

利用SPSS进行科学数据分析

中科院文献情报中心

王辉

主要内容

- 统计学简介
- SPSS简介
- 建库
- 数据编辑
- 统计性描述
- 相关性分析
- 降维分析
- 多元回归分析
- 图形构建程序

统计学简介

- 统计学是一门收集、整理和分析数据的方法科学，其目的是探索数据的**内在数量规律性**，以达到对客观事物的科学认识。
- 试验设计、数据收集、数据获取、数据准备、数据分析、数据报告、模型发布

统计研究对象的特点

- 数量性---统计学研究的对象是客观现象的**数量特征和规律性**。
- 总体性---统计学研究的是客观现象**总体的**数量特征与规律性，而不是个体的量。
- 具体性---统计的对象是**一定时间、地点、条件下**事物的量，而不是抽象对象的量，这是统计学和数学的一个重要区别。
- 差异性---组成统计研究对象总体的**个体是有差异的**，否则就不需要进行统计分析。统计研究中需要对总体中大量的个体进行观察并进行综合分析，由此才能获得总体的数量分布特征。

SPSS简介

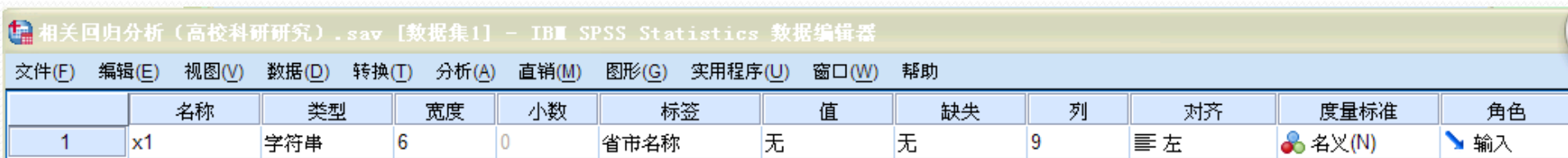
- SPSS-----Statistical Package for Social Science （社会科学统计软件包）现改名为Statistical Product and Service Solutions(统计产品与服务解决方案)
- SPSS for Windows的命令语句、子命令及选择项大部分由“菜单”、“图标按钮”、“对话框”的操作完成，操作简单、使用方便。
- 具有完整的数据输入、编辑、统计分析、报表、图形制作等功能。能更快速地读取并分析大量数据。
- SPSS for Windows与其它软件有数据转换接口。

SPSS简介

- 使用SPSS时，我们需要**具备一定的统计学知识**，只有掌握了相应的统计学方法后才能使用该软件包及对结果做出合理解释。
- 与另一著名统计软件SAS相比，更适用于统计初学者或非统计学专业人员。

建库

- 打开SPSS, **文件-新建-数据**。
- 数据编辑窗口有两个标签
 - 一个是**变量视图 (Variable View)** , **变量视图**用于定义和编辑变量的数据格式, 如变量的**类型**、**宽度**、**缺失值**等。变量: 事物的特征, 是运用统计方法所分析的对象。

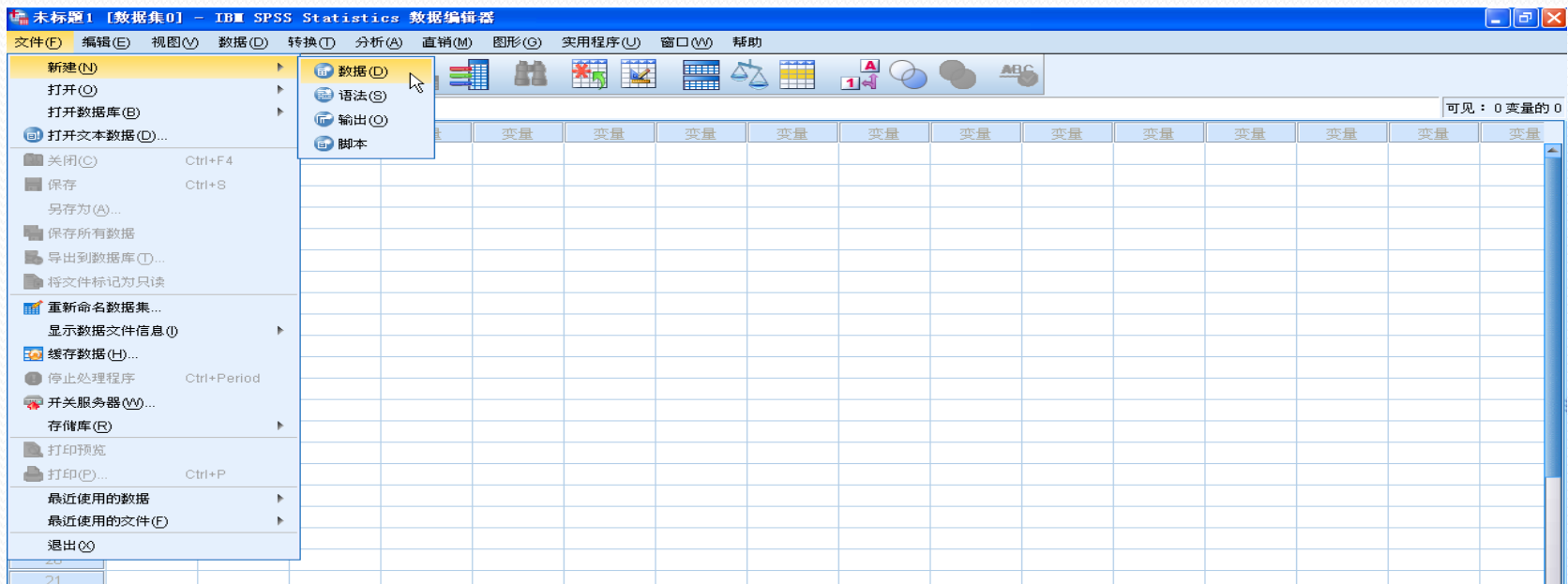


	名称	类型	宽度	小数	标签	值	缺失	列	对齐	度量标准	角色
1	x1	字符串	6	0	省市名称	无	无	9	左	名义(N)	输入

- 另一个是**数据视图 (Data View)** , **数据视图**用于输入数据。数据: 是与变量相关的值。变量可能随时间变动, 这些不同的值就是与变量相关的数据, 或者简单地说, 就是统计所要分析的“数据”。

建库

- 新建数据
- 打开已有数据
 - SPSS的数据接口非常多，除了在SPSS中新建一个数据库以外，我们也可以从其它数据库软件中导入现成的数据，如DBF文件、ASCII文件以及Excel电子表格等13种文件。



数据编辑

- 对数据进行处理时，一种统计分析方法需要数据具有一定的数据格式，因此需要对原来数据文件进行编辑加工。如：
 - 变量的增加和删减
 - 观察值的增加和修改
 - 数据的定位、排序、对数据进行转换
 - 重新编码

*相关回归分析 (高校科研研究).sav [数据集1] - IBM SPSS Statistics 数据编辑器

文件(F) 编辑(E) 视图(V) 数据(D) **转换(T)** 分析(A) 直销(M) 图形(G) 实用程序(U) 窗口(W) 帮助

1: s2n 13590.00 可见: 13 变量

	x1	x2	x3	x4	x5	x6	x7	x8	VAR00001	PRE_1	FAC
1	北京	6795.0	3737.0	33980300.0	3						
2	天津	1649.0	939.0	4539200.0							
3	河北	2367.0	1039.0	4063100.0							
4	山西	1460.0	658.0	4966100.0							
5	内蒙	455.0	231.0	700100.0							
6	辽宁	3664.0	1591.0	7030100.0	1						
7	吉林	2514.0	1208.0	4415400.0							
8	黑龙江	1430.0	797.0	947700.0							
9	上海	3783.0	1833.0	11629200.0	2						
10	江苏	5480.0	2436.0	13841800.0	3						
11	浙江	2765.0	1238.0	4432000.0	1						
12	安徽	2157.0	982.0	4967200.0							

计算变量

目标变量(T): s2n

数字表达式(E): x2 * 2

类型与标签(L)...

