Stata 软件基本操作和数据分析入门

赵耐青 张文彤

第一讲 Stata 操作入门

第一节 概况

Stata 最初由美国计算机资源中心（Computer Resource Center）研制，现在为 Stata 公司的产品，其最新版本为 7.0 版。它操作灵活、简单、易学易用，是一个非常有特色的统计分析软件，现在已越来越受到人们的重视和欢迎，并且和 SAS、SPSS 一起，被称为新的三大权威统计软件。

Stata 最为突出的特点是短小精悍、功能强大，其最新的 7.0 版整个系统只有 10M 左右，但已经包含了全部的统计分析、数据管理和绘图等功能，尤其是他的统计分析功能极为全面，比起 1G 以上大小的 SAS 系统也毫不逊色。另外，由于 Stata 在分析时是将数据全部读入内存，在计算全部完成后才和磁盘交换数据，因此运算速度极快。

由于 Stata 的用户群始终定位于专业统计分析人员，因此他的操作方式也别具一格，在 Windows 席卷天下的时代，他一直坚持使用命令行 / 程序操作方式，拒不推出菜单操作系统。但是，Stata 的命令语句极为简洁明快，而且在统计分析命令的设置上又非常有条理，它将相同类型的统计模型均归在同一个命令族下，而不同命令族又可以使用相同功能的选项，这使得用户学习时极易上手。更为令人叹服的是，Stata 语句在简洁的同时又拥有着极高的灵活性，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。

除了操作方式简洁外，Stata 的用户接口在其他方面也做得非常简洁，数据格式简单，分析结果输出简洁明快，易于阅读，这一切都使得 Stata 成为非常适合于进行统计教学的统计软件。

Stata 的另一个特点是他的许多高级统计模块均是编程人员用其宏语言写成的程序文件（ADO 文件），这些文件可以自行修改、添加和下载。用户可随时到 Stata 网站寻找并下载最新的升级文件。事实上，Stata 的这一特点使得他始终处于统计分析方法发展的最前沿，用户几乎总是能很快找到最新统计算法的 Stata 程序版本，而这也使得 Stata 自身成了几大统计软件中升级最多、最频繁的一个。

由于以上特点，Stata 已经在科研、教育领域得到了广泛应用，WHO 的研究人员现在也把 Stata 作为主要的统计分析工作软件。

第二节　Stata 操作入门

一、Stata 的界面

图 1 即为 Stata 7.0 启动后的界面，除了 Windows 版本的软件都有的菜单栏、工具栏，状态栏等外，Stata 的界面主要是由四个窗口构成，分述如下：

1．结果窗口：位于界面右上部，软件运行中的所有信息，如所执行的命令、执行结果和出错信息等均在这里列出。窗口中会使用不同的颜色区分不同的文本，如白色表示命令，红色表示错误信息。

2．命令窗口：位于结果窗口下方，相当于 DOS 软件中的命令行，此处用于键入需要执行的命令，回车后即开始执行，相应的结果则会在结果窗口中显示出来。

3．命令回顾窗口：即 review 窗口，位于界面左上方，所有执行过的命令会依次在该窗口中列出，单击后命令即被自动拷贝到命令窗口中；如果需要重复执行，用鼠标双击相应的命令即可。

4．变量名窗口：位于界面左下方，列出当前数据及中的所有变量名称，。

除以上四个默认打开的窗口外，在 Stata 中还有数据编辑窗口、程序文件编辑窗口、帮助窗口、绘图窗口、Log 窗口等，如果需要使用，可以用 Window 或 Help 菜单将其打开。

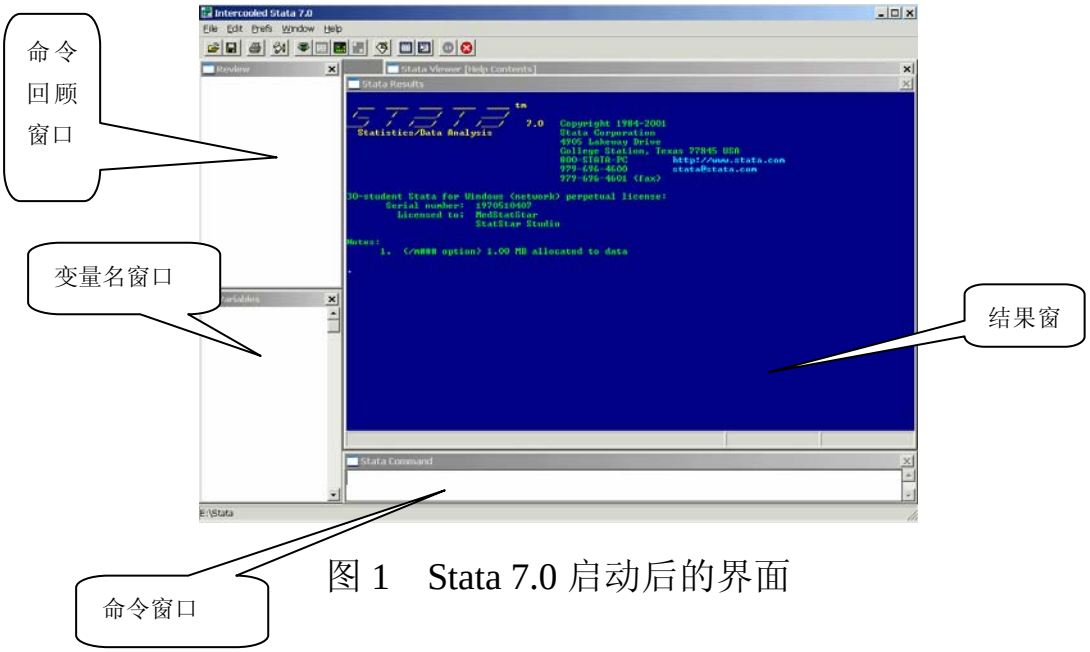


图 1　Stata 7.0 启动后的界面

二、数据的录入与储存

Stata 为用户提供了简捷，但是非常完善的数据接口，熟悉它的用法是使用 Stata 的第一步，在 Stata 中读入数据可以有三种方式：直接从键盘输入、打开已有数据文件和拷贝、粘贴方式交互数据。

1)从键盘输入数据

在 Stata 中可以使用命令行方式直接建立数据集，首先使用 input 命令制定相应的变量名称，然后一次录入数据，最后使用 end 语句表明数据录入结束。

例 1　在某实验中得到如下数据，请在 Stata 中建立数据集。

观测数据

X	1	3	5	7	9
Y	2	4	6	8	10

解：此处需要建立两个变量 X、Y，分别录入相应数值，Stata 中的操作如下，其中划线部分为操作者输入部分。

```
. drop _all

. input x y
        x     y
  1. 1     2
  2. 3     4
  3. 5     6
  4. 7     8
  5. 9    10
  6. end
```

2)用 stata 的数据编辑工具

①进入数据编辑器

进入 stata 界面，在命令栏键入 edit 或在 stata 的 window 下拉菜单中单击 data editor 或点击编辑图标  (注意：  是浏览图标，点击后只能浏览，不能编辑)即可进入 stata 数据编辑器。（stata 界面如下图 2）

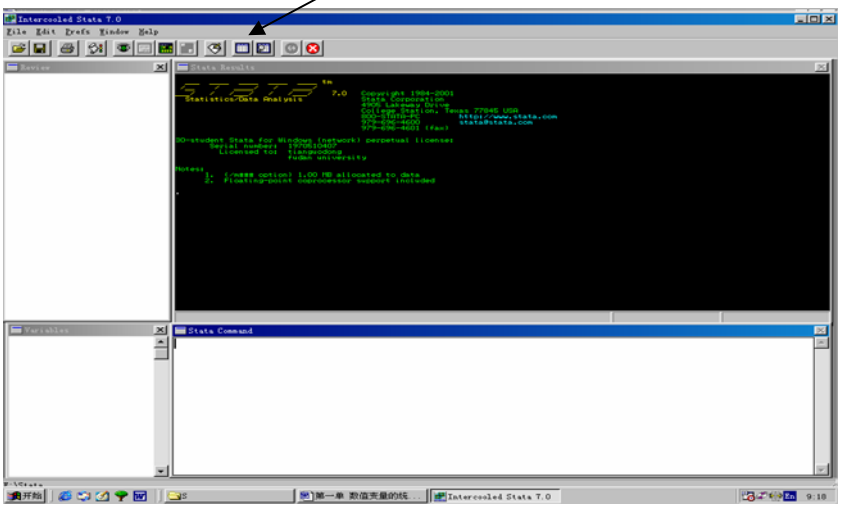


图 2

②数据编辑

stata 数据编辑器界面：此时进入了数据全屏幕编辑状态。

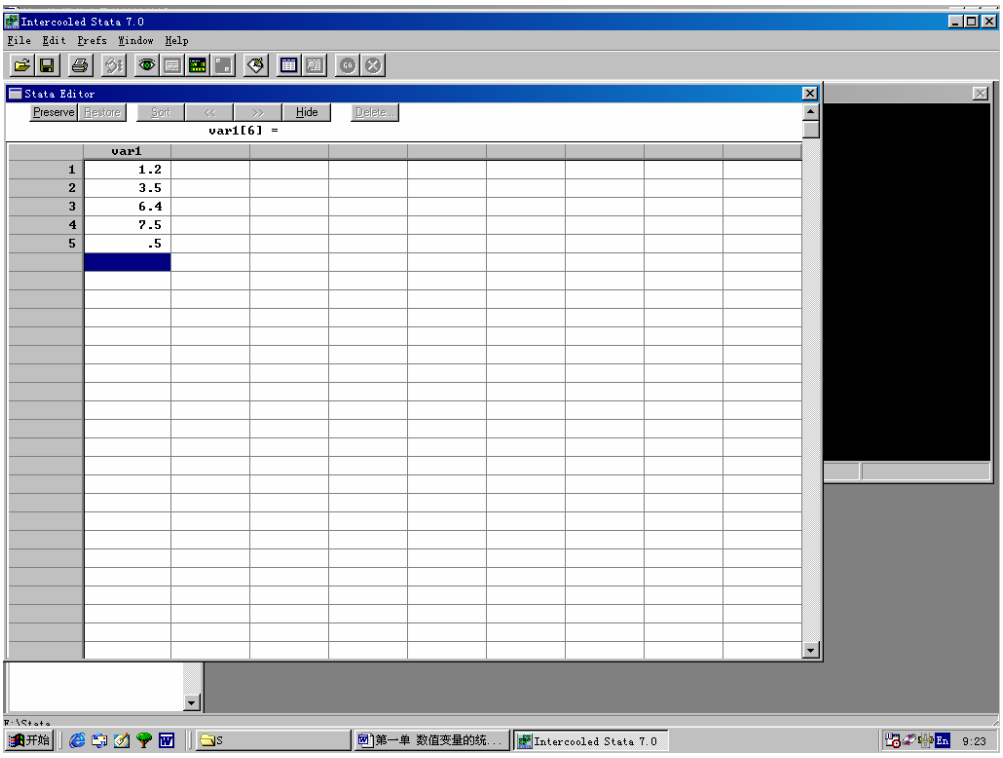


图 3

在第一列输入数据后，Stata 第一列自动命名为 var1；在第二列输入数据后，第二列自动命名为 var2……依次类推。在输入数据后，双击纵格顶端的变量名栏(如：Var1 或 Var2 处)，可以更改变量名，并可以在 label 栏中注释变量名的含义,点击 OK 确认(如图 4 所示)。仍沿用上例，双击观察值所在列顶端的变量名栏，更改变量名为 x，并在 label 栏中注明 “7 岁男童身高（cm）”。

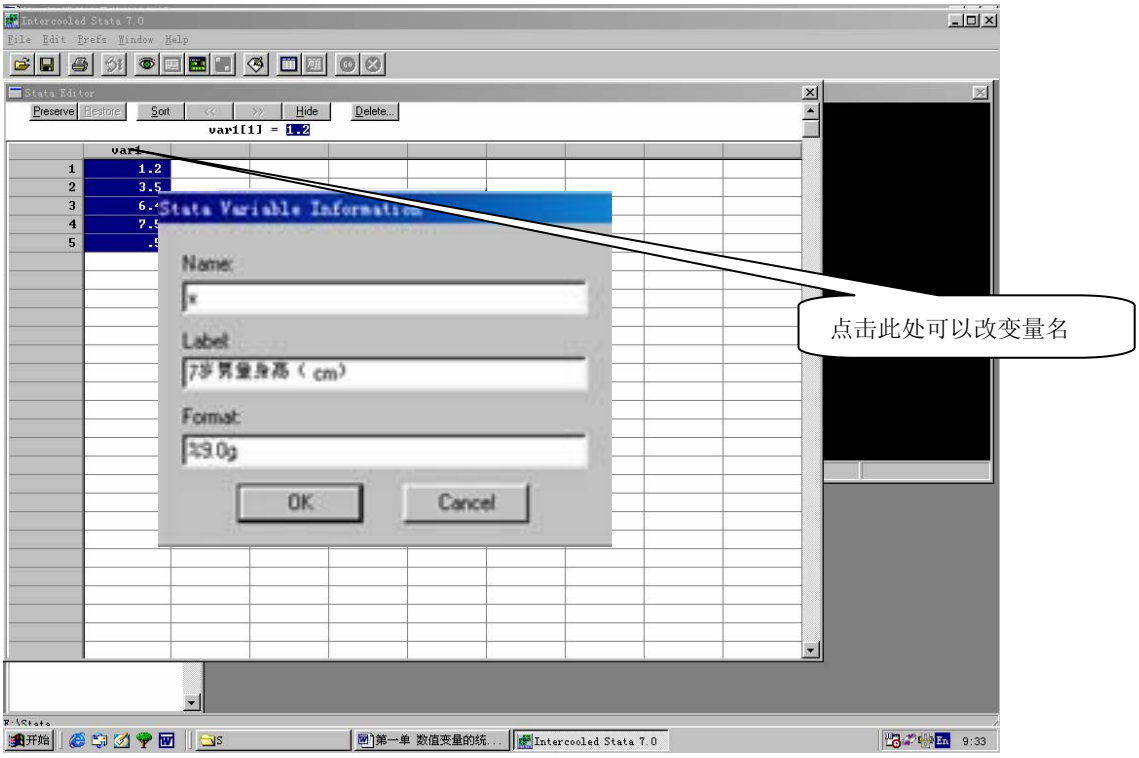


图 4

数据输入完毕后，单击 preserve 键确认所输数据，按关闭键 ✕ 即可退出编辑器。

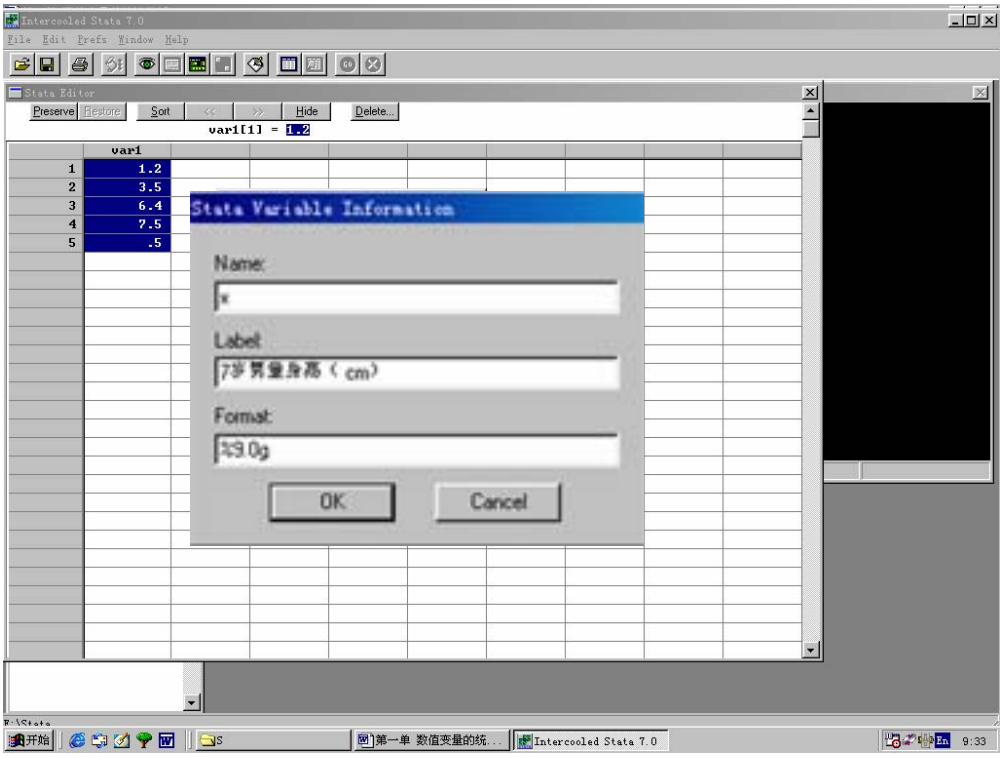


图 5

数据输入完毕后，单击 preserve 键确认所输数据，按关闭键  即可退出编辑器。

3)拷贝、粘贴方式交互数据

Stata 的数据编辑窗口是一个简单的电子表格，可以使用拷贝、粘贴方式直接和 EXCEL 等软件交互数据，在数据量不大时，这种方式操作极为方便。

例 2　现在 EXCEL 中已录入了三个变量，共五条记录，格式见下图，请将数据读入 Stata。

解：首先将 EXCEL 中的 A1～C6 全部 18 个单元格选中，选择菜单编辑➔复制，将数据拷贝到剪贴板上；然后切换到 Stata，选择菜单 Window➔Data Editor，打开数据编辑窗口；再选择 Edit➔Paste，相应的数据就会被直接粘贴如数据编辑窗口中，并且变量名、记录数、变量格式等均会被自动正确设置，见图 6 和图 7。

