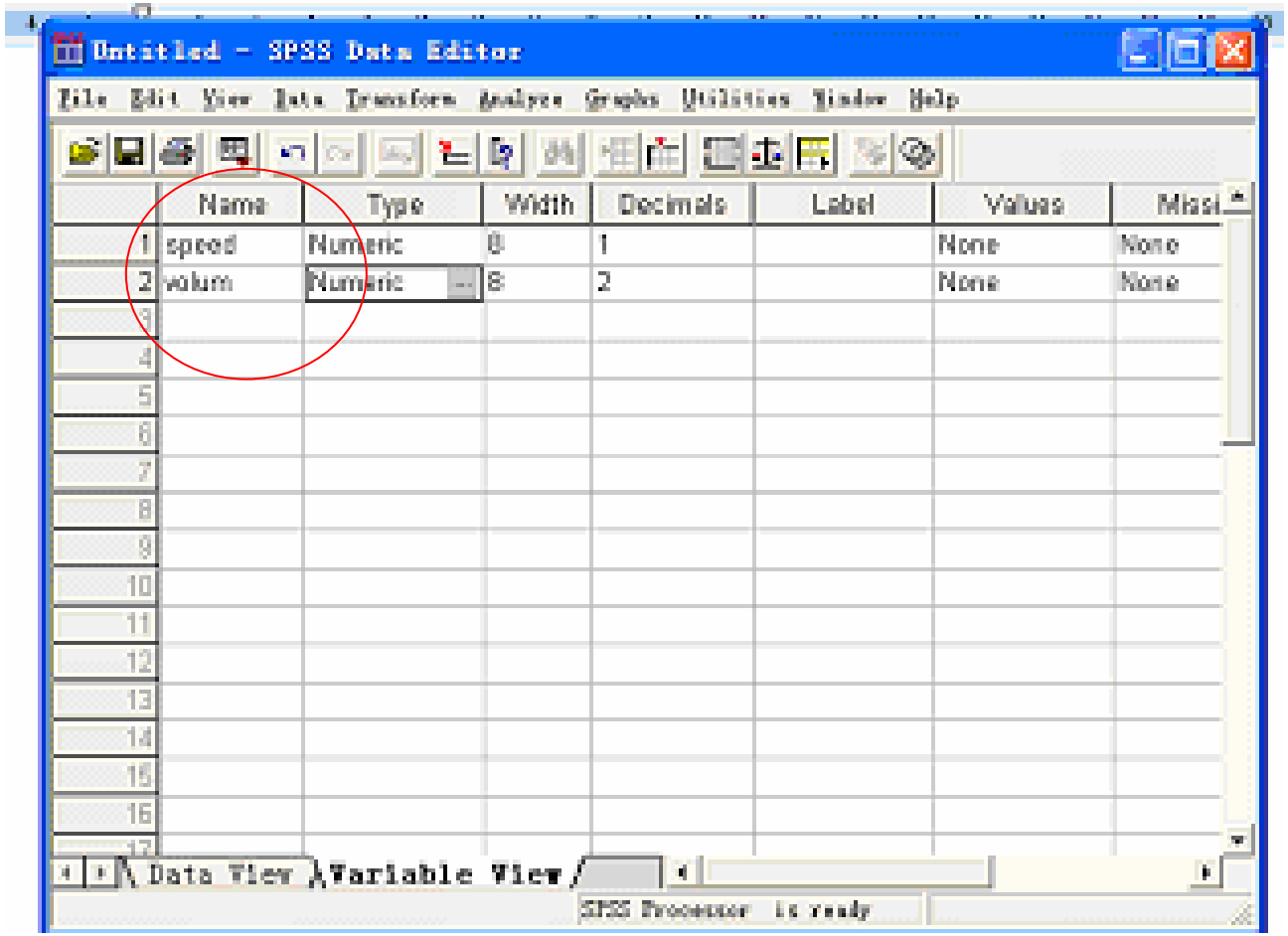

Untitled - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window
New
Open
Open Database
Read Text Data
Save Ctrl+S
Save As...
Display Data Info...
Cache Data...
Print... Ctrl+P
Print Preview
Switch Server...
Stop Processor Ctrl+.
Recently Used Data
Recently Used Files
Exit
Data...
Syntax...
Output...
Script...
Other...
var
13


