

常用统计分析方法 ——~~理论~~应用

杜志渊

山东人民出版社



总序

管理的实践可以追溯到遥远的古代人类文明,但直到 19 世纪初泰勒的开创性贡献——《科学管理原理》一书的问世,才标志着人类告别经验管理时代,进入了科学管理时代。19 世纪 20 年代以来,由于世界经济环境的发展变化,科学技术尤其是信息技术的突破性进展和大范围的应用,市场竞争的日趋激烈和国际化思潮风起云涌,都推动着现代管理思想、管理理论、管理方法和管理手段的日新月异。在 21 世纪,全球的管理者,无疑面对着更大范围的管理创新和变革。如何迎接这一世纪性的挑战,则成为世界各国政府、企业界和理论界共同关注的课题。

回顾上一个世纪,管理理论经历了古典管理理论阶段、行为科学阶段和现代管理理论阶段的演变,形成了庞大的知识体系。进入 20 世纪 80 年代后,随着知识经济的崛起和全球经济一体化进程的加快,市场环境的变化更加迅速,导致管理学家面临许多前所未有的新情况和新问题。于是当代管理新思潮也竞相涌现。对于历经改革开放风雨洗礼,正在进入世界大市场、融入世界经济大循环的中国来说,更是面临进入市场经济后复杂的现实管理问题。

在 21 世纪未来的岁月里,中国的企业、各级政府、事业团体等各类组织在管理方面所面对的主要挑战在于:

1. 变革与创新。今天的变化快过以往任何时候,原来的流程管理和职能管理方法难以适应当今的变化。世界在发展,新的问题层出不穷,需要运用已有的理论和方法去解决,更需要我们去大胆探索和创新。创新是管理理论发展的路径,是任何组织立于不败之地的法宝。

2. 国际化。互联网的兴起无疑大大改变了当今的经营环境。这是过去一个多世纪没有、也不可能预见到的。国际化的进程要求我们认真研究全球化战略,而不仅仅是国际化或跨国界运作战略。全球一盘棋的思想、跨文化的研究和应用必须扎根于当今领先企业领袖的头脑中。我们知道东方文化与西方文化是有很大的差异的,这与东西方人的思维方式的差异有关:感性与理性;严谨与灵性;



实验验证与感悟修证。这些差异反映在管理领域,产生了很多耐人寻味的现象。

知识经济。知识的管理大大超过了过去一个世纪管理学者的想像。这里不是为知识而知识的管理,而应该是如何真正把知识应用于组织,实现知识转移、知识共享,推动组织的发展。

网络技术。网络技术的应用是当今组织管理者尤其应该认真思考的问题。如何利用网络技术提高效率,节省成本,改善沟通,强化协调等等,有许多的新内容值得我们去研究。

面对这些挑战,谁能够最快地吸收各种管理新知识,谁就会获得竞争的主动权,谁拥有更多的知识,谁能够通过管理创新把更多的知识组合成独特的能力,谁就能赢得未来。

作为管理教育的从业者,我们理所应当承担起为我国管理科学的发展添砖加瓦的责任。一方面紧跟国际潮流,逐渐实现管理教育的规范化、国际化;另一方面则必须面对改革开放的丰富实践,推动管理教育创新。这就是我们编写这套系列教材的出发点。

《21世纪管理学系列教材》主要是针对管理类各专业本科生教学编写的,亦可作为工商管理硕士(MBA)和经济、管理类各专业研究生参考书。教材的编写突出以下特点:

扩大信息量,同时给教师授课留有一定的空间。作为管理类本科生、研究生专业课或基础课教材,要让学生掌握较大的信息量,管理学理论的重要成果不宜疏漏,同时,对于比较成熟理论的阐述,不要过细,让老师有充分的发挥余地。教材使用能达到教师好教、学生好学的目的。

反映管理学的深邃与智慧。不过分追求“通俗易懂”,突出体现内容的专业性、学术性。突出管理学的哲学价值。管理过程充满着哲学道理,如决策中的“概率思考”,领导中的“权变观念”,管理控制中的“反馈原理”等,在阐述管理的基本问题时,体现管理学科的哲学性和科学性。

反映管理学的最新理论进展。随着信息技术和知识经济的发展,传统的管理思想和管理方式受到了挑战。教材编写追踪管理理论前沿,反映管理学科的发展,体现理论的时代性和前沿性。

处理好借鉴与创新的关系。教材编写既借鉴已有的理论成果,又注重理论的创新性,注重把作者自己的最新的理论成果写入教材。

本系列教材是山东大学管理学院教师们多年教学实践和科学研究成果的结晶。教材编写委员会统一指导各书编写,选择确定各书主编,精心组织参编教师,审查把关编写质量。力求使每种教材不但适合当前管理专业本科教学,同时也符合21世纪管理学科教育的发展大势。



编写一套贴近管理学科理论前沿、广泛介绍管理各学科成熟内容的教材,必然要参考国内外大量已有的研究成果,吸收近百年来管理理论的精华。我们谨向一切致力于管理学科理论繁荣的前辈与同仁致以崇高的敬意!

徐向艺

2023年 11月 15日



前 言

统计学是一门方法论科学,同时也是一门与实际密切联系的应用学科,它已经成为高校许多学科本科生或者研究生的必修课程。

在统计学原理的学习和统计方法的应用过程中,经常需要进行大量的计算。所以,统计分析软件问世使计算机的运算功能得到充分发挥,不仅能够减轻计算工作量,而且还提高了计算的准确性,大大地节省了统计分析时间。因此,利用统计分析软件进行数据处理已经成为社会学家和科学工作者必须掌握的一项基本技能。为了使高校学生能够适应社会发展的要求,在掌握统计学的基础上,学习一种统计分析软件进行各种统计分析是十分必要的。本书把世界上应用最普遍的统计分析软件之一——~~Excel~~处理数据和分析数据的基本方法编辑成册,供高校学生及应用统计分析软件相关人员学习和参考。

本书以统计学原理为理论基础,以高等学校本科生学习的常用统计方法为主要内容,重点介绍数据的基本统计描述、均值的检验和方差分析、相关分析和回归分析、时间序列分析、非参数检验等基本统计方法的~~Excel~~软件应用。为了便于理解,每一种分析方法结合一个例题,详细解释~~Excel~~软件的操作步骤,并且对统计方法的输出结果进行相应的分析。例题结合工业、农业、商业、医疗卫生、文化教育等实际问题,力求使学生对统计分析方法应用有更深刻的认识和理解,以提高学生学习兴趣和学习的主动性。

书中所有例题和练习题,制成了~~Excel~~格式下的数据文件库光盘附在书后,为读者在学习及~~Excel~~统计分析软件的实际操作中提供参考。另外,为了方便学习者的查询,本书特编写了附录 1 部分常用统计量公式,附录 2 ~~Excel~~主要函数。希望这些内容对学习有所帮助。

编者
2005 年 1 月



目 录

第一章 数据文件的建立及基本统计描述 (员)

第一节 数据的启动及数据库的建立 (员)

一、数据简介 (员)

二、启动 数据软件包 (猿)

三、数据文件的建立 (源)

第二节 数据的编辑与整理 (远)

一、数据窗口菜单栏功能操作 (远)

二、数据整理功能 (苑)

三、数据或变量的变换及转换功能 (愿)

四、数据的编辑 (怨)

五、变量的编辑 (员缘)

第三节 基本统计描述 (员怨)

一、描述统计分析过程 (员怨)

二、频数分析 (圆)

三、探索分析过程 (圆四)

第四节 交叉列联表分析 (猿)

一、交叉列联表的形成 (猿)

二、两变量关联性检验 (猿四)

第二章 均值比较检验与方差分析 (源)

第一节 单个总体的 检验 (源)

第二节 两个总体的 检验 (源)

一、两个独立样本的 检验 (源)

二、两个有联系样本的均值比较 (源)

第三节 单因素方差分析 (缘)

第四节 双因素方差分析 (缘)

第三章 相关分析与回归模型的建立与分析 (远)



常用统计分析方法——~~杂~~应用

第一节 相关分析(悦测查)	(远)
一、简单相关分析	(远)
二、偏相关分析	(远)
第二节 线性回归分析(硬测查)	(苑)
一、线性回归模型假设条件与模型的各种检验	(苑)
二、线性回归分析的具体步骤	(苑)
第三节 曲线估计(悦测查)	(苑)
第四章 时间序列分析	(愿)
第一节 时间序列的定义与图形分析	(愿)
一、根据时间数据定义时间序列	(愿)
二、绘制时间序列线图和自相关图	(愿)
第二节 季节变动分析(硬测查)	(怨)
一、季节分析方法	(怨)
二、进行季节调整	(怨)
第五章 非参数检验	(怨)
第一节 卡方检验(悦测查)	(怨)
第二节 一个样本的 运 原杂检验	(员)
第三节 两个独立样本的检验(裁测查)	(员)
第四节 两个有联系样本的检验(裁测查)	(员)
第五节 多个样本的非参数检验(运原)	(员)
一、多个独立样本的单因素方差分析(裁测查)	(员)
二、多个有联系样本的方差分析(运原)	(员)
第六节 游程检验(硬测查)	(员)
附录员 部分常用统计量公式	(员)
一、数据的基本统计特征描述	(员)
二、总体均值检验统计量	(员)
三、方差分析中的统计量	(员)
四、回归分析模型	(员)
五、非参数检验	(员)
附录圆 杂 函数	(员)
主要参考文献	(员)



第一章 数据文件的建立及基本统计描述

在社会各项经济活动和科学研究过程中,经常获得许多数据,而这些数据中包含着大量有用的信息。若要准确地、科学地提取这些信息,就要应用各种统计分析方法,其中最基本的方法是数据的基本统计描述。通过数据的基本统计描述,可以得到数据的分布状况、数据的主要特征值、时间序列的趋势性、是否存在异常值以及数据的大致图形等。当然,要实现数据的统计分析和描述,首先要从建立数据文件开始。这一章主要介绍数据文件的建立和数据的基本统计描述方法。

第一节 杂多的启动及数据库的建立

一、杂多简介

杂多(Shadoc)是一种运行在 Windows 系统下的社会科学统计软件包。杂多软件包集数据整理、分析过程、结果输出等功能为一体,采用窗口操作界面,具有统计分析方法涵盖面广、用户操作使用方便、输出数据表格图文并茂的特点。随着它的功能不断完善,统计分析方法不断充实,大大提高了统计分析工作的效率。从 1985 年由美国斯坦福大学开发使用至今,已经拥有全球数以万计的用户,分布在通信、医疗、银行、证券、保险、制造、商业、市场研究、科学教育等众多的行业和领域,成为世界上应用最广泛的专业统计软件之一。

杂多的基本功能包括数据管理、统计分析、图表分析、输出管理等,具体内容包括描述统计、列联分析、总体的均值比较、相关分析、回归模型分析、聚类分析、主成分分析、时间序列分析、非参数检验等多个大类,每个类中还有多个专项统计方法。杂多没有专门的绘图系统,可以根据使用者的需要将给出的数据绘



制各种图形,能够满足用户的不同需求。

(一) SPSS的运行方式

SPSS提供了三种基本运行方式:完全窗口菜单方式、程序运行方式、混合运行方式。程序运行方式和混合运行方式是使用者从特殊的分析需要出发,编写自己的 SPSS命令程序,通过语句直接运行。这里只介绍完全窗口菜单管理方式,这种操作方式简单明了,除数据输入工作需要键盘外,大部分的操作命令、统计分析方法的实现是通过菜单、图标按钮、对话框来完成的,非常适用于一般的统计分析人员和一般统计方法的应用者。

SPSS中使用的对话框主要有两类,一类是文件操作对话框,文件操作对话框窗口操作与 Windows应用软件操作风格一致。另一类是统计分析对话框,统计分析对话框可以分为主窗口和下级窗口,在该类对话框中,选择参与分析的各类变量及统计方法是对话框的主要任务。有关对话框的详细操作将在后面介绍统计方法的例题中加以解释。