	A	B	C
1	x	y	z
2	1	2	q
3	3	4	wqw
4	5	6	e
5	7	8	dfw
6	9	10	f

图 6　在 EXCEL 中的数据格式

	x	y	z
1	1	2	q
2	3	4	wqw
3	5	6	e
4	7	8	dfw
5	9	10	f

图 7　粘贴入 Stata 后的数据格式

4)、打开已有的数据文件

Stata 能够直接打开的数据文件只能是自身专用格式或者以符号分隔的纯文本格式，后者第一行可以是变量名，分述如下：

1．点击图标  ，然后选择路径和文件名，可以打开 Stata 专用格式的数据文件，并且扩展名为.dta。

2.打开 Dta 数据文件：该格式文件是 Stata 的专用格式数据文件，也使用 use 命令即可打开，例如要打开数据文件“C:\data1.dta”，则命令为：

. use c:\data1

即扩展名可以省略，如果 Stata 中已经修改或者建立了数据集，则需要使用 clear 选项清除原有数据，命令为：

. use c:\data1 , clear

3．读入文本格式数据：需要使用 insheet 命令实现，例如需要读入已建立好的文本格式数据文件“C:\data1.txt”，则命令为：

. insheet using c:\data1.txt

该命令会自动识别第一行是否为变量名，以及变量列间的分隔符是 tab、逗号还是其他字符。如果 Stata 中已经修改或者建立了数据集，则需要使用 clear 选项清除原有数据，命令为：

. insheet using c:\data1.txt , clear

5)数据文件的保存

为了方便以后重复使用，输入 Stata 的数据应存盘。Stata 实际上只能将数据存为自身专用的数据格式或者纯文本格式，分述如下：

1．点击图标  ，然后选择路径和文件名，点击保存。

2.存为 dta 格式：可以直接使用文件菜单，也可以使用 save 命令操作，如欲将上面建立的数据文件存入“C:\”中，文件名为 Data1.dta，则命令为：

. save c:\data1

file c:\data1.dta saved

该指令将在 C 盘根目录建立一个名为“data1.dta”的 Stata 数据文件，后缀 dta 可以在命令中省略，会被自动添加。该文件只能在 Stata 中用 use 命令打开。如所指定的文件已经存在，则该命令将给出如下信息：file c:\data1.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 c:\data1.dta , replace

file c:\data1.dta saved

2．存为文本格式：需要使用 outsheet 命令实现，该命令的基本格式如下。

outsheet [变量名列表] using 文件名 [, nonames replace]

其中变量名列表如果省略，则将全部变量存入指定文件。

如欲将上面建立的数据文件存入文本文件“C:\data1.txt”中，则命令为：

. outsheet using c:\data1.txt

此时建立的文件 data1.txt 第一行为变量名，第 2～6 行为变量值。变量列间用 Tab 键分隔。如果不希望在第一行存储变量名，则可以使用 nonames 选项。如果文件已经存在，则需要使用 replace 选项。

第二讲 统计描述入门

一、调查某市 1998 年 110 名 19 岁男性青年的身高（cm）资料如下，计算均数、标准差、中位数、百分位数和频数表。

173.1	167.8	173.9	176.9	173.8	171.5	175.1	175.2	176.7	174.5
169.2	174.7	185.4	175.8	173.5	175.9	175.9	173.2	174.8	177.2
171.9	*166.0*	177.3	175.2	179.8	175.7	180.8	171.4	178.9	172.6
166.9	170.8	168.7	175.0	183.7	171.6	172.9	173.6	177.7	172.4
181.2	178.1	173.3	177.5	173.0	174.3	174.5	172.5	171.3	174.0
177.9	170.7	175.2	178.5	177.6	183.3	173.1	170.9	180.5	176.8
179.6	180.6	176.6	174.3	168.7	175.2	179.5	172.5	173.0	174.2
169.5	177.0	183.6	170.3	178.8	181.1	182.9	177.8	164.1	169.1
176.3	169.4	171.1	172.9	177.0	179.8	178.2	174.4	169.2	176.4
178.3	165.0	175.8	181.0	177.6	177.4	178.7	175.1	181.8	171.3
174.8	181.7	177.3	178.5	179.3	177.0	175.8	181.8	177.5	180.2

Stata 数据结构

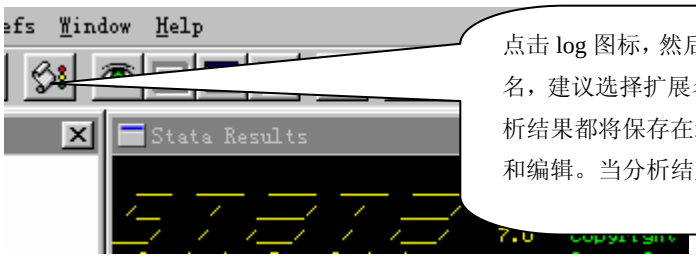
	x
1	173.1
2	169.2
3	171.9
4	166.9
5	181.2
6	177.9
7	179.6
8	169.5
9	176.3
10	178.3
11	174.8
12	167.8
13	174.7
14	166
15	170.8
16	178.1
17	170.7
18	180.6
19	177
20	169.4
21	165
22	181.7
23	173.9
24	185.4
25	177.3

26	168.7
27	173.3
28	175.2
29	176.6
30	183.6
31	171.1
32	175.8
33	177.3
34	176.9
35	175.8
36	175.2
37	175
38	177.5
39	178.5
40	174.3
41	170.3
42	172.9
43	181
44	178.5
45	173.8
46	173.5
47	179.8
48	183.7
49	173
50	177.6
51	168.7
52	178.8
53	177
54	177.6
55	179.3
56	171.5
57	175.9
58	175.7
59	171.6
60	174.3
61	183.3
62	175.2
63	181.1
64	179.8
65	177.4
66	177
67	175.1
68	175.9

69	180.8
70	172.9
71	174.5
72	173.1
73	179.5
74	182.9
75	178.2
76	178.7
77	175.8
78	175.2
79	173.2
80	171.4
81	173.6
82	172.5
83	170.9
84	172.5
85	177.8
86	174.4
87	175.1
88	181.8
89	176.7
90	174.8
91	178.9
92	177.7
93	171.3
94	180.5
95	173
96	164.1
97	169.2
98	181.8
99	177.5
100	174.5
101	177.2
102	172.6
103	172.4
104	174
105	176.8
106	174.2
107	169.1
108	176.4
109	171.3
110	180.2

(读者可以把数据直接粘贴到 Stata 的 Edit 窗口)

在介绍统计分析命令之前，先介绍打开一个保存统计分析结果的文件操作：



点击 log 图标，然后选择路径和输入保存结果的文件名，建议选择扩展名为 log 的文件，这样以后统计分析结果都将保存在这个文件中并且可以用 word 打开和编辑。当分析结束时，仍点击该图标，关闭文件。

计算样本的均数、标准差、最大值和最小值

命令 1：su 变量名　　（可以多个变量：即：su 变量名 1 变量名 2 … 变量名 m）

命令 2：su 变量名,d　　（可以多个变量:即：su 变量名 1 变量名 2 … 变量名 m,d）

本例命令 su x

变量　　样本量　　均数　　标准差　　最小值　　最大值

Variable	Obs	Mean	Std. Dev.	Min	Max
x	110	175.3655	4.222297	164.1	185.4

本例命令. su x,d

x

Percentiles		Smallest		
1%	165	164.1		
5%	168.7	165		
10%	169.45	166	Obs	110
25%	172.9	166.9	Sum of Wgt.	110
50%	175.2		Mean	175.3655
		Largest	Std. Dev.	4.222297
75%	178.1	183.3		
90%	180.9	183.6	Variance	17.82779
95%	181.8	183.7	Skewness	-.1756947
99%	183.7	185.4	Kurtosis	2.895843

结果说明

Smallest	最小值	Obs	110	样本量
164.1	第 1 最小值	Sum of Wgt.	110	加权和(即每个记录的权是 1)
165	第 2 最小值			
166	第 3 最小值	Mean	175.3655	均数
166.9	第 4 最小值	Std. Dev.	4.222297	标准差
Largest	最大值	Variance	17.82779	方差
183.3	第 4 最大值	Skewness	-.1756947	偏度系数
183.6	第 3 最大值	Kurtosis	2.895843	峰度系数
183.7	第 2 最大值			
185.4	第 1 最大值			

	Percentiles	百分位数
1%	165	=P_1
5%	168.7	=P_5
10%	169.45	=P_{10}
25%	172.9	=P_{25}
50%	175.2	=P_{50}
75%	178.1	=P_{75}
90%	180.9	=P_{90}
95%	181.8	=P_{95}
99%	183.7	=P_{99}

百分位数 P_X 表示样本中 X%的数据小于等 P_X 并且(100－X)%的数据大于等于 P_X。
特别：P_{50} 就是中位数，表示一半的数据小于等于它，另一半的数据大于等于它。本例：P_{50}=175.2
样本量 obs=110，因此有 55 个数据小于等于 175.2，另有 55 个数据大于等于 175.2

计算百分位数还可以用专用命令 centile。
centile 变量名(可以多个变量)，centile(要计算的百分位数) 例如计算 $P_{2.5}$，$P_{97.5}$ 等
centile 变量名,centile(2.5 97.5)
本例计算 $P_{2.5}$，$P_{97.5}$，P_{50}，P_{25}，P_{75}。
本例命令. centile x,centile(2.5 25 50 75 97.5)

Variable	Obs	Percentile	Centile	-- Binom. Interp. -- [95% Conf.	Interval]
x	110	2.5	165.775	164.1	168.7*
		25	172.825	171.3314	173.6267
		50	175.2	174.5	176.6789
		75	178.125	177.3	179.4371
		97.5	183.6225	181.8	185.4*

* Lower (upper) confidence limit held at minimum (maximum) of sample

结果说明

Percentile	Centile	百分位数
2.5	165.775	=$P_{2.5}$
25	172.825	=P_{25}
50	175.2	=P_{50}(中位数)
75	178.125	=P_{75}
97.5	183.6225	=$P_{97.5}$

制作频数表，组距为 2，从 164 开始，
gen f=int((x-164)/2)*2+164　　其中 int( )表示取整数
tab f　频数汇总和频率计算

频数　　频率　　累积频率

f	Freq.	Percent	Cum.
164	2	1.82	1.82
166	3	2.73	4.55
168	7	6.36	10.91
170	11	10.00	20.91
172	16	14.55	35.45
174	23	20.91	56.36
176	20	18.18	74.55
178	13	11.82	86.36
180	10	9.09	95.45

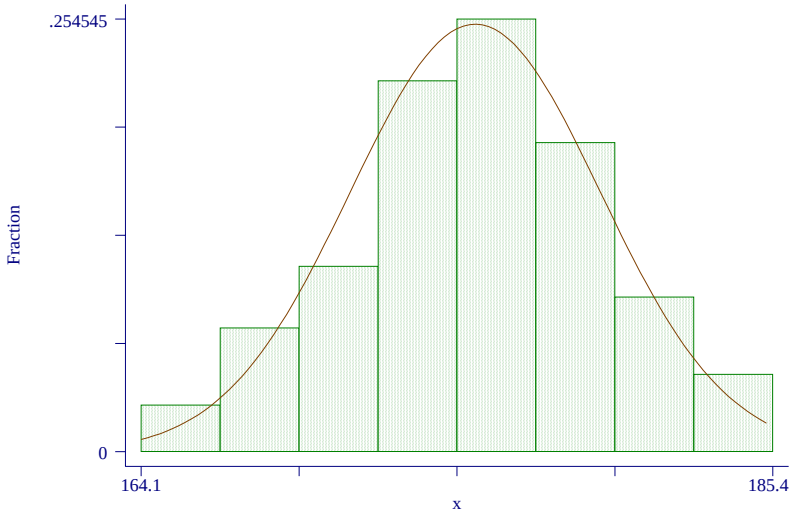
182	4	3.64	99.09
184	1	0.91	100.00
Total	110	100.00	

作频数图

命令 graph 变量,bin(#) norm

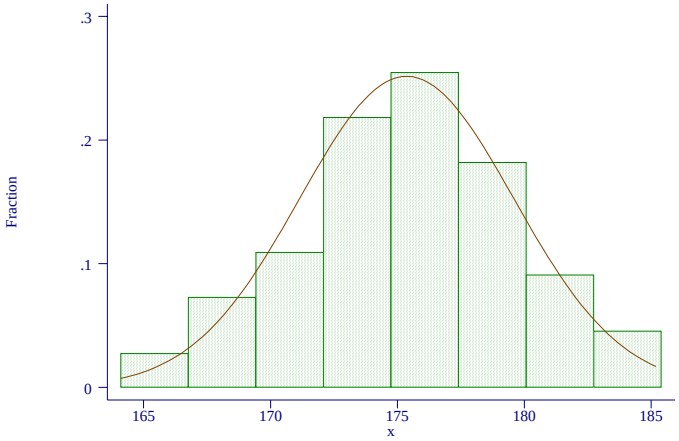
其中#表示频数图的组数;norm 表示画一条相应的正态曲线(可以不要)

