```
  *================================
   *第2讲  数据文件建立和管理（初级篇）
   *================================

*--------------------------------
*->2.1数据文件的建立和读取
*--------------------------------
*-2.1.1.直接录入（适用于小样本少变量的数据文件新建）
*    直接在stata中录入：打开程序，调用数据编辑窗口，直接录入数据，如excel中操作。
*       调用数据窗口方式：
*            (a)在command窗口中输入edit命令
*            (b)点工具栏上的编辑文件图标
*    命令行输入：在command窗口，用input命令。
*操作实例->
    input  x  y
    1   2
    2   3
    end
*-2.1.2.读取dta数据文件
    *    鼠标或菜单操作：只能读取本地磁盘上的Stata数据文件
*       直接双击stata数据文件
*       菜单操作：在工具栏上直接点击打开文件图标，打开指定文件，或File>Open
*    use命令
    *       语法：use [varlist] [if] [in] using filename [,clear nolabel]
        *操作实例->
            cd "C:\data"
*    读取数据文件的指定变量信息
use make mpg using auto.dta, clear  // clear：在读取新数据时清除程序已读数据
*    读取数据文件的全部变量信息
            use auto.dta, clear
            use auto.dta, clear nolabel  // nolabel：在读取数据时清除数据中变量的值标签
         *    读取数据文件中符合指定条件的样本信息
            use auto.dta if price>7000, clear
*逻辑符号
*     ==：等于，!=：不等于
*     > ：大于，>=：大于等于
*     < ：小于，<=：小于等于
*     & ：且  ，| ：或
         *     读取数据文件中符合指定范围的样本信息
            use auto.dta in 1/20, clear
* 范围语句的定义：
*     in #：第#个观测值
*     in #1/#2：从#1个观测值到第#2个观测值
*     in f/#：从第1个观测值到第#个观测值
*     in #/l：从第#个观测值到最后一个观测值
*    use命令的扩展：sysuse/webuse/netuse
*-2.1.3.导入其它格式的数据文件
*    支持导入的数据类型包括：（1）Excel数据：*.xls, *xlsx;
//（2）用spreadsheet建立的ASCII(txt)数据：*.raw, *.txt, *.csv;
//（3）固定列宽的ASCII(txt)数据：*.dct;
//（4）以dictionary格式建的ASCII(txt)数据：*.dct, *.raw;
//（5）无固定格式的ASCII(txt)数据：  *.txt,  *.raw;
//（6）SAS XPORT数据：*.xpt;
//（7）ODBC数据源：包括access数据源，*.mdb，dBase数据源，*.dbf;
//（8）xml数据：*.xml
*    import excel命令：导入Excel数据，文件后缀名是*.xls, *.xlsx
*       语法1：import excel [using] filename [, import_excel_options]  //导入整个Excel数据工作表
*       语法2：import excel extvarlist using filename [, import_excel_options]
//导入Excel数据工作表的部分变量
*操作实例->
cd "C:\data"
sysuse auto
export excel auto, firstrow(variables)
import excel auto.xls, firstrow clear
describe
import excel auto.xls, cellrange(:D70) firstrow clear
describe
import excel make mpg weight price using auto.xls, firstrow clear
```

```
describe
import excel make=A mpg=B price=D using auto.xls, firstrow clear
describe
*    insheet命令：导入以tab-/逗号（comma-）/指定字符分隔的数据，文件后缀名通常是*.raw, *.txt, *.csv。
//空格分隔的数据无法直接用insheet命令，需要在参数设定delimiter(“ ”)。
*       语法：insheet [varlist] using filename [, options]
*操作实例->
insheet using auto.raw, tab
insheet using auto.txt, comma clear
insheet mpg price using auto.raw, comma
insheet mpg price using auto.raw, delimiter(“/”)
*    infile命令：导入以空格（space）-/tab-/逗号（comma-）分隔的数据，文件后缀名通常.dct, *.raw, *.txt。注意：在raw，txt格式的数据文件导入，一定要在命令中写出导入后数据的对应变量名。
*操作实例->
infile v1 v2 v3 using d21.txt, clear
infile x1 x2 using infile.txt, clear

*-----------------------------
*->2.2数据的保存或导出
*-----------------------------
*-2.2.1.数据保存
    *    save命令
    *        语法：save [filename] [, replace/nolabel/orphans /emptyok]
*操作实例->
save auto1a, replace
save auto1b, replace nolabel
save auto1c, replace orphans  //保存所有值标签
*-2.2.2.数据导出
    *   export excel/outfile/outsheet/fdasave/xmlsave命令

*---------------------------------
*->2.3数据清理的常见操作
*---------------------------------
*-2.3.1.变量属性，如名称、标签、值标签
*    变量名：由英文字符、数字、中文字符和_组成，最多不超过32个字母。
*            字母大小写表示的含义不同！！！
*            数字不可以单独做变量名，也不可放在变量名的最前面。
*            建议不要使用_作为变量的第一个字母。因为许多stata的内部变量都是以_开头的，如，_n, _N, _cons, _b等等。
*    标签：对变量的含义进行解释
*    值标签：对分类变量值的含义进行解释
*例子：性别：女=0，男=1
*-2.3.2.变量属性的显示
*    describe命令
*        语法1：describe [varlist], [simple/short]