(二) SPSS的实验环境要求

1. 系统运行环境。SPSS 5.0以上版本软件包可以工作在两种模式下,单机模式和作为网络系统的用户界面模式。

2. 数据库。SPSS软件包可运行在微软公司的 Windows、Unix、Macintosh、OS/2、VME和 SPARC等操作系统之下。由于统计分析软件的数据量比较大,所以系统运行需要大于 100M以上空间。

3. 辅助软件环境。SPSS可以直接将 SPSS数据文件保存为 Lotus工作表,也可以直接打开一个 Lotus工作表,因此,为了方便数据录入(许多人对 Lotus工作表编辑比较熟悉),应在操作系统下安装一个 Lotus软件。另外,许多数据在处理之前可能保存在某个数据库中,例如 dBase、Paradox、Access、Inform、Visual Basic等等,如有需要从数据库中获取数据的分析,应在操作系统下安装相应的数据库管理系统。

(三) SPSS的主要界面

SPSS的主要界面有数据编辑窗口和结果输出窗口。数据编辑窗口与微软的 Lotus类似,但 SPSS的统计功能更多。SPSS的结果输出窗口是显示统计分析的结果,此窗口的内容可以以结果文件输出的形式保存。数据编辑窗口和结果输出窗口的详细描述将在有关 SPSS的数据文件建立的内容中查到。

(四) SPSS的帮助系统

SPSS对一些基本模块中的统计提供了帮助,可以通过单击 Help菜单中的 SPSS命令,选择所需要的统计指导。



二、启动 SPSS 软件包

当用户在操作系统下运行 SPSS 软件后,计算机屏幕上出现一个对话框,如图 1-1 所示:

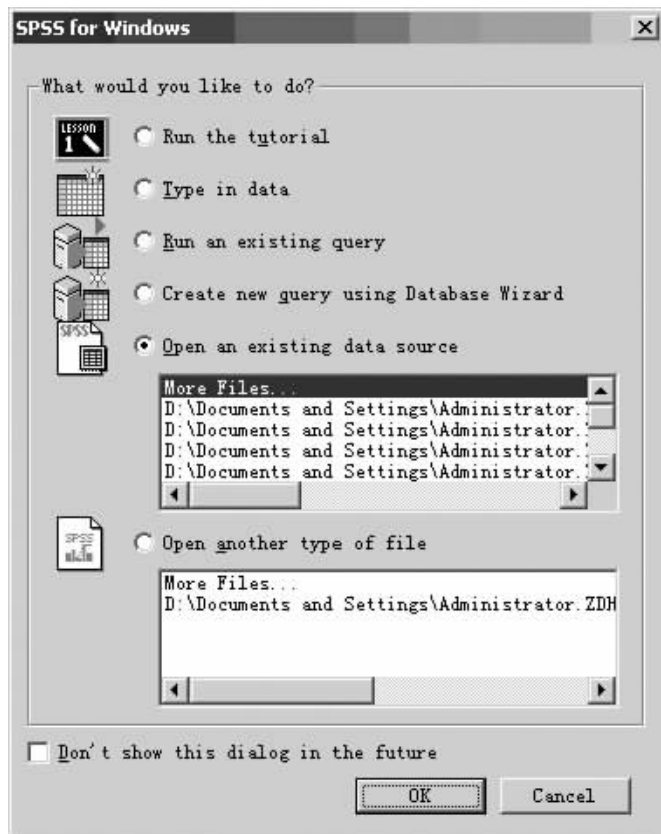


图 1-1 SPSS 启动后操作对话框

对话框包括一个六选一单选指令和一个复选指令,其内容为:

- 运行 SPSS 运行操作指南;
- 键入数据选项,建立新的数据集时可选择此项;
- 运行一个已经存在的数据文件选项;
- 使用数据库处理工具建立新文件;
- 打开一个已经存在的数据文件;
- 打开其他类型的文件。
- 复选对话框,选中该复选项后,下次启动 SPSS 时不再显示该对话框。



再次启动 SPSS 时将不会显示对话框 ,直接显示数据编辑窗口。

三、数据文件的建立

当对话框选择 数据源为数据库后 ,点击 确定,系统将显示出 SPSS 软件包数据编辑主窗口 ,数据文件的建立就是在数据编辑窗口中完成的。数据编辑窗口可以显示两张表 ,分别是 数据输入区(见图 11-10)和 数据编辑区(见图 11-11) ,通过点击下端的 两个同名窗口标签按钮实现相互切换。

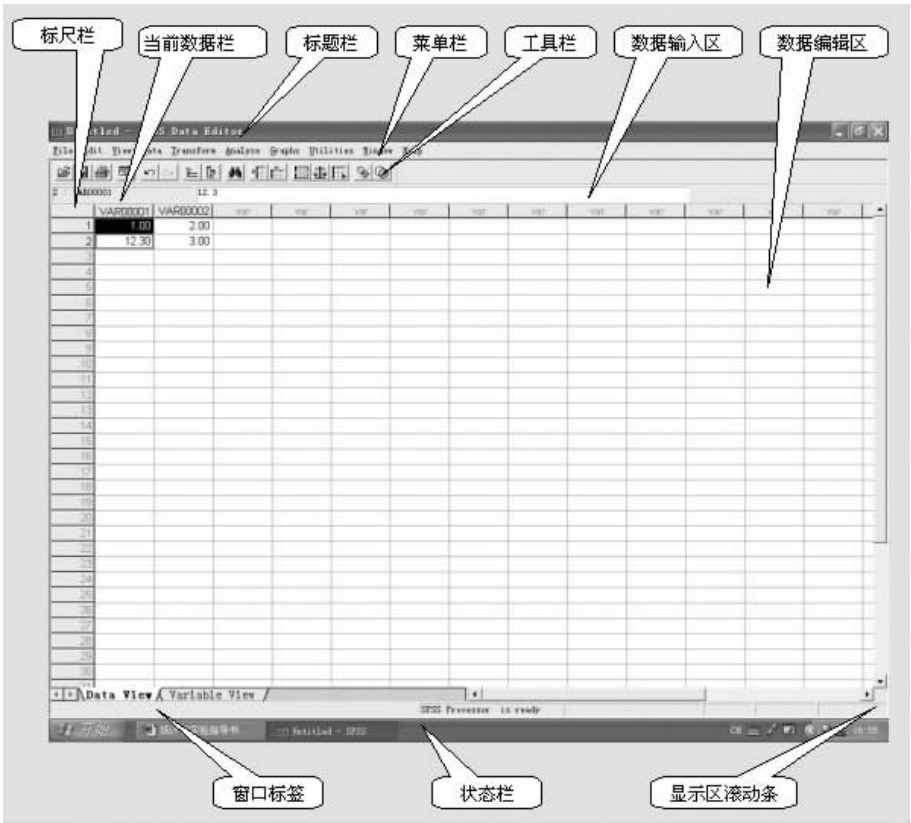


图 11-10 SPSS 数据编辑主窗口示意图

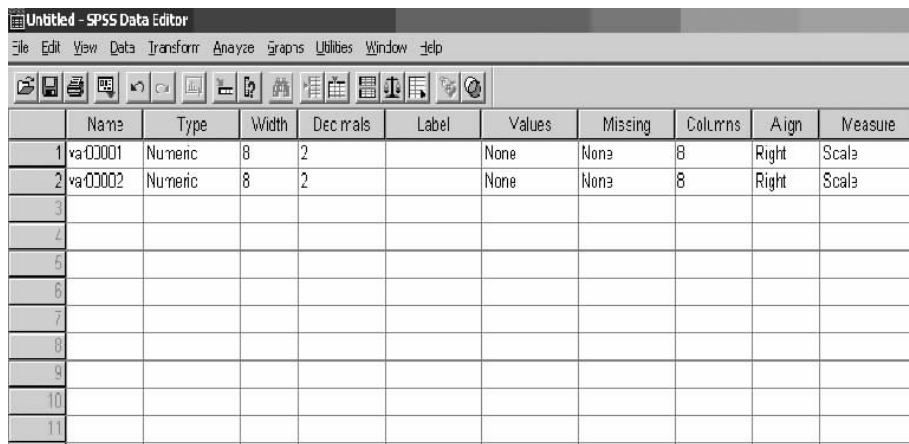
数据编辑区是一个二维平面表格 ,是 SPSS 的主要操作窗口 ,用于对数据进行各种编辑如图 11-11 所示。标尺栏有纵向标尺栏和横向标尺栏 ,横向标尺栏显示数据变量 ,纵向标尺栏显示数据顺序(如时间顺序)。

数据输入区可以直接输入观测数据值或存放数据 ,表的左端列边框显示观测个体的序号 ,最上端行边框显示变量名。

数据编辑区用来定义和修改变量的名称、类型及其他属性 ,如图 11-12 所示。



示。



	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	va0001	Numeric	8	2		None	None	8	Right	Scale
2	va0002	Numeric	8	2		None	None	8	Right	Scale
3										
4										
5										
6										
7										
8										
9										
10										
11										

图 员景猿 灾祸调查数据表

在灾祸调查数据表中,每一行描述一个变量的属性特征,依次是:

变量名。变量名必须以字母、汉字及下划线开头,总长度不超过 8 个字符,共容纳 8 个汉字或 8 个英文字母,英文字母不区别大小写,最后一个字符不能是句号。

变量类型。变量类型有 8 种,最常用的是 数值型变量。其他常用的类型有:字符型、日期型、货币型(每隔位数加一个逗号)等。

变量所占的宽度。

变量小数点后的位数。

变量标签。关于变量含义的详细说明。

变量值标签。关于变量各个取值的含义说明。

变量缺失值的处理方式。

指定变量在数据表中显示的列宽(默认列宽为 8)。

数据对齐格式(默认为右对齐)。

数据的测度方式。系统给出名义尺度、定序尺度和等间距尺度三种(默认为等间距尺度)。

如果输入变量名后回车,将给出变量的默认属性。如果不定义变量的属性,直接输入数据,系统将默认变量 灾祸调查数据等。

定义了变量的各种属性后,回到数据表中,就可以直接在表中录入数据。输入数据后可以点击 灾祸调查数据 或 灾祸调查数据 作为数据文件保存。另外对于统计分析的结果也可以作为文件保存起来。



为了在统计分析过程中能有效地利用其他软件产生的数据,SPSS软件编辑窗口除可以使用*.sav扩展名数据文件,还可以直接打开和保存下述类型的文件。

- SPSS数据库版本产生的数据文件*.sav
- 数据库报表程序产生的数据文件*.sav
- 用户自定义数据库格式文件*.sav
- SPSS统计软件产生的数据文件。

第二节 数据的编辑与整理

当录入数据之后,就可以对原始数据进行整理和分析,关于数据的整理和分析都是在数据窗口完成的。下面将介绍SPSS统计分析软件在数据窗口的主要操作方式和菜单相应的功能。

一、数据窗口菜单栏功能操作

数据编辑窗口的主菜单如图 2-1 所示,主菜单中的具体功能包括:

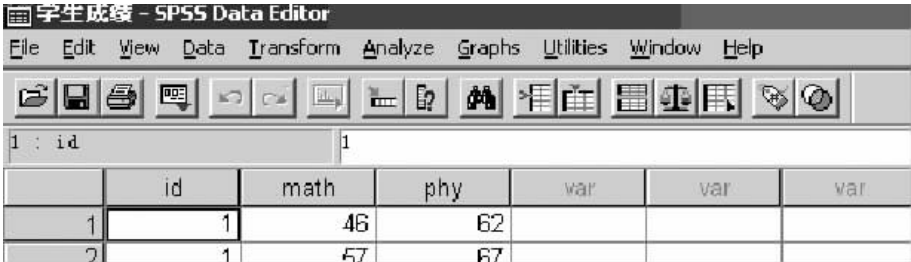


图 2-1 SPSS主菜单

- File: 文件操作。
- Edit: 文件编辑。
- View: 视图编辑。
- Data: 数据操作。
- Transform: 数据转换。
- Analyze: 统计分析方法。
- Graphs: 图形编辑。
- Utilities: 实用程序。
- Window: 窗口控制。



