第一章 Stata 概貌

§1.1 Stata的功能、特点和背景

Stata是一个用于分析和管理数据的功能强大又小巧玲珑的实用统计分析软件，由美国计算机资源中心（Computer Resource Center）研制。从1985至1998的十四年时间里，已连续推出1.1，1.2，1.3，1.4，1.5，……及2.0，2.1，3.0，3.1，4.0，5.0，6.0等多个版本，通过不断更新和扩充，内容日趋完善。它同时具有数据管理软件、统计分析软件、绘图软件、矩阵计算软件和程序语言的特点，又在许多方面别具一格。Stata融汇了上述程序的优点，克服了各自的缺点，使其功能更加强大，操作更加灵活、简单，易学易用，越来越受到人们的重视和欢迎。

Stata的突出特点是只占用很少的磁盘空间，输出结果简洁，所选方法先进，内容较齐全，制作的图形十分精美，可直接被图形处理软件或字处理软件如WORD等直接调用。

一、 Stata的数据管理能力

⑴ Stata的数据管理空间受计算机的操作系统和计算机扩展内存的影响。对640k内存的微机，3.1版本的Stata可以管理2400个记录×99个变量，并随计算机扩展内存的增加而增加；对4.0的WINDOWS版本，Stata可以管理4800个记录×99个变量；对WINDOWS 95下的5.0版本，可根据计算机的配置情况设置变量数和记录数，如32M扩展内存的计算机，可处理2千万个数据。变量数和记录数可以互相交易（trade），即减少记录数可以增加变量数，减少变量数可以增加记录数。

⑵ 可以将分组变量转换成指示变量(哑变量)，将字符串变量映射成数字代码。

⑶ 可以对数据文件进行横向和纵向链接，可以将行数据转为列数据，或反之。

⑷ 可以恢复、修改执行过的命令。

⑸ 可以利用数值函数或字符串函数产生新变量。

⑹ 可以从键盘或磁盘读入数据。

二、 Stata的统计功能

Stata的统计功能很强，除了传统的统计分析方法外，还收集了近20年发展起来的新方法，如Cox比例风险回归，指数与Weibull回归，多类结果与有序结果的logistic回归，Poisson回归、负二项回归及广义负二项回归，随机效应模型等。具体说，Stata具有如下统计分析能力：

⑴ 数值变量资料的一般分析：参数估计，t检验，单因素和多因素的方差分析，协方差分析，交互效应模型，平衡和非平衡设计，嵌套设计，随机效应，多个均数的两两比较，缺项数据的处理，方差齐性检验，正态性检验，变量变换等。

⑵ 分类资料的一般分析：参数估计，列联表分析(χ^2检验，列联系数，确切概率)，流行

第二章 Stata的函数和变量

§2.1 Stata的函数

Stata具有丰富的函数功能。它不仅提供了一般计算机语言和统计软件包所具有的数学函数和字符串函数，而且还提供了很多有用的统计函数、特殊函数，以及许多独具特色的系统变量。借助于这些函数和系统变量，用户可以得心应手地使用Stata，充分发挥自己的聪明才智，提高工作效率。

为讲述方便，先引入指令display。display使我们的计算机行使简单的计算功能，例如，要计算 $3+2^2$：

```
.dislplay 3+2^2
 7
```

结果是7。

有了这个命令后，下面的讲述和练习就容易了。

一、 数学函数

1. abs(x) /*绝对值函数
2. exp(x) /*指数函数
3. log(x) /*自然对数
4. log10(x) /*常用对数
5. sqrt(x) /*平方根函数
6. sin(x) /*正弦函数
7. cos(x) /*余弦函数
8. atan(x) /*反正切函数
9. lngamma(x) /*整数x的Γ函数之对数ln[(x-1)!]
10. mod(x,y) /*模数函数获得x除以y的余数，如display mod(25,3)，结果将是1。

二、 统计函数

1. normprob(df,x) /*正态分布的下侧概率函数
2. invnorm(p) /*正态分布的分位数
3. Binomial(n,k,p) /*二项分布函数，表示n次试验中成功次数≥k的概率，p为成功概率
4. invbinomial(n,k,p) /*二项分布的逆函数，p示n次试验中成功次数≥k的概率，本函数给出的是每次成功的概率。
   当p<0.5时，概率p满足Pr(x≥k)=p
   当p>0.5时，概率p满足Pr(x≥k)=1-p
5. tprob(df,t) /*自由度为df的t分布双侧累积概率Pr(|t|>t)
6. invt(df,P) /*自由度为df的t分位数：invt(df,1-tprob(df,t))=t

7. fprob(df1,df2,f) /*自由度为df1,df2的F分布的上侧累积概率
8. invfprob(df1,df2,p) /*F分布的分位数。
   如果fprob(df1,df2,F)=p,则invfprob(df1,df2,p)=F
9. chiprob(df,x) /*自由度为df的χ^2分布的上侧累积概率
10. nchi(df,L,x) /*非中心χ^2分布的上侧概率。1<=df<=200,0<=L<=1000
11. invnchi(df,L,p) /*非中心χ^2分布的分位数。
    如果nchi(df,L,x)=p，则invnchi(df,L,p)=x
12. gammap(a,x) /*不完全gamma函数P(a,x)
13. invgammap(a,p) /*不完全gamma函数P(a,x)的逆函数：
    如果gammap(a,x)=p,则invgammap(df,p)=x
14. ibeta(a,b,x) /*不完全beta函数I_x(a,b)
15. uniform( ) /*产生(0,1)内的均匀分布的伪随机机数。每次使用时最好用命令“set seed“设置随机数种子，以打乱伪随机数的固有顺序。
16. invnorm(uniform( )) /*产生均数为0，标准差为1的标准正态分布随机数。
17. μ+σ×invnorm(uniform()) /*产生均数为μ，标准差为σ的正态分布随机数。

三、 字符串函数

以下用s表示一个字符串， n表示一个数值。

1. length(s) /*长度函数，计算s的长度，如，disp length("ab")的结果是2
2. substr(s,n1,n2) /*子串函数，获得从s的n1个字符开始的n2个字符组成的字符串，如，disp substr("abcdef",2,3)的结果是"bcd"
3. string(n) /*将数值n转换成字符串函数，如，disp string(41)+"f"的结果是"41f"
4. real(s) /*将字符串s转换成数值函数，如，disp real("5.2")+1的结果是6.2
5. upper(s) /*转换成大写字母函数，如，disp upper("this")的结果是"THIS"
6. lower(s) /*转换成小写字母函数，如disp lower("THIS")的结果是"this"
7. index(s1,s2) /*子串位置函数，计算s2在s1中第一次出现的起始位置，如果s2不在s1中，则结果为0。如，disp index("this","is")的结果是3，而index("this","it")的结果是0
8. trim(s) /*去除字符串前面和后面的空格
9. ltrim(s) /*去除字符串前面的空格
10. rtrim(s) /*去除字符串后面的空格

四、 特殊函数

1. **符号函数**sign(x) x>0**时取**1, x<0**时取**-1, x=0**时取**0。

2. **取整函数**int(x) **去掉**x**的小数部分，得到整数。**int(x+0.5)**是对**x**四舍五入取整，**int(x+sign(x)/2)**产生与**x**最近的一个整数。**

3. **求和函数**sum(x) **很常用，获得包括当前记录及以前的所有记录的**x**的和。缺失值**(missing value)**当**0**处理。**

4. **最大值函数**max(x1,x2,...,Xn) **忽略缺失值。**

5. **最小值函数**min(x1,x2,...,Xn) **忽略缺失值。**

6. **分组函数**group(x) **将数据分成大小近似相等的**x**个子样本，分别给予组编号** 1，2,...,x。

7. **浮点转换函数**float(x) **将**x**转换成浮点表示法。**Stata**是用浮点形式存储数据的， 因此，在将变量与具体数值进行比较时，最好先将具体数值转换成浮点形式。例如， 当**x**为**1.1**时，表达式**x==1.1**的结果可能不真， 因为表达式左边的**x**是浮点形式， 右边的**1.1**是双精度形式，二者相差**0.00000002384，而改写为**x==float(1.1)**后，结果就正确了。当某个数值没有有限的二进制表达时， 常常会出现这种情况。**

8. **条件函数**cond(x,a,b) x**可以是一个条件，**x**非**0**(条件成立)时取**a，x**为**0**(条件不成立)时取**b。

9. **归组函数**recode(x,x_1,x_2,...,x_n) **其结果可表达如下**:

$$\text{recode}(x,x_1,x_2,\cdots,x_n) = \begin{cases} x_1 & \text{如果} x<=x_1 \\ x_2 & \text{如果} x_1<x<=x_2 \\ x_3 & \text{如果} x_2<x<=x_3 \\ \cdots\cdots & \\ x_{n-1} & \text{如果} x_{n-2}<x<=x_{n-1} \\ x_n & \text{如果} x>x_{n-1} \\ \text{缺失值} & \text{如果} x\text{为缺失值。} \end{cases}$$

10．**自动归组函数**autocode(x,ng,xmin,xmax) **自动将区间（**xmin,xmax**）分成**ng**个等长的小区间，其结果是包含**x**值那个小区间的上界值。其作用与归组函数相同。**

§2.2 Stata的格式文件、变量和系统变量

一、 文件名和文件类型

Stata**的格式文件命名规则与**Dos**中文件的命名规则相同，文件名以字母开头，不超过**8**个字符，不能用标点符号，及**Dos**中的通配符。**Stata**共有六种格式文件，其默认的后缀(文件扩展名)见表**2.1**。**

表2.1 Stata的文件类型

文件扩展名	文件特性
dct	ASCⅡ数据字典文件
raw	ASCⅡ数据文件
do	命令文件
dta	Stata数据文件
log	Stata结果文件
gph	Stata图形文件
xp	Stata的xp数据文件

二、 变量名和变量类型

与文件名一样，Stata**的变量名可以是英文字母(**A－Z**和**a-z**)，数字(**0－9**)，下划线(** _ **)，可**

区分的有效长度<=8。Stata中英文字母的大小写是有区别的。此外，以下是Stata的关键字或系统变量，不得用作用户变量名：

_all _n _N _skip _b _coef _cons _pi _pred _rc _weight double float long int in if using with

Stata的用户变量有数值变量和字符串变量两种。字符串长度可以多达254，但只有前面80个才有效。

三、 系统变量

1．_coef[变量名]或_b[变量名]系统函数 拟合方差分析、回归分析、Cox、logit或probit等模型后，利用系统函数可得到指定变量在当前拟合模型中的系数。在方差分析后，中括号[ ]内的变量可以是某种处理的某一水平。例如，_coef[drug[2]]表示药物的第二水平的系数，coef[drug[2]*disease[1]]表示药物的第二水平与第一种疾病的交互作用项的系数；在多类结果的logistic回归后，中括号[ ]内的变量可以是变量在某一类中的回归系数(见第十五章)。

2．_cons 常数函数 直接使用时总是1，而_b[_cons]]的结果是当前拟合模型的常数项。

3．_N 数据库中观察值的总个数。

4．_n 当前观察值的位置。

5．_pi 圆周率π的数值。

6．_rc 最近一次capture命令返回代码的数值

四、 结果变量

除此以外，Stata还提供了一个独具特色的结果变量_result(#)，该变量实际上是一个系统变量，但由于其特殊性和重要性，专门把它列为一节讲述。

第一章已讲到，Stata有许多其它软件所没有的优点，其中一个优点是它的显示结果非常简明、清晰，并将用户可能用于构造新变量的分析结果存于系统变量_result(#)下，这就为用户编制批处理命令文件进行连续分析处理数据创造了条件。

_result(#)括号内的#可以是一个具体数值，也可以是一个算术表达式，变化范围根据使用时的环境而变化，如在回归分析之后，n可以是1,2,3,4或5，每一个数值对应一个统计指标，详见表格2.2。

表2.2 _result(n)的使用环境及相应的统计指标

correlate	
1．观察值个数	4. 第一和第二个变量的相关性或协方差
2.	5. 1或第二个变量的方差
3．1或第一个变量的方差	
count	
1．计数结果	
inspect	
1．观察值个数	5.为负整数的观察值个数
2．为负数的观察值个数	6. 为正整数的观察值个数
3．变量个数	7.唯一数值或缺失值的个数
4．为正数的观察值个数	8.不能辨别是数值还是缺失值的个数

表2.2(续) _result(n)的使用环境及相应的统计指标

describe or describe using

1．观察值个数
2．变量个数
3．当前宽度
4.观察值的最大个数
5.变量的最大个数
6.最大宽度

cox,logit,or probite

1.观察值个数
2．似然对数
3.模型的自由度
4.χ^2值

anova,regress,or stepwise

1．观察值个数
2．模型的平方和
3．模型的自由度
4．残差平方和
5．残差的自由度,
6.F统计量
7.R^2
8.调整R^2
9.误差均方根

Factor

1．观察值个数
2．保留因素的个数
3．相对于无因子时的χ^2检验
4．相对于无因子时的χ^2检验的自由度
5．相对于更多因子时的χ^2检验
6．相对于更多因子时的χ^2检验的自由度
7．第1个特征根
8．第2个特征根
…
…
19．第13个特征根

Oneway

1．观察值个数
2．组间平方和
3．组间自由度
4．组内平方和
5.组内自由度
6.F统计量
7.Bartlettχ^2检验
8.Bartlettχ^2检验的自由度

Summarize

1．观察值个数
2．权重之和
3．均值
4．方差
5．最小值
6．最大值
7．第5百分位点（选detail时才有效）
8．第10百分位点（选detail时才有效），
9.第25百分位点（选detail时才有效）
10.第50百分位点（选detail时才有效）
11.第75百分位点（选detail时才有效）
12.第90百分位点（选detail时才有效）
13.第95百分位点（选detail时才有效）
14.偏度系数（选detail时才有效）
15.峰度系数（选detail时才有效）

Tabulate

1．观察值个数
2．行数
3．列数
4．Pearsonχ^2检验
5．Pearsonχ^2检验的显著性
6．l.r. χ^2检验
7．l.r. χ^2检验的显著性,
8.Fisher确切概率
9.Cramer V 统计量
10.gamma 统计量
11.gamma 统计量的ASSE
12.tau-b 统计量
13.tau-b 统计量的ASSE

Test

2．模型的平方和
3．模型的自由度
4．残差平方和（只在anova后才有效），
5.残差的自由度
6.F统计量

§2.3 Stata的算术运算和关系运算

一、 算术运算

stata的加、减、乘、除及乘方运算符依次是+、－、*、/ 和^ 。如：

$\frac{x/y^{(x-y)}}{xy}$ 应表达为（x/y^ (x-y)）/(x*y)

二、 字符串运算

字符串只有“加”运算，如“this”+“is”，结果是“thisis”。

三、 关系运算

Stata的关系运算符有：>（大于）、>=（大于等于）、<（小于）、<=（小于等于）、~=（不等于）、= =（等于），在Stata的条件语句中“等于”要用两个等于号表示。关系运算的取值是真（取值1）或假（取值0）。关系运算不仅对数值有效，也可用于字符串。字符串的关系运算是比较字符串在ASCII码中的先后顺序而不是数值的大小。此外，Stata规定所有的大写字母的位置都在小写字母的前面，而缺失值在所有非缺失值的后面。

四、 逻辑运算

&(“与”)、|（“或”）、~(“非”)是Stata的三个逻辑运算符。逻辑运算的结果也是真（取值1）和假（取值0）。

五、 运算优先顺序

-（负），~（非），^，/，×，－（减），+，~=，>,<,<=，>=,==，&，|。

六、 实例

表2.3给了一些运算实例。在Stata状态键入disp再空一格键入表中第一列的表达式并按回车键，便可得到表中第二列的结果。

表2.3 运算实例

表达式	结果
2+”this”	错误
substr(string(20+2),1,1)+upper(substr(“ rf”,1+1,1)	“2F”
“Zebra”>”cat”	0
(2==2)+1	2
(3>2)+1	2
3>2+1	0
3>2&5>4	1
(3>2)&(5>4)	1

最后，值得一提的是：本章所述变量和函数均不能单独使用，而必须与generate、replace、display**指令结合起来使用。**generate**和**replace**的使用详见第三章。**

病学表格分析等。

⑶ 等级资料的一般分析：秩变换，秩和检验，秩相关等。

⑷ 相关与回归分析：简单相关，偏相关，典型相关，以及多达数十种的回归分析方法，如多元线性回归，逐步回归，加权回归，稳键回归，二阶段回归，百分位数(中位数)回归，残差分析、强影响点分析，曲线拟合，随机效应的线性回归模型，等。

⑸ 危险度分析：条件和非条件的logistic 回归，多类结果与有序结果的logistic 回归，Probit回归，及其他广义线性模型，随机效应的logistic 回归，随机效应的Poisson回归，等。

⑹ 生存分析：基线生存曲线的估计、相对危险度的估计，Kaplan-Meier生存曲线、寿命表分析，对数秩检验，Mantel-Haenszel检验，Wilcoxon-Gehan检验，Cox比例风险模型，正态截尾及Tobit回归，指数回归和Weibull回归，等。

⑺ 其它方法：质量控制，整群抽样的设计效率，诊断试验评价，kappa，等。

三、 Stata的作图功能

Stata的作图模块，主要提供如下八种基本图形的制作：直方图(histogram)，条形图(bar)，百分条图(oneway)，百分圆图(pie)，散点图(twoway)，散点图矩阵(matrix)，星形图(star)，分位数图。这些图形的巧妙应用，可以满足绝大多数用户的统计作图要求。在有些非绘图命令中，也提供了专门绘制某种图形的功能，如在生存分析中，提供了绘制生存曲线图，回归分析中提供了残差图等。详见第五章。

四、 Stata的矩阵运算功能

矩阵代数是多元统计分析的重要工具，Stata提供了多元统计分析中所需的矩阵基本运算，如矩阵的加、积、逆、Cholesky 分解、Kronecker内积等；还提供了一些高级运算，如特征根、特征向量、奇异值分解等；在执行完某些统计分析命令后，还提供了一些系统矩阵，如估计系数向量、估计系数的协方差矩阵等。

尽管Stata的容量最大只容许400×400的矩阵(默认为40×40)，用它来完成日常工作中的统计分析显然不现实，但用它来做一些练习，提高多元统计分析的教学效率，无疑是很有帮助。详见第十八章。

五、 Stata的程序设计功能

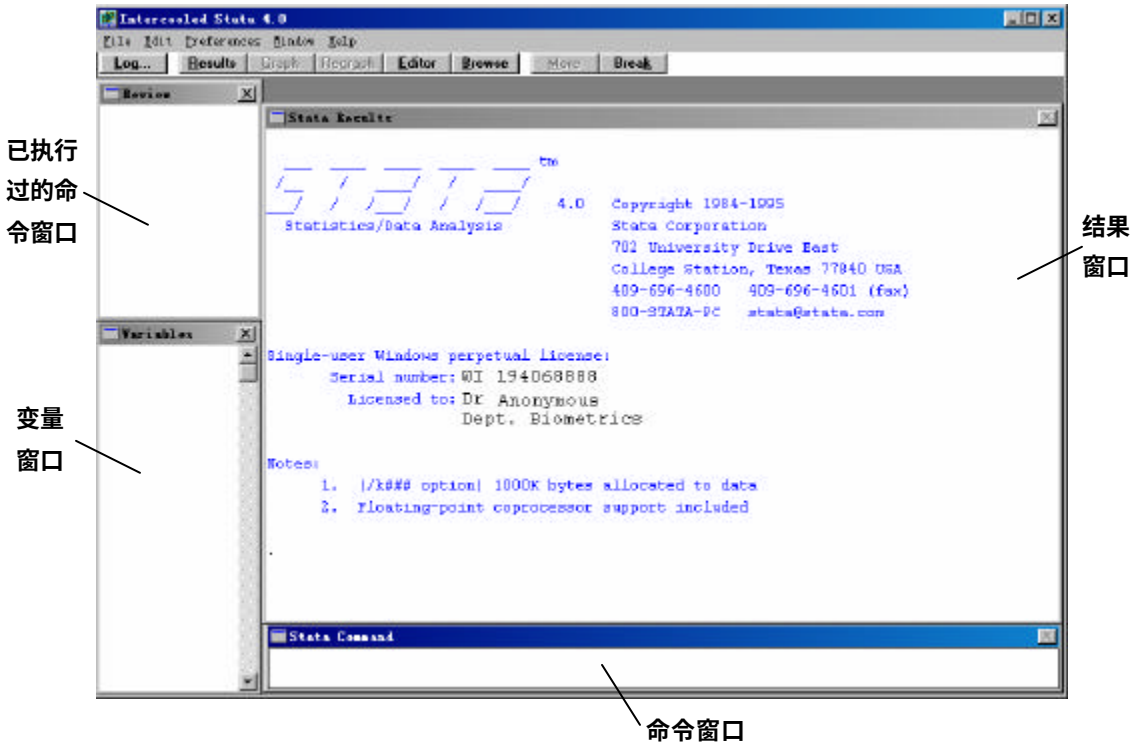
Stata是一个统计分析软件，但它也具有很强的程序语言功能，这给用户提供了一个广阔的开发应用的天地，用户可以充分发挥自己的聪明才智，熟练应用各种技巧，真正做到随心所欲。事实上，Stata的ado文件(高级统计部分)都是用Stata自己的语言编写的。下面这段程序是笔者自行编写的，用于产生n个参数为λ的Poisson分布的随机数。

```
prog  define  rp                          /*  定义程序名
set obs `2'                               /*  定义数据库的最大记录数
set  seed   `3'                           /*  设置随机数种子，
gen  rp=.                                 /*  定义变量rp，用于存放Poisson分布随机数
local lamda0=exp(`1')                     /*  计算lamda0=exp(λ)
local  j=1                                /*  j=1
while `j'<`2'+1  {                        /*  对 j<n循环，j表示产生的第j个Poisson分布随机数
```

```
    local i=1                                  /*  i=1
    local r0=1                                 /*  r0=1
        while `i'>0 {                          /*  循环
            local r1=uniform()                 /*  r1=均匀分布的随机数
            local r0=`r1'*`r0'                 /*  r0=r1*t0
                if `r0'< `lamda0'{             /*  如果 r0<lamda0
                local n0= `i'-1                /*  n0= i-1
                local i=-1                     /*  i=-1
                }
            local i= `i'+1                     /*  i循环
            }
  quiet replace rp=`n0'if _n==`j'              /*  第j个rp＝n0
  local j= `j'+1                               /*  循环
  }
end
```

§1.2 Stata的界面

Windows版本的Stata的界面上有一级菜单行，二级菜单窗口，命令窗口，结果窗口，图形窗口，变量名窗口，已执行过的命令窗口，帮助窗口等。窗口的大小、位置可根据用户需要进行调整。



§1.3 进入和退出Stata

一、 DOS版本的Stata的进入和退出

前已述及，要将Stata程序所在的路径放入autoexec.bat中，我们可在DOS下任何目录位置进入Stata，但我们假定d:\盘上进行。

D:\>Stata

进入Stata后，屏幕显示Stata的版本号，公司所在地等信息，Dos版本下的Stata即出现圆点提示符。这时即可键入Stata的各种命令。

若已在Stata状态读入了数据，并且已将数据按Stata指令存盘，或读入的数据虽经分析，但对数据及数据结构等未作任何修改，则只须键入：

. exit

即可退出Stata。

如未将数据按Stata指令存盘，或读入的数据或数据结构已被修改(Stata的有些命令会自动修改数据结构，如按某变量排序等)，这时，Stata将拒绝退出Stata状态。若确实不需要存盘而退出Stata，可键入：

. e，clear

(e为exit的简写)即可强行退出Stata。或分两步，即先放弃所有数据，

. drop _all

再退出Stata，

. exit

二、 WINDOWS版本的Stata的进入和退出

在桌面上双击Wstata的图标：

即可进入 Stata，并出现命令窗口。

在Stata的菜单中选 File ,再选 exit ,如数据已经存盘，则可退出Stata。如数据未存盘，则Stata给出如下提示：“ Data has changed without being saved. Do you really want to exit? ”(数据已改变，但未存盘，是否真的要退出？)如要退出，则按 确定 ，否则按 取消 。将数据存盘后再退出。

在WINDOWS下，亦可用DOS的命令退出Stata。

§1.4 Stata的数据输入与储存

Stata可以从键盘输入数据，也可以从文件读入数据。WINDOWS下的Stata还可以用Stata的数据编辑器输入、修改和管理数据。这里简单介绍如何从键盘输入数据，有关更详细的数据读

入方式将在第三章中讲述。

一、 从键盘输入数据

例1.1 某实验得到如下数据

x	1	2	3	4	5
y	4	5.5	6.2	7.7	8.5

进入Stata后，操作过程如下，其中划线部分为操作者输入部分。

```
. input x y

            x         y
  1. 1  4
  2. 2  5.5
  3. 3  6.2
  4. 4  7.7
  5. 5  8.5
  6. end
```

用list命令可以看到输入的数据。

```
. list

          x         y
  1.      1         4
  2.      2       5.5
  3.      3       6.2
  4.      4       7.7
  5.      5       8.5
```

二、 保存数据

为了方便以后应用，输入Stata的数据应存盘。如欲将上述数据存入d:\mydata\子目录中，文件名为ex1.dta，命令为：

```
. save d:\mydata\ex1
  file d:\temp\ex1replace.dta saved
```

该指令在d:盘的mydata子目录中建立了一个名为“ex1.dta”的Stata格式的数据文件。后缀dta是Stata内定的数据格式文件。该格式文件只能在Stata中用use命令打开：

```
. use d:\mydata\ex1
```

如目标盘及子目录中已有相同文件名的文件存在，则该命令将给出如下信息：file d:\mydata\ex1.dta already exists，告诉用户在该目标盘及子目录中已有相同的文件名存在。如欲覆盖已有文件，则加选择项replace。命令及结果如下：

```
. save d:\mydata\ex1 , replace
  file d:\temp\ex1.dta saved
```

这样，Stata在d:盘的mydata子目录中建立了一个名为“ex1.dta”的Stata格式数据文件,并替换了原有文件。

§1.5 Stata的结果文件

Stata在屏幕上显示的运行结果有两种，一种是纯字符型的(如方差分析结果，回归分析结果等)，一种是图形。

若要将操作过程和纯字符型结果记录下来，需事先打开一个log文件：

```
. log using 文件名
```

设结果文件名为result1，则Stata自动加上后缀“.log”,亦可由用户自己加上其他后缀。执行该指令后的所有操作指令和文字结果(除help下显示的结果)将记录在结果文件“result1.log”中。若执行某一指令后的结果没有必要记录下来，则可事先用指令“log off”暂停记录，需要记录时再用“log on”继续记录，最后用“log close”关闭文件。

如果结果文件“result1.log”已经存在，用“log using result1”不能打开已有文件result1.log。如要覆盖文件result1.log，则加选择项replace。即键入：

```
. log using result1, replace
```

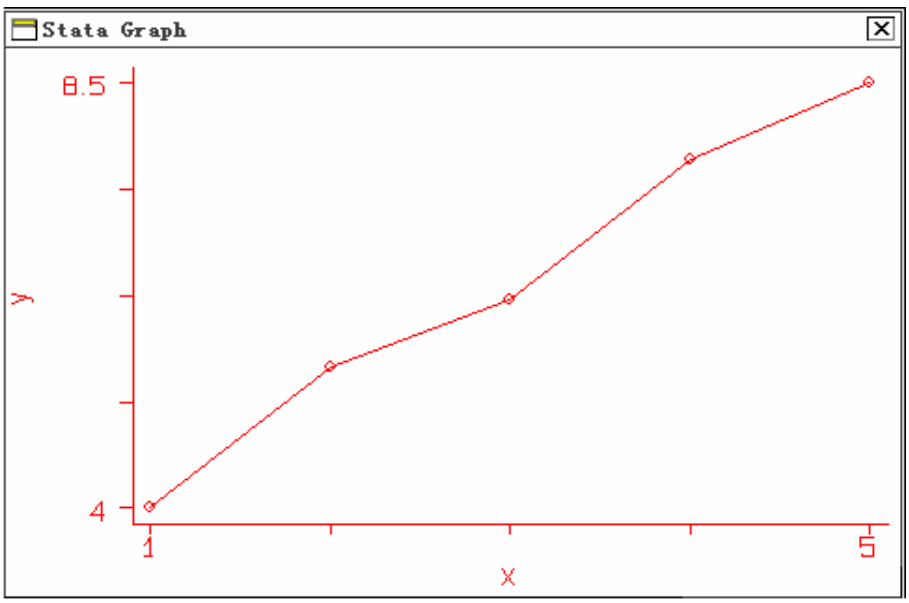
如要在其后进行添加，则键入：

```
. log using result1, append
```

文件“result1.log”可在EDIT、PE2、WPS或WORD等字处理软件下编辑、打印，也可在DOS下用type或print命令通过显示器浏览或打印机输出硬拷贝。

若要将图形结果打印下来，需要在绘图指令中加上“saving”选择项。例如，画例1.1中x与y的散点图并存入文件“ex1.gph”，可用下述指令：

```
. graph y x , c(l) saving(d:\mydata\ex1)
```



