第一讲 SPSS概览

- 什么是SPSS
- SPSS窗口
- 变量定义
- 数据输入、输出

什么是SPSS？

- SPSS是软件英文名称的首字母缩写，原意为Statistical Package for the Social Sciences，即“社会科学统计软件包”
- 随着SPSS产品服务领域的扩大和服务深度的增加，SPSS公司于2000年正式将英文全称更改为Statistical Product and Service Solutions，意为“统计产品与服务解决方案

- SPSS已有40年历史
- SPSS是应用最广的定量数据分析和管理的统计软件
- SPSS既可以指SPSS软件，又可以指SPSS公司

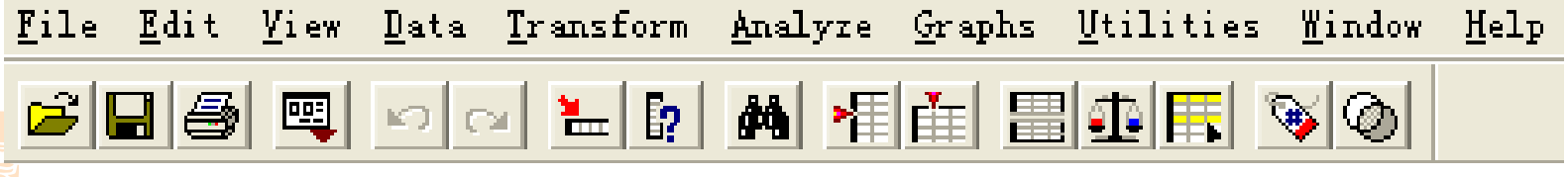
SPSS窗口

SPSS有三个窗口：

- 数据编辑窗口 Data Editor Window
  - 数据显示窗口 Data View
  - 变量显示窗口 Variable View
- 结果输出窗口 Output Viewer Window
- 命令编辑窗口 Syntax Editor Window

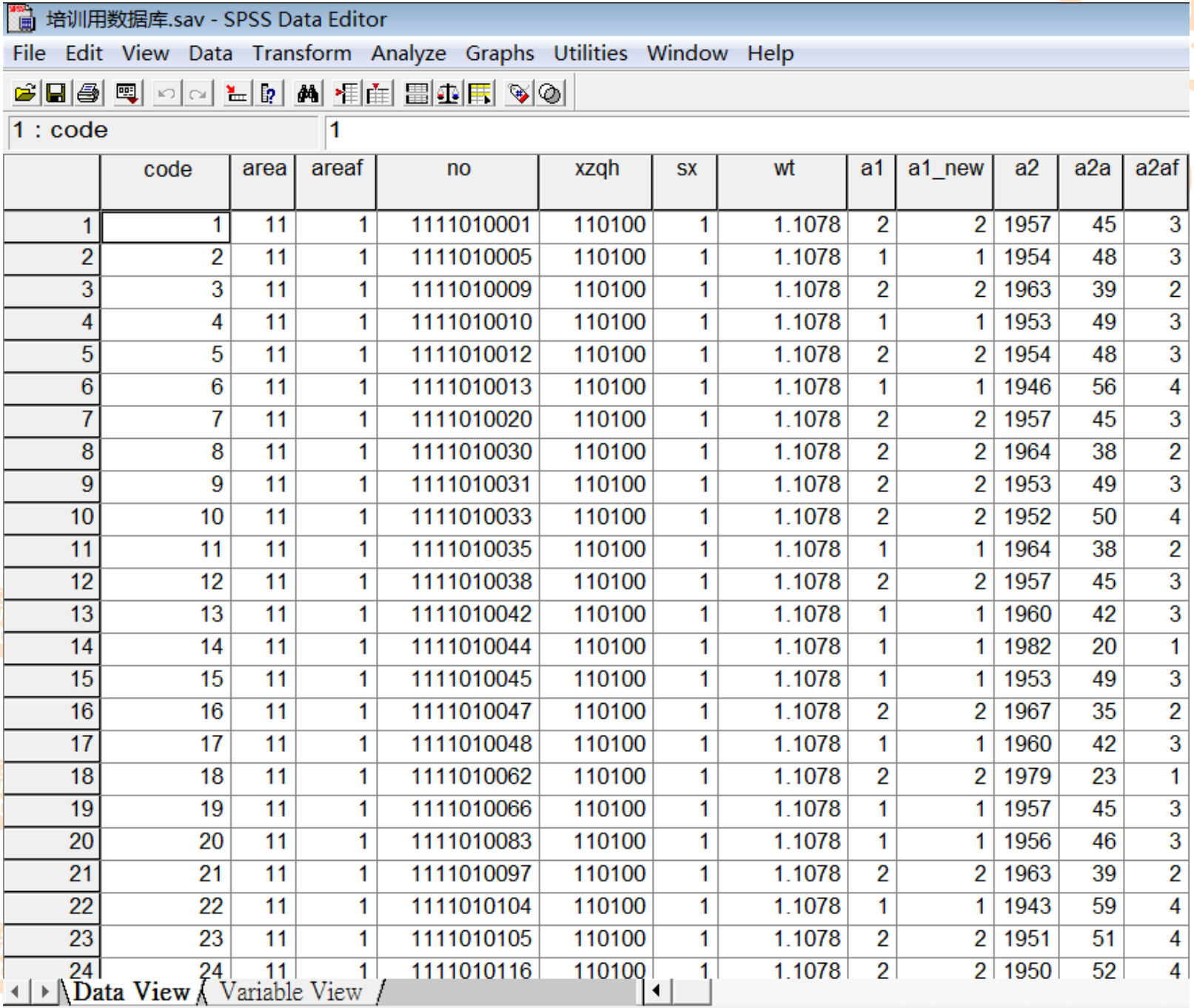
主菜单

- 菜单引导
  - 与Windows Office其他软件类似
- 10个菜单：



数据编辑窗口

- 数据显示窗口
- 变量显示窗口



	code	area	areaf	no	xzqh	sx	wt	a1	a1_new	a2	a2a	a2af
1	1	11	1	1111010001	110100	1	1.1078	2	2	1957	45	3
2	2	11	1	1111010005	110100	1	1.1078	1	1	1954	48	3
3	3	11	1	1111010009	110100	1	1.1078	2	2	1963	39	2
4	4	11	1	1111010010	110100	1	1.1078	1	1	1953	49	3
5	5	11	1	1111010012	110100	1	1.1078	2	2	1954	48	3
6	6	11	1	1111010013	110100	1	1.1078	1	1	1946	56	4
7	7	11	1	1111010020	110100	1	1.1078	2	2	1957	45	3
8	8	11	1	1111010030	110100	1	1.1078	2	2	1964	38	2
9	9	11	1	1111010031	110100	1	1.1078	2	2	1953	49	3
10	10	11	1	1111010033	110100	1	1.1078	2	2	1952	50	4
11	11	11	1	1111010035	110100	1	1.1078	1	1	1964	38	2
12	12	11	1	1111010038	110100	1	1.1078	2	2	1957	45	3
13	13	11	1	1111010042	110100	1	1.1078	1	1	1960	42	3
14	14	11	1	1111010044	110100	1	1.1078	1	1	1982	20	1
15	15	11	1	1111010045	110100	1	1.1078	1	1	1953	49	3
16	16	11	1	1111010047	110100	1	1.1078	2	2	1967	35	2
17	17	11	1	1111010048	110100	1	1.1078	1	1	1960	42	3
18	18	11	1	1111010062	110100	1	1.1078	2	2	1979	23	1
19	19	11	1	1111010066	110100	1	1.1078	1	1	1957	45	3
20	20	11	1	1111010083	110100	1	1.1078	1	1	1956	46	3
21	21	11	1	1111010097	110100	1	1.1078	2	2	1963	39	2
22	22	11	1	1111010104	110100	1	1.1078	1	1	1943	59	4
23	23	11	1	1111010105	110100	1	1.1078	2	2	1951	51	4
24	24	11	1	1111010116	110100	1	1.1078	2	2	1950	52	4

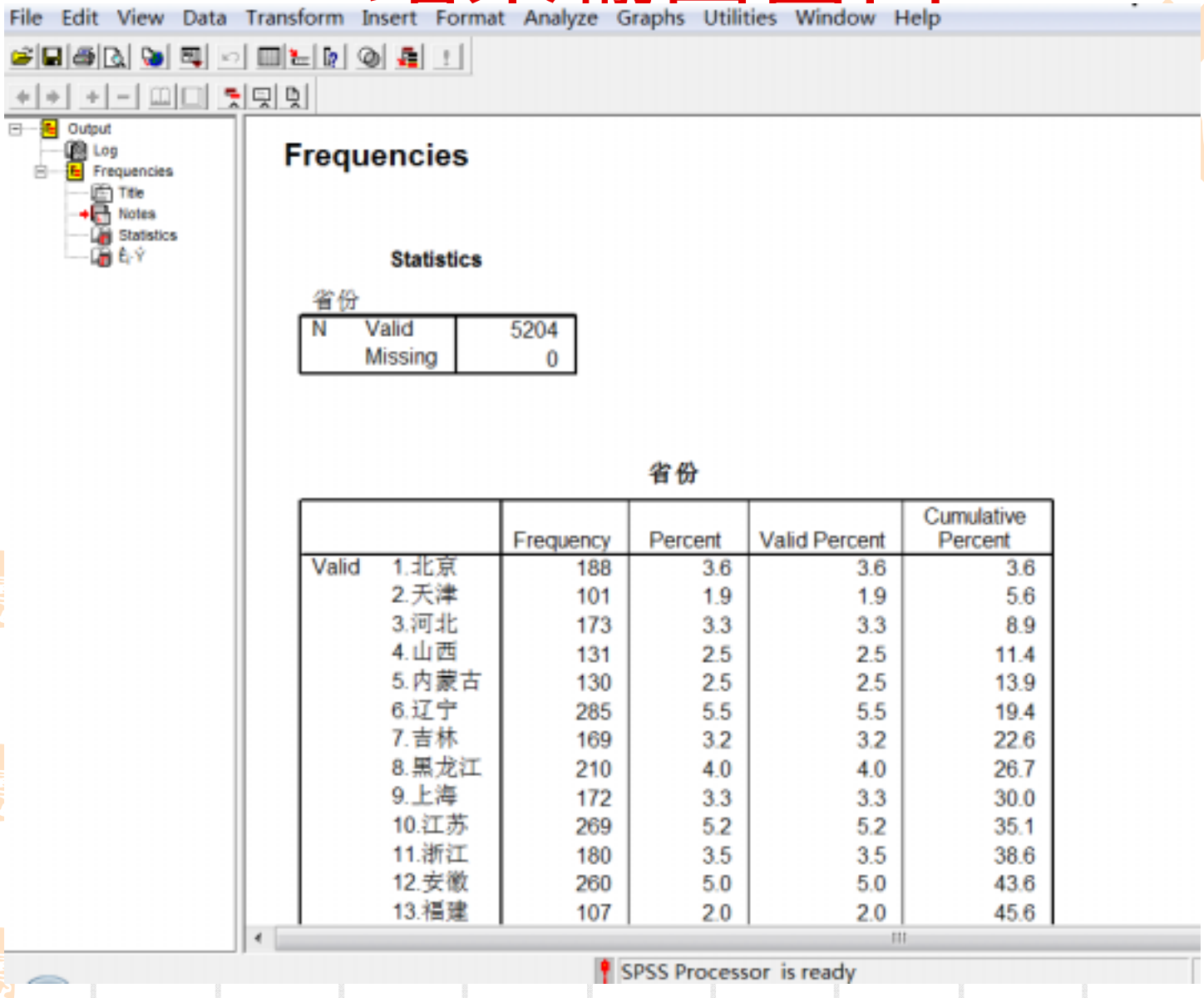
File Edit View Data Transform Analyze Graphs Utilities Window Help

	Name	Type	Width	Decimals	Label	Values	Missing
1	code	Numeric	8	0		None	None
2	area	Numeric	2	0	省份	{11, 1.北京}...	None
3	areaf	Numeric	8	0	区域划分	{1, 1.东部地区}...	None
4	no	Numeric	11	0	问卷编号	None	None
5	xzqh	Numeric	6	0	行政区划代码	None	None
6	sx	Numeric	1	0	地(市)/县(市)	{1, 地市}...	None
7	wt	Numeric	8	4	分地县权数	None	None
8	a1	Numeric	11	0	您的性别	{1, 男}...	None
9	a1_new	Numeric	11	0	您的性别	{1, 男}...	None
10	a2	Numeric	11	0	您的出生年份	None	None
11	a2a	Numeric	2	0	年龄	None	None
12	a2af	Numeric	2	0	年龄分组	{1, <=29岁}...	None
13	a2an	Numeric	8	0	年龄分组	{1, 16-25岁}...	None
14	a2a_new	Numeric	8	0		{1, 1.35岁及以下	None
15	a3	Numeric	11	0	您目前的户口情况	{1, 1.本地非农业	None
16	a3_new	Numeric	11	0	您目前的户口情况	{1, 1.本地非农业	None
17	a3f	Numeric	8	0	您目前的户口情况	{1, 非农业户口}..	None
18	a4	Numeric	11	0	您目前的婚姻状况	{1, 1.未婚}...	None
19	a5	Numeric	11	0	您目前的健康状况	{1, 1.良好}...	None
20	a6	Numeric	11	0	您目前的政治面貌	{1, 1.中共党员}...	None
21	a7	Numeric	11	0	您目前的宗教信仰	{1, 不信仰任何宗	None
22	a8	Numeric	11	0	您的民族	{1, 汉族}...	None
23	a9	Numeric	11	0	您目前的文化程度	{1, 不识字或识字	None
24	a9n	Numeric	8	0	受教育年限	None	None
25	a9f	Numeric	8	0	文化程度(归类)	{1, 初等文化程度	None
26	a10	Numeric	11	0	您目前的就业状况	{1, 有工作}	None