函数组(G): 全部, 算术, CDF 与非中心 CDF, 转换, 当前日期时间

60%

数据编辑

例如：

- 投入科研事业费X4和获奖数X8两列数据，数据在值上的差异比较大，如果就这两列数据进行分析，就需要对数据进行预处理。
- 可以采用数据标准化的方法消除变量间的量纲关系，将他们放在同一个标准的数据上，从而使数据具有可比性。
- 一般标准化采用的是Z标准化，即每一变量值与其平均值之差除以该变量的标准差，标准化后均值为0，方差为1。
- 当然也有其他标准化，比如离差标准化，是对原始数据的线性变换，使结果落到[0,1]区间等等。

统计性描述

- 在SPSS中，在数据视图窗口
- 点击分析-描述统计- 将相比指标加入到右边选区，将图片左下角处打钩，即进行标准化处理-确定，数据表中自动列出标准化后的数据。
- 点击分析-描述统计- 描述-将相比指标加入到右边选区，在选项选择要描述的统计值，点击继续，确定。在output窗口即可看到结果。

描述统计量

	N	全距	极小值	极大值	均值		标准差	方差
	统计量	统计量	统计量	统计量	统计量	标准误	统计量	统计量
投入人年数	31	6720.0	75.0	6795.0	2144.387	293.5428	1634.3769	2671187.845
投入高级职称的人年数	31	3713.0	24.0	3737.0	1035.032	149.4795	832.2668	692668.032
投入科研事业费	31	33980300.0	.0	33980300.0	5563245.161	1209847.608	6736146.396	4.538E13
课题总数	31	3244.0	17.0	3261.0	960.000	150.5431	838.1887	702560.333
专著数	31	2717.0	6.0	2723.0	486.161	94.2478	524.7494	275361.940
论文数	31	12153.0	117.0	12270.0	4530.258	596.9628	3323.7483	11047302.79
获奖数	31	540.0	.0	540.0	133.806	25.2369	140.5132	19743.961
有效的 N (列表状态)	31							

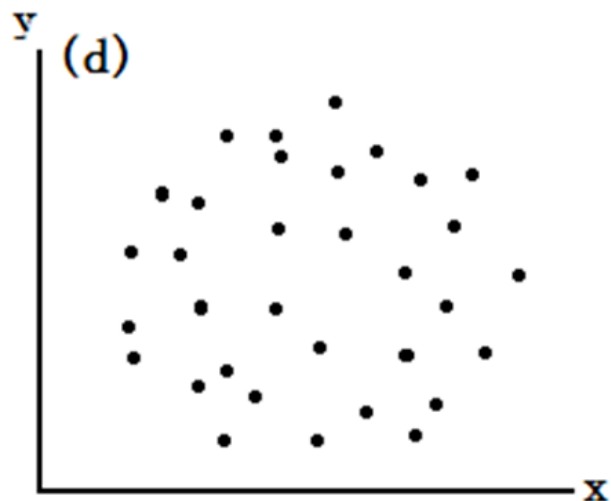
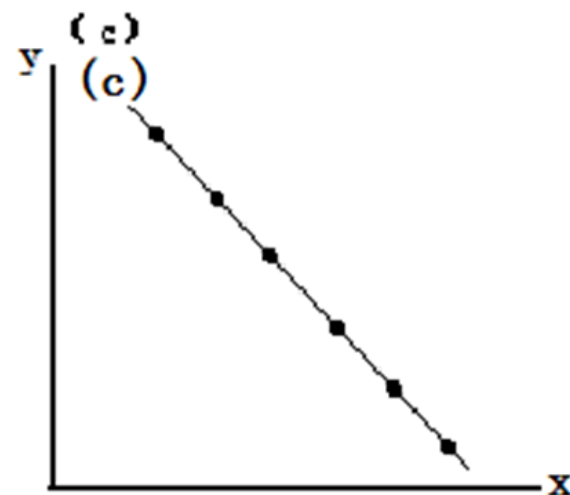
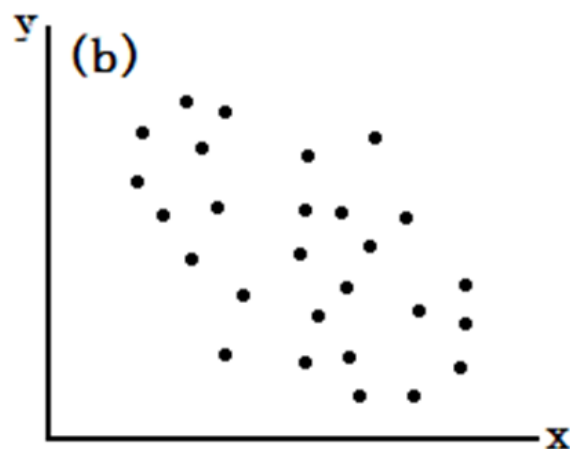
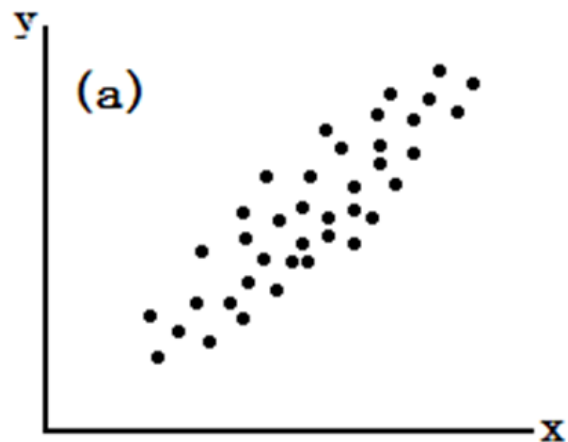
- 全距：最大值和最小值之差。
- 方差和标准差反映的信息在离散指标中是最全面、最可靠的变异描述指标。
- 方差：用于不同含量样本数据分布离散程度的比较。
- 标准差：标准差度量了偏离平均数的大小，相当于平均偏差，可以直接地、概括地、平均地描述数据变异的大小。
- 标准误：样本均数的标准差。
- 方差和标准差越大，数据分布离散程度越大。

相关性分析

- 相关分析可以发现变量间的**共变关系**（包括**正向的**和**负向的**共变关系），一旦发现了共变关系就意味着变量间可能存在两种关系中的一种
 - 第一，**因果关系**（两个变量中一个为因、一个为果）
 - 第二，**存在公共因子**（两变量均为果，有潜在的共因）
- 目的：寻找因果关系，或者是寻找公共因子。