这时屏幕上显示y与x的散点图，并将被存入d:\mydata\子目录中，文件名为“ex1.gph”(gph是Stata内定的图形文件后缀，用户亦可自己定义后缀名)。该图形可在Stata状态用“ graph

using d:\mydata\ex1”重新显示在屏幕上，可在 File 的 Print Graph 打印，也可用打印命令“gphdot”打印。

DOS版本的Stata可在DOS提示符下用“gphdot” 命令打印：

```
D:\MYDATA>gphdot ex1.gph
```

更详细的内容见第五章。

§1.6 Stata的操作方式

Stata的操作有交互式操作和非交互式操作两种形式。

一、 交互式操作

在Stata状态直接键入指令，每输入一个指令，Stata执行一个，这种方式称为交互式操作。

例1.2 用例1.1数据建立回归方程。

```
. use ex1
. reg y x
```

二、 非交互式操作

若分析内容很多，有时甚至涉及到多个数据库，有几十个甚至成百个分析内容，若仍采取交互式操作，不仅要将许多时间花在等待运算结果上，而且容易漏掉一些主要的分析内容或做一些无益的重复劳动。这时最好在EDIT，PE2，WORD等文字处理下将这些指令写入一个以“do”为扩展名的命令文件(文本格式，即ASCII码)，并仔细核对分析内容、命令格式，直至组织数据文件的合理性等，修改好后再在Stata状态执行该命令文件。

例1.3 用非交互式操作对例1.1数据进行相关和回归分析。

第一步，在字处理软件下写入如下指令，并以文件名“ex1.do”存入磁盘d:\mydata\子目录中。

```
set more 1                          /* 指定结果窗口中，当输出结果满一屏后，不再显示--more--，直接显示下一屏
log using d:\mydata\ex1.log         /* 打开结果文件ex1.log
use d:\mydata\ex1.dta               /* 调用数据文件d:\mydata\ex1.dta
gra y x,saving(d:\mydtata\ex1)      /* 作y与x的散点图，并存入d:\mydtata\ex1.gph
cor y x                             /* 作y与x的相关
reg y x                             /* 作y与x的回归
log close                           /* 关闭结果文件ex1.log
set more 0                          /* 指定结果窗口中，当输出结果满一屏后，显示-- more--，直到按任意键后，再显示下一屏
```

第二步，在Stata状态键入：

```
. do d:\mydtata\ex1.do
```

Stata将首先打开一个名为“ex1.log”的结果文件，然后打开数据文件“ex1.dta”，画散点图并将图形存入文件“ex1.gph”，进行相关分析、回归分析，最后关闭结果文件。此时，Stata执行这些命令是自动的，不间断的。

§1.7 Stata的帮助功能

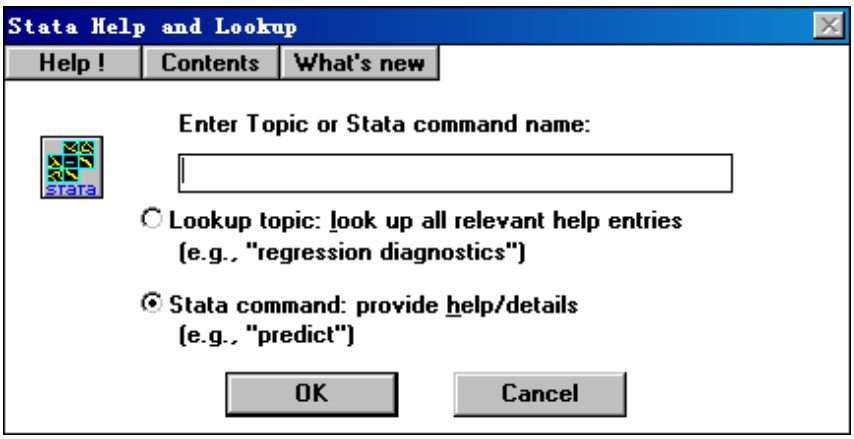
Stata具有很强的帮助功能。帮助功能的使用有两种方式。

一是在Stata状态，需要了解某个指令的格式和功能，这时只需键入help(或按功能键F1)，然后空一格键入该指令即可。例如，若需了解回归分析的指令格式，则：

```
. help regress
```

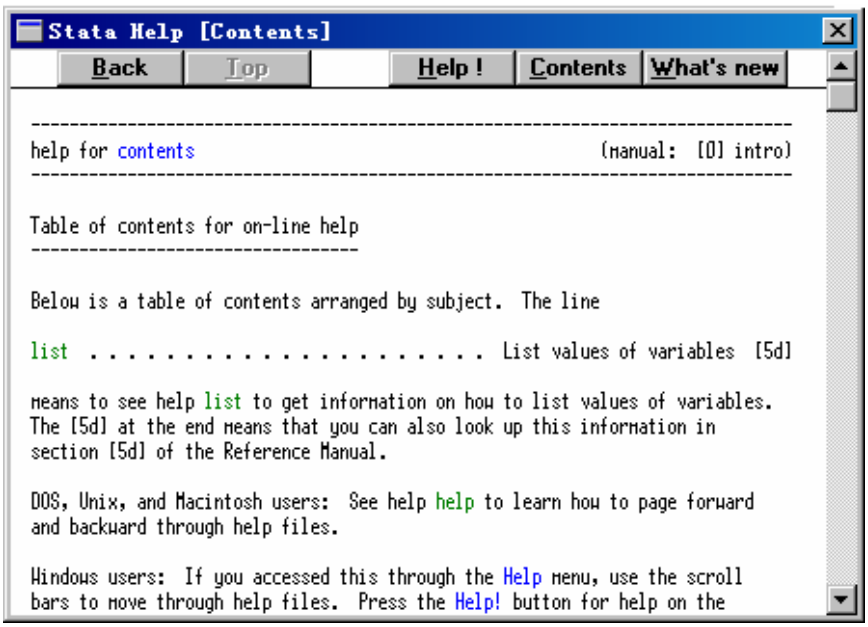
则可得到帮助。

二是利用菜单，在Stata的菜单上按 Help ，出现帮助窗口。



此时输入需要帮助的命令关键词，如regress，按 OK 即可得到帮助。

如需了解Stata的全部命令，可键入help contents，可得到Stata的全部命令及其简单解释；或在帮助窗口按 Contents ,则出现如下的帮助内容窗口。



在知道所要帮助的命令时，在命令窗口键入help加命令，即可获得帮助；在不知道所要帮助的命令时，用菜单操作更好。Stata的常用命令见附录。

下面以多元线性回归命令为例，介绍Stata的命令的格式。多元线性回归命令为regress，欲得到命令格式，键入help regress即可得到：

```
[by varlist:]  regress [depvar [varlist1 [(varlist2)]]]
                       [weight] [if exp] [in range] [, level(#)
                        beta hascons noconstant noheader eform(string)
                        depname(varname) mse1 ]
```

命令中，[ ]内为选择项，括号外为必选项。

这里介绍命令的公共选择部分，该命令的专用选择项将在相应章节作介绍。

(1) by varlist，是指定按变量varlist的取值逐一作多元线性回归。如变量名为group，且取值为1，2，3，4，则“by group:”是指定Stata分别按group=1, group=2, group=3和group=4的观察值分别作4个回归方程。在选用该选择项前，要对变量排序，即先执行sort,如：

. sort group

(2) weight，是指本命令允许使用加权或频数，有[fw=频数变量]和[aw=加权变量]两种形式。

(3) if exp, 用条件语句指定条件。如，下列条件是合法的：

```
if group==1                 /*  对满足group=1条件的观察值进行分析
if group>2                  /*  对满足group>2条件的观察值进行分析
if group==1 | group==2      /*  对满足group=1或group=2条件的观察值进行分析
if group~=3                 /*  对满足group不等于3条件的观察值进行分析
if group==1 & sex==0        /*  对满足group=1,同时sex=0条件的观察值进行分析
```

(4) in range，指定观察值的范围，对在范围内的观察值作分析。下列语句是合法的：

```
in 1/25                     /*  对观察值范围为1~25号的观察值作分析
in 26/44                    /*  对观察值范围为26~44号的观察值作分析
in 26/l                     /*  对观察值范围为26~最后(last)的观察值作分析
in –5/l                     /*  对最后5个观察值进行分析
```

这些公共选择项在很多命令中都可选用，本书在介绍各命令时将省去这些公共选择项。

另外一个选择项，也可用于很多命令，它就是 for。例如，在作回归分析时，自变量为 $x1,x2,\ldots,x22$ 共 22 变量，而因变量有 $y1,y2,\ldots,y10,z1,\ldots,z5$ 共 15 个变量。欲分别建立每个因变量 *yi* 和 *zi* 与 $x1,x2,\ldots,x22$ 的回归，则需要写 15 个命令。而用 for 选择项只需一个命令即可：

```
for y1-y10 z1-z5 : regress @ x1-x22
```

命令中，for 后面的变量是选定的，regress 是作回归分析，@是替换符，Stata 自动用 for 语句指定的变量逐一替换作为因变量，而自变量为 x1-x22。

又如，

```
for y* : summ @,detail
```

表示，对以y字母开始的变量作详细的统计描述。

第三章 Stata的数据库操作技巧

数据库管理是统计分析软件的基础，熟练地掌握数据库的操作是进行统计分析的前提，特别是对实际资料进行分析时，数据库操作技巧尤显重要。本章是Stata的基础部分，对需要深入了解Stata或进行复杂的数据库操作的读者，是必不可少的。

§3.1 Stata数据库的建立

Stata数据库的建立有4种方法，即从命令行键盘输入、用Stata的数据编辑器输入、从ASCII数据文件读入，以及从dbase或Foxbase数据库，SAS，SPSS等数据文件中转入。

一、 从键盘输入数据

从键盘输入数据适用于数据量比较少的情况。用input命令。

例3.1 表3.1为一配对试验数据，试从键盘输入Stata，并保存为Stata格式文件。

表3.1 配对试验数据

x0	x1
3550	2450
2000	2400
3000	1800
3950	3200
3800	3250
3750	2700
3450	2500
3050	1750

进入Stata后，键入input及变量名x0 x1，Stata即进入数据输入状态。然后依次输入数据x0和x1，所输数据的顺序与变量名一致，数据间用空格分开，每输完一组键入回车键 Enter ，数据输完后键入“ end ”，Stata将自动退到圆点提示符状态。

```
. input x0 x1
      x0     x1
1.  3550   2450
2.  2000   2400
3.  3000   1800
4.  3950   3200
5.  3800   3250
6.  3750   2700
7.  3450   2500
8.  3050   1750
```

```
9. end
```

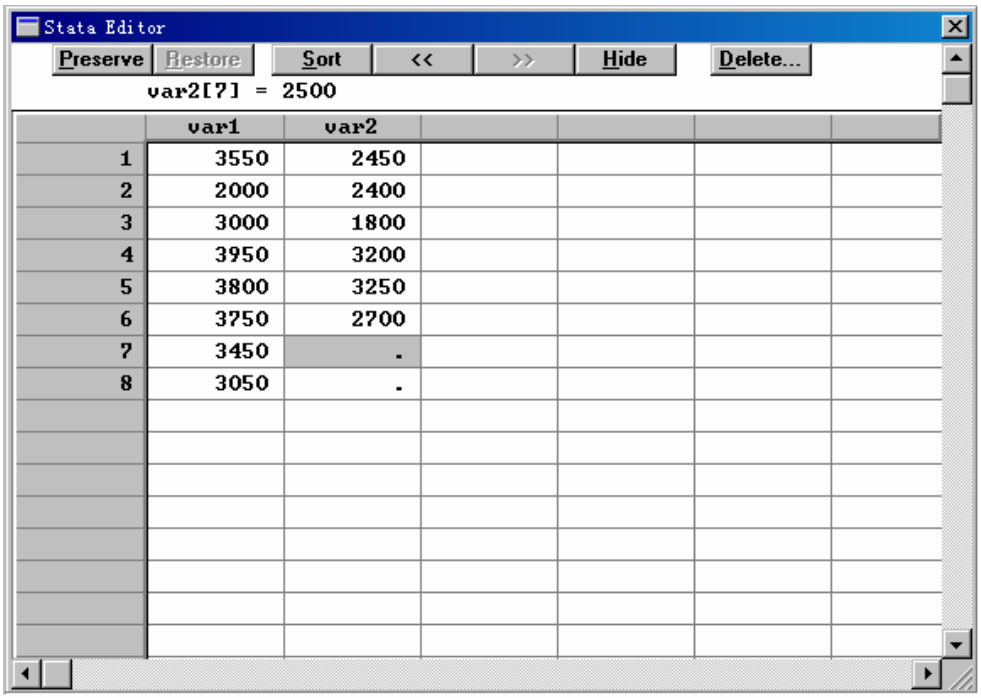
至此，数据输入完毕。可用list命令查看。要将数据存成Stata的格式文件，用命令“save”：

```
. save d:\mydata\ex3-1
```

该指令在d:\盘的mydata子目录中建立了一个名为“ex3-1.dta”的Stata数据文件。后缀dta是Stata内定的数据格式文件。该文件只能在Stata中用“use”命令打开。

二、 用Stata的数据编辑工具

Windows版本的Stata提供了数据编辑工具editor，给数据的输入提供了方便。在Stata中键入edit或在Stata的菜单中单击 Editor ,即可进入Stata的数据编辑器，按变量输入数据。Stata将第1列自动命名为var1,第2列命名为var2，依次类推。在数据输入完后，单击 Preserve 键确认所输数据，按“关闭”即可退出编辑器。此时数据已输入内存。再用“rename”命令将变量更名。




```
. ren var1 x0
. ren var2 x1
. list
          x0        x1
  1.    3550      2450
  2.    2000      2400
  3.    3000      1800
  4.    3950      3200
```

5. 3800 3250
6. 3750 2700
7. 3450 2500
8. 3050 1750

如发现数据有误，可在任何时候进入编辑器对数据进行修改。按前述方法可将数据存盘。

三、 从ASCII文件读入数据

Stata可以直接将ASCII码数据文件读入内存。用“ infile 变量名 using ASCII数据文件名”。具体操作如下。

第一步 用WPS,PE2,EDIT,WORD或WINDOWS中的写字板等字处理软件输入表3.1数据，格式如表3.2，每个变量一列，不加变量名，纯粹是数据。并以文件名ex3-1.txt存入磁盘d:\mydata\子目录中(ASCII码文件)；

表3.2 表3.1数据在PE2字处理软件下的排列格式

3550	2450
2000	2400
3000	1800
3950	3200
3800	3250
3750	2700
3450	2500
3050	1750

第二步 在Stata状态输入以下命令:

```
. infile x0 x1 using d:\mydata\ex3-1.txt
```

此时数据已被读入内存。用“ save ”命令存盘。注意，infile后的变量顺序需与数据文件中的顺序一致。

四、 从dbase、foxbase文件或其它数据格式读入数据

transfer是DOS下的一个程序，该程序可将多种格式的数据文件，包括dbase及foxbase文件，转换成Stata格式文件或其它格式文件。现以将foxbase转换为Stata格式文件为例，来说明transfer的用法。

在dbase或foxbase下输入例3.1数据，并以文件名为ex3-1.dbf，存入d:\mydata\子目录下,欲将ex3-1.dbf转换成Stata的格式文件，按transfer的菜单指示进行操作，十分方便。步骤如下：

第一步，在DOS状态键入transfer，出现transfer的页面。

```
                    S T A T  /  T R A N S F E R

                        (C) Copyright 1986-1991
                          All Rights Reserved
                          Circle Systems, Inc.
                             (206) 682-3783

                              Version 2.0

                              3220-9815-47
                      Licensed to: Nantong Medical College
                              User: Dr. Feng Chen


                          Press Any Key to Continue
```

按提示，压任意键，进入下一页。

```
 What kind of file would you like to transfer FROM?

                        1-2-3 V1-2.x Worksheet
                        1-2-3 V3.x or Windows Worksheet
                        Alpha Four File
                        Clipper File
                        dBASE II File
                        dBASE III or IV File
                        Excel Worksheet
                        FoxBASE File
                        Gauss System File
                        Paradox Table
                        Quattro Pro Worksheet
                        SAS Transport File
                        SPSS Export File
                        Stata System File
                        Symphony Worksheet
                        SYSTAT System File

      Move the menu pointer to your selection and press [ENTER].
                Press [ESCAPE] to leave Stat/Transfer.
```

第二步，选择原文件的格式。transfer**可以将很多格式的文件转为其它格式。原文件的格式有：**1-2-3，dbaseII，dbaseIII，Foxbase，SAS，SPSS，SYSTAT**和**Stata。**本例选**Foxbase。**选好原文件格式后，程序自动进入下一个页面。**

```
   Transferring FROM: FoBASE File

    What kind of file would you like to transfer TO?

                     Alpha Four File
                     Clipper File
                     dBASE II File
                     dBASE III or IV File
                     FoxBASE File
                     Gauss System File
                     Paradox Table
                     SAS Transport File
                     SPSS Export File
                     Stata System File
                     SYSTAT System File


          Move the menu pointer to your selection and press [ENTER].
            Press [ESCAPE] choose a different source file type.
```

第三步，选择目标文件格式。本例为Stata格式文件。
第四步，选择原文件名，则出现如下页面。

```
                        S T A T  /  T R A N S F E R?
You may select individual variables or
press * to select all of the variables.

                                                      ▸ X        byte
                                                      ▸ Y        float
Use the up and down arrow keys to move the cursor.
[SPACE] toggles the selection marker.
The period key [.] anchors a range of variables.
[+] selects a range of output variables.
[-] removes a range of output variables.
Selected variables are marked with an arrow ' '.

To change the target type of numeric variables,
type:  'b' -> byte
       'i' -> integer
       'l' -> long
       'f' -> float
       'd' -> double
       'D' -> Date

Press [ENTER] when you are finished.
Press [ESCAPE] to select a different file.
```

该页面是让用户选择要转换的变量，变量列于页面的右则。选择变量的方法有3种。

1）如选择所有的变量，则按键盘上的 * 键。

2）如选择部分变量，则在要选的变量上按[SPACE]键，选中的变量前会出现符号‣，[SPACE]键是一个开关键，如变量选错，再按[SPACE]键，则可消除变量的符号‣，即放弃该变量。

3）如果需选的变量较多，且这些变量是紧挨着的，则可用在第一个变量上按小数点键 . ，再用加号键 + 选择要转换的变量，直到最后一个变量被选入。当然，用减号键 – 可将选入的变量消除。

选好变量后，按回车[Enter]键，即完成转换。此时，在d:\mydata\中就有一个文件名为ex3-1.dta的Stata格式文件。注意，transfer转换后的文件自动存放在与原文件相同的子目录中。该文件可在Stata中用“ use 文件名 ”命令打开。

五、 直接读入Stata格式数据

如果数据已按Stata格式存盘，则在下次调用时，直接键入“ use 文件名 ”即可。本例：

```
. use d:\mydata\ex3-1
```

后缀dta被自动省略。此时，数据被读入内存。

§3.2 数据库、变量、数值的说明

一、 描述数据库

用命令describe可描述数据库，查看数据库的基本情况。假定数据库ex3-1.dta已调入内存，键入describe或 F3 即可。

```
. drop _all
. use d:\mydata\ex3-1
. describe
Contains data from d:\mydata\ex3-1.dta
  Obs:       8 (max=  4720)
  Vars:      2 (max=    99)                     25 Dec 1998 10:20
  Width:     4 (max=   200)
  1. x0           int    %8.0g
  2. x1           int    %8.0g
Sorted by:
```

结果显示，该数据文件在d:\mydata\子目录中，当前数据库ex3-1.dta可容许最多4720个记录和最多99个变量(不同版本的Stata及计算机内存，最大记录数和最大变量数不同，见1.1)，实际情况是，2个变量 x0,x1和8个记录。变量x0和x1是宽度为8的g格式数值变量。g格式是变量的显示格式之一，还有e格式，f格式和字符串格式，详见§3.4。结果中还显示了变量创建的时间。

二、 对变量作说明

实际分析资料时，为了方便，总将观察指标用变量名来代替，但当变量较多时，可能会搞不清楚每个变量代表的是哪一个指标。此时，可用label var命令对变量加一些说明，以示区别。

```
. label var x0 "before treat"
. label var x1 "after  treat"
```

对变量作了说明后，在任何时候，当我们对变量的含义不太清楚时，就可用describe命令来回忆。

```
. des
Contains data from d:\mydata\ex3-1.dta
  Obs:     8 (max=  4712)
 Vars:     2 (max=    99)                    23 Dec 1998 09:20
Width:     4 (max=   200)
  1. x0           int    %8.0g                before treat
  2. x1           int    %8.0g                after  treat
Sorted by:
```

结果中给出了x0与x1的说明。

三、 对数值作说明

处理分类资料常常要对其取值作说明。作为例子，在数据文件ex3-1.dta中增加一个变量group，1表示回答“yes”(是)，0表示回答“no”(否)，并规定，结论为奇数时取1，偶数时取0。要对取值作说明，用label define命令。

为便于说明，我们先产生变量group，规定偶数记录取1，奇数记录取0。

```
. gen group=1-mod(_n,2)
. list
          x0        x1       group
  1.    3550      2450           0
  2.    2000      2400           1
  3.    3000      1800           0
  4.    3950      3200           1
  5.    3800      3250           0
  6.    3750      2700           1
  7.    3450      2500           0
  8.    3050      1750           1
```

现对group的数值作说明:

```
. lab define group  1 "yes" 0 "no"
```

要查看变量取值的说明，用lab list命令：

```
. lab list group
group:
           1 yes
           0 no
```

四、 对数据库作说明

对数据库作说明，用label data命令。

```
. lab data "paired design, 23 Dec 1998"
. describe
Contains data from d:\mydata\ex3-1.dta
  Obs:     8 (max=  4712)                    paired design, 23 Dec 1998
 Vars:     3 (max=    99)                    23 Dec 1998 09:20
Width:     8 (max=   200)
  1. x0           int    %8.0g               before treat
  2. x1           int    %8.0g               after  treat
  3. group        float  %9.0g
Sorted by:
Note:  Data has changed since last save
```

对变量、数值或文件作说明后，最后要将数据存盘，执行以下命令：

```
. save ,replace
file d:\mydata\ex3-1.dta saved
```

§3.3 数据库的维护

Stata提供了很强的数据库维护功能，可以对数据排序，可以增删数据记录或变量，可以横向或纵向链接等。

一、 数据库的排序

排序的命令为：

sort 变量清单

该指令将按变量数值的上升序列重排数据，变量清单表示可有若干变量。例：

```
. use d:\mydata\ex3-1
. sort x0
. list
          x0         x1
  1.    2000       2400
  2.    3000       1800
  3.    3050       1750
```

```
  4.     3450      2500
  5.     3550      2450
  6.     3750      2700
  7.     3800      3250
  8.     3950      3200
```

二、 删除变量或记录

删除变量或记录的命令为drop。如：

```
. drop x1 x2            删除变量x1和x2
. drop x1-x5            删除数据库中介于x1和x5间的所有变量(包括x1和x5)
. drop if x<0           删去x<0的所有记录
. drop in 10/20         删去第10~20个记录
. drop if x==.          删去x为缺失值的所有记录
. drop if x==.|y==.     删去x或y之一为缺失值的所有记录
. drop if x==.&y==.     删去x和y同时为缺失值的所有记录

. drop _all             删掉数据库中所有变量和数据
```

三、 保留变量或记录

保留变量或记录的命令是keep。keep是drop的逆操作，语法结构与drop完全一样，只是其结果刚好相反。如：

```
. keep in 10/20         保留第10～20个记录
. keep x1-x5            保留数据库中介于x1和x5间的所有变量(包括x1和x5)，其余变量删除
. keep if x>0           保留x>0的所有记录
```

四、 纵向链接数据库

将两个变量相同的数据库纵向联接起来。命令是“ append using 文件名 ”。

例3.2 将下列文件ex3-2.dta与ex3-1.dta上下联接。

```
. use d:\mydata\ex3-2
. list
          x0        x1       g
  1.    2450      1450       2
  2.    2100      2400       2
  3.    2300      3800       2
  4.    1590      4200       2
```

该文件有4条记录，3个变量，其中变量x0,x1与文件ex3-1.dta相同，而变量g只有ex3-2.dta中有，ex3-1.dta中无该变量，将该文件与ex3-1.dta联接：

```
. append using d:\mydata\ex3-1
. list
```

	x0	x1	g
1.	2450	1450	2
2.	2100	2400	2
3.	2300	3800	2
4.	1590	4200	2
5.	3550	2450	.
6.	2000	2400	.
7.	3000	1800	.
8.	3950	3200	.
9.	3800	3250	.
10.	3750	2700	.
11.	3450	2500	.
12.	3050	1750	.

结果，新产生的数据库，记录数为12(两数据库记录数之和)，变量数为3(两数据库变量的并集)。相同的变量上下相接，ex3-1.dta中没有的变量，在新数据库中用缺失值代替。

五、 横向链接数据库

将两个文件按关键变量横向联接。命令是“merge 关键变量 using 文件名”。其中，两个库中均必须要有相同变量名的“关键变量”，并按“关键变量”排序，才可将两个数据库横向联接。

例3.3 设有如下两个文件：

ex3-3.dta

bh	x0	x1
1	12	24
2	15	26
3	16	49
4	18	57
5	20	68

ex3-4.dta

bh	y0	y1	x0
1	35	79.2	2
3	45	47.4	8
4	52	34.6	6
6	66	28.0	9

ex3-3.dta中，bh为1～5，而在ex3-4.dta中，bh取1,3,4,6，这些记录号与ex3-3.dta中不一样，比ex3-3.dta中少了2和5，但多了bh=6。不妨将ex3-3.dta称为被动库，将ex3-4,dta称为主动库，将主动库中的数据加到被动库中。步骤如下：

```
. drop _all                                  /*  清空数据库
. use d:\mydata\ex3-4                        /*  调用ex3-4.dta文件
. sort bh                                    /*  按bh从小到大排序
. save ,replace                              /*  将排好序的ex3-4.dta存盘，并替换原有文件
file d:\mydata\ex3-4.dta saved
. use d:\mydata\ex3-3,replace                /*  调用ex3-3.dta文件
. sort bh                                    /*  按bh从小到大排序
```

```
. merge bh using d:\mydata\ex3-4          /*  按bh将两个数据库左右相拼
. list
        bh        x0        x1        y0        y1    _merge
 1.      1        12        24        35      79.2         3
 2.      2        15        26         .         .         1
 3.      3        16        49        45      47.4         3
 4.      4        18        57        52      34.6         3
 5.      5        20        68         .         .         1
 6.      6         .         .        66        28         2
```

结果，Stata自动产生了新变量_merge，该变量的取值为1，2，3，当该记录只在被动库中出现而未在主动库中出现时，该记录的_merge＝1；当该记录只在主动库中出现而未在被动库中出现时，该记录的_merge＝2；在两个库中均出现时，_merge=3。记录空缺的用缺失值(小数点)来代替。

六、 产生新变量和替换变量数值

产生新变量用：

generate 新变量＝表达式

替换原变量的取值：

replace 变量＝表达式

如：

```
. generate bh=_n                          /*  将数据库的内部编号赋给变量bh。
. generate group=int((_n-1)/5)+1          /*  按当前数据库的顺序，依次产生5个1，5个2，5个
                                              3……。直到数据库结束。
. generate block=mod(_n,6)                /*  按当前数据库的顺序，依次产生1,2,3,4,5,0。
                                              Mod是模数运算函数。
. replace  bolck=6 if block==0            /*  将block＝0的数全部替换为6。
. generate y=log(x) if x>0                /*  产生新变量y，其值为所有x>0的对数值log(x)，当x<=0
                                              时，用缺失值代替。
. replace  z=. if z==9                    /*  将所有z=9的值用缺失值代替。
```

要改变某个已存在变量的数值，只能用命令replace。

七、 变量更名

变量的更名用：

rename 原变量名 新变量名

如：

```
. rename x y                              /*  将变量名x改为y
```

八、 增加空白记录

当需要增加一个或多个记录时，可先用指令des查看当前数据库的最大记录数，再在此基础上增加若干空白记录。

例如，若当前最大记录数是40，需增加2个空白记录，这时的操作为：

```
. set obs 42
```

九、 更新数据库

将修改后的数据用原文件名存盘，替代原文件：

```
. save 数据文件名，replace
```

§3.4 规定显示格式

用指令des后，已多次看到 "float %9.0g"，它表示变量以Stata的g格式显示，这是Stata变量的默认格式。Stata的另外两种显示格式是e格式和f格式。为显示这三种格式的差别，现更名x为g_fmt，并产生存贮内容与g_fmt完全相同的新变量e_fmt和 f_fmt。

```
. drop _all                    /*  清空数据库
. use ex3-1                    /*  调用ex3-1.dta文件
. ren x g_fmt                  /*  将变量名x改为g_fmt
. gen e_fmt=g_fmt              /*  产生新变量e_fmt=g_fmt
. gen f_fmt=g_fmt              /*  产生新变量f_fmt=g_fmt
. des                          /*  描述数据库
Contains data from ex3-1.dta
  Obs:     5 (max=  2333)
 Vars:     3 (max=    99)
    1. g_fmt        float  %9.0g
    2. e_fmt        float  %9.0g
    3. f_fmt        float  %9.0g
```

可以看到，g_fmt和e_fmt都是g格式，现用指令format改变变量e_fmt和f_fmt的显示格式:

```
. format e_fmt  %9.2e
. format f_fmt  %9.2f
. des
Contains data from ex3-2.dta
  Obs:     5 (max=  2333)
 Vars:     3 (max=    99)
    1. g_fmt        float  %9.0g
    2. e_fmt        float  %9.2e
    3. f_fmt        float  %9.2f
Sorted by:
Note:  Data has changed since last save
```