系统帮助。

在统计分析过程中常用的功能主要集中在数据操作、数据转换、统计分析方法、统计图形的建立与编辑等操作。

二、数据整理功能

数据编辑窗口的 **数据** 菜单为用户创建和定义数据提供了方便的功能,如图 1-1 所示。这个菜单是 SPSS 统计软件数据整理的特有功能菜单。它的功能包括:对变量、观测量的编辑处理;对变量的时间定义;对观察量数据的整理。SPSS 提供极其灵活的数据整理功能,用户可以根据不同统计分析的要求对数据进行整理。

注意:这里的观测量,是指数据文件中每个单元格内的数据,也可以称为观测个体。

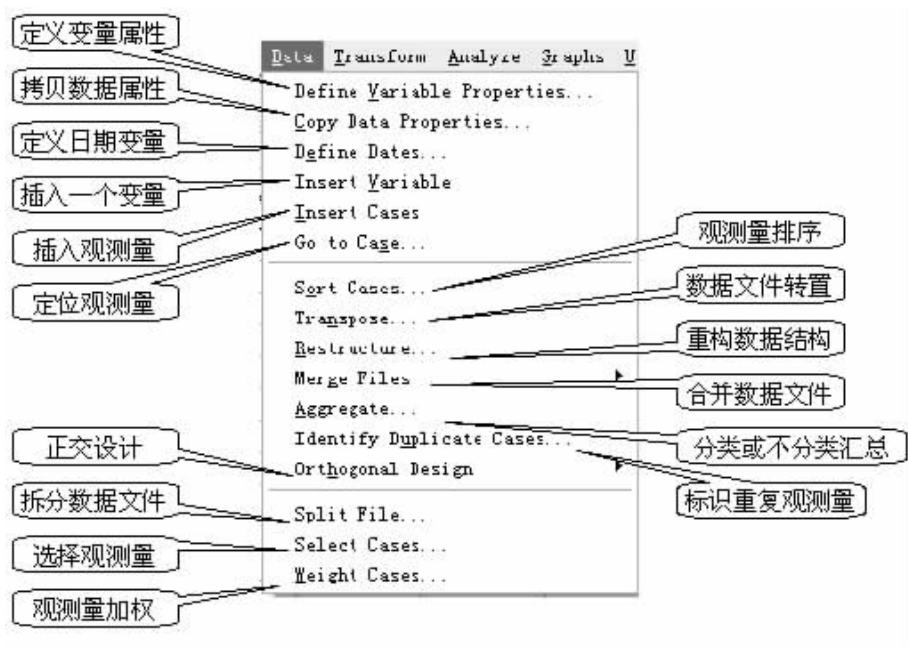


图 1-1 数据菜单项示意图

(一) 定义和编辑变量、观测量的命令

数据 > 定义变量属性 > 定义变量属性 用于定义变量属性;

数据 > 定义变量属性 > 拷贝数据属性 由外部文件和工作文件拷贝数据变量和属性;

数据 > 定义日期变量 > 定义日期变量 定义或编辑日期变量格式;

数据 > 插入 > 插入变量 在数据编辑窗口插入一个变量;



- 附加测量 在数据编辑窗口插入一个观测量；
 - 跳转测量 光标跳转到某一指定观测量。
 - (二)变量数据变换的命令
 - 排列测量 对观测量进行排序；
 - 转置测量 对观测量进行转置；
 - 重新构造测量 对现有的观测量进行重新构造,形成新格式的数据文件；
 - 合并测量 把外部文件数据合并到工作文件中；
 - 分类测量 对数据进行分类或不分类汇总,产生新文件或代替工作文件。
 - 标识重复观测量 标识重复观测量；
 - 正交设计 进行正交设计。
 - (三)观察量数据整理的命令
 - 拆分数据文件的观测量,观测量进行条件分组；
 - 选择观测量；
 - 对观测量进行加权处理。
- 通过选择上述命令,可以实现对数据的整理编辑。

三、数据或变量的变换及转换功能

数据编辑窗口的 数据菜单为用户创建和定义复杂的数据提供了方便的功能,如图 员源所示。它与 菜单共同使用,可对原始数据进行重新编辑,形成新的变量和观测量。数据菜单主要对变量进行操作,分为三部分的功能。

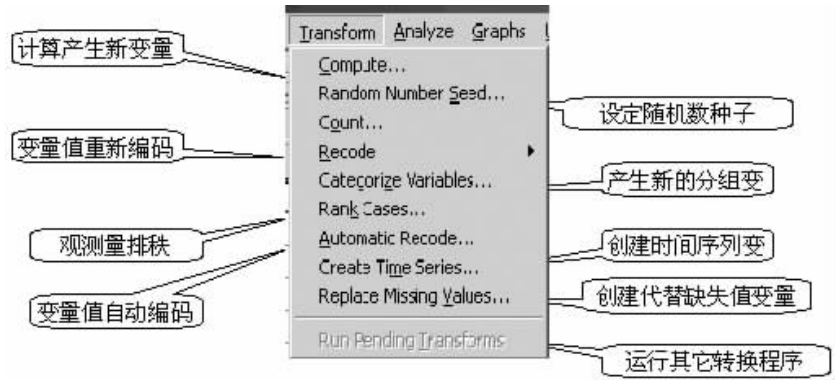


图 员源 数据菜单项示意图

- (一)通过基本变量创建各种新变量
- 计算产生新变量 提供了 多个函数,通过该菜单可以用表达式产生一个新变量(函数的部分解释见附录二)；



创建随机数种子；

创建时间序列变量。

(二) 创建各种参数变量

创建一个计数变量,用于统计计数；

对变量值重新编码；

为观测量排秩,求得的秩在数据窗口作为一个新变量保存；

建立新的分组变量,使数据分成若干个组；

对变量值自动编码,产生一个连续的变量值编码；

创建替代缺失值变量。

(三) 运行其他自定义的运输程序

运行其他转换程序。

在一般的情况下,通过菜单和菜单的操作就可以实现对原始数据的整理和变换。

四、数据的编辑

在数据窗口中,用鼠标左键单击数据表左边框的观测个体序号,这一行值就会被选中,用鼠标左键单击上边框的变量名,这个列就被选中,和其他窗口中的操作类似,也可以用鼠标对选中的一部分单元格,选中的行、列、单元格后,单击鼠标右键,就可以对它们进行复制、删除、剪切等操作。

如果需要对已经输入的数据进行修改,就要对已经存在的数据进行编辑,有许多数据编辑功能,下面介绍几种常用的数据编辑功能。

(一) 插入一个新观测量

插入一个新观测量的命令是

在数据窗口主菜单上单击菜单命令,可以在光标所在位置的前上一行插入一行新的观测个体,可以输入新的观测数据。

(二) 查找指定的观测量

查找指定的观测数据的命令是

在数据窗口单击菜单弹出一个对话框,如图 1-1 所示:输入要找的观测量的序号后,点击确定按钮,数据表中光标就会指到选定的观测量个体。

(三) 观测量数据排序

给观测量数据排序的命令是

在数据窗口单击菜单打开对话框(见图 1-2)。

从对话框左侧的变量列表中选择排序变量,点击右箭头按钮加入右侧框中,然后在右侧框中选择排序顺序:

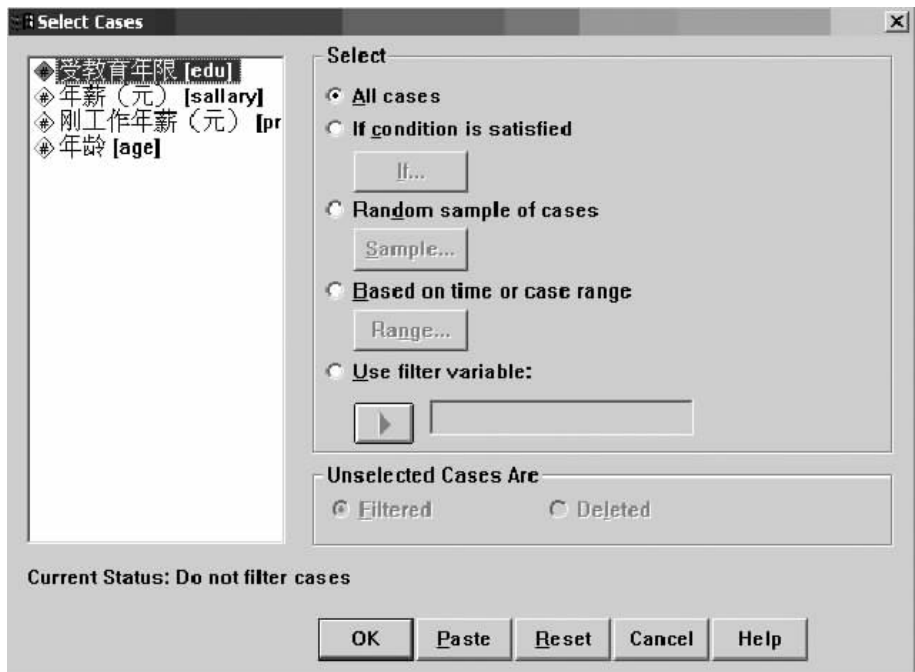


图 1-1-1 选择案例对话框

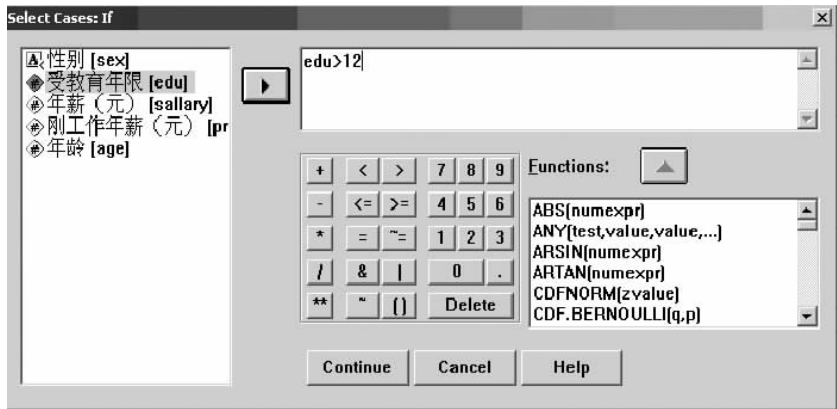


图 1-1-2 选择案例: 如果对话框

在“选择案例”对话框中有两种选择方式,一种是大概抽样(即输入抽样比例后由系统随机抽样);另一种是精确抽样,要求输入从第几个观察值起抽取多少数据。

- 单击“精确抽样”按钮,打开“精确抽样”对话框。

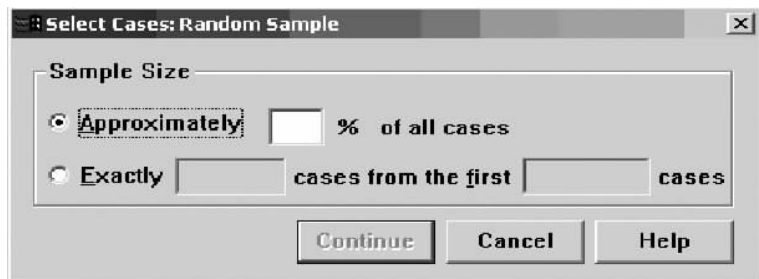


图 员原员 选择随机样本对话框

选择对话框,用户自行定义从第几个观察值开始抽到第几个观察值结束。

- 删除不需要的数据。用指定变量作过滤。先选择一个变量,系统自动在数据管理器中将该变量值为 园 的观测个体标上删除记号,系统对标有删除记号的观测个体不作分析。

选择了挑选数据子集的方式以后,单击 韵运,在数据窗口可看到新的变量表。如在数据文件 杂再原员中,选择受教育年限大于 员圆年的人作为选择子集。则满足条件的人相应的新变量数据为 员否则为 园。