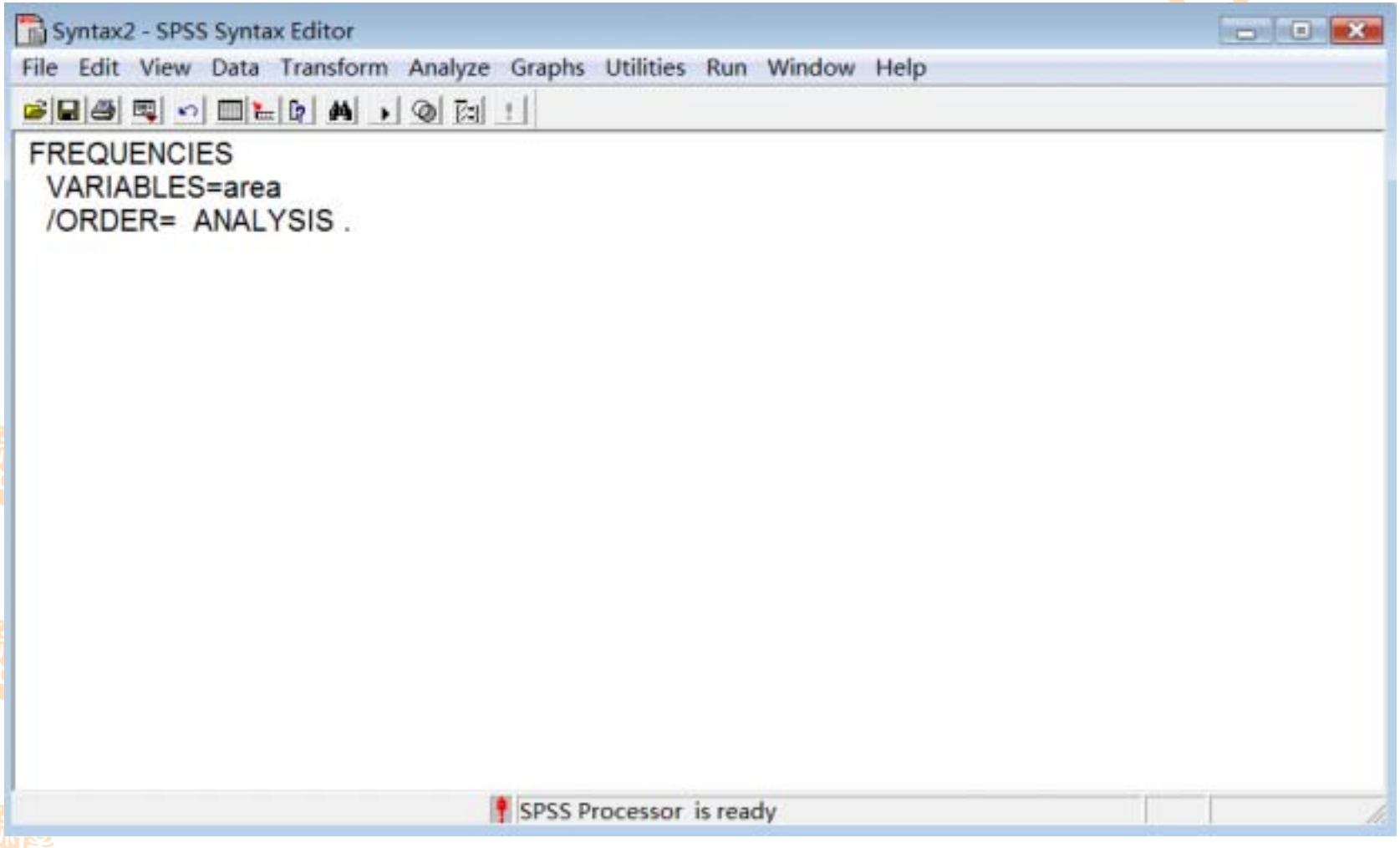
Data View | Variable View

SPSS Processor is ready

结果输出窗口


File Edit View Data Transform Insert Format Analyze Graphs Utilities Window Help
Output
Log
Frequencies
Title
Notes
Statistics
Frequencies
Statistics
省份
N Valid 5204
Missing 0
省份
Frequency Percent Valid Percent Cumulative Percent
Valid 1.北京 188 3.6 3.6 3.6
2.天津 101 1.9 1.9 5.6
3.河北 173 3.3 3.3 8.9
4.山西 131 2.5 2.5 11.4
5.内蒙古 130 2.5 2.5 13.9
6.辽宁 285 5.5 5.5 19.4
7.吉林 169 3.2 3.2 22.6
8.黑龙江 210 4.0 4.0 26.7
9.上海 172 3.3 3.3 30.0
10.江苏 269 5.2 5.2 35.1
11.浙江 180 3.5 3.5 38.6
12.安徽 260 5.0 5.0 43.6
13.福建 107 2.0 2.0 45.6
SPSS Processor is ready

命令编辑窗口


Syntax2 - SPSS Syntax Editor
File Edit View Data Transform Analyze Graphs Utilities Run Window Help
FREQUENCIES
VARIABLES=area
/ORDER= ANALYSIS .
SPSS Processor is ready

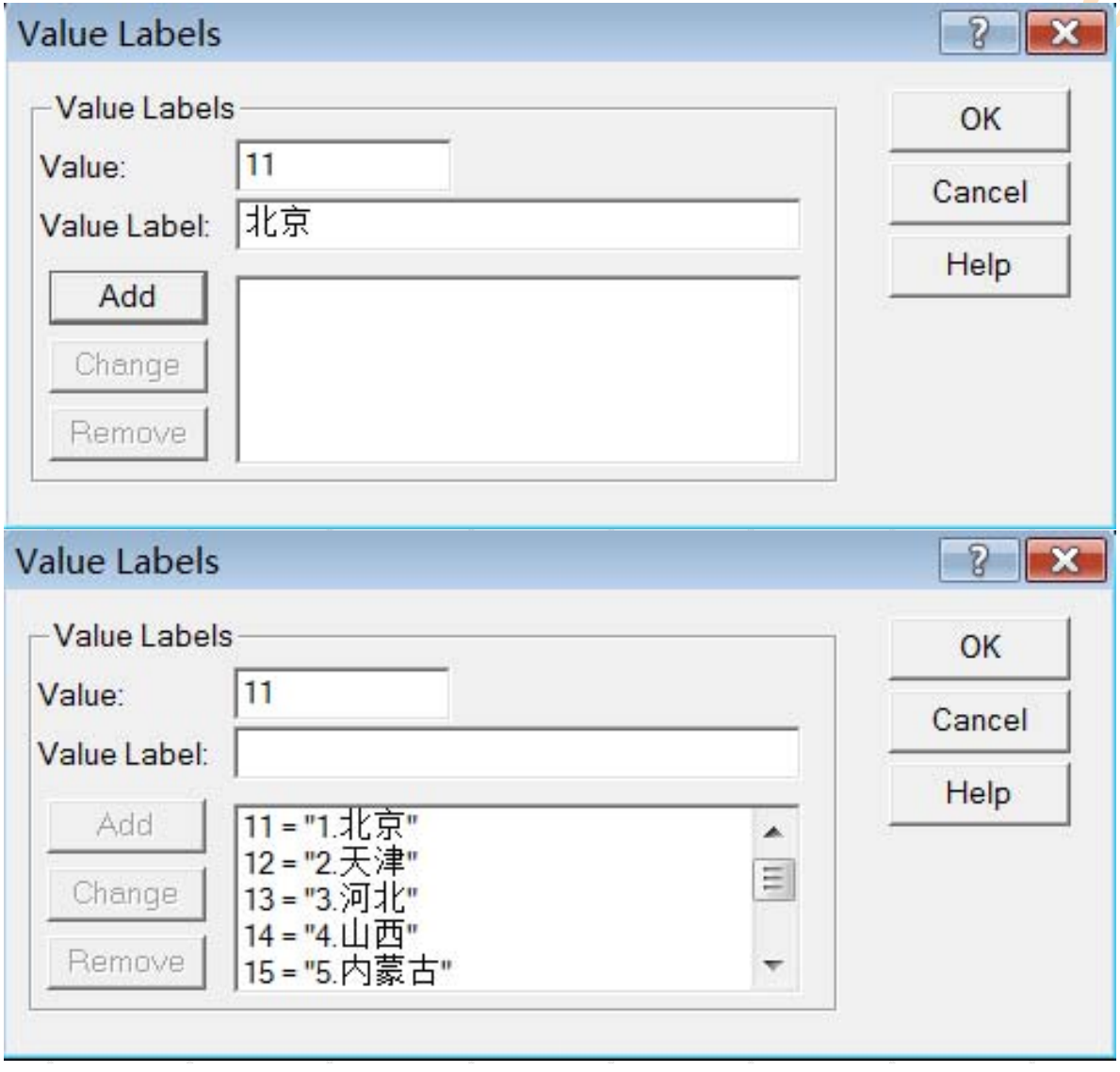
数据输入

- 可以直接输入
- 可以从Excel导入
- 使用数据录入软件，例如EpiData，可以免费下载 http://www.epidata.dk/download.php

变量定义

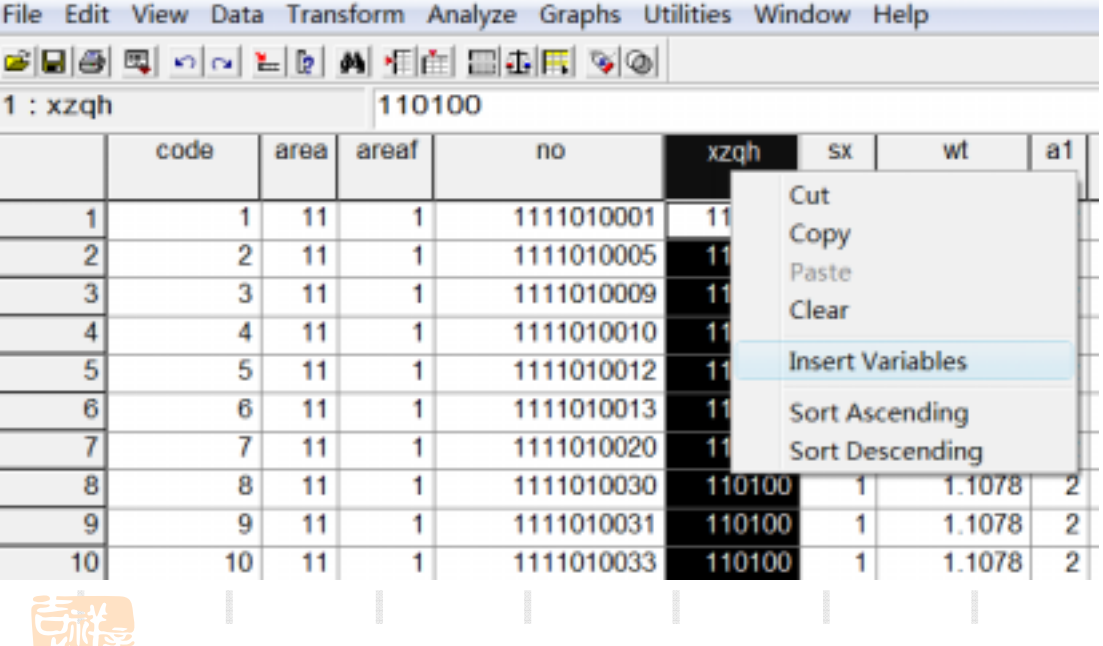
- 变量名
  - 以英文字母开头，中文也可以
  - 不超过64个字符
  - 唯一性
  - 空格或特殊符号不能使用
- 变量标签
- 变量值标签

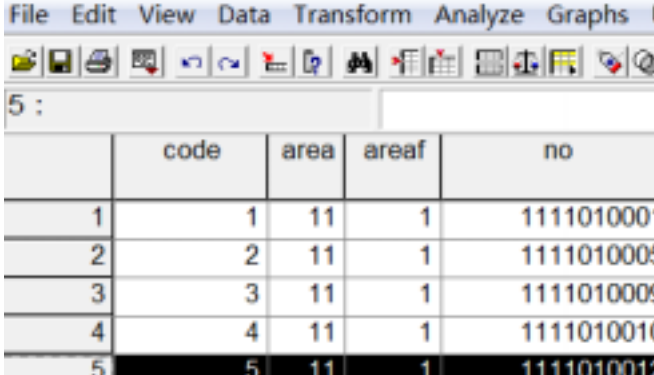
变量值标签


Value Labels
Value Labels
Value:
11
Value Label:
北京
Add
Change
Remove
OK
Cancel
Help
Value Labels
Value Labels
Value:
11
Value Label:
Add
Change
Remove
11 = "1.北京"
12 = "2.天津"
13 = "3.河北"
14 = "4.山西"
15 = "5.内蒙古"
OK
Cancel
Help

变量（案例）的增删

- 增加 Insert variables (cases)
- 删除





数据保存、输出

- 保存Save、另存为Save As
  - 经常需要Save As
- 输出为其他类型格式的数据
- 不同类型格式数据之间的转换

使用软件Stat/Transfer

作业

- 从本培训班收集以下数据:
  (1) 姓名
  (2) 性别
  (3) 文化程度
  (4) 家乡与北京的距离

进行以下联系

- 将数据输入 Excel
- 导入 SPSS
- 定义变量
- 保存SPSS数据
- 计算平均距离
- 退出 SPSS

第二讲 单变量分析

- 中心趋势测量
  （平均数、中位数）
- 离散程度测量
  （方差、标准差）

引言

- 频数分布和绘图是数据分析最基本但很有用的方法。对某个变量的总体情况进行了解，不能准确刻画变量的特征，因此，需要计算一些指标来反映变量的特征。
- 这些指标包括测量变量值的平均水平和变量分布的离散程度

中心趋势测量

- 对于某一变量，其值的平均水平或代表性值
- 常用的有两个指标：中位数（Median）和平均数（Mean）

中位数

- 把一个变量的值由小到大或由大到小排列起来，处于中心的那个值就是中位数。即中位数将变量的分布分成前后相等的两部分，其中一半的值低于中位数，另一半的值高于中位数。
- 中位数适用于序次变量和间距变量

平均数

- 简单算术平均数是使用最广泛的平均数。其计算方法就是把所有案例的该变量值都加起来，然后除以案例数。
- 平均数只适用于间距变量。

比较平均数和中位数的变化

4 8 12

平均数=(4+8+12)/3=8

中位数=8

4 8 120

平均数=(4+8+120)/3=44

中位数=8

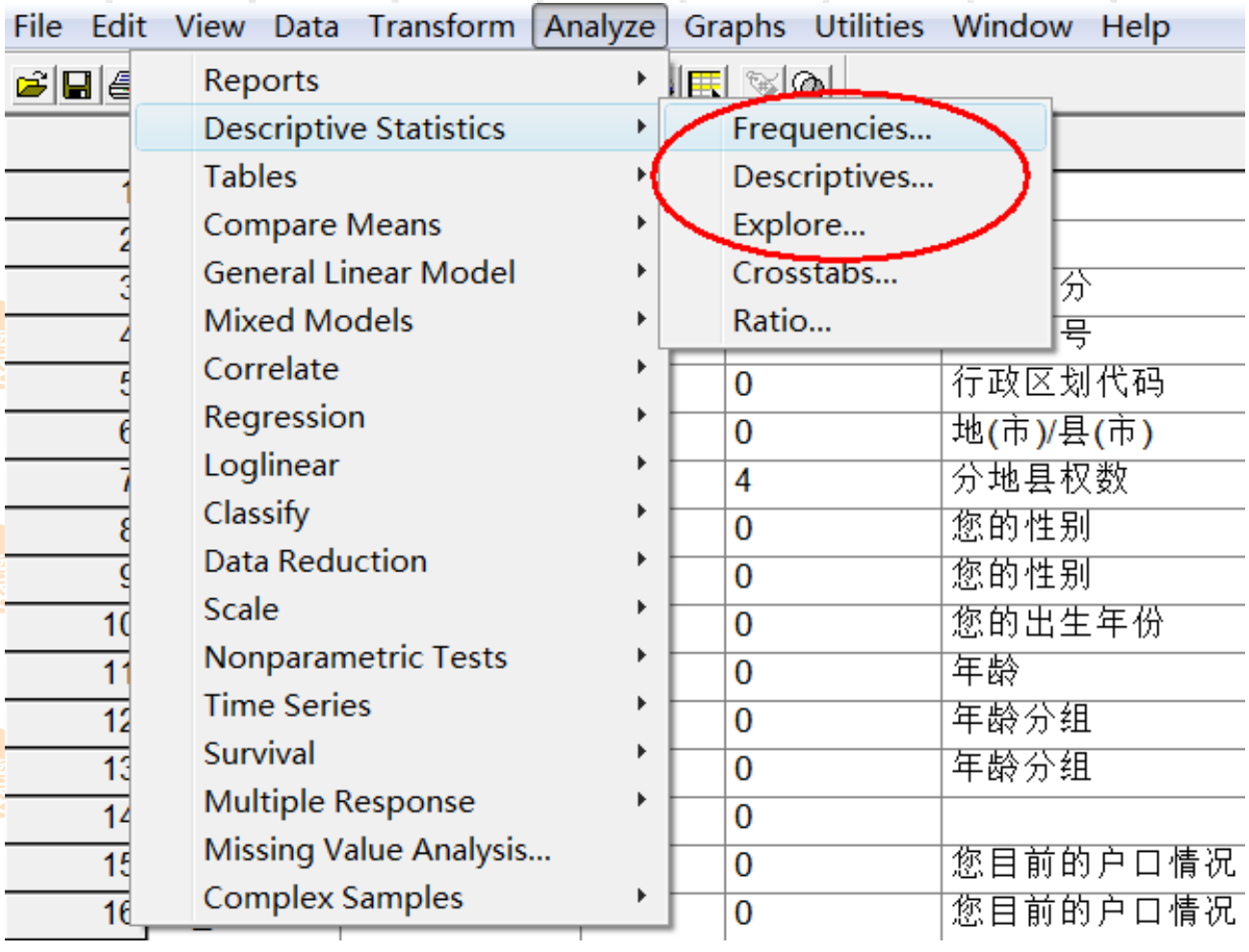
- 当平均数比较接近时，标准差的大小基本反映了差异的大小。即标准差大的，差异大；标准差小的，差异小。
- 当平均数有较大差异时，标准差大小本身不能准确说明差异大小。