File Edit View Data Transform Analyze Graphs Utilities Window Help
var
Open File
查找范围(I):
SPSS
de
en
es
fr
Help
it
ja
Looks
MapData
Maps
Scripts
tutorial
文件名(N):
文件类型(T):
SPSS (*.sav)
打开(O)
Paste
取消
SPSS (*.sav)
SPSS/PC+ (*.sys)
Systat (*.syd)
Systat (*.sys)
SPSS Portable (*.por)
Excel (*.xls)
Lotus (*.w*)
SYLK (*.slk)
dBase (*.dbf)
SAS Long File Name (*.sas7bdat)

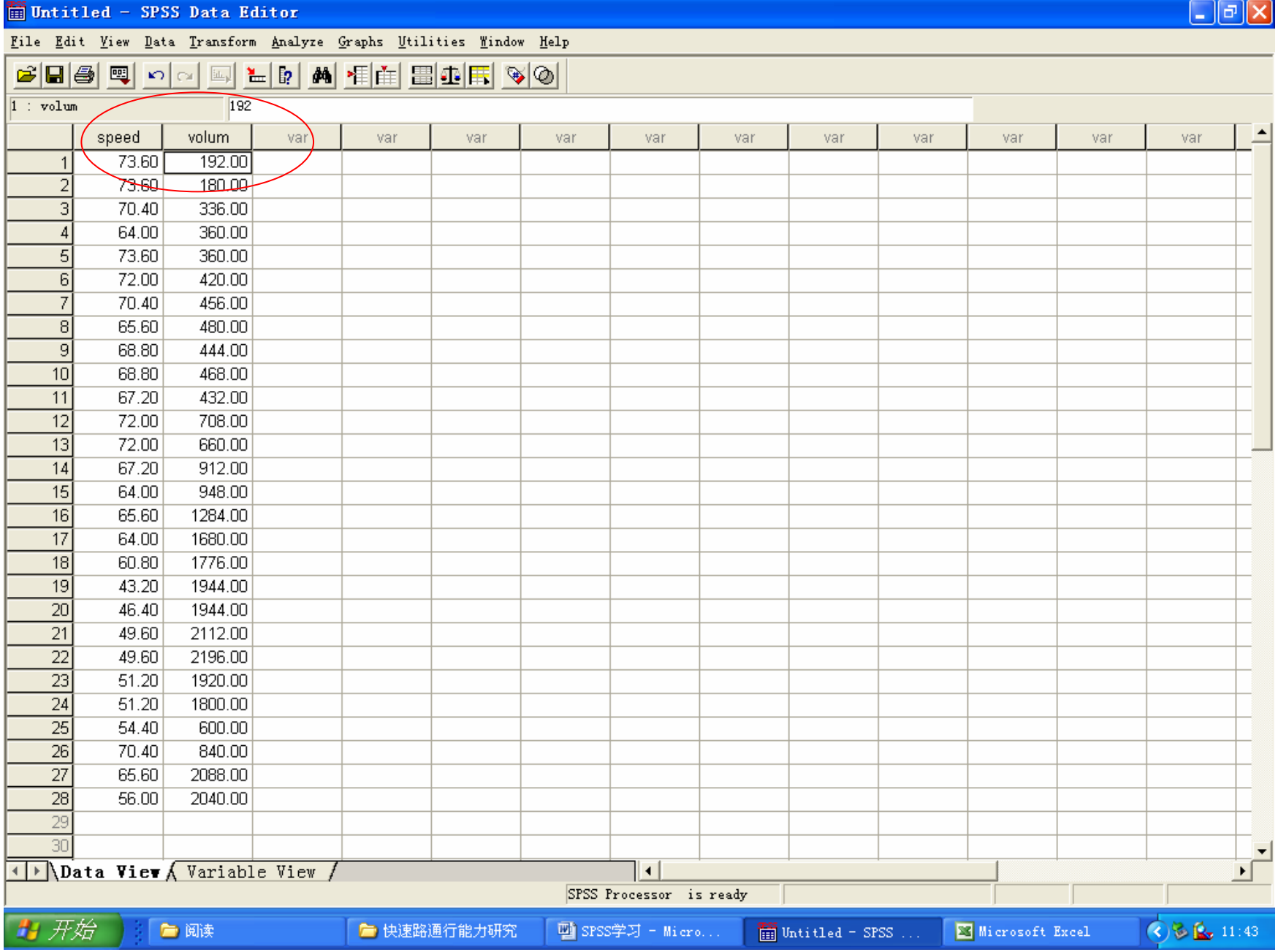
新建数据文件



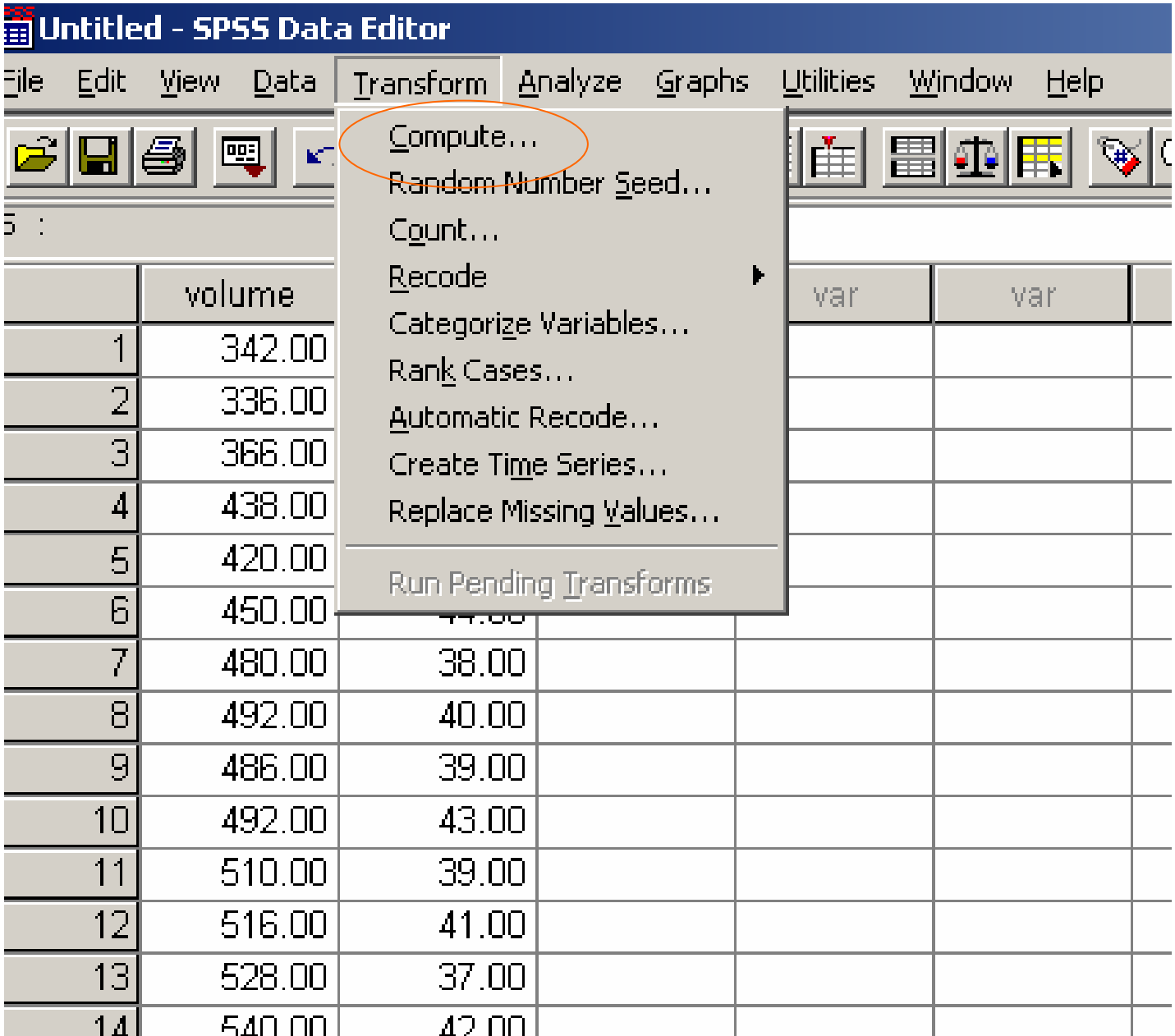
在File 下拉菜单里选择“新建”，可得到上图显示，然后，单击”Variable View”