相关性分析

- 分析的角度主要包括下述三个方面
 - 相关的方向
 - 相关的形式
 - 相关的程度或强度



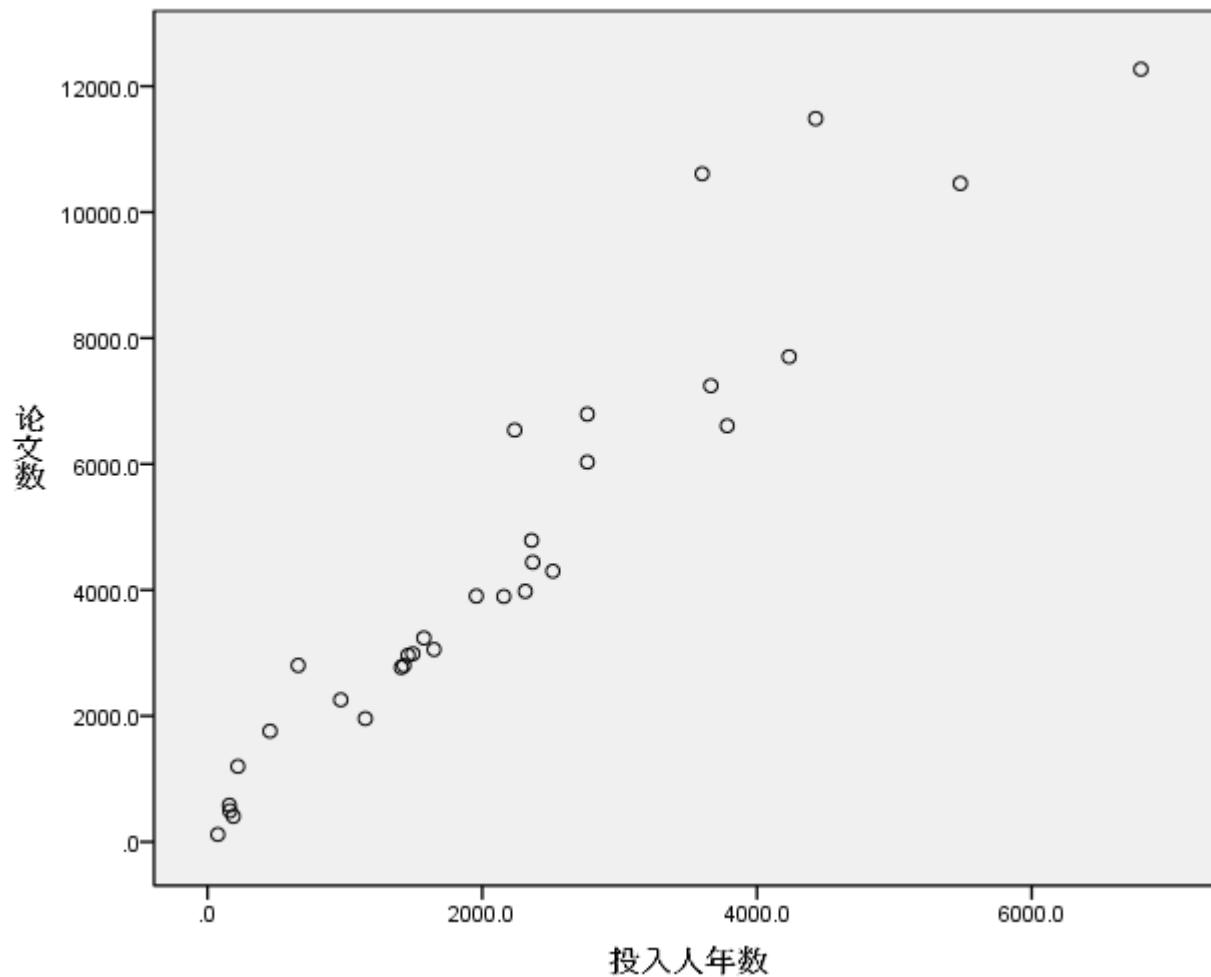
- (a) 一个较高的正相关, r 的近似值为 $+0.90$
- (b) 一个较低的负相关, r 的近似值为 -0.40
- (c) 一个高度的负相关, $r = -1.00$
- (d) 无线性相关, $r = 0$

数据的几种相关关系图

- 相关系数 r ：在直线相关条件下，反映两变量间线性相关关系的统计指标
- $|r|=1$ ，表示 x 与 y 变量为完全的线性相关，也即为确定的函数关系；
 $|r|=0$ ，表示两变量不存在线性相关；
 $0<|r|<1$ ，表示两变量存在不同程度的线性相关。
 $0<|r|\leq 0.3$ 为微弱相关；
 $0.3<|r|\leq 0.5$ 为低度相关；
 $0.5<|r|\leq 0.8$ 为显著相关；
 $0.8<|r|<1$ 为高度相关。
- 相关系数 r 的显著性检验(t 检验)：两变量之间是否真正存在显著的线性相关关系，即相关系数高的样本是否可能来自一个不存在线性相关的总体，可通过对**相关系数的显著性检验来作出判断**。

相关性分析-散点图

- 1) 可以用散点图来直观表达两个变量的变化关系
- 操作：在SPSS中，在数据视图窗口
 - 图形-散点/点状图，选择简单散点图，输入x（年投入人数），y（论文数），得到散点图。



- 散点图分析

- 从散点图上我们可以看出，这两组数据呈非线性的正相关关系。

相关性分析-相关分析

- 双变量分析 ----确定两个变量之间的相关性

相关性

		投入人年数	论文数
投入人年数	Pearson 相关性	1	.952 ^{**}
	显著性 (双侧)		.000
	N	31	31
论文数	Pearson 相关性	.952 ^{**}	1
	显著性 (双侧)	.000	
	N	31	31

** . 在 .01 水平 (双侧) 上显著相关。

相关性分析-偏相关(Partial Corr.)分析

- 直接的相关分析所得到的两个变量间的共变关系，它反映了这两个变量间相互作用的关系或共同受到某一潜在因素影响**的强弱**，但是这种关系未必纯粹。一般来说，简单**相关系数和偏相关系数**相比，前者有夸大的成分，后者更符合实际。
- 投入人数与论文数的相关不是纯粹反映投入人数与论文数的关系的，因为投入人数可能还与投入科研事业费这一“第三者”有关。为了在剔除投入科研事业费影响的情况下，找到投入人数与论文数的相关性，这时就要使用偏相关(Partial Corr.)分析方法以对“第三者”施加“管制”。
- 为了在剔除投入科研事业费情况下，找到投入人数和论文数相关性，使用**偏相关(Partial Corr.)分析方法**。

相关性分析-偏相关(Partial Corr.)分析

- 操作：在SPSS中，在数据视图窗口
 - 分析-相关-偏相关分析，选择两个变量：年投入人数和论文数，选择控制变量投入科研事业费，点击确定，在output窗口看输入结果。

相关性

		投入人年数	论文数	投入科研事业费
投入人年数	Pearson 相关性	1	.952**	.856**
	显著性 (双侧)		.000	.000
	N	31	31	31
论文数	Pearson 相关性	.952**	1	.781**
	显著性 (双侧)	.000		.000
	N	31	31	31
投入科研事业费	Pearson 相关性	.856**	.781**	1
	显著性 (双侧)	.000	.000	
	N	31	31	31