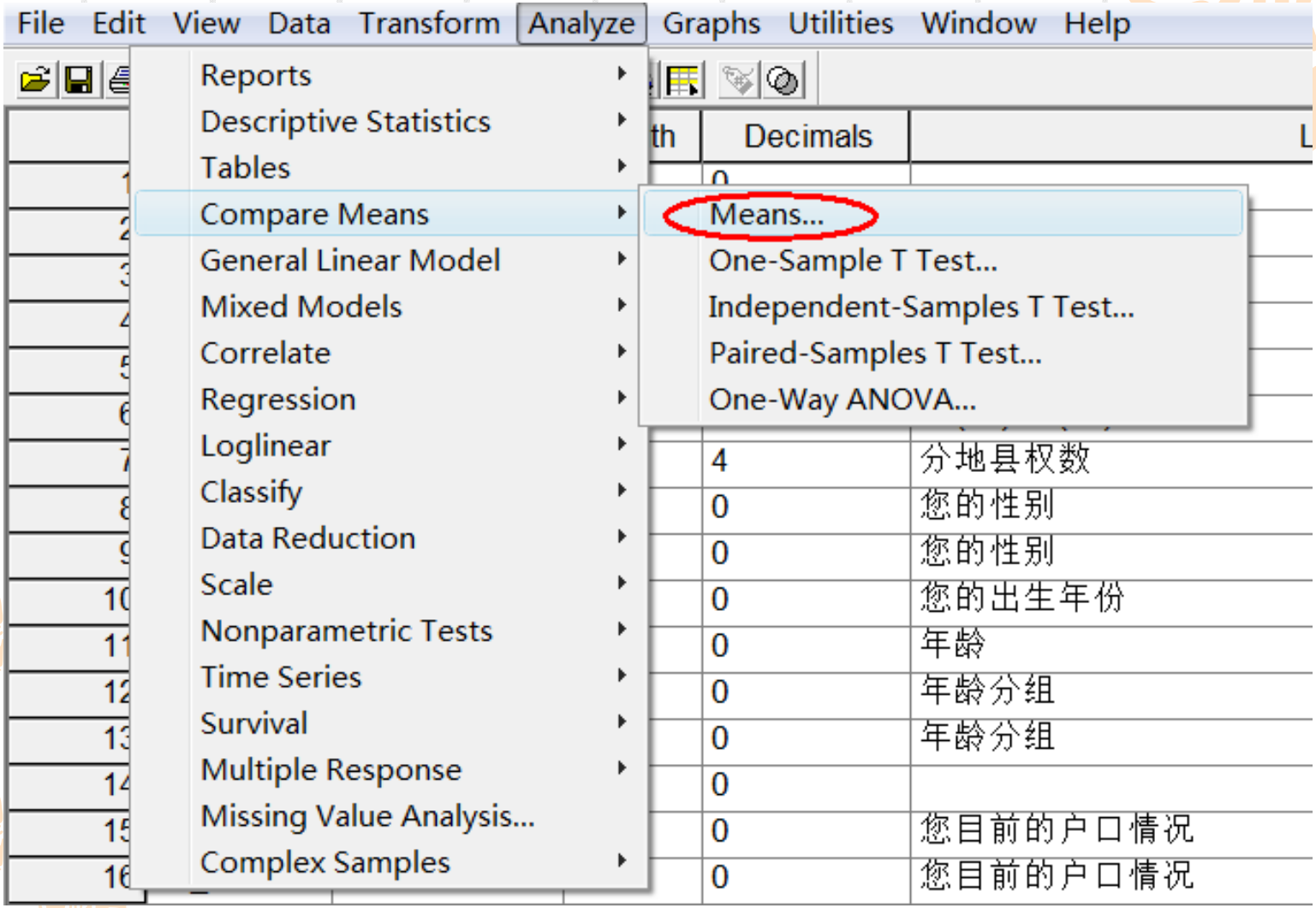
男女收入

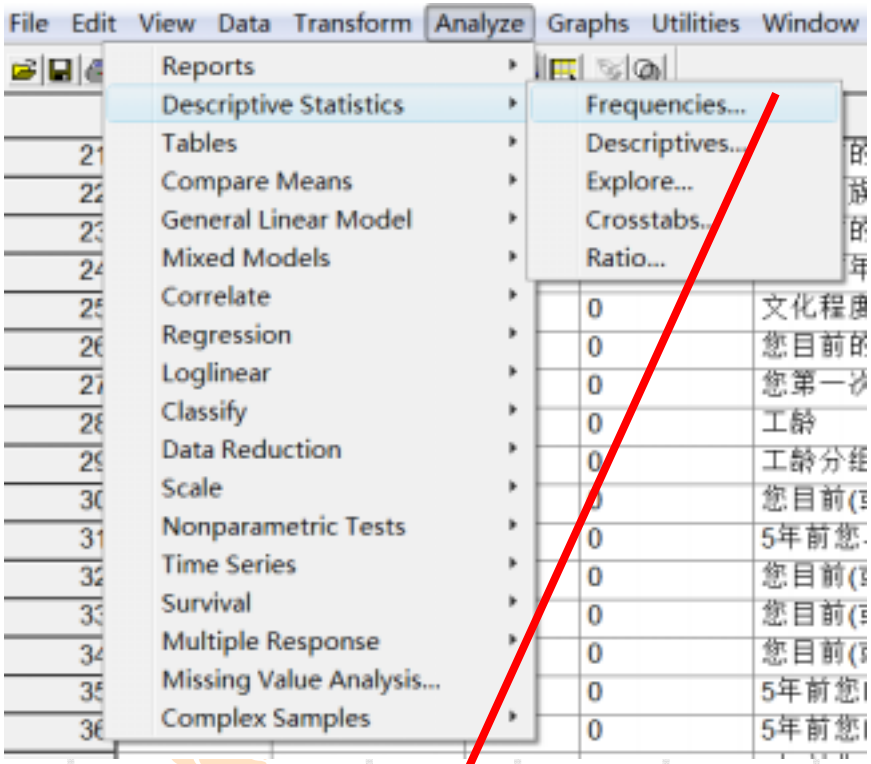
性别	平均数	样本量	标准差	离散系数
男	1117.950	2446	743.267	0.665
女	934.794	2024	713.996	0.764
合计	1035.017	4470	735.749	0.711

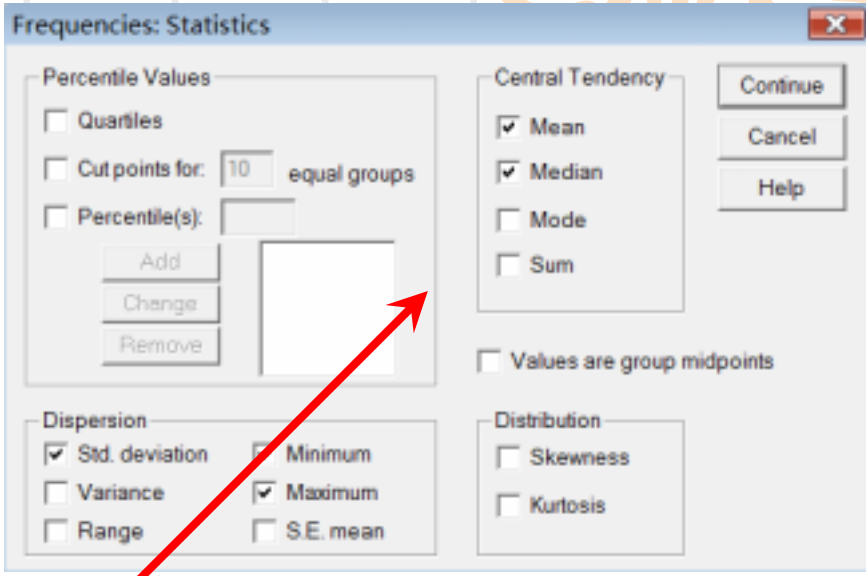
- 男性收入标准差更大，但是平均数也更大，男性收入离散系数小于女性，说明男性的收入差异小于女性。

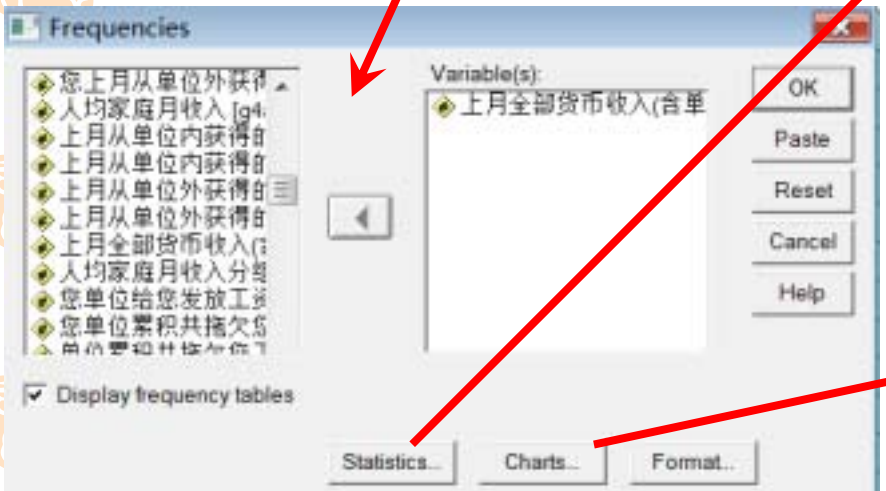
SPSS在下列功能块中都可以计算平均数、标准差

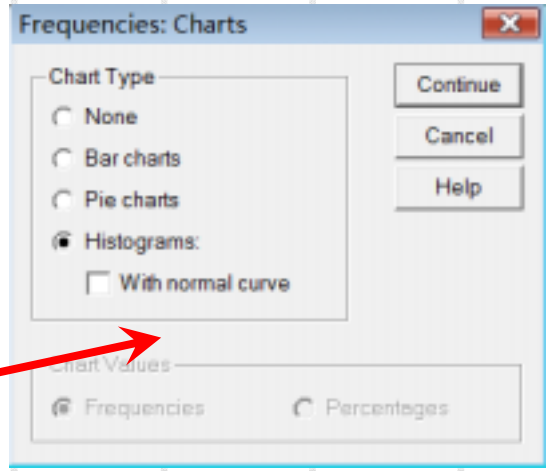


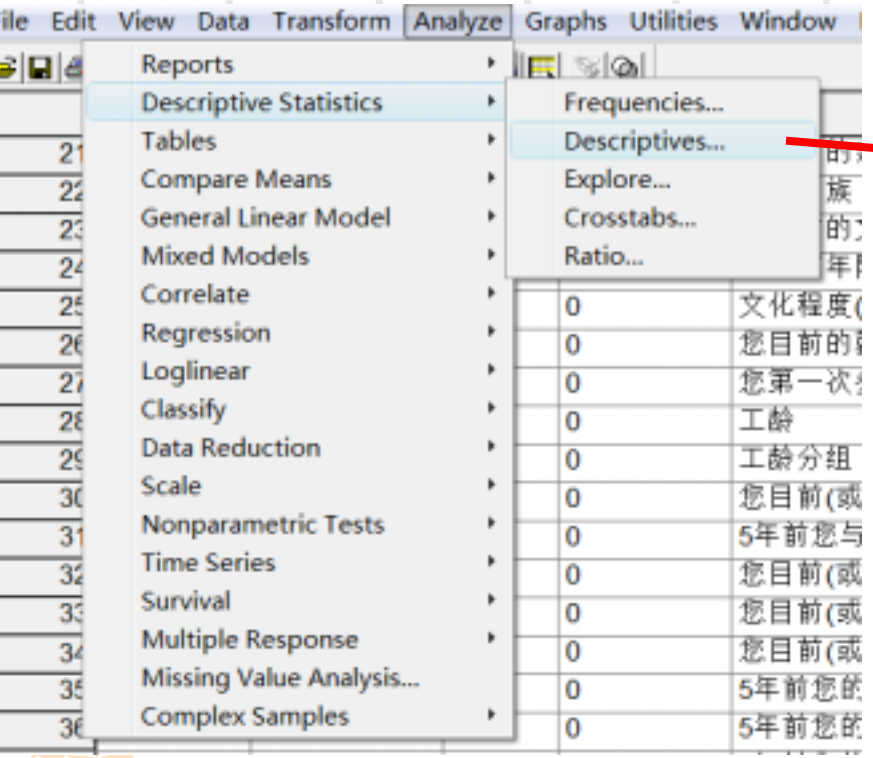

File Edit View Data Transform Analyze Graphs Utilities Window Help
Reports
Descriptive Statistics
Tables
Compare Means
General Linear Model
Mixed Models
Correlate
Regression
Loglinear
Classify
Data Reduction
Scale
Nonparametric Tests
Time Series
Survival
Multiple Response
Missing Value Analysis...
Complex Samples
Means...
One-Sample T Test...
Independent-Samples T Test...
Paired-Samples T Test...
One-Way ANOVA...
th
Decimals
L
0
4 分地县权数
0 您的性别
0 您的性别
0 您的出生年份
0 年龄
0 年龄分组
0 年龄分组
0
0 您目前的户口情况
0 您目前的户口情况


File Edit View Data Transform Analyze Graphs Utilities Window
Reports
Descriptive Statistics
Tables
Compare Means
General Linear Model
Mixed Models
Correlate
Regression
Loglinear
Classify
Data Reduction
Scale
Nonparametric Tests
Time Series
Survival
Multiple Response
Missing Value Analysis...
Complex Samples
Frequencies...
Descriptives...
Explore...
Crosstabs...
Ratio...