本例命令为 graph x,bin(8) norm



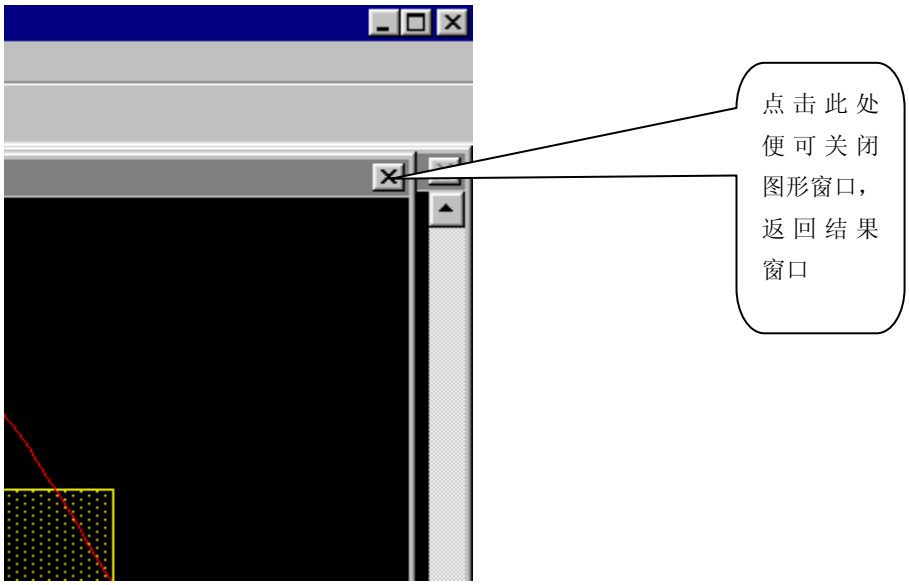
为了使坐标更清楚地在图上显示，可以输入下列命令

graph x,bin(8) xlabel norm ylabel



图形可以从 Stata 中复制到 word 中来，操作如下：

然后到 Word 中粘贴和编辑，便可以得到所需要的图形。



计算几何均数可以用 means 变量名(可以多个变量：即:means 变量 1 …变量 m)

```
means x
```

Variable	Type	Obs	Mean	[95% Conf.	Interval]
x	Arithmetic	110	175.3655	174.5676	176.1634
	Geometric	110	175.3149	174.5168	176.1166
	Harmonic	110	175.2642	174.4657	176.07

Arithmetic(算术均数) Geometric(几何均数) 调和均数(Harmonic)

作 Pie 图描述构成比：每一类的频数用一个变量表示，命令：

graph 各类频数变量名,pie

例：下列有 2 个地区的血型频数分布数据，请用 Pie 描述：

	频数			
地区	A	B	O	AB
第 1 地区 area=1	100	120	240	75

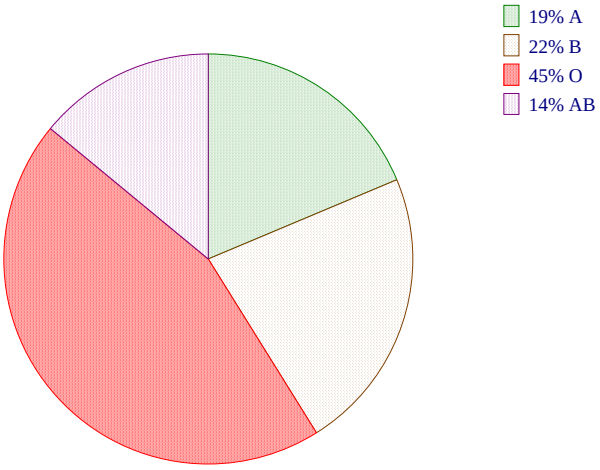
第 2 地区 area=2	80	70	200	50

Stata 数据格式

	a	b	o	ab	area
1	100	120	240	75	1
2	80	70	200	50	2

第 1 地区血型构成比的 Pie 图的命令和图

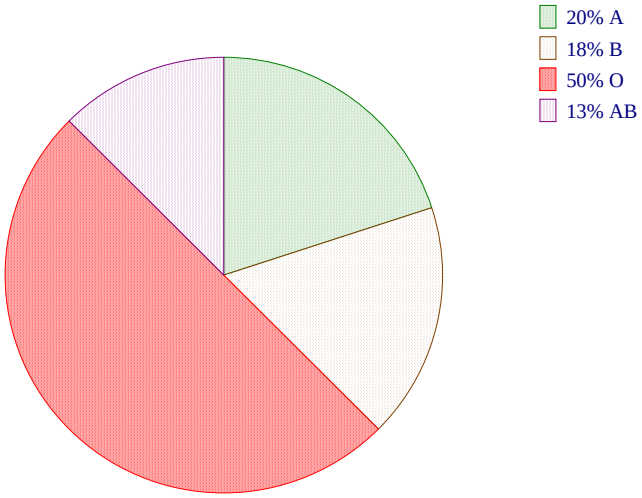
```
graph a b o ab  if area==1,pie
```



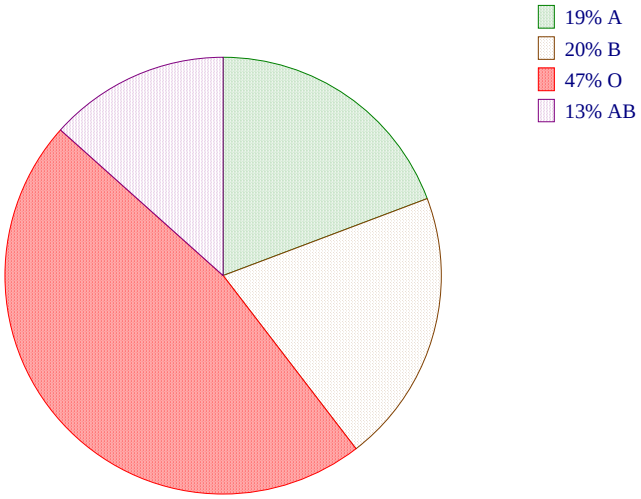
注意逻辑表达式中 if area==1 是两个等号。

第 2 地区血型构成比的 Pie 图的命令和图

```
graph a b o ab  if area==2,pie
```



两个地区合并后的血型构成比的 Pie 图的命令和图



正态性检验. sktest 变量名 1 变量名 2 … 变量名 m
在上例中的 110 名 19 岁男性青年的身高资料正态性检验如下：
sktest x

Skewness/Kurtosis tests for Normality

Variable	Pr(Skewness)	Pr(Kurtosis)	joint adj chi2(2)	joint Prob>chi2
x	0.398	0.451	1.31	0.5198

无效假设 H_0：资料服从正态分布
备选假设 H_1：资料不服从正态分布
设α=0.05（样本比较大时，α取 0.05，样本很小时，α取 0.1）

Prob>z	P 值
.5198	=P 值>0.05

因此可以认为资料近似服从正态分布。

计量资料统计描述的主要策略。
若资料近似正态分布，则用均数±标准差描述
若资料偏态分布(频数图明显不对称)，则用中位数(P_{25}——P_{75})描述
P_{25}——P_{75} 称为四分位数范围(Inter-quartile range,IQR)

但在一些临床试验资料统计分析时，往往给出样本均数、标准差、中位数、四分位数范围、最小值和最大值，但对结果的主要解释按照上述策略进行进行。

第三讲 概率分布和抽样分布

概率分布累积函数

1. 标准正态分布累积函数 norm(X)
2. t 分布右侧累积函数 ttail(df，X) ，其中 df 是自由度
3. χ^2 分布累积函数 chi2(df，X) ，其中 df 是自由度
4. χ^2 分布右侧累积函数 chi2tail(df，X) ，其中 df 是自由度
5. F 分布累积函数 F(df1，df2，X)，df1 为分子自由度，df2 为分母自由度
6. F 分布右侧累积函数 F(df1，df2，X)，df1 为分子自由度，df2 为分母自由度

累积函数的计算使用

正态分布计算

X 服从 N(0,1)，计算概率 P(X<1.96)

```
. display norm(1.96)
.9750021
```

即概率 P(X<1.96)＝0.9750021

display 可简写为 di，如： di norm(1.96)，同样可以得到上述结果。

X 服从 N(0,1)，计算概率 P(X>1.96)，则

```
. di  1- norm(1.96)
.0249979
```

即概率 P(X>1.96)＝0.0249979

X 服从 $N(\mu,\sigma^2)$，则 $Y=\frac{X-\mu}{\sigma}\sim N(0,1)$，因此对其他正态分布只要在函数括号中插入一个上述表达式就可以得到相应概率。

例如：X 服从 $N(100,6^2)$，计算概率 P(X<111.76)，则操作如下

```
. di    norm((111.76-100)/6)
.9750021        即：概率 P(X<111.76)=0.9750021
```

又如 X 服从 $N(100,6^2)$，计算概率 P(X>90)，操作如下

```
. di 1-norm((90-100)/6)
.95220965
```

χ^2 分布累积概率计算

设 X 服从自由度为 1 的χ^2 分布，计算概率 P(X>3.84)，则操作如下

```
. di    1-chi2(1,3.84)
.05004353               概率 P(X>3.84)=0.05004353
```

设 X 服从自由度为 3 的χ^2 分布，计算概率 P(X<5)，则操作如下

```
. di    chi2(3,5)
.82820288           概率 P(X<5)=0.82820288
```

χ^2 分布右侧累积概率计算

设 X 服从自由度为 1 的χ^2 分布，计算概率 P(X>3.84)，则操作如下

```
. di chi2tail(1,3.84)
.05004353               概率 P(X>3.84)=0.05004353
```

设 X 服从自由度为 3 的χ^2 分布，计算概率 P(X<5)，则操作如下

```
.di    chi2(3,5)
.82820288           概率 P(X<5)=0.82820288
```

t 分布右侧累积概率计算

设 t 服从自由度为 10 的 t 分布，计算概率 P(t>2.2)，操作如下

```
. di ttail(10,2.2)
.02622053       概率 P(t>2.2)=0.02622053   (注意：这是右累积函数)
```

设 t 服从自由度为 10 的 t 分布，计算概率 P(t<－2)，操作如下

```
. di 1-ttail(10,-2)
.03669402       概率 P(t<－2)=0.03669402
```

F 分布累积概率计算

设 F 服从 F(3,27)，计算概率 P(F<1)，操作如下：

```
. di F(3,27,1)          注意这里的函数是大写 F，stata 软件中是区分大小写的
.59208514           概率 P(F<1)=0.59208514
```

设 F 服从 F(4,40)，计算概率 P(F>3)，操作如下：

```
. di    1-F(4,40,3)
.02954694           概率 P(F>3)=0 .02954694
```

F 分布右侧累积概率计算

设 F 服从 F(3,27)，计算概率 P(F<1)，操作如下：

```
. di    1-Ftail(3,27,1)          注意这里的函数是大写 F，stata 软件中是区分大小写的
.59208514           概率 P(F<1)=0.59208514
```

设 F 服从 F(4,40)，计算概率 P(F>3)，操作如下：

```
. di    Ftail(4,40,3)
```

.02954694　　概率 P(F>3)=0 .02954694

概率分布的临界值计算

正态分布的临界值计算函数 invnorm(P)

例如：双侧 $U_{0.05}$(即：左侧累积概率为 0.975)，操作如下

. di invnorm(0.975)
1.959964　　即 $U_{0.05}$＝1.959964

t 分布的临界值计算函数 invchi2tail(df,P)

例如计算自由度为 28 的右侧累积概率为 0.025 的临界值 $t_{28,\alpha}$，操作如下

. di invttail(28,0.025)
2.0484071　　临界值 $t_{28,\alpha}$=2.0484071

χ^2 分布的临界值计算函数 invchi2(df,P) 或 invchi2tail(df,P)

例如：计算自由度为 1 的χ^2 右侧累积概率为 0.05 的临界值$\chi^2_{0.05}$，操作如下：

. di invchi2(1,0.95)
3.8414591　　临界值$\chi^2_{0.05}$=3.8414591

或者操作如下：

. di invchi2tail(1,0.05)
3.8414591　　临界值$\chi^2_{0.05}$=3.8414591

F 分布的临界值计算函数 invF(df1,df2,P) 或 invF(df1,df2,P)

例如计算分子自由度为 3 和分母自由度 27 的右侧累积概率为 0.05 的临界值，操作如下：

```
. di invF(3,27,0.95)
2.9603513          临界值 F0.05(3,27)= 2.9603513
```

或者操作为：

```
. di invFtail(3,27,0.05)
2.9603513          临界值 F0.05(3,27)= 2.9603513
```

产生随机数

计算机所产生的随机数是通过一串很长的序列数模拟随机数，故称为伪随机数，在实际应用这些随机数时，这些随机数一般都能具有真实随机数的所有概率性质和统计性质，因此可以产生许许多多的序列伪随机数，一个序列的第一个随机数对应一个数，这个数称为种子数(seed)，因此可以利用种子数，使随机数重复实现。

设置种子数的命令为 set seed 数。每次设置同一种子数，则产生的随机序列是相同的。

产生(0,1)区间上的均匀分布的随机数 uniform()

例如产生种子数为 100 的 20 个在(0,1)区间上的均匀分布的随机数，则操作如下：

```
clear              清除内存
set seed 100       设置种子数为 100
```

set obs 20　　设置样本量为 20

gen r=uniform()　产生 20 个在(0，1)区间上均匀分布的随机数。

list　　显示这些随机数

结果如下

	r
1.	.7185296
2.	.1646728
3.	.9258041
4.	.1833736
5.	.0067327
6.	.7413361
7.	.3599943
8.	.1634543
9.	.445553
10.	.6489049
11.	.3799431
12.	.5964895
13.	.0251346
14.	.2164402
15.	.6848479
16.	.1270018
17.	.6466258
18.	.1869288
19.	.4522384
20.	.067132

利用均匀分布随机数进行随机分组：

例：某实验要把 20 只大鼠随机分为 2 组，每组 10 只，请制定随机分组方案和措施。

第一步、把 20 只大鼠编号，1，2，3，4，5，6，7，8，9，10，11，12，13，14，15，16，17，18，19，20。并且标明。

第二步、用 Stata 软件制定随机分组方案，操作如下：

clear	清除内存
set seed 200	设置种子数为 200
set obs 20	设置样本量为 20
range no 1 20	建立编号 1 至 20
gen r=uniform()	产生在(0,1)均匀分布的随机数
gen group=1	设置分组变量 group 的初始值为 1
sort r	对随机数从小到大排序
replace group=2 in 11/20	设置最大的 10 个随机数所对应的记录为第 2 组，即：最小的 10 个随机数所对应的记录为第 1 组
sort no	按照编号排序
list	显示随机分组的结果

结果如下：

	no	r	group
1.	1	.9512007	2
2.	2	.5249876	2
3.	3	.5129986	1
4.	4	.126439	1
5.	5	.5866161	2
6.	6	.7059209	2
7.	7	.2633286	1
8.	8	.5644688	2
9.	9	.1171033	1
10.	10	.954065	2
11.	11	.4822863	1
12.	12	.3347736	1
13.	13	.5678902	2
14.	14	.7994431	2
15.	15	.1180503	1
16.	16	.9834299	2
17.	17	.2807874	1
18.	18	.095245	1
19.	19	.9446051	2
20.	20	.3467524	1

随机分组整理如下

第一组										
编号	3	4	7	9	11	12	15	17	18	20
第二组										
编号	1	2	5	6	8	10	13	14	16	19

产生服从正态分布 N(μ，σ^2)的随机数 invnorm(uniform())*σ+μ。例如产生 10 个服从正态分布 N(100，6^2)的随机数，操作如下:

clear	清除内存
set seed 200	设置种子数为 200
set obs 10	设置样本量为 10
gen x=invnorm(uniform())*6+100	产生服从 N(100，6^2)的随机数
list	显示随机数

结果如下：

	x
1.	109.9397
2.	100.3761
3.	100.1955
4.	93.13968
5.	101.3131
6.	103.249
7.	96.2013
8.	100.9739
9.	92.86244
10.	110.1137

教学应用：考察样本均数的分布。

由于个体变异的原因，样本均数 $\bar{X}$ 的抽样误差(其定义为样本均数与总体均数的差值)是不可避免的，并且样本均数的抽样误差是呈随机变化的。对于一次抽样而言，无法考察样本均数的抽样误差的规律性，但当大量地重复抽样，计算每次抽样的样本均数 $\bar{X}$，考察样本均数 $\bar{X}$ 的随机分布规律性和统计特征。举例如下：