可以看到，显示格式已经发生改变。 用命令list可以看到三种显示格式的差别：

```
. list g_fmt e_fmt f_fmt
        g_fmt       e_fmt       f_fmt
  1.      2.8    2.80e+00        2.80
  2.  3962323    3.96e+06  3962322.50
  3.     4.85    4.85e+00        4.85
  4. 5.60e-06    5.60e-06        0.00
  5.     6.26    6.26e+00        6.26
```

e格式是用科学记数法，如在“ %9.2e ”格式下32.5显示为3.25e+01，3.25显示为3.25e+00；f格式是通常表示法，严格按照规定的小数位数显示，如0.0000056在“ %9.2f ”格式显示为的0.00；g格式根据可读性自动选择e或f格式显示，并不受规定的小数位数的限制，如1.1在“ %9.2g ”格式下显示为“ 1.10 ”，在“ %9.1g ”格式下显示为1.1，而在"%9.0g"格式下还是显示为1.1。命令“ format x %9s ”规定x是一个可以接收9个字符的字符串变量。

第四章　统计描述及区间估计

本章介绍资料的统计描述和统计量的区间估计。

§4.1　统计资料的一般描述

统计描述在统计分析过程中占有相当重要的地位，必须给予充分重视。通过统计描述，我们不仅可以对整个数据的概貌、分布状况有个大致的了解，对各因素或变量间的相互关系有个初步的结论，而且还可发现数据中的异常现象，为进一步分析选择方法提供依据。而数据的可靠性，正是保证统计分析正确揭示客观规律的前提条件。因此，在进行任何统计分析之前，必须对分析数据进行全面的描述。

Stata 具有很强的统计描述功能，可用统计量(数值)描述，也可用图形描述。本章介绍统计量描述，图形描述见第五章。部分专用统计描述指令穿插在有关章节讲述。如描述指标间的相关性安排在第九章，生存率的描述安排在第十六章等。

注意：统计描述是对分析数据进行描述，而第三章中的指令 describe 是对数据库的结构进行描述。

一、　数值变量资料的描述

对一组数值变量资料的描述，最常用的统计量有均数、标准差、百分位数、偏度系数与峰度系数、变异系数等。主要命令有 summary 与 centile。

```
summarize    [变量名]  [, detail ]
centile [变量名] [, centile(# [# ...]) cci normal meansd level(#) ]
```

这里的选择项分别表示：

```
detail        /*  详细描述，缺失时为简单描述
meansd        /*  指定百分位数用近似正态法，缺失时为直接算法
cci           /*  指定百分位数的可信区间用保守算法
normal        /*  指定百分位数的可信区间用近似正态法
level(#)      /*  指定百分位数的可信区间的可信限
```

下面看一个例子。

例 4.1　某市 1982 年 110 名 7 岁男童的身高(cm)资料如下:

112.4	117.2	122.7	123.0	113.0	110.8	118.2	108.2	118.9	118.1	123.5
118.3	120.3	116.2	114.7	119.7	114.8	119.6	113.2	120.0	119.7	116.8
119.8	122.5	119.7	120.7	114.3	122.0	117.0	122.5	119.8	122.9	128.0
121.5	126.1	117.7	124.1	129.3	121.8	112.7	120.2	120.8	126.6	120.0
130.5	120.0	121.5	114.3	124.1	117.2	124.4	116.4	119.0	117.1	114.9
129.1	118.4	113.2	116.0	120.4	112.3	114.9	124.4	112.2	125.2	116.3
125.8	121.0	115.4	121.2	117.9	120.1	118.4	122.8	120.1	112.4	118.5

```
113.0   120.8   114.8   123.8   119.1   122.8   120.7   117.4   126.2   122.1   125.2
118.0   120.7   116.3   125.1   120.5   114.3   123.1   122.4   110.3   119.3   125.0
111.5   116.8   125.6   123.2   119.5   120.5   127.1   120.6   132.5   116.3   130.8
```

首先对资料作简单描述。设数据已被存入 d:\mydata\ex4-1.dta。

```
. drop _all
. use ex4-1
. summ
Variable |     Obs        Mean    Std. Dev.       Min        Max
---------+--------------------------------------------------------
       x |     110    119.7273    4.741325      108.2      132.5
```

这里，只用了 summ 命令，没有加任何选择项。结果中给出了变量 *x* 的样本含量(Obs)、均数(Mean)、标准差(Std.Dev.)、最小值(Min)、最大值(Max)。

要得到更多的信息，需要加选择项“detail”(或 d)：

```
. summ  x , d
                            x
-------------------------------------------------------------
      Percentiles      Smallest
 1%         110.3          108.2
 5%         112.3          110.3
10%         113.1          110.8       Obs                 110
25%         116.4          111.5       Sum of Wgt.         110

50%         119.9                      Mean           119.7273
                        Largest        Std. Dev.      4.741325
75%         122.8          129.3
90%         125.7          130.5       Variance       22.48017
95%           128          130.8       Skewness       .1524946
99%         130.8          132.5       Kurtosis       2.921794
```

除样本含量，均数，标准差外，结果中还给出了 9 个百分位数(Percentiles)，即 1%，5%，10%，25%，50%，75%，90%，95%和 99%，他们依次是：110.3，112.3，113.1，116.1，119.9，122.8，125.7，128.0 和 130.8；给出了 4 个最小数和 4 个最大数；方差(Varance)，偏度系数 (Skewness)与峰度系数(Kurtosis)。对正态分布来说，偏度系数=0，峰度系数=3。偏度系数为 0 时称为对称分布，大于 0 为正偏态，小于 0 为负偏态；峰度系数为 3 时称为正态峰，大于 3 为尖峭峰，小于 3 为平阔峰。

如欲得到更多的百分位数，则用命令“centile”。

```
. centile x , centile(2.5,50,97.5)

                                                  -- Binom. Interp. --
```

```
Variable |     Obs  Percentile      Centile        [95% Conf. Interval]
---------+-------------------------------------------------------------
       x |     110        2.5      110.6875           108.2     112.389*
         |                 50         119.9        118.9211    120.5789
         |               97.5      130.5675        127.1988       132.5*
* Lower (upper) confidence limit held at minimum (maximum) of sample
```

我们在选择项 centile 中指定了 3 个百分位数，即 2.5%，50%和 97.5%。结果中除给出了百分位数，同时还给出了百分位数的 95%可信区间。如 2.5%分位数为 110.6875，其 95%的可信区间为(108.2,112.389)，这里的*号表示可信区间的下限已达到所给数据的最小值(108.2)。

这里，百分位数的可信区间是按二项分布用插值法求出的。也可用近似正态法，只需加上选择项 norm。

```
. centile x , centile(2.5,50,97.5) norm
                            -- Normal, based on observed centiles --
Variable |     Obs  Percentile      Centile        [95% Conf. Interval]
---------+-------------------------------------------------------------
       x |     110        2.5      110.6875        108.5527    112.8223
         |                 50         119.9        118.7888    121.0112
         |               97.5      130.5675        125.8348    135.3002
```

加上选择项 norm 后，所得百分位数相同，但可信区间不同。Stata 还提供了另一种保守的基于二项分布的百分位数可信区间算法 cci(conservative confidence interval)。

```
. centile x , centile(2.5,50,97.5) cci
                                                -- Binomial Exact --
Variable |     Obs  Percentile      Centile        [95% Conf. Interval]
---------+-------------------------------------------------------------
       x |     110        2.5      110.6875           108.2       112.4*
         |                 50         119.9           118.9       120.6
         |               97.5      130.5675           127.1       132.5*
* Lower (upper) confidence limit held at minimum (maximum) of sample
```

该法所得可信区间一般比插值法要宽。

上述百分位数用直接法计算的，Stata 提供了正态分布算法，即按公式：

$$\bar{x} + u_a s \tag{4.1}$$

如本例，$\bar{x}$ =119.7273，s=4.741325，故 2.5%分位数为：

119.7273-1.96×4.741325=110.4343

这只需在 centile 命令中增加选择项 meansd。

```
. centile x , centile(2.5) meansd
                            -- Normal, based on mean and std. dev.--
Variable |     Obs  Percentile      Centile        [95% Conf. Interval]
---------+-------------------------------------------------------------
```

```
       x |     110         2.5      110.4344        108.9156    111.9533
```

此时，百分位数的可信区间的算法也是基于正态分布的。

二、 分类变量资料的描述

对分类资料一般用率、构成比、比来描述某事物的发生强度、频率或构成，相应的命令为：

tabulate **变量名** [, generate(**新变量**) missing nofreq nolabel plot]

tab1 **变量** 1 **变量** 2 **变量** 3…… [, missing nolabel plot]

tabulate **变量** 1 **变量** 2 [, cell column row missing nofreq]

tab2 **变量** 1 **变量** 2 **变量** 3 ……[, tabulate_options]

其中，前两个命令用于单变量的分类描述，后两个命令用于两个变量的交叉分类描述。选择项的意义：

generate(**新变量**)	/*	**按分组变量产生哑变量**
nofreq	/*	**不显示频数**
nolabel	/*	**不显示数值标记**
plot	/*	**显示各组频数图示**
missing	/*	**包含缺失值**
cell	/*	**显示各小组的构成比(小组之和为** 1)
column	/*	**按栏显示各组之构成(各栏总计为** 1)
row	/*	**按行显示各组之构成(各行总计为** 1)

例 4.2 有三组(group)**患者，男女**(sex)**若干人，**sex=1 **表示男性，**sex=0 **表示女性。测得其血红蛋白浓度**(x1,%)**和红细胞计数**(x2,**万**/mm^3)**，资料存入** d:\mydata\ex4-2.dta**。试对其进行描述。**

```
. use d:\mydata\ex4-2
. list
```

	x1	x2	group	sex
1.	3.9	210	1	0
2.	4.2	190	1	0
3.	3.7	240	1	0
4.	4	170	1	0
5.	4.4	220	1	0
6.	5.2	230	1	0
7.	2.7	160	1	0
8.	2.4	260	1	0
9.	3.6	240	1	1
10.	5.5	180	1	1
11.	2.9	220	1	1
12.	3.3	300	1	1
13.	4.8	270	2	0
14.	4.7	180	2	0

```
 15.      5.4       230         2         0
 16.      4.5       245         2         0
 17.      4.6       270         2         1
 18.      4.4       220         2         1
 19.      5.9       290         2         1
 20.      5.5       290         2         1
 21.      4.3       220         2         1
 22.      5.1       310         2         1
 23.      4.4       250         2         1
 24.      3.7       305         3         1
 25.      2.9       330         3         1
 26.      4.5       240         3         1
 27.      3.3       195         3         1
 28.      4.5       275         3         0
 29.      3.8       310         3         0
 30.      3.7       240         3         0
```

首先看看各组的频数。

```
. tab group
      group|      Freq.     Percent        Cum.
-----------+-----------------------------------
         1 |         12       40.00       40.00
         2 |         11       36.67       76.67
         3 |          7       23.33      100.00
-----------+-----------------------------------
     Total |         30      100.00
```

结果显示，各组的样本含量分别为：12，11，7。**产生组变量的哑变量，分别以** g1,g2,g3 **表示：**

```
. tab group , gen(g)
      group|      Freq.     Percent        Cum.
-----------+-----------------------------------
         1 |         12       40.00       40.00
         2 |         11       36.67       76.67
         3 |          7       23.33      100.00
-----------+-----------------------------------
     Total |         30      100.00
```

这样，Stata **自动产生** group **的** 3 **个哑变量(**group **有** 3 **组)，命令中用** g **表示哑变量，**Stata **自动以** g1,g2,g3 **表示，结果如下：**

```
. list group g1-g3
```

```
      group         g1         g2         g3
 1.       1          1          0          0
 2.       1          1          0          0
 3.       1          1          0          0
 4.       1          1          0          0
 5.       1          1          0          0
 6.       1          1          0          0
 7.       1          1          0          0
 8.       1          1          0          0
 9.       1          1          0          0
10.       1          1          0          0
11.       1          1          0          0
12.       1          1          0          0
13.       2          0          1          0
14.       2          0          1          0
15.       2          0          1          0
16.       2          0          1          0
17.       2          0          1          0
18.       2          0          1          0
19.       2          0          1          0
20.       2          0          1          0
21.       2          0          1          0
22.       2          0          1          0
23.       2          0          1          0
24.       3          0          0          1
25.       3          0          0          1
26.       3          0          0          1
27.       3          0          0          1
28.       3          0          0          1
29.       3          0          0          1
30.       3          0          0          1
```

这一命令在广义线性回归中是很有用的。

再看看各组性别分布情况。

```
. tab group sex

           | sex
     group |         0          1 |     Total
-----------+----------------------+----------
         1 |         8          4 |        12
         2 |         4          7 |        11
```

```
        3 |         3          4 |         7
----------+----------------------+----------
    Total |        15         15 |        30
```

欲了解各组男女构成，在命令中加 row **选择项：**

```
. tab group sex, row
           | sex
     group |         0          1 |     Total
-----------+----------------------+----------
         1 |         8          4 |        12
           |     66.67      33.33 |    100.00
-----------+----------------------+----------
         2 |         4          7 |        11
           |     36.36      63.64 |    100.00
-----------+----------------------+----------
         3 |         3          4 |         7
           |     42.86      57.14 |    100.00
-----------+----------------------+----------
     Total |        15         15 |        30
           |     50.00      50.00 |    100.00
```

欲了解各组构成，在命令中加 cell **选择项：**

```
. tab group sex, cell
           | sex
     group |         0          1 |     Total
-----------+----------------------+----------
         1 |         8          4 |        12
           |     26.67      13.33 |     40.00
-----------+----------------------+----------
         2 |         4          7 |        11
           |     13.33      23.33 |     36.67
-----------+----------------------+----------
         3 |         3          4 |         7
           |     10.00      13.33 |     23.33
-----------+----------------------+----------
     Total |        15         15 |        30
           |     50.00      50.00 |    100.00
```

三、　分类变量与连续变量资料的综合描述

欲了解某数值变量资料在各组的均数、标准差等，用综合描述命令：

tab **分组变量** ， summ(**数值变量**)

tab **分组变量** 1 **分组变量** 2 ， summ(**数值变量**)

前者用于按一个变量分类，后者用于按两个变量分类。summ **后每次只能指定一个数值变量。**

例 4.3　对例 4.2 **资料，计算血红蛋白浓度和红细胞计数在各组的均数、标准差。**

```
. tab group, sum(x1)
           |            Summary of x1
     group |       Mean   Std. Dev.       Freq.
-----------+------------------------------------
         1 |  3.8166667   .93889033          12
         2 |  4.8727273   .52932203          11
         3 |  3.7714286   .58513326           7
-----------+------------------------------------
     Total |  4.1933333   .88236879          30

. tab group, sum(x2)
           |            Summary of x2
     group |       Mean   Std. Dev.       Freq.
-----------+------------------------------------
         1 |  218.33333   39.962103          12
         2 |  252.27273   38.299062          11
         3 |  270.71429   48.082271           7
-----------+------------------------------------
     Total |         243   45.383424          30
```

若按分组变量和性别变量交叉分组，则得各交叉分类时血红蛋白浓度的均数：

```
. tab group sex , sum(x1) nofreq
      Means and Standard Deviations of x1
           | sex
     group |         0          1      Total
-----------+----------------------+----------
         1 |    3.8125      3.825 | 3.8166667
           | .90307009  1.1528949 | .93889033
-----------+----------------------+----------
         2 |      4.85  4.8857143 | 4.8727273
           |  .3872984   .6256425 | .52932203
-----------+----------------------+----------
         3 |         4        3.6 | 3.7714286
           | .43588989  .68313003 | .58513326
```

```
-----------+---------------------+----------
     Total | 4.1266667       4.26 | 4.1933333
           |  .8224238   .9627342 | .88236879
```

§4.2 可信区间估计

统计推断有两个重要内容，其一是假设检验，其二是参数的可信区间估计。Stata 提供了均数(正态分布)，率(二项分布)和事件数(Poisson 分布)的可信区间的估计。用于可信区间估计的命令是：

ci 变量 [, level(#) binomial poisson exposure(观察数变量) by(分组变量) total]

Stata 还提供了已知 $n, \bar{x}, s$ 时均数的可信区间估计，已知 n,x 时率的可信区间估计，以及已知 n,x(事件数)时的总体事件数的可信区间估计。相应的命令为：

cii 观察数 均数 标准差 [, level(#)] /* 正态分布
cii 观察数 阳性数 [, level(#)] /* 二项分布
cii 观察数 事件数 , poisson [level(#)] /* Poisson 分布

其中选择项：

level(#) /* 指定可信度，缺失时为 95(%)
binomial/poisson /* 指定总体分布。只能选其中之一，缺失时为正态分布
exposure(观察数变量) /* 指定观察数变量，仅用于 Poisson 分布时
by(分组变量) /* 指定按分组变量分别估计均数的可信区间
total /* 指定除按分组变量估计可信区间外，还对整个数据估计，仅用于 by(分组变量)时

例 4.4 对例 4.2 中资料分别估计各组血红蛋白浓度和红细胞计数均数的可信区间。

```
. use d:\mydata\ex4-2
. sort group                                   /* 在用 by(分组变量)前，必须对分组变量排序
. ci x1 x2, by(group)
-> group=1
Variable |     Obs        Mean    Std. Err.       [95% Conf. Interval]
---------+--------------------------------------------------------------
      x1 |      12    3.816667    .2710343        3.220124    4.413209
      x2 |      12    218.3333    11.53607        192.9426     243.724
-> group=2
Variable |     Obs        Mean    Std. Err.       [95% Conf. Interval]
---------+--------------------------------------------------------------
      x1 |      11    4.872727    .1595966        4.517124    5.228331
      x2 |      11    252.2727     11.5476        226.5431    278.0024
-> group=3
Variable |     Obs        Mean    Std. Err.       [95% Conf. Interval]
```

```
---------+--------------------------------------------------------
      x1 |       7     3.771429    .2211596        3.230271    4.312587
      x2 |       7     270.7143    18.17339        226.2456     315.183
```

结果中给出了各组各变量的样本含量，均数，均数的标准误(Std. Err.)，以及 95%的可信区间(95% conf. Interval)。

如果已知各组均数，则可用 cii 命令直接估计，如第一组血红蛋白浓度均数的可信区间：

```
. cii 12 3.816667 0.2710343
Variable |     Obs        Mean    Std. Err.       [95% Conf. Interval]
---------+--------------------------------------------------------
         |      12     3.816667    .0782409         3.64446    3.988874
```

结果与从原始资料估计所得结果相同。估计 90%的可信区间的命令为：

```
. cii  12 3.816667 0.2710343, level(90)
```

(结果略)

例 4.5 某地抽查了 10 名献血员的乙肝表面抗原(HBsAg)携带情况，阳性人数为 2，试估计该地 HBsAg 阳性率。

直接用 cii 命令：

```
  . cii 10  2
                                                    -- Binomial Exact --
Variable |     Obs        Mean    Std. Err.       [95% Conf. Interval]
---------+--------------------------------------------------------
         |      10          .2    .1264911        .0251953       .55625
```

结果，阳性率为 0.2，标准误为 0.1265，阳性率的 95%可信区间为：(0.0252, 0.5563)。

例 4.6 将一个面积为 $100cm^2$ 的培养皿置于某病房，1 小时后取出，培养 24 小时，查得 8 个菌落，求该病房平均每 $100cm^2$ 的面积细菌数的 95%可信区间。

```
. cii 1 8 , poisson
                                                    -- Poisson  Exact --
Variable | Exposure        Mean    Std. Err.       [95% Conf. Interval]
---------+--------------------------------------------------------
         |       1           8    2.828427          3.454    15.76225
```

这里的 1 表示 1 个 $100cm^2$ 的面积。即病房平均每 $100cm^2$ 的面积细菌数的 95%可信区间为(3.5,15.8)。

第五章 Stata 的绘图功能

统计作图是 Stata **又一强大的功能。**Stata **的作图命令简洁，图形精美，应用者可随心所欲，并可充分发挥想象力。**Stata **的作图命令** graph **主要提供如下八种基本图形的制作：直方图**(histogram)**，条形图**(bar)，**百分条图**(oneway)**，百分圆图**(pie)**，散点图**(twoway)**，散点图矩阵**(matrix)**，星形图**(star)**，分位数图。在有些非绘图命令中，也提供了专门绘制某种图形的功能，如在生存分析中，提供了绘制生存曲线图等。

§5.1 几种常见的统计图

首先展示几种 Stata 绘制的常见的统计图。读者可以先有一个直观印象。

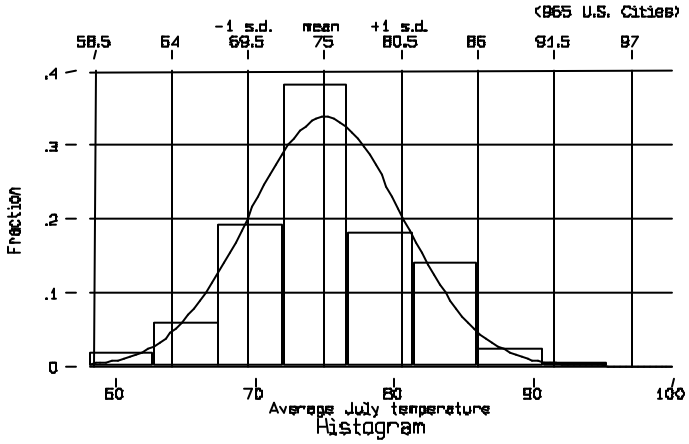


图 5.1 直方图

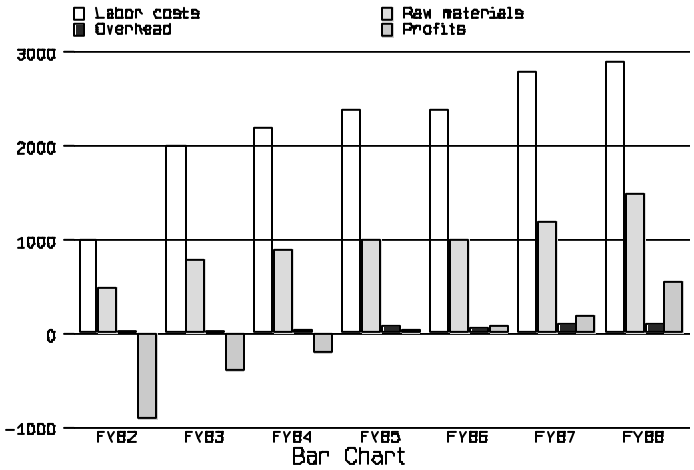


图 5.2 直条图

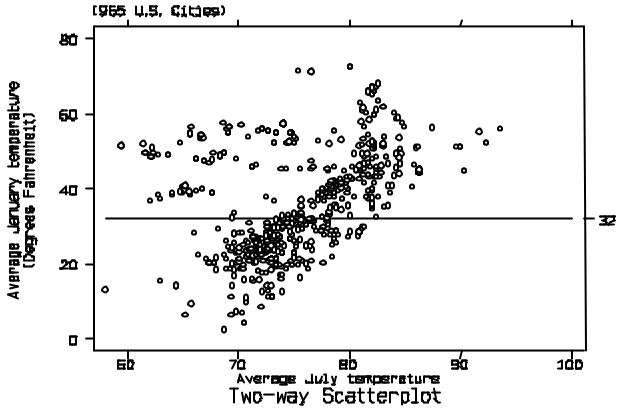


图 5.3　　散点图

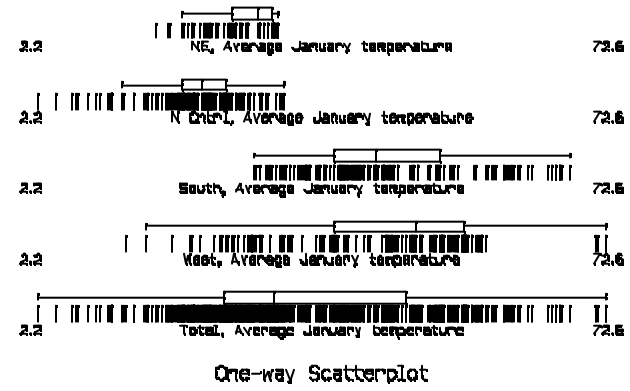


图 5.4　　单变量散点图与箱式图

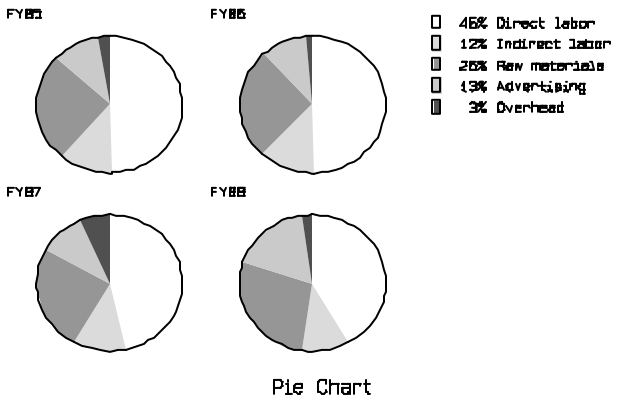


图 5.5　　百分圆图

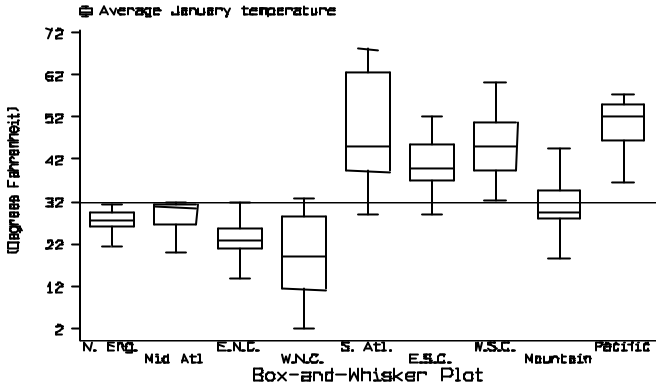


图 5.6 Box-Whisker 的箱式图

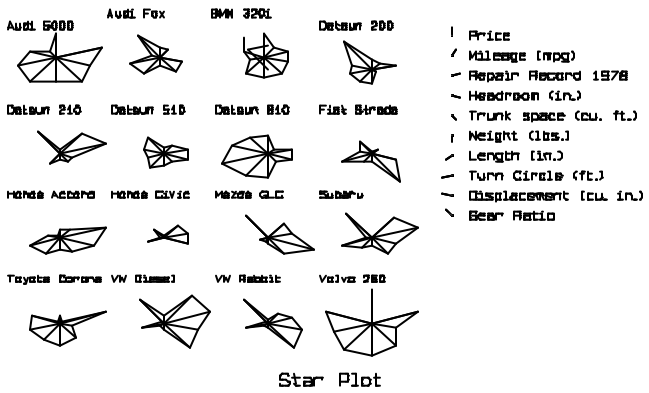


图 5.7 星状图

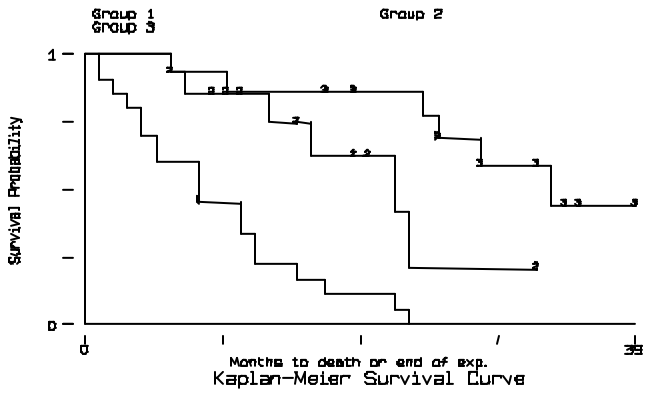


图 5.8 Kaplan-Meier 生存曲线图

§5.2 几种常见统计图的绘制

这里重点讲述直方图(histogram)，条图(bar)和，圆饼图(pie)，一维、二维散点图和线图(twoway)、箱式图(box)的制作。其它统计图安排在相应的章节结合实例讲述。需要时可用命令 help graph 了解详细内容。

一、 直方图的制作

直方图主要用于表示数值变量资料的分布。常以横轴表示被观察对象，纵轴表示频数或频率。绘制直方图的命令为：

graph 变量名] [, 选择项]

这里的选择项有：

bin(#)	/* 将数据分为几组，#为数字，缺省值为 5。
freq	/* 指定以频数为纵轴画图，缺省时为以频率为纵轴。
normal[(#,#)]	/* 在直方图上加上正态分布曲线，N(#,#)，前一个#为均数，后一个为方差，缺省值为原资料的均数与方差。
density(#)	/* 与 normal 合用，表示在画正态曲线时的光滑程度。缺省值为 100。
shading(s)	/* 定义直方图的阴影。范围在 1～4，缺省值为 3。
axis/ noaxis	/* 画/不画坐标轴，缺省值为画坐标轴。
border/noborder	/* 画/不画边框，缺省值为不画。
xlable/ylable/tlable/rlable[(#,…,#)]	/* lable 是指在坐标轴上画上坐标点及相应的数据，xlable,ylable, tlable, rlable 分别表示 x 轴、y 轴、上边的轴、右边的轴。下同。缺省时只画 x 和 y 轴的最小值和最大值。
xtick/ytick/ttick/rtick[(#,…,#)]	/* tick 是指在相应的坐标轴上画上坐标点，但不画数据。
xline/yline/tline/rline[(#,…,#)]	/* line 表示以相应的坐标画线。
xscale/yscale[(#,,#)]	/* 分别指定 x 轴和 y 轴的最小和最大坐标点。
title(["][字符串]["])	/* 给图加上总标题。Stata 不接受中文字符。
b1/l1/t1/rl(["][字符串]["])	/* 给各坐标轴加上标题，b 表示底轴(x 轴),l 表示左轴(y 轴)，t 表示上边 的轴，r 表示右边的轴。下同。
b2/l2/t2/r2(["][字符串]["])	/* 给各坐标轴加上副标题。
gap(#)	/* 调整标题与坐标轴的间距。范围为 1～8，缺省值为 8。
saving(文件名[,replace])]	/* 将图形存盘，replace 表示替换原有图形文件。