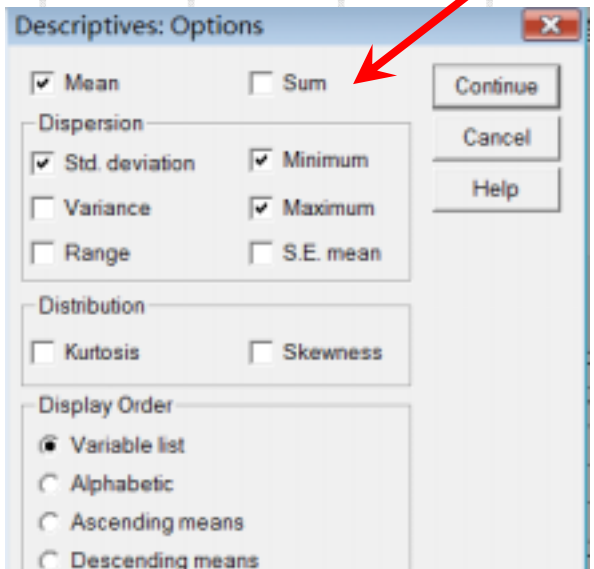

Frequencies: Statistics
Percentile Values
Quartiles
Cut points for: 10 equal groups
Percentile(s):
Add
Change
Remove
Central Tendency
Mean
Median
Mode
Sum
Continue
Cancel
Help
Values are group midpoints
Dispersion
Std. deviation
Minimum
Variance
Maximum
Range
S.E. mean
Distribution
Skewness
Kurtosis

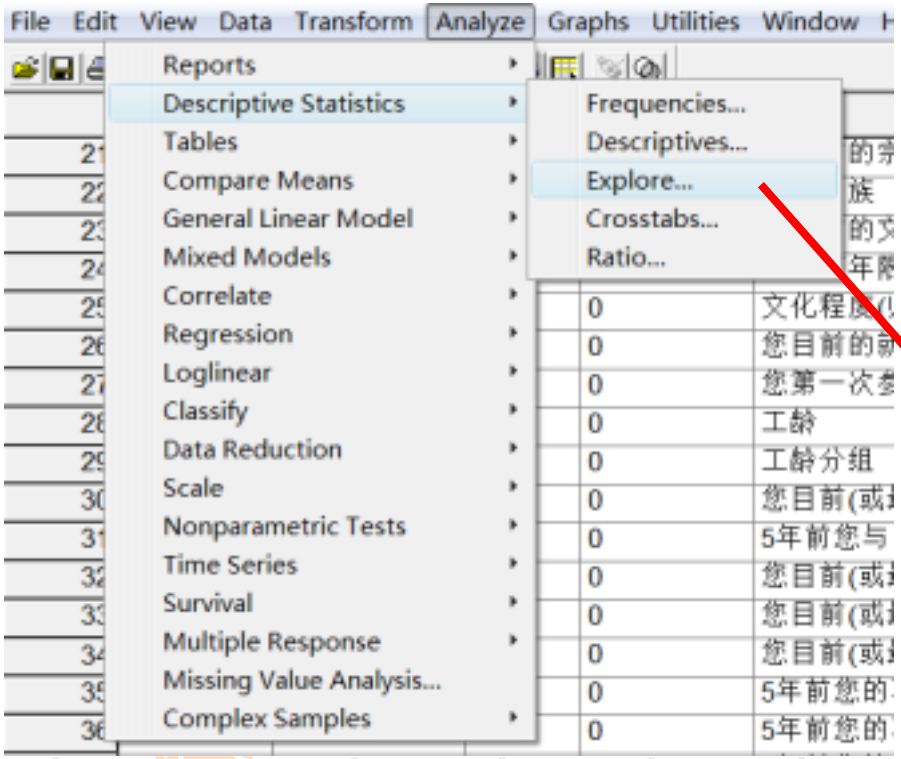

Frequencies
Variable(s):
上月全部货币收入(含单
OK
Paste
Reset
Cancel
Help
Display frequency tables
Statistics...
Charts...
Format...

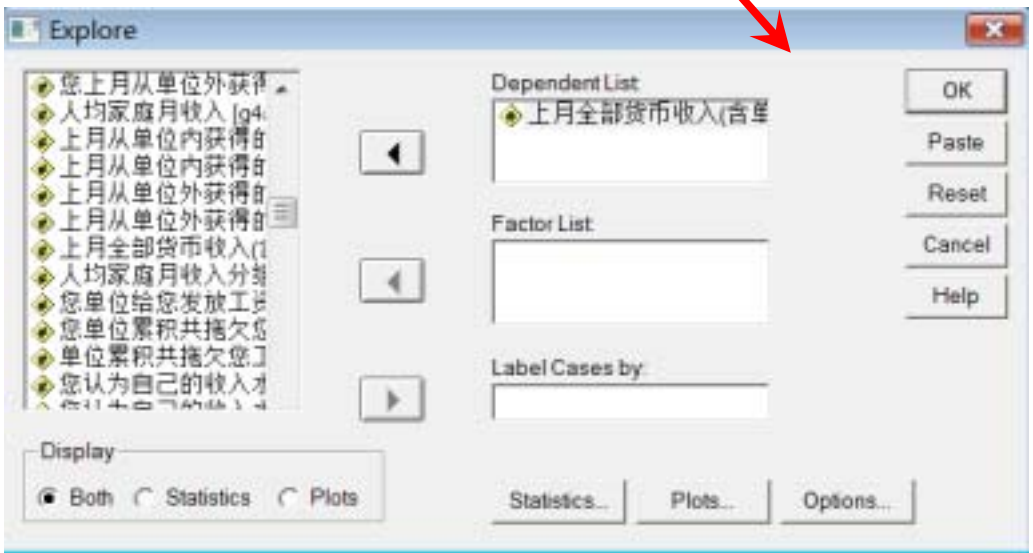

Frequencies: Charts
Chart Type
None
Bar charts
Pie charts
Histograms:
With normal curve
Continue
Cancel
Help
Chart Values
Frequencies
Percentages


File Edit View Data Transform Analyze Graphs Utilities Window
Reports
Descriptive Statistics
Tables
Compare Means
General Linear Model
Mixed Models
Correlate
Regression
Loglinear
Classify
Data Reduction
Scale
Nonparametric Tests
Time Series
Survival
Multiple Response
Missing Value Analysis...
Complex Samples
Frequencies...
Descriptives...
Explore...
Crosstabs...
Ratio...
文化程度(
您目前的
您第一次
工龄
工龄分组
您目前(或
5年前您与
您目前(或
您目前(或
您目前(或
5年前您的
5年前您的

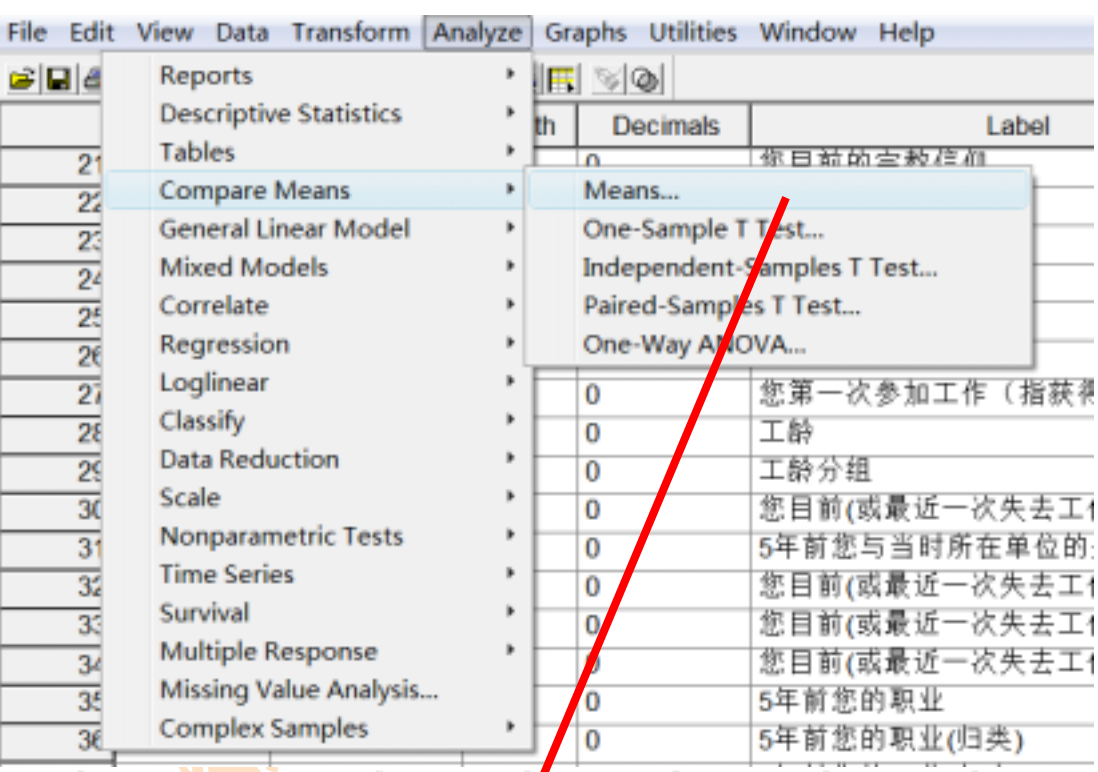

Descriptives
经济补偿总金额分组
您与单位协议解除劳
您与单位协议解除劳
您上月从本单位（或
单位内月收入五等分
其中：工资(含基本工
您上月从单位外获得
人均家庭月收入 [q4a
上月从单位内获得的
Variable(s):
上月全部货币收入(含单
OK
Paste
Reset
Cancel
Help
Save standardized values as variables
Options...

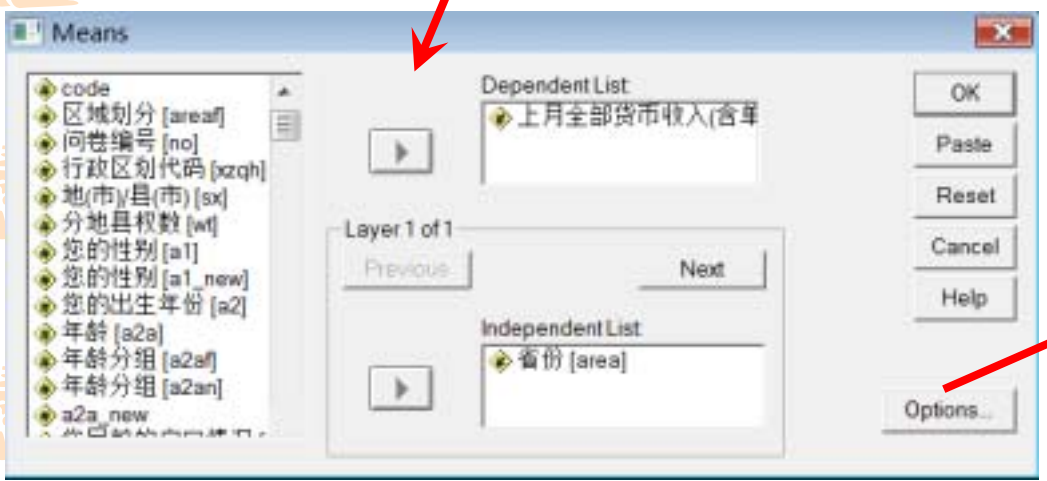

Descriptives: Options
Mean
Sum
Continue
Cancel
Help
Dispersion
Std. deviation
Minimum
Variance
Maximum
Range
S.E. mean
Distribution
Kurtosis
Skewness
Display Order
Variable list
Alphabetic
Ascending means
Descending means

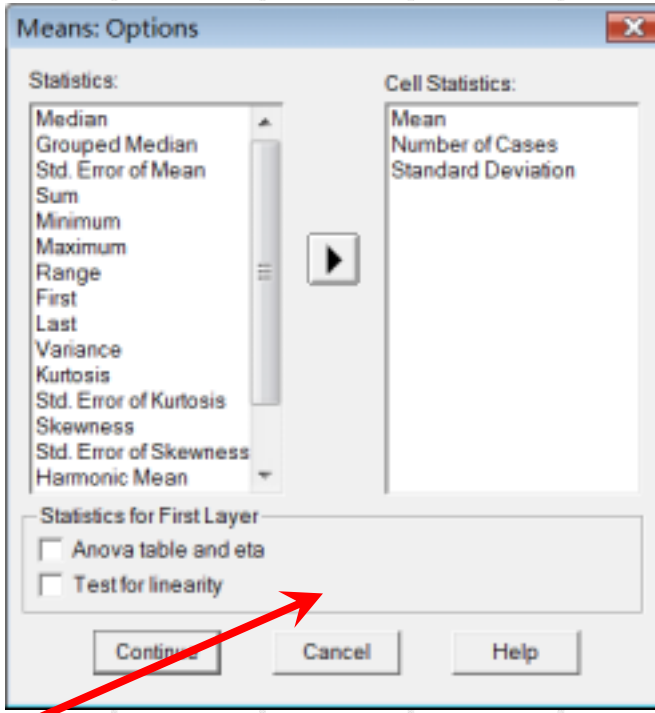

File Edit View Data Transform Analyze Graphs Utilities Window
Reports
Descriptive Statistics
Tables
Compare Means
General Linear Model
Mixed Models
Correlate
Regression
Loglinear
Classify
Data Reduction
Scale
Nonparametric Tests
Time Series
Survival
Multiple Response
Missing Value Analysis...
Complex Samples
Frequencies...
Descriptives...
Explore...
Crosstabs...
Ratio...


Explore
Dependent List:
Factor List:
Label Cases by:
OK
Paste
Reset
Cancel
Help
Display
Both
Statistics
Plots
Statistics...
Plots...
Options...

计算各省的平均收入


File Edit View Data Transform Analyze Graphs Utilities Window Help
Reports
Descriptive Statistics
Tables
Compare Means
General Linear Model
Mixed Models
Correlate
Regression
Loglinear
Classify
Data Reduction
Scale
Nonparametric Tests
Time Series
Survival
Multiple Response
Missing Value Analysis...
Complex Samples
Means...
One-Sample T Test...
Independent-Samples T Test...
Paired-Samples T Test...
One-Way ANOVA...
Decimals
Label
0 您第一次参加工作（指获得
0 工龄
0 工龄分组
0 您目前(或最近一次失去工
0 5年前您与当时所在单位的
0 您目前(或最近一次失去工
0 您目前(或最近一次失去工
0 您目前(或最近一次失去工
0 5年前您的职业
0 5年前您的职业(归类)


Means
code
区域划分 [areaf]
问卷编号 [no]
行政区划代码 [xzqh]
地(市)/县(市) [sx]
分地县权数 [wt]
您的性别 [a1]
您的性别 [a1_new]
您的出生年份 [a2]
年龄 [a2a]
年龄分组 [a2af]
年龄分组 [a2an]
a2a_new
Dependent List:
上月全部货币收入(含草
Layer 1 of 1
Previous
Next
Independent List:
省份 [area]
OK
Paste
Reset
Cancel
Help
Options...


Means: Options
Statistics:
Median
Grouped Median
Std. Error of Mean
Sum
Minimum
Maximum
Range
First
Last
Variance
Kurtosis
Std. Error of Kurtosis
Skewness
Std. Error of Skewness
Harmonic Mean
Cell Statistics:
Mean
Number of Cases
Standard Deviation
Statistics for First Layer
Anova table and eta
Test for linearity
Continue
Cancel
Help

第三讲 线性回归

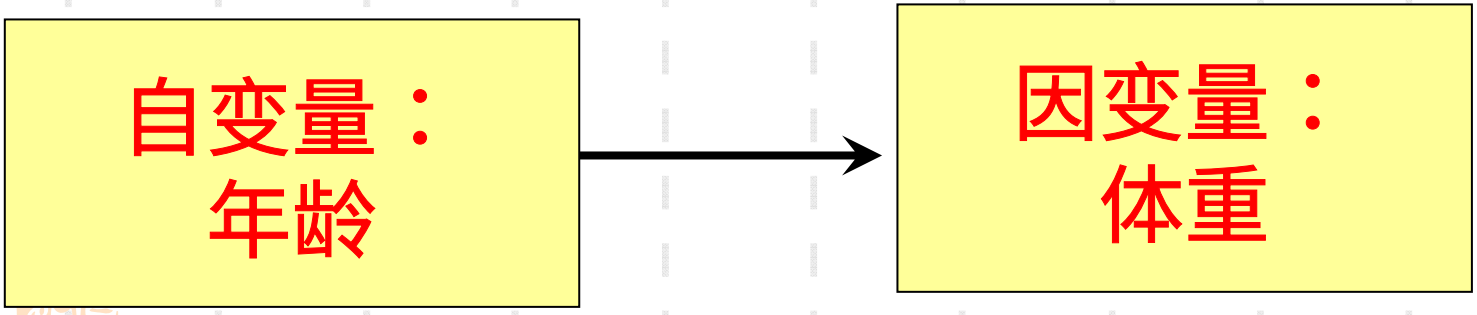
- 一元线性回归
- 相关系数
- 多元线性回归

二个间距（连续）变量的关系

- 相关和回归分析可以用来检验二个间距变量之间的关系。（例如婴儿年龄和体重）

研究问题

- 婴儿的年龄和体重是否存在关系?