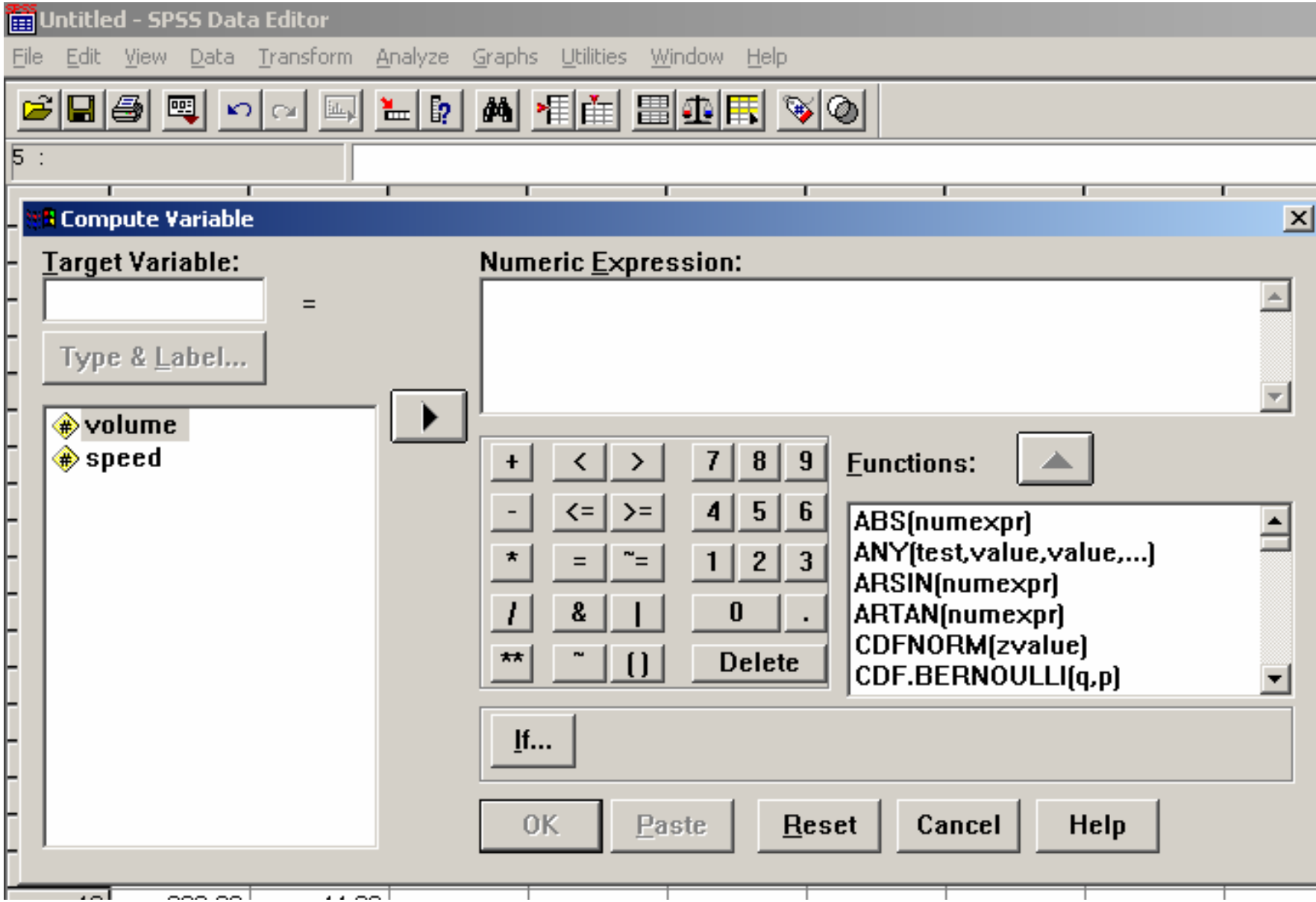


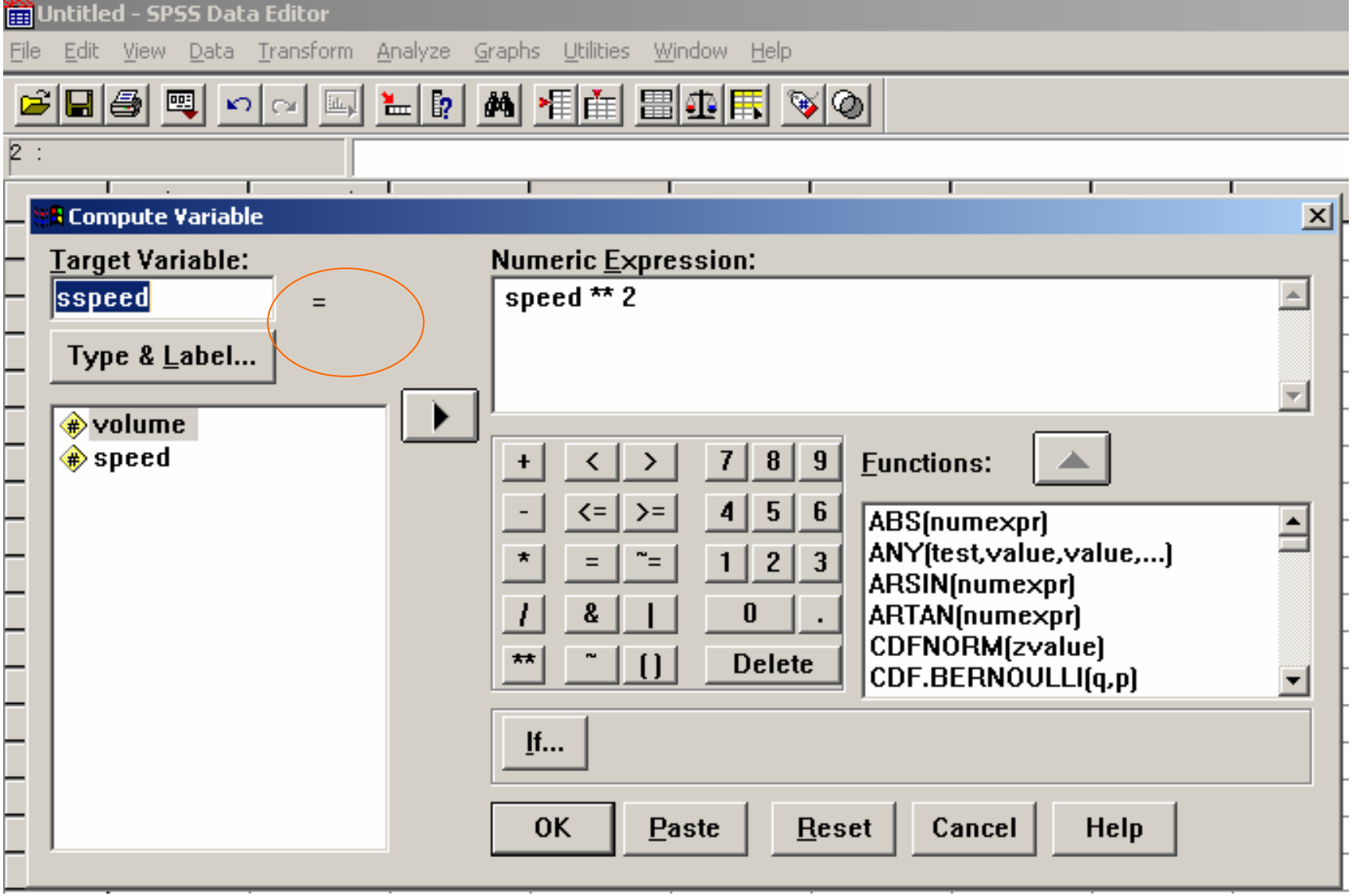
在Variable view 中输入要定义变量的名字，数值类型中选择输入变量类型（数值型或字符型等）。如定义两个数值型变量；速度（speed）和流量(volum)。