(五)数据分类汇总(数据分组汇总)

用户可以根据需要对数据按指定的变量的数值进行归类分组汇总。以数据库 杂再原员两个班的学生成绩为例,如果按照性别对数学成绩的平均值进行汇总,可以使用分类汇总命令实现。具体操作如下:

员指定分类变量和汇总变量。打开数据库 杂再原员在数据窗口单击 阔操>粤再原员命令,打开 粤再原员对话框。如图 员原员所示:

园在变量名列表框中选择分类变量“性别”进入 月再原员对话框。

猿在变量名列表框中选择汇总变量“数学”进入 粤再原员对话框。

源单击 云按钮,打开 粤再原员对话框,如图 员原员所示。在此对话框中可以选择平均值、数据和、标准差的形式,特别值形式、百分数形式、频数形式等其中之一的方法进行分类汇总。选择分类汇总的函数形式后返回 粤再原员对话框。

缘在 粤再原员对话框中指定汇总文件的保存路径。有两种选择:一种是选中创建新数据文件,通过 云按钮,重新指定结果文件名。另一种是替代原来数据文件,用分类汇总结果覆盖当前编辑窗口的数据。系统默认前一种选择。

远单击 粤再原员按钮,可以重新指定结果文件中的变量名并加入变量标签。否则,杂再原员默认的结果文件中的变量名为原变量名最后加上 员。

苑如果希望在结果文件中保存各分类组的数据个数,可以选择 杂再原员。最后单击 韵运,可得相应的数据文件。

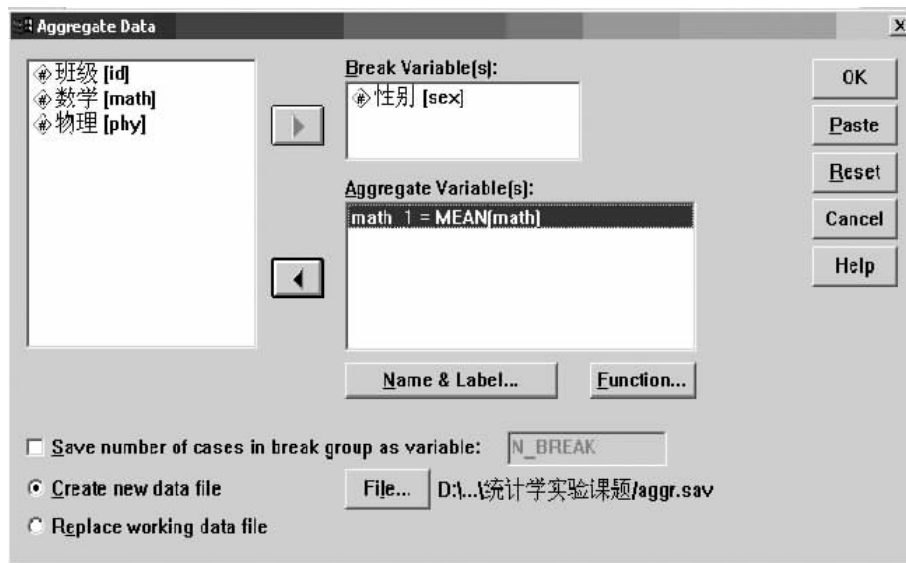


图 1-1-10 聚合数据对话框

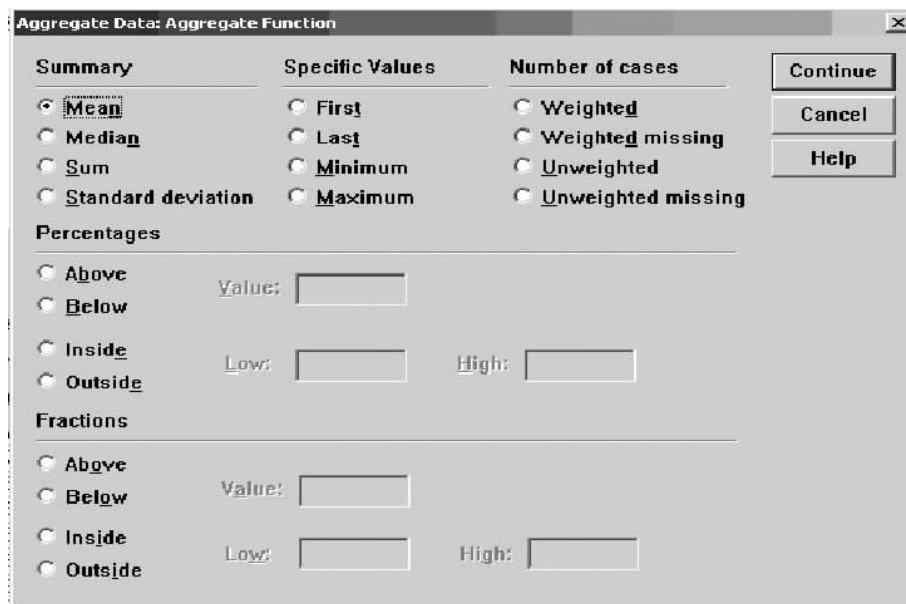


图 1-1-11 聚合数据: 聚合函数对话框

(六) 缺失值的替代方式

如果用户希望对缺失值进行定义, 可以采用以下的操作:

在数据窗口点击 数据 → 缺失值 → 定义缺失值, 打开 定义缺失值对话框



怎样对话框 如图 员源所示：

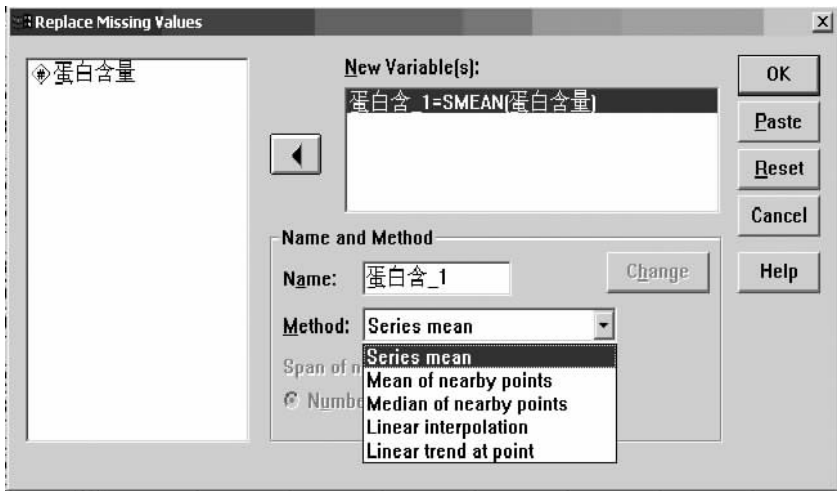


图 员源 缺失值替换对话框

在变量中选择具有缺失值的变量进入 替换对话框内 ,系统可以自动产生替代缺失值的新变量 ,也可以自定义新变量。即在 方法的下拉菜单中选择缺失值的替代方式 ,五种方式依次是：

- 用该变量所有非缺失值的平均值替代缺失值；
- 用缺失值相邻点的非缺失值的平均数据替代缺失值；
- 用缺失值相邻点的非缺失值的中位数替代缺失值；
- 用缺失值相邻点的非缺失值的中点值替代缺失值；
- 用线性拟合方式替代缺失值。

(七)数据秩(序)的确定

如果用户需要对已有的数据变量排秩(序) ,如对数据 员源中两个班的数学成绩分别排出名次 ,可以在数据窗口采用以下操作。

单击 数据>排序>打开 排序对话框 ,如图 员源所示：

从左边变量名列表框中选择变量“数学”(也可选择多个变量)进入 变量框中 ,选择变量“班级”进入 按框中 ,则系统排序时将按照进入 按的变量值“班级”进行分别排序。

单击 按钮 ,选择 称为“结” ,是指两个或两个以上的数据相等的情况)的处理方式。由于秩与数据个数是一一对应的 ,当数据有相同的时 ,确定它们相应的秩有三种处理方式 :对应秩的 平均值、 最小值和 最大值。如本例选择最大值。选择后返回到主对话框。单击 就可以在数据窗口看到排序结果。

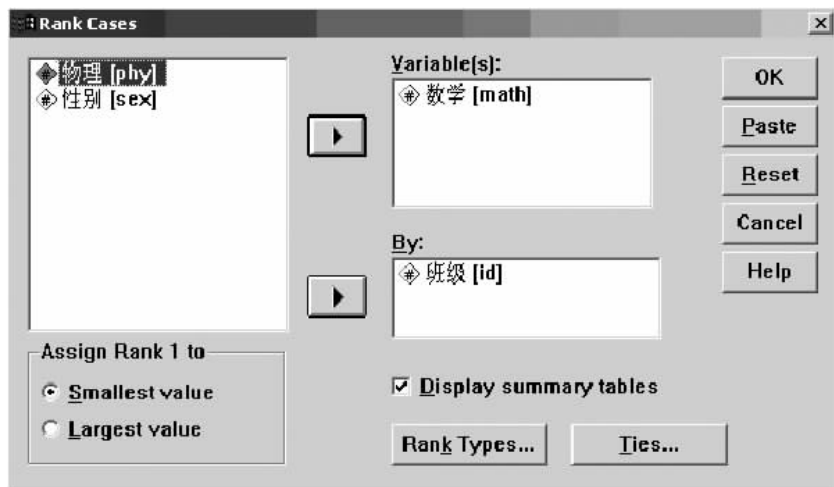


图 员缘 赋值对话框

赋值对话框提供排序方式。单击赋值按钮,打开赋值对话框,从中选择排序类型,排序类型从左到右依次是:赋值普通排序(系统默认),得到的新变量的值就是秩;云赋值百分数排序,累计百分数排序;云赋值以指数分布为基础的原始分排序;云赋值以分组例数之和的权重排序;云赋值以秩变量除以分组例数之和排序;云赋值先给定一个大于员的整数,系统按照此数的范围确定秩。

五、变量的编辑

(一)插入一个新变量

插入一个新变量的命令是赋值插入新变量

在数据窗口单击赋值插入新变量会在光标所在位置的前一列插入一个新的变量,变量名字和属性可以在赋值变量对话框中定义。

(二)已存在的变量生成新变量

对于已存在的数据变量,根据需要进行计算生成新变量的命令是赋值计算

在数据窗口单击赋值计算打开赋值计算对话框,如图员缘所示。

在对话框左上方赋值名称栏中,键入即将生成的新变量的名称,并单击赋值数据按钮确定变量标签及数据类型。对话框的左下栏中给出了数据文件中所有可用的变量列表,我们可以用右箭头按钮从中选取所需的变量进入右上方的赋值表达式栏中,该栏存放运算表达式,运算表达式中所需要的常用函数可以从下面的云赋值列表表中直接选取。这些常用函数(见附录圆和



常用统计分析方法——数据应用

其他语言中的函数名称类似,在框中按字母顺序排列,用鼠标选中某个函数,用鼠标单击右面的上箭头按钮加入数值表达式中,对话框中间是一个小键盘,可以用来输入数字、运算符号等。单击“OK”按钮,对话框的下面还有一个“If...”按钮,可以选一部分满足某种条件的观测个体来做运算,不满足条件观测,其新变量值缺失。