- 14个婴儿的数据
- 年龄：月数
- 体重：斤

	Baby	Age	Weight
1	A	1	10.0
2	B	4	10.0
3	C	0	5.0
4	D	6	15.9
5	E	5	17.2
6	F	6	12.7
7	G	0	6.8
8	H	3	14.1
9	I	2	7.3
10	J	5	14.5
11	K	2	9.5
12	L	1	8.2
13	M	4	12.7
14	N	3	11.3

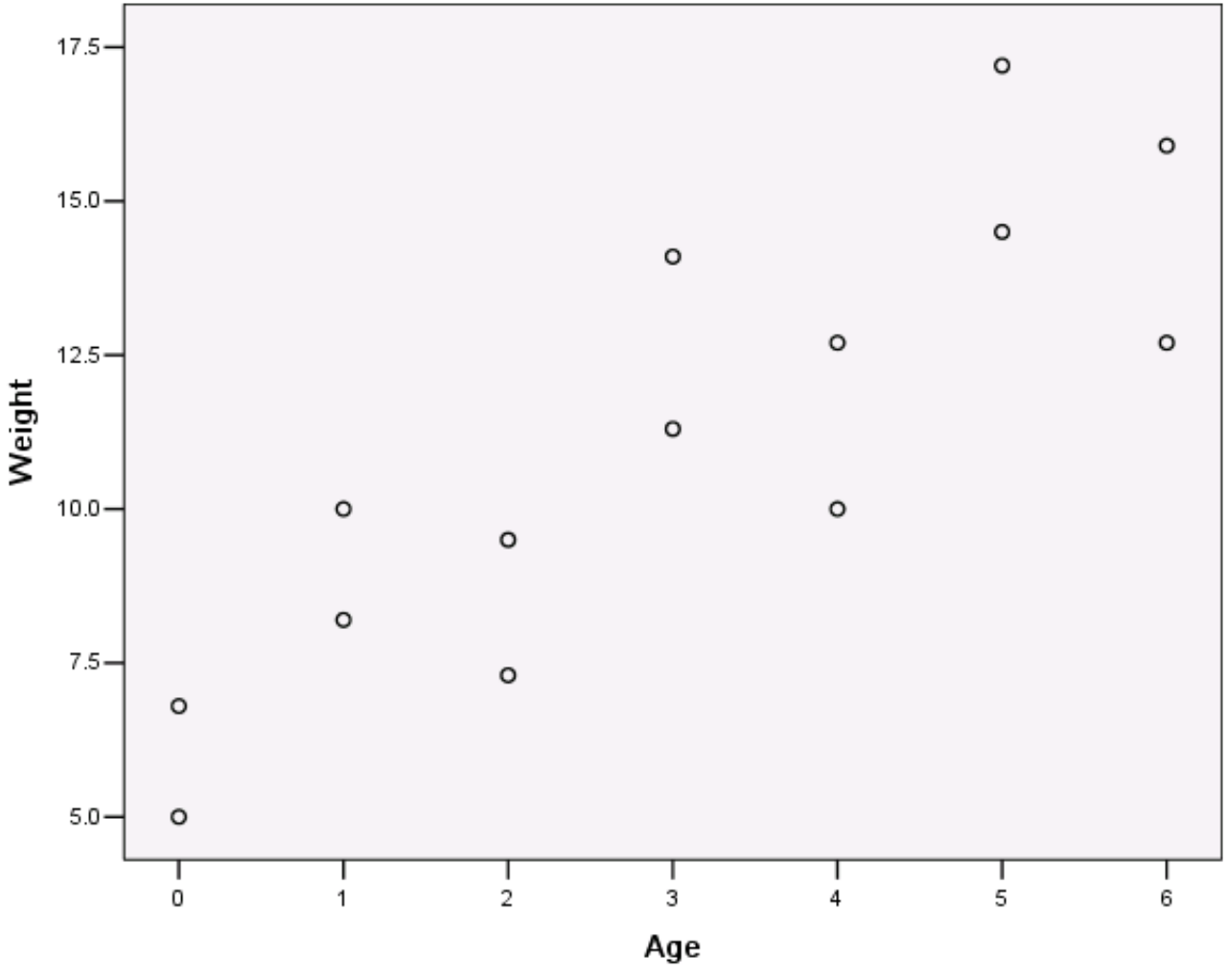
三种方式进行考察

- 散点图：图示两者的关系
- 双变量回归：用两个数概况两者的关系
- 相关系数：用一个数概括两者的关系

双变量关系的特征

- 关系是否存在？
- 关系的强度？
- 关系的方向？

婴儿年龄和体重的散点图



“读”散点图

- 负关系: 随着自变量值上升（下降），因变量值下降（上升）
- 正关系: 随着自变量值上升（下降），因变量值也上升（下降）
- 没有关系：随着自变量值增加或减少，因变量值不变

- 关系是否存在：如果这些点表现出上升或下降趋势，表明存在关系
- 关系的强度：如果这些点围绕这种上升或下降趋势而分布紧密，则存在较强的关系
- 关系的方向：趋势向上倾斜，表明存在正向关系；趋势向下倾斜，表明存在负向关系

