

SPSS ZAI SHEHUI JINGJI FENXI ZHONG DE YINGYONG

普通高等学校“十一五”规划教材·经济科学

SPSS 在社会经济分析中的应用

王伏虎 编著

中国科学技术大学出版社

普通高等学校“十一五”规划教材·经济科学

SPSS在社会经济分析中的应用

王伏虎 编著

中国科学技术大学出版社

·合肥·

内 容 简 介

本书以 SPSS 16.0 版本为基础,主要介绍 SPSS 统计软件在社会经济领域中的应用。作者从统计专业用户的角度出发,通过生动的社会经济案例,讲述如何利用 SPSS 统计软件分析问题。书中不仅简要概述了 SPSS 统计软件的主要模块所涉及的统计分析原理和方法,而且介绍了 SPSS 统计软件的基本操作和常用技巧,重点在于如何利用 SPSS 统计软件分析实际的社会经济问题。

全书共 10 章,内容涉及社会学、经济学、心理学等众多领域,适合从事 SPSS 应用以及各类数据分析工作的读者阅读和使用。

图书在版编目(CIP)数据

SPSS 在社会经济分析中的应用/王伏虎编著. —合肥:中国科学技术大学出版社,2009.8

ISBN 978-7-312-02466-5

I. S… II. 王… III. 社会统计—统计分析—软件包 IV. F222.1

中国版本图书馆 CIP 数据核字(2009)第 043836 号

责任编辑:张善金 吴月红

出版者:中国科学技术大学出版社

地址:安徽省合肥市金寨路 96 号 邮编:230026

网址:<http://press.ustc.edu.cn>

电话:发行 0551-3602905 邮购 3602906 编辑 3602910

印刷者:合肥学苑印务有限公司

发行者:中国科学技术大学出版社

经销者:全国新华书店

开本:710mm×960mm 1/16

印张:24

字数:458 千

版次:2009 年 8 月第 1 版

印次:2009 年 8 月第 1 次印刷

定价:36.00 元



前 言

SPSS 是世界上最常用的统计软件之一,其应用领域涵盖了自然科学和社会经济科学的各个领域,由于其友好的用户界面和简单易学的操作方法深受广大用户的喜爱。

由于工作的原因,笔者在过去的几年间和许多使用 SPSS 统计软件的用户有所接触,并且了解到他们在使用 SPSS 统计软件时遇到的问题,主要有下面三个:第一,很多人都在使用 SPSS 统计软件处理日常工作中的一些数据,但是面对各种各样的实际问题和 SPSS 提供的众多统计分析方法无从下手,不知道应该选择 SPSS 的哪个模块来解决问题。第二,SPSS 统计软件功能非常强大,作为一个权威和严肃的专业统计软件,它为各种统计方法的检验以及数据的分析提供了众多的输出统计量,很多用户对输出统计量只知其然而不知其所以然,无法梳理出输出统计量之间的关系,没有办法探寻出社会经济现象之间的数量规律性。第三,目前 SPSS 统计软件方面的书籍不少,但是要寻找到一本既深入分析实际的社会经济问题,又介绍 SPSS 强大统计分析功能的书籍非常困难。正是从试图解决这三个问题的目标出发,笔者才有了撰写这本书的规划并花大力气完成了这一规划。

呈献给广大读者的本书共分 10 章,通过对具体的社会经济案例的分析介绍了 SPSS 在社会经济中的应用。本书最大的特点是将 SPSS 的基本操作、常用技巧和统计分析方法融合在一起,侧重于统计分析结果的解读,着力于读者在阅读本书以后能够举一反三,选择合适的统计分析方法解决实际问题,能够得心应手地使用 SPSS 统计软件处理日常工作,能够增强使用 SPSS 统计软件分析社会经济问题的能力。

本书在撰写过程中,得到江苏大学京江学院路正南教授、江苏大学



财经学院赵喜仓教授、江苏大学科技处董洁副教授的热心帮助和大力支持,在此一并感谢!