利用计算机模拟产生 100000 个服从正态分布 N(100,6^2)的样本，样本量分别为 n=4，n=9，n=16，n=36，每个样本计算样本均数。这

里关键处是要清楚什么是样本量(每次抽样所观察的对象个数，也就是每个样本的个体数 n)、什么是样本个数(指抽样的次数)，现以 n=4 为例，一条记录存放一个样本，样本量 n=4，也就是每个样本的第 1 个数据放在第 1 列，第 2 个数据放在第 2 列，第 3 个数据放在第 3 列，第 4 个数据放在第 4 列，因此第 1 行是第一个样本，第 2 行是第 2 个样本，第 100000 行是第 100000 个样本，计算样本均数放在第 5 列，因此共有 100000 个样本均数。具体操作如下：

clear	清除内存
set memory 60m	扩大虚拟内存为 60M
set obs 100000	设置记录数为 100000
set seed 200	设置种子数为 200
gen x1=invnorm(uniform())*6+100	产生第 1 个随机数据
gen x2=invnorm(uniform())*6+100	产生第 2 个随机数据
gen x3=invnorm(uniform())*6+100	产生第 3 个随机数据
gen x4=invnorm(uniform())*6+100	产生第 4 个随机数据
gen mean=(x1+x2+x3+x4)/4	计算平均数，并且存放在变量名为 mean
su mean	以样本均数为数据，计算其平均值和标准差

结果

Variable	Obs	Mean	Std. Dev.	Min	Max
mean	100000	99.98388	3.002225	87.97424	112.0461

现共有 100000 个样本，每个样本计算一个样本均数，因此有 100000 个样本均数，现在把一个样本均数 $\bar{X}$ 视为一个数据，把 100000 个样本均数视为一个样本量为 100000 的新样本(这个样本里有 100000 个

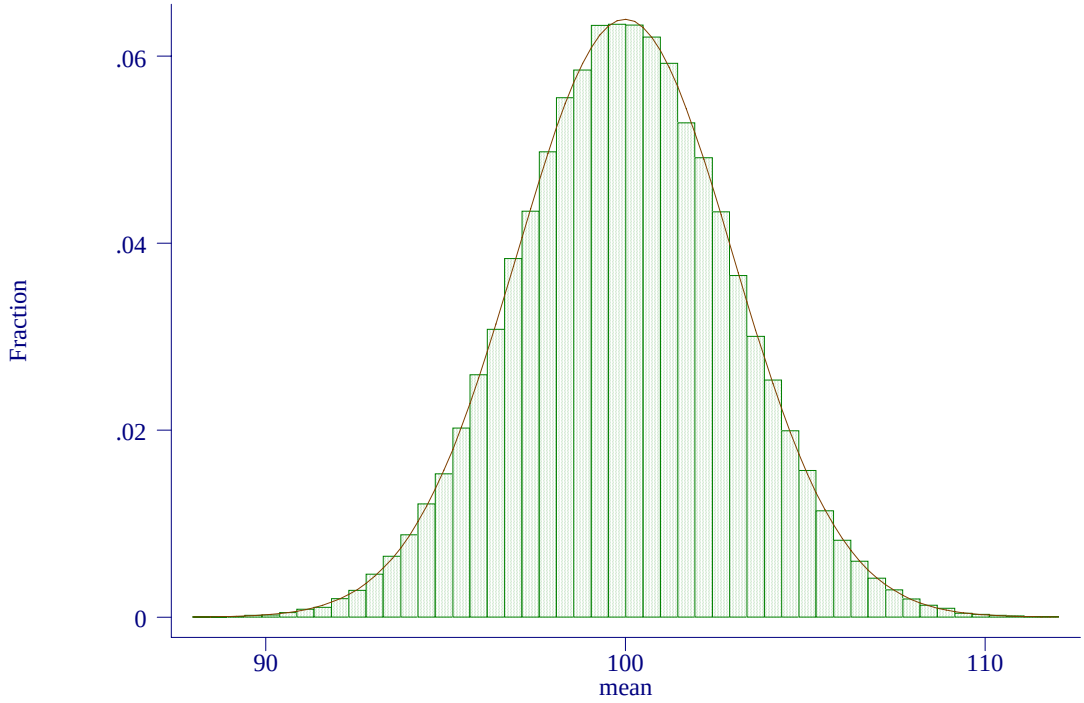
$\bar{X}$)，计算这 100000 个 $\bar{X}$ 的平均值和标准差：得到：

这 100000 个 $\bar{X}$ 的平均值＝99.98388 非常接近总体均数μ=100

这 100000 个 $\bar{X}$ 的标准差＝3.002225$\approx\frac{\sigma}{\sqrt{n}}=\frac{6}{\sqrt{4}}=3$(理论上可以证明样本均数的总体均数与样本所在的总体的总体均数相同，样本均数的标准差=$\frac{\text{样本所在总体的总体标准差}}{\sqrt{n}}$)

再考察这 100000 个 $\bar{X}$ 的频数图

graph mean,bin(50) xlabel ylabel norm



可以发现正态分布的样本均数仍呈正态分布，峰的位置在μ＝100。再考察这 100000 个 $\bar{X}$ 的百分位数

Variable	Obs	Percentile	Centile	-- Binom. Interp. -- [95% Conf.	Interval]
mean	100000	2.5	94.11224	94.05934	94.15675
		5	95.04831	95.00758	95.08677
		50	99.97672	99.95568	100.0002
		95	104.9248	104.8881	104.9571

	97.5	105.8656	105.8161	105.9181

比较理论上的百分位数

百分位数	Stata 操作	理论百分位数	模拟百分位数
$P_{2.5}$	di 100+invnorm(0.025)*3	94.120108	94.11224
P_{5}	di 100+invnorm(0.05)*3	95.065439	95.04831
P_{50}	di 100+invnorm(0.5)*3	100	99.97672
P_{95}	di 100+invnorm(0.95)*3	104.93456	104.9248
$P_{97.5}$	di 100+invnorm(0.975)*3	105.87989	105.8656

可以发现理论上的百分位数与模拟数据的百分位数非常接近。可以证明：样本量越大，这种 $\bar{X}$ 的误差小的可能性越大。

由于在实际研究中，只有一个样本，因此只有一个样本均数，无法如模拟数据一样计算样本均数的标准差，但是一个样本的数据可以计算样本的标准差 S 近似σ，利用样本均数的标准差 $\sigma_{\bar{X}}=\frac{\sigma}{\sqrt{n}}$ 关系，间接估计得到样本均数的标准差估计为 $S_{\bar{X}}=\frac{S}{\sqrt{n}}$，为了区分样本的标准差和样本均数的标准差，故称 $S_{\bar{X}}=\frac{S}{\sqrt{n}}$ 为标准误。

为了帮助大家方便地进行模拟实习，特地编制的相应的 stata 模拟程序:模拟正态分布的样本均数分布的模拟程序 simumean.ado 复制到 stata 软件安装的目录下的子目录 ado\base。例如：stata 软件安装在 D:\stata，则 simumean.ado 复制到 d:\stata\ado\base

然后启动 stata 软件后，输入连接命令:net set ado d:\stata\ado\base

若 stata 安装在其他目录下，则相应改变上述路径便是(这是一次性操作，以后无需再重复进行)。这是模拟抽 10000 个正态分布的样本，具体说明如下：

举例说明

simumean　样本量　均数 标准差

例如模拟抽 10000 个正态分布的样本，样本量为 4、总体均数是 20、标准差为 6，则操作如下：

simumean　4　20　6

得到下列结果(随机的)

Variable	Obs	Mean	Std. Dev.	Min	Max
mean	10000	19.99352	2.990616	8.344506	31.40937
ssd	10000	5.511469	2.346368	.258496	15.51934

即 10000 个样本均数(视为一个新的样本数据)的平均值为 19.99352≈总体均数 20，10000 个样本均数的标准差＝2.990616≈$\frac{6}{\sqrt{4}}=\frac{\text{总体标准差}}{\sqrt{n}}=3$。

变量　样本量　%　百分位数

Variable	Obs	Percentile	Centile	-- Binom. Interp. -- [95% Conf.	Interval]
mean	10000	2.5	14.19629	14.01392	14.31436
		5	15.08899	14.96281	15.2017
		50	19.96537	19.88963	20.03251
		95	24.91111	24.78268	25.05202
		97.5	25.92742	25.75092	26.05995

理论上，样本均数 $\bar{X}$ 的 95%范围是$\mu\pm1.96\frac{\sigma}{\sqrt{n}}$＝20±1.96×3=(14.12,25.88)

比较 10000 个样本均数的 95％百分位数=(14.196,25.927)

模拟习题

1)运行正态分布的样本均数模拟程序 simumean.ado，考察不同样本量情况下，$\bar{X}$ 的标准差与$\frac{\sigma}{\sqrt{n}}$的差异，95％范围的比较。

样本量 n	9	16	25	36	49
总体均数μ	100	100	100	100	100
总体标准差σ	6	6	6	6	6
$\bar{X}$ 的标准差					
$\frac{\sigma}{\sqrt{n}}$					
$\mu\pm1.96\frac{\sigma}{\sqrt{n}}$					

$P_{2.5}-P_{97.5}$					

考察频数图的变化

graph 变量名,xlabel bin(40)

考察原始资料：graph x1,xlabel bin(40)

考察样本均数(变量名为 mean) graph mean,xlabel bin(40)

考察：原始资料和样本均数的峰的位置，离散程度。

考察非正态分布情况下，样本均数

可以运行下列程序

双峰分布的样本均数分布程序：simubpeak.ado

自由度为 1 的χ^2分布的样本均数模拟程序 simuchi.ado

把上述程序复制到 路径:\stata\ado\base

连接：net set ado 路径:\stata\ado\base

操作：simubpeak.ado 样本量

simuchi.ado 样本量

考察原始资料的分布和样本均数的分布变化，

原始资料所在总体分布的频数图：graph x1,bin(40) xlabel

样本均数的抽样分布的频数图：graph meanx ,bin(40) xlabel

考察原始资料 x1,x2 的标准差和样本均数 meanx 的标准差

样本量 n	9	16	25	36	100

考察不同样本量对样本均数分布的影响。

可以证明：样本量较大时，样本均数的分布趋向于正态分布(称为中心极限定理)，并且样本均数的总体均数(理论均数)仍与样本所在总体相同，样本均数的总体标准差(标准误)$=\frac{\text{样本所在总体的总体标准差}\sigma}{\sqrt{n}}$

第四讲 两组计量资料平均水平的统计检验

一、配对设计的平均水平检验

统计方法选择原则：

如果配对的差值服从近似正态分布(小样本)或大样本，则用配对 t 检验

小样本的情况下，配对差值呈明显偏态分布，则用配对秩符号检验(matched-pairs signed-ranks test)。

例 1 10 例男性矽肺患者经克矽平治疗，其血红蛋白（g/dL）如下：

表 10 例男性矽肺患者血红蛋白值（g/dL）

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

问：治疗前后的血红蛋白的平均水平有没有改变

这是一个典型的前后配对设计的研究(但不提倡，因为对结果的解释可能会有问题)

Stata 数据输入结构

X1	X2
11.3	14
15	13.8
15	14
13.5	13.5
12.8	13.5
10	12
11	14.7
12	11.4
13	13.8
12.3	12

操作如下：

gen d=x1-x2　　产生配对差值的变量 d

swilk　d　　正态性检验

正态性检验结果如下：

```
. sktest d
                    Skewness/Kurtosis tests for Normality
                                                       ------- joint ------
    Variable | Pr(Skewness)   Pr(Kurtosis)  adj chi2(2)    Prob>chi2
-------------+-------------------------------------------------------
           d |       0.279          0.774          1.43        0.4885
```

正态性检验的无效假设为：资料正态分布

相应的备选假设为：资料非正态分布

α=0.05，由于正态性检验的 P 值=0.40189>>α，故可以认为资料近似服从正态分布。

ttest　d=0　　配对 t 检验：　H_0:μ_d=0　vs　H_1:$\mu_d\neq0$，

α=0.05

结果如下：

```
One-sample t test
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       d |      10   -.6799999    .5204272    1.645735   -1.857288    .4972881
------------------------------------------------------------------------------
Degrees of freedom: 9

                            Ho: mean(d) = 0

     Ha: mean < 0                 Ha: mean ~= 0                 Ha: mean > 0
       t =  -1.3066                 t =  -1.3066                 t =  -1.3066
   P < t =   0.1119           P > |t| =   0.2237           P > t =   0.8881
```

P 值=0.2237>α，故认为治疗前后的血红蛋白的平均数差异没有统计学意义。即：没有足够的证据可以认为治疗前后的血红蛋白的总体平均数不同。

如果已知差值的样本量，样本均数和样本标准差，可以用立即命令如下(如，已知样本量为 10，差值的样本均数为-0.66，差值的标准差为 1.65，则输入命令如下：

ttesti　样本量 样本均数 样本标准差 0

本例为： ttesti 10 -0.66 1.65 0

得到下列结果如下：

```
. ttesti 10 .66 1.65 0
One-sample t test
------------------------------------------------------------------------------
         |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       x |      10         .66    .5217758        1.65   -.5203389    1.840339
------------------------------------------------------------------------------
Degrees of freedom: 9

                     Ho: mean(x) = 0

     Ha: mean < 0                 Ha: mean ~= 0                 Ha: mean > 0
       t =   1.2649                 t =   1.2649                 t =   1.2649
   P < t =   0.8812             P > |t| =   0.2377             P > t =   0.1188
```

结果解释与结论同上述相同。

如果对于小样本的情况下，差值不满足正态分布，则用 Match-Sign-rank test，操作如下：

signrank　差值变量名=0

假如本例不满足正态分布(为了借用上例资料，而假定的，实际上本例满足正态分布)则

H_0：差值的中位数＝0

(其意义是治疗前的血红蛋白配大于治疗后的血红蛋白的概率＝治疗前的血红蛋白小于治疗后的血红蛋白的概率)

H_1：差值的中位数≠0

$\alpha=0.05$

本例为 signrank　d=0

```
Wilcoxon signed-rank test
        sign |      obs   sum ranks    expected
-------------+--------------------------------
    positive |        4          18          27
    negative |        5          36          27
        zero |        1           1           1
-------------+--------------------------------
         all |       10          55          55
unadjusted variance       96.25
adjustment for ties        0.00
adjustment for zeros      -0.25
                     ----------
adjusted variance         96.00
Ho: d = 0
             z =  -0.919
    Prob > |z| =   0.3583
```

P 值=0.3583>>α，故没有足够的证据说明两个总体不同。

二、平行对照设计的两组资料平均水平统计检验

统计方法选择原则：

如果两组资料的方差齐性和相互独立的，并且每组资料服从正态分布(大样本资料可以忽略正态性问题)，则用成组 t 检验，否则可以用成组 Wilcoxon 秩和检验。

例 2 为研究噪声对纺织女工子代智能是否有影响，一研究人员在某纺织厂随机抽取接触噪声 95dB（A）、接触工龄 5 年以上的纺织女工及同一单位、条件与接触组相近但不接触噪声的女职工，其子女（学前幼儿）作为研究对象，按韦氏学前儿童智力量表（中国修订版）测定两组幼儿智商，结果如下。问噪声对纺织女工子

代智能有无影响？(接触组 group=0，不接触组 group=1)

资料及其结果如下：

group	x
0	79
0	93
0	91
0	92
0	94
0	77
0	93
0	74
0	91
0	101
0	83
0	73
0	88
0	102
0	90
0	100
0	81
0	91
0	83
0	106
0	84
0	78
0	87
0	95
0	101
1	101
1	100
1	114
1	86
1	106
1	107
1	107
1	94
1	89
1	104
1	98
1	110