例 5.1 对例 4.1 资料绘制直方图。

最简单的命令是：

```
. use d:\mydata\ex4-1
. gra x
```

gra 是 graph 的简写形式。因为无选择项，所以 Stata 给出最简单的图形，即分 5 组，以频率表示，给出了 x 轴的最小、最大值，y 轴标有 0(直方图必须从 0 开始)，和 5 组中的最大频率；并各标有另外 3 个等间隔的坐标点。见图 5.9。(为方便印刷，图中阴影部分在用 WORD 处理时去掉了)。

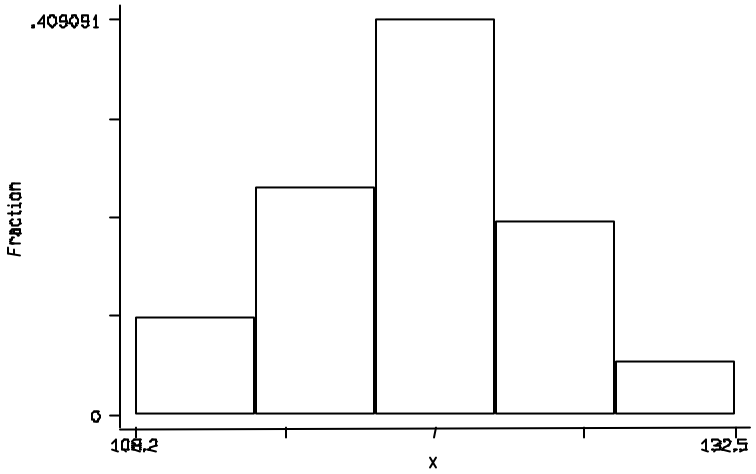


图 5.9 例 4.1 资料的直方图

适当选用选择项可以使图形更精细。如：

```
. gra x, bin(9) freq xlab(108,111,114,117,120,123,126,129,132,135) ylab(0,5,10, 15,20,25,30,35)
      norm gap(4) b2("height  (CM)")
```

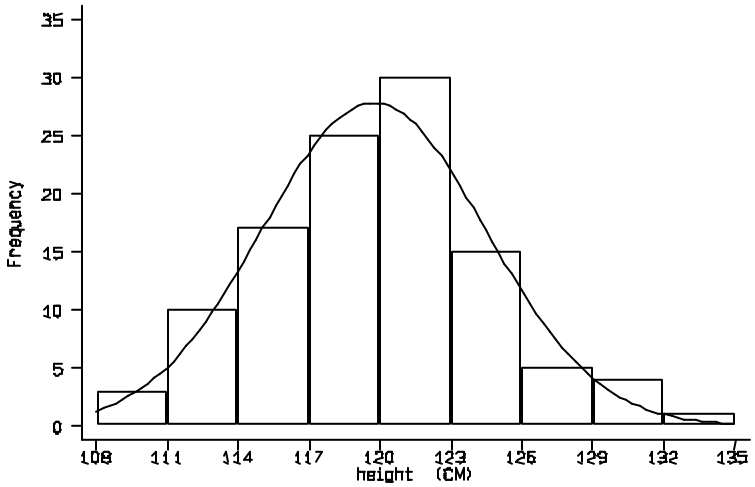


图 5.10 例 4.1 资料的直方图及正态分布曲线

这里，Stata 按要求将资料分为 9 组(bin(9))；用频数作为纵坐标(freq)；画上了正态曲线(norm)，其均数和标准差直接取 x 的均数和标准差，分别为 119.7273 和 22.4802；并在纵坐标和横坐标上分别加上等间隔的刻度(xlab,ylab)；将坐标轴的标题与图形间的距离缩小(gap)，并定义横坐标为：“height (CM)”。这样画出的图形就较精确、美观。

Stata 允许在图形的每一边(上，下，左，右)最多各加两条说明或标题，相应的选择项分别是 t1，t2，b1，b2，l1，l2 及 r1 和 r2，然后紧跟括号，并在括号内写上相应的说明或标题名，若说明中本身就有括号，则需用引号将说明语句引起来。

二、 条图的制作

用等宽直条的长短来表示相互独立的各指标的取值大小。有单式条图和复式条图。绘制条图的命令格式如下：

graph 变量 1 [变量 2[...]], bar [选择项]

这里的选择项有：

means	/* 用均数而不是用总数来表示该指标的值。缺省值为总数
stack	/* 将各指标的堆积起来，而不是并排。缺省值为并排
accumulate	/* 将各指标的值逐次累加
totle	/* 增加各指标的总和直条
shading(s)	/* 定义直方图的阴影。范围在 1～4，缺省值为 3
axis/ noaxis	/* 画/不画坐标轴，缺省值为画坐标轴
[no]alt	/* 将横坐标的刻度错开排放。缺省值为放在一排
border/noborder	/* 同直条图

其它一些选择项的意义与直方图是一样的，有：xlable/ylable/tlable/rlable[(#,...,#)], ytick[(#,...,#)], yline[(#,...,#)], yscale[(#,,#)] , title([''][字符串]['']), b1/l1/t1/rl([''][字符串]['']), b2/l2/t2/r2([''][字符串]['']), gap(#), saving(文件名[,replace])。

这里要注意的是，tick，line，scale 只对 y 轴有效，其它轴无效。

例 5.2 某地二年三种疾病的死亡率如表 5.1，请绘制复式条图(每种疾病为一组，每组有两个直条，分别代表两个年度，条图的纵轴必须从 0 开始)。

表 5.1 某地二年三种疾病的死亡率(1/10 万)

死因	1952 年	1972 年
肺结核	163.2	27.4
心脏病	72.5	83.6
恶性肿瘤	57.2	178.2

我们面临的首要问题是将表 5.1 的数据转化为 Stata 制作条形图所要求的数据格式。根据要求，我们需引入死因变量 D 和年度死亡率变量 P52 和 P72，并定义如下：

D= { 1 肺结核
     2 心脏病
     3 恶性肿瘤

P52=1952 年的死亡率

P72=1972 年的死亡率

数据输入过程：

```
. drop _all
. input d p52 p72
            d        p52         p72
  1. 1 163.2  27.4
  2. 2  72.5  83.6
  3. 3  57.2 178.2
```

```
  4. end
. save d:\mydata\ex5-1
file ex5-1.dta saved
```

数据已以文件名 ex5-1.dta 存入磁盘。现对数据库结构进行描述：

```
. des
Contains data from ex5-1.dta
  Obs:      3 (max=  4719)
 Vars:      3 (max=    99)
Width:     12 (max=   200)
  1. d            float  %9.0g
  2. p52          float  %9.0g
  3. p72          float  %9.0g
Sorted by:
```

为使图形更具有可读性，还可对变量及其取值给予必要的说明：

```
. lab var d "Reasons of die"
. lab var p52 "Rate of die in 1952"
. lab var p72 "Rate of die in 1972"
. lab define d 1 "tuberculosis" 2 "heart disease" 3 "tumour"
. des
Contains data from ex5-1.dta
   Obs:      3 (max=  4719)
  Vars:      3 (max=    99)
 Width:     12 (max=   200)
  1. d            float  %9.0g            Reasons of die
  2. p52          float  %9.0g            Rate of die in 1952
  3. p72          float  %9.0g            Rate of die in 1972
Sorted by:
. save d:\mydata\ex5-1, replace
file ex5-1.dta saved
```

有了以上的准备，Stata 就可以为我们作出精美的条形图了：

```
. gra p52 p72, bar by(d) sh(31) l1("Rate of die(1/100000)") b1(Comparison of rate of die)
```

指令 sh(31)是 shading(31)的缩写，3 和 1 分别指示 1952 年和 1972 年的条形图的明暗度,3 与 1 之间无空格及其它符号。“gra”后跟了几个变量，sh 的括号内就应有几个数字与之对应。[注:为便于排版，图 5.11 中的阴影在用 WORD 编辑时作了处理]。

图 5.11　　例 5.2 资料的直条图(直条并排)

增加不同的选择项，会出现不同的效果。

```
. gra p52 p72 , bar by(d) ylab  stack  total  alt  l1("Rate of Die (1/100000)")  gap(4)
```

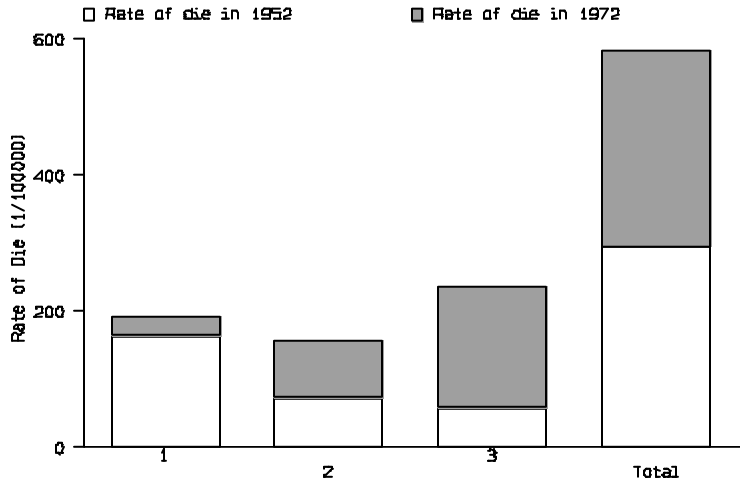


图 5.12　　例 5.2 资料的直条图(直条堆积)

这里，选择项 stack 要求 Stata 将各指标叠起排放；选择项 total 是要求 Stata 给出三者的总和；选择项 alt 要求横坐标的 4 个标目错开排列；选择项 gap(4) 要求缩小坐标的标题与图的间隔。

三、　圆饼图的制作

圆饼图(pie)主要用于表示全体中各部分的比重。绘制圆饼图命令同直条图，只是将 bar 换成 pie 就行了。

graph **变量** 1 **变量** 2 [...]　,pie [**选择项**]

选择项比直条图少，有：shading(#,...#)，by(变量)，accumulate，totle，以及标题等：title(["][字符串]["])，t|l|b|r1|2title(["][字符串]["])，saving(文件名[,replace])。其意义同上。

例 5.3 将例 5.2 资料作圆饼图：

```
. gra p52 p72, pie by(d) sh(31) total
```

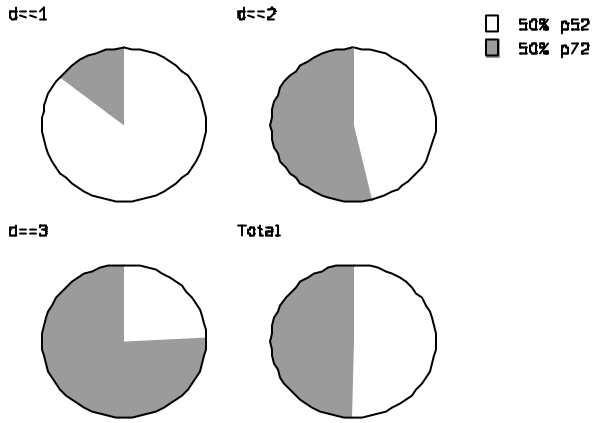


图 5.13 例 5.2 资料的圆饼图

选择项 total 要求 Stata 给出合计的圆饼图。

四、 二维散点图与线图的制作

散点图(scatter graph)用于反映两个或多个变量间的关系，将各散点用连线连接起来，就是线图(line graph)，主要用于反映事物的变化趋势等。用于散点图和线图绘制的命令为：

graph y变量 [y变量2[...]] x变量, [选择项]

绘制散点图和线条的两个主要的选择项为：

connect(c...c)
或简写为 c(c...c)

/* 连接各散点的方式，c 表示：

.	不连接 (缺省值)
l	用直线连接
L	沿 x 方向只向前不向后直线连接
m	计算中位数并用直线连接
s	用三次平滑曲线连接
J	以阶梯式直线条连接
‖	用直线连接在同一纵向上的两点
II	同 ‖，只是线的顶部和底部有一个短横

Symbol(s...s)
或简写为 s(s...s)

/* 表示各散点的图形，s 表示：

O	大圆圈（缺省值）
S	大方块
T	大三角形
o	小圆圈
d	小菱形

p	**小加号**
.	**小点**
i	**无符号**
[varname]	**用变量的取值代码表示**
[_n]	**用点的记录号表示**

其它选择项还有：

axis/ noaxis，border/noborder，twoway box，sort

xlable/ylable/tlable/rlable[(#,...,#)]，xtick/ytick/ttick/rtick[(#,...,#)]

xline/yline/tline/rline[(#,...,#)]，xscale/yscale[(#,,#)]

title(["][**字符串**]["])，b1/l1/t1/rl(["][**字符串**]["])，b2/l2/t2/r2(["][**字符串**]["])

gap(#)，saving(**文件名**[,replace])]

意义同前。

例 5.4　某地三岁儿童 10 人的体重与体表面积测量值如下，试作散点图，描述两者间的关系。

体　重 x,(kg)	**11**	**12.3**	**12**	**11.8**	**13.1**	**14.4**	**13.7**	**14.9**	**15.2**	**16**
体表面积 y,(10^3cm)	**5.283**	**5.292**	**5.358**	**5.299**	**5.602**	**5.83**	**6.014**	**6.102**	**6.075**	**6.411**

```
.gra y x , s(o) xlab ylab
```

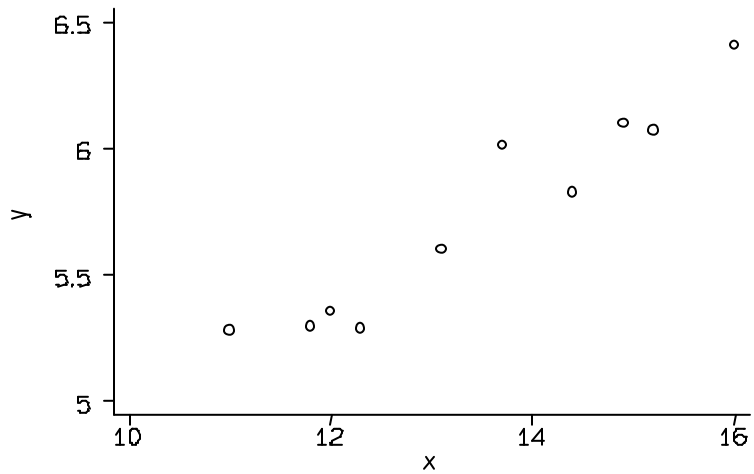


图 5.14　10 名三岁儿童的体重与体表面积散点图

这里有两个变量，前者为纵轴，后者为横轴。s(o)**表示散点用圆圈表示。**gra **是** graph **的缩写。**

下面 4 **个命令都是画线图的，但不同的选择，所作图形之效果就不同。**

```
. gra y x, xlab ylab c(l) s(d) b2("(a)")        gap(4)  saving(d:\mydata\ex54a)
. gra y x, xlab ylab c(l) s(p) b2("(b)")  sort  gap(4)  saving(d:\mydata\ex54b)
. gra y x, xlab ylab c(J) s(.) b2("(c)")  sort  gap(4)  saving(d:\mydata\ex54c)
. gra y x, xlab ylab c(L) s(T) b2("(d)")        gap(4)  saving(d:\mydata\ex54d)
```

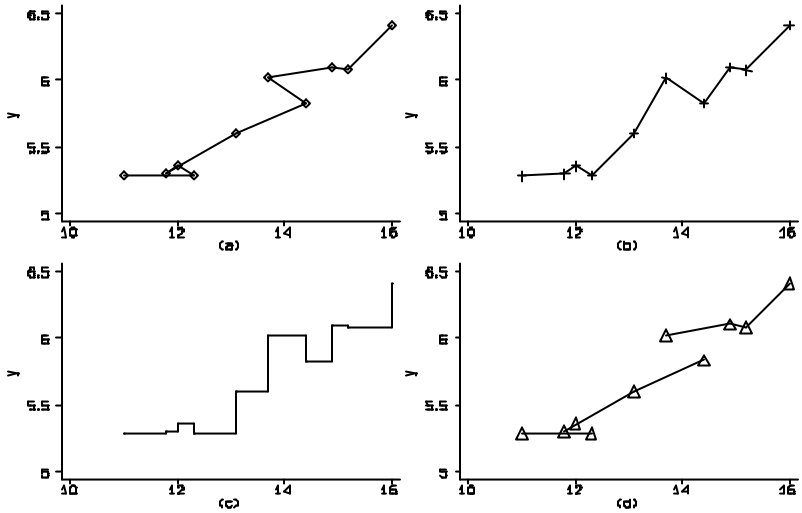


图 5.15 例 5.4 资料的四种线图

这里给出的 4 个图各有特点，分别说明了几个选项的作用：

symbol(s)：(a)、(b)、 (c)、 (d)4 **个图中，分别用了** s(d)、s(p)、s(.)、s(T)，**分别表示菱形、加号、小点号和三角形。**

connect(c)：(a)**和**(b)、 (c)、 (d)4 **个图中，分别用了** c(l)、c(J)、c(L)，**分别表示直线、阶梯式直线、不向后直线三种连接方式。**

sort：**图**(a)**和图**(b)**中，都是用直线相连的，但连线的顺序不同，前者无** sort **选择项，而后者选用了** sort**。故前者是按数据的顺序连接的，而后者是按横轴从小到大的顺序连接的。这就是** sort **的作用。**

用 s()，c()**及** sort **的不同组合，可以得到不同的效果，读者不妨一试。**

一个图中可画多条趋势线，作图命令不变，所有变量列在关键词 graph **后，最后一个变量是横轴变量。变量的顺序须与** c(c…c)**及** s(s…s)**中的连线及符号相对应。**

下面通过一个实例来看一看散点图和线图的巧妙应用。

例 5.5 **两组病人某指标的动态变化(均数±标准差)结果如下，要求在同一个图中画两条趋势曲线，并标上标准差。**

时间(h)	0	2	4	6	8	10	12
实验组	11.8±1.4	18.5±2.3	25±3.4	27.3±2.9	30.4±3.2	31.7±3.4	32.5±3.5
对照组	10.6±1.5	16.3±2	18.9±2.1	22.6±1.9	22.8±1.8	22.6±2.3	22.9±2.2

先输入资料：

```
. input time y1 sd1 y2 sd2
1.  0  11.8   1.4  10.6   1.5
2.  2  18.5   2.3  16.3   2
3.  4  25     3.4  18.9   2.1
4.  6  27.3   2.9  22.6   1.9
```

```
5.  8  30.4   3.2  22.8   1.8
6.  10 31.7   3.4  22.6   2.3
7.  12 32.5   3.5  22.9   2.2
8.  end
```

先画一组，以说明如何在趋势线图上加上标准差，以实验组为例。资料要作一些整理。

首先要产生两个新变量，以表示标准差的上下两点的位置。

```
. gen z1=y1+sd1
. gen z2=y1-sd1
```

则 z1,z2 分别表示标准差的上下两点的位置。用下列命令可得到所要作的图。

```
. gra y1 z1 z2 t , s(T..) c(lll) xlab(0,2,4,6,8,10,12,14) ylab
```

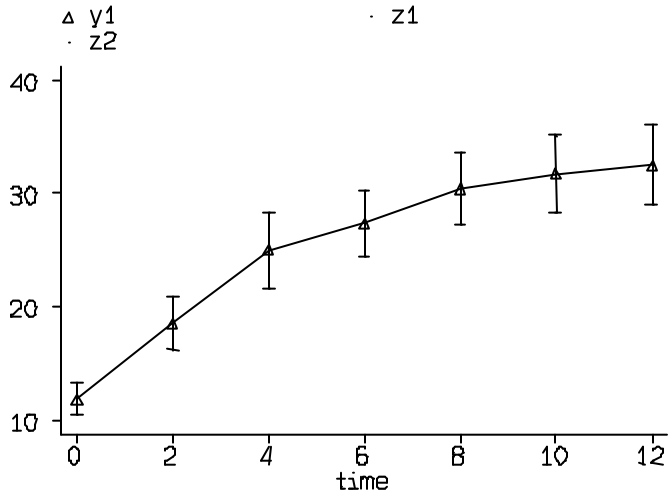


图 5.16　例 5.5 资料实验组的趋势线图

这里，有三组散点，即(y1,t),(z1,t),(z2,t)，各 7 各点。s(T..)表示第一组散点用三角表示，另两组用点表示。其中，第一组要用直线相连，以表示 y1 随时间的变化趋势，而另两组的点不要连接。但要将对应的 y1+sd1 和 y1-sd1 连起来，以反映变异的大小。故这里选用了 c(lll)。

再将两组资料合在一个图中。当然可按上述步骤，增加几个变量即可。但因两组资料的横轴刻度是一样的，当将两组资料合在一个图中时，就有很多重叠的部分，因而分不清两组资料。为了更清楚地表示，可将两组的横轴错开。所谓错开，即是两组的横坐标错开，如第 2 组的横坐标向右移 0.4。但这样两组的横坐标就不一样了，而 graph 命令中横坐标只能有一个，因而原来的横坐标数据个数就增加一倍，变成 14 个点。步骤如下：

```
. set obs 14                              /* 将观察点扩展到 14。
. replace time=time[_n-7]+0.4 in 8/l      /* 将第 2 组资料的时间加上 0.4，并放在 8～14 号。
. replace y2=y2[_n-7] in 8/l              /* 将第 2 组资料的观察值放在 8～14 号。(l 表示 last)
. replace  y2=. in 1/7                    /* 将第 2 组资料的 1～7 号观察值置为缺省值
. replace sd2=sd2[_n-7] in 8/l            /* 将第 2 组资料的标准差放在 8～14 号
. replace sd2=. in 1/7                    /* 将第 2 组资料 1～7 号的标准差置为缺省值
. replace z1=y2+sd2 in 8/l                /* 计算第 2 组资料的标准差上面点的值，并放在 8～14 号
```

```
. replace z2=y2-sd2 in 8/l                    /* 计算第 2 组资料的标准差下面点的值，并放在 8～14 号
```

看一下现在的数据库：

```
. list time y1 y2 z1 z2
```

	time	y1	y2	z1	z2
1.	0	11.8	.	13.2	10.4
2.	2	18.5	.	20.8	16.2
3.	4	25	.	28.4	21.6
4.	6	27.3	.	30.2	24.4
5.	8	30.4	.	33.6	27.2
6.	10	31.7	.	35.1	28.3
7.	12	32.5	.	36	29
8.	.4	.	10.6	12.1	9.1
9.	2.4	.	16.3	18.3	14.3
10.	4.4	.	18.9	21	16.8
11.	6.4	.	22.6	24.5	20.7
12.	8.4	.	22.8	24.6	21
13.	10.4	.	22.6	24.9	20.3
14.	12.4	.	22.9	25.1	20.7

共有 14 个点，其中 y2 和 sd2 被放在 8～14 号上，z1 和 z2 表示标准差的上下两点的位置，前面 7 个数据是实验组的，后面 7 个数据是对照组的。做好这样的准备后，即可画图。

```
. gra y1 y2 z1 z2 time, c(llll) s(T0..) xlab(0,2,4,6,8,10,12)  ylab
```

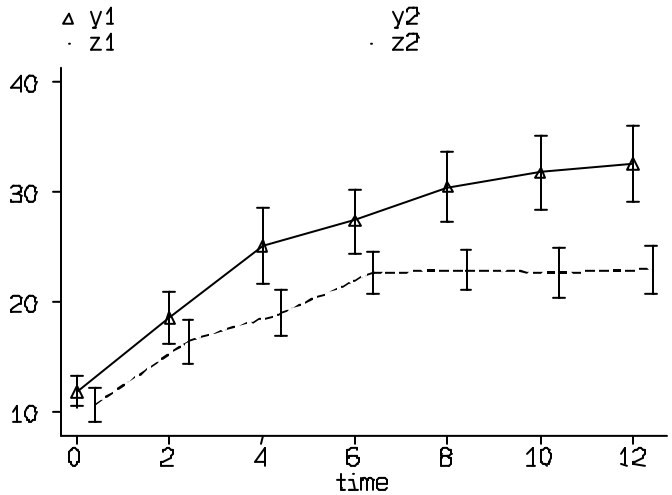


图 5.17　例 5.5 两组资料的趋势线图

两组资料一目了然。

注意，这里的两个 l l 必须写在一起，z1,z2 也必须排在一起。用||代替 l l 时，标准差的上下

两点无小横线段。

五、 一维散点图与箱式图的制作

一维散点图用于反映一个变量各观察点的分布位置。箱式图绘制 box-whisker 图，用于描述一组资料的中位数、四分位数及最大值、最小值的分布位置。

用于绘制一维散点图和箱式图的命令为：

gra 变量名 , oneway

gra 变量名 , oneway box

所作图形如图 5.4。注意，这里的 oneway 不是选择项，如无该项，则 Stata 绘制的是直方图。所指定的变量不超过 6 个。各变量的取值范围参差太大时，看不出效果。

六、 二维箱式图的制作

二维箱式图绘制两个变量(一个是纵轴变量，一个是横轴变量)的箱式图，分别平行于对应的轴。

用于绘制二维箱式图的命令是在为：

gra y变量 x变量, twoway box

注意，这里的 twoway 不是选择项，如无该项，则 Stata 给出的是两个单变量的箱式图。其余选择项与二维散点图的选择项相同。

例 5.6 将例 5.2 的散点图上加上箱式图。命令为：

```
. gra y x, two box s(T) ylab xlab(11,12,13,14,15,16) gap(4)
```

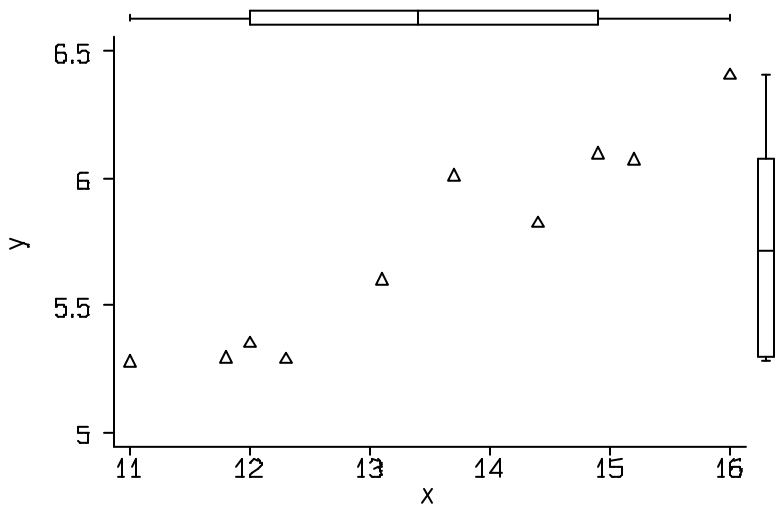
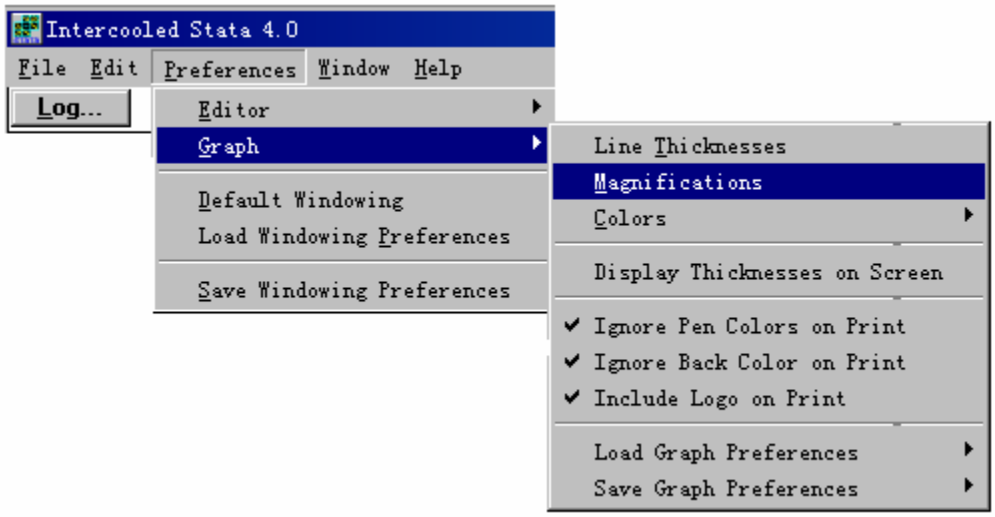


图 5.18 例 5.2 两组的二维散点图与箱式图

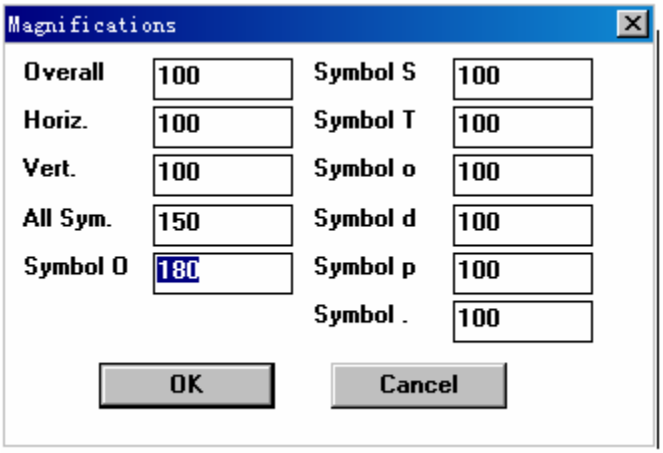
§5.3 怎样使图形更美观

Stata 画出的图比其它统计软件画出的图精美，且可直接被 WINDOWS 中的画笔及 WORD 等软件直接调用。但要精益求精，使画出的图形更美观，还有几个技巧。现分述如下：

⑴ 调整散点的大小，使画出的散点大小适中。这可在 Stata 的菜单中选 Preferences，再选 Graph 中的 Magnifications。见下图：



这时，Stata 弹出 Magnifications 菜单：



用户可在这个菜单中对各种散点的符号的大小进行调整。各散点的大小 Stata 默认值为 100，要使散点变大，只需增加各散点的取值，一般取 150～200。要使散点变小，只需减少各散点的取值。选定后，按 OK 即可。

⑵ 调整图形中文字的大小。4.0 以下版本中，Stata 图形中显示的数字和文字实际上是一种图形，且不能被文字或图形处理软件所识别；5.0 以上版本中，Stata 图形中显示的数字及文字与操作系统中的文字一致。要使图形中文字的大小适中，可通过命令 set text 来定义。

```
set text #
```

这里＃表示取值。Stata 默认值为 100，要使文字变大，则需加大其取值，一般取 150 左右。如：

set text 150。这一命令用在 graph 命令之前，输入该命令后的所有作图命令的文字都将变大。直到再用该命令对文字的大小进行修改。

(3) 调整打印出来的图形大小。在 DOS 版本的 Stata 中，用命令 gphdot 可将图形打印出来；在 WINDOWS 版本的 Stata 中可在菜单中选 Print Graph ,将图形印出来，也可在命令行输入 DOS 命令 gphdot。但直接打印的图形较大，要使打印出的图形大小适中，须增加一些选择项。命令如下：

gphdot 文件名 [/r#] [/rx#] [/ry#] [/n]

其中：

```
/r#     /* 定义打印图的大小，同时缩小或放大长度与宽度。默认为 100。# 的取值越大，打印出的图
           就越大，反之亦然。
/rx#    /* 定义打印图的宽度。默认为 100。
/ry#    /* 定义打印图的长度。默认为 100。
/n      /* 图的下方不打印 stata 标记。
```

(4) 用 WORD 等软件对图形进行编辑，如定义线条的粗细、线型，改变填充色或填充图案等，使图形更美观。

第六章　数值变量资料的统计分析

数值变量资料又称计量资料，通常是指每个观察单位某项指标量的大小，一般具有计量单位。这类资料按分析的内容一般可分为两种：一种是比较几种处理之间的效应，简单地讲就是比较各处理组观察值均数、方差的大小；另一种是寻找指标间的关系，即某个（或某些）指标的取值是否受其它指标的影响。本章主要介绍不同设计类型的数值变量资料的比较。

§6.1　样本均数与总体均数比较的 *t* 检验

t 检验亦称 student's *t* 检验，主要用于下列三种情况：(1)样本均数与总体均数比较；(2)配对数值变量资料的比较；(3)两样本均数的比较。

Stata 用于样本均数与总体均数比较的 *t* 检验的命令是 ttest：

```
ttest  变量名= #val
```

这里，#val 表示总体均数。命令中可以选用 if 语句和 in 语句对要分析的内容加一些条件限制。

对已知样本含量、均数和标准差的资料，欲将其与某总体均数进行比较，Stata 还提供了更为简洁的命令 ttesti：

```
ttesti  #obs  #mean  #sd  #val
```

这里，#obs 表示样本含量，#mean 表示样本均数，#sd 表示样本标准差，#val 表示总体均数。

先看一个实例。

例 6.1　10 例男性矽肺患者的血红蛋白(g/dl)如下：

病例号:	1	2	3	4	5	6	7	8	9	10
血红蛋白: (x,g/dl)	11.3	15.0	15.0	13.5	12.8	10.0	11.0	12.0	13.0	12.3

已知男性健康成人的血红蛋白正常值为 14.02(g/dl)，问矽肺患者的血红蛋白是否不同于一般?