由于作者水平所限,书中不妥之处在所难免,恳请同行专家、学者和广大读者批评指正。如遇不解或疑难之处可与作者联系,联系方法:fhwang@ujss.edu.cn。

王伏虎

2009年3月于江苏大学财经学院



目 录

前言	(I)
第 1 章 数据文件的报表分析	(1)
1.1 在线分层分析(OLAP Cube)	(1)
1.2 个案汇总(Case Summaries)分析	(8)
第 2 章 SPSS 描述性统计分析	(14)
2.1 频数分析	(14)
2.2 描述性统计分析	(23)
2.3 探索性分析	(30)
2.4 列联表分析	(37)
2.4.1 名义变量的列联表分析	(37)
2.4.2 定序变量的列联表分析	(44)
2.4.3 列联表中的相对风险分析	(50)
2.4.4 列联表中的一致性检验	(57)
2.5 比率(Ratio)分析	(60)
第 3 章 均值比较	(65)
3.1 均值分析	(65)
3.2 单样本均值检验	(68)
3.3 独立样本均值检验	(71)
3.4 配对样本均值检验	(77)
3.5 单因素方差分析(One-Way ANOVA)	(79)
第 4 章 单变量方差分析	(89)
4.1 单变量的双因素方差分析	(89)
4.2 单变量的协方差分析	(98)



4.3	单变量的随机效应方差分析	(107)
4.4	单变量的混合效应模型	(112)
第5章	相关分析	(118)
5.1	双变量相关分析	(118)
5.2	偏相关分析	(129)
第6章	回归模型分析	(133)
6.1	一元线性回归模型	(133)
6.2	曲线估计	(143)
6.3	非线性回归	(148)
6.4	多元线性回归模型	(154)
6.5	二元逻辑斯蒂回归模型	(165)
6.6	多元逻辑斯蒂回归模型	(179)
第7章	聚类分析和判别分析	(191)
7.1	两步聚类分析	(191)
7.2	K-Mean 聚类法	(198)
7.3	系统聚类法	(206)
7.4	判别分析	(212)
第8章	主成分分析和因子分析	(231)
8.1	主成分分析	(231)
8.2	因子分析	(239)
第9章	时间序列分析	(258)
9.1	指数平滑法	(258)
9.1.1	简单指数平滑模型	(263)
9.1.2	霍特模型	(269)
9.1.3	温特斯模型	(273)
9.1.4	模型的预测	(277)
9.2	自回归模型	(282)
9.3	ARIMA 模型	(290)



9.3.1 ARIMA 模型的构成	(290)
9.3.2 ARIMA 模型的建模过程	(292)
9.4 季节分解	(307)
第 10 章 非参数检验	(315)
10.1 卡方检验	(315)
10.2 二项检验	(319)
10.3 游程检验	(324)
10.4 单样本科尔莫哥洛夫—斯米洛夫检验	(329)
10.5 两个独立样本的非参数检验	(334)
10.5.1 曼—惠特尼检验	(334)
10.5.2 双样本的科尔莫哥洛夫—斯米洛夫检验	(338)
10.5.3 Moses Extreme Reaction(Moses 极端反应)检验	(343)
10.5.3 Wald-Wolfowitz 游程	(346)
10.6 多个独立样本的检验	(350)
10.6.1 克鲁斯卡尔—沃利斯检验	(350)
10.6.2 中位数检验	(354)
10.6.3 Jonckheere-Terpstra 多样本检验	(357)
10.7 两个相关样本的非参数检验	(359)
10.7.1 威斯康星符号秩检验	(360)
10.7.2 符号检验	(362)
10.7.3 麦克内曼变化显著性检验	(365)
10.8 多个相关样本的非参数检验	(367)
10.8.1 弗里德曼检验	(368)
10.8.2 肯德尔协和系数	(370)
10.8.3 Cochran's Q 检验	(373)
参考文献	(376)



第1章 数据文件的报表分析

SPSS 的 Analyze 菜单的 Reports 中包含了 OLAP(在线分层分析)、Case Summaries(个案汇总分析)、Reports Summaries in Rows(行报告汇总分析)和 Reports Summaries in Columns(列报告汇总分析),后面两个过程的功能比较相似,分别按行和按列的形式输出报表,可以生成比较复杂的统计报表,但是输出的表格是文本形式的,因此本章主要介绍 OLAP 和 Case Summaries。