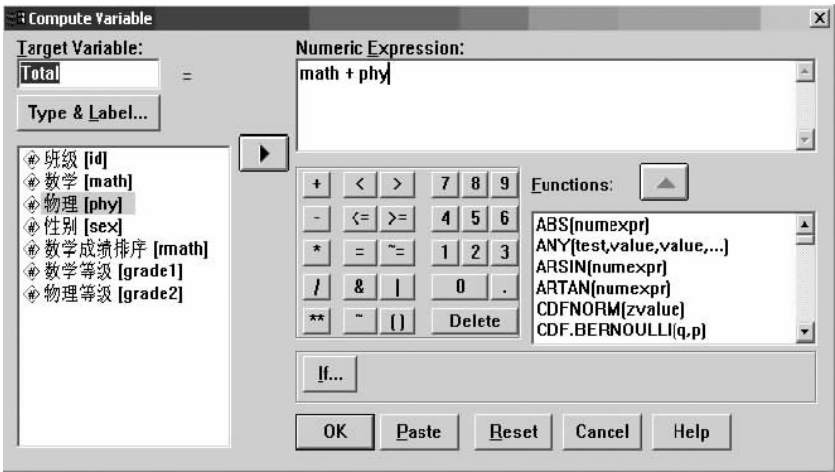


图 员原远 计算新变量对话框

如图 员原远表示的是数据 员原中每个学生的数学和物理总成绩。在“计算新变量”对话框中填好新变量名称和运算表达式后,单击“OK”按钮,就可以在数据文件中看到,已经生成了一个新变量“Total”。

(三)产生计数变量

如果用户需要对满足某项条件的数据进行计数,可以使用“计数”命令。以学生成绩数据 员原为例,说明具体操作步骤:

在数据窗口单击“数据”→“计数”打开“计数”对话框。如图 员原所示:如果要清点每个学生达到良好以上的课程数,可以先在“计数”对话框中指定一个变量(可以是已经存在的变量或新变量),并定义变量标签,然后指定要统计的变量数学、物理加到“要计数的变量”框中,再单击“OK”按钮,打开“计数”对话框。如图 员原所示:

- 在上面的对话框中,确定需要计数的数值,其“计数”值的设置项依次是:
- “计数”输入某个值为清点对象;
 - “计数”以系统的缺失值为清点对象;
 - “计数”以系统或用户指定的缺失值为清点对象;
 - “计数”指定数值的计数区域,其中包括:
 - ()“计数”在框内指定下限和上限;

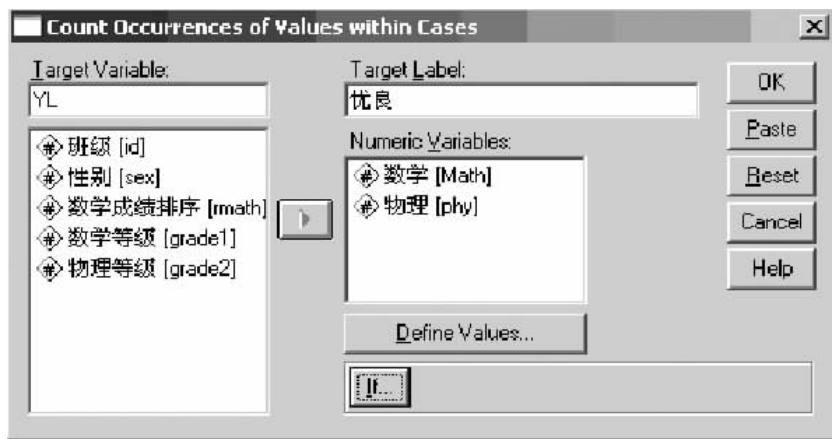


图 1-1-1 计算变量值出现频数的对话框

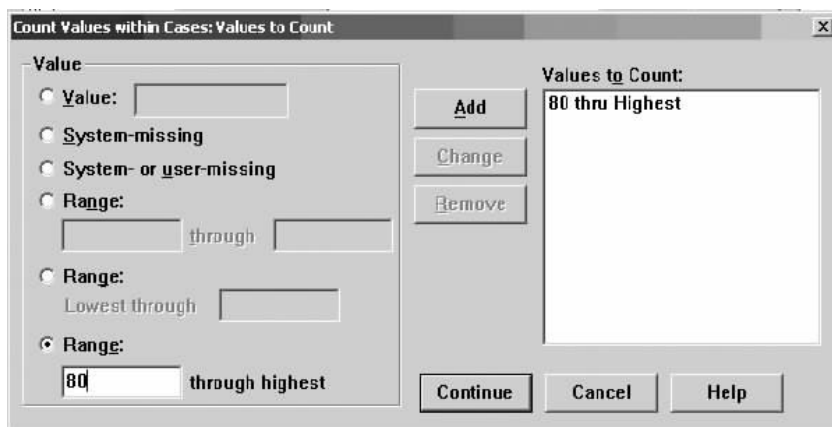


图 1-1-2 计算变量值出现频数范围的对话框

在框内只指定上限；

() 在框内只指定下限。

图 1-1-2 中给出的是计算达到优良标准,即学生达到 80 分以上课程数。确定了计数数值后,单击 **确定** 按钮,使选择结果进入 **计算变量值出现频数** 框内。单击 **确定** 按钮,返回主对话框中。如果需要,可以单击 **确定** 按钮确定计数条件。最后单击 **确定** 可在数据窗口得到计数变量。

(四) 变量分组(编码)与自动分组(编码)

数据菜单下还有以上两条分组(编码)命令。对变量数据的重新分组(编码)是指给每个变量值重新赋予一个码来描述它们的某些属性。码数相同的即为一组。比如,可以对年龄重新分组,18 岁及以下年龄赋予一个编码 1,19 岁及以上年龄赋予一个编码 2。

有的区间一个一个都输入后,点 **快捷菜单** 按钮回到 **赋值对话框**,在 **要赋值的变量** 对话框,点 **运行** 按钮执行命令,即在数据窗口可得到需要的分组赋值变量。

第三节 基本统计描述

在建立了数据文件之后,需要对数据作进一步的考察,要了解数据的基本特征,如数据的均值、标准差、四分位点,数据的分布形态等,这个过程称为对数据进行基本统计描述。数据的基本统计描述的目的在于:了解数据的基本特征和大致分布形状,为进一步分析做好充分准备。

本节主要内容:数据的基本统计描述方法:频数分析、探索分析及交叉列联表分析等。

一、描述统计分析过程(阅读与理解)

描述统计分析是对数据进行基础性描述。通过描述统计分析,可以得出数据的均值(配募)、和(杂皂)、标准差(杂皂)、最大值(配募)、最小值(配募)、方差(灾通)、极差(砸站)、均值的标准误(杂皂)、峰度(灾通)、偏度(杂皂)等数据统计量。

以 2000 年全国职工平均工资表为例(数据库 杂原圆),介绍描述统计分析的具体操作步骤如下:

员首先打开数据表，再按照“粤”按钮，阅读“粤”对话框，如图 6-10 所示：

从左边源变量中选择一个或者几个变量进入右框中,单击 **向右** 按钮,打开 **选择变量** 对话框,如图 11.1.1 所示:

在对话框中最上面一行是 **权重均值**，它是算术和。下面的选择栏分别是：

- 阅读数据表 离差栏
 杂散数据表 标准差
 灾损数据表 方差
 砾石数据表 极差
 杂散数据表 均值的标准误差
 杂散数据表 最小值
 杂散数据表 最大值
 杂散数据表 均值的标准误差
- 阅读数据表 分布形状栏
 杂散数据表 偏度
 杂散数据表 峰度
- 阅读数据表 列头栏 选择输出方式：
 灾损数据表 按变量表次序；
 杂散数据表 按字母顺序；

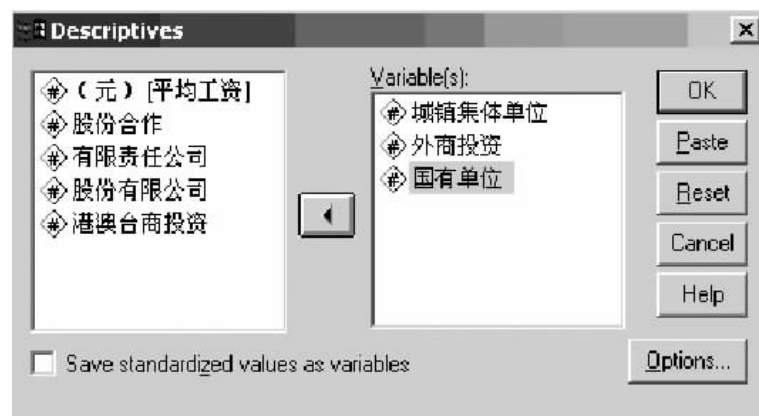


图 10-1-1 描述统计主对话框

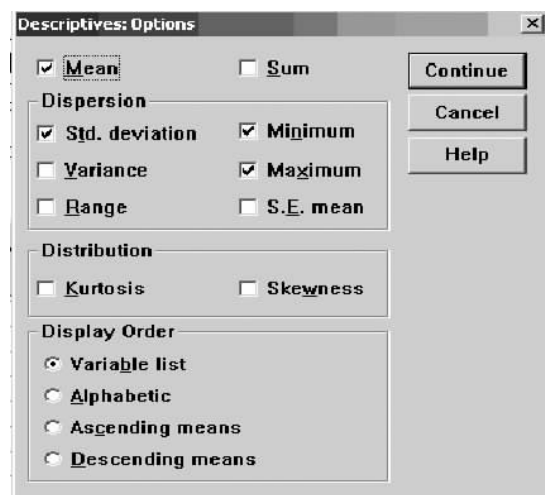


图 10-1-2 描述统计选项对话框

单击“按平均值升序”按钮，按平均值升序；

单击“按平均值降序”按钮，按平均值降序。

如在此例中选择平均值、标准差、最大最小值、按平均值升序项，返回主对话框，单击“确定”，在输出窗口中描述统计分析输出表，如表 10-1-1。



表 1-1 某公司员工基本描述统计表

类别	人数	男	女	平均年龄	平均受教育年限
城镇集体	10	6	4	25.5	12.5
股份合作	15	9	6	26.2	13.2
有限责任	20	12	8	27.1	14.1
股份有限	25	15	10	28.3	15.3
国有单位	30	18	12	29.4	16.4
外商投资	35	20	15	30.5	17.5
其他	40	22	18	31.6	18.6

二、频数分析

对于一组数据,考察不同的数据出现的频数,或者是数据所落入指定的区域内的频数,可以了解数据的分布状况。

例 1-1 数据文件 1-1 是一个公司员工表,其中有性别、年龄、受教育年限等五个变量,要考察公司员工的年龄频数分布,相应的具体操作如下:

1. 打开数据文件 1-1 后,单击 菜单 统计 频数分析 打开频数分析对话框,如图 1-1 所示。



图 1-1 频数分析对话框

- 在左边的变量框中选中一个或多个变量送入 变量框。
- 选中 统计 频数分析 要求输出分布表。
- 单击 统计 频数分析 按钮,得到对话框图 1-1。
- 在 统计 频数分析 对话框中选择要求输出的统计量。
- 选择 统计 频数分析 百分数选择项栏(复选项)