算得该 10 例矽肺患者的血红蛋白均数为 12.59(g/dl)，显然不等于 4.02。造成这种差别的原因可能有两种，其一：矽肺患者的血红蛋白确实不同于健康人（本质上的差异）；其二：抽样误差。由于每个人的血红蛋白不尽相同，即使从正常人中抽检 10 个人，所得的样本均数亦不会恰好等于 14.02，这种差别称为抽样误差。只要个体之间存在差异，抽样误差就不可避免，但抽样误差是有规律的，这种规律是可以被认识和掌握的！欲想知道矽肺患者的血红蛋白均数与 14.02 的差别到底是本质上的差异还是纯粹的抽样误差，需进行假设检验。

假设检验就是首先根据设计和研究目的提出某种假设，然后根据现有资料提供的信息，推断此假设应当拒绝还是不拒绝。

结合本例，检验假设 H_0 是：(所有)矽肺患者的血红蛋白均数=14.02。 然后根据样本的含量

n=10、均数 $\bar{x}$ =12.59、标准差 s=1.63 构造一反映差别大小的检验统计量 t，如果 H_0 成立，即样本均数与总体均数的差别仅是抽样误差，则这种差别一般不会太大，即｜t｜值不会太大，如｜t｜值很大，超过了事先规定的界值，则就有理由怀疑 H_0 的成立。但｜t｜值与 n 有关，故将对｜t｜值的判断改为对概率 P 判断，P 是根据样本均数与总体均数的抽样误差规律(t分布)，由 t及 n-1(自由度)求得的，其涵义是：在 H_0 成立的条件下，纯粹由抽样得到现有 t这么大的误差或比 t更大的误差，有多大的可能性或称概率。显然 P 越小，越有理由怀疑 H_0 的成立， 因而拒绝 H_0；而 P 大，就没有理由拒绝 H_0。一般以α=0.05 作为拒绝与不拒绝的界限，α称为检验水准。P<α称差异有显著性，否则称差异无显著性。

这一过程可由 ttest 命令完成。

```
. input x
              x
     1.  11.3
     2.  15.0
         .....
    10.  12.3
    11.  end
```

将数据存入 d:\mydata\ex5-1.dta:

```
. save ex5-1
```

进行 t 检验:

```
. ttest x=14.02

Variable |   Obs        Mean    Std. Dev.       Min        Max
---------+--------------------------------------------------------
       x |    10       12.59    1.632619         10         15

         Test: mean of x = 14.02
               t-statistic = -2.77 with 9 d.f.
               Prob > |t|  =  0.0218
```

运算结果给出了几个基本统计量，有观察数(Obs)，均数(Mean)，标准差(Std.Dev)，最小值(Min)，最大值(Max)；检验的内容(Test)，即假设检验 H_0；t 值(t-statistics)；自由度(d.f.)及检验概率(Prob > |t|)。本例 t=-2.77，P=0.0218<0.05。故按α=0.05 水准，拒绝 H_0，可认为矽肺患者的血红蛋白含量低于正常值 14.02(g/dl)。

如已知该资料样本含量#obs=10，均数#mean=12.59,标准差#sd=1.632629，欲将其与总体均数#val=14.02 比较，可用下列命令 ttesti：

```
. ttesti 10 12.59 1.632619 14.02

      Variable |      Obs        Mean    Std. Dev.
      ---------+--------------------------------------
             x |       10       12.59    1.632619
```

```
Ho:  mean = 14.02
        t = -2.77 with 9 d.f.
 Pr > |t| = 0.0218
```

结果与从原始资料计算是等价的。

注意，ttesti 命令中必需要 4 个数据，且 4 个数据在命令中的顺序不能变，各数据间用空格分开。

§6.2　两样本均数比较的 t 检验

一、　配对设计 t 检验

医学研究中常将受试对象配成对子，对每对中的两个受试对象分别给予两种不同的处理，观察两种处理的结果是否一致，称为配对(设计)研究。有时以同一个受试对象先后给予两种不同的处理，观察两种处理的结果是否相同，这种配对称为自身配对。配对设计的优点是能消除或部分消除个体间的差异，使比较的结果更能真实地反映处理的效应。

配对 t 检验首先计算每对结果之差值，再将差值均数与 0 作比较。如两种处理的效应相同，则差值与 0 没有显著性差异。检验假设 H_0 为：两种处理的效应是相同，或总体差值均数为 0。

Stata 用于配对样本 t 检验的命令是 ttest：

ttest　变量 1=变量 2

这里，这里“变量 1”和“变量 2”是成对输入的配对样本。ttest 命令容许使用[if 表达式]和[in 范围]条件限制。

例 6.2　续例 6.1，10 例矽肺患者经克矽平治疗，其血红蛋白(g/dl)如下：

病例号	1	2	3	4	5	6	7	8	9	10
治疗前(x0)	11.3	15.0	15.0	13.5	12.8	10.0	11.0	12.0	13.0	12.3
治疗后(x1)	14.0	13.8	14.0	13.5	13.5	12.0	14.7	11.4	13.8	12.0

欲了解治疗对血红蛋白含量有无作用，需作配对 t 检验。先清除内存，然后输入数据：

```
. input x
              x0       x1
     1.  11.3  14.0
     2.  15.0  13.8
          ......
    10.  12.3  12.0
    11.  end
```

将数据存入 d:\mydata\ex6-2.dta：

```
. save ex6-2
. ttest x0=x1

  Variable |      Obs        Mean   Std. Dev.
  ---------+--------------------------------
        x0 |       10       12.59    1.632619
        x1 |       10       13.27    1.080175
  ---------+--------------------------------
```

```
diff. |       10    -.6799999     1.645735

      Ho:  diff = 0  (paired data)
             t = -1.31 with 9 d.f.
       Pr > |t| = 0.2237
```

本例差值均数为-.68，t=-1.31，自由度为 9，P=0.2237，按α=0.05 水准，尚不能认为治疗对血红蛋白含量的增加有作用。

按配对设计的思路，上述问题亦可作如下处理：即先求出差值 d，然后对差值 d 进行 *t* 检验。具体步骤如下：

```
. gen d=x1-x0
. ttest d=0
```

这样所得结果相同，请读者自己完成。

注意，这里的两个变量 x1 和 x0 必须成对输入。样本含量必须相等，如有缺省值，则用小数点表示，但与之对应的记录在计算时被省略。

二、 成组设计 t 检验

有时无法将受试对象逐个配成对，可将受试对象随机分为两组，每组接受不同的处理，检验两组的均数，以达到比较的目的。

t 检验要求两样本来自方差相同的正态总体，即各组资料达到或接近正态，两组的方差达到齐性。如两组资料偏态或方差不齐，则需要对原始数据作变量变换，如变换后仍未达到正态，可用秩和检验；如未达到方差齐性，则需用 *t* ' 检验，或用秩和检验。

Stata 提供了三种资料形式的两样本均数比较的 *t* 检验的命令，即：

```
ttest   变量 1=变量 2, unpaired [ unequal  welch ]
ttest   变量,  by(分组变量)  [unequal   welch]
ttesti  #obs1 #mean1 #sd1 #obs2 #mean2 #sd2  [, unequal  welch ]
```

这里：

第一个命令的数据格式是将两组数据用两个变量“变量 1”和“变量 2”分别输入，如两组的样本含量不等，则先输入样本含量大的变量，再输入样本含量少的变量，不足部分，Stata 将自动生成缺省值(用小数点表示)。也可同时输入，缺失部分用小数点表示。unpaired 是必选项，如不选，则 Stata 将作配对 *t* 检验。

第二个命令的数据格式是将两组数据用一个“变量”输入，再用另一个分组变量，以区分两组资料，如用 1 表示第 1 组资料，用 2 表示第 2 组资料。by(分组变量)是必选项。

第三个命令是针对已知两组资料的样本含量、均数和标准差的资料进行比较的简洁命令。这里有 6 个数据，#obs 表示样本含量，#mean 表示样本均数，#sd 表示样本标准差，1 表示第 1 组，2 表示第 2 组。

第一个命令和第二个命令允许加[权数]及[in 范围] 和[if 表达式]条件。

选择项 unequal 表示假设两组方差不齐，如不选表示假设两组方差达到齐性。

选择项 welch 表示用 Welch 方法对自由度进行校正，如不选此项，则用 Satterthwaite 方法对自由度进行校正。welch 选择项只有在选择了 unequal 才有效。

例 6.3 分别测得 14 例老年人慢性支气管炎病人及 11 例正常人的尿中 17 酮类固醇排出量

(mg/dl)**如下，试比较两组的均数有无差别。**

病人：	2.90	5.41	5.48	4.60	4.03	5.10	4.97	4.24	4.36	2.72	2.37	2.09	7.10	5.92
健康人：	5.18	8.79	3.14	6.46	3.72	6.64	5.60	4.57	7.71	4.99	4.01			

这里用三种不同的数据格式对资料进行分析，所得结果等价。

(1) **用两个变量表示两组资料：**

```
.input x1 x2
1.  2.90   5.18
2.  5.41   8.79
    ……
11. 2.37   4.01
12. 2.09   .
13. 7.10   .
14. 5.92   .
15  end
```

将数据存入文件 d:\mydata\ex6-3.dta。*t* **检验的命令为：**

```
. ttest x1=x2 , unpaired

Variable |      Obs        Mean    Std. Dev.
---------+--------------------------------
      x1 |       14    4.377857    1.449892
      x2 |       11    5.528182    1.735401
---------+--------------------------------
combined |       25       4.884    1.653227
         Ho:  mean(x) = mean(y)  (assuming equal variances)
                    t = -1.81 with 23 d.f.
                Pr > |t| = 0.0839
```

t＝-1.81，**自由度为** 23，P=0.0839>0.05，**故按** α=0.05 **水准，尚不能认为老年慢性支气管炎病人与正常人的尿中** 17 **酮类固醇排出量**(mg/dl) **有何不同。**

命令中无 unequal **选择项，故** Stata **自动假设两组方差齐**(assuming equal variances)。**如果两组方差不齐，则可用下列命令进行** *t′* **检验。**

```
. ttest x1=x2,unp une w

Variable |      Obs        Mean    Std. Dev.
---------+--------------------------------
      x1 |       14    4.377857    1.449892
      x2 |       11    5.528182    1.735401
---------+--------------------------------
combined |       25       4.884
         Ho:  mean(x) = mean(y)  (assuming unequal variances)
                    t = -1.77 with 21.19 d.f.
                Pr > |t| = 0.0917
```

这时，Stata **告知用户，是按方差不齐计算的**(assuming unequal variances)。**因选择了** welch **选择项**(w **是缩写**)，Stata **按** Welch **方法对自由度进行校正。所得自由度是非整数。本例自由度为** 21.19。**不选** w，**则** Stata **按** Satterthwaite **方法对自由度进行校正。**

```
. ttest x1=x2,unp une
```

```
Variable |      Obs        Mean    Std. Dev.
---------+--------------------------------------
      x1 |       14    4.377857    1.449892
      x2 |       11    5.528182    1.735401
---------+--------------------------------------
combined |       25       4.884

         Ho:  mean(x) = mean(y)  (assuming unequal variances)
                   t = -1.77 with 19.47 d.f.
              Pr > |t| = 0.0929
```

按 Satterthwaite **方法所得自由度为** 19.47**，而按** Welch **方法所得自由度为** 21.19**。本例结论相同。**

(2) **用一组变量表示观察值，用另一个分组变量表示两个不同的组。数据输入格式如下：**

```
. input y group
           y      group
  1. 2.90 1
  2. 5.41 1
     ……
 13. 7.10 1
 14. 5.95 2
     ……
 25. 4.01 2
 26. end
```

相应的命令为：

```
. ttest y , by(g)

Variable |      Obs        Mean    Std. Dev.
---------+--------------------------------------
       1 |       14    4.377857    1.449892
       2 |       11    5.528182    1.735401
---------+--------------------------------------
combined |       25       4.884    1.653227

         Ho:  mean(x) = mean(y)  (assuming equal variances)
                   t = -1.81 with 23 d.f.
              Pr > |t| = 0.0839
```

结果同上。这种格式的命令也可用 unequal **和** welch **选择项。**

(3) **如果已知两组资料的样本含量、均数和标准差，则可方便地用简洁命令** ttesti**，结果同上。**

```
. ttesti 14 4.377857 1.449892 11 5.528182 1.735401

Variable |      Obs        Mean    Std. Dev.
---------+--------------------------------------
       x |       14    4.377857    1.449892
       y |       11    5.528182    1.735401
---------+--------------------------------------
combined |       25       4.884    1.653227
```

```
Ho:  mean(x) = mean(y)  (assuming equal variances)
           t = -1.81 with 23 d.f.
    Pr > |t| = 0.0839
```

该命令要求各组资料输入的顺序不能变，即：样本含量、均数、标准差，但两组资料的顺序可任意。

该命令也可加选择项 unequal 和 welch。留作练习。

两组均数的比较还可通过方差分析来实现(见下节)。

§6.3　单因素方差分析及方差齐性检验

一、　单因素方差分析

根据某一试验因素，将受试对象随机分为若干处理组(各组样本含量可以相等亦可不等)，即为单因素试验设计。比较此多个样本均数的目的是推断各处理的效应有无差异。常用单因素方差分析。

单因素方差分析的假设检验 H_0：各处理效应相同(或各组总体均数相等)。并根据各组样本含量、均数、组内离均差平方和、组间离均差平方和等构造检验统计量 F，F 是反映各组差别大小的统计量，F 越大说明各组均数差别就越大。同样 F 与处理组数、样本含量的大小有关。

如单因素方差分析拒绝检验假设 H_0，只说明各组总体均数不等或不全相等，到底是哪些组间有差别，需进一步作均数间的两两比较。两两比较的方法很多，Stata 提供的两两比较方法有 Bonferroni 法、Scheffe 法、Sidak 法。

Stata 用于单因素方差分析及两两比较的命令为：

oneway 响应变量 分组变量，[选择项]

这里选择项有：

noanova	/*不打印方差分析表
nolabel	/*不打印分组变量的取值标签
missing	/*将缺省值作为单独的一组
wrap	/*两两比较的表格不分段
tabulate	/*打印各组的基本统计量表
[no]means	/*[不]打印均数
[no]standard	/*[不]打印标准差
[no]freq	/*[不]打印各组观察例数。这三项只有在选择了 tabulate 才有效。

Stata 还提供了三种两两比较方法。

scheffe	/*Scheffe 法
bonferroni	/*Bonferroni 法
sidak	/*Sidak 法

[注]　Stata 提供的三种两两比较方法均偏于保守(见薛禾生，陆守曾：计量资料多个样本均数间两两比较方法的评价，中国卫生统计，4(2):24～27，1987)

例 6.4　以小鼠研究正常肝核糖核酸(RNA)对癌细胞的生物学作用，试验分为对照组(生理盐水)，水层 RNA 组和酚层 RNA 组，分别用此三种不同处理诱导肝癌细胞的 FDP 酶活力，数据列于表 6.1，试比较三组均数有无差别。

表 6.1　三组小鼠的 FDP 酶活力

对照组	水层 RNA 组	酚层 RNA 组
2.79	3.83	5.41
2.69	3.15	3.47
3.11	4.70	4.92
3.47	3.97	4.07
1.77	2.03	2.18
2.44	2.87	3.13
2.83	3.65	3.77
2.52	5.09	4.26

要分析的数据需按特定的形式输入计算机，即新定义一分组变量 g，而将三个处理组的观察值均以某变量 X 表示，将表 6.1 整理成下列形式：

表 6.2　表 6.1 资料数据格式

X	g	X	g	X	g
2.79	1	3.83	2	5.41	3
2.69	1	3.15	2	3.47	3
3.11	1	4.70	2	4.92	3
3.47	1	3.97	2	4.07	3
1.77	1	2.03	2	2.18	3
2.44	1	2.87	2	3.13	3
2.83	1	3.65	2	3.77	3
2.52	1	5.09	2	4.26	3

其中 g 为分组变量，g=1 表示对照组；g=2 表示水层 RNA 组；g=3 表示酚层 RNA 组。然后进行方差分析。

```
. input x g
            x         g
  1. 2.79 1
  2. 2.69 1
  3. 3.11 1
  4. 3.47 1
  5. 1.77 1
  6. 2.44 1
  7. 2.83 1
  8. 2.52 1
  9. 3.83 2
 10. 3.15 2
 11. 4.70 2
 12. 3.97 2
 13. 2.03 2
 14. 2.87 2
 15. 3.65 2
 16. 5.09 2
```

```
 17. 5.41 3
 18. 3.47 3
 19. 4.92 3
 20. 4.07 3
 21. 2.18 3
 22. 3.13 3
 23. 3.77 3
 24. 4.26 3
 25. end

. oneway x g, t sch

           |            Summary of x
         g |        Mean   Std. Dev.       Freq.
-----------+------------------------------------
         1 |      2.7025   .50013569           8
         2 |     3.66125   .98508069           8
         3 |   3.9012501   1.0164425           8
-----------+------------------------------------
     Total |   3.4216667   .98273207          24

                      Analysis of Variance
    Source              SS         df      MS            F     Prob > F
------------------------------------------------------------------------
Between groups      6.43680897      2   3.21840449      4.28     0.0275
 Within groups      15.7757246     21   .751224983
------------------------------------------------------------------------
    Total           22.2125336     23   .965762331

Bartlett's test for equal variances:  chi2(2) =   3.4559  Prob>chi2 =   0.178

                           Comparison of x by g
                                (Scheffe)
Row Mean-|
Col Mean |          1          2
---------|----------------------
       2 |     .95875
         |      0.111
         |
       3 |    1.19875        .24
         |      0.038      0.859
```

由于选用了选择项 t(tabulate)，Stata **首先给出了三个组的及总的均数、标准差和观察频数**(Freq.)。

结果中给出了完整的方差分析表及 Bartlett **的方差齐性检验，**Chi2 **表示**χ^2**统计量，括号中数表自由度，**Prob>Chi2=0.178 **表示大于现有**χ^2**值的概率** P=0.178。

选择项 sch(scheffe)**提供了三组之间的两两比较**(Comparison of x by g)。**其中上行表示相比较的两组均数之差，下行表示检验的概率。如第一组**(g=1)**与第二组**(g=2)**相比较**,x1-x2=0.95875,

检验概率 P=0.111。

如不用选择项 tabulate 及 scheffe，则结果中只有方差分析表及 Bartlett 的方差齐性检验。

本例经方差分析，得 F＝4.28，P=0.0275，按 α=0.05 水准，可认为三组的 FDP 酶活力不同。从两两比较的结果来看，这种差别主要来自对照组与酚层 RNA 组，而水层 RNA 组与酚层 RNA 组似无差别。对照组与水层 RNA 组是否有差异暂不能下结论，需进一步研究。

二、 方差齐性检验

无论是进行 *t* 检验还是方差分析，资料都必需满足一定的条件，即①正态性，②方差齐性，③独立性。而以方差齐性条件最为重要。因此，在进行 *t* 检验和方差分析之前，必须进行方差齐性检验。即检验各处理组数据的变异(方差)是否相同。一般情况下进行方差齐性检验都不希望拒绝 H_0，此时，为提高检验把握度，检验水准应定得高一些，比如：α＝0.10，0.20 等。

Stata 用于样本方差与总体方差的比较，以及两样本方差齐性检验的命令为：

```
sdtest  变量名 = #val
sdtest  变量名 1 = 变量名 2
sdtest  变量名，by(分组变量)
sdtesti #obs {#mean | . } #sd #val
sdtesti #obs1 {#mean1 | . } #sd1 #obs2 {#mean2 | . } #sd2
```

这里，第一个命令用于检验某变量的方差是否来自方差为#val 的总体；第二、三个命令是用于检验两样本对应的总体方差是否相同，但两个命令要求的数据输入形式不同，第二个命令用于每组各一个变量，第三个命令用于有分组变量的情形。第四、五个命令用于已知样本含量和样本标准差的情形，其中，第四个命令用于样本方差与总体方差的比较，第五个命令用于两样本方差齐性检验。样本均数可以输入，亦可缺失并用小数点表示。

(1) 两个方差的比较

两样本方差的齐性检验一般用 *F* 检验，*F* 值反映的是两样本方差之比，如相应的总体方差相等，则 *F* 应接近 1。

例 6.5 检验例 6.3 资料中，病人与正常人尿中 17 酮类固醇排出量(mg/dl)之方差是否达到齐性，命令及输出结果如下：

```
. use d:\mydata\ex6-3
. sdtest y ,by(group)

  Variable |      Obs        Mean    Std. Dev.
  ---------+--------------------------------
         1 |       14    4.377857    1.449892
         2 |       11    5.528182    1.735401
  ---------+--------------------------------
  combined |       25       4.884    1.580377

           Ho:  sd(1) = sd(2)  (two-sided test)
   Lower tail:  F1(10,13) =  0.70
   Upper tail:  F2(10,13) =  1.43
    (Pr < F1) + (Pr > F2) =  0.5555
```

Stata 给出了左尾的 F 比值，即小方差/大方差:$1.449892^2/1.735401^2=0.69802603\approx0.7$，右尾的

F 比值，即大方差/小方差:$1.735401^2/1.449892^2=1.4326113\approx1.43$。得相应的双尾概率之和为P=0.5555，故按α=0.20 水准，可认为两组方差达到齐性。

也可用下列命令，得到同样结果：

```
. sdtesti 14 4.377857 1.229892 11 5.528182 1.735401
```

因方差齐性检验中，未涉及样本均数，故命令中的均数可以不输入，但它(们)的位置必须留着，且用小数点表示。即

```
. sdtesti 14  . 1.449892  11  .  1.735401
```

(2) 样本方差与总体方差的比较

样本方差与总体方差的比较一般用χ^2 检验。

例 6.6　检验例 6.3 资料中，健康人的尿中 17 酮类固醇排出量是否来自方差为 4^2 的总体。这是样本方差与总体方差的比较。命令及结果如下：

```
. sdtesti 11 . 1.735401  4
  Variable |      Obs        Mean    Std. Dev.
  ---------+--------------------------------
         x |       11           .     1.735401

           Ho:  std. dev. = 4  (two-sided test)
                chi2(10) =  1.88
           2*(Pr <= chi2) =  0.0057
```

命令中前 3 个数字分别为样本含量、均数(用小数点表示)和标准差，最后一个数字为总体标准差。结果中给出了χ^2 分布的双侧面积为：2*(Pr<=chi2)=0.0057。按α=0.20 水准，不能认为健康人的尿中 17 酮类固醇排出量来自方差为 16 的总体。

[注]　$c^2=(n-1)\times s^2/s^2$,　n = n -1。在 Stata3.1 以下版本中，给出的是单侧概率，如本例给出的概率 probability=0.0028 是χ^2 分布的单侧面积，意指样本标准差(Std. Dev.)之平方大于已知标准差σ之平方的概率，其双侧检验概率为 P=0.0028×2=0.0057。下结论时相应的概率须乘以 2。

(3) 多个方差的齐性检验

多个方差的齐性检验是检验每个处理组相应的总体方差是否全部相等，检验假设为 H_0：$\sigma_1^2=\sigma_2^2=\cdots\cdots=\sigma_K^2$。当拒绝检验假设时，则可认为至少有两个方差不等。用χ^2 检验。该检验由 oneway 命令给出，见例 6.4。

§6.4　两因素的方差分析

两因素的方差分析一般是指配伍组方差分析，和不考虑交互作用与考虑交互作用的 a×b 析因分析。Stata 的命令为：

anova 因变量　分组变量 1　分组变量 2

anova 因变量　分组变量 1　分组变量 2　交互作用项

oneway 命令只适用于单因素方差分析，而要进行两因素、多因素的方差分析需用 anova 命令。anova 命令亦能用于单因素情形，但却不如 oneway 命令方便，因为 anova 不能提供方差齐

性检验和多重比较。因此在进行单因素方差分析时，建议用 oneway 命令。

anova 命令只适合于平衡资料，对非平衡资料需要用 glm(广义线性模型)命令。在 Stata 中用 help glm 可获得帮助。

例 6.7 (配伍组设计) 四种抗癌药物抑瘤效果的配伍组方差分析。

表 6.3 四种抗癌药物抑瘤效果

配伍组	a1	a2	a3	a4
b1	0.08	0.36	0.17	0.28
b2	0.74	0.50	0.42	0.36
b3	0.31	0.20	0.38	0.25
b4	0.48	0.18	0.44	0.22
b5	0.76	0.26	0.28	0.13

按配伍组设计的资料，必需定义两个分组变量，用以描述观察值 X 所在的处理组和配伍组，这里以 a=1，…，4 表示四种药物，b=1，…，5 表五个配伍组。数据按如下形式输入：

```
. input x a b
            x         a         b
  1.       .80  1  1
  2.       .74  1  2
  3.       .31  1  3
  4.       .48  1  4
  5.       .76  1  5
  6.       .36  2  1
  7.       .50  2  2
  8.       .20  2  3
  9.       .18  2  4
 10.       .26  2  5
 11.       .17  3  1
 12.       .42  3  2
 13.       .38  3  3
 14.       .44  3  4
 15.       .28  3  5
 16.       .28  4  1
 17.       .36  4  2
 18.       .25  4  3
 19.       .22  4  4
 20.       .13  4  5
 21.       end

. format x %9.4f      /* 定义 x 的数据格式保留 4 位小数
. tab a,summ(x)

           |            Summary of x
         a |        Mean   Std. Dev.       Freq.
-----------+------------------------------------
         1 |      0.6180      0.2134           5
         2 |      0.3000      0.1319           5
         3 |      0.3380      0.1123           5
         4 |      0.2480      0.0841           5
```

```
------------+------------------------------------
      Total |      0.3760      0.1975          20

. tab b,summ(x)

           |            Summary of x
         b |       Mean   Std. Dev.       Freq.
-----------+------------------------------------
         1 |      0.4025      0.2762           4
         2 |      0.5050      0.1668           4
         3 |      0.2850      0.0777           4
         4 |      0.3300      0.1519           4
         5 |      0.3575      0.2765           4
-----------+------------------------------------
      Total |      0.3760      0.1975          20

. anova x a b
                   Number of obs =      20      R-square     =  0.7058
                   Root MSE      = .134818      Adj R-square =  0.5341

          Source |  Partial SS    df       MS           F     Prob > F
-----------------+----------------------------------------------------
           Model |  .523170004     7  .074738572       4.11     0.0157
                 |
               a |  .410839998     3  .136946666       7.53     0.0043
               b |  .112330006     4  .028082501       1.55     0.2514
                 |
        Residual |  .218110001    12  .018175833
-----------------+----------------------------------------------------
           Total |  .741280005    19  .039014737
```