1.1 在线分层分析(OLAP Cube)

SPSS 的 Online Analytical Processing(在线分层分析)主要用来分析尺度变量(Scale Variable),也就是所谓的数值变量,在 SPSS 的 OLAP Cube 对话框中,它们被称为 Summary Variable(汇总变量)。我们可以选择输出尺度变量的多种统计量,从简单的计数、求和、比例等统计量到集中趋势、离散趋势等复杂的描述性统计量。利用在线分层分析时必须选择分组变量(Grouping Variable),输出结果中的 OLAP Cube(在线分层表)中,分组变量以层字段(Layer)的方式显示在表格的左上角,待分析的尺度变量是作为行字段(Row),以行的形式显示出来,描述性统计量作为列字段(Column),以列的形式显示出来。通过 SPSS 提供的 Pivoting Tray 功能,在 OLAP Cube 中可以很方便地实现层字段、列字段和页字段的转化,这一功能非常类似于 Excel 中的数据透视表。我们还可以在描述性统计量表中设置表格的标题和脚注。

【例 1.1】 电信公司客户流失情况分析。

一家电信公司想减少客户流失率,也就是客户转向其他电信供应商的比例。客户流失率和客户的年龄、家庭收入、使用电信公司业务的数量及时间以及月消费金额都可能存在一定的关系。

打开“电信公司客户流失率.sav”。使用在线分析过程对客户流失率进行分析。数据文件中使用的变量名、变量标签值及变量类型见表 1.1。



表 1.1 数据文件中使用的变量名、变量标签值及变量类型

变量名 Variable name	变量标签值 Value label	变量类型
区域编号	1=区域 1,2=区域 2 3=区域 3,4=区域 4 5=区域 5	名义变量
使用电信业务的时间		尺度变量
年龄		尺度变量
家庭收入		尺度变量
长途电话消费额		尺度变量
免费电话消费额		尺度变量
设备租赁消费额		尺度变量
电话卡消费额		尺度变量
无线业务消费额		尺度变量
多线业务	0=没有使用 1=使用	名义变量
语音信箱业务	同上	名义变量
传呼业务	同上	名义变量
互联网业务	同上	名义变量
来电显示业务	同上	名义变量
呼叫等待业务	同上	名义变量
呼叫转移业务	同上	名义变量
三方通话业务	同上	名义变量
电子帐单业务	同上	名义变量
客户流失	0=流失 1=没有流失	名义变量

数据来源:SPSS16.0 Samples 中的 telco.sav

在分析之前首先要生成一个变量来测度每位客户到底使用了电信商提供的哪些服务。单击 Transform→Count, 打开 Count Occurrences of Values within Cases 对话框, 如图 1.1 所示。在对话框中, Target Variable 下输入“电信业务”, 在 Target Label 下输入“实际使用的电信业务”, 将“多线业务”到“电子账单业务”



之间的9个变量移入右侧的 Numeric Variables 方框下,再单击其下方的 Define Values 按钮,弹出 Count Values within Cases: Values to Count 对话框(见图 1.2)。在 Data View 中可以发现,上述9个变量对应的数值为1或者0,也就是说使用其中一项业务,其对应的数值为1,否则为0。因此要对客户使用的电信业务进行计数,需要在 Value 后的文本框中输入1,单击右侧的 Add 按钮,则 Values to Count 下方显示1,单击 Continue 按钮,回到 Count Occurrences of Values within Cases 对话框。单击 OK 按钮,返回到 Data View 窗口会发现生成了一个新变量“电信业务”。

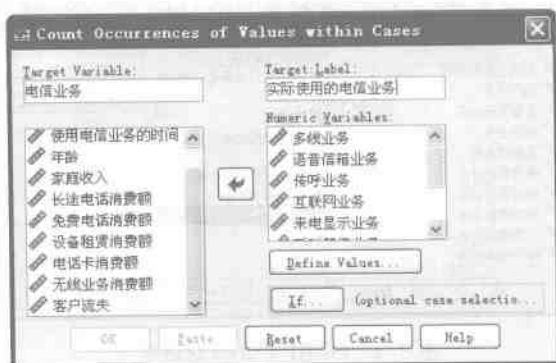


图 1.1 Count 对话框的设置



图 1.2 Define Values 对话框



图 1.3 Compute Variable 对话框

进一步还需要测度客户每月使用各类电信业务的平均消费金额。单击 Transform→Compute, 打开 Compute Variable 对话框(见图 1.3), 在 Target Variable 下方文本框中输入“月消费金额”。