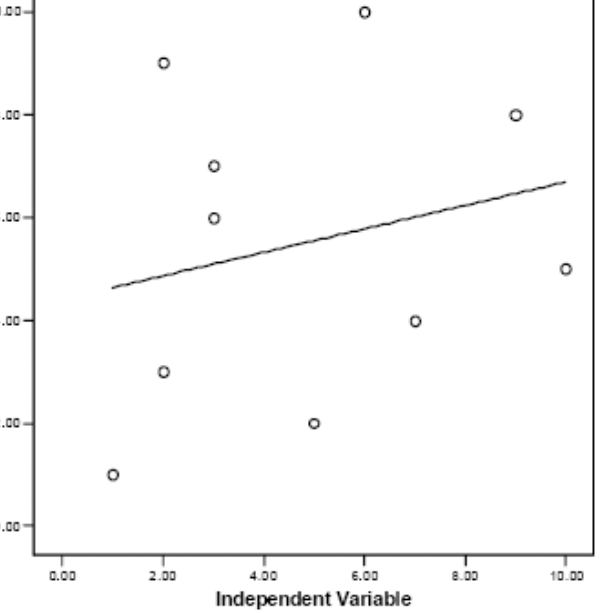


Figure 2. Weak positive relationship

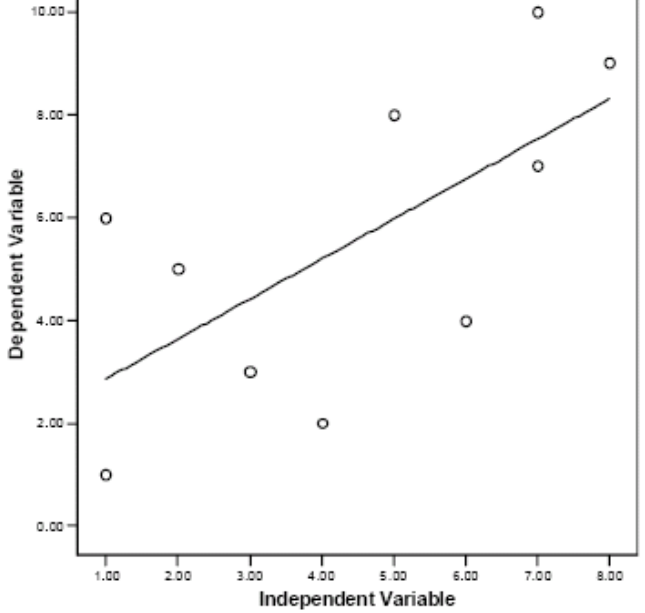


Figure 3. Strong positive relationship

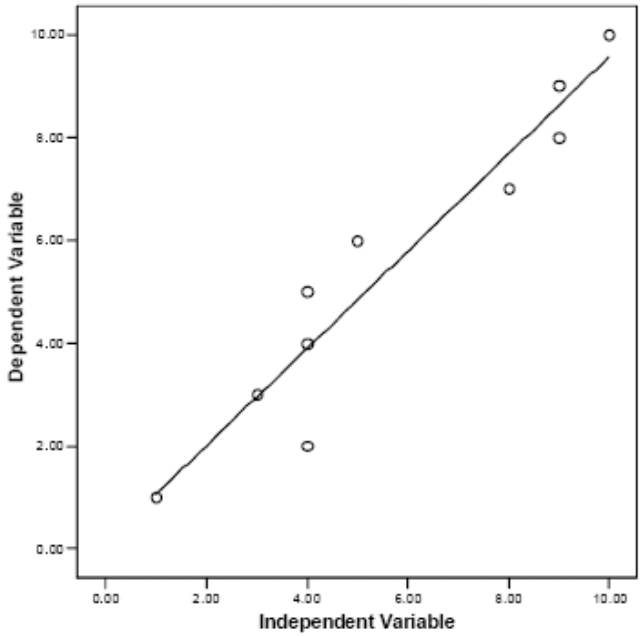


Figure 4. Very strong positive relationship

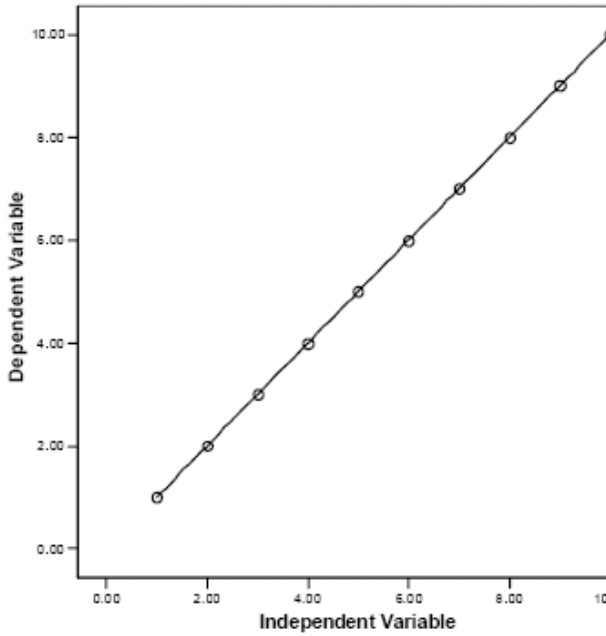


Figure 5. Perfect positive relationship

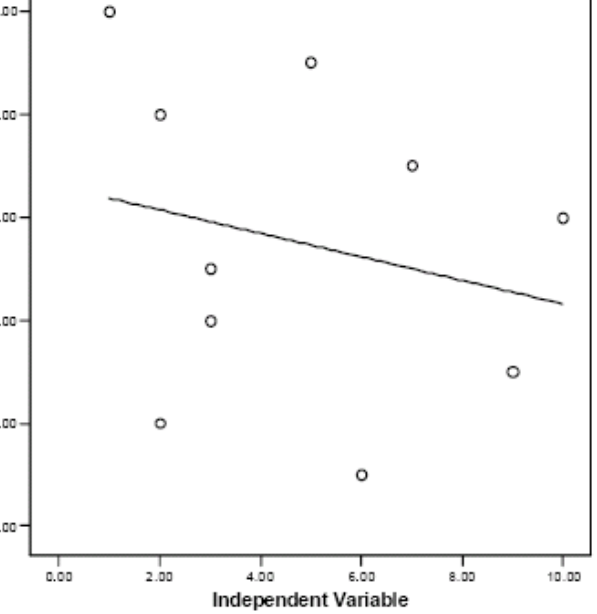


Figure 6. Weak negative relationship

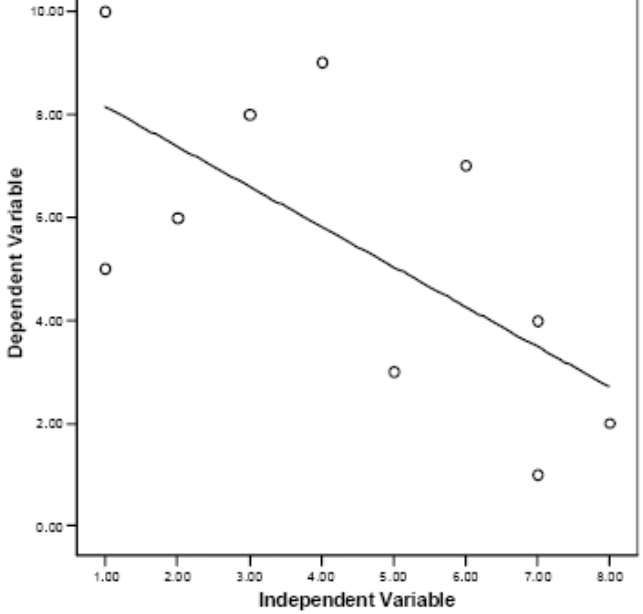


Figure 7. Strong negative relationship

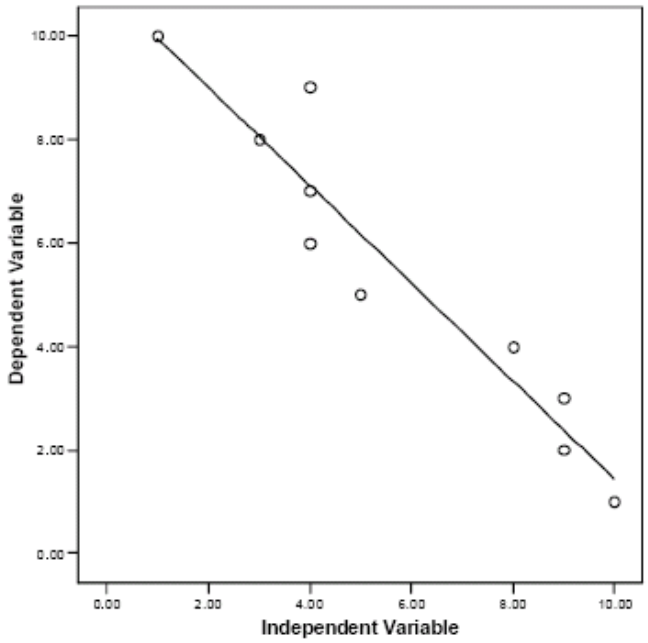


Figure 8. Very strong negative relationship

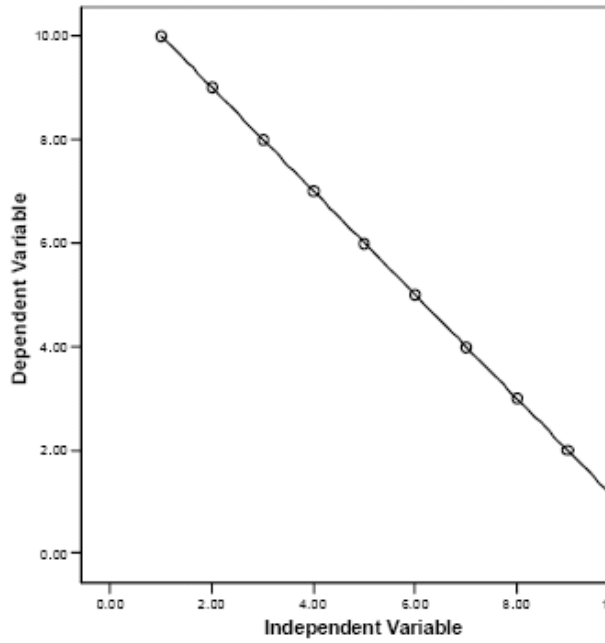


Figure 9. Perfect negative relationship

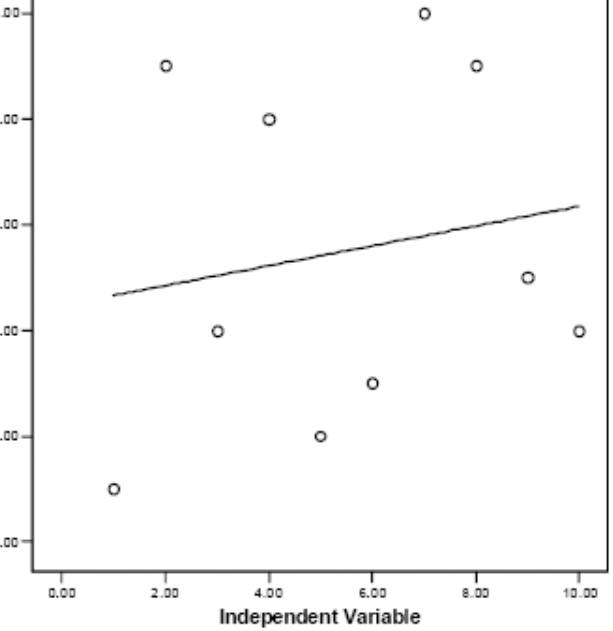


Figure 10. Almost no relationship

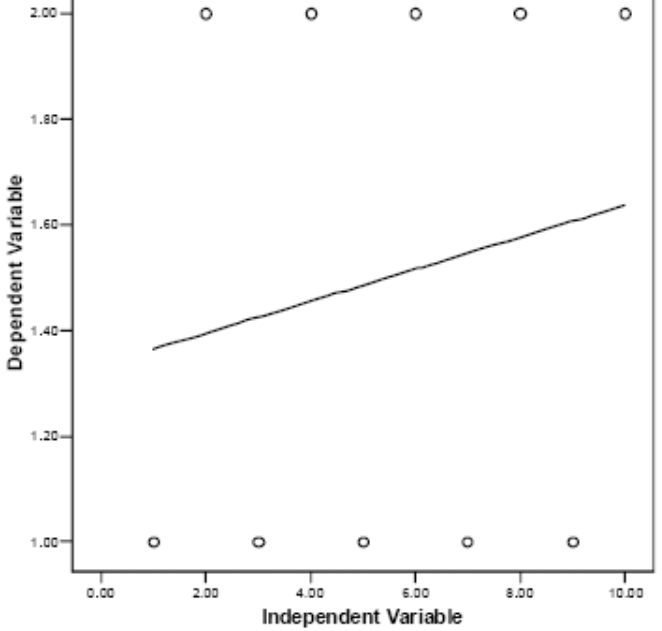
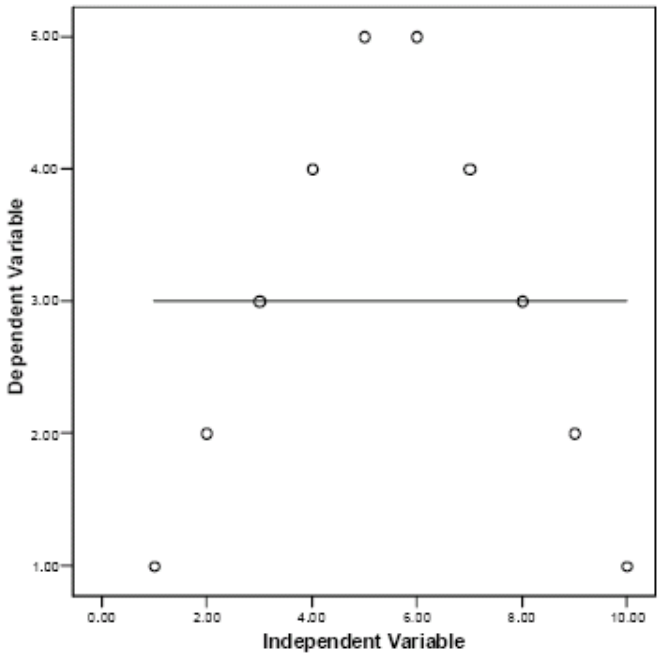
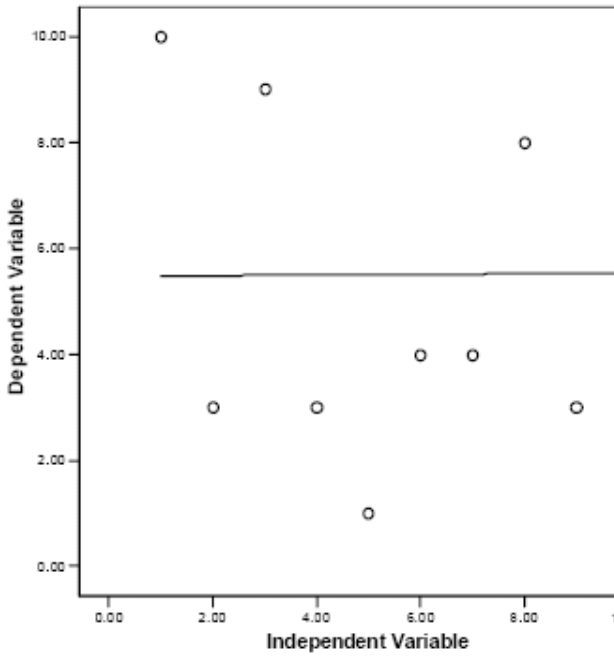


Figure 11. Almost no relationship





线性关系

- 散点图中的这些点，可以近似地用一条直线来拟合，那么这两个变量就存在一种线性关系
- 没有线性关系：这些点可以用一条水平线拟合，或者呈现曲线关系

什么是直线?

- 直线可以用下列方程来表示:

$$y = a + bx$$

– y 代表因变量，是纵轴上的值
– x 代表自变量，是横轴上的值
– a 为截距 (当 x = 0时y的值)
– b 为斜率 (当 x 变化 1 个单位, y 将变化 “b” 个单位)

寻找这条直线？

- 如果两个变量存在线性关系，我们想找到一条直线，用这条直线的方程来表示两个变量之间的关系
- 但是通过这些散点的直线，可以有很多条，那么我们如何找到最佳的那条直线？

最佳拟合线（The best-fitting line）

- 最佳拟合线就是产生误差最小的直线
- 误差：观测值与预测值之间差

$$e = Y - \hat{Y}$$

最佳拟合线

- 我们想找到一条直线能够是所有人的误差都最小。但是任何一条直线，都会使一些人的误差达到最小，而使另一些人的误差增大。因此，我们就想使所有人的误差总和达到最小。

总误差

- 统计学家使用误差平方和来计算总误差

$$\sum (Y - \hat{Y})$$

$$\sum \left| Y - \hat{Y} \right|$$

$$\sum (Y - \hat{Y})^2$$

- 将所有人的误差进行平方之后加总得到总误差：

$$\sum e^2 = \sum (Y - \hat{Y})^2$$

- 最佳拟合线就是使总误差最小的那条直线

最小二乘线（The least-squares line）

- 最佳拟合线称作最小二乘线，计算出这条直线的方法叫最小二乘法（least squares method）。也就是说要求出下列方程中的a、b值,

$$\hat{Y} = a + bX$$

使 $\sum e^2$ 达到最小值

回归方程

$$\hat{Y} = a + bX$$

- $\hat{Y}$ = Y的预测值
- X = 自变量值
- a = y截距 (X = 0时 Y的值)，有时候截距值没有什么意义（因为自变量不会是0）
- b = 斜率 (自变量X 每增加一个单位， Y值的预测增量)

计算 a、b 值

- 满足 最小化的a、b值, 通过下列公式计算：

$$b = \frac{S_{YX}}{S_X^2}$$

$$a = \bar{Y} - b\bar{X}$$

$S_X^2 =$ X 的方差:

$$S_X^2 = \frac{\sum (X - \bar{X})^2}{N - 1}$$

$S_{YX} =$ X 和 Y 的协方差:

$$S_{YX} = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{N - 1}$$

b 和 a 的计算公式

- 一般我们利用计算机软件计算。如果要用计算器或Excel计算，可以用下面格式计算 b 值：

$$b = \frac{N(\sum XY) - (\sum X)(\sum Y)}{N(\sum X^2) - (\sum X)^2}$$

Baby	Age X	Weight Y	X^2	XY
A	1	10.0	1	10.0
B	4	10.0	16	40.0
C	0	5.0	0	0.0
D	6	15.9	36	95.4
E	5	17.2	25	86.0
F	6	12.7	36	76.2
G	0	6.8	0	0.0
H	3	14.1	9	42.3
I	2	7.3	4	14.6
J	5	14.5	25	72.5
K	2	9.5	4	19.0
L	1	8.2	1	8.2
M	4	12.7	16	50.8
N	3	11.3	9	33.9
Sum	42	155.2	182	548.9
Mean	3	11.086		

$$b = \frac{N(\sum XY) - (\sum X)(\sum Y)}{N(\sum X^2) - (\sum X)^2}$$

$$= \frac{14 * 548.9 - 42 * 155.2}{14 * 182 - 42^2}$$

$$= 1.488$$

Baby	X	Y	$(X-\bar{X})$	$(Y-\bar{Y})$	$(X-\bar{X})^2$	$(X-\bar{X})(Y-\bar{Y})$
A	1	10.0	-2	-1.086	4	2.171
B	4	10.0	1	-1.086	1	-1.086
C	0	5.0	-3	-6.086	9	18.257
D	6	15.9	3	4.814	9	14.443
E	5	17.2	2	6.114	4	12.229
F	6	12.7	3	1.614	9	4.843
G	0	6.8	-3	-4.286	9	12.857
H	3	14.1	0	3.014	0	0.000
I	2	7.3	-1	-3.786	1	3.786
J	5	14.5	2	3.414	4	6.829
K	2	9.5	-1	-1.586	1	1.586
L	1	8.2	-2	-2.886	4	5.771
M	4	12.7	1	1.614	1	1.614
N	3	11.3	0	0.214	0	0.000
Sum	42	155.2			56	83.300
Mean	3	11.086			4.308	6.408