1	89
1	103
1	89
1	121
1	94
1	95
1	92
1	109
1	98
1	98
1	120
1	104
1	110

方差齐性检验

H_0：$\sigma_1=\sigma_2$ vs H_1:$\sigma_1\neq\sigma_2$

α=0.1

两组方差齐性的检验命令(仅适合两组方差齐性检验)

sdtest x,by(group)

```
Variance ratio test
------------------------------------------------------------------------------
   Group |      Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |       25       89.08    1.822928    9.11464    85.31766    92.84234
       1 |       25      101.52    1.900982    9.504911    97.59657    105.4434
---------+--------------------------------------------------------------------
combined |       50        95.3    1.577456    11.1543    92.12998    98.47002
------------------------------------------------------------------------------
                            Ho: sd(0) = sd(1)
              F(24,24) observed   = F_obs             =     0.920
              F(24,24) lower tail = F_L   = F_obs     =     0.920
              F(24,24) upper tail = F_U   = 1/F_obs   =     1.087
   Ha: sd(0) < sd(1)          Ha: sd(0) ~= sd(1)          Ha: sd(0) > sd(1)
 P < F_obs = 0.4195      P < F_L + P > F_U = 0.8389      P > F_obs = 0.5805
```

P 值=0.8389>>α，因此可以认为两组方差齐性的。

正态性检验：H_0：资料服从正态分布 vs H_1：资料偏态分布

α=0.05

每一组资料正态性检验

```
. swilk  x  if group==1
                    Shapiro-Wilk W test for normal data
    Variable |    Obs        W          V          z      Prob>z
-------------+-------------------------------------------------
           x |     25    0.97403      0.722     -0.667   0.74747

. swilk  x  if group==0
                    Shapiro-Wilk W test for normal data
    Variable |    Obs        W          V          z      Prob>z
-------------+-------------------------------------------------
           x |     25    0.97199      0.778     -0.513   0.69588
```

P 值均大于α，因此可以认为两组资料都服从正态分布

H_0：$\mu_1=\mu_2$　vs　H_1：$\mu_1\neq\mu_2$

α=0.05

ttest　x,by(group)

```
Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |      25       89.08    1.822928    9.11464    85.31766    92.84234
       1 |      25      101.52    1.900982   9.504911    97.59657    105.4434
---------+--------------------------------------------------------------------
combined |      50        95.3    1.577456    11.1543    92.12998    98.47002
---------+--------------------------------------------------------------------
    diff |             -12.44    2.633781               -17.73557   -7.144429
------------------------------------------------------------------------------
Degrees of freedom: 48
                     Ho: mean(0) - mean(1) = diff = 0
     Ha: diff < 0                 Ha: diff ~= 0                 Ha: diff > 0
       t =  -4.7232                 t =  -4.7232                 t =  -4.7232
   P < t =   0.0000             P > |t| =   0.0000             P > t =   1.0000
```

P 值(<0.0001)<α,并且有$\mu_0-\mu_1$的 95%可信区间为(-17.73557,-7.144429)

可以知道，不接触组幼儿的平均智商高于接触组的幼儿平均智商，并且差别有统计学意义。

如果已知两组的样本量、样本均数和样本标准差，也可以用立即命令

进行统计检验

ttesti　样本量 1 样本均数 1 样本标准差 1 样本量 2 样本均数 2 样本标准差 2

例如：本例第 1 组 n1=25 均数 1=89.08 标准差 1=9.115

第 2 组 n2=25 均数 2=101.52 标准差 2=9.505

则 ttesti　25　89.08 9.115 25 101.52 9.505

```
Two-sample t test with equal variances
------------------------------------------------------------------------------
         |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       x |      25       89.08       1.823       9.115    85.31751    92.84249
       y |      25      101.52       1.901       9.505    97.59653    105.4435
---------+--------------------------------------------------------------------
combined |      50        95.3    1.577482    11.15448    92.12993    98.47007
---------+--------------------------------------------------------------------
    diff |             -12.44    2.633843               -17.7357   -7.144303
------------------------------------------------------------------------------
Degrees of freedom: 48

                    Ho: mean(x) - mean(y) = diff = 0

     Ha: diff < 0                 Ha: diff ~= 0                 Ha: diff > 0
       t =  -4.7231                 t =  -4.7231                 t =  -4.7231
   P < t =   0.0000         P > |t| =   0.0000         P > t =   1.0000
```

结果解释同上。

方差不齐的情况，(小样本时，资料正态分布)还可以用 t’检验

命令：ttest　观察变量名，by(分组变量名) unequal

立即命令为 ttesti 样本量 1 均数 1 标准差 1 样本量 2 均数 2 标准差 2,unequal

假定本例的资料方差不齐(实际为方差不齐的)，则要用 t’检验如下

ttest x,by(group) unequal

```
Two-sample t test with unequal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |      25       89.08    1.822928    9.11464    85.31766    92.84234
       1 |      25      101.52    1.900982    9.504911   97.59657    105.4434
---------+--------------------------------------------------------------------
combined |      50        95.3    1.577456    11.1543    92.12998    98.47002
---------+--------------------------------------------------------------------
    diff |             -12.44    2.633781               -17.73581   -7.144189
------------------------------------------------------------------------------
Satterthwaite's degrees of freedom:  47.9159
                     Ho: mean(0) - mean(1) = diff = 0
     Ha: diff < 0                 Ha: diff ~= 0                 Ha: diff > 0
       t =  -4.7232                 t =  -4.7232                  t =  -4.7232
   P < t =   0.0000           P > |t| =   0.0000          P > t =   1.0000
```

结果解释同上。

t’检验有许多方法，这里介绍的 Satterthwaite 方法，主要根据两个样本方差差异的程度校正相应的自由度，由于本例的两个样本方差比较接近，故自由度几乎没有减少(t 检验的自由度为 48,而本例 t’自由度为 47.9159)。由于 t 检验要求的两组总体方差相同(称为方差齐性)，以及由于抽样误差的原因，样本方差一般不会相等，但是方差齐性的情况下，样本方差表现为两个样本方差之比≈1。(注意：两个样本方差之差很小，仍可能方差不齐。如：第一个样本标准差为 0.1，样本量为 100,第 2 个样本标准差为 0.01，样本量为 100，两个样本标准差仅差 0.09，但是两个样本方差之比为 100。故用方差齐性检验的结果如下：

方差齐性的立即命令为 sdtesti 样本量1 . 标准差1 样本量2 . 标准差2

sdtesti 100 . 0.1 100 . 0.01

```
Variance ratio test
------------------------------------------------------------------------------
         |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       x |     100           .         .01          .1           .           .
       y |     100           .        .001         .01           .           .
---------+--------------------------------------------------------------------
combined |     200           .           .           .           .           .
------------------------------------------------------------------------------
                            Ho: sd(x) = sd(y)
              F(99,99) observed   = F_obs             =   100.000
              F(99,99) lower tail = F_L   = 1/F_obs =     0.010
              F(99,99) upper tail = F_U   = F_obs   =   100.000

   Ha: sd(x) < sd(y)          Ha: sd(x) ~= sd(y)          Ha: sd(x) > sd(y)
  P < F_obs = 1.0000     P < F_L + P > F_U = 0.0000     P > F_obs = 0.0000
```

P 值<0.0001，因此认为两组的方差不齐。故方差齐性是考察两个样本方差之比是否接近 1。

如果本例的资料不满足 t 检验要求(注：实际是满足的，只是想用本例介绍成组秩和检验)，则用秩和检验(Wilcoxon Ranksum test)。

H_0:两组资料所在总体相同

H_1：两组资料所在总体不同

$\alpha=0.05$

命令：ranksum 观察变量名,by(分组变量)

本例为 ranksum x,by(group)

```
. ranksum x,by(group)

Two-sample Wilcoxon rank-sum (Mann-Whitney) test

       group |      obs    rank sum    expected
-------------+---------------------------------
           0 |       25         437       637.5
           1 |       25         838       637.5
-------------+---------------------------------
    combined |       50        1275        1275

unadjusted variance     2656.25
adjustment for ties       -3.70
                        ---------
adjusted variance       2652.55

Ho: x(group==0) = x(group==1)
             z =  -3.893
    Prob > |z| =   0.0001
```

P 值<0.0001<α，故认为两个总体不同

练习题

一、 某地随机抽样调查了部分健康成人红细胞数和血红蛋白量，结果如下，请就此资料统计分析：

指标	性别	例数	均数	标准差	标准值
红细胞数（10^{12}/L）	男	360	4.66	0.58	4.84
	女	255	4.18	0.29	4.33
血红蛋白（g/L）	男	360	134.50	7.10	140.20
	女	255	117.60	10.20	124.70

（1） 该地健康成年男女血红蛋白含量有无差别？

（2） 该地男女两项血液指标是否均低于上表的标准值（若测定方法相同）？

二、 为了解聋哑学生学习成绩与血清锌含量的关系，某人按年龄、性别和班级在聋哑学校随机抽取成绩优、差的 14 对学生进行配对研究，得其结果如下。问聋哑学生学习成绩与血清锌含量有无关系？

表 14 对学生的血清锌含量（μg/mL）

编号	优生组	差生组	编号	优生组	差生组

1	1.20	1.31	8	0.80	0.86
2	0.99	1.34	9	0.84	0.72
3	1.03	1.10	10	0.85	0.88
4	0.90	0.72	11	1.05	0.81
5	1.22	0.92	12	1.08	1.30
6	0.90	1.34	13	1.15	0.85
7	0.97	0.98	14	0.90	0.80

教学应用：考察影响t检验结果的各种因素

1. 首先把程序ttest2.ado和程序ttestexp.ado复制到stata所在的目录下\ado\base(例如：Stata软件安装在D:\stata，则把这两个程序复制到d:\stata\ado\base目录下。然后输入连接命令:在STATA环境下，输入 net set ado 路径\stata\ado\base。(路径表示Stata所在的盘符和目录)
2. 程序ttest2.ado是模拟在正态总体中随机抽10000个样本，每个样本有2组，两组的样本量、正态分布的总体均数和标准差由读者选择输入，考察α=0.05的情况下，考察当两个总体均数相同时拒绝H_0的比例(拒绝的频率估计第一类错误)是否接近0.05和当两个总体均数不同时接受H_0的比例(估计发生第二类错误的概率)。

   运行ttest2.ado的输入命令为：

   ttest2 样本量1 均数1 标准差1 样本量2 均数2 标准差2

   例如：考察两组样本量均为30，总体均数均为100，标准差均为6的拒绝$H_0(\mu_1=\mu_2)$比例，结果如下：

```
. ttest2 30  100 6  30 100 6
两样本t检验模拟程序
输入 样本量1 均数1 标准差1 样本量2 均数2 标准差2

        sig |      Freq.     Percent        Cum.
------------+-----------------------------------
    receive |        9506       95.06       95.06
     refuse |         494        4.94      100.00
------------+-----------------------------------
      Total |       10000      100.00

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
    average1 |     10000    99.99388    1.083106   95.77671   104.2778
         sd1 |     10000    5.942067    .7764423   3.245709   8.692573
    average2 |     10000    99.99675    1.086406   95.91508   103.8237
         sd2 |     10000    5.949536    .7776711   3.276635   9.546211
           t |     10000    -.003644      1.0035   -4.32787   3.602131

                                                      -- Binom. Interp. --
    Variable |       Obs  Percentile      Centile        [95% Conf. Interval]
-------------+-------------------------------------------------------------
           t |     10000         2.5    -2.001922       -2.077161   -1.955956
             |                    50    -.0115932       -.0389369    .0137221
             |                  97.5      1.992317        1.933308    2.033179
    average1 |     10000         2.5     97.85904        97.79236    97.93009
             |                    50     99.98936        99.96717    100.0172
             |                  97.5     102.1116        102.0614    102.1734
    average2 |     10000         2.5     97.86119        97.80749    97.91781
             |                    50      99.9868        99.96412    100.0107
             |                  97.5     102.1835        102.1131    102.2403
```

在随机抽10000个样本中，计算了10000个t值，结果有494次拒绝$H_0(\mu_1=\mu_2)$，因此非常接近$\alpha=0.05$。

建议读者运行程序ttest2考察下列情况

目的1：$\mu_1\neq\mu_2$时，不同的样本量，考察下列不同情况下的接受H_0的比例(估计β)以及两组样本量之比不同的情况对检验结果的影响。

	两组的总体标准差$\sigma=2$

	μ_1=100 μ_2=99	μ_1=100	μ_2=98	μ_1=100	μ_2=97
n_1： n_2	10:10	10:10		10:10	
n_1： n_2	20:20	30:30		20:20	
n_1： n_2	30:30	10:50		30:30	
n_1： n_2	40:40	40:40		40:40	
n_1： n_2	30:50	30:50		30:50	
n_1： n_2	20:60	20:60		20:60	
n_1： n_2	10:70	10:70		10:70	

目的2：考察方差不齐对t检验(不是t’检验)结果的影响

	μ_1=100 μ_2=100	μ_1=100	μ_2=98	μ_1=100	μ_2=97
	σ_1=1 σ_2=9	σ_1=9	σ_2=1	σ_1=5	σ_2=5
n_1： n_2	40:10	40:10		40:10	
n_1： n_2	10:40	10:40		10:40	
n_1： n_2	60:30	60:30		60:30	
n_1： n_2	30:60	30:60		30:60	
n_1： n_2	30:30	30:30		30:30	
n_1： n_2	40:40	40:40		40:40	
n_1： n_2	40:40	40:40		40:40	

目的3：通过运行程序ttestexp.ado，考察资料非正态分布对结果的影响。

3. 程序ttestexp.ado是模拟在指数分布总体中随机抽10000个样本，

每个样本有2组，两组的样本量和总体均数由读者选择输入，考察α=0.05的情况下，考察当两个总体均数相同时拒绝H_0的比例(拒绝的频率估计第一类错误)是否接近0.05和当两个总体均数不同时接受H_0的比例(估计发生第二类错误的概率)。

运行ttestexp.ado的输入命令为：

ttestexp　样本量1 均数1　样本量2 均数2

例如：考察两组样本量均为10，总体均数均为1的拒绝$H_0(\mu_1=\mu_2)$的比例，结果如下：

```
. ttestexp  5 1 5 1
指数分布
输入 样本量1  均数1  样本量2  均数2
    Variable |      Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
    average1 |    10000    .9942006    .444696   .1223783    3.46752
         sd1 |    10000    .8637844   .5004927   .0310705   4.281092
    average2 |    10000    1.007233   .4560518   .0613991   3.577513
         sd2 |    10000    .8707893   .5053219   .0353676   4.620248
           t |    10000   -.0177069   1.115122  -5.645559   6.235384

        sig |      Freq.     Percent        Cum.
------------+-----------------------------------
    receive |        9630       96.30       96.30
     refuse |         370        3.70      100.00
------------+-----------------------------------
      Total |       10000      100.00

                                                      -- Binom. Interp. --
    Variable |      Obs  Percentile      Centile        [95% Conf. Interval]
-------------+------------------------------------------------------------
           t |    10000         2.5    -2.169495       -2.23945   -2.096289
             |                   50    .0088744       -.0182028    .0357137
             |                 97.5    2.089225        2.030593    2.155895
    average1 |    10000         2.5    .3240474        .3139804    .3349038
             |                   50    .9310558        .9198599    .9414931
             |                 97.5    2.041828        2.010877    2.081691
    average2 |    10000         2.5    .3262316        .3131719    .3370006
             |                   50    .9381162        .9254703    .9481275
```