Stata 的方差分析是用线性模型的思维求解的，故结果中除给出了完整的方差分析表外，还给出了方差分析模型的 R^2(R-square)，调整 R^2(Adj R-square)，均方根(Root MSE)，及模型(Model)的方差分析。

本例，检验四种药物(a 因素)的 $F=7.53$，$P=0.0043$，按 $\alpha=0.05$ 水准，可认为四种药物的抑瘤效果不同。而不同配伍组(b 因素)间的差异无显著性，$F=1.55$，$P=0.2514$。模型检验的 Partials *SS*，*df* 是 a 因素和 b 因素之和，*MS* 为 *SS* 与 *df* 之商，*F* 是相应的 *MS* 与误差的 *MS* 之比。

例 6.8(2×2 析因设计)　分别用新旧两法提取某食物中甲乙两化合物，观察其回收率，结果如下，用 2×2 析因分析。

表 6.4　2×2 析因试验的回收率(%)

测定次数	a1　新法		a2　旧法	
	b1,甲物	b2,乙物	b1,甲物	b2,乙物
1	52	84	52	47
2	48	88	44	64
3	44	90	40	52
4	44	88	26	45

首先定义分组变量 a=1，2 表新旧两法；b=1，2 表甲乙两物。并计算各组的均数，标准差，最后进行方差分析。

```
. format x %9.2f
. tabu  b, summ(x) nofreq
```

```
                Means and Standard Deviations of x
           | b
         a |         1          2      Total
-----------+----------------------+----------
         1 |     47.00      87.50 |     67.25
           |      3.83       2.52 |     21.86
-----------+----------------------+----------
         2 |     40.50      52.00 |     46.25
           |     10.88       8.52 |     10.94
-----------+----------------------+----------
     Total |     43.75      69.75 |     56.75
           |      8.31      19.85 |     19.91
```

其中，上行为均数，下行为标准差。

```
. anova x a b a*b
```

```
                  Number of obs =      16      R-square     =  0.8930
                  Root MSE      = 7.28011      Adj R-square =  0.8663

      Source |  Partial SS    df       MS          F     Prob > F
-------------+-----------------------------------------------------
       Model |     5309.00     3  1769.66667      33.39     0.0000
             |
           a |     1764.00     1     1764.00      33.28     0.0001
           b |     2704.00     1     2704.00      51.02     0.0000
         a*b |      841.00     1      841.00      15.87     0.0018
             |
    Residual |      636.00    12       53.00
-------------+-----------------------------------------------------
       Total |     5945.00    15  396.333333
```

无论新、旧法间，甲、乙两化合物间，以及方法与化合物间的交互作用，经方差分析 P 值均较小，按 $\alpha=0.05$ 水准，可认为新法回收率高于旧法；乙化合物高于甲化合物；而乙化合物用新法提取效果更佳。

§6.5 多因素的方差分析

多因素的试验即是同时考虑多种因素影响的试验。相应的方差分析称为多因素的方差分析，仍用 anova 命令，只是分组变量多于两个。例如：

anova x a b c …… a*b b*c a*b*c ……

这里，a,b,c…表示分组变量，a*b,b*c,a*b*c…表示交互作用项。

例 6.9(拉丁方设计)　以下是五名受试者(a)在五个不同日期(b) 穿五种不同防护服(c)时测得的脉博数。试作分析。

表 6.5　　穿五种不同防护服时测得的脉博数(次/分)

试验日期 (b)	受试者 (a)				
	甲 (a=1)	乙 (a=2)	丙 (a=3)	丁 (a=4)	戊 (a=5)
1	D 133.4	B 98.0	A 114.0	E 110.8	C 110.6
2	B 144.4	E 132.8	D 113.2	C 119.2	A 115.2
3	C 143.0	A 123.0	E 115.8	D 118.0	B 103.8
4	A 129.8	D 104.0	C 114.8	B 116.2	E 100.6
5	E 142.8	C 120.0	B 105.8	A 110.6	D 109.8

假定该数据已存入文件 ex6-9.dta**，可直接调用，其中，**a=1,2,…,5 **表示** 5 **个不同的受试者，**b **表示** 5 **个不同日期，**c=1 **表** A **防护服，** c=2 **表** B **防护服，余类推。**

```
. drop _all
. use d:\mydata\ex6-10
. list
          x        a        b        c
  1.   133.4        1        1        4
  2.   144.4        1        2        2
  3.      143        1        3        3
  4.   129.8        1        4        1
  5.   142.8        1        5        5
  6.       98        2        1        2
  7.   132.8        2        2        5
  8.      123        2        3        1
  9.      104        2        4        4
 10.      120        2        5        3
 11.      114        3        1        1
 12.   113.2        3        2        4
 13.   115.8        3        3        5
 14.   114.8        3        4        3
 15.   105.8        3        5        2
 16.   110.8        4        1        5
 17.   119.2        4        2        3
 18.      118        4        3        4
 19.   116.2        4        4        2
 20.   110.6        4        5        1
 21.   110.6        5        1        3
 22.   115.2        5        2        1
 23.   103.8        5        3        2
 24.   100.6        5        4        5
 25.   109.8        5        5        4

. sort c
. tab c,summ(x)

           |                Summary of x
```

```
          c|       Mean   Std. Dev.       Freq.
-----------+------------------------------------
          1 |     118.52   7.7699439           5
          2 |     113.64   18.409451           5
          3 |     121.52   12.584594           5
          4 |     115.68   11.141451           5
          5 |     120.56   17.034907           5
-----------+------------------------------------
      Total |    117.984   13.079745          25
```

```
. anova x a b c
```

```
                   Number of obs =      25      R-square     =  0.8719
                   Root MSE      = 6.62156      Adj R-square =  0.7437

          Source |  Partial SS    df       MS           F     Prob > F
-----------------+----------------------------------------------------
           Model |  3579.77248    12  298.314373       6.80     0.0011
                 |
               a |   2853.6733     4  713.418324      16.27     0.0001
               b |  508.073488     4  127.018372       2.90     0.0684
               c |  218.025697     4  54.5064243       1.24     0.3445
                 |
        Residual |  526.140783    12  43.8450652
-----------------+----------------------------------------------------
           Total |  4105.91326    24  171.079719
```

按α＝0.05 水准，可认为各受试对象(a 因素)间的脉搏数有差别，但受试日期(b 因素)及穿不同防护服(c 因素)对脉搏数的影响尚看不出差别。

按拉丁方设计的资料如每个因素的不同组合没有重复试验，则不能进行交互作用的分析，下列命令是错误的：

```
. anova x a b c a*b
too many variables or values
r(146);
```

例 6.10(3×2×2 析因设计)　就表 6.6 资料分析三种基础液(a)中的钩端螺旋体计数(conut)有无差别，兔血清与胎盘血清(b)的计数有无差别，两种浓度(c)　间的计数有无差别，各因素间有无交互作。

```
. anova count a b c a*b a*c c*b a*b*c
```

```
                   Number of obs =      48      R-square     =  0.5656
                   Root MSE      =  400.421      Adj R-square =  0.4328

          Source |  Partial SS    df       MS           F     Prob > F
-----------------+----------------------------------------------------
           Model |  7514726.92    11  683156.992       4.26     0.0005
```

```
           |
         a |  107712.792      2   53856.3958       0.34     0.7169
         b |  6588972.00      1   6588972.00      41.09     0.0000
         c |  573781.333      1   573781.333       3.58     0.0666
       a*b |   95267.375      2   47633.6875       0.30     0.7448
       a*c |  47553.2917      2   23776.6458       0.15     0.8627
       c*b |  10502.0833      1   10502.0833       0.07     0.7995
     a*b*c |  90938.0417      2   45469.0208       0.28     0.7547
           |
  Residual |  5772117.00     36   160336.583
-----------+---------------------------------------------------
     Total |  13286843.9     47   282698.807
```

表 6.6　　钩端螺旋体计数

加入维生素的基础液(a)	血清种类(b)			
	兔血清(b=1)		胎盘血清(b=2)	
	血清浓度（c）		血清浓度(c)	
	5%,c=1	8%,c=2	5%,c=1	8%,c=2
缓冲液	1426	1260	604	1108
(a=1)	1183	1599	1081	886
	2000	1410	487	831
	1612	2416	624	1159
蒸馏水	684	875	867	1115
(a=2)	1430	2250	771	698
	1165	1871	403	791
	2022	1962	370	559
自来水	1182	1220	1243	1283
(a=3)	1512	1095	1115	1142
	1450	1700	416	677
	1385	2372	533	534

本例分析了三种因素及其所有交互作用不同水平间的差别。结果表明，兔血清与胎盘血清(b 因素)的钩端螺旋体计数有差别。而三种基础液(a 因素)间，两种浓度(c 因素)间的计数无差别，各因素间亦无交互作用。

§6.6　协方差分析

协方差分析是在扣除协变量的影响后再对（修正后的）主效应进行方差分析，是把直线回归或多元线性回归与方差分析结合起来的一种方法。协变量一般是连续性变量，并假设协变量与响应变量间存在线性关系，且在各处理组这种线性关系一致。用于协方差分析的命令是在 anova 命令后再加选择项 continuous(协变量名)，或 category(分组变量名)。

```
anova y a b c a*b b*c a*b*c…… x1 x2 ……, continuous(x1 x2 ……)
```

其中，y 为响应变量，a ,b……为分组变量，x1 x2 ……为协变量，加选择项 continuous(x1 x2……)的意思是指明 x1,x2……为连续性变量(协变量)，从而 Stata 自动以 x1,x2……为协变量进行协方差分析。在不指定连续性变量时，Stata 视所有变量为分组变量(响应变量除外)。亦可指定分组变量，则其余变量将视为是连续的，相应的选择项应改为 categroy( )，如

```
anova y a b c a*b b*c a*b*c…… x1 x2 ……, categroy(a b c ……)
```

与上述命令是等价的。

当有一个协变量时，称为一元协方差分析，当有两个或多个协变量时，称为多元协方差分析。

例 6.11(配伍组的协方差分析) 以下资料是三组小白鼠的进食量(x)与所增体重(y)，由于体重增加受进食量的影响，故在分析体重的增加时， 必须扣除进食量的影响。即以进食量为协变量，对三组的增加体重进行分析。这里，协变量为一个。

```
. use ex6-11
. list
```

	x	y	a	b
1.	256.9	27	1	1
2.	271.6	41.7	1	2
3.	210.2	25	1	3
4.	300.1	52	1	4
5.	262.2	14.5	1	5
6.	304.4	48.8	1	6
7.	272.4	48	1	7
8.	248.2	9.5	1	8
9.	242.8	37	1	9
10.	342.9	56.5	1	10
11.	356.9	76	1	11
12.	198.2	9.2	1	12
13.	260.3	32	2	1
14.	271.1	47.1	2	2
15.	214.7	36.7	2	3
16.	300.1	65	2	4
17.	269.7	39	2	5
18.	307.5	37.9	2	6
19.	278.9	51.5	2	7
20.	256.2	26.7	2	8
21.	240.8	41	2	9
22.	340.7	61.3	2	10
23.	356.3	102.1	2	11
24.	199.2	8.1	2	12
25.	544.7	160.3	3	1
26.	481.2	96.1	3	2
27.	418.9	114.6	3	3
28.	556.6	134.8	3	4
29.	394.5	76.3	3	5
30.	426.6	72.8	3	6
31.	416.1	99.4	3	7
32.	549.9	133.7	3	8

```
 33.     580.5       147         3         9
 34.     608.3     165.8         3        10
 35.     559.6     169.8         3        11
 36.     371.9      54.3         3        12
```

这里，a=1，2，3 表示相应的三个处理组，b=1，…，12 表示配伍组，x 为协变量。

```
. anova y a b x, cont(x)
```

```
                   Number of obs =      36      R-square     =  0.9706
                   Root MSE      = 10.2942      Adj R-square =  0.9511

      Source |  Partial SS    df       MS          F     Prob > F
-----------+----------------------------------------------------
       Model |  73560.9936    14  5254.35669      49.58     0.0000
             |
           a |   463.94765     2  231.973825       2.19     0.1369
           b |  3765.32692    11  342.302448       3.23     0.0101
           x |  6174.24838     1  6174.24838      58.26     0.0000
             |
    Residual |  2225.36425    21  105.969726
-----------+----------------------------------------------------
       Total |  75786.3579    35  2165.32451
```

在控制了进食量(x)的影响后，三组小白鼠所增体重无差别。方差分析中，进食量 x 的 P=0.0000，说明控制进食量 x 的影响是必要的。

也可用下列命令，结果是等价的。

```
. anova y a b x, categroy(a b)
```

例 6.12(2×3 析因试验的协方差分析)　分别以不同来源和成分的蛋白质喂养六组小白鼠，每组 10 只。资料见表 6.7。

表 6.7　六组公鼠的食物消耗量(x，10cal)及所增体重(y，g)

高蛋白,a=1						低蛋白,a=2					
牛肉 b=1		谷类 B=2		猪肉 b=3		牛肉 b=1		谷类 b=2		猪肉 b=3	
x	y	x	y	x	y	x	y	x	y	x	y
108	73	99	98	194	94	165	90	124	107	140	49
136	102	117	74	198	79	164	76	95	95	177	82
138	118	90	56	196	96	161	90	116	97	189	73
159	104	141	111	198	98	159	64	112	80	142	86
146	81	106	95	210	102	175	86	123	98	216	81
141	107	112	88	196	102	135	51	110	74	200	97
175	100	110	82	230	108	132	72	137	74	255	106
149	87	117	77	222	91	190	90	105	67	173	70
174	117	111	86	220	120	145	95	135	89	153	61
176	111	122	92	228	105	142	78	126	58	160	82

```
. use ex6-12
. list
```

```
          x          y          a          b
```

```
  1.       108         73          1          1
  2.       136        102          1          1
           ......
           ......
 60.       160         82          2          3

. anova y a b a*b x,cont(x)
                      Number of obs =      60      R-square     =  0.4694
                      Root MSE      = 12.7349      Adj R-square =  0.4093

             Source |  Partial SS    df       MS           F     Prob > F
        ------------+----------------------------------------------------
              Model |  7603.55945     6  1267.25991       7.81     0.0000
                    |
                  a |  2343.46252     1  2343.46252      14.45     0.0004
                  b |  1673.30508     2  836.652542       5.16     0.0090
                a*b |    933.8117     2   466.90585       2.88     0.0650
                  x |  2990.62611     1  2990.62611      18.44     0.0001
                    |
           Residual |  8595.37389    53  162.176866
        ------------+----------------------------------------------------
              Total |  16198.9333    59  274.558192
```

结果表明：在控制了食物消耗量（x）的影响后，用高蛋白与用低蛋白（a 因素）喂养小白鼠所增体重不同，用高蛋白喂养比用低蛋白喂养体重增加多；用牛肉、谷类、猪肉（b 因素）喂养小白鼠所增体重亦不同；但尚不能认为有交互作用。如不考虑协变量的影响，结论就不同了。请读者自行验算。

也可用下列命令，结果是等价的。

```
. anova y a b x, categroy(a b)
```

例 6.13(多元协方差分析)　某地测得 30 名初生至 3 周岁儿童的身高，体重及体表面积如表 6.8。欲比较男女体表面积是否相同，此时身高，体重为协变量，为二元协方差分析。

```
. input y x1 x2 sex
. ......
. sort sex
. by sex : summ y x1 x2

-> sex=        1
Variable |   Obs        Mean    Std. Dev.       Min        Max
---------+--------------------------------------------------------
       y |    15    4099.327    1592.838     1928.4     6410.6
      x1 |    15        75.2    18.30671       50.5         99
      x2 |    15    8.583333    4.804821       2.25         16

-> sex=        2
Variable |   Obs        Mean    Std. Dev.       Min        Max
---------+--------------------------------------------------------
       y |    15     3790.76    1543.524     1632.5     6074.9
```

```
      x1 |    15    73.16667   16.93229         51         94
      x2 |    15    8.116667   4.408663       2.25         15
```

表 6.8　　30 名儿童的身高(x1,cm).体重(x2,kg)及体表面积(y,cm2)

男,sex=1			女,sex=2		
x1	x2	y	x1	x2	y
54.0	3.00	2446.2	54.0	3.00	2117.3
50.5	2.25	1928.4	53.0	2.25	2200.2
51.0	2.50	2094.5	51.5	2.50	1906.2
56.5	3.50	2506.7	51.0	3.00	1850.3
52.0	3.00	2121.0	51.0	3.00	1632.5
76.0	9.50	3845.9	77.0	7.50	3934.0
80.0	9.00	4380.8	77.0	10.00	4180.4
74.0	9.50	4314.2	77.0	9.50	4246.1
80.0	9.00	4078.4	74.0	9.00	3358.8
76.0	8.00	4134.5	73.0	7.50	3809.7
96.0	13.50	5830.2	91.0	12.00	5358.4
97.0	14.00	6013.6	91.0	13.00	5601.7
99.0	16.00	6410.6	94.0	15.00	6074.9
92.0	11.00	5283.3	92.0	12.00	5299.4
94.0	15.00	6101.6	91.0	12.50	5291.5

计算男婴组，女婴组及合并的体表面积与体重、身高的相关系数。

```
. by sex :corr y x1 x2
-> sex=        1  (obs=15)
        |        y       x1       x2
--------+---------------------------
      y |   1.0000
     x1 |   0.9895   1.0000
     x2 |   0.9902   0.9809   1.0000

-> sex=        2  (obs=15)
        |        y       x1       x2
--------+---------------------------
      y |   1.0000
     x1 |   0.9901   1.0000
     x2 |   0.9806   0.9827   1.000

. corr y x1 x2
(obs=30)
        |        y       x1       x2
--------+---------------------------
      y |   1.0000
     x1 |   0.9887   1.0000
     x2 |   0.9842   0.9817   1.0000
```

由此可见，无论是男婴组还是女婴组，体表面积都与身高、体重高度相关，故应扣除两者的影响，再比较性别间体表面积有无差别。用协方差分析。

```
. anova y sex x1 x2,cont(x1 x2)
                      Number of obs =      30      R-square     =  0.9847
```

```
                Root MSE      = 202.428      Adj R-square =  0.9829

      Source |  Partial SS    df       MS           F     Prob > F
-----------+----------------------------------------------------
       Model |  68523073.3     3  22841024.4     557.41     0.0000
             |
         sex |   139769.36     1   139769.36       3.41     0.0762
          x1 |  938153.264     1  938153.264      22.89     0.0001
          x2 |  368955.082     1  368955.082       9.00     0.0059
             |
    Residual |  1065399.93    26  40976.9205
-----------+----------------------------------------------------
       Total |  69588473.2    29  2399602.52
```

方差分析的结果表明，根据现有资料，在扣除了身高、体重的影响后，男婴女婴的体表面积之差别无显著性，P＝0.0762。

该命令与下面的命令是等价的

```
. anova y sex x1 x2,categroy(sex)
```

（结果略）

§6.7 正态性检验与变量变换

正态性是很多传统统计方法的应用条件之一，如 *t* 检验，方差分析等均要求资料服从正态分布。如资料不服从正态分布，则需作适当的变量变换，以使资料达到或接近正态。

本节介绍几种正态性检验方法和几种常见的正态化和对称化变换。

一、正态性检验

用于正态性检验的命令为：

```
sktest    变量
```

该命令要求资料的样本含量至少为 8。先看一个实例。

例 6.14　某市 200 名正常成人的血铅含量(μg/100g)如下，试对其进行正态性检验。

```
 3  4  4  4  4  4  5  5  5  5  5  5  5  5  5  5  6  6  6  6  6  6  6  7  7
 7  7  7  7  7  7  7  7  7  7  8  8  8  8  8  8  8  8  8  8  8  9  9  9
 9  9  9  9 10 10 10 10 10 10 10 10 10 10 11 11 11 11 11 12 12 12 12 12 12 12
13 13 13 13 13 13 13 13 13 13 13 13 14 14 14 14 14 14 14 14 14 14 15 15 15
15 15 15 15 16 16 16 16 16 16 17 17 17 17 17 17 17 17 17 17 17 17 18 18 18
18 18 19 19 19 19 19 19 20 20 20 20 20 20 20 20 20 21 21 21 21 21 22 22 22 22
22 22 23 23 23 24 24 24 24 24 24 25 25 26 26 26 26 26 27 27 28 28 29 29 30
30 31 31 31 31 32 32 32 32 32 32 33 33 36 38 38 39 40 41 41 43 47 50 53 60
```

首先用 summ 命令计算偏度系数和峰度系数：

```
. summ x,d
                            x
-------------------------------------------------------------
      Percentiles      Smallest
 1%           4              3
```

```
 5%          5          4
10%          6          4       Obs                 200
25%          9          4       Sum of Wgt.         200

50%         15                  Mean              17.085
                     Largest    Std. Dev.       10.33984
75%         22          47
90%       31.5          50      Variance        106.9123
95%         38          53      Skewness        1.215245
99%       51.5          60      Kurtosis        4.734997
```

对 x 的偏度系数和峰度系数进行假设检验：

```
. sktest x
                Skewness/Kurtosis tests for Normality
                                                 ------- joint -------
 Variable |  Pr(Skewness)   Pr(Kurtosis)  adj chi-sq(2)  Pr(chi-sq)
----------+-----------------------------------------------------------
        x |      0.000          0.001          34.93        0.0000
```

结果中给出了偏度系数检验的 P 值[Pr(Skewness)]，峰度系数检验的 P 值[Pr(Kurtosis)]，以及偏度系数和峰度系数联合检验的校正χ2[adj chi-sq(2)]及检验概率 Pr(chi-sq)。结果表明，该资料不服从正态分布。这从资料的分布亦可判断。

```
. set tex 150
. gra x , bin(13) xlab(0,5,10,15,20,25,30,35,40,45,50,55,60,65) ylab(0,.1,.15,.2,.25) gap(3)
```

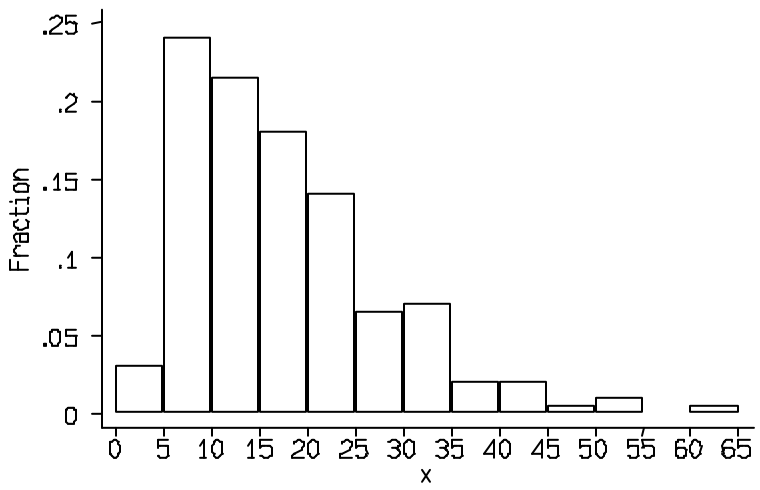


图 6.1　例 6.14 资料的频数分布图

现对 x 作对数变换，计算其对数值的偏度系数和峰度系数，并对其进行假设检验。

```
. gen lnx=ln(x)
. summ lnx ,d
                          lnx
```

```
-------------------------------------------------------------
      Percentiles      Smallest
 1%     1.386294       1.098612
 5%     1.609438       1.386294
10%     1.791759       1.386294       Obs                 200
25%     2.197225       1.386294       Sum of Wgt.         200

50%      2.70805                      Mean           2.658423
                        Largest       Std. Dev.      .6167802
75%     3.091043       3.850147
90%     3.449862       3.912023       Variance       .3804178
95%     3.637586       3.970292       Skewness      -.1735798
99%     3.941157       4.094345       Kurtosis       2.418212
```

```
. sktest lnx
              Skewness/Kurtosis tests for Normality
                                                 ------- joint -------
 Variable |  Pr(Skewness)   Pr(Kurtosis)  adj chi-sq(2)  Pr(chi-sq)
----------+-------------------------------------------------------
      lnx |      0.303          0.029           5.72        0.0574
```

结果中给出了对数值 $\ln x$ **的偏度系数检验的** P **值，峰度系数检验的** P **值，以及偏度系数和峰度系数联合检验的校正**χ^2 **及检验概率。结果表明，该资料经对数变换后，该资料已基本对称，但其峰度比正态峰扁平。按**$\alpha=0.10$ **水准，对数变换后的资料仍不服从正态分布。**

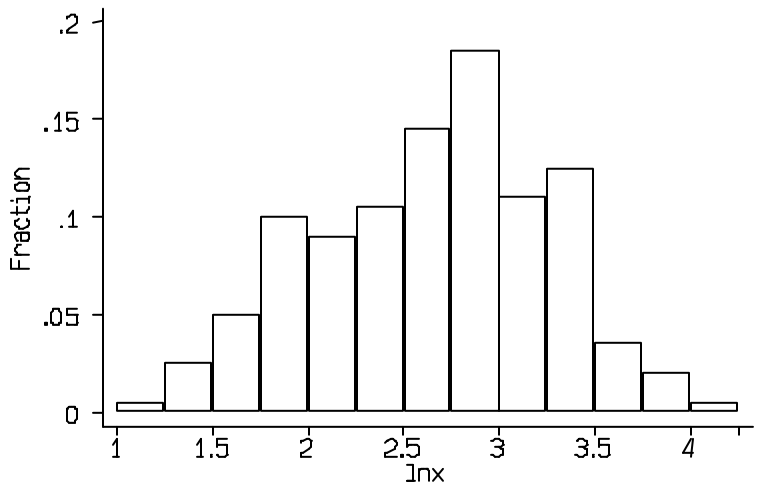


图 6.2　例 6.14 资料对数值的频数分布图

二、Box-Cox 正态性变换

所谓 Box-Cox **变换是指对变量** x **作变换：**

$$y=\begin{cases}\dfrac{x^{l}-1}{l} & l\neq 0\\ \ln(x) & l=0\end{cases}$$

Box-Cox 正态性变换就是寻找参数λ，使变换后的资料最接近正态分布。

用于寻找 Box-Cox 正态性变换的命令为：

boxcox 原变量 , generat(新变量)

例 6.15 对例 6.14 **资料作** Box-Cox **正态性变换。**

```
. boxcox x ,g(y)
(note: iterations performed using zero =.001)

   Iteration      Lambda         Zero     Variance        LL
   ------------------------------------------------------------
         0        1.0000    -73.90087   107.166828   -467.43868
         1        0.0491      5.98333   76.7498368   -434.05513
         2        0.1180     -0.00210   76.6147889   -433.87901
         3        0.1180      0.00000   76.6147808   -433.87900
   ------------------------------------------------------------

Transform:  (x^L-1)/L
                 L       [95% Conf. Interval]     Log Likelihood
            ---------------------------------------------------
               0.1180      (not calculated)             -433.879

     Test:  L == -1      chi2(1) =   106.69    Pr>chi2 =  0.0000
            L ==  0      chi2(1) =     1.31    Pr>chi2 =  0.2521
            L ==  1      chi2(1) =    65.64    Pr>chi2 =  0.0000
```

参数λ是用迭代的方法求出的，Stata **给出了迭代的中间步骤。结果，**λ=0.1180**。结果中还给出了分别与**λ=-1**，**λ=0**，及**λ=1(**不作变换**)时的**χ^2 **检验，表明，**λ=0.1180 **的变换与**λ=0(**对数变换**)无显著性，而比原资料有较大的改进。**

```
. gen lnx=log(x)
. sktest lnx y
                Skewness/Kurtosis tests for Normality
                                                 ------- joint -------
 Variable |  Pr(Skewness)   Pr(Kurtosis)  adj chi-sq(2)  Pr(chi-sq)
----------+-------------------------------------------------------
      lnx |      0.303         0.029            5.72       0.0574
        y |      0.898         0.028            4.89       0.0869
```

可见，作λ=0.1180 **的** Box-Cox **变换后的偏度系数，较作对数变换**(λ=0)**有所改善，而对资料的峰度系数则两种变换相差不大。**

Stata **还提供了其它检验正态分布的检验方法：**Shapiro-Wilk **法和** Shapiro-Francia **法。 命令为：**swilk **和** sfrancia **。**

三、对称性变换

所谓对称性变换，即寻找变换，使资料接近对称，或偏度系数接近 0。Stata **提供了两种对称性变换，其一是** Box-Cox **对称性变换，即寻找** Box-Cox **中的**λ**，使变换后资料的偏度系数接近** 0**；其二是对数对称性变换，即寻找一** k **值，作变换：**

$$y=\ln\left(\pm x-k\right)$$

使变换后资料 *y* **的偏度系数接近** 0**。相应的两个命令为：**

lnskew0 **新变量 =** ±**原变量**
bcskew0 **新变量 =** ±**原变量**

x **前面的正负号将根据其具体取值，由用户自己定义。**

例 6.16 **对例** 6.14 **资料作对称性变换。**

```
. lnskew0 ltx=x

      Transform |         k     [95% Conf. Interval]       Skewness
----------------+----------------------------------------------------
        ln(x-k) |  -2.090607      (not calculated)         .0006012
```