** 在 .01 水平 (双侧) 上显著相关。

相关性

控制变量		投入人年数	论文数
投入科研事业费	投入人年数	相关性	1.000
		显著性 (双侧)	.878
		df	.000
论文数	论文数	相关性	0
		显著性 (双侧)	.28
		df	.000

曲线拟合

- 用连续曲线近似地刻画或比拟平面上离散点组所表示的坐标之间的函数关系。在模型汇总结果中，根据R方和F值选择最优拟合曲线方程。
- 操作：在SPSS中，在数据视图窗口
 - 分析-回归-曲线估计

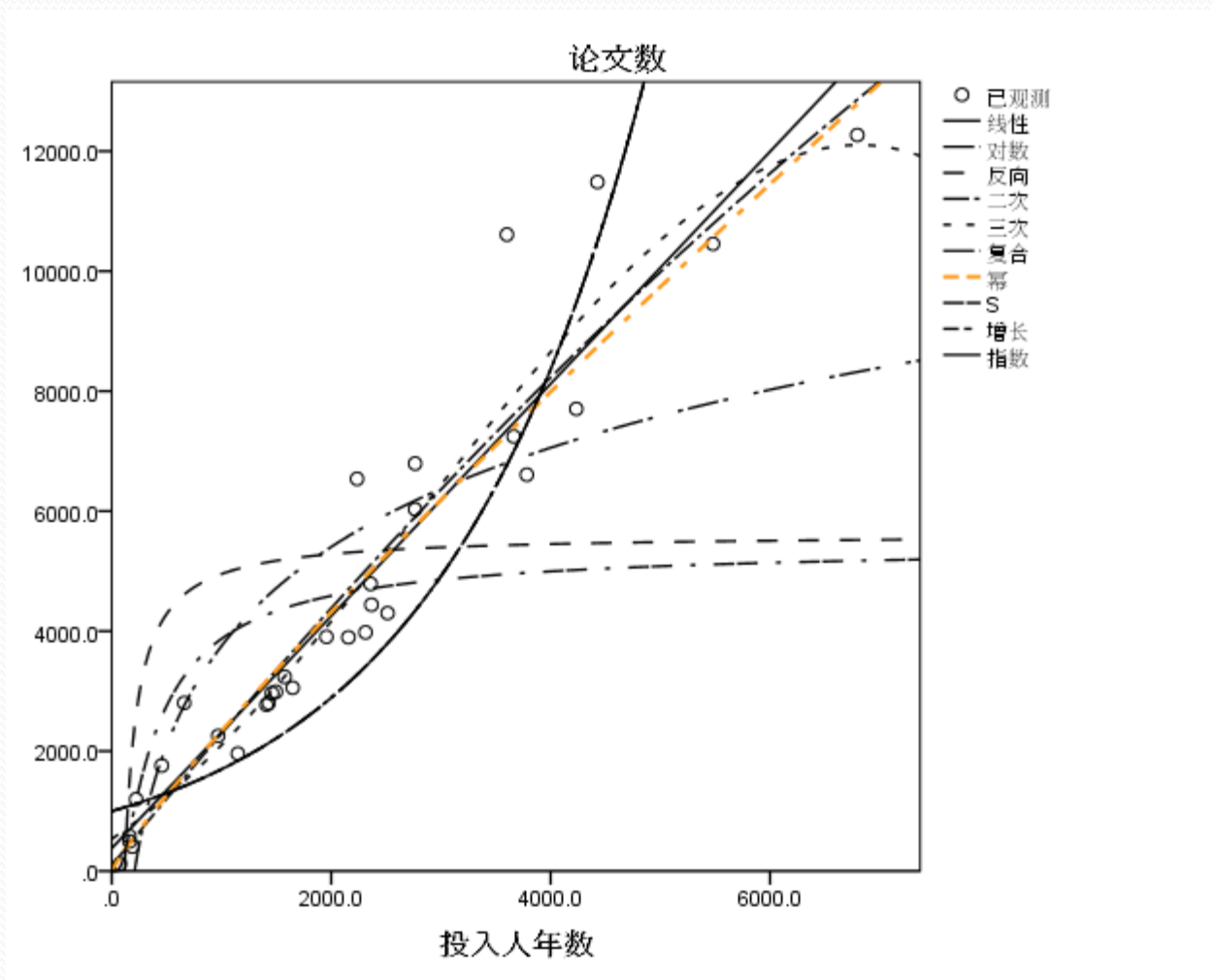
曲线拟合

模型汇总和参数估计值

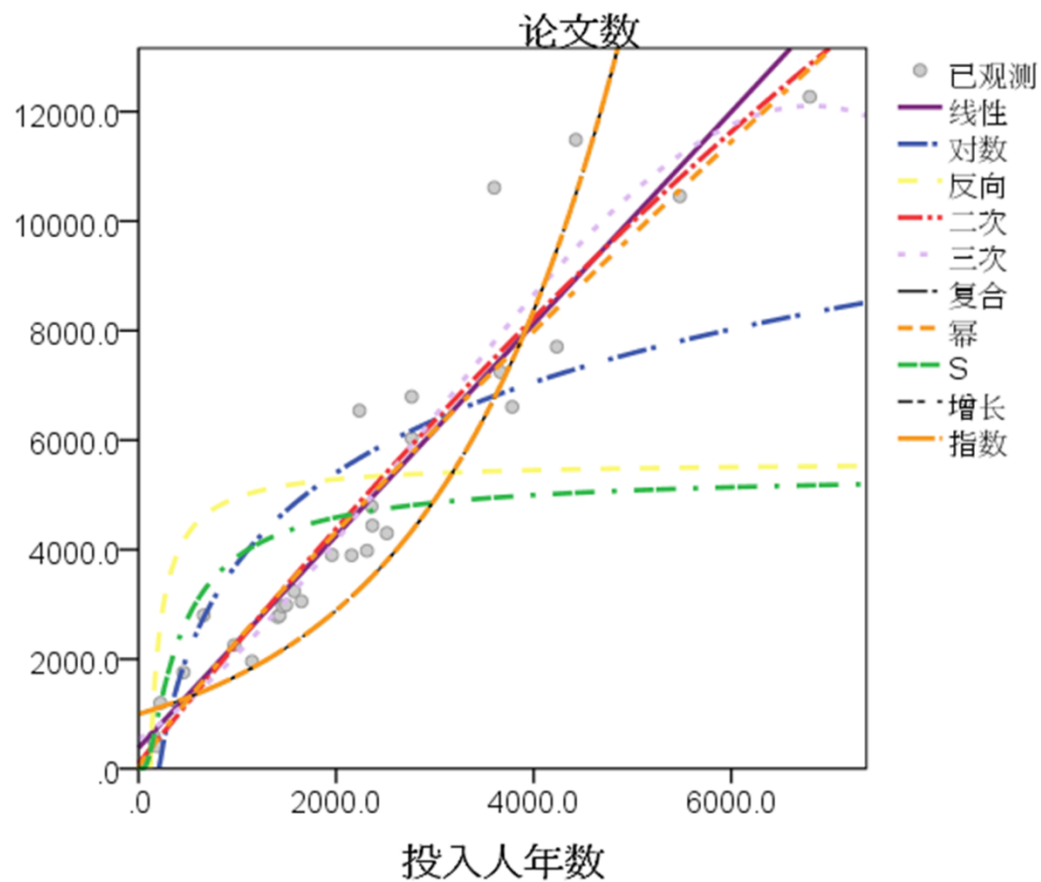
因变量:论文数

方程	模型汇总					参数估计值			
	R 方	F	df1	df2	Sig.	常数	b1	b2	b3
线性	.906	279.857	1	29	.000	379.107	1.936		
对数	.682	62.065	1	29	.000	-12724.444	2384.992		
倒数	.313	13.236	1	29	.001	5617.247	-665312.336		
二次	.909	140.684	2	28	.000	91.652	2.255	-5.535E-5	
三次	.918	100.195	3	27	.000	536.153	1.222	.000	-4.686E-8
复合	.675	60.321	1	29	.000	997.600	1.001		
幂	.933	404.383	1	29	.000	5.067	.888		
S	.810	123.775	1	29	.000	8.601	-340.312		
增长	.675	60.321	1	29	.000	6.905	.001		
指数	.675	60.321	1	29	.000	997.600	.001		
Logistic	.675	60.321	1	29	.000	.001	.999		

自变量为 投入人年数。



曲线拟合



降维分析

- 在各个领域的科学研究中，往往需要对反映事物的 **多个变量** 进行大量的观测，收集大量数据以便进行 **分析寻找规律**。
- 多变量大样本无疑会为科学研究提供丰富的信息，但也在一定程度上增加了 **数据采集的工作量**。更重要的是在大多数情况下，**许多变量之间可能存在相关性**而增加了问题分析的复杂性，同时对分析带来不便。如果分别分析每个指标，分析又可能是孤立的，而不是综合的。盲目减少指标会损失很多信息，容易产生错误的结论。