Untitled - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
1 : volum
192
speed volum var var var var var var var var var var var var
1 73.60 192.00
2 73.60 180.00
3 70.40 336.00
4 64.00 360.00
5 73.60 360.00
6 72.00 420.00
7 70.40 456.00
8 65.60 480.00
9 68.80 444.00
10 68.80 468.00
11 67.20 432.00
12 72.00 708.00
13 72.00 660.00
14 67.20 912.00
15 64.00 948.00
16 65.60 1284.00
17 64.00 1680.00
18 60.80 1776.00
19 43.20 1944.00
20 46.40 1944.00
21 49.60 2112.00
22 49.60 2196.00
23 51.20 1920.00
24 51.20 1800.00
25 54.40 600.00
26 70.40 840.00
27 65.60 2088.00
28 56.00 2040.00
29
30
Data View
Variable View
SPSS Processor is ready
开始
阅读
快速路通行能力研究
SPSS学习 - Micro...
Untitled - SPSS ...
Microsoft Excel
11:43

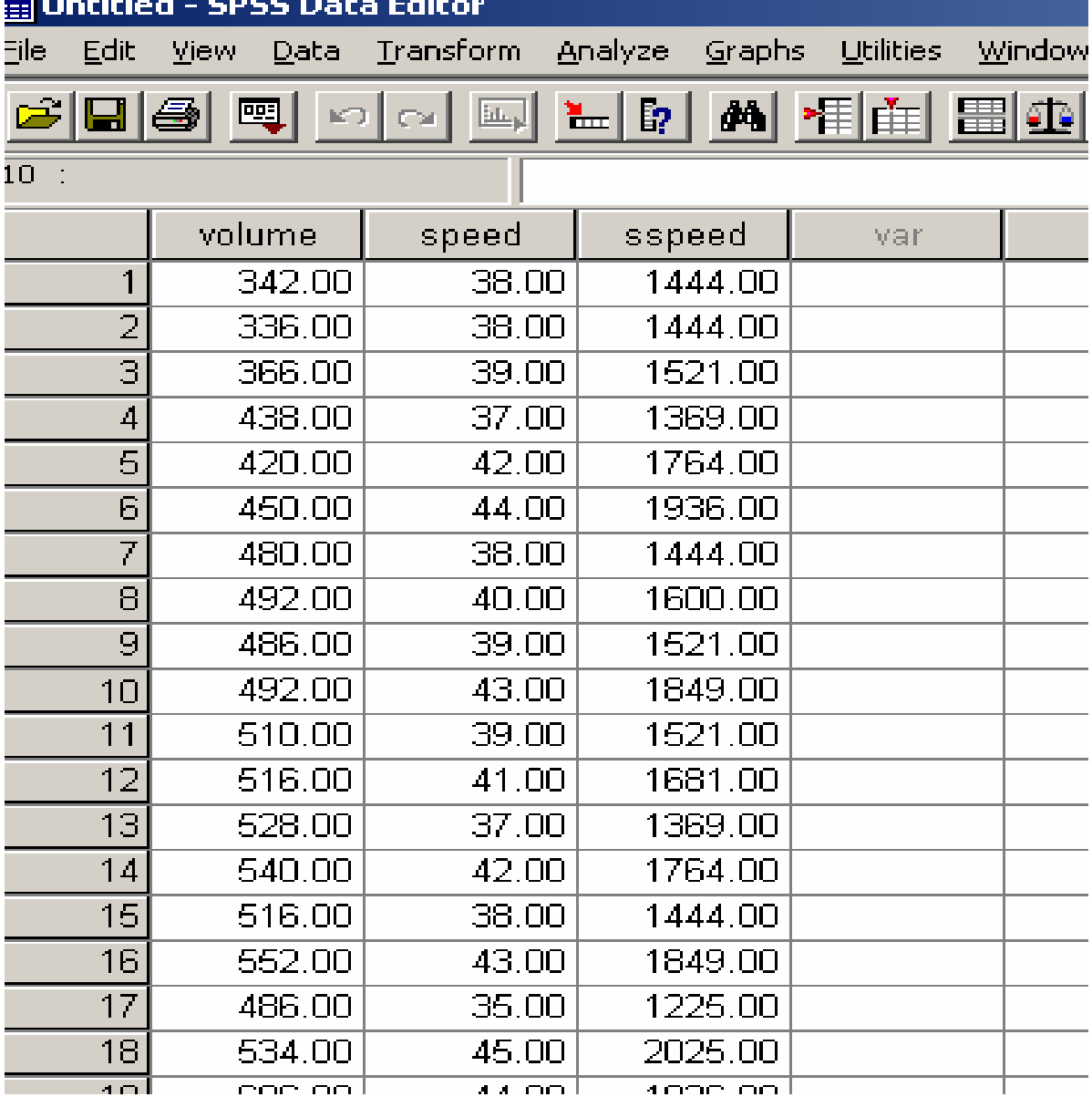
数据变换


Untitled - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
Compute...
Random Number Seed...
Count...
Recode
Categorize Variables...
Rank Cases...
Automatic Recode...
Create Time Series...
Replace Missing Values...
Run Pending Transforms
volume
var
var
1 342.00
2 336.00
3 366.00
4 438.00
5 420.00
6 450.00
7 480.00 38.00
8 492.00 40.00