结果 k=-2.09**。此时的偏度系数为：**0.0006012**。在执行完命令后，**Stata **产生了一个新变量** ltx**，其取值为** ln(*x*+2.090607)**。**

```
. bcskew0 bcx=x,lev(95)

      Transform |         L     [95% Conf. Interval]       Skewness
----------------+----------------------------------------------------
      (x^L-1)/L |   .1349801    -.1275427      .38839        .0005503
```

结果λ=0.135**。此时的偏度系数为：**0.0005503**。在执行完命令后，**Stata **产生了一个新变量** bcx**，其取值为**$[x^{0.1349801}-1]/0.1349801$**。**

第七章 分类资料的统计分析

分类资料又称定性资料，或计数资料，其取值是定性的，表现为互不相容的类别或属性。按类别间的关系，又分为有序分类和无序分类。有序分类资料又称等级资料。等级资料的统计分析将在第八章介绍，本章介绍无序分类资料的统计分析。

§7.1 率、构成比的比较

率与构成比的资料形式一般都是行列表形式。Stata 用于处理分类资料的命令是双向(二维)tabulate 命令(参见第四章)。

tabulate var1 var2 [fw=频数变量] [，选择项]

其中 var1，var2 分别表示行变量和列变量，[fw=频数变量]只在变量以频数形式存放时选用。选择项有:

chi2	/* (Pearson)χ^2 检验
lrchi2	/* 似然比χ^2 检验
gamma	/* Goodman-Kruskal 的γ系数
taub	/* Kendall 的相关系数τ_b
V	/* Cramer 的列联系数 V
all	/* 同时给出以上五种结果
exact	/* Fisher 的确切概率
cell	/* 打印每个格子的频数占总频数的百分比
column	/* 打印每个格子的频数占相应列合计的百分比
row	/* 打印每个格子的频数占相应行合计的百分比
nofreq	/* 不打印频数

以上命令可同时选用。

分类资料的一个特点是重复数较多，在报告结论时，一般都将数据整理成频数表。但收集资料时都是未整理的原始形式。Stata 对这两种形式的资料都可以进行分析，所得结果相同，只是命令稍有区别。下面以两种数据形式，三种命令格式对四格表资料进行分析，以说明 tabulate 命令的应用。

例 7.1(两个率的比较，四格表) 试比较甲乙两种疗法对某病的治疗效果。

表 7.1 甲乙两种疗法对某病的治疗效果

	无效	有效	合计	有效率(%)
甲法	6	4	10	p1=40.0
乙法	11	33	44	p2=75.0
合计	17	37	54	p_C=68.5

(1) 频数形式

记 a=0 表甲法，a=1 表乙法；b=0 表无效，b=1 表有效；freq 表相应的频数，数据结构如下:

```
. use d:\mydata\ex7-1
```

```
. list
          a         b      freq
  1.      0         0         6
  2.      0         1         4
  3.      1         0        11
  4.      1         1        33
```

数据是以频数的形式存放的，频数变量为 freq。相应的命令格式为：

```
. tab a b [fw=freq ], row all exact
```

```
           | effect
    mathod |         0          1 |     Total
-----------+----------------------+----------
         0 |         6          4 |        10
           |     60.00      40.00 |    100.00
-----------+----------------------+----------
         1 |        11         33 |        44
           |     25.00      75.00 |    100.00
-----------+----------------------+----------
     Total |        17         37 |        54
           |     31.48      68.52 |    100.00

         Pearson chi2(1) =   4.6273   Pr = 0.031
 likelihood-ratio chi2(1) =   4.3274   Pr = 0.038
              Cramer's V =   0.2927
                   gamma =   0.6364  ASE = 0.218
         Kendall's tau-b =   0.2927  ASE = 0.143
          Fisher's exact =                 0.056
  1-sided Fisher's exact =                 0.041
```

由于选用了 all 和 exact 项，结果中给出了包括 Fisher 确切概率在内的全部六种统计量，注意，all 选择中不包括 Fisher 的确切概率。各统计量的计算见后。结论：本例样本含量较小，且有理论频数小于 5，应以 Fisher 的确切概率下结论。按 $\alpha=0.05$ 水准，尚不能认为两种疗效有差别。

(2) 原始资料形式

分类资料在收集资料时都是未整理的原始形式。Stata 对这种资料可直接以进行分析，所得结果相同。如例 7.1 资料用原始资料形式存放即为：

```
. drop _all
. use d:\mydatat\ex7-1-1
. list
          a         b
  1.      0         0  ┐
  2.      0         0  │
          .....        │ 6
          .....        │
  6.      0         0  ┘
  7.      0         1  ┐
          .....        │
                       ┘
```

```
          ......          4
10.       0       1
11.       1       0
          ......
          ......          11
21.       1       0
22.       1       1
          ......
          ......          33
54.       1       1
```

此时，相应的命令为：

```
. tab  a  b , all exact row
```

命令中没有“ [fw=freq] **”选择项，但所得结果相同。**

(3) **直接输入频数**

对频数表资料还可用“ tabi **”命令直接输入频数，按行输入，各行数据间用“** \ **”分开。因该法较之上两法更为简单，故推荐使用。命令格式如下：**

```
. tabi 6 4\11 33 ,row all exact
```

结果相同，略。

[注]　设有下列 R×C 表表：

	1	2	3	…	j	…	C	
1	n_{11}	n_{12}	n_{13}	…	n_{1j}		n_{1C}	$n_{1.}$
2	N_{21}	n_{22}	n_{23}	…	n_{2j}		n_{2C}	$n_{2.}$
…	…	…	…	…	…	…	…	…
i	n_{i1}	n_{i2}	n_{i3}	…	n_{ij}		n_{IC}	$n_{i.}$
…	…	…	…	…	…	…	…	…
R	n_{R1}	n_{R2}	n_{R3}	…	n_{Rj}	…	n_{RC}	$n_{R.}$
	$n_{.1}$	$n_{.2}$	$n_{.3}$	…	$n_{.j}$	…	$n_{.C}$	$n..$

记：

$$m_{ij} = n_{i}.\times n._{j} / n..$$

$$A_{ij} = \sum_{k>i}\sum_{l>j} n_{kl} + \sum_{k<i}\sum_{l<j} n_{kl}$$

$$D_{ij} = \sum_{k<i}\sum_{l>j} n_{kl} + \sum_{k>i}\sum_{l<j} n_{kl}$$

$$P = \sum\nolimits_{i}\sum\nolimits_{j} n_{ij}A_{ij} \qquad Q = \sum\nolimits_{i}\sum\nolimits_{j} n_{ij}D_{ij}$$

则：

(1) (Pearson)χ^2:

$$Q_p = \sum\nolimits_{i}\sum\nolimits_{j} (n_{ij} - m_{ij}) / m_{ij}$$

(2) **似然比**χ^2:

$$G=\sum_i\sum_j n_{ij}\ln(n_{ij}/m_{ij})$$

(3) Cramer **列联系数**:

$$V=\begin{cases}(n_{11}n_{22}-n_{12}n_{21})/(n_1.n_2.n._1n._2)^{1/2} & \text{对 } 2\times2 \text{ 表}\\ [(QP/n)/\min(R-1,\ C-1)]^{1/2} & \text{其 他}\end{cases}$$

(4) Goodman-Kruskal **的** γ :

$$\text{gamma}=(P-Q)/(P-Q)$$

(5) Kendall **的列联系数**:

$$\tau_b=(P-Q)/(w_Rw_C)^{1/2}$$

$$w_r=n^2-\Sigma in_{i.}^2 \qquad w_c=n^2-\Sigma_j n_{j.}^2$$

例 7.2(多个率的比较) 用免疫法观察鼻咽癌患者(a=1)**、头颈部其他恶性肿瘤患者**(a=2)**及正常成人组**(a=3)**的血清** EB **病毒壳抗原的免疫球蛋白** A(VCA-IgA)**抗体的反应情况，资料如下。三组阳性率有无差别**?

表 7.2 三组人群中 EB 病毒 VCA-IgA 抗体阳性率

分 组	阳性例数	阴性例数	合 计	阳性率(%)
a=1	188	16	204	92.3
a=2	10	23	33	30.3
a=3	49	333	382	12.8
合 计	247	372	619	39.9

按频数形式输入原始数据。

```
. list
          a          b        pop
  1.      1          1        188
  2.      2          1         10
  3.      3          1         49
  4.      1          0         16
  5.      2          0         23
  6.      3          0        333
```

```
. tab a b [fw=pop] , row chi2 lrchi2 exact
```

```
           | b
         a |         0          1 |     Total
-----------+----------------------+----------
         1 |       188         16 |       204
           |     92.16       7.84 |    100.00
-----------+----------------------+----------
         2 |        10         23 |        33
           |     30.30      69.70 |    100.00
```

```
-----------+----------------------+----------
         3 |         49         333 |       382
           |      12.83       87.17 |    100.00
-----------+----------------------+----------
     Total |        247         372 |       619
           |      39.90       60.10 |    100.00

          Pearson chi2(2) = 350.3259   Pr = 0.000
 likeli hood-ratio chi2(2) = 387.3664   Pr = 0.000
            Fisher's exact =                 0.000
```

也可直接用以下命令：

```
. tabi 188 16\10 23\49 333,row chi2 lrchi2 exact
```

所得结果同上。结论：无论是卡方检验还是似然比检验，按α＝0.05 水准可认为三组阳性率不同。鼻咽癌患者的反应阳性率最高，正常成人组的反应阳性率最小。

如在 DOS 版本上使用，当总例数大于 170 时，即使命令中选用“ exact ”，也不能给出 Fisher 的确切概率。3.0 以上的版本无此限制。

例 7.3(多组构成比的比较)　就下表资料分析三个民族的血型分布(构成比)是否相同。

表 7.3　傣族、佤族、土家族居民的 ABO 血型分布

		A (xx=1)	B (xx=2)	O (xx=3)	AB (xx=4)	合计
傣　族	(mz=1)	112	150	205	40	507
佤　族	(mz=2)	200	112	135	73	520
土家族	(mz=3)	362	219	310	69	960
合　计		674	481	650	182	1987

```
. tabi 112 150 205 40\200 112 135 73\362 219 310 69,nofreq row chi2 lrchi2
          | xx
       mz |         1          2          3          4 |     Total
----------+---------------------------------------------+----------
        1 |     22.09      29.59      40.43       7.89 |    100.00
        2 |     38.46      21.54      25.96      14.04 |    100.00
        3 |     37.71      22.81      32.29       7.19 |    100.00
----------+---------------------------------------------+----------
          |     33.92      24.21      32.71       9.16 |    100.00

         Pearson chi2(6) =  71.5186   Pr = 0.000
 likelihood-ratio chi2(6) =  72.2521   Pr = 0.000
```

按命令要求，结果中给出了 Pearson 的χ^2检验和似然比χ^2检验。

结论：卡方检验与似然比检验的 P 值均较小，可认为三个民族的血型分布不同。其中傣族以O型为主，而佤族与土家族均以A型为多。

例 7.4(计数相关)　就下列资料分析人群中 ABO 血型与 MN 血型有无相关关系。

表 7.4　6094 人 MN 血型与 ABO 血型的分布

ABO 血型	MN 血型

	M	N	MN	合计
A	431	490	902	1823
B	388	410	800	1598
O	495	587	950	2032
AB	137	179	325	641
合　计	1451	1666	2977	6094

以 a 表 ABO 血型，b 表 MN 血型，c 为相应的频数。

```
. tabi 431 490 902\388 410 800\495 587 950\137 179 325,nofreq all

         Pearson chi2(6) =   8.5952   Pr = 0.198
 likelihood-ratio chi2(6) =   8.6689   Pr = 0.193
              Cramer's V =   0.0266
                   gamma =  -0.0078  ASE = 0.017
         Kendall's tau-b =  -0.0053  ASE = 0.011
```

结论：从列联系数来看，Cramer 的 V，Goodman-Kruskal 的 gamma(γ)，以及 Kendall 的 t_b 均较小；从 P 值来看，无论是卡方检验还是似然比检验，P 值均较大，尚不能认为两种血型间有相关关系。

§7.2　流行病学表格分析

在流行病学资料分析中，经常要计算某事件的发生率(如发病率、死亡率等)、率差、相对危险度(RR)、比数比(OR)及它们的可信区间等。用该软件可以非常方便地解决此类问题。

Stata 用于处理流行病简单表格资料的命令有：ir，cs，cc，mcc 等。他们分别适用于定群研究，病例对照研究和配比病例对照研究。详细说明请查阅帮助：help　epitab。

一、定群研究资料

定群研究又称队列研究，前瞻性研究，随访研究或纵向研究。在定群研究时，根据以往有无暴露经历，研究者将研究人群分为暴露和非暴露，在一定时间内，随访观察和比较两组人群的发病率或死亡率。对定群研究的资料，Stata 提供了 ir 和 cs 命令。

ir　病例变量　暴露变量　时间变量　[,选择项]

cs　病例变量　暴露变量　　[,选择项]

这里选择项有：

```
level(#)            /# 指定可信区间的可信度
tb                  /# 以检验方法为基础，作可信区间的估计
by(varname)         /# 指定分层变量
fast                /# 不计算层内 OR 或可信区间
estandard           /# 指定用外在权数计算标准化估计，与 by()一起用
istandard           /# 指定用内在权数计算标准化估计，与 by()一起用
standard(varname)   /# 指定按变量为权数计算标准化估计，与 by()一起用
ird                 /# 指定计算标准化率之差，用于 estandard，istandard 或 standard
                       选择项后
```

```
nocrude          /# 不计算合并资料的指标。用于 by()选择项后
pool             /# 直接加权估计，与 by()一起用
nohet            /# 不做层间的齐性检验
```

ir 命令适用于发病率(发病密度或人-时资料)，主要用于估计发病密度比和差；而 cs 适用于随访时间相同、随访资料的分母是观察对象数而不是人时数的资料。这两种类型的频数资料都能直接用快速命令 iri 或 csi，格式如下：

```
iri    #a #b #N1 #N2   [, level(#) tb ]
csi    #a #b  #c  #d   [, level(#) exact or tb woolf ]
```

例 7.5　就表 7.5 资料进行流行病学分析。

表 7.5　暴露和不暴露 X 线患结核病妇女乳腺癌病例发生数和观察人年数

	暴露	不暴露	合计
病例数	41(a)	15(b)	56(M)
人年数	28,010(N1)	19,017(N2)	47,027(T)

凡此种含有时间变量的资料，应采用 iri 或 ir 命令分析之。

```
. iri 41 15 28010 19017

                |   Exposed   Unexposed  |     Total
----------------+------------------------+----------
          Cases |        41          15  |        56
    Person-time |     28010       19017  |     47027
----------------+------------------------+----------
                |                        |
  Incidence Rate |  .0014638    .0007888  |  .0011908
                |                        |
                |         Pt. Est.       |  [95% Conf. Interval]
                |------------------------+----------------------
 Inc. rate diff. |          .000675       |  .0000749    .0012751
 Inc. rate ratio |         1.855759       |  1.005815    3.611192  (exact)
 Attr. frac. ex. |         .4611368       |  .0057813    .7230831  (exact)
 Attr. frac. pop |          .337618       |
                +-------------------------------------------------
                  (midp)   Pr(k>=41) =                  0.0177  (exact)
                  (midp) 2*Pr(k>=41) =                  0.0355  (exact)
```

解释： *RD*(**率差**) = 0.000675 = 6.75/**万**，　95%CI（0.749/**万**,12.751/**万**）

RR(**相对危险度**) = 1.855759，　95%CI（1.005815, 3.611192）

ARP(**归因危险度百分比**) = 0.4611368，　95%CI（0.0057813, 0.7230831）

PARP(**人群归因危险度百分比**) = 0.337618

P = 0.0177　**（单侧）**

根据 Stata 输出的结果，暴露 X 线患结核病妇女发生乳腺癌的危险性为非暴露者的 1.86 倍；暴露者中有 46% 的乳腺癌是由暴露 X 线所致；人群中乳腺癌的 33.8%是由接触 X 线所致。

注意：该命令中数据的输入顺序必须正确，依次为暴露组病例数、非暴露组病例数、暴露

组观察人时数、非暴露组观察人时数。一旦数据输入顺序有误，则结果将大相径庭，请读者自己验证。

也可用 ir 命令，首先输入数据：

```
. input case exposed time
        case    exposed       time
  1. 41 1 28010
  2. 15 0 19017
  3. end

. ir case exposed time

                |   Exposed   Unexposed  |     Total
----------------+------------------------+----------
          Cases |        41          15  |        56
    Person-time |     28010       19017  |     47027
----------------+------------------------+----------
                |                        |
  Incidence Rate |  .0014638    .0007888  |  .0011908
                |                        |
                |         Pt. Est.       |  [95% Conf. Interval]
                |------------------------+----------------------
 Inc. rate diff. |          .000675       |  .0000749    .0012751
 Inc. rate ratio |         1.855759       |  1.005815    3.611192  (exact)
 Attr. frac. ex. |         .4611368       |  .0057813    .7230831  (exact)
 Attr. frac. pop |          .337618       |
                +-------------------------------------------------
                  (midp)   Pr(k>=41) =                    0.0177  (exact)
                  (midp) 2*Pr(k>=41) =                    0.0355  (exact)
```

结果与前完全相同。

例 7.6　就表 7.6 资料计算妇女乳腺癌 *RR* 及 90%可信区间。

表 7.6　母亲乳汁中 IgG 抗体滴度高低与 6 个月以上婴儿患呼吸道疾病的关系

	高滴度	低滴度
发病	5 (a)	16 (b)
不发病	10 (c)	7 (d)
合计	15	23

最简单的 csi 命令为：csi #a #b #c #d，请注意数据输入顺序。

```
. csi 5 16 10 7,level(90)

                |   Exposed   Unexposed  |     Total
----------------+------------------------+----------
          Cases |         5          16  |        21
       Noncases |        10           7  |        17
----------------+------------------------+----------
          Total |        15          23  |        38
                |                        |
           Risk |  .3333333    .6956522  |  .5526316
```

```
                |                      |
                |        Pt. Est.      |  [90% Conf. Interval]
                |----------------------+----------------------
Risk difference |       -.3623188      | -.6172448   -.1073928
     Risk ratio |        .4791667      |  .2521484    .9105775
Prev. frac. ex. |        .5208333      |  .0894225    .7478516
Prev. frac. pop |        .2055921      |
                +-------------------------------------------
                          chi2(1) =     4.82  Pr>chi2 = 0.0281
```

解释：选择项 level(#)指定可信区间的可信度，该例中选 90%，若缺省则为 95%。

RR = 0.4791667， 90% CI: (0.2521484, 0.9105775)

χ^2 = 4.82 P = 0.0281

从而可认为：母亲乳汁中 IgG 抗体滴度高低与婴儿呼吸道疾患发生率有关，滴度较低者其婴儿较易患呼吸道疾病。

若用 cs 命令，资料输入形式如下：

```
. input d e pop
        d          e        pop
  1.    1          1          5
  2.    1          0         16
  3.    0          1         10
  4.    0          0          7
  5.    end

. cs d e [freq=pop]
```

结果同上，略。

二、病例-对照研究资料

病例-对照研究，又称回顾性调查研究。它是先按疾病状态确定调查对象，分为病例和对照两组，然后利用已有的记录，或用询问，或调查等方法，了解其以往(发病前)的暴露情况，并进行比较，推测疾病与暴露之间的联系。

Stata **用于简单病例对照-研究资料分析的命令是：**cc **和** cci**，用于配比病例对照研究资料分析的命令是：**mcc **和** mcci**。**

```
cc      病例变量  暴露变量  [，选择项]
mcc     暴露-病例数  暴露-对照数  [, level(#) tb ]
cci     #a #b #c #d  [,level(#) exact tb woolf]
mcci    #a #b #c #d  [, level(#) tb ]
```

这里选择项有：

```
level(#)：          /# 指定可信区间的可信度
tb                  /# 以检验方法为基础，作可信区间的估计
by(varname)         /# 指定分层变量
exact               /# 指定计算 Fisher 的确切概率
fast                /# 不计算层内 OR 或可信区间
estandard           /# 指定用外在权数计算标准化估计，与 by()一起用
```

```
istandard            /# 指定用内在权数计算标准化估计，与 by()一起用
standard(varname)    /# 指定按变量为权数计算标准化估计，与 by()一起用
ird                  /# 指定计算标准化率之差，用于 estandard，istandard 或 standard
                        选择项后
nocrude              /# 不计算合并资料的指标。用于 by()选择项后
pool                 /# 直接加权估计，与 by()一起用
nohet                /# 不做层间的齐性检验
```

例 7.7(病例对照研究)　某单位研究胸膜间皮瘤与接触石棉的关系，资料见表 7.7。试对其进行分析。

表 7.7　　胸膜间皮瘤与接触石棉的关系

组别	以往接触过石棉	未接触过石棉	合计
间皮瘤病例	40	36	76
对照	9	67	76
合计	49	103	152

处理该类资料的命令有 cci 及 cc 两种。

```
. cci 40 36 9 67

                                                    Proportion
                |   Exposed   Unexposed  |     Total     Exposed
----------------+------------------------+---------------------
          Cases |        40          36  |        76      0.5263
       Controls |         9          67  |        76      0.1184
----------------+------------------------+---------------------
          Total |        49         103  |       152      0.3224
                |                        |
                |         Pt. Est.       |  [95% Conf. Interval]
                |------------------------+---------------------
     Odds ratio |         8.271605       |  3.650693    18.67752  (Cornfield)
 Attr. frac. ex. |         .8791045       |  .7260794    .9464597  (Cornfield)
 Attr. frac. pop |         .4626866       |
                +-------------------------------------------------
                           chi2(1) =    28.94  Pr>chi2 = 0.0000
```

解释：*OR* = 8.271605,　　**其** 95% **可信区间**（3.650693　18.67752）

ARP = 87.91045%　　**其** 95% **可信区间**（0.7260794　0.9464597）

PARP = 46.26866%

χ^2 = 28.94　　P < 0.0001

由此可见，接触石棉者发生间皮瘤的危险性为未接触者的 8.27 **倍；接触石棉工人中有** 87.91% **的间皮瘤是由接触石棉所致；人群中间皮瘤的** 46.26%**是因接触石棉所致。**

若用 cc **命令，则需将资料整理成频数形式。结果相同。**

例 7.8(*M-H* **分层病例对照资料)　在吸烟与肺癌发生关系的研究中，年龄是一个混杂因素，试根据下列资料计算调整年龄后吸烟与肺癌发生关系的比数比,并作假设检验。**

表 7.8　　吸烟与肺癌的病例对照研究

结果	年龄（岁）			
	<40（age=0）		≥40 岁（age=1）	
	>1 包/天(e=1)	<1 包/天(e=0)	>1 包/天(e=1)	<1 包/天(e=0)
病例(d=1)	58	73	50	111
对照(d=0)	100	280	41	380

```
. input d e pop age
          d         e        pop        age
  1.      1         1         58          0
  2.      1         0         73          0
  3.      0         1        100          0
  4.      0         0        280          0
  5.      1         1         50          1
  6.      1         0        111          1
  7.      0         1         41          1
  8.      0         0        380          1

. cc d e [freq=pop],by(age)

             age |       OR      [95% Conf. Interval]    M-H Weight
-----------------+-------------------------------------------------
               0 |   2.224658     1.473033   3.360231      14.28571 (Cornfield)
               1 |   4.174907      2.63077    6.62562      7.819588 (Cornfield)
-----------------+-------------------------------------------------
           Crude |   2.747456      2.03801   3.704017               (Cornfield)
    M-H combined |   2.914544     2.143308   3.963297
-----------------+-------------------------------------------------
Test for heterogeneity (M-H)     chi2(1) =     3.944  Pr>chi2 = 0.0470

                    Test that combined OR = 1:
                                Mantel-Haenszel chi2(1) =     49.43
```

结果中给出了每一层的 *OR* 及其 95%可信区间，以及未调整年龄时两组合并(Crude)的 *OR*＝2.747456；调整年龄后，OR_{MH}＝2.914544，χ^2_{MN}=49.43，df=1，*P*=0.0000。故可认为调整年龄后，吸烟 1 包/天以上者与 1 包/天以下者相比，患肺癌的危险度为 2.91，且有统计学意义。

例 7.9(配对病例-对照研究)　某单位为研究软组织肉瘤与接触苯氧乙酸或氯酚的关系，作了一次病例-对照研究，结果见表 7.9 资料。

表 7.9　　软组织肉瘤与接触苯氧乙酸或氯酚的配对病例对照研究

		对照	
		暴露过	未暴露过
软组织肉瘤	暴露过	3	16
	未暴露过	3	30

直接用 mcci **命令：**

```
. mcci 3 16 3 30

                           Controls
Cases              |   Exposed   Unexposed  |     Total
-------------------+------------------------+----------
           Exposed |         3          16  |        19
         Unexposed |         3          30  |        33
-------------------+------------------------+----------
             Total |         6          46  |        52

McNemar's chi2(1) =      8.89        Pr>chi2 =  0.0029

Proportion with factor
        Cases       .3653846
        Controls    .1153846     [95% conf. interval]
                   ---------     --------------------
        difference       .25      .0811853   .4188147
        ratio       3.166667      1.422659   7.048617
        rel. diff.  .2826087      .1253028   .4399146

        odds ratio  5.333333      1.528067   28.47879   (exact)
```

可知，OR=5.33**，其** 95% CI **为**(1.528067，28.47879)，χ^2 = 8.89**，自由度为** 1，P = 0.0029**，故可认为接触苯氧乙酸或氯酚者发生软组织肉瘤的危险性为不接触者的** 5.33 **倍。**

注意：数据输入顺序必须正确。

若用 mcc **命令，则先整理资料如下：**

```
. input case cntl  pop
            case       cntl          pop
  1. 1 1  3
  2. 1 0 16
  3. 0 1  3
  4. 0 0 30
  5. end
```

命令为：

```
. mcc case cntl [fw=pop]
```

（结果同上，略）

第八章 等级资料的统计分析

等级资料是一类常见的资料，如临床上的无效(-)，有效(+)，显效(++)，痊愈(+++)等。处理这类资料时，常将它们用数值来代替，如以 0 代－，以 1 代＋，以 2 代++，以 3 代+++等。这里，数值之间的关系仅仅是等级关系，例如，3 比 2 大一个等级，1 亦比 0 大一个等级，而不能认为等级 3 与等级 2 的差等于等级 1 与等级 0 的差。这类资料的统计分析常用秩和检验、等级相关等。另外，数值变量资料在不满足 *t* 检验、方差分析、相关分析等的条件时，亦可用秩和检验或等级相关。

Stata 用于等级资料分析的命令有:

```
genrank      /# 编秩
signtest     /# 符号检验
signrank     /# 符号秩和检验(Wilcoxon)
ranksum      /# 两样本秩和检验(Wilcoxon-Mann-Whitney)
wilcoxon     /# 两样本秩和检验(Wilcoxon)
kwallis      /# 多样本秩和检验(Kruskal-Wallis)
spearman     /# 等级相关(Spearman)
ktau         /# 等级相关(Kendall)
```

详见以下各节。

§8.1 秩变换

将一组数据按从小到大的顺序编成秩次，称为秩变换。Stata 用于秩变换的命令为 genrank:

```
genrank var = original_var
```

genrank 命令首先对原变量(original_var)按从小到大的顺序排列，然后进行编秩，数据相同者编以等秩，缺失值不参加编秩，并将秩次赋予新变量(var)。

例 8.1 对数据 0 -2 -3 4 ? 5 47 0 进行编秩，其中?为缺失值。

```
. genrank rankx=x
. list
        x      rankx
 1.    -3          1
 2.    -2          2
 3.     0        3.5
 4.     0        3.5
 5.     4          5
 6.     5          6
 7.    47          7
 8.     .          .
```

原数据中有两个 0，排在第三、四位，由于取值相等，故秩次亦取平均值。原数据中的缺

失值排在最后，由于取值未知，故不参加编秩。

§8.2 配对样本的比较

Stata 用于配对等级资料检验的命令有 signtest 和 signrank，其形式与 ttest 命令相似:

```
signtest  var = 常数
signtest  var1 = var2
signrank  var1 = var2
```

其中第一条指令用于样本中位数与总体中位数的比较；第二条指令用于符号检验；第三条指令用于 Wilcoxon 符号秩和检验。

例 8.2 12 名宇航员航行前及返航后 24 小时的心率(次/分)如下，问宇航对心率有无影响?