- **减少分析指标的同时，尽量减少原指标包含信息的损失**，对所收集的资料作全面的分析。
- 由于各变量间存在一定的相关关系，因此有可能 **用较少的综合指标分别综合存在于各变量中的各类信息**。

降维分析

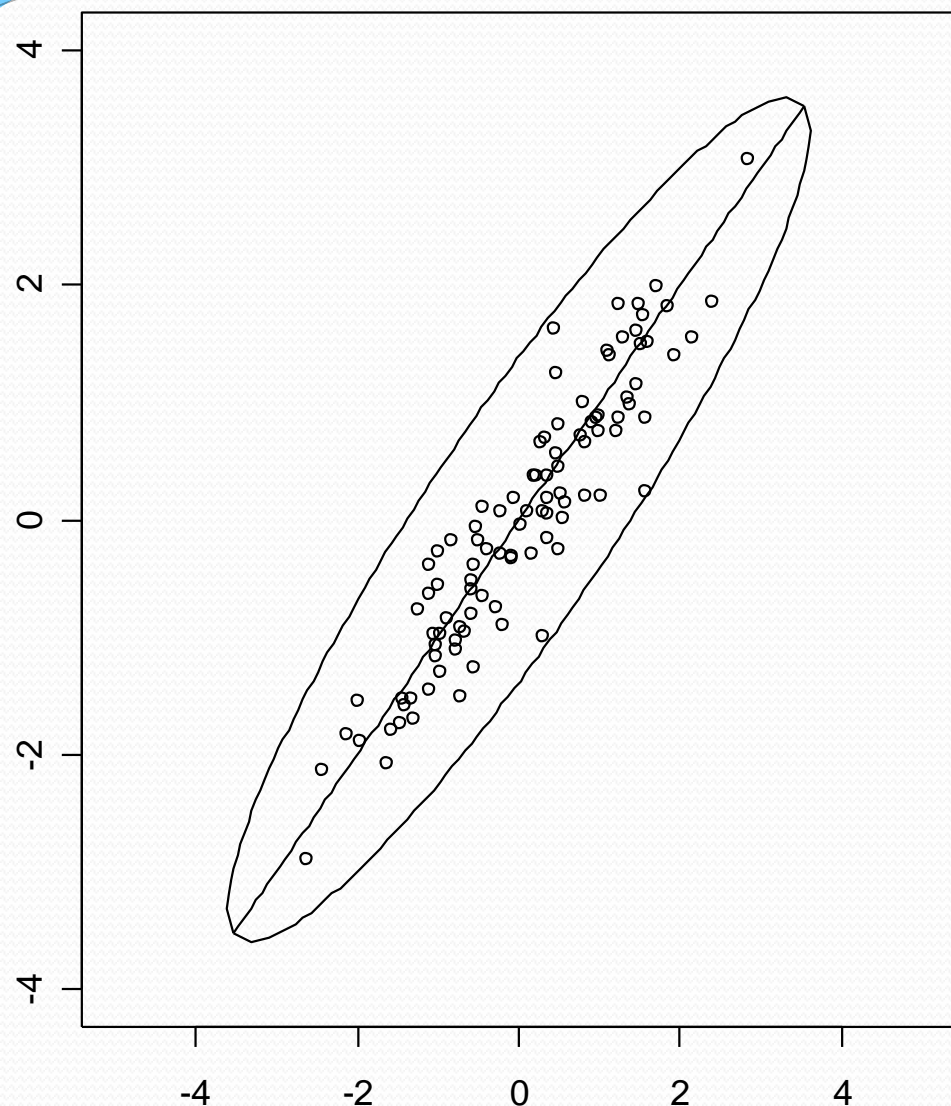
- 每个人都会遇到**有很多变量的数据**。比如全国或各个地区的带有许多经济和社会变量的数据；各个学校的研究、教学等各种变量的数据等等。
- 这些数据的共同特点是变量很多，在如此多的变量之中，有很多是相关的。人们希望能够找出它们的**少数“代表”**来对它们**进行描述**。
 - 主成分分析 (principal component analysis)
 - 因子分析 (factor analysis)
- 主成分分析与因子分析是将多个**实测变量**转换为**少数几个不相关的综合指标**的多元统计分析方法。**综合指标**往往是**不能直接观测到的**，但它更能反映事物的本质。

降维分析----主成分分析

- 目前的问题是，能不能把这个数据的7个变量用一两个综合变量来表示呢？
- 这一两个综合变量包含有多少原来的信息呢？
- 能不能利用找到的综合变量来对各省科研状况排序呢？

降维分析----主成分分析

- 例中的的数据点是7维的；也就是说，每个观测值是7维空间中的一个点。我们希望把7维空间用低维空间表示。
- 先**假定只有二维**，即只有**两个变量**，它们由横坐标和纵坐标所代表；因此每个观测值都有相应于这两个坐标轴的两个坐标值；如果这些数据形成一个椭圆形状的点阵（这在变量的二维正态的假定下是可能的）
- 那么这个椭圆有一个长轴和一个短轴。在短轴方向上，数据变化很少；在极端的情况，短轴如果退化成一点，那只有在长轴的方向才能够解释这些点的变化了；这样，由二维到一维的降维就自然完成了。



- 当坐标轴和椭圆的长短轴平行，那么代表长轴的变量就描述了数据的主要变化，而代表短轴的变量就描述了数据的次要变化。
- 但是，坐标轴通常并不和椭圆的长短轴平行。因此，需要寻找椭圆的长短轴，并进行变换，使得新变量和椭圆的长短轴平行。
- 如果长轴变量代表了数据包含的大部分信息，就用该变量代替原先的两个变量（舍去次要的一维），降维就完成了。
- 椭圆（球）的长短轴相差得越大，降维也越有道理。