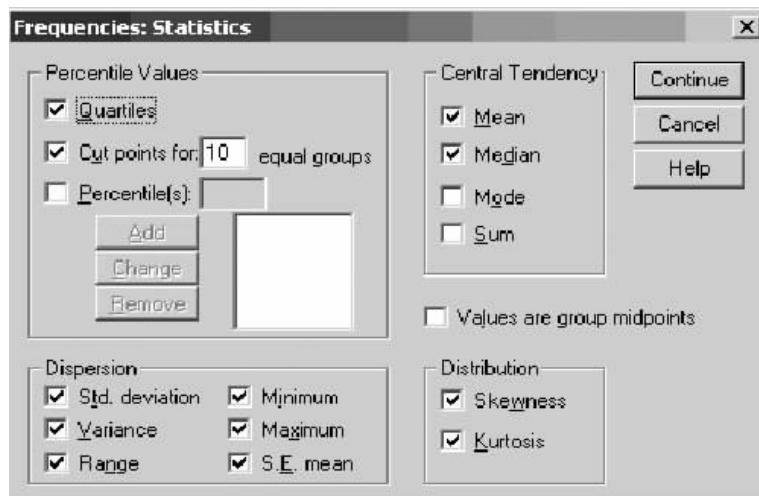


图 10-1-1 频数统计: 统计量对话框

图 10-1-1 频数统计: 统计量对话框

图 10-1-1 频数统计: 统计量对话框

图 10-1-1 频数统计: 统计量对话框

图 10-1-1 频数统计: 统计量对话框

图 10-1-1 频数统计: 统计量对话框

图 10-1-1 频数统计: 统计量对话框

图 10-1-1 频数统计: 统计量对话框

在本例中选择四分位点、10等分的百分位点、标准差、方差、最大、最小值，全距、均值、均值的标准误、中位数、偏度、峰度等复选项。

单击“继续”按钮，得到“频数统计: 统计量”对话框，如图 10-1-2 在对话框中有

图 10-1-2 频数统计: 统计量对话框

图 10-1-2 频数统计: 统计量对话框

图 10-1-2 频数统计: 统计量对话框

单击“继续”按钮，得到对话框图 10-1-3

图 10-1-3 频数统计: 统计量对话框

图 10-1-3 频数统计: 统计量对话框

百分比。

单击“继续”按钮，得到对话框图 10-1-4

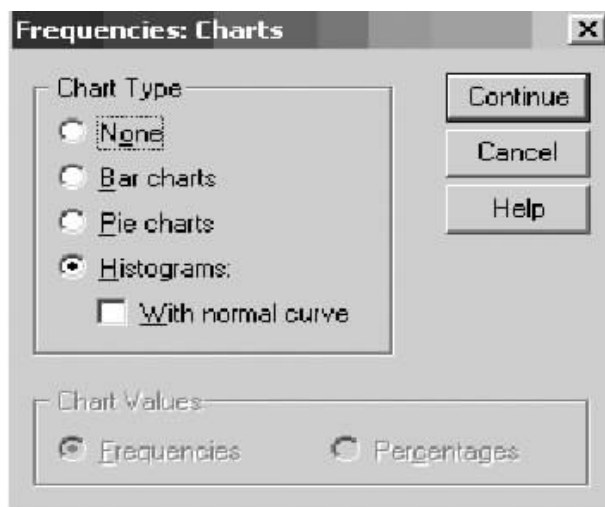


图 1-1-1 频数分布表对话框

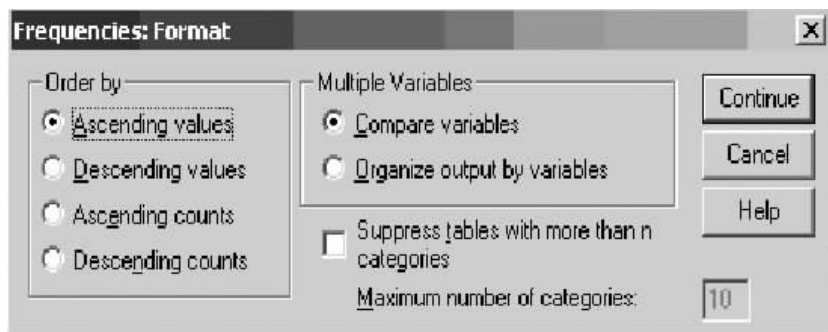
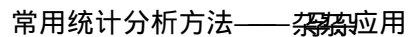


图 1-1-2 频数分布表格式对话框

在“频数分布表”对话框中：

- “按变量值升序排列”表示频数分布表的排列顺序(单选)。系统默认按变量值升序排列(系统默认)。
- “按变量值降序排列”按变量值降序排列。
- “按变量各种取值发生的频数升序排列”按变量各种取值发生的频数升序排列。
- “按变量各种取值发生的频数降序排列”按变量各种取值发生的频数降序排列。
- 如果设置了直方图, 频数表将按照变量值顺序排列。
- “多变量输出表格设置(单选)”。将所有变量的结果输出在一个表中。
- “为每一个变量输出一个表”。
- “控制频数表输出的分类数复选项”。



本例中均选择系统默认项。点击 **韵运**, 得到一组输出结果如表 10-9-2 和

表 员原圆葬

多元回归统计分析表

晕	灾有有效值 配并缺失值	苑园
配并均值		源园
灾有均值的标准误		员园
灾有标准差		员园
灾有偏度		源
灾有偏度的标准误		源
灾有峰度		源
灾有峰度的标准误		源
灾有最小值		源
灾有最大值		源
灾有	源 源 源 源 源 源 源 源 源 源 源	源 源 源 源 源 源 源 源 源 源 源

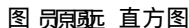
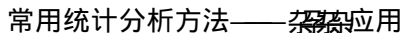


表 员原圆遭
年龄

频数表

	云源灾源	云源灾源	云源灾源	云源灾源
	频数	豫	灾源灾源灾源	灾源灾源灾源
灾源	圆	灾源	灾源	灾源
灾源	圆	灾源	灾源	灾源
灾源	圆	灾源	灾源	灾源
灾源	源	灾源	灾源	灾源
灾源	圆	灾源	灾源	灾源
灾源	源	灾源	灾源	灾源
灾源	苑	灾源	灾源	灾源
灾源	源	灾源	灾源	灾源
灾源	远	灾源	灾源	灾源
灾源	缘	灾源	灾源	灾源
灾源	猿	灾源	灾源	灾源
灾源	员	灾源	灾源	灾源
灾源	圆	灾源	灾源	灾源
灾源	员	灾源	灾源	灾源
灾源	猿	灾源	灾源	灾源
灾源	员	灾源	灾源	灾源
灾源	圆	灾源	灾源	灾源
灾源	圆	灾源	灾源	灾源
灾源	员	灾源	灾源	灾源
灾源	员	灾源	灾源	灾源
灾源	圆	灾源	灾源	灾源
灾源	圆	灾源	灾源	灾源
灾源	员	灾源	灾源	灾源
灾源	员	灾源	灾源	灾源
灾源	猿	灾源	灾源	灾源
灾源	圆	灾源	灾源	灾源
灾源	圆	灾源	灾源	灾源
灾源	员	灾源	灾源	灾源
灾源	灾源	灾源	灾源	灾源

从上面的表和直方图中可以观察到该公司 灾源岁至 灾源岁之间的人数最多，占到总人数的 灾源%以上。



三、探索分析过程(耕普农)

(一)观察数据的分布特征 通过绘制箱图和茎叶图等图形直观地反映数据的分布形式和数据的一些规律性,包括考察数据中是否存在异常值等。

(二)正态分布检验 检验观测数据是否服从正态分布。

(三)方差齐性的检验:用 Levene 检验比较各组的方差是否相等。

例 14.4 数据库 杂再原员 提供两个班的学习成绩数据。对两个班的数学成绩按照性别考察数据的分布、按照性别检验其数学成绩的方差是否相等。打开数据库 杂再原员 具体操作步骤：

单击 **粤教云课** 阅读资源 **粤教云课** 打开 **粤教云课** 主对话框,如图 员原所示:

(员) 从左侧的变量列表中选出变量“数学”，送入 阅读成绩 栏；

(圆) 选择“性别”作为因子变量, 送入 **非条件模型** 栏。有了因子变量, 就不会把所有的观测个体按照因子变量的取值分成相应组, 再分组考察模型中的各个变量, 如果不选择因子变量, 则会对全部观测来做探索分析。

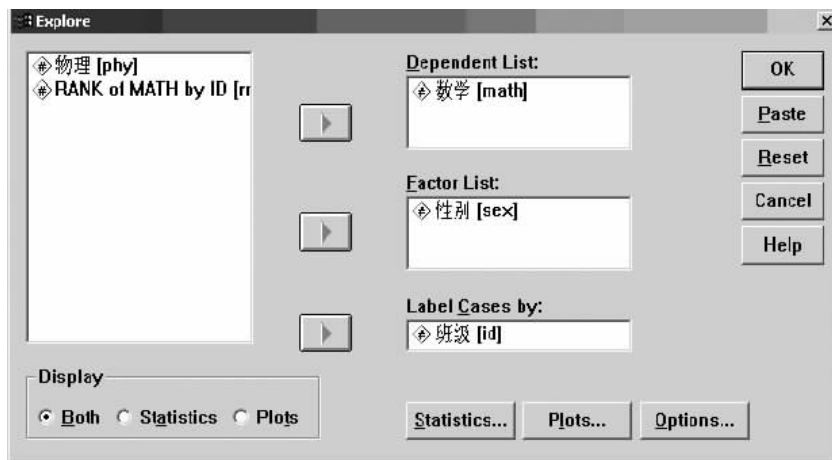


图 1-1-1 探索分析主对话框

选择“班级”标识变量送入 标识变量 栏,当输出涉及到观测个体时,使用该变量值标识各观测个体。

在 显示 栏中选择输出项,依次是 月 选择项、输出图形与描述统计量(系统默认), 只输出描述统计量和 只输出图形。本例中选择默认项。

单击 统计量 按钮,打开 统计量 对话框,选择统计输出量。有四个选择项,分别是:

选择 基本统计描述。同时指定均值的置信区间的置信度,系统默认为 95%。

选择 估计(估计在计算时对所有观测量赋予权重,随观测量距分布中心的远近而变化);

选择 输出分析数据中五个最大值和五个最小值;

选择 输出百分数。

本例中选择 基本统计描述 和 输出百分数 后,返回主对话框。如图 1-1-2 所示。

单击 图形 按钮,打开 图形 对话框,如图 1-1-3 所示。

对话框中有四个选择栏:

- 月 箱图选择栏(单选项)。

选择 按因素水平分组(系统默认);

选择 所有因变量生成一个并列箱图(本例中选择项);

选择 不显示箱图。

箱图中,最底部的水平线段是数据的最小值(奇异点除外),顶部的水平线段是数据的最大值(奇异点除外),中间矩形箱子的底所在位置是数据的第一个四

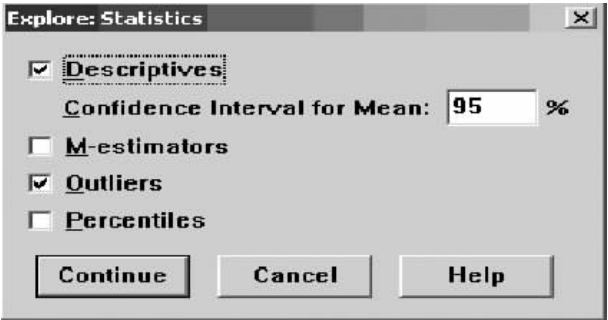


图 员原愿 探索分析 统计对话框

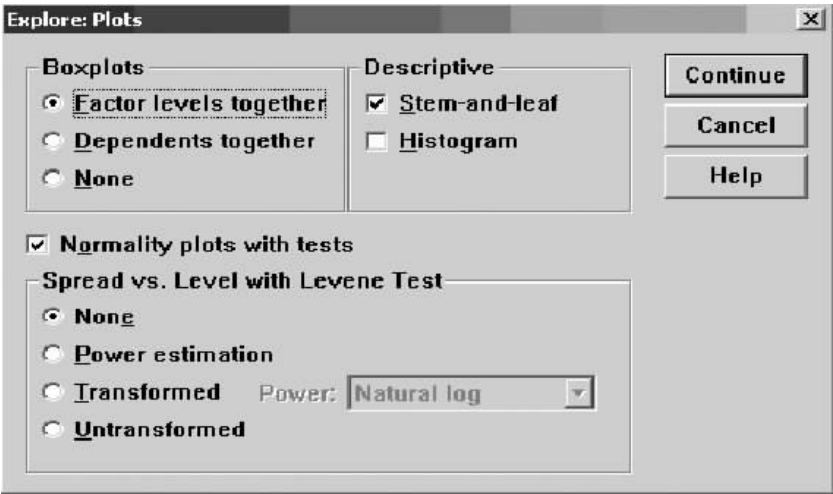


图 员原愿 探索分析 图形对话框

分位数(即 孕) 箱子顶部所在位置是数据的第三个四分位数数据(即 孕)。箱子中间的水平线段刻画的是数据的中位数(即 孕)。奇异值是指大于 孕垣员豫孕(原孕)或小于 孕原员豫孕(原孕)的数值 标记为 韵而对于大于 孕垣员豫孕(原孕)或小于 孕原员豫孕(原孕)的数值称为极值 标记为 *。