宇航员:	A	B	C	D	E	F	G	H	I	J	K	L
航行前(x):	76	71	70	61	80	59	74	62	79	72	84	63
航行后(ax):	93	68	65	65	93	78	83	79	98	78	90	60

(1) 用 signtest 命令:

```
. signtest x=ax

Test: Equality of medians (Matched-Sample Sign Test)
Result of x - (ax)
 ------------------
 Positive          3
 Negative          9
 ------------------
 Total            12
one-sided binomial Pr(k >= 9) = 0.0730
two-sided binomial Pr(k >= 9) = 0.1460
```

signtest 命令给出了 x 与 ax 之差大于 0 的个数(Positive)及小于 0 的个数(Negative)，如出现差数为 0，则正负各半。结果中还给出了单双侧的确切概率(二项分布)。由于双侧概率 P=0.1460，故尚不能认为宇航对心率有影响。

(2) 用 signrank 命令:

```
. signrank x=ax

Test:  Equality of distributions (Wilcoxon Signed-Ranks)
Result of x - (ax)
 Sum of Positive Ranks = 7
 Sum of Negative Ranks = 71
 z-statistic  -2.51
 Prob > |z|    0.0121
```

signrank 命令给出了正、负秩次之和及近似正态检验的结果。

欲检验宇航前的心率之中位数是否为 76 次/分，命令如下:

```
. signtest x = 76
```

结果略。

§8.3　两样本比较

Stata 用于两等级资料比较的命令为 ranksum，wilcoxon:

```
ranksum   var ，by(group_var)
wilcoxon  var ，by(group_var)
```

这两个命令所得结果是等价的，其中 group_var 为分组变量，注意，by()是必选项。

例 8.3　测得铅作业与非铅作业工人的血铅值(μg/100g)如下，试检验两组血铅值有无差别。

非铅作业组(a=1):	5	6	7	9	12	13	15	18	21
铅作业组(a=2):	17	18	20	25	34	43	44		

以 x 表示血铅值，a 表示分组。

(1) 用 ranksum 命令:

```
. ranksum x, by(a)

Test: Equality of medians (Two-Sample Wilcoxon Rank-Sum)
Sum of Ranks: 86.5 (a == 2)
 Expected Sum: 59.5
z-statistic  2.86
 Prob > |z|   0.0043
```

ranksum 命令首先对 x 进行编秩，其方法与 genrank 命令相同。结果中给出样本含量较小组的秩和，本例第二组样本含量较小，故给出 a=2 时的秩和: Sum of Ranks：86.5 (a==2)；结果中还给出了两组中位数相同时，该组的期望秩和：Expected sum：59.5，及正态近似检验。结论：按α＝0.05 水准，可认为铅作业工人的血铅值大于非铅作业工人。

(2) 用 wilcoxon 命令:

```
. wilcoxon x,by(a)
Test:  a==1 has longer survival time
       Wilcoxon-Gehan statistic =       -54
                              z =     -2.86
                         Pr>|z| =    0.0042
```

wilcoxon 命令主要用于两组生存率的 Wilcoxon-Gehan 检验(见第十六章)，当命令中省略 deadvar 时，wilcoxon 命令进行的是 Wilcoxon-Mann-Whitney 秩和检验。本例两个命令所得结论相同。

§8.4　多样本比较

Stata 用于多组等级资料比较的命令是 kwallis:

```
kwallis var ，by(group_var)
```

该命令对多组等级资料进行 Kruskal－Wallis 检验，其中 group_var 是分组变量，by( )是必选项。

例 8.4　试检验下表中三组人的血浆总皮质醇含量有无差别。

表 8.1　三组人的血浆总皮质醇测定值(μg/L)

正常人 (a=1)	单纯性肥胖 (a=2)	皮质醇增多症 (a=3)
0.4	0.6	9.8
1.9	1.2	10.2
2.2	2.0	10.6
2.5	2.4	13.0
2.8	3.1	14.0
3.1	4.1	14.8
3.7	5.0	15.6
3.9	5.9	15.6
4.6	7.4	21.6
7.0	13.6	24.0

```
. input y a

1.   0.4  1
2.   1.9  1
     ……
30.  24.0  3
31.  end

. kwallis y , by(a)

Test: Equality of populations (Kruskal-Wallis Test)
        a       _Obs   _RankSum
        1        10      96.50
        2        10     117.50
        3        10     251.00
 chi-square =    18.122 with 2 d.f.
probability =     0.0001
```

kwallis 命令首先对 x 进行编秩，方法与 genrank 命令相同。结果中给出了各组的秩和（_Ranksum）及χ^2检验。按α＝0.05 水准，认为三组人的血浆总皮质醇测定值不同。

§8.5　等级相关

Stata 用于等级相关分析的命令为 spearman 和 ktau:

```
spearman  x y
ktau  x y
```

分别进行 Spearman 等价相关和 Kendall 等价相关分析。

例 8.5　在肝癌病因研究中，调查某地 10 个公社的肝癌死亡率(y，1/10 万)与某种食物中黄曲霉毒素相对含量（x，以最高为 10），结果如下，试推断两者间有无相关关系。

肝癌死亡率：　3.7　1.0　1.7　0.7　4.0　5.1　5.5　5.7　5.9　10.0

黄曲霉毒素相对含量：　46.5　18.9　14.4　21.5　27.3　64.6　46.3　34.2　77.6　55.1

```
. spearman x y
Number of obs =      10
Spearman's rho =      0.7455        /#  Spearman ρs
```

```
Test of Ho: x and y independent
      Pr > |t| =       0.0133
. ktau x y
  Number of obs =      10
Kendall's tau-a =       0.5111      /#  Kendall τa
Kendall's tau-b =       0.5111      /#  Kendall τb
Kendall's score =       23          /#  Kendall 记分
    SE of score =       11.180      /#  Kendall 记分的标准误
Test of Ho: x and y independent
       Pr > |z| =       0.0491  (continuity corrected)
```

spearman 命令及 ktau 命令均首先求出等级相关系数，然后进行假设检验。尽管 Kendall 等级相关系数τ_b 与 Spearman 等级相关系数ρ_S 不相等，但两者有函数关系。结论：肝癌死亡率与某种食物中黄曲霉毒素相对含量有关，黄曲霉毒素相对含量愈多，肝癌死亡率就愈大。

§8.6 秩变换检验

所谓秩变换检验是指将原变量的秩次视为数值变量，对其用相应的参数方法对秩次进行检验或分析。当样本含量较大时，其结果与相应的非参数方法基本一致。但样本含量较小时，宜选用非参数方法。

例 8.6 对例 8.4 资料用秩变换检验。

```
. use d:\mydata\ex8-4
```

首先对 y 进行秩变换，记相应的秩次为 ry，然后对 ry 进行方差分析:

```
. genrank ry=y
. anova ry a
```

```
                  Number of obs =      30      R-square     = 0.6252
                  Root MSE      = 5.58454      Adj R-square = 0.5974
     Source |  Partial SS    df       MS           F    Prob > F
------------+----------------------------------------------------
      Model |     1404.45     2     702.225      22.52    0.0000
            |
          a |     1404.45     2     702.225      22.52    0.0000
   Residual |      842.05    27   31.187037
------------+----------------------------------------------------
      Total |     2246.50    29  77.4655172
```

结论同例 8.4。

例 8.7(等级相关) 用秩变换检验的思想方法分析例 8.5 资料。

首先分别对 x，y 进行编秩，记为 rx，ry，再对 rx，ry 进行相关分析:

```
. genrank rx=x
. genrank ry=y
. corr rx ry
```

```
(obs=10)
             |       rx       ry
-------------+------------------
          rx |   1.0000
          ry |   0.7455   1.0000
```

结论同例 8.5。**事实上，**x **与** y **的** Spearman **相关系数等于** rx **与** ry **的** Pearson **相关系数，但检验方法不同。**

例 8.8（ 配伍组的秩变换检验 ） 每隔两个月抽样检查三个作坊生产的黄豆芽中维生素 C **的含量**(mg/100g)**，结果如下，问三个作坊的黄豆芽中** VC **含量有无不同**?

表 8.2 三个作坊的黄豆芽中维生素 C 的含量(mg/100g)

采样日期	甲作坊	乙作坊	丙作坊
a	b=1	b=2	b=3
2 月(a=2)	11.4	5.8	3.5
4 月(a=4)	6.4	8.6	7.5
6 月(a=6)	11.2	7.0	9.8
8 月(a=8)	13.8	10.8	10.4
10 月(a=10)	7.3	8.8	9.3
12 月(a=12)	8.3	6.2	2.5

首先对变量 x **进行秩变换，然后对秩次** rankx **进行配伍组方差分析**:

```
. genrank rankx=x
. anova rankx a b
```

```
                   Number of obs =      18      R-squared     = 0.5769
                   Root MSE      = 4.52769      Adj R-squared = 0.2807
          Source |  Partial SS    df       MS           F     Prob > F
-----------------+----------------------------------------------------
           Model |       279.50     7  39.9285714       1.95    0.1635
                 |
               a |  223.166667     5  44.6333333       2.18    0.1382
               b |  56.3333333     2  28.1666667       1.37    0.2970
                 |
        Residual |       205.00    10       20.50
-----------------+----------------------------------------------------
           Total |       484.50    17       28.50
```

尚不能认为三个作坊的黄豆芽中维生素 C **的含量有何不同。**

例 8.9（ 交叉设计的秩变换检验 ） 用 12 **名高血压病人研究** A，B **两种治疗方案的疗效，试验用交叉设计进行，结果如下，试用秩变换检验进行分析。**

表 8.3 12 名高血压病人用 A，B 两法治疗的血压下降值(mmHg)

阶 段	病人编号(a)											
(stage)	1	2	3	4	5	6	7	8	9	10	11	12
I	B	B	A	B	A	A	A	A	B	B	B	A
	23	10	33	14	24	28	31	8	8	17	26	18
II	A	A	B	A	B	B	B	B	A	A	A	B
	21	11	28	27	20	12	20	13	11	14	26	13

```
. use exam8_9
. list
```

	x	a	method	stage
1.	23	1	2	1
2.	10	2	1	1
3.	33	3	2	1
	.....			
	.....			
24.	13	12	2	2

首先对 x **进行编秩，然后对秩次** xrank **进行方差分析**:

```
. genrank xrank=x
. anova xrank a method stage
```

```
                    Number of obs =      24      R-square     = 0.8352
                    Root MSE      = 4.34693      Adj R-square = 0.6209

      Source |  Partial SS    df       MS           F    Prob > F
-------------+----------------------------------------------------
       Model |  957.541667    13  73.6570513       3.90    0.0188
             |
           a |      872.00    11  79.2727273       4.20    0.0158
      mathod |       64.00     1       64.00       3.39    0.0955
         stg |  12.0416667     1  12.0416667       0.64    0.4432
             |
    Residual |  188.958333    10  18.8958333
-------------+----------------------------------------------------
       Total |     1146.50    23  49.8478261
```

按 $\alpha=0.05$ **水准，尚不能认为用** A，B **两法治疗高血压病人的血压下降值有差别。**

第九章　直线相关与回归分析

直线相关与回归是处理两变量(其中至少有一个是随机变量)间线性依存关系的统计方法。一般是先作散点图(详见第五章)，当确认两变量有线性相关趋势时，才能进一步计算相关系数和回归方程。若两变量呈某种曲线关系，则需用曲线表示两者间的非线性回归关系，详见第十二章。

§9.1　相关分析

相关系数是表达两变量线性相关程度和方向的一个指标，一般用 r 表示，其值在-1～+1之间。r=0 表示两变量无相关；r>0 表示两变量是正相关，即随一个变量的增加，另一个变量随之增加，反之亦然；r<0 表示两变量是负相关，即随一个变量的增加，另一个变量在减少；反之亦然。r 越接近 0，表示关系越不密切，r 越接近+1 或-1 表示关系越密切。Stata 用于计算相关系数的命令为 correlate，格式入下：

correlate 变量 [, 选择项]

这里“变量”可以是两个亦可超过两个。correlate 命令给出的是变量间两两的简单相关系数。选择项有：

```
means         /*  同时输出均数、标准差等统计量；
covariance    /*  不输出相关系数矩阵，而输出协方差矩阵；
wrap          /*  相关系数矩阵打印时不分段。
```

例 9.1　测得某地 10 名 3 岁儿童的体重与体表面积如下，试对该资料进行分析。

体重 x（kg）	11.0	11.8	12.0	12.3	13.1	13.7	14.4	14.9	15.2	16.0
体表面积 $y(\times 10^3 cm^2)$	5.283	5.299	5.358	5.292	5.602	6.014	5.830	6.102	6.075	6.411

先作散点图(见图 9.1)，从散点图可以看出，x 与 y 呈线性趋势。故可进一步作线性相关与回归分析。计算相关系数：

```
. corr y x
(obs=10)
             |        y        x
-------------+------------------
           y |   1.0000
           x |   0.9579   1.0000
```

corr 是 correlate 的缩写，算得相关系数 r=0.9579。

当变量多于两个时，输出的结果是多个变量间的相关系数矩阵(见第十章)。

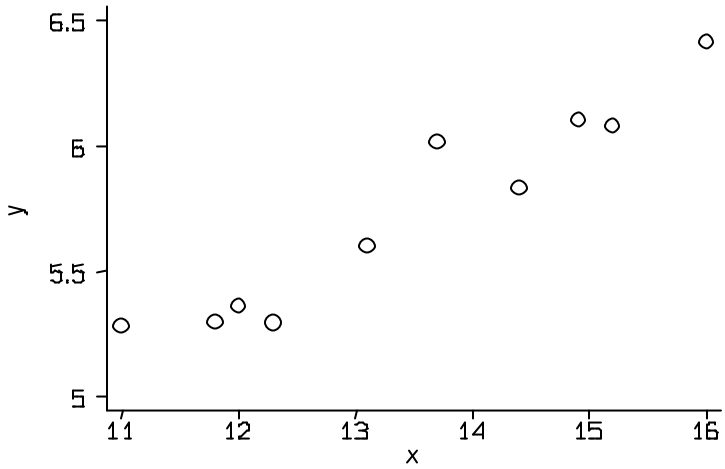


图 9.1　10 名 3 岁儿童的体重与体表面积散点图

相关系数 r **是总体相关系数**ρ**的估计值，由于抽样误差，即使从**$\rho=0$ **的总体中抽样，所得的** r **也常不等于零。因此，判断** x **与** y **间是否有线性关系，还须进行**$\rho=0$ **的假设检验，检验统计量用式**(9.1)**计算。**

$$t=\frac{r}{\sqrt{\frac{1-r^2}{n-2}}},\quad \text{自由度}\nu=n-2 \tag{9.1}$$

Stata **没有给出相关系数的假设检验结果，因此要用户自行计算。本例，**$r=0.9579$**，**$n=10$**，代入式**(9.1):

```
. disp 0.9579/sqrt((1-0.9579^2)/8)
9.4368894
```

这就是 t **值，按自由度** 8 **计算大于等于现有统计量** t **的概率：**

```
. disp tprob(8,9.4368894)
.00001306
```

即 $P<0.0001$**，故拒绝假设**$\rho=0$**，认为** x **与** y **存在线性关系。简单相关系数的假设检验与直线回归中自变量的假设检验结果是等价的，故也可用回归系数的假设检验来代替相关系数的假设检验。**

以上所计算的相关系数又称作积差相关系数(coefficient of product-moment correlation)**，适合** x **与** y **均为连续变量且散点图显示出线性趋势的情形，若散点图不呈现线性趋势，或原始数据是等级资料的，这时应计算等级相关**(rank correlation)**系数，常用的等级相关有** Spearman **等级相关和** Kandel**等级相关(详见第八章)。**

§9.2　回归分析

由散点图和相关系数的计算确认 x **与** y **存在线性关系后，可进一步研究二变量间的线性依存关系，即求出直线回归方程。**Stata **中用于估计直线回归方程的命令为：**

regress **因变量 自变量** [,**选择项**]

其中，选择项为：

```
level(#)      /* 选定检验水准
beta          /* 指定输出标准回归系数
noconstant    /* 指定方程中不包含常数项。默认为包含常数项
noheader      /* 不打印方差分析结果及复相关系数等
```

例 9.2　对例 9.1 资料，求回归方程。

```
. reg y x

  Source |       SS       df       MS              Number of obs =      10
---------+------------------------------           F(  1,     8) =   89.01
   Model |   1.4165746     1   1.4165746           Prob > F      =  0.0000
Residual |  .127317917     8   .01591474           R-squared     =  0.9175
---------+------------------------------           Adj R-squared =  0.9072
   Total |  1.54389252     9  .171543614           Root MSE      =  .12615

------------------------------------------------------------------------------
       y |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
       x |   .2384983   .0252793      9.435   0.000       .1802041    .2967925
   _cons |   2.521183    .342088      7.370   0.000       1.732327     3.31004
------------------------------------------------------------------------------
```

根据估计结果，获得了回归方程式：

$$\hat{y}=2.521183+0.2385x$$

其中，_cons **表示常数项。**

结果中还给出了对方程检验的方差分析的结果。F=89.01**，相应的概率为** P=0.0000**。**R^2 **称为复相关系数，又称决定系数。在这里表示回归的** SS **占总** SS **的比重，即：**

$$R^2=\frac{SS_{回归}}{SS_{总}}=1-\frac{SS_{误差}}{SS_{总}} \tag{9.2}$$

本例 R^2**＝**1.4165746/1.52389252=0.9175**。在直线回归中，**R^2 **实际上是相关系数** r **的平方。调整** R^2 **表示**:

$$R^2_{arj}=1-\frac{MS_{误差}}{MS_{总}} \tag{9.3}$$

本例校正 R^2**＝**0.01591474/0.171543614=0.9072**。效果较满意。调整** R^2 **又称校正** R^2**。**

结果中的 Root MSE **表示误差均方，又称剩余标准差，等于：** $s_{剩}=\sqrt{MS_{误差}}=\sqrt{0.1591474}$ = 0.12615**。结果中还给出了各系数与** 0 **比较的** t **检验之结果，和各系数的** 95**%可信区间。在直线回归中，自变量的显著性与方程的显著性是一样的。事实上，** $\sqrt{F}=\sqrt{89.01}=9.4345=t$**。**

注意到 x **的回归系数的假设检验结果与对应的相关系数的假设检验之结果是等价的。**

§9.3 估计与预测

求出回归方程后，可立即用该回归方程进行回代预测，并求出预测值的标准误及绘出 95%可信区间曲线。Stata 给出了 y 的估计值，残差，标准残差，残差的标准误等。相应的命令为：

predict 新变量 [, 选择项]

这里，选择项有：

```
cooksd       /* 计算 cook 的检验统计量 D
residuals    /* 计算残差
rstandard    /* 计算标准化残差
rstudent     /* 计算 student 残差
stdr         /* 计算残差的标准误
stdp         /* 计算估计值 y 的标准误
stdf         /* 估计预测值 y 的标准差
```

例 9.3 根据例 9.2 所得方程，计算 y 的估计值：

```
. pred yhat
```

执行该命令后，Stata 将产生一个变量 yhat，并将根据回归方程估计的 $\hat{y}$ 值写入该变量中。

```
. l
          x          y       yhat
 1.      11      5.283   5.144664
 2.    11.8      5.299   5.335463
 3.      12      5.358   5.383162
 4.    12.3      5.292   5.454712
 5.    13.1      5.602   5.645511
 6.    13.7      6.014    5.78861
 7.    14.4       5.83   5.955558
 8.    14.9      6.102   6.074807
 9.    15.2      6.075   6.146357
10.      16      6.411   6.337155
```

根据 y,yhat 及 x 作回归线图。

```
. gra y yhat x , c(.l) s(0i) xlab(11,12,13,14,15,16,17) ylab
```

见图 9.2。进一步计算估计值的 95%可信区间：

```
. pred seyhat, stdp                    /* 计算预测值的标准误
```

该命令将各 y 估计值的标准误写入变量 seyhat。

```
. gen l1=yhat-invt(8,0.95)*seyhat      /* 计算 95%可信区间下界,invt(8,0.95)是自由度
                                          为 9 的下侧累计概率为 0.95 的 t 分布之分位数
. gen l2=yhat+invt(8,0.95)*seyhat      /* 计算 95%可信区间上界
```

计算估计值的 95%容许区间：

```
. pred sey, stef                       /* 计算预测值的标准差
```

该命令将各 y 估计值的标准差写入变量 sey。

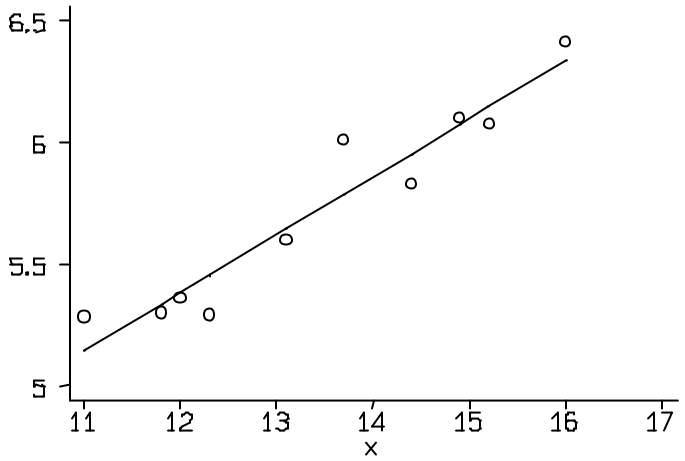


图 9.2　10 名 3 岁儿童的体重与体表面积线性回归

```
. gen l3=yhat-invt(8,0.95)*seyhat          /*计算 95%容许区间下界
. gen l4=yhat+invt(8,0.95)*seyhat          /*计算 95%容许区间上界
```

结果如下：

```
. list y yh l1 l2 l3 l4
        y       yhat         l1         l2         l3         l4
 1.  5.283   5.144664    4.97527   5.314059   4.808029     5.4813
 2.  5.299   5.335463   5.202788   5.468138   5.015726     5.6552
 3.  5.358   5.383162   5.258626   5.507699   5.066716   5.699609
 4.  5.292   5.454712   5.341225   5.568199   5.142449   5.766975
 5.  5.602   5.645511   5.551406   5.739615   5.339758   5.951263
 6.  6.014    5.78861   5.695375   5.881844   5.483123   6.094096
 7.   5.83   5.955558   5.847879   6.063237   5.645359   6.265758
 8.  6.102   6.074807   5.949481   6.200133   5.758049   6.391565
 9.  6.075   6.146357   6.008556   6.284158   5.824459   6.468255
10.  6.411   6.337155   6.161846   6.512465   5.997505   6.676806
```

绘制估计值的 95%可信区间及 95%容许区间曲线：

```
. gra y yhat l1 l2 l3 l4 x, c(.lssss) s(Oiiiii) t1(" ")
                          xlab(11,12,13,14,15,16,17) ylab(4.5,5,5.5,6,6.5,6)
```

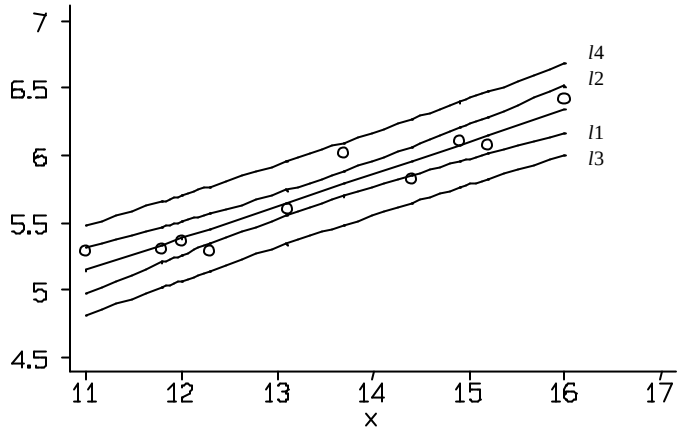


图 9.3　10 名 3 岁儿童的体重与体表面积线性回归及容许区间和可信区间估计

图 9.3 **中有** 5 **条线，中间一条是回归线；最上面一条** *l*4 **和最下面一条** *l*3 **是** *y* **的容许区间，而另外两条** *l*1,*l*2 **是** *y* **的可信区间。**

如 x=12kg **时，观察值为** $5.358\times10^3\text{cm}^2$,**相应的估计值** $5.383\times10^3\text{cm}^2$**，其** 95**%可信区间为**$(5.259,\ 5.508)\times10^3\text{cm}^2$,95**%的容许区间为**$(5.067,\ 5.700)\times10^3\text{cm}^2$**。意即：对所有体重＝**12kg **的** 3 **岁男童，估计其平均体表面积为** $5.383\times10^3\text{cm}^2$**，该均数的** 95**%可信区间为**$(5.259,\ 5.508)\times10^3\text{cm}^2$**；估计约有** 95**%的体重为** 12kg **的** 3 **岁男童，其体表面积在** $5.067\times10^3\text{cm}^2$**～**$5.700\times10^3\text{cm}^2$ **之间。**

§9.4　过定点的直线回归

医学研究中应用直线回归常遇到这样一个问题，即所估计的直线除了要根据观察值进行最佳拟合外，还要求所拟合的直线通过某定点(y_0,x_0)**。这些情况在应用光电比色，荧光分析，火焰光度测定以及同位素测定等实验方法来绘制标准直线时经常遇到。**

要使直线通过原点(0,0)**，只需在回归命令中增加选择项** noconstant **即可。而要直线通过任意一点，只需一点小小的技巧。**

例 9.4(**过原点的直线回归**) **下面的资料为进行光电比色分析时，所得总维生素** C **浓度**(μg/ml)**与光密度之间的相应关系，目的是要建立标准直线，理论上此直线要过**(0,0)**点。试求回归方程。**

总维生素 C **浓度**	0	2	4	6	8	10	12
光密度	0	0.051	0.081	0.109	0.150	0.186	0.244

```
. reg y x ,nocons

  Source |       SS       df       MS                  Number of obs =       7
```

```
---------+------------------------------           F(  1,     6) = 3288.00
   Model |  .128081281     1  .128081281           Prob > F      =  0.0000
Residual |  .000233725     6  .000038954           R-squared     =  0.9982
---------+------------------------------           Adj R-squared =  0.9979
   Total |  .128315006     7  .018330715           Root MSE      =  .00624

------------------------------------------------------------------------------
       y |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
       x |   .0187582   .0003271     57.341   0.000       .0179578    .0195587
------------------------------------------------------------------------------
```

由于增加了选择项 nocons，结果中未给出常数项的系数。回归方程为：

$$\hat{y} = 0.0187582x$$

该直线在 $x=0$ 时，$\hat{y}=0$，故直线经过$(0,0)$点。

```
. gra y yhat x , c(.l0 s9o.) xlab(0,2,4,6,8,10,12) ylab(0,0.05,0.1,0.15,0.2,0.25)
```

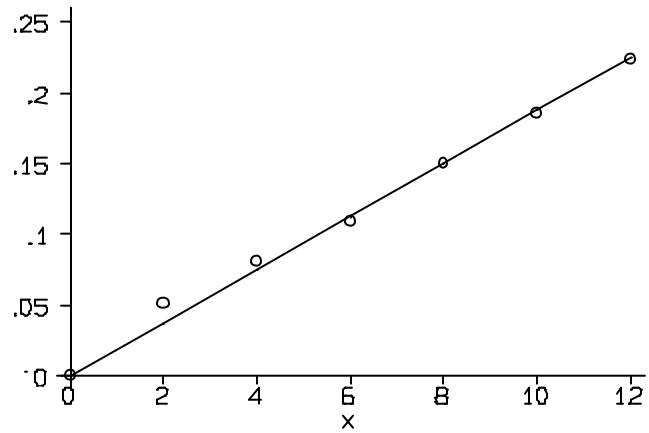


图 9.4　　过原点的直线回归

例 9.5(过任意定点的直线回归)　以例 9.1 资料来说明过任意定点的直线的回归。假设该直线需通过点(5.4，12)，则步骤入下：

1. 所有 y 减去 5.4，记为 $y1$；
2. 所有 x 减去 12，记为 $x1$；
3. 则要求直线需通过点(5.4，12)，实际上是要求根据 $y1$,$x1$ 建立的回归方程经过(0,0)，故用例 9.4 方法建立 $y1$ 与 $x1$ 的回归方程，并使直线通过(0,0)；
4. 将 $y1$,$x1$ 还原到 y，x，所得方程即为所求。

命令如下：

```
. use d:\mydata\ex9-1
. gen y1=y-5.4
. gen x1=x-12
. reg y1 x1 , nocons

  Source |       SS       df       MS              Number of obs =      10
---------+------------------------------           F(  1,     9) =  173.32
```

```
   Model |  2.48170291     1  2.48170291             Prob > F      =  0.0000
Residual |  .128864898     9  .014318322             R-squared     =  0.9506
---------+------------------------------             Adj R-squared =  0.9452
   Total |  2.61056781    10  .261056781             Root MSE      =  .11966

------------------------------------------------------------------------------
      y1 |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
      x1 |   .2331858   .0177122     13.165   0.000        .193118    .2732537
------------------------------------------------------------------------------
```

得回归方程：

$$\hat{y}\,1=0.2331858x1$$

将 y1=y-5.4,x1=x-12 代入上式，得：

$$\hat{y}=2.60177+0.2332x$$

此即为所求。读者不妨验算一下。

```
. gen yhat=2.60177+0.2332x
. gra y yaht x , c(.l) s(0.) xlab(11,12,13,14,15,16) ylab(5,5.4,5.5,6,6.5,) xline(12)
  yline(5.4)
```

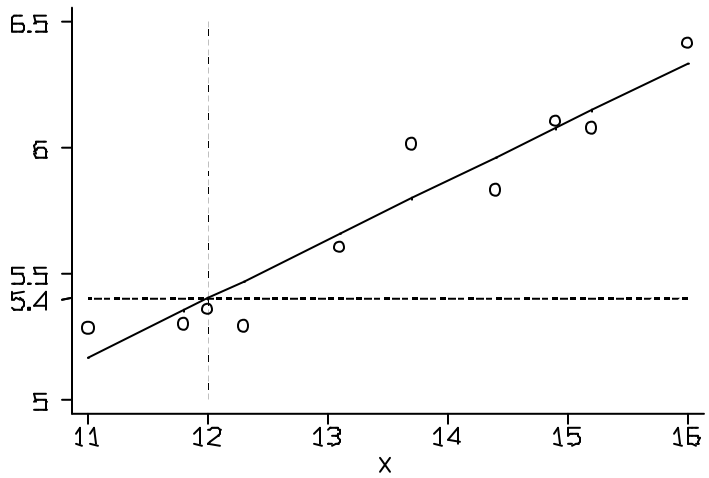


图 9.5　　过定点的直线回归