9 486.00 39.00
10 492.00 43.00
11 510.00 39.00
12 516.00 41.00
13 528.00 37.00


Untitled - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
5 :
Compute Variable
Target Variable:
=
Type & Label...
volume
speed
Numeric Expression:
+ < > 7 8 9
- <= >= 4 5 6
* = ~= 1 2 3
/ & | 0 .
** ~ ( ) Delete
Functions:
ABS(numexpr)
ANY(test,value,value,...)
ARSIN(numexpr)
ARTAN(numexpr)
CDFNORM(zvalue)
CDF.BERNOULLI(q,p)
If...
OK
Paste
Reset
Cancel
Help



在target variable 筐中输入变换后得到的变量名,对本例用（sspeed）表示速度的平
在numeric expression 中输入变换公式。公式中的变量从左边列出的
量中选出，运算符号可以点击右边的符号即可。



	volume	speed	sspeed	var	
1	342.00	38.00	1444.00		
2	336.00	38.00	1444.00		
3	366.00	39.00	1521.00		
4	438.00	37.00	1369.00		
5	420.00	42.00	1764.00		
6	450.00	44.00	1936.00		
7	480.00	38.00	1444.00		
8	492.00	40.00	1600.00		
9	486.00	39.00	1521.00		
10	492.00	43.00	1849.00		
11	510.00	39.00	1521.00		
12	516.00	41.00	1681.00		
13	528.00	37.00	1369.00		
14	540.00	42.00	1764.00		
15	516.00	38.00	1444.00		
16	552.00	43.00	1849.00		
17	486.00	35.00	1225.00		
18	534.00	45.00	2025.00		

练习（应用SPSS软件）：

1）打开数据文件、建立新的数据集和变量

2）对给定的数据进行对数变换

4）倒数变换

5）标准化变换