Stata 软件基本操作和数据分析入门

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

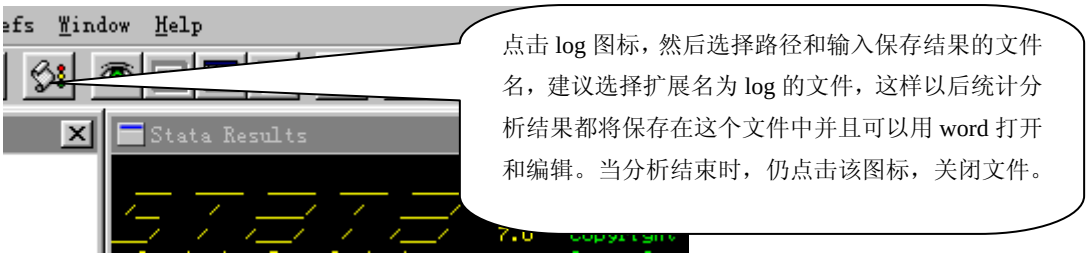
24	185.4
25	177.3
26	168.7
27	173.3
28	175.2
29	176.6
30	183.6
31	171.1
32	175.8
33	177.3
34	176.9
35	175.8
36	175.2
37	175
38	177.5
39	178.5
40	174.3
41	170.3
42	172.9
43	181
44	178.5
45	173.8
46	173.5
47	179.8
48	183.7
49	173
50	177.6
51	168.7
52	178.8
53	177
54	177.6
55	179.3
56	171.5
57	175.9
58	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：



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

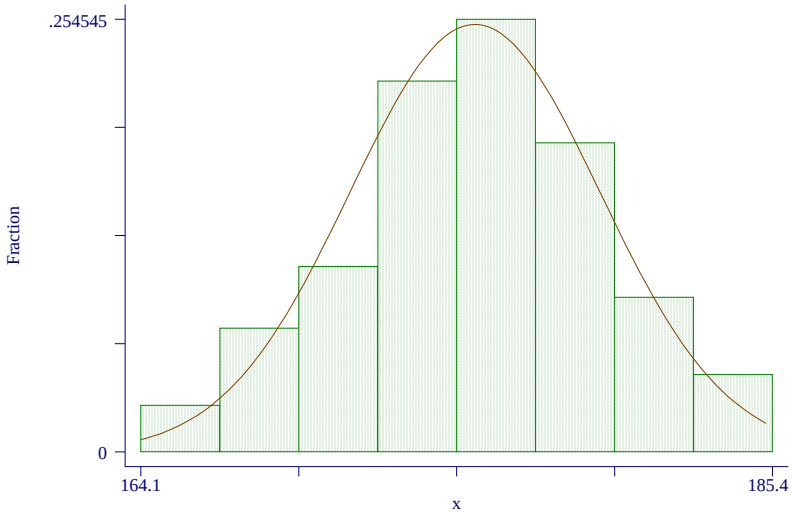
178	13	11.82	86.36
180	10	9.09	95.45
182	4	3.64	99.09
184	1	0.91	100.00
Total	110	100.00	

作频数图

命令 graph 变量,bin(#) norm

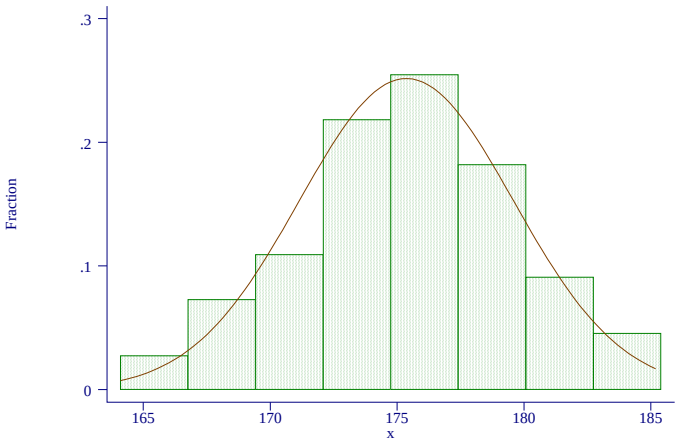
其中#表示频数图的组数;norm 表示画一条相应的正态曲线(可以不要)