在 Numeric Expression 中输入表达式“(长途电话消费额 + 免费电话消费额 + 设备租赁消费额 + 电话卡消费额 + 无线业务消费额) / 使用电信业务的时间”，单击 OK 按钮，生成名为“月消费金额”的变量。

单击 Analyze → Reports → OLAP Cubes，打开 OLAP Cubes 对话框(见图 1.4)，从左侧的变量列表框中选择“使用电信业务的时间”、“年龄”、“家庭收入”、“电信业务”和“月消费金额”这五个尺度变量，移入右侧的 Summary Variable 方框中。选择“区域编号”和“客户流失”作为分组变量，移入右侧的 Grouping Variable 方框中。



图 1.4 OLAP Cubes 对话框



图 1.5 在 OLAP Cubes: Statistics 对话框中选择输出的统计量



图 1.6 在 OLAP Cubes: Title 对话框设置出图表的标题

单击 Statistics 按钮，弹出 OLAP Cubes: Statistics 对话框(见图 1.5)，从 Cell Statistics(单元格统计量)列表框中移出系统默认的 Sum、Percent of Total Sum 和 Percent of Total N 这三个统计量，从左侧的 Statistics 框中选择 Median，移入



Cell Statistics 列表框中。

单击 Continue 按钮,回到 OLAP Cubes 对话框中(见图 1.6),单击 Title 按钮,弹出 OLAP Cubes: Title 对话框,在 Title 下方输入汇总表的标题“电信公司客户流失分析”,在 Caption 下方输入标题说明“根据客户是否流失及区域编号分析”,单击 Continue 按钮,回到在线分析对话框,单击 OK 按钮,输出在线分析结果,见表 1.2。

表 1.2 在线分析输出的汇总表
电信公司客户流失分析

区域编号	Total				
客户流失	Total				
	没有流失	N	Mean	Std. Deviation	Median
使用电信业务	流失	1000	35.53	21.360	34.00
年龄	Total	1000	41.68	12.559	40.00
家庭收入		1000	77.535 0	107.044 16	47.000 0
实际使用的电信业务		1000	3.740 0	2.601 38	3.000 0
月消费金额		1000	62.323 0	50.041 73	48.895 6

根据客户是否流失及区域编号分析。

从输出的汇总表中可以看出,两个分组变量“区域编号”和“客户流失”显示在表格的左上角(Layer,即页字段)。如果分析时需要了解某个分组变量中某个类别的描述性统计量,比如要分析客户“流失”的统计量,可以双击表格,将它激活。单击分组变量“客户流失”右侧的下拉箭头,从弹出的下拉列表中选择“流失”就可以实现这一目的。

表 1.3 显示“流失”客户的汇总表^①

区域编号:Total

客户流失:流失

	N	Mean	Std. Deviation	Median
使用电信业务的时间	274	22.43	17.87	17.00
年龄	274	36.52	10.522	35.00
家庭收入	274	61.627 7	60.970 78	41.000 0
实际使用的电信业务	274	4.219 0	2.774 07	4.000 0
月消费金额	274	60.990 7	48.558 70	48.966 7

根据客户是否流失及区域编号分析。

表 1.3 显示,在所有区域中有 274 位客户流失。在流失的客户中,变量“使用

^① 本书的表格本应统一为表 1.1 的形式,但由于 SPSS 软件本身就设置为如表 1.1 和表 1.3 两种类型,为保持与软件设置一致,方便读者使用,本书未强调表格形式统一。相似之处不再说明。