	97.5	2.092387	2.0545	2.14614

拒绝H_0($\mu_1=\mu_2$)的比例为3.7%，离开$\alpha=0.05$，较远。考察下列样本量情况与偏态分布造成的影响之间的关系。

	n_1： n_2	n_1： n_2	n_1： n_2	n_1： n_2	n_1： n_2
$\mu_1=1, \mu_2=1$	5:5	10:10	30:30	20:40	60:60
$\mu_1=1.5, \mu_2=1$	5:5	10:10	30:30	20:40	60:60
$\mu_1=2, \mu_2=1$	5:5	10:10	30:30	20:40	60:60

您能从上述模拟结果可以得到下列结论

1)当$\mu_1\neq\mu_2$时且方差齐性的正态分布情况下，$n_1=n_2$时，拒绝H_0的比例比较高，可以证明t检验中，两组样本量为n_1和n_2，则其检验效能等价于每组样本量相同$n=\frac{2}{\frac{1}{n_1}+\frac{1}{n_2}}$。特别当两组样本量之比为n：kn时，则样本量等价于$\frac{2}{\frac{1}{n}+\frac{1}{kn}}=\frac{2n}{1+\frac{1}{k}}<2n$，也就是说，如果一组的样本量为10，另一组的样本量再大，其检验效能也不会超过两组样本量相同且为20的统计检验效能。

2)当方差不齐时，且$\mu_1=\mu_2$，拒绝H_0的比例偏离α，但是$n_1=n_2$时，方差不齐对结果的影响将下降。

3)资料偏态分布，则小样本时，偏态分布对结果有影响，大样本时，偏态分布对结果基本无影响。

第五讲　多组平均水平的比较

一、复习和补充两组比较的统计检验

1. 配对设计资料(又称为 Dependent Samples)

a)对于小样本的情况下，如果配对的差值资料服从正态分布，用配对 t 检验(ttest　差值变量=0)

b)大样本的情况下，可以用配对 t 检验

c)小样本的情况下，并且配对差值呈偏态分布，则用配对符号秩检验(signrank 差值变量=0)

2. 成组设计(Two Independent Samples)

a)如果方差齐性并且大样本情况下，可以用成组 t 检验(ttest　效应指标变量,by(分组变量))

b)如果方差齐性并且两组资料分别呈正态分布，可以用成组 t 检验

c)如果方差不齐，或者小样本情况下偏态分布，则用秩和检验(Ranksum test)

group	x
0	79
0	93
0	91
0	92
0	94
0	77
0	93
0	74
0	91
0	101
0	83
0	73
0	88
0	102
0	90
0	100
0	81
0	91
0	83
0	106
0	84
0	78
0	87
0	95
0	101

1	101
1	100
1	114
1	86
1	106
1	107
1	107
1	94
1	89
1	104
1	98
1	110
1	89
1	103
1	89
1	121
1	94
1	95
1	92
1	109
1	98
1	98
1	120
1	104
1	110

二、多组比较

1. 完全随机分组设计(要求各组资料之间相互独立)

a)方差齐性并且独立以及每一组资料都服从正态分布(小样本时要求)，则采用完全随机设计的方差分析方法(即：单因素方差分析，One Way ANOVA)进行分析。

b)方差不齐或小样本情况下资料偏态，则用 Kruskal Wallis 检验(H 检验)

例5.1 为研究胃癌与胃粘膜细胞中DNA含量（A.U）的关系，某医师测得数据如下，试问四组人群的胃粘膜细胞中平均DNA含量是否相同？

组别	group	DNA 含量（A.U）									
浅表型胃炎	1	9.81	12.73	12.29	12.53	12.95	9.53	12.6	8.9	12.27	14.26	10.68
肠化生	2	14.61	17.54	15.1	17	13.39	15.32	13.74	18.24	13.81	12.63	14.53	16.17
早期胃癌	3	23.26	20.8	20.6	23.5	17.85	21.91	22.13	22.04	19.53	18.41	21.48	20.24
晚期胃癌	4	23.73	19.46	22.39	19.53	25.9	20.43	20.71	20.05	23.41	21.34	21.38	25.70

由于这四组对象的资料是相互独立的，因此属于完全随机分组类型的。检验问题是考察四组 DNA 含量的平均水平相同吗。如果每一组资料都正态分布并且方差齐性可以用 One way-ANOVA 进行分析，反之用 Kruskal Wallis 检验。

STATA 数据输入格式

g	x
1	9.81
1	12.73
1	12.29
1	12.53
1	12.95
1	9.53
1	12.6
1	8.9
1	12.27
1	14.26
1	10.68
2	14.61
2	17.54
2	15.1
2	17
2	13.39
2	15.32
2	13.74
2	18.24
2	13.81
2	12.63
2	14.53
2	16.17
3	23.26
3	20.8
3	20.6
3	23.5
3	17.85
3	21.91
3	22.13
3	22.04
3	19.53
3	18.41
3	21.48
3	20.24
4	23.73
4	19.46
4	22.39
4	19.53
4	25.9
4	20.43
4	20.71
4	20.05

4	23.41
4	21.34
4	21.38
4	25.7

分组正态性检验，α=0.05

```
. sktest x  if g==1

                    Skewness/Kurtosis tests for Normality
                                                       ------- joint ------
    Variable | Pr(Skewness)   Pr(Kurtosis)  adj chi2(2)    Prob>chi2
-------------+-------------------------------------------------------
           x |       0.491         0.485           1.07       0.5861

. sktest x  if g==2

                    Skewness/Kurtosis tests for Normality
                                                       ------- joint ------
    Variable | Pr(Skewness)   Pr(Kurtosis)  adj chi2(2)    Prob>chi2
-------------+-------------------------------------------------------
           x |       0.482         0.541           0.96       0.6201

. sktest x  if g==3

                    Skewness/Kurtosis tests for Normality
                                                       ------- joint ------
    Variable | Pr(Skewness)   Pr(Kurtosis)  adj chi2(2)    Prob>chi2
-------------+-------------------------------------------------------
           x |       0.527         0.750           0.52       0.7704

. sktest x  if g==4

                    Skewness/Kurtosis tests for Normality
                                                       ------- joint ------
    Variable | Pr(Skewness)   Pr(Kurtosis)  adj chi2(2)    Prob>chi2
-------------+-------------------------------------------------------
           x |       0.260         0.616           1.75       0.4166
```

上述结果表明每一组资料都服从正态分布。

单因素方差分析的 STATA 命令：oneway　效应指标变量　分组变量,t　b

其中 t 表示计算每一组均数和标准差,b 表示采用 Bonferroni 统计方法进行两两比较。

本例命令为 oneway x group,t b

```
. oneway x g,t b
                |            Summary of x
              g |        Mean   Std. Dev.       Freq.
----------------+------------------------------------
              1 |   11.686364   1.6884388          11
              2 |   15.173333    1.749173          12
              3 |   20.979167   1.7668279          12
              4 |     22.0025   2.2429087          12
----------------+------------------------------------
          Total |   17.583191   4.6080789          47

                        Analysis of Variance
    Source              SS         df      MS            F     Prob > F
------------------------------------------------------------------------
Between groups      824.942549      3     274.98085     77.87      0.0000
 Within groups      151.839445     43    3.53114987
------------------------------------------------------------------------
    Total           976.781994     46    21.2343912
Bartlett's test for equal variances:  chi2(3) =   1.1354  Prob>chi2 = 0.769
```

方差齐性的检验为：卡方=1.1354，自由度=3,P 值=0.769，因此可以认为方差是齐性的。

H_0：$\mu_1=\mu_2=\mu_3=\mu_4$ 四组总体均数相同

H_1：μ_1，μ_2，μ_3，μ_4不全相同

α=0.05，相应的统计量 F=77.87 以及相应的自由度为 3 和 43，P 值<0.0001，因此 4 组均数的差别有统计学意义。

```
                        Comparison of x by g
                              (Bonferroni)
Row Mean-|
Col Mean |          1          2          3
---------+------------------------------
       2 |    3.48697(第 2 组样本均数－第 1 组样本均数)
         |      0.000(H0:μ1=μ2 检验的 P 值)
         |
       3 |     9.2928    5.80583(第 3 组样本均数－第 2 组样本均数)
         |      0.000      0.000(H0:μ3=μ2 检验的 P 值)
         |
       4 |    10.3161    6.82917    1.02333(第 4 组样本均数-第 3 组样本均数)
         |      0.000      0.000      1.000(H0:μ3=μ4 检验的 P 值)\
```

上述输出为两两比较的结果，在表格的每个单元中，第一行为两组均数的差值，第二行为两组均数比较检验的 P 值。

根据上述结果可以知道，第 2 组、第 3 组和第 4 组的 AU 均数均大于第 1 组的 AU 均数，并且差别有统计学意义。说明肠化生患者和胃癌患者的 DNA 的 AU 含量平均水平均高于正常人的 AU 平均水平，并且差别有统计学意义。

第 3 组和第 4 组的 AU 均数也大于第 2 组的 AU 平均水平，并且差别有统计学意义。说明胃癌患者的 DNA 的 AU 含量平均水平均高于肠化生患者的 AU 平均水平，并且差别有统计学

意义。
第 3 组和第 4 组两组均数的差别没有统计学意义，说明没有足够的证据可以 DNA 的 AU 含量与癌症的早期与晚期有关系。
假如本例的资料不满足方差分析的要求，则用 Kruskal Wallis 检验，数据结构同上。命令为：
kwallis 效应指标变量, by(分组变量)
本例的命令为 kwallis x,by(g)
H_0：4 组的 AU 总体分布相同
H_1：4 组的 AU 总体分布不全相同
α=0.05

结果如下：

```
Test: Equality of populations (Kruskal-Wallis test)
       g          _Obs    _RankSum
       1            11       72.00
       2            12      205.00
       3            12      411.50
       4            12      439.50
chi-squared =    37.814 with 3 d.f.
probability =     0.0001

chi-squared with ties =    37.816 with 3 d.f.
probability =     0.0001
```

说明：4 组 AU 的总体分布不全相同，然后秩和检验，但α应取小一些(多重比较时，会增大第一类错误的概率)。根据 Sidak 检验的建议：$\alpha' = 1-(1-\alpha)^{\frac{1}{k}}$，其中 k 为要比较的次数，α为多组比较总的检验水平(一般为 0.05)，α'为两两比较时的检验水平。

如本例：4 组两两比较共比 $C_4^2 = 6$ 次，因此 $a' = 1-(0.95)^{\frac{1}{6}} = 0.0085$，

对于比较第 1 组和第 2 组的 AU 分布差别的操作命令为：
先计算中位数
sort g 组别变量排序
by g:centile x,centile(50) 计算各组中位数

```
-> g = 1
                                                    -- Binom. Interp. --
    Variable |     Obs  Percentile      Centile        [95% Conf. Interval]
-------------+-------------------------------------------------------------
           x |      11          50         12.29        9.729564      12.7932
-> g = 2
                                                    -- Binom. Interp. --
    Variable |     Obs  Percentile      Centile        [95% Conf. Interval]
-------------+-------------------------------------------------------------
           x |      12          50         14.855        13.74745     16.91172
-> g = 3
                                                    -- Binom. Interp. --
    Variable |     Obs  Percentile      Centile        [95% Conf. Interval]
-------------+-------------------------------------------------------------
           x |      12          50          21.14        19.60552     22.12043
-> g = 4
```

Variable	Obs	Percentile	Centile	-- Binom. Interp. -- [95% Conf.	Interval]
x	12	50	21.36	20.09042	23.69596

得到这 4 组中位数分别为：M_1=12.29，M_2=14.855，M_3=21.14 和 M_4=21.36

ranksum x if g==1 | g==2,by(g)

Two-sample Wilcoxon rank-sum (Mann-Whitney) test

g	obs	rank sum	expected
1	11	72	132
2	12	204	144
combined	23	276	276

unadjusted variance 264.00
adjustment for ties 0.00
adjusted variance 264.00
Ho: x(g==1) = x(g==2)
z = -3.693
Prob > |z| = 0.0002

P 值<α’，因此第 2 组 AU 的平均水平要高于第 1 组的平均水平(M_2>M_1)，并且差别有统计学意义。

第 1 组与第 3 组比较

ranksum x if g==1 | g==3,by(g)

Two-sample Wilcoxon rank-sum (Mann-Whitney) test

g	obs	rank sum	expected
1	11	66	132
3	12	210	144
combined	23	276	276

unadjusted variance 264.00
adjustment for ties 0.00
adjusted variance 264.00
Ho: x(g==1) = x(g==3)
z = -4.062
Prob > |z| = 0.0000

P 值<α’，因此第 3 组 AU 的平均水平要高于第 1 组的平均水平(M_3>M_1)，并且差别有统计学

意义，其他比较类似进行。

要注意的问题：

- 在方差分析中，要求每一组资料服从正态分布(小样本时)，并不是要求各组资料服从一个正态分布(因为这就意味各组的总体均数相同，失去统计检验的必要性)，所以不能把各组的资料合在一起作正态性检验。总的讲，方差分析对正态性具有稳健性，即：偏态分布对方差分析的结果影响不会太大，故正态性检验的α取 0.05 也就可以了。
- 样本量较大时，方差分析对正态性要求大大降低(根据中心极限定理可知：样本均数近似服从正态分布)。并且由于大多数情况下，样本资料只是近似服从正态分布而不是完全服从正态分布。由于在大样本情况下，用正态性检验就变为很敏感，对于不是完全服从正态分布的资料往往会拒绝正态性检验的 H_0：资料服从正态分布。因为正态性检验不能检验资料是否近似服从正态分布，而是检验是否服从正态分布。故在大样本情况下，考察资料的近似正态性，应用频数图进行考察。
- 方差齐性问题对方差分析相对比较敏感，并且并不是随着样本量增大而方差齐性对方差分析减少影响的。但是当各组样本量接近相同或相同时，方差齐性对方差分析呈现某种稳健性。即：只有当各组样本量相同时，方差齐性对方差分析结果的影响大大降低。这时随着样本量增大，影响会进一步降低。相反，如果各组样本量相差太大时，方差齐性对方差分析结果的影响很大。这时随着样本量增大，影响会进一步加大。

2. 随机区组设计(处理组之间可能不独立)

a)残差(定义为：$e_{ij} = X_{ij} + \bar{\bar{X}} - \bar{X}_{i.} - \bar{X}_{.j}$，也就是随机区组方差分析中的误差项)的方差齐性且小样本时正态分布，则用随机区组的方差分析(无重复的两因素方差分析,Two-way ANOVA)。

b)不满足方差齐性或小样本时资料偏态，则对用秩变换后再用随机区组的方差分析也可以直接用非参数随机区组的秩和检验 Fredman test)。

例2下表是某湖水中8个观察地点不同季节取样的氯化物含量测定值，请问在不同季节该湖水中氯化物的含量有无差别？

表2　某湖水中不同季节的氯化物含量测定值（mg/L）

location no	春	夏	秋	冬
1	21.28	18.33	17.27	14.91
2	22.78	19.81	16.55	14.85
3	20.90	18.93	16.36	16.30
4	19.90	21.23	17.86	15.73
5	21.49	19.09	15.11	17.05
6	22.38	17.92	16.57	14.34
7	21.67	19.39	17.19	16.31
8	22.06	19.65	16.58	14.33

显然同一地点不同季节的氯化物含量有一定的相关性，故不能采用完全随机设计的方差分析方法对4个季节的氯化物含量进行统计分析。可以把同一地点的4个季节氯化物含量视为一个区组，因此可以用随机区组的方差分析进行统计分析。

设第8个地点在冬季的氯化物总体均数为μ_0，同样在冬季，第i个地点的氯化物总体均数与第8个地点在冬季的氯化物总体均数相差β_i，i=1，2，3，4，5，6，7。因此在冬季的这8个地点在冬季的氯化物总体均数可以表示为

地点编号	1	2	3	4	5	6	7	8

冬季氯化物均数	$\mu_0+\beta_1$	$\mu_0+\beta_2$	$\mu_0+\beta_3$	$\mu_0+\beta_4$	$\mu_0+\beta_5$	$\mu_0+\beta_6$	$\mu_0+\beta_7$	μ_0