$$b = \frac{S_{YX}}{S_X^2} = \frac{6.408}{4.308} = 1.488$$

$$a = \bar{Y} - b\bar{X} = 11.086 - 1.488 * 3 = 6.623$$

回归方程

- 婴儿年龄 X （月），体重Y （斤）
- 回归方程如下：$\hat{Y} = 6.623 + 1.488\,X$

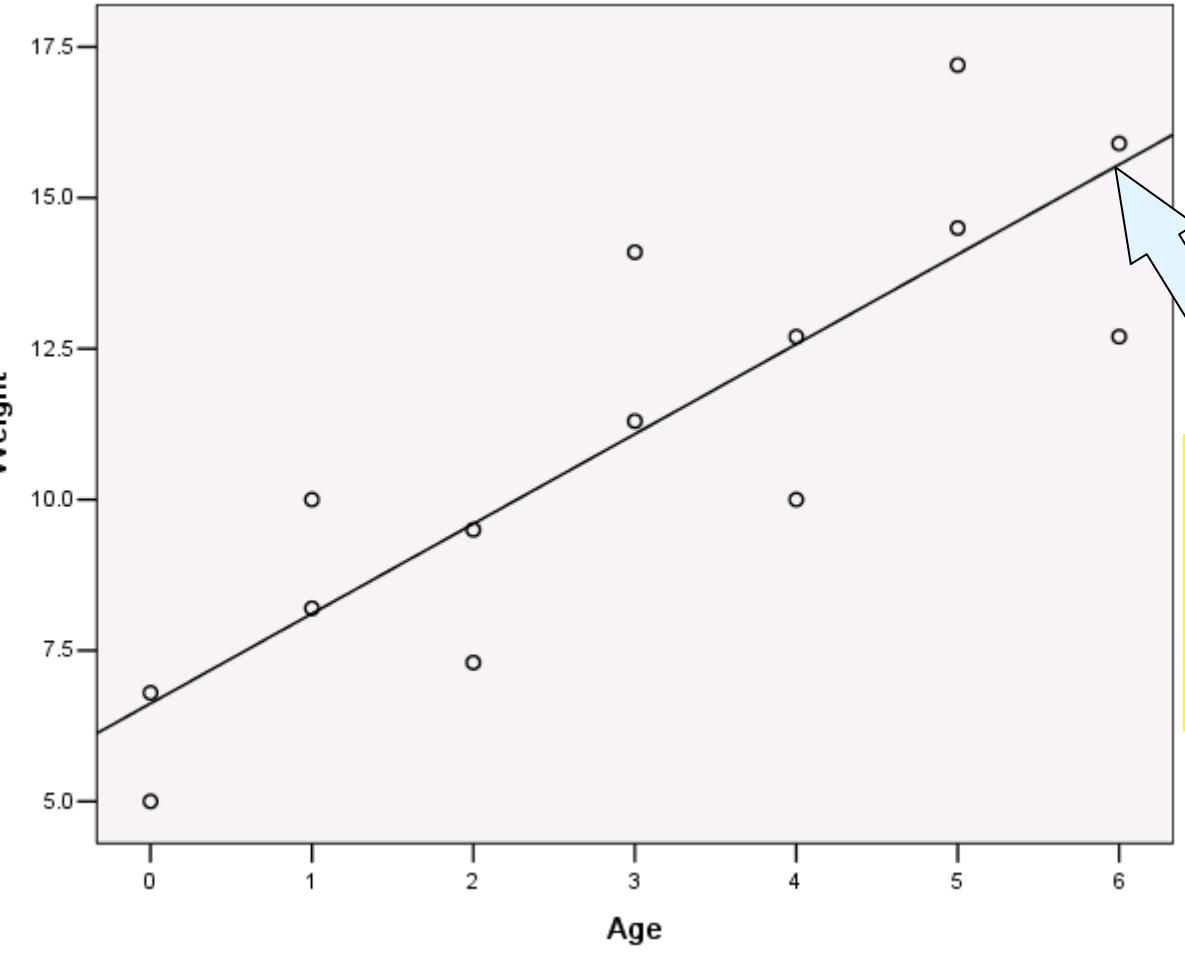
– 截距 (a) = 6.623

• 婴儿年龄为0个月（新生儿），其预测的体重为6.623斤

– 斜率 (b) = 1.488

• 婴儿年龄每增加一个月，体重增加1.488 斤

回归线



这是回归方程

$$\hat{Y} = 6.623 + 1.488X$$

- 通过这个回归方程，我们可以预测某个年龄的婴儿体重。
- 例如：婴儿3个月时体重多少斤？

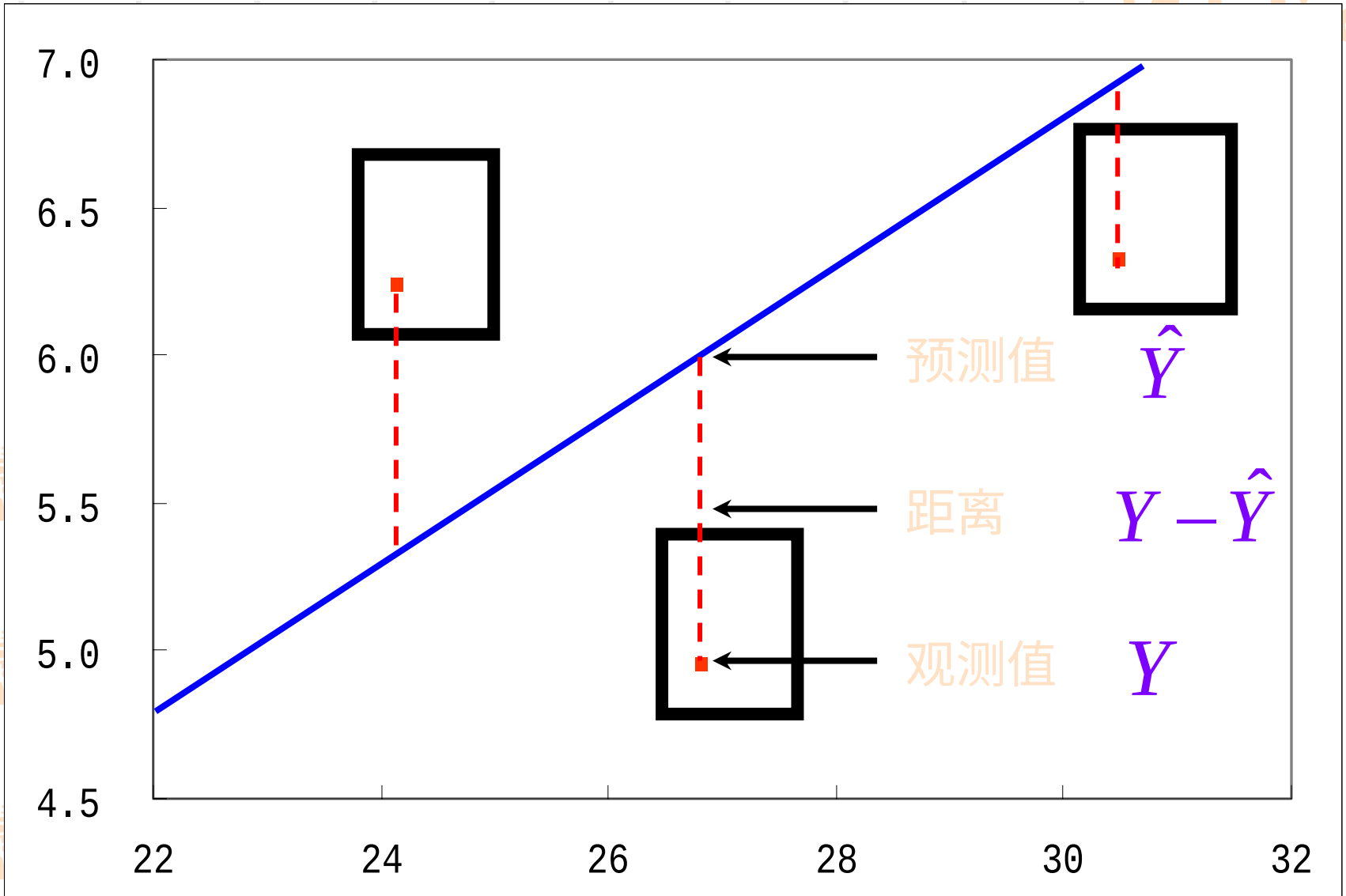
$$\hat{Y} = a + bX = 6.623 + 1.488 * 3 = 11.086$$

回归模型的拟合优度

- 回归模型的预测准确性如何？
- 确定系数 (r^2)：回归模型对数据的拟合程度（优度）。

预测误差：

误差 = Y 的观测值 - Y 的预测值



确定系数 (r^2)

- r^2 测量回归方程预测的准确性
- 方差被解释的比例：就是因变量(Y) 的总差异被自变量(X)所解释的比例。

- 因变量的总差异

$$\sum (Y - \bar{Y})^2$$

- 没有被自变量所解释的差异

$$\sum (Y - \hat{Y})^2$$

确定系数

- 因变量总差异被自变量所解释的比例：

$$r^2 = \frac{\sum (Y - \bar{Y})^2 - \sum (Y - \hat{Y})^2}{\sum (Y - \bar{Y})^2}$$

确定系数

- 确定系数的值范围：0-1
- 如果 $r^2 = 0$
- X 没有解释因变量的任何差异，X 与 Y无关
- 如果 $r^2 = 1$
- X 解释了100% 的因变量差异，即 X 与 Y完美相关
- 散点图中所有的点都落在回归线上
- 婴儿的年龄与体重：73.1% 的体重差异被年龄所解释

SPSS 的回归结果

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.855[a]	.731	.709	1.9491

a. Predictors: (Constant), Age

这是截距（a）

这是确定系数r^2

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	6.623	.939		7.053	.000
	Age	1.488	.260	.855	5.711	.000

a. Dependent Variable: Weight

这是斜率（b）

r^2的计算

- 计算 r^2 的一个简单的公式是：

$$r^2 = \frac{(X和Y的协方差)^2}{(X的方差)(Y的方差)} = \frac{S_{XY}^2}{S_X^2 S_Y^2}$$

皮尔逊相关系数 (r)

- 两个间距变量之间线性关系强弱的测量指标为皮尔逊r，就是确定系数r^2的平方根。

$$r = \sqrt{r^2}$$

- 相关系数 r 可以使用下列公式计算：

$$r = \frac{(X\text{和}Y\text{的协方差})}{(X\text{的标准差})(Y\text{的标准差})} = \frac{S_{XY}}{S_X S_Y}$$

- r 值的范围：-1 to +1
- 0 表示没有线性关系
- -1 表示完全负相关
- +1 表示完全正相关

相关系数 (r)

相关系数	强度解释	
1.00	完美正相关	Perfect Positive
0.80 to 0.99	非常强的正相关	Very Strong Positive
0.60 to 0.79	强的正相关	Strong Positive
0.40 to 0.59	一般正相关	Moderate Positive
0.20 to 0.39	弱正相关	Weak Positive
0.01 to 0.19	非常弱的正相关	Very Weak Positive
0.00	没有线性关系	No (Linear) Relationship
-0.01 to -0.19	非常弱的负相关	Very Weak Negative
-0.20 to -0.39	弱负相关	Weak Negative
-0.40 to -0.59	一般负相关	Moderate Negative
-0.60 to -0.79	强负相关	Strong Negative
-0.80 to -0.99	非常强的负相关	Very Strong Negative
-1.00	完美负相关	Perfect Negative

- 婴儿的例子：

$$r = \sqrt{r^2} = \sqrt{0.731} = 0.855$$

– 因为回归系数 b 是正的，所以r也是正的

- r 的计算：

$$r = \frac{S_{XY}}{S_X S_Y} = \frac{6.408}{2.075 * 3.611} = 0.855$$

$$r = \frac{N(\sum XY) - (\sum X)(\sum Y)}{\sqrt{\left[N(\sum X^2) - (\sum X)^2\right]\left[N(\sum Y^2) - (\sum Y)^2\right]}}$$

$$= \frac{14 * 548.9 - 42 * 155.2}{\sqrt{(14 * 182 - 42^2)(14 * 1890 - 155.2^2)}}$$

$$= 0.855$$

prov	edu	training	income	insurance
北京	12.51	1.79	1648.28	2.14
天津	11.91	1.39	1074.12	2.06
河北	12.01	1.58	903.75	1.91
山西	12.37	1.61	837.72	1.48
内蒙古	12.40	1.48	896.07	1.73
辽宁	11.83	1.40	912.80	2.00
吉林	12.26	1.42	966.45	1.52
黑龙江	12.16	.91	804.78	1.40
上海	12.00	1.43	1671.06	3.12
江苏	11.95	1.35	1062.69	2.24
浙江	11.19	1.90	1522.84	2.64
安徽	12.12	.94	806.29	1.89
福建	12.14	1.67	1213.70	1.92
江西	12.25	1.11	784.86	1.20
山东	12.98	1.53	1064.08	2.50
河南	12.38	1.21	915.28	1.98
湖北	12.84	1.34	994.05	1.95
湖南	12.61	1.07	1022.19	1.82
广东	12.36	1.77	1767.90	2.90
广西	12.49	1.82	1069.78	2.37
海南	12.32	1.06	987.39	2.20
重庆	12.06	2.39	976.83	2.24
四川	11.89	1.62	857.83	2.24
贵州	12.25	1.31	788.00	1.47
云南	12.34	1.51	1052.47	2.64
西藏	13.30	.63	2289.58	1.17
陕西	12.18	.83	797.81	1.70
甘肃	12.37	1.00	833.06	1.36
青海	12.23	1.05	1018.07	1.94
宁夏	11.87	1.90	836.79	1.49
新疆	12.96	1.23	1101.20	2.13

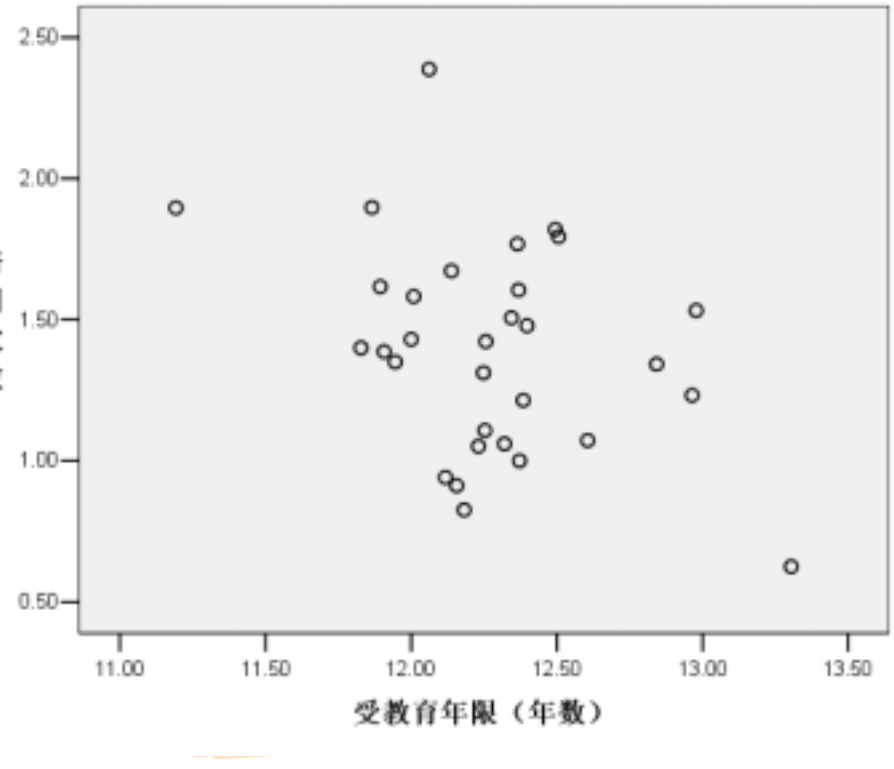
2002年调查

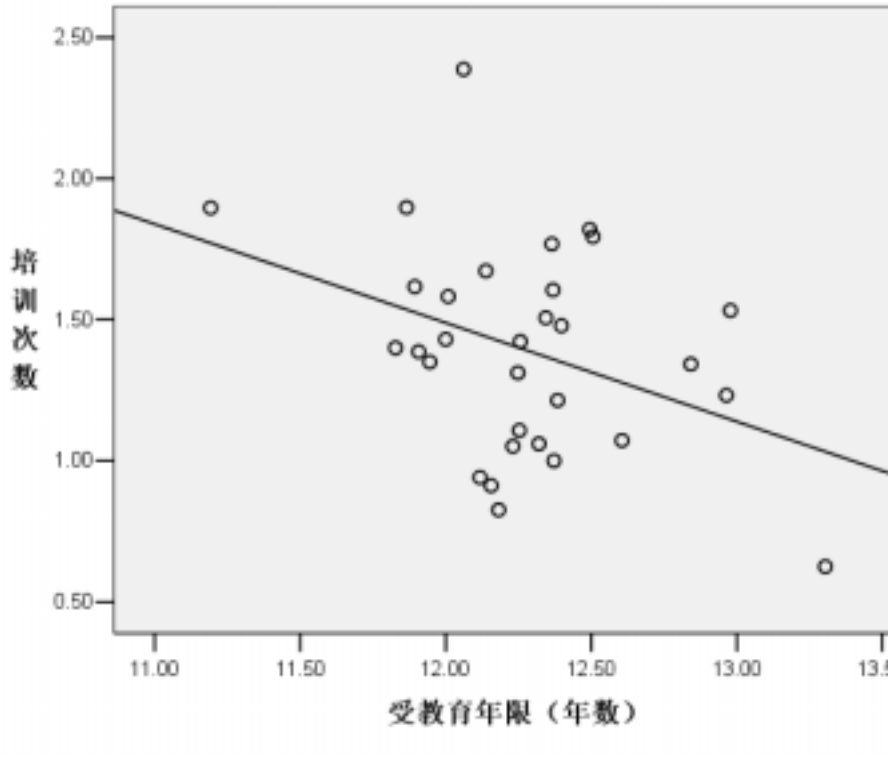
- 分析单位：省份
- edu：受教育年限
- training：培训次数
- income：收入
- insurance：参加保险种数

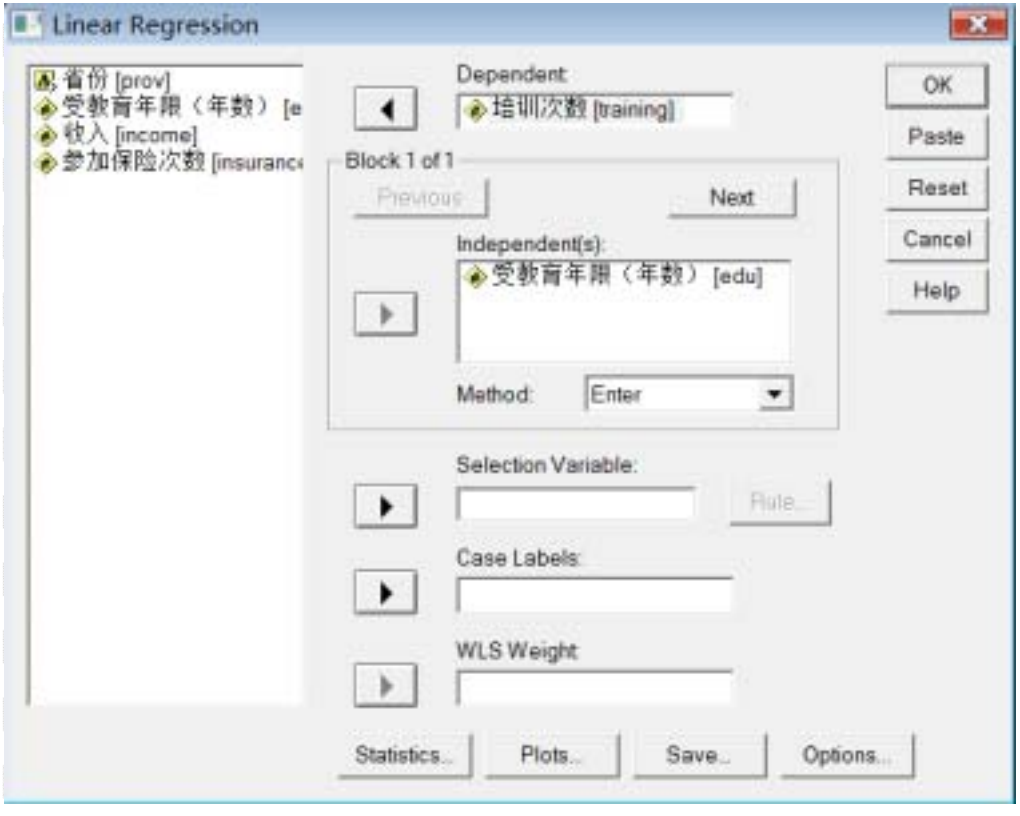
例1：文化程度和职业技能培训

- 自变量 (X)：受教育年限
- 因变量 (Y)：职业技能培训次数
- 职业技能培训次数与受教育年限之间是否存在显著的关系?