电信业务的时间”、“家庭收入”和“月消费金额”的均值都明显地大于中位数,有一个向右的尾部,这三个变量应该服从右偏分布。

在输出的 OLAP Cubes 报表中,可以通过 SPSS 提供的 Pivoting Trays 功能(见图 1.7),实现表格中层字段、行字段和列字段的相互转换。双击 OLAP Cubes 报表将其激活会弹出 Pivoting Trays 对话框(如果没有弹出,在 SPSS Viewer 中单击 Pivot→Pivoting Trays)。将分组变量“客户流失”的图标从 Layer 托盘拖动到下方的 Row 托盘上,将 Statistics 图标从右侧的 Column 托盘中拖动到 Layer 托盘中,将分组变量“区域编号”的图标从 Layer 托盘中拖动到 Column 托盘中,单击 Pivoting Trays 对话框右上角的按钮,关闭该对话框。

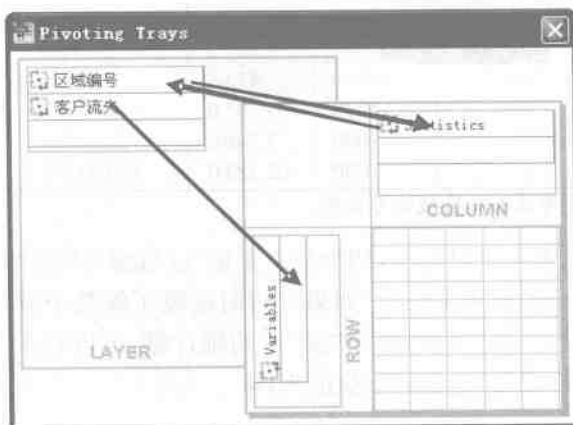


图 1.7 在 Pivoting Trays1 对话框中改变行变量、列变量和分层变量的位置

在编辑后的报表中,单击统计量右侧下拉箭头,从弹出的下拉列表清单中选择 Mean(均值),输出结果见表 1.4。表中最后的“Total”列显示,根据地理区域分类的客户中,“流失”客户使用电信业务的平均时间为 22.43 个月,几乎是“没有流失”客户使用电信业务平均时间的一半,这表明使用电信业务的时间越长,流失的可能性越小。最后一列显示“没有流失”客户的平均年龄是 43.63 岁,而“流失”客户的平均年龄是 36.52 岁,说明年轻客户相对容易流失。“没有流失”客户的平均家庭收入为 83.5386,高于“流失”客户的家庭平均收入,家庭收入越高,客户的流失倾向越小。“没有流失”客户实际使用电信业务的平均值为 3.5592 项,小于“流失”客户的平均值,这表明客户使用的电信业务越多,反而越有可能流失。“没有流失”客户月平均消费金额为 62.8259,“流失”客户月平均消费金额为



60.9907,这两类客户的月平均消费金额相差不大,但是在前面对“流失”客户进行分析时发现“使用电信业务的时间”、“家庭收入”和“月消费金额”这三个变量都是右偏的分布,使用均值来对比分析未必非常合适。因此,有必要分析分组变量的中位数。

表 1.4 行变量、列变量和分层变量位置发生改变后的描述性统计量表
电信公司客户流失分析

Mean		区域编号			
	客户流失	区域 1	区域 2	区域 3	Total
使用电信业务的时间	没有流失	39.85	40.25	41.24	40.47
	流失	23.00	23.05	21.21	22.43
	Total	35.14	35.41	36.00	35.53
年龄	没有流失	42.89	43.88	44.08	43.63
	流失	37.59	36.32	35.66	36.52
	Total	41.41	41.75	41.88	41.68
家庭收入	没有流失	77.4569	85.8375	86.9213	83.5386
	流失	63.0889	62.3191	59.4444	61.6277
	Total	73.4410	79.2186	79.7326	77.5350
实际使用的电信业务	没有流失	3.5733	3.5000	3.6024	3.5592
	流失	4.2111	4.2234	4.2222	4.2190
	Total	3.7516	3.7036	3.7645	3.7400
月消费金额	没有流失	62.3930	63.1913	62.8760	62.8259
	流失	61.4338	60.2713	61.2990	60.9907
	Total	62.1249	62.3695	62.4634	62.3230