假定在同一地区，春季的氯化物总体均数与冬季的氯化物总体均数相差α_1，因此春节和冬季的氯化物总体均数可以表示为

地点编号	1	2	3	4	5	6	7	8
冬季氯化物均数	$\mu_0+\beta_1$	$\mu_0+\beta_2$	$\mu_0+\beta_3$	$\mu_0+\beta_4$	$\mu_0+\beta_5$	$\mu_0+\beta_6$	$\mu_0+\beta_7$	μ_0
春季氯化物均数	$\mu_0+\alpha_1+\beta_1$	$\mu_0+\alpha_1+\beta_2$	$\mu_0+\alpha_1+\beta_3$	$\mu_0+\alpha_1+\beta_4$	$\mu_0+\alpha_1+\beta_5$	$\mu_0+\alpha_1+\beta_6$	$\mu_0+\alpha_1+\beta_7$	μ_0

如果$\alpha_1=0$说明在同一地点，冬季和春季的氯化物总体均数相同；$\alpha_1>0$说明春季的氯化物含量平均高于冬季氯化物含量，反之$\alpha<0$，说明春季氯化物含量均数低于冬季氯化物含量。

同理假定在同一地区，夏季和秋季的氯化物总体均数与冬季的氯化物总体均数分别相差α_2和α_3，则四个季节的氯化物总体均数可以表示为

地点编号	1	2	3	4	5	6	7	8
冬季氯化物均数	$\mu_0+\beta_1$	$\mu_0+\beta_2$	$\mu_0+\beta_3$	$\mu_0+\beta_4$	$\mu_0+\beta_5$	$\mu_0+\beta_6$	$\mu_0+\beta_7$	μ_0
春季氯化物均数	$\mu_0+\alpha_1+\beta_1$	$\mu_0+\alpha_1+\beta_2$	$\mu_0+\alpha_1+\beta_3$	$\mu_0+\alpha_1+\beta_4$	$\mu_0+\alpha_1+\beta_5$	$\mu_0+\alpha_1+\beta_6$	$\mu_0+\alpha_1+\beta_7$	μ_0
夏季氯化物均数	$\mu_0+\alpha_2+\beta_1$	$\mu_0+\alpha_2+\beta_2$	$\mu_0+\alpha_2+\beta_3$	$\mu_0+\alpha_2+\beta_4$	$\mu_0+\alpha_2+\beta_5$	$\mu_0+\alpha_2+\beta_6$	$\mu_0+\alpha_2+\beta_7$	μ_0
春季氯化物均数	$\mu_0+\alpha_3+\beta_1$	$\mu_0+\alpha_3+\beta_2$	$\mu_0+\alpha_3+\beta_3$	$\mu_0+\alpha_3+\beta_4$	$\mu_0+\alpha_3+\beta_5$	$\mu_0+\alpha_3+\beta_6$	$\mu_0+\alpha_3+\beta_7$	μ_0

根据上述总体均数表示，可以知道：在四个季节中的氯化物总体均数(同一地点)无变化就是H_0：$\alpha_1=\alpha_2=\alpha_3=0$(在随机区组方差分析中称为无处理效应，但不能称4组的总体均数相同，因为在同一季节中不同地点的总体均数可能不同)。

H_1：α_1，α_2，α_3不全为0

Stata 数据输入格式

t	id	x
1	1	21.27589
1	2	22.77649
1	3	20.89943
1	4	19.9043
1	5	21.4929
1	6	22.38085
1	7	21.67344
1	8	22.06133
2	1	18.33405
2	2	19.80538
2	3	18.92919
2	4	21.22814
2	5	19.09215
2	6	17.9237
2	7	19.38569
2	8	19.64971
3	1	17.27141
3	2	16.54567
3	3	16.36019
3	4	17.85548
3	5	15.11296

3	6	16.56507
3	7	17.18734
3	8	16.58279
4	1	14.90559
4	2	14.85127
4	3	16.29782
4	4	15.7286
4	5	17.05169
4	6	14.34088
4	7	16.31367
4	8	14.33015

其中 id 表示观察地点编号，t＝1，2，3，4 对应表示春节、夏季、秋季和冬季。
Stata 操作命令：
anova x t id
. anova x t id

```
                   Number of obs =        32     R-squared     =  0.8923
                   Root MSE      =  1.01769     Adj R-squared =  0.8410

          Source |  Partial SS    df       MS           F     Prob > F
-----------------+----------------------------------------------------
           Model |  180.214326    10  18.0214326       17.40     0.0000
                 |
               t |  177.344737     3  59.1149122       57.08     0.0000
              id |  2.86958916     7  .409941308        0.40     0.8942
                 |
        Residual |   21.749618    21   1.0356961
-----------------+----------------------------------------------------
           Total |  201.963944    31  6.51496593
```

处理效应 H_0：$\alpha_1=\alpha_2=\alpha_3=0$ 的检验对应的统计量 $F=\frac{MS_{处理}}{MS_{误差}}=\frac{18.021}{1.036}=57.08$

相应的 P 值<0.0001(计算机输出值是 0.0000)，所以拒绝无效假设，可以认为 4 个季节的氯化物总体均数不全相同。
不同季节中的两两比较用 LSD 方法检验如下：
在输入 anova　x　t　id 命令后，再输入 regress 命令便得到下列结果

Source	SS	df	MS	Number of obs =	32
Model	180.214326	10	18.0214326	F(10, 21) =	17.40
Residual	21.749618	21	1.0356961	Prob > F =	0.0000
Total	201.963944	31	6.51496593	R-squared =	0.8923
				Adj R-squared =	0.8410
				Root MSE =	1.0177

x	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
_cons (μ_0)	15.37992	.5966746	25.78	0.000	14.13906	16.62077
t						
α_1= 1	6.080619	.5088458	11.95	0.000	5.022417	7.138822
α_2= 2	3.816041	.5088458	7.50	0.000	2.757838	4.874244
α_3= 3	1.20765	.5088458	2.37	0.027	.1494472	2.265853
4	(dropped)					
id						
β_1= 1	-.2092595	.7196166	-0.29	0.774	-1.705784	1.287265
β_2= 2	.3387067	.7196166	0.47	0.643	-1.157818	1.835231
β_3= 3	-.034339	.7196166	-0.05	0.962	-1.530864	1.462186
β_4= 4	.5231357	.7196166	0.73	0.475	-.973389	2.01966
β_5= 5	.0314307	.7196166	0.04	0.966	-1.465094	1.527955
β_6= 6	-.353369	.7196166	-0.49	0.628	-1.849894	1.143156
β_7= 7	.4840407	.7196166	0.67	0.509	-1.012484	1.980565
8	(dropped)					

其中$\hat{\alpha}_1 = 6.081$，对应的假设检验 H_0：$\alpha_1=0$ 的统计量 t=11.95，P 值<0.001，95%可信区间为(5.022，7.139)，因此可以认为春季的氯化物平均高于冬季，差别有统计学意义。

$\hat{\alpha}_2 = 3.816$，对应的假设检验 H_0：$\alpha_2=0$ 的统计量 t=7.50，P 值<0.001，95%可信区间为(2.758，4.874)，因此可以认为夏季的氯化物平均高于冬季，差别有统计学意义。

$\hat{\alpha}_3 = 1.208$，对应的假设检验 H_0：$\alpha_3=0$ 的统计量 t=2.37，P 值=0.027，95%可信区间为(0.1494，2.266)，因此可以认为秋季的氯化物平均高于冬季，差别有统计学意义。

对于春季氯化物平均数($\mu_0+\alpha_1+\beta_i$)与夏季的氯化物平均数($\mu_0+\alpha_2+\beta_i$)比较对应为$\alpha_1>\alpha_2$、$\alpha_1=\alpha_2$和$\alpha_1<\alpha_2$的问题。因此需要检验 H_0：$\alpha_1=\alpha_2$ vs H_1：$\alpha_1\neq\alpha_2$ ，相应的 STATA 命令(anova x t id 命令和 regress 命令后)为 test b[t[1]]=_b[t[2]]，得到下列结果

```
( 1)  t[2] - t[3] = 0.0

       F(  1,     21) =   26.28
            Prob > F =    0.0000
```

相应的统计量 F=26.28，P 值<0.0001，差别有统计学意义。由于α_1的估计值>α_2的估计值，所以可以认为春季氯化物平均高于夏季的氯化物含量。

同理检验 H_0：$\alpha_1=\alpha_3$ vs H_1：$\alpha_1\neq\alpha_3$，只需输入命令 test b[t[1]]=_b[t[3]]
检验 H_0：$\alpha_2=\alpha_3$ vs H_1：$\alpha_2\neq\alpha_3$，只需输入命令 test b[t[2]]=_b[t[3]]
此处不在详细叙述了。

由于随机区组方差分析要求残差($e_{ij}=X_{ij}+\bar{\bar{X}}-\bar{X}_{i.}-\bar{X}_{.j}$)服从正态分布，再输入 regress 以后，只要输入 predict 残差变量名,residual，就可以得到残差计算值。
本例用 e 表示残差变量名，因此输入 predict e,residual
就可以得到残差计算值 e，然后对残差进行正态性检验(sktest 残差变量名)
本例输入命令为: sktest e
结果如下（H_0：残差服从正态分布 vs H_1：残差偏态分布，$\alpha=0.05$）

```
                    Skewness/Kurtosis tests for Normality
                                                    ------- joint ------
    Variable | Pr(Skewness)   Pr(Kurtosis)  adj chi2(2)     Prob>chi2
-------------+-------------------------------------------------------
           e |       0.699          0.586          0.46        0.7948
```

P 值=0.93349>>α，因此可以认为资料近似服从正态分布。(大样本时，可以不考虑正态性问题)

如果资料呈偏态分布，可以对资料进行秩变换(Rank Transform)后，然后把变换后的秩视为原始数据进行随机区组的方差分析。
秩变换的 STATA 命令为 egen 秩变量名＝rank(观察变量名),by(区组变量)
为了说明上述操作分析的过程，故借用本例资料进行秩变换操作说明如下(本例资料正态分布，无需用秩变换，只是说明操作而言).
设用 r 表示秩变量名，则本例操作为
egen r=rank(x) ,by(id) 产生秩 r
anova 命令 anova r t id 结果如下

```
              Number of obs =      32     R-squared     =  0.9125
              Root MSE      = .408248     Adj R-squared =  0.8708

    Source | Partial SS    df       MS           F     Prob > F
-----------+----------------------------------------------------
     Model |       36.50    10         3.65       21.90     0.0000
           |
         t |       36.50     3  12.1666667       73.00     0.0000
        id |        0.00     7         0.00        0.00     1.0000
           |
  Residual |        3.50    21  .166666667
-----------+----------------------------------------------------
     Total |       40.00    31  1.29032258
```

命令 regress　　结果如下

```
regress

      Source |       SS       df       MS              Number of obs =      32
-------------+------------------------------           F( 10,     21) =   21.90
       Model |       36.50     10        3.65          Prob > F      =  0.0000
    Residual |        3.50     21  .166666667          R-squared     =  0.9125
-------------+------------------------------           Adj R-squared =  0.8708
       Total |       40.00     31  1.29032258          Root MSE      =  .40825

------------------------------------------------------------------------------
           r        Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------------------------------------------------------------------------
_cons               1.125   .2393568     4.70   0.000     .6272303     1.62277
t
           1         2.75   .2041241    13.47   0.000     2.325501     3.174499
           2            2   .2041241     9.80   0.000     1.575501     2.424499
           3          .75   .2041241     3.67   0.001     .3255006     1.174499
           4    (dropped)
id
           1            0   .2886751     0.00   1.000    -.6003328     .6003328
           2            0   .2886751     0.00   1.000    -.6003328     .6003328
           3            0   .2886751     0.00   1.000    -.6003328     .6003328
           4            0   .2886751     0.00   1.000    -.6003328     .6003328
           5            0   .2886751     0.00   1.000    -.6003328     .6003328
           6            0   .2886751     0.00   1.000    -.6003328     .6003328
           7            0   .2886751     0.00   1.000    -.6003328     .6003328
           8    (dropped)
------------------------------------------------------------------------------
```

进一步两两比较

```
test _b[t[1]]=_b[t[2]]
 ( 1)  t[1] - t[2] = 0.0
           F(  1,    21) =   13.50
                Prob > F =    0.0014

. test _b[t[1]]=_b[t[3]]
 ( 1)  t[1] - t[3] = 0.0
           F(  1,    21) =   96.00
                Prob > F =    0.0000

. test _b[t[2]]=_b[t[3]]
 ( 1)  t[2] - t[3] = 0.0

           F(  1,    21) =   37.50
                Prob > F =    0.0000
```

解释如同上述，不再重复。

第六讲 线性相关和回归

在实际研究中，经常要考察两个指标之间的关系，即：相关性。现以体重与身高的关系为例，分析两个变量之间的相关性。要求身高和体重呈双正态分布，既：在身高和体重平均数的附近的频数较多，远离身高和体重平均数的频数较少。

样本相关系数计算公式(称为 Pearson 相关系数)：

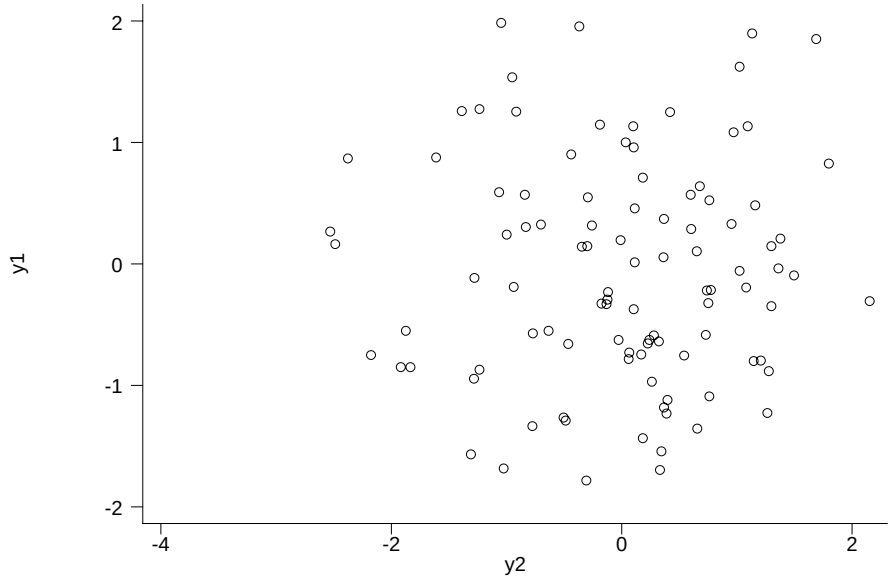
$$r=\frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sqrt{\sum(X-\bar{X})^2}\sqrt{\sum(Y-\bar{Y})^2}}=\frac{L_{XY}}{\sqrt{L_{XX}}\sqrt{L_{YY}}} \qquad (1)$$