从散点图看，两者存在负相关








```
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT training
  /METHOD=ENTER edu  .
```

相关系数

确定系数

回归系数

Regression

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	受教育年限（年数）[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: 培训次数

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.374[a]	.140	.110	.35362

a. Predictors: (Constant), 受教育年限（年数）

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	.589	1	.589	4.709	.038[a]
	Residual	3.626	29	.125		
	Total	4.215	30			

a. Predictors: (Constant), 受教育年限（年数）

b. Dependent Variable: 培训次数

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	5.689	1.980		2.874	.008
	受教育年限（年数）	-.350	.161	-.374	-2.170	.038

a. Dependent Variable: 培训次数

回归方程：$y = 5.689 - 0.350x$

- a = 5.689：受教育程度为0的职工，接受职业技能培训5.689次
- b = -0.350：受教育年限每增加1年，接受培训次数将下降0.35次
- $r^2 = 0.14$：受教育年限解释了14%的培训次数差异

- r = 0.374：受教育年限和培训次数的相关系数为0.374 (弱相关)

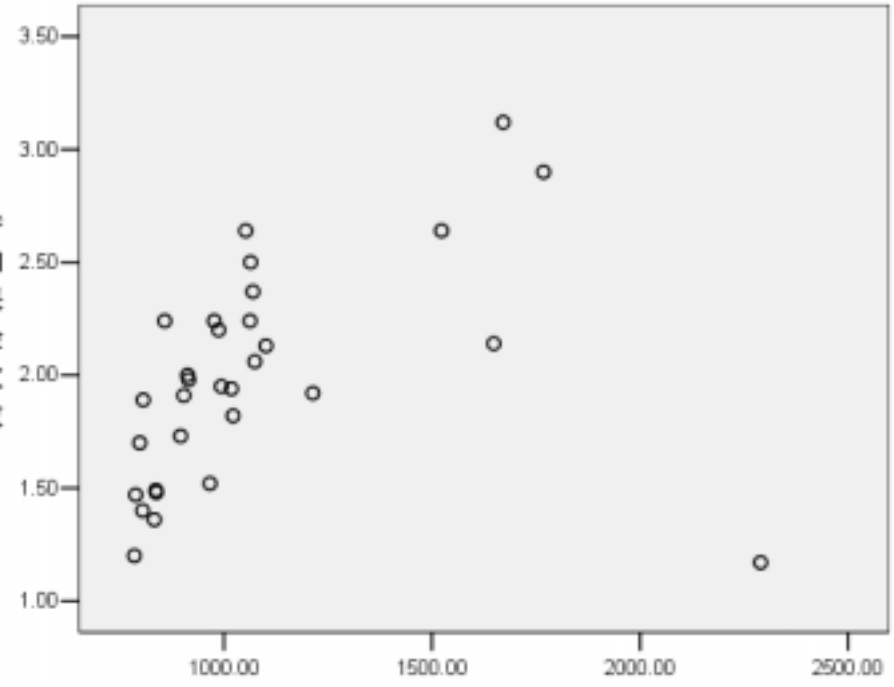
- 大学毕业者会培训几次？

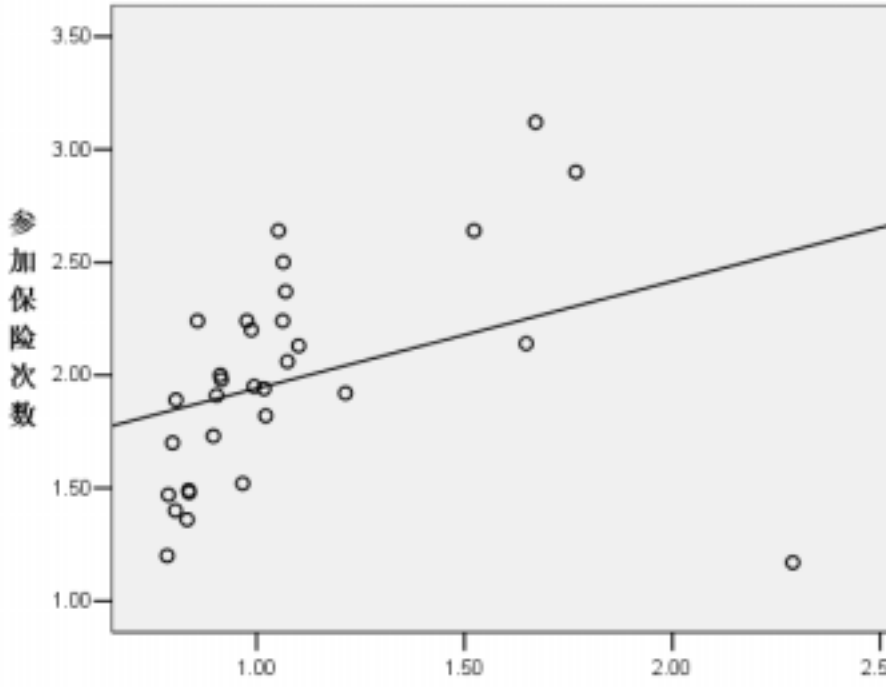
$$y=5.689-0.35*16=0.089$$

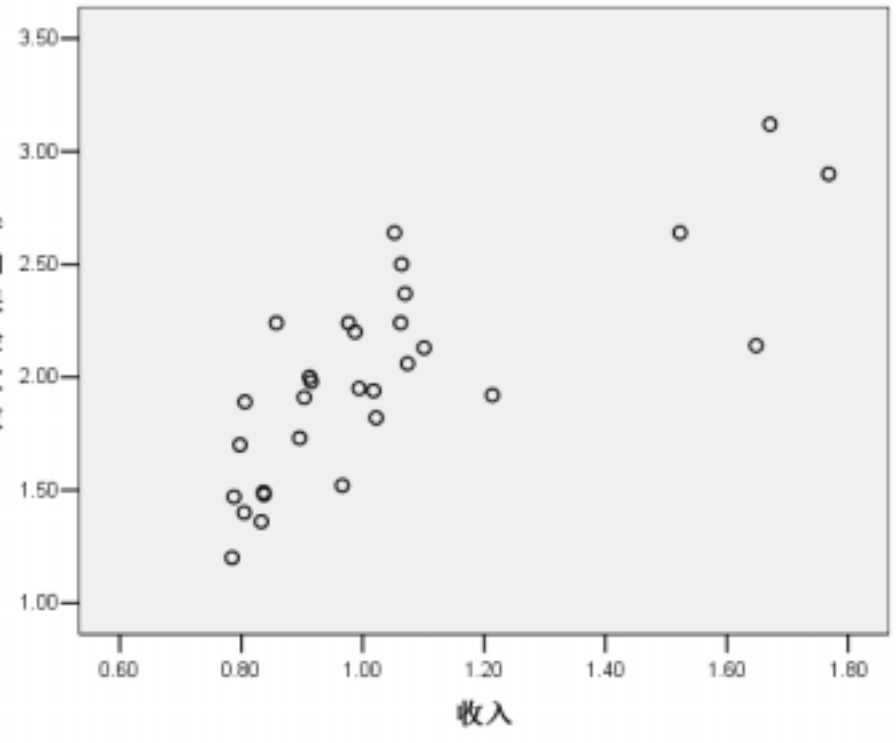
例2：收入与参加保险种数

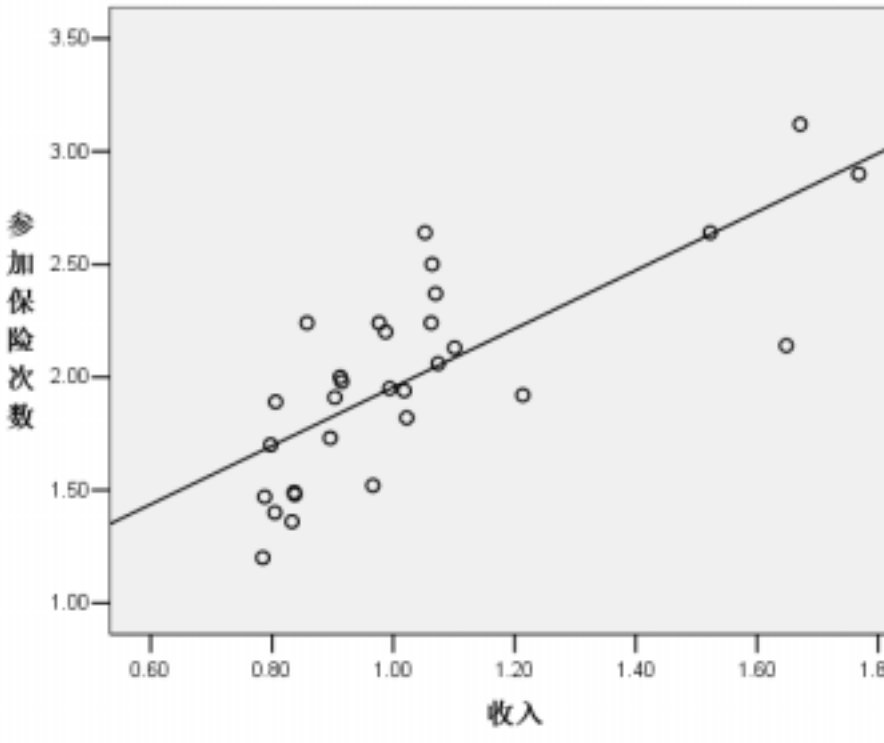
- 自变量 (X)：收入
- 因变量 (Y)：参加保险种数
- 参加保险种数与收入之间是否存在显著的关系?

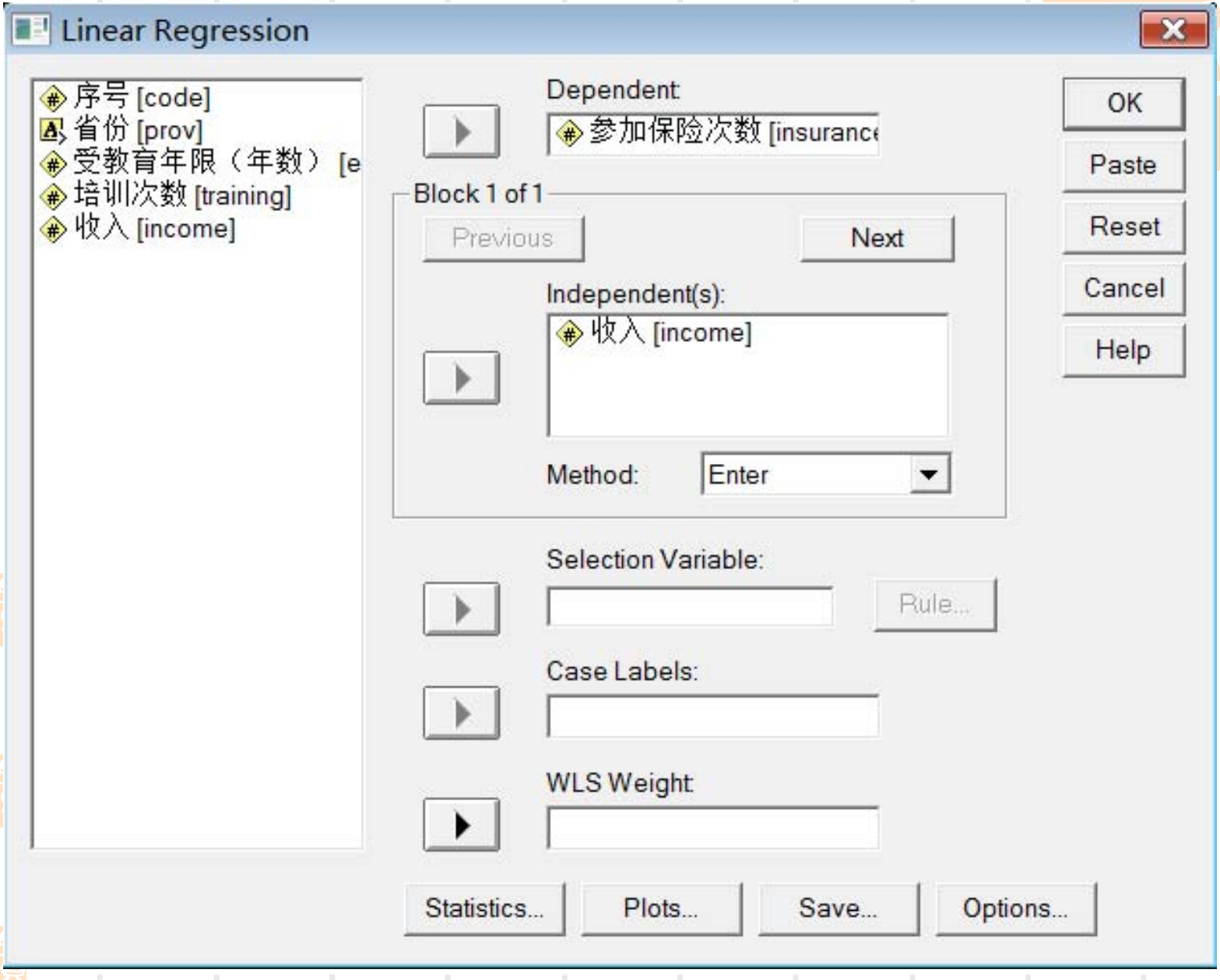
从散点图看，两者存在正相关






3.50
3.00
2.50
2.00
1.50
1.00
0.60
0.80
1.00
1.20
1.40
1.60
1.80
收入


参加保险次数
3.50
3.00
2.50
2.00
1.50
1.00
0.60
0.80
1.00
1.20
1.40
1.60
收入


Linear Regression
序号 [code]
省份 [prov]
受教育年限（年数） [e
培训次数 [training]
收入 [income]
Dependent:
参加保险次数 [insurance
Block 1 of 1
Previous
Next
Independent(s):
收入 [income]
Method:
Enter
Selection Variable:
Rule...
Case Labels:
WLS Weight:
OK
Paste
Reset
Cancel
Help
Statistics...
Plots...
Save...
Options...

```
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT insurance
  /METHOD=ENTER income  .


temporary.
select if code~=26.
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT insurance
  /METHOD=ENTER income  .
```

回归方程：y =1.466+0.475 *x*

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.345[a]	.119	.089	.45711

a. Predictors: (Constant), 收入

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	1.466	.272		5.387	.000
	收入	.475	.240	.345	1.980	.057

a. Dependent Variable: 参加保险次数

回归方程：y =0.659+1.296 *x*

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.756[a]	.571	.556	.30827

a. Predictors: (Constant), 收入

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	.659	.228		2.895	.007
	收入	1.296	.212	.756	6.105	.000

a. Dependent Variable: 参加保险次数

多元线性回归

- 解释社会现象很少只用一个原因，往往社会现象受众多因素影响
- 将更多的自变量放入回归模型，可以更充分地解释自变量的差异，更准确预测因变量值
- 多元回归中，自变量起到相互控制作用，可以反映自变量的独立影响


```
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT insurance
  /METHOD=ENTER edu income  .
```

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.467[a]	.218	.162	.43822

a. Predictors: (Constant), 收入, 受教育年限（年数）

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	6.125	2.485		2.465	.020
	受教育年限（年数）	-.390	.207	-.326	-1.885	.070
	收入	.001	.000	.429	2.481	.019

a. Dependent Variable: 参加保险次数

回归系数

标准化回归系数

- 谢谢！
- 相关学习内容补充请见课程网站学习专栏