根据客户是否流失以区域编号分析。

双击报表将其激活,单击统计量右侧下拉箭头,从弹出的下拉列表清单中选择 Median(见表 1.5)。

重新编辑的报表显示“年龄”和“实际使用的电信业务”这两个变量反映的信息和前面基本一致。“没有流失”客户和“流失”客户在使用电信业务的平均时间上仍旧存在较大差距,而且中位数显示出来的差距似乎更大。从表中还可以看出,居住在“区域 1”的“流失”客户,他们的收入中位数为 51,高于没有流失的客户(47),这和均值方面反映的信息正好相反。最后我们还可以发现,在使用电信业务的时间方面,“流失”客户和他们居住的区域之间似乎存在相互作用的关系。从“区域 1”到“区域 3”,“流失”客户使用电信业务的时间从原先的 20.50 逐渐减小到 12.50。因此在分析那些流失客户时,地区是一个值得考虑的因素。



表 1.5 显示各组中位数的描述性统计量表

电信公司客户流失分析

Median

		区域编号			
		区域 1	区域 2	区域 3	Total
使用电信业务的时间	客户流失				
	没有流失	40.50	42.00	42.00	41.50
	流失	20.50	17.50	12.50	17.00
	Total	33.00	34.50	35.00	34.00
年龄	没有流失	42.00	43.00	43.00	43.00
	流失	35.00	35.00	32.50	35.00
	Total	40.00	41.00	40.00	40.00
家庭收入	没有流失	47.0000	52.0000	49.0000	49.0000
	流失	51.0000	40.0000	37.5000	41.0000
	Total	47.5000	48.0000	46.0000	47.0000
实际使用的电信业务	没有流失	3.0000	3.0000	3.0000	3.0000
	流失	4.0000	3.0000	4.0000	4.0000
	Total	3.0000	3.0000	4.0000	3.0000
月消费金额	没有流失	49.6790	50.8417	46.8825	48.8956
	流失	49.2440	50.5370	46.3302	48.9667
	Total	49.5103	50.7997	46.5078	48.8956

根据客户是否流失及区域编号分析。

1.2 个案汇总(Case Summaries)分析

SPSS 的个案汇总也是对尺度变量进行分析的,使用个案汇总分析时,我们可以选择分组变量(Grouping Variable),也可以不选择分组变量。个案汇总的功能比在线分层处理要强大得多,它可以计算数值变量的描述性统计量,选择分组变量以后,还可以对分组变量进行分类汇总。在默认状态下它还会输出个案的相关信息。

【例 1.2】产品订单金额分析。

一家公司为了了解该公司全年的订单金额情况,收集整理了公司 9 名销售员的数据。分析人员使用个案汇总命令生成销售员的订单金额报表数据,对销售员的业绩进行分析。本例中使用的变量名、变量标签值及变量类型见表 1.6。



表 1.6 例 1.2 使用的变量名、变量标签值及变量类型

变量名 Variable name	变量标签值 Value label	变量类型
国家和地区		名义变量
销售人员	1=付乐宜,2=金永昌,3=董向阳 4=戴润林,5=苏术平,6=曹汉 7=金克农,8=潘金,9=李卫林	名义变量
月份		名义变量
订单号		尺度变量
订单金额		尺度变量

打开“产品订单金额.sav”。单击 Analyze→Reports→Case Summaries, 打开个案汇总分析对话框(见图 1.8)。将“订单金额”移入右侧的 Variables 框中, 将“销售人员”移入右侧的 Grouping Variables 方框中。我们首先输出分组汇总表, 不需要显示个案列表, 此时取消左下角 Display cases 复选框前的勾号“√”。

单击 Statistics 按钮, 在统计量对话框中, 从左侧的统计量列表框中选择 Mean、Median、Minimum、Maximum 移入右侧的 Cell Statistics 方框中(见图 1.9), 其中 Number of Cases 是默认的统计量。单击 Continue 按钮, 回到个案汇总对话框。



图 1.8 Summarize Cases 对话框中的设置



图 1.9 在统计量对话框中选择输出的统计量

单击 Options 按钮, 打开报表选项对话框(见图 1.10)。在 Title 下方文本框