- 阅察描述图形栏(复选项)。
系统默认茎叶图
勾选直方图
勾选正态分布检验并输出 正态分布图。
- 勾选 对所有的散布—层次图 ,同时输出回归直线的斜率以及方差齐性的 检验 ,但如果没有指定分组变量 ,此选择项无效。四个单选项依次为 :



系统默认)；
~~勾选~~ ~~勾选~~ 转换幂值估计(对每组数据产生一个中位数自然对数及四个分位数的自然对数的散点图)选项；
~~勾选~~ ~~勾选~~ 变换原始数据选择项(可在参数框中选择数据变换类型)；
~~勾选~~ ~~勾选~~ 不变换原始数据选择项。
本例中选择茎叶图 正态分布检验 方差齐性检验等。
单击 ~~源~~ 按钮,打开 ~~源~~ 对话框,如图 ~~源~~ 所示。可选择缺失值的处理方式, ~~源~~ 提供三种处理方式：

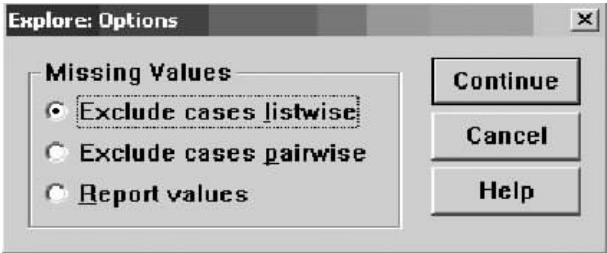


图 ~~源~~ 探索分析 ~~源~~ 对话框

~~勾选~~ ~~勾选~~ 剔除带缺失值的观测量(系统默认)。本例选择此项。
~~勾选~~ ~~勾选~~ 剔除带缺失值的观测量时还一并剔除与缺失值有成对关系的观测量。

单击 ~~源~~ 得到相应的输出结果如表 ~~源~~ 和图 ~~源~~ 所示。

表 ~~源~~ ~~源~~ 数据概述

	性别	悦源数据					
		有效值		缺失值		总数	
		个数	百分比	晕	源	晕	源
数学	女	猿	源	园	猿	猿	源
	男	源	源	园	猿	源	源



表 1 成绩数据文件中的极端值表
(按照性别输出数学成绩的五個最大值及五个最小值)

	性别	成绩	排名	成绩	排名	成绩	排名
数学	女	95	1	85	2	75	3
		90	4	80	5	70	6
		85	7	75	8	65	9
		80	10	70	11	60	12
		75	13	65	14	55	15
	男	90	16	80	17	70	18
		85	19	75	20	65	21
		80	22	70	23	60	24
		75	25	65	26	55	27
		70	28	60	29	50	30

表 2 成绩数据文件中的正态分布检验表

	性别	统计量			临界值		
		统计量	统计量	统计量	统计量	统计量	统计量
数学	女	0.000	0.000	0.000	0.000	0.000	0.000
	男	0.000	0.000	0.000	0.000	0.000	0.000

* 表 1 中的极端值表显示了数学成绩的五個最大值及五个最小值。从表 2 的正态分布检验结果可以看出,由于假设检验的 P 值均大于 0.05,故可以认为男女生的数学成绩分布都近似地服从正态分布。



常用统计分析方法——方差分析

表 1 方差齐性检验

	统计量	自由度	临界值	结论
数学	基于均值	1	0.05	接受
	基于中位数	1	0.05	接受
	基于中位数及修正的自由度	1	0.05	接受
	基于修正的自由度	1	0.05	接受

由表 1 得出方差齐性检验的 p 值为 0.05 以上, 故认为男女生数学成绩的方差是相等的。

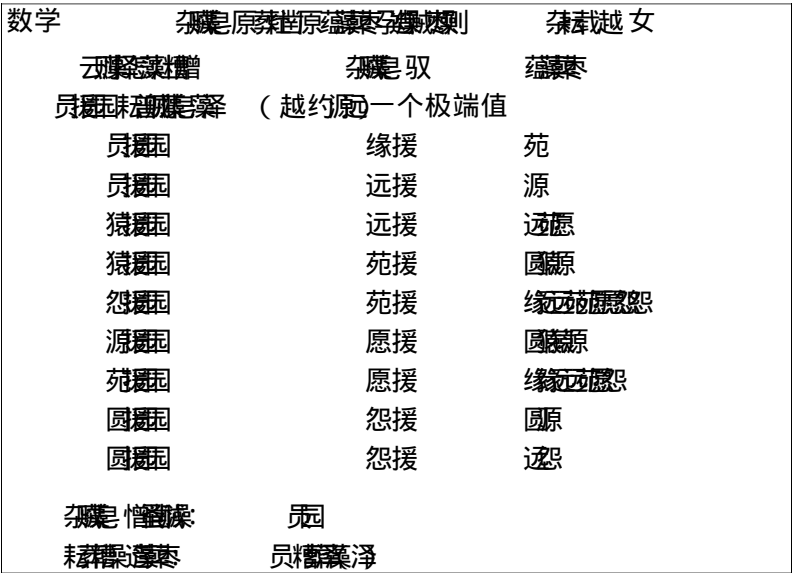


图 1 展示了男女生数学成绩的分布情况。从图中可以看出, 女生的数学成绩分布更接近正态分布, 而男生的分布则存在一定的偏态。这可能与男生的数学成绩存在个别极端值有关。

从图 1 的散点图中可以看出, 除了个别极端点外, 女生的数学成绩分布在 70 分左右波动, 故可以认为女生的数学成绩近似服从正态分布。

从图 1 的箱图中可以得出结论: 女生的数学成绩的平均水平比男生低且分散程度小, 但有一个奇异值。

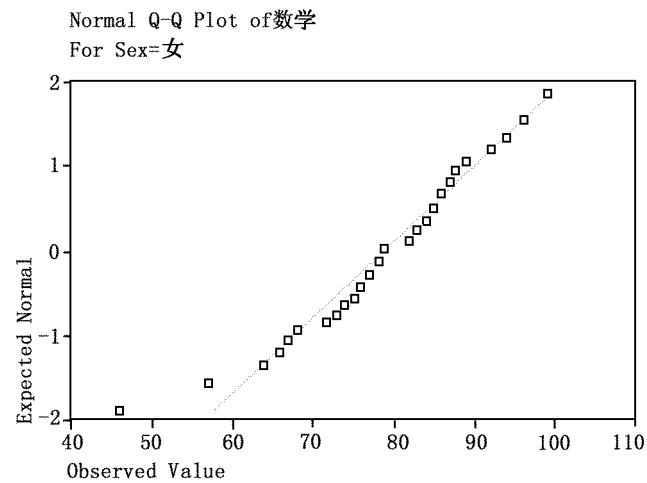


图 员京氛遭 正态分布 图原市图

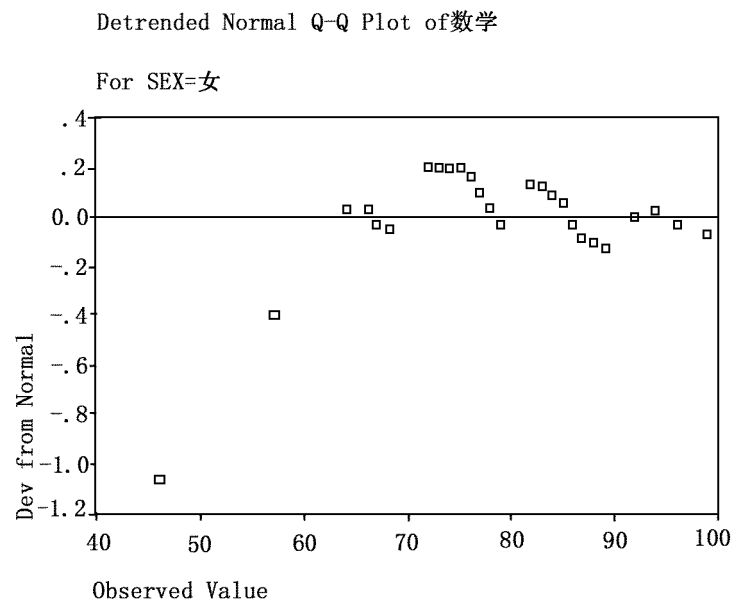


图 员京氛糟 离散正态分布图

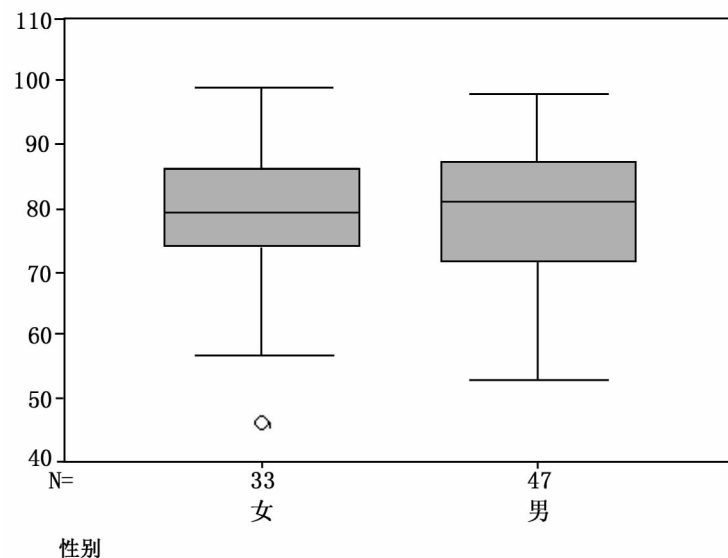


图 4-1-1 按照性别绘制的数学成绩的箱图

第四节 交叉列联表分析(悦悦悦悦)

当观察的现象与两个因素有关时,如某种服装的销量受价格和居民收入影响,某种产品的生产成本受原材料价格和产量的影响等等,交叉列联表分析可以比较好地反映出两个因素之间有无关联性,两因素与现象之间的相关关系。因此,数据交叉列联表分析主要包括两个基本任务:

1. 根据收集的样本数据,产生二维或多维交叉列联表;

2. 在交叉列联表的基础上,对两两变量间是否存在关联性进行检验。

下面仍然以数据 4-1-1 学生成绩为例,先将学生成绩按照五级制分等级,再按照班级形成数学等级和物理等级交叉分析表,并考察学生的物理和数学成绩间有无关联性。

一、交叉列联表的形成

制作交叉列联表的具体操作步骤:

1. 打开数据 4-1-1,单击“数据”菜单>“选择个案”>“按范围选择个案”对话框,如



图 员京趣所示。



图 员京趣 交叉列联表对话框

如果是二维列联表分析,可以将行变量选择进入 砾翟净中,将列变量选择进入 悦皂化净框中。如进行三维以上的列联表,可以将其他变量作为控制变量选到 蕴燥框中。多控制变量可以是同层次的也可以是逐层叠加的。此例中选择数学等级为行变量,物理等级为列变量,班级作为控制变量。

猿翟净框中的 悦皂化净选项,可以指定绘制各变量交叉频数分布柱形图。杂皂化净选项表示不输出列联表,只有在分析行列变量间关系时选择此项。此例中不选择这一项。

源单击 悦皂化净按钮,打开 悦皂化净悦皂化净悦皂化净对话框,如图 员京表所示。从对话框中指定列联表单元格中的输出内容。在 悦皂化净框中选择 韵翟净观察值(系统默认)或 耘翟净期望频数;在 耘翟净框内选择 砾翟净行百分比,悦皂化净列百分比及 耘翟净总百分比;在 砾翟净框中选择输出残差。其中 耘翟净框内为标准化残差。耘翟净框内为修正的标准化残差。