本例命令为 graph x,bin(8) norm

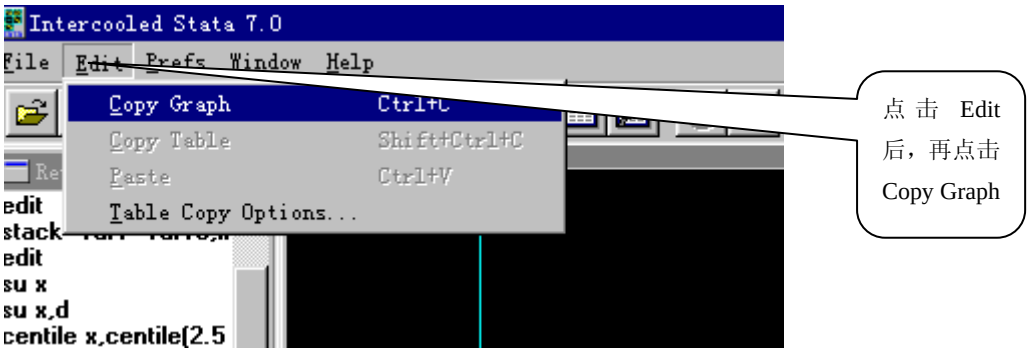


为了使坐标更清楚地在图上显示，可以输入下列命令

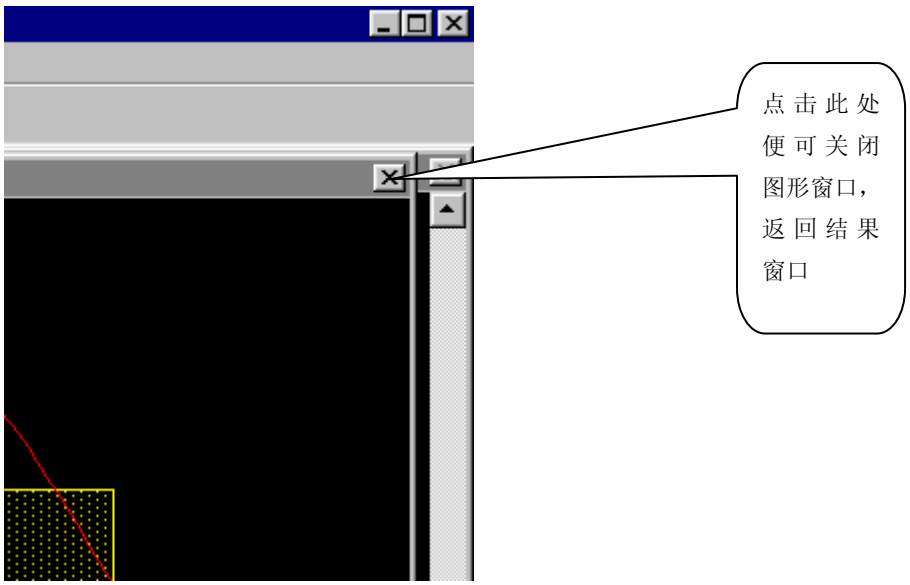
graph x,bin(8) xlabel norm ylabel



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：

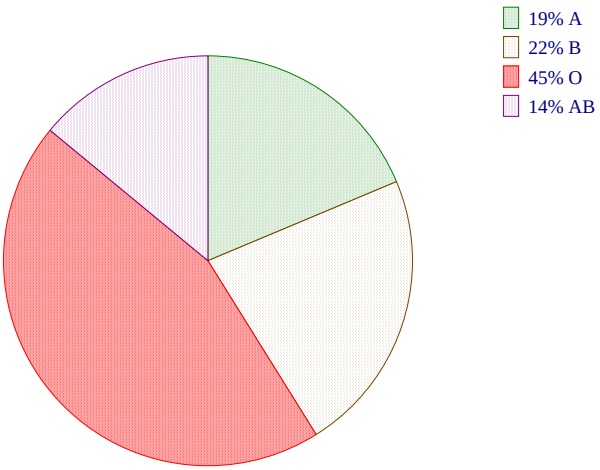
	频数			
地区	A	B	O	AB
第 1 地区 area=1	100	120	240	75
第 2 地区 area=2	80	70	200	50

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

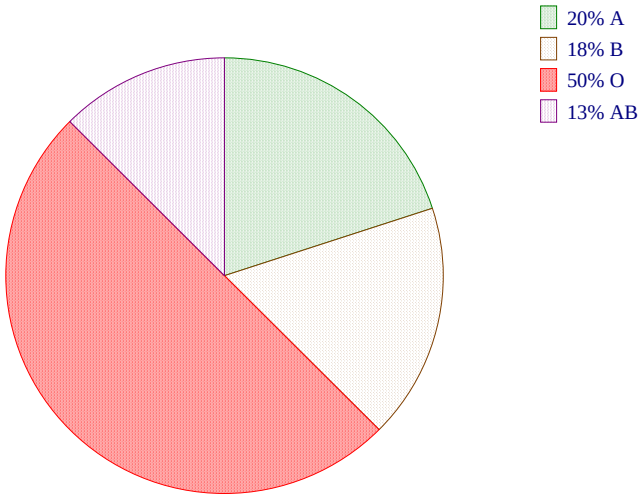
```
graph a b o ab  if area==1,pie
```



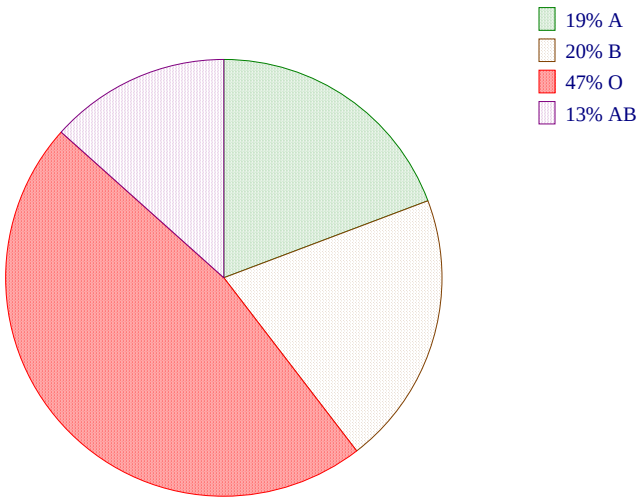
注意逻辑表达式中 if area==1 是两个等号。

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。