降维分析----主成分分析

- 对于多维变量的情况和二维类似，也有高维的椭球，只不过无法直观地看见罢了。
- 首先把高维椭球的主轴找出来，再用**代表大多数数据信息的最长的几个轴作为新变量**；这样，主成分分析就基本完成了。
- 注意，和二维情况类似，高维椭球的主轴也是互相垂直的。**这些互相正交的新变量是原先变量的线性组合，叫做主成分 (principal component analysis(PCA))。**

降维分析----因子分析

- 主成分分析从原理上是寻找椭球的所有主轴。因此，原先有几个变量，就有几个主成分。
- 而因子分析是**事先确定要找几个成分**，这里叫**因子**（factor）（比如两个），那就找两个。
- 这使得在数学模型上，因子分析和主成分分析有不少区别。而且**因子分析的计算也复杂得多**。根据因子分析模型的特点，它还多一道工序：**因子旋转（factor rotation）**；**这个步骤可以使结果更好。**

降维分析----因子分析

- 操作：在SPSS中，在数据视图窗口
 - 分析-降维-因子分析，选择旋转算法，做主成分分析。
- 结果分析：
 - 因子方差表，提取因子后因子方差的值均很高，表明提取的因子能很好的描述这7个指标。

公因子方差

	初始	提取
投入人年数	1.000	.968
投入高级职称的人年数	1.000	.987
投入科研事业费	1.000	.916
课题总数	1.000	.922
专著数	1.000	.952
论文数	1.000	.931
获奖数	1.000	.967

提取方法：主成份分析。

降维分析----因子分析

在数学变换中保持变量的总方差不变，使第一变量具有最大的方差，称为第一主成分，第二变量的方差次大，并且和第一变量不相关，称为第二主成分。

方差分解表 表明前两个因子能够解释7个指标的**94.889%**。

解释的总方差

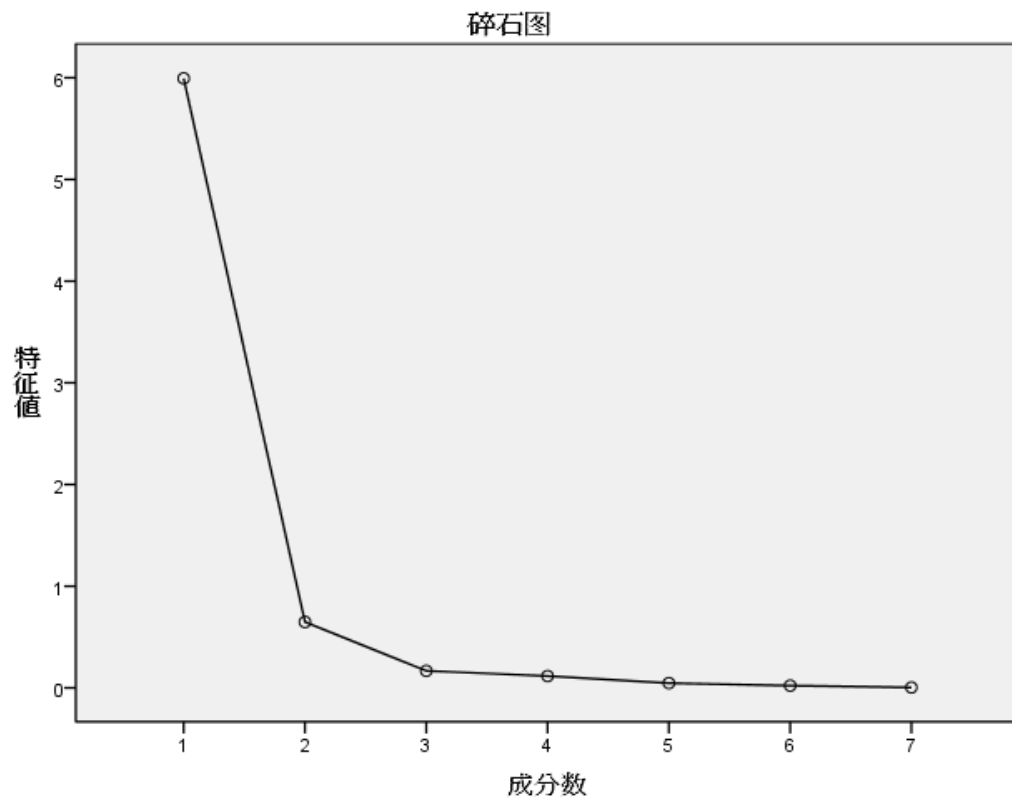
成份	初始特征值			提取平方和载入			旋转平方和载入		
	合计	方差的 %	累积 %	合计	方差的 %	累积 %	合计	方差的 %	累积 %
1	5.994	85.631	85.631	5.994	85.631	85.631	5.938	84.835	84.835
2	.648	9.258	94.889	.648	9.258	94.889	.704	10.054	94.889
3	.167	2.392	97.281						
4	.117	1.669	98.950						
5	.047	.673	99.623						
6	.022	.317	99.940						
7	.004	.060	100.000						

提取方法：主成份分析。

降维分析----因子分析

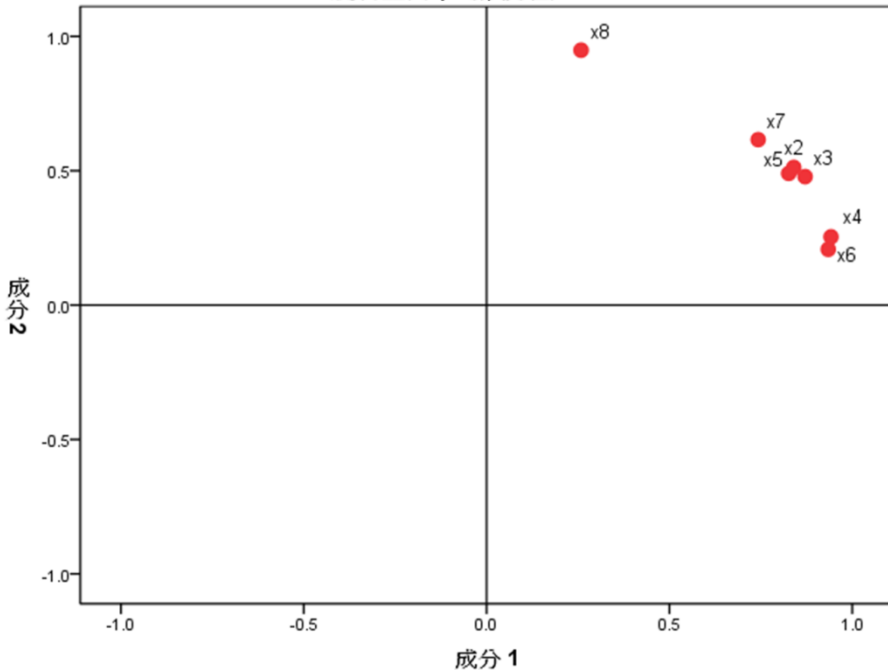
- 结果分析:

- **碎石图**表明, 从第三个因子开始, 特征值差异很小。



降维分析----因子分析

旋转空间中的成分图



旋转成份矩阵^a

	成份	
	1	2
投入人年数	.840	.512
投入高级职称的人年数	.871	.478
投入科研事业费	.934	.208
课题总数	.826	.490
专著数	.942	.254
论文数	.743	.616
获奖数	.258	.949

提取方法 :主成份。

旋转法 :具有 Kaiser 标准化的正交旋转法。

a. 旋转在 3 次迭代后收敛。

- 用少数几个因子去描述许多指标或因素之间的联系，即将相关比较密切的几个变量归在同一类中，每一类变量就成为一个因子，以较少的几个因子反映原资料的大部分信息。

降维分析---注意

- 可以看出，因子分析和主成分分析都依赖于原始变量，也只能反映原始变量的信息。所以原始变量的选择很重要。
- 另外，如果原始变量都本质上独立，那么降维就可能失败，这是因为很难把很多独立变量用少数综合的变量概括。数据越相关，降维效果就越好。
- 在得到分析的结果时，并不一定会得到清楚的结果。这与问题的性质，选取的原始变量以及数据的质量等都有关系。
- 在用因子得分进行排序时要特别小心，特别是对于敏感问题。由于原始变量不同，因子的选取不同，排序可以很不一样。

多元回归分析

- 一个被解释变量往往受多个解释变量的影响，利用具有因果关系的变量样本观测量，按照一定的实现原理来建立能够使被解释变量的计算值与实际值误差最小的回归方程，作为研究对象总体模型的估计参数。

多元回归分析

- 操作：在SPSS中，在数据视图窗口
 - 分析-回归-线性回归-保存-未标准化预测值
- 结果分析：
 - 判定系数为0.934，说明用自变量解释因变量变异的程度的结果较好。

模型汇总

模型	R	R 方	调整 R 方	标准 估计的误差
1	.967 ^a	.934	.924	915.2427

a. 预测变量：(常量)，课题总数，投入科研事业费，投入高级职称的人年数，投入人年数。

多元回归分析

Anova ^b						
模型		平方和	df	均方	F	Sig.
1	回归	3.096E8	4	77409921.707	92.411	.000 ^a
	残差	21779397.106	26	837669.119		
	总计	3.314E8	30			

a. 预测变量: (常量), 课题总数, 投入科研事业费, 投入高级职称的人年数, 投入人年数。

b. 因变量: 论文数

● 结果分析:

- 回归模型的方差分析表, F值为92.411, 显著性概率是0.000, 表明回归极显著。

多元回归分析

系数^a

模型	非标准化系数		标准系数	t	Sig.
	B	标准 误差	试用版		
1 (常量)	225.736	293.671		.769	.449
投入人年数	.487	.886	.239	.550	.587
投入高级职称的人年数	4.438	1.666	1.111	2.664	.013
投入科研事业费	.000	.000	-.301	-2.319	.029
课题总数	-.527	.764	-.133	-.691	.496