*        语法2：describe [varlist] using filename, [simple/short]
*操作实例->
sysuse auto, clear
describe
describe make price mpg
describe course using "c:\data\myscore.dta"
*-2.3.3.变量属性的修改
*    rename命令：变量名更改
*       语法1：rename old_var new_var
*       语法2：rename (old1 old2 ...) (new1 new2 ...) [, options1]
*操作实例->
    sysuse auto, clear
rename price pri
rename (make price mpg) (mk pr mp)
rename t* t1*
*    label命令：定义和管理数据、变量或变量值的标签
*       语法1：label data “label”  //定义数据标签
*       语法2：label var varname “label”  //定义变量标签
*       语法3：
*label define lblname # "label" [# "label" ...] [, add modify replace nofix]  //定义值标签
* label values varlist [lblname|.] [, nofix]  //附加值标签到指定分类变量上
*操作实例->
```

```
sysuse auto, clear
label data “auto in American”  //数据标签
label var foreign “car type”  //变量标签
label define origin 0 “domestic” 1 “foreign” //值标签
label values foreign origin //附加值标签origin到分类变量foreign上
*    语法4：label dir  //显示值标签名
*    语法5：label list  //显示值标签名和内容
*    语法6：label drop  //剔除值标签
*    语法7：label copy  //复制值标签
*    语法8：label save  //保存值标签到do文件里
*-2.3.4.变量的存储类型
*    整数（数值型变量）的存储类型：
        * byte     字节型     (-100, +100)
        * int      一般整数型 (-32000, +32000)
        * long     长整数型   (-2.14*10^10, +2.14*10^10)，即，正负21亿
*    小数（数值型变量）的存储类型：
        * float    浮点型    8位有效数字
        * double   双精度   16位有效数字
*    字符型变量的存储类型：  str#
*变量信息由字母、特殊符号和数字组成
*保存为str格式，str后面的数字代表最大字符长度。比如，str18
*用在英文状态下的引号””标注
*   日期型变量，如19870815或15081987
*
！！！注意：数据及存储类型设置不当，如类型设置过小就会使得一些数据无法正常输入，甚至在两个变量或多个变量的观测值进行数学计算时结果出错。
*    recast命令：更改数值型变量的存储类型
*操作实例->
sysuse auto, clear
list gear_ratio in 1/5
display gear_ratio
recast int gear_ratio, force  //recast命令用于更改变量的存储类型
display gear_ratio
list gear_ratio in 1/5
*    compress命令：精简资料的存储格式
*-2.3.5.定义变量的显示格式
*    format命令：
*        语法1：format varlist %fmt  /  format %fmt varlist

*读懂%fmt的常见格式含义。如%-18s靠左列印于屏幕上；若%18s，则靠右列印；若%~18s,则居中列印。
*        语法2：format [varlist]
*操作实例->
   help format
list price gear_ratio in 1/5
format price %6.1f
format gear_ratio %6.4f
list price gear_ratio in 1/5
*-2.3.6.数值型分类变量和字符变量的转换（分类变量要定义值标签）
*    encode命令：将字符变量转换为分类数值变量。
*        语法：encode varname [if] [in], generate(newvar) [label (name) noextend]
*操作实例->
webuse hbp2, clear
describe sex
encode sex, generate(gender) label(sexlbl)
describe gender
*    decode命令：将分类数值变量转换为字符变量。注意：无值标签的数值变量不适用。
*        语法：decode varname [if] [in], generate(newvar) [maxlength(#)]
*操作实例->
webuse hbp2, clear
describe sex
label define gender 1 "female" 2 "male"
replace sex = "other" in 2
encode sex, generate(gender)
label list gender
decode gender, generate(sex2)
tostring gender, generate(sex3)
*-2.3.7.包含数值数据的字符型变量与数值型变量转换
*    destring命令：包含数值数据的字符型变量转换为数值型变量
```

```
*        语法：destring [varlist], [generate (newvarlist) | replace] [options]
*           常用参数：ignore (“chars”) 删除字符变量中的非数值字符
*                     force将非数值字符转换为缺失值
*                     replace：转换后的变量值替代原变量值
*操作实例->
webuse destring2, clear
describe date
list date
destring date, ignore(" ") replace
destring price, generate(p1) ignore(" $ ,")
encode price, generate(p2)
*    tostring命令：将数值变量转换为字符变量
*       语法：tostring varlist, [generate (newvarlist) | replace]
*操作实例->
webuse tostring, clear
describe
list
tostring year day, generate(y1 d1)
decode year, generate(y2)
*-2.3.8.新变量生成
*    generate/egen命令：
*        语法1：generate [type] newvar=exp [if] [in]
*        语法2：egen [type] newvar=fcn(arguments) [if] [in] [, options]
*操作实例->
sysuse auto, clear
gen id=.