1. 考察随机模拟相关的情况。

显示两个变量相关的散点图程序 simur.ado（本教材配套程序,使用见前言）。命令为 simur　样本量　总体相关系数

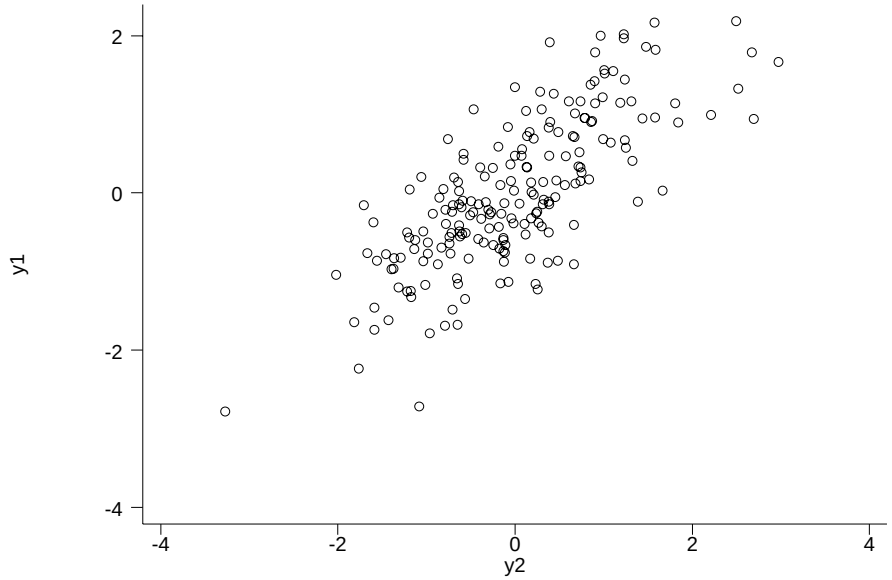
如显示样本量为 100，ρ=0 的散点图

本例命令为 simur　100　0



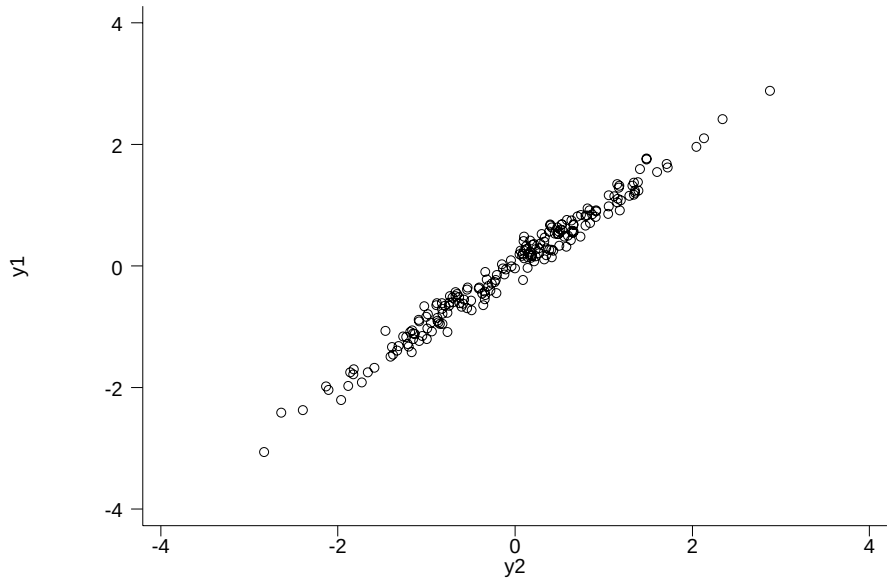
如显示样本量为 200，ρ=0.8 的散点图

本例命令为 simur　200　0.8



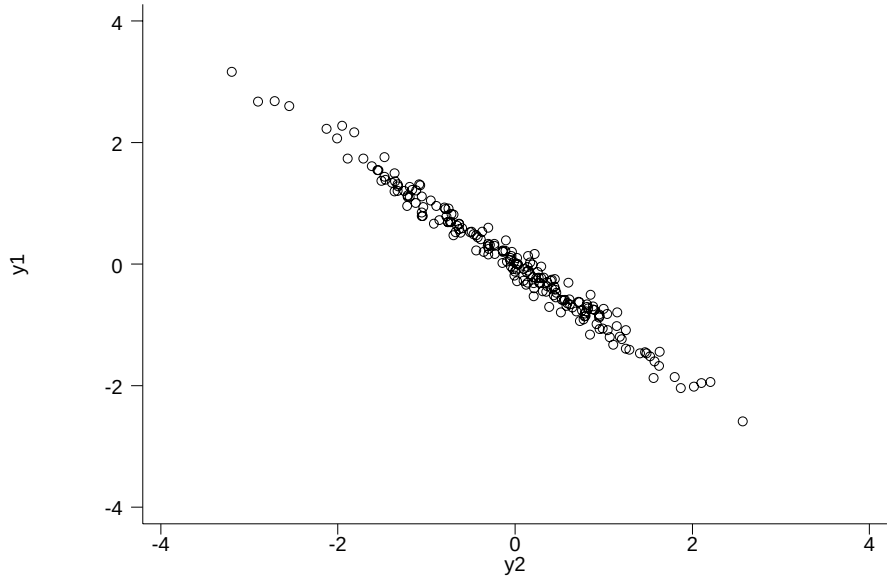
如显示样本量为 200，ρ=0.99 的散点图

本例命令为 simur　200　0.99



如显示样本量为 200，ρ=-0.99 的散点图

本例命令为 simur　200　-0.99



例 1.　测得某地 15 名正常成年男子的身高 x（cm）、体重 y（kg）如试计算 x 和 y 之间的相关系数 r 并检验 H_0：$\rho=0$ vs　H_1: $\rho\neq0$。

$\alpha=0.05$

数据格式为

X	Y
171.0	58.0
176.0	69.0
175.0	74.0
172.0	68.0
170.0	64.0
173.0	68.5
168.0	56.0
172.0	54.0
170.0	62.0
172.0	63.0
173.0	67.0
168.0	60.0
171.0	68.0
172.0	76.0
173.0	65.0

Stata 命令　pwcorr　变量 1　变量 2 …　变量 m，sig

本例命令　pwcorr　x　y,sig

pwcorr x y,sig

```
             |        x          y
-------------+------------------
           x |   1.0000
             |
             |
           y |   0.5994    1.0000
             |   0.0182
             |
```

Pearson 相关系数=0.5994，P 值=0.0182<0.05，因此可以认为身高与体重呈正线性相关。

注意：Pearson 相关系数又称为线性相关系数并且要求 X 和 Y 双正态分布，通常在检查中要求 X 服从正态分布并且 Y 服从正态分布。

如果不满足双正态分布时，可以计算 Spearman 相关系数又称为非参数相关系数。

Spearman 相关系数的计算基本思想为：用 X 和 Y 的秩代替它们的原始数据，然后代入 Pearson 相关系数的计算公式并且检验与 Pearson 相关系数类同。

Stata 实现

spearman x y

Number of obs = 15
Spearman's rho = 0.6552
Test of Ho: x and y are independent
Prob > |t| = 0.0080

stata 计算结果与手算的结果一致。结论为身高与体重呈正相关，并且有统计学意义。

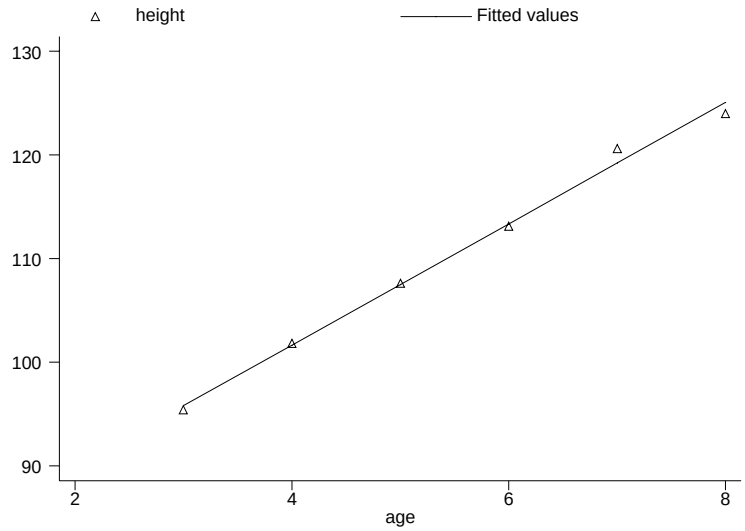
直线回归

例 2 为了研究 3 岁至 8 岁男孩身高与年龄的规律，在某地区在 3 岁至 8 岁男孩中随机抽样，共分 6 个年龄层抽样：3 岁，4 岁，…，8 岁，每个层抽 10 个男孩，共抽 60 个男孩。资料如下：

60 个男孩的身高资料如下

年龄	3 岁	4 岁	5 岁	6 岁	7 岁	8 岁
身高	92.5	96.5	106.0	115.5	125.5	121.5
	97.0	101.0	104.0	115.5	117.5	128.5
	96.0	105.5	107.0	111.5	118.0	124.0
	96.5	102.0	109.5	110.0	117.0	125.5
	97.0	105.0	111.0	114.5	122.0	122.5
	92.0	99.5	107.5	112.5	119.0	123.5
	96.5	102.0	107.0	116.5	119.0	120.5
	91.0	100.0	111.5	110.0	125.5	123.0
	96.0	106.5	103.0	114.5	120.5	124.0
	99.0	100.0	109.0	110.0	122.0	126.5
平均身高	95.4	101.8	107.6	113.1	120.6	124.0

由于男孩的身高与年龄有关系，不同的年龄组的平均身高是不同的，由平均身高与年龄作图可以发现：年龄与平均身高的点在一条直线附近。



考虑到样本均数存在抽样误差，故有理由认为身高的总体均数与年龄的关系可能是一条直线关系 $\mu_y = \alpha + \beta x$，其中 y 表示身高，x 表示年龄。由于身高的总体均数与年龄有关，所以更正确地标记应为

$$\mu_{y|x} = \alpha + \beta x$$

表示在固定年龄情况下的身高总体均数。

上述公式称为直线回归方程。其中β为回归系数（regression coefficient），或称为斜率（slope）；α称为常数项（constant），或称为截距（intercept）。回归系数β表示 *x* 变化一个单位 *y* 平均变化β个单位。当 x 和 y 都是随机的，*x*、*y* 间呈正相关时β>0，*x*、*y* 间呈负相关时β<0，*x*、*y* 间独立时β=0。

一般情况而言，参数α和β是未知的。对于本例而言，不同民族和不同地区，α和β往往是不同的，因此需要进行估计的。由于不同年龄的身高实际观察值应在对应的身高总体均数附近(即：实际观察值与总体均数之间仅存在个体变异的差异)，故可以用年龄和实际身高观察值的资料对未知参数α和β进行估计。得到样本估计的回归方程

$$\hat{y} = a + bx$$

二、直线回归方程的建立

直线回归分析的 Stata 实现：

数据结构：

x	y
3	92.5
3	97
3	96
3	96.5
3	97
3	92
3	96.5
3	91
3	96
3	99
4	96.5
4	101
4	105.5
4	102
4	105
4	99.5
4	102
4	100
4	106.5
4	100
5	106
5	104
5	107
5	109.5
5	111
5	107.5
5	107
5	111.5
5	103
5	109
6	115.5
6	115.5
6	111.5
6	110

6	114.5
6	112.5
6	116.5
6	110
6	114.5
6	110
7	125.5
7	117.5
7	118
7	117
7	122
7	119
7	119
7	125.5
7	120.5
7	122
8	121.5
8	128.5
8	124
8	125.5
8	122.5
8	123.5
8	120.5
8	123
8	124
8	126.5

多重线性回归命令为

regress　因变量　自变量 1　自变量 2 ……自变量 m

直线回归命令 regress　因变量　自变量

本例为 regress　y　x，得到下列结果：

```
      Source |       SS       df       MS              Number of obs =      60
-------------+------------------------------           F(  1,    58) =  777.41
       Model |  5997.71571     1  5997.71571           Prob > F      =  0.0000
    Residual |  447.467619    58  7.71495895           R-squared     =  0.9306
-------------+------------------------------           Adj R-squared =  0.9294
       Total |  6445.18333    59  109.240395           Root MSE      =  2.7776

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
```

x	5.854286	.2099654	27.88	0.000	5.433994	6.274577
_cons	78.18476	1.209202	64.66	0.000	75.76428	80.60524

得到回归系数 b=5.854286，常数项 a=78.18746，回归系数的检验统计量 t_b=27.88，P 值<0.0001，可以认为 Y 与 X 呈直线回归关系。

来源	平方和 SS	自由度 df	均方 MS	F	P 值
回归	5997.71571	1	5997.71571	777.41	<0.0001
残差	447.467619	58	7.71495895		
合计	6445.18333	59			

称 $R^2=1-\frac{SS_{残差}}{SS_{合计}}$ 为决定系数(本例 Stata 计算结果 R-squared=0.9306)，因此 $0\le R^2\le 1$，因此残差平方和 SSE 越小，决定系数 R^2 就越接近 1。特别当所有的残差为 0 时，SSE=0，相应的决定系数 $R^2=1$。决定系数 R^2 表示 y 被 x 所解释的部分所占的百分比，R^2 越接近于 1 说明 x 对 y 的解释越充分。

残差=应变量观察值（y）-预测值($\hat{y}$)

Stata 的残差计算命令

在输入回归命令 regress　y　x 后，再

输入 predict　e,residual 计算残差并用变量 e 表示残差

输入 sktest　e　　残差的正态性检验

输入 predict　yy　　计算预测值。

残差正态性检验(H_0:残差正态分布,α=0.05)

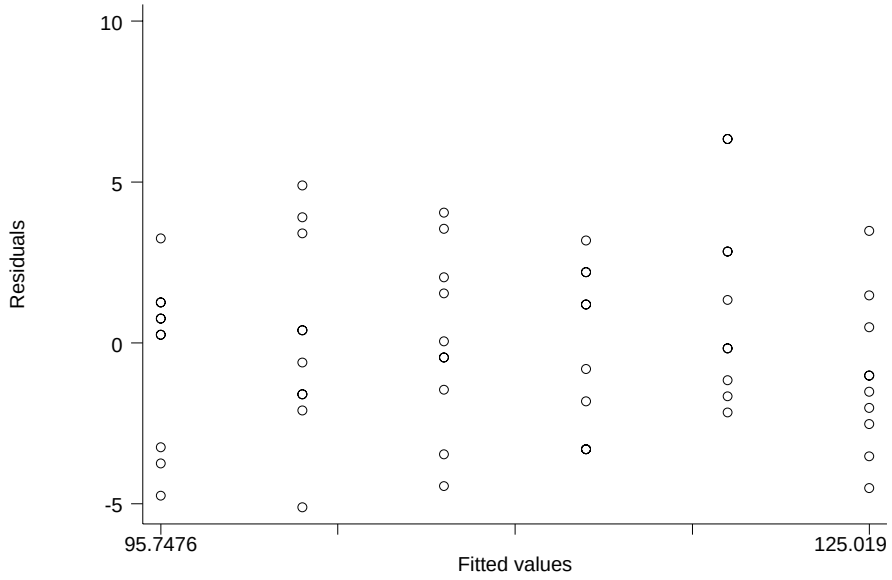
```
sktest e

                    Skewness/Kurtosis tests for Normality
                                                  ------- joint ------
    Variable | Pr(Skewness)   Pr(Kurtosis)  adj chi2(2)    Prob>chi2
-------------+-------------------------------------------------------
           e |       0.459          0.441          1.18        0.5534
```

P 值=0.5534>>0.05，可以认为残差呈正态分布。

所建立的回归方程是否有意义，仅凭借假设检验的结论或 R^2 的大小还不能充分说明问题。残差 $\boldsymbol{e}=\boldsymbol{Y}-\hat{\boldsymbol{Y}}$ 的大小直接反应回归方程的优劣，经常采用图示的方法，以 e 做纵轴，$\hat{\boldsymbol{Y}}$ 为横轴作图来考察残差的变化，如果残差比较均匀地散布在 e=0 的周围，没有明显的散布趋势和明显的离群点，则说明所建回归方程比较理想，否则要借助统计软件做进一步诊断。

graph　残差　预测值

本例　graph　e　yy



说明残差比较均匀地散布在 e=0 的周围，没有明显的散布趋势和明显的离群点，故说明所建回归方程比较理想。