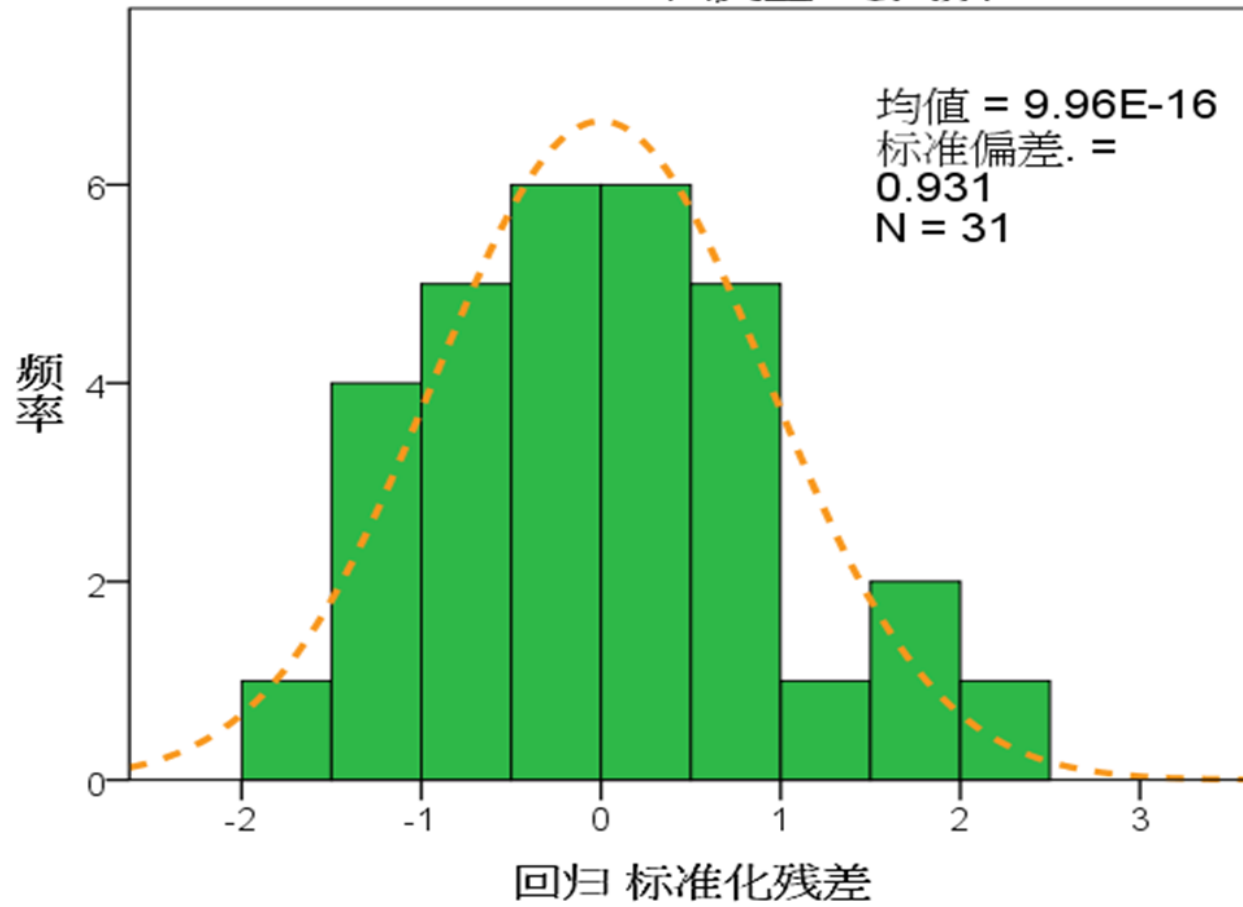
a. 因变量：论文数

● 结果分析：

- 多元回归模型的建立 $y = B_0 + B_1X_1 + B_2X_2 + B_3X_3 + B_4X_4$
- 回代检验

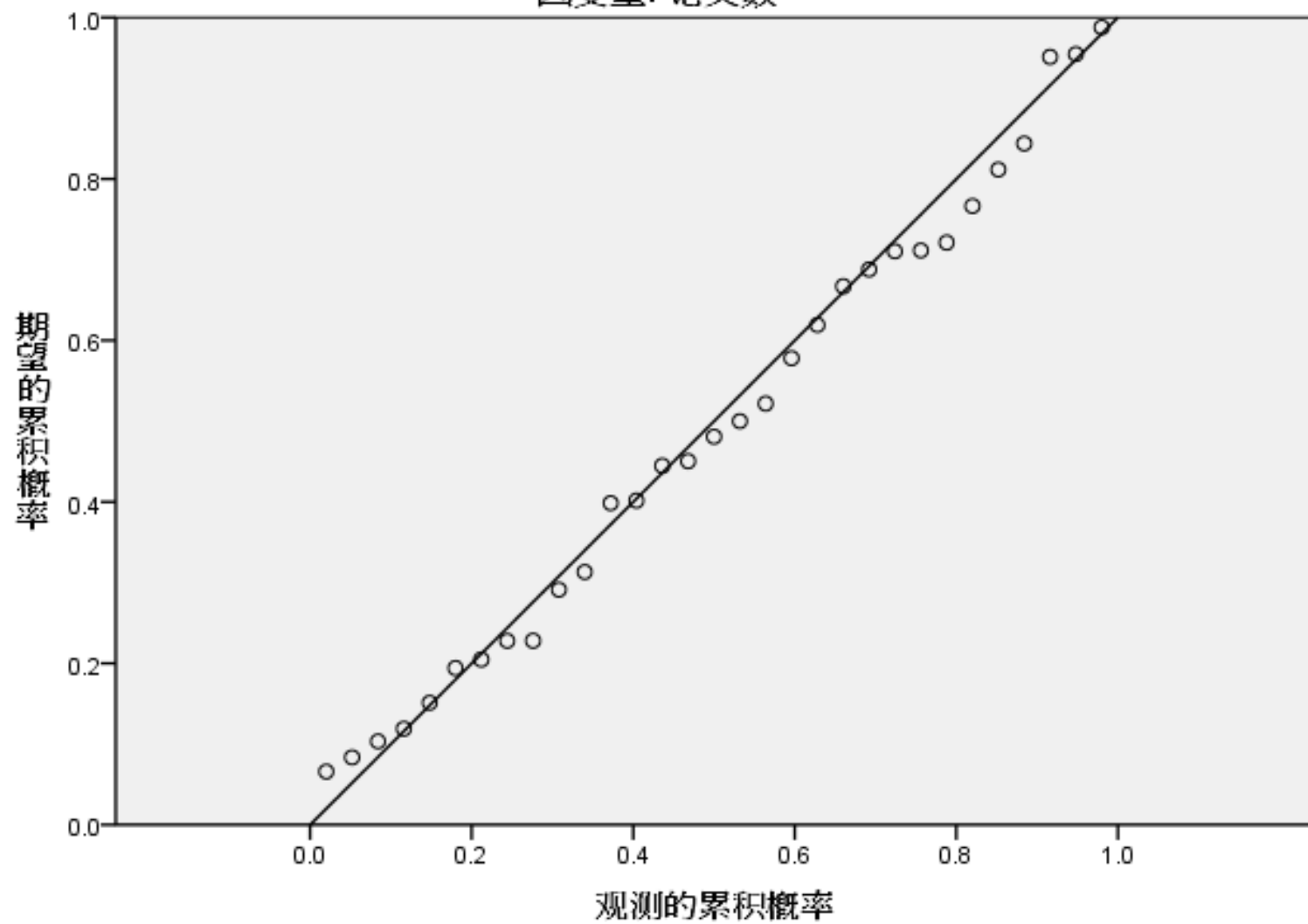
直方图

因变量: 论文数



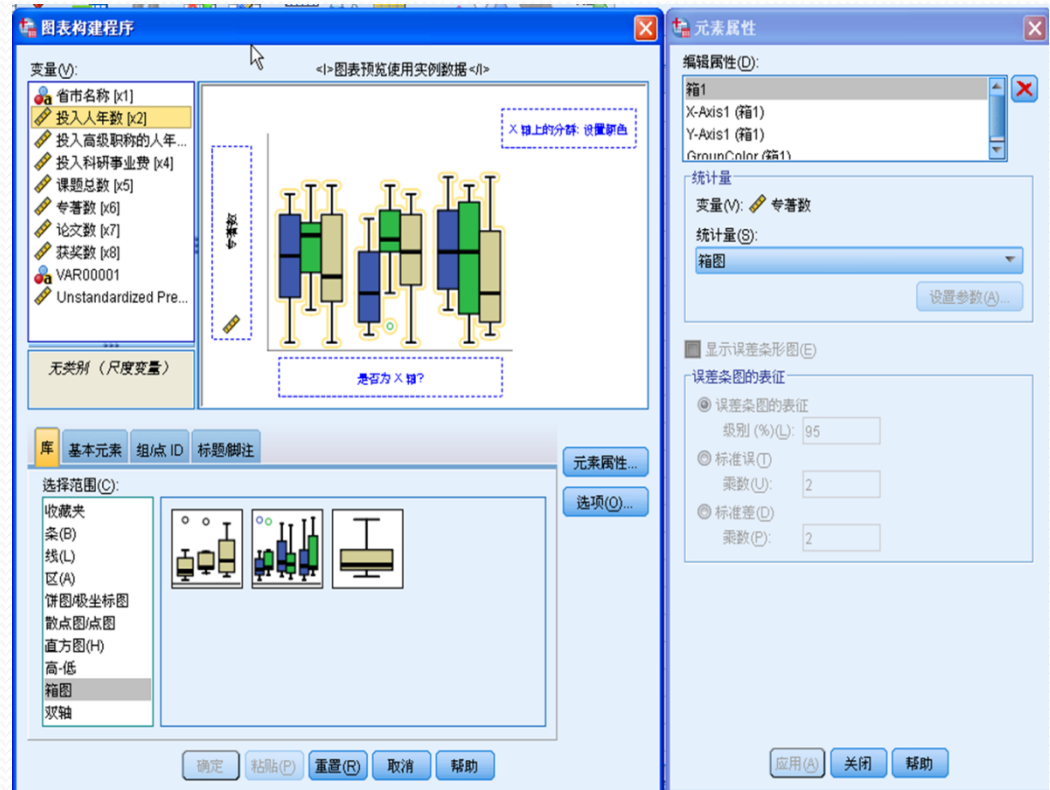
回归 标准化残差 的标准 P-P 图

因变量: 论文数



图形构建程序

- 操作：在SPSS中，在数据视图窗口
 - 图形-图表构建程序-选择x轴,y轴和元素
- 结果输出页面中，右键单击，可复制。



- 除了我们今天介绍的案例之外, SPSS还提供了从简单的统计描述到复杂的多因素统计分析方法, 比如
 - 数据的探索性分析、统计描述、列联表分析、二维相关、秩相关、偏相关、方差分析、非参数检验、多元回归、生存分析、协方差分析、判别分析、因子分析、聚类分析、非线性回归、Logistic回归等。

- 还有SAS和SYSTAT等优秀的统计软件供我们选择。SAS是功能非常齐全的软件，仍然需要一定的训练才可以进入，对于基本统计课程则不那么方便。
- STATA: 这是众多统计软件的后起之秀，具有数据管理软件、统计分析软件、绘图软件、矩阵计算软件和程序语言的特点。占用计算机系统资源少，绘图漂亮，对有简单编程基础者来讲非常容易上手，有专门出版的专业刊物。