gen lowprice=1 if price<4500
replace lowprice=0 if lowprice==.
gen lowprice2=(price<4500)
//注意：如果price中有缺失值时，该命令会将price为缺失值的样本的lowprice赋值为0，因为stata会将缺失值视为最大正值。
webuse egenxmpl2, clear
by dcode, sort: egen medstay = median(los)
gen slos1=sum(los)
egen slos2=sum(los)
webuse egenxmpl3, clear
egen byte differ = diff(inc1 inc2 inc3)
webuse egenxmpl4, clear
egen hsum = rowtotal(a b c)
egen hnonmiss = rownonmiss(a b c)
egen hsd = rowsd(a b c)
*        语法3：generate newvar=recode(varname, num1,num2, num3, …., numk)
*操作实例->
gen priceg=recode(price, 2000, 4000, 6000)
//在新变量priceg中，price<=2000的样本赋值为2000，>2000&<=4000）赋值为4000，>4000赋值为6000。
*    recode命令：分类变量再编码
*        语法4：  recode varlist (rule) [(rule) ...] [, generate(newvar)]
*操作实例->
webuse fullauto, clear
recode rep77 rep78 (1 2 = 1 "Below average") (3 = 2 Average) (4 5 = 3 "Above average"), pre(new)
label(newrep)
label list repair newrep
*    replace命令：变量值的修改
*        语法5：replace oldvar=exp [if] [in] [, nopromote]
*操作实例->
webuse genxmpl2, clear
generate lastname=word(name,2)
list
replace lastname=word(name,1)
list
*-2.3.9.缺失值的处理
*    在调查中，经常用88,
99,888,999,….等来表示不知道或不清楚。而这些信息在数据处理时，均属于非有效信息，是无法进入统计分析的。因此，这些都可视为观测缺失值。Stata中通常是用”.”来表示这类观测缺失值。
*    mvencode命令：用特定含义的数值来表示观测缺失值。
*        语法：mvencode varlist [if] [in], mv(#|mvc=# [\ mvc=#...] [\ else=#]) [override]
*    mvdecode命令：将特定含义的数值处理为缺失值
*        语法：  mvdecode varlist [if] [in], mv(numlist | numlist=mvc [\ numlist=mvc...])
```

```
*操作实例->
sysuse auto,clear
list rep78 foreign if rep78 == .
mvencode _all, mv(999) override
mvencode rep78 if foreign == 0, mv(998)
mvdecode rep78, mv(998=. \ 999=.a)
*-2.3.10.变量（观测值）的剔除、保留和显示
*    drop/keep命令：
*        语法1：drop/keep varlist  //变量
*        语法2：drop/keep if exp  //观测值
*操作实例->
sysuse census
drop pop*
drop if medage > 32
*    list命令：
*        语法：list varlist [if] [in] [,options]
*-2.3.11.数据的排序
*    sort命令/gsort命令
*        语法1：sort varlist [in]
[,stable]，用于升序排序。参数stable指如果两个观测值相同，则按其内存中的存放顺序来定排序。
*        语法2：gsort [-] varname [[-] varname …] [, generate (newvar)
mfirst]，可同时升、降序排序。参数mfirst表示将缺失值放在前面。
*操作实例->
sysuse auto
keep make mpg weight
sort make mpg weight, stable
list in 1/10
gsort + make -mpg - weight
list in 1/10

*--------------------------------
*->2.4数据集的合并和附加
*--------------------------------
*-2.4.1.数据集的合并
*    merge命令：1对1匹配合并，1对多（多对1）匹配合并，多对多匹配合并，按观测值1对1匹配合并
*        语法1：merge 1:1 varlist using filename [, options] //
*        语法2：merge m:1 varlist using filename [, options]
*        语法3：merge 1:m varlist using filename [, options]
*操作实例->
webuse overlap2, clear
merge 1:m id using https://www.stata-press.com/data/r16/overlap1, update replace
list
*-2.4.2.数据集的附加
*    append命令：append using filename [filename ...] [, options]
*操作实例->
sysuse auto, clear
keep if foreign == 0
save domestic
sysuse auto, clear
keep if foreign == 1
keep make price mpg rep78 foreign
append using domestic, keep(make price mpg rep78 foreign) generate(app1)
*============================over=========================================*

```