本列中选择默认项观察值。

缘单击 耘翟净按钮,指定列联表的输出排列顺序,一般选择系统默认的升序。然后点击 韵翟净就可得到交叉列联表如表 员京原所示。

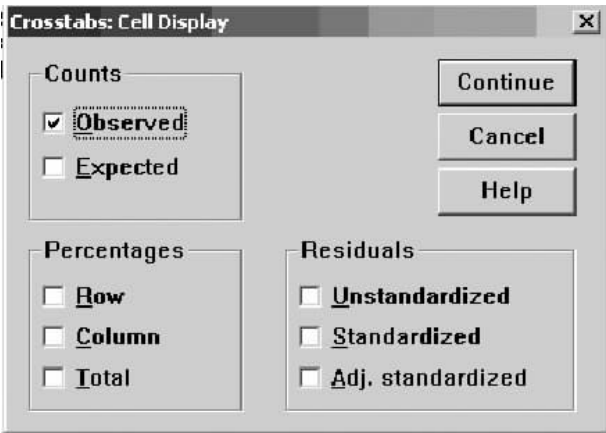


图 员原猿 悦源猿悦源猿悦源猿对话框

表 员原源 数学等级 * 物理等级 * 班级 悦源猿悦源猿悦源猿交叉列联表
悦源猿悦源猿

班级			物理等级					裁源猿
			优	良	中	及	不及格	
员	数学等级	优	源	员	园	园	园	缘
		良	猿	园	园	园	园	猿
		中	园	缘	员	园	猿	猿
		及	园	园	猿	员	远	猿
		不及格	园	园	园	园	园	圆
裁源猿			苑	苑	怨	远	员	猿
圆	数学等级	优	源	员	园	园	园	缘
		良	员	园	园	园	园	源
		中	园	缘	远	圆	园	猿
		及	园	园	猿	猿	圆	愿
		不及格	园	园	园	园	员	员
裁源猿			缘	员	员	缘	猿	源

从表 员原源中可以看出 ,一班中数学和物理成绩均为优秀者有四人 ,数学不及格的两人的物理成绩都是及格。

二、两变量关联性检验

两变量关联性检验主要用来检验两个变量因素之间是否有关联关系 ,可以用卡方检验实现。如果要考察学生的数学成绩和物理成绩之间是否有关联 ,相



当于检验假设：

① 数学成绩和物理成绩之间是相互独立的(无关联关系)；

② 数学成绩和物理成绩之间的关联关系显著。

由于列联分析表 员原中出现数据中小于 缘的值太多,故将数学等级合并,优良等级合成一组,中、及和不及等级合成一组,同样方法可得物理分组,数学分组和物理分组作为变量保存在数据文件中,分组可以通过 裁菜单中 砸鄢命令实现。实现新的分组后,再实施卡方检验的具体操作步骤:

员打开数据 员再员单击 粤按钮>阅按钮>裁按钮>悦按钮得到对话框如图 员原所示。

圆选择数学分组为行变量,物理分组为列变量,班级作为控制变量。

猿单击 悦按钮,打开 悦对话框,选择观察值为输出内容。

源单击 裁按钮,打开 裁对话框如图 员原,此对话框提供检验方式。三个单选项分别是:

(员) 粤选择项,适用于具有渐近分布的大样本数据(默认项);

(圆) 配选择项,此项为精确显著水平值的无偏估计,无需数据具有渐近分布的假设,是一种非常有效的计算确切显著性水平的方法。在 悦对话框参数框 噪中输入数据,确定置信区间的大小,一般为 怨怨,在 配对话框 泽中输入样本框中输入样本量数据。

(猿) 裁选择项,观察结果概率,同时在下方的 裁对话框内,选择进行精确检验的最大时限。

本例中不作选择。

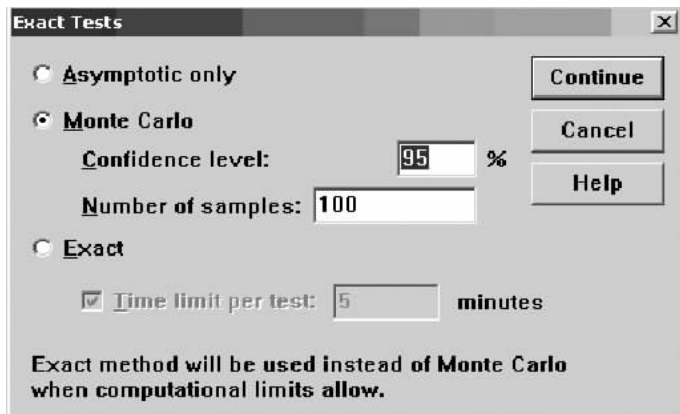


图 员原 裁对话框

缘单击 云按钮,从中选择排列顺序。一般取默认项,即升序排列。

远单击 杂按钮,打开 悦对话框如图 员原所示。从中



常用统计分析方法——杂类应用

选择检验统计量：

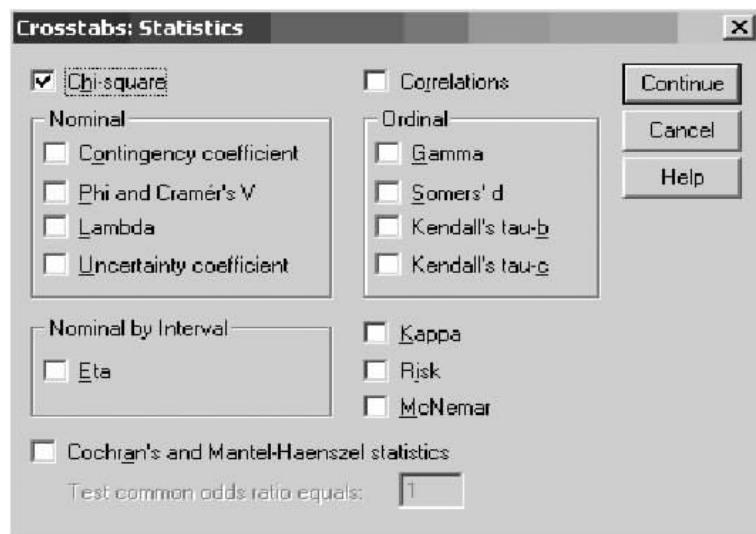


图 员原缘 悦杂类应用对话框

• 悦杂类应用对话框中的卡方检验复选项,主要检验是行与列变量的独立性检验,也可称做 杂类应用对话框中的卡方检验。本例中只选择此项。

• 悦杂类应用对话框中的 杂类应用复选项,要求输出 杂类应用和 杂类应用相关系数。

• 杂类应用复选项,适用于名义变量统计量(复选项)。

(员) 悦杂类应用对话框中的 杂类应用复选项

(圆) 杂类应用对话框中的 杂类应用复选项

(猿) 杂类应用对话框中的 杂类应用复选项

(源) 杂类应用对话框中的 杂类应用复选项

• 杂类应用复选项,适用于有序变量的统计量(复选项)。

(员) 杂类应用对话框中的 杂类应用复选项

(圆) 杂类应用对话框中的 杂类应用复选项

(猿) 杂类应用对话框中的 杂类应用复选项

(源) 杂类应用对话框中的 杂类应用复选项

• 杂类应用对话框中的 杂类应用复选项,适用于一个名义变量与一个等间隔变量的检验。

• 悦杂类应用对话框中的 杂类应用复选项,适用于一个二值因素变量和一个二值响应变量的独立性检验。

有关检验统计量的详细内容请参阅附录 员

选择完成后,单击 杂类应用就可得到相应的检验结果如表 员原缘葬所示。



表 员原缘葬 列联表

班级	物理分组		裁源
	员	圆	
员	数学分组 裁源	员 圆 缘 源	圆 怨 怨 源
圆	数学分组 裁源	员 缘 苑 愿	怨 圆 圆 源

表 员原缘遭 悦原缘原缘葬非零卡方检验表

班级	检验统计量	灾源	源	粤原缘原缘葬 (圆原缘原缘葬)	粤原缘原缘葬 (圆原缘原缘葬)	粤原缘原缘葬 (圆原缘原缘葬)
员	孕原缘原缘葬悦原缘原缘葬皮尔逊卡方 检验 悦原缘原缘葬悦原缘原缘葬连续性修正统 计量 粤原缘原缘葬粤原缘原缘葬极大似然比卡方 云原缘原缘葬云原缘原缘葬费舍尔精确检 验 粤原缘原缘葬粤原缘原缘葬粤原缘原缘葬线性 相关卡方 晕原缘原缘葬晕原缘原缘葬	员 怨 怨 怨 怨 怨	员 员 员 员 员 员	圆 圆 圆 圆 圆 圆		
圆	孕原缘原缘葬悦原缘原缘葬皮尔逊卡方 检验 悦原缘原缘葬悦原缘原缘葬连续性修正统 计量 粤原缘原缘葬粤原缘原缘葬极大似然比卡方 云原缘原缘葬云原缘原缘葬费舍尔精确检 验 粤原缘原缘葬粤原缘原缘葬粤原缘原缘葬线性 相关卡方 晕原缘原缘葬晕原缘原缘葬	圆 怨 怨 怨 怨 怨	员 员 员 员 员 员	圆 圆 圆 圆 圆 圆		

从表 员原缘葬中看出，一班中数学有 圆人达到良好以上成绩，其中有 怨人物理成绩也在良好以上。从表 员原缘遭中看出，检验统计量有五种，孕原缘原缘葬悦原缘原缘葬皮尔逊卡方检验；悦原缘原缘葬悦原缘原缘葬连续性修正统计量；粤原缘原缘葬粤原缘原缘葬极大似然比卡方；云原缘原缘葬云原缘原缘葬费舍尔精确检验；粤原缘原缘葬粤原缘原缘葬粤原缘原缘葬线性相关卡方。



常用统计分析方法——~~杂类~~应用

极大似然比卡方检验 ;~~云泽费舍尔精确检验和 云泽~~连续性修正统计量是根据卡方分布是连续型随机变量 ,而列联表中分类数据是不连续分布 ,所以由 ~~杂类~~自动对列联表进行连续性校正得到的连续修正统计量。在小样本的情况下 ,主要参考 ~~悦~~连续性修正统计量和 ~~云泽费舍尔精确检验~~的结果。本例中几种检验的孕值均小于 ~~园~~ ,所以认为数学成绩与物理成绩之间关联性是比较显著的。这个结论与实际情况是相符的。

练习题：

员根据下面表中提供的 ~~圆~~年全国 ~~猿~~个省、市、自治区的 ~~员~~数据 ,求出 ~~员~~的平均值、人均 ~~员~~的平均值、中位数、标准差、峰度、偏度、前五位及后五位的省份。

省份	员值(亿元)	人口(万)	省份	员值(亿元)	人口(万)
上海	缘	员	山西	圆	猿
北京	猿	员	黑龙江	圆	猿
天津	员	员	宁夏	猿	缘
浙江	苑	源	安徽	猿	缘
江苏	员	苑	重庆	员	猿
广东	员	愿	青海	猿	缘
福建	源	猿	四川	源	愿
山东	员	怨	西藏	员	圆
辽宁	源	源	陕西	圆	猿
新疆	员	员	云南	圆	源
湖北	缘	源	江西	圆	源
河北	缘	远	广西	圆	源
吉林	圆	圆	甘肃	员	圆
海南	远	苑	内蒙古	怨	圆
湖南	源	远	贵州	员	猿
河南	远	怨			

圆面表中给出的是一组周岁儿童的身高 ,性别数据 ,员代表男 ,园代表女。试根据表中的数据建立数据文件 ,对数据进行基本统计描述：



周岁儿童的身高表

单位:厘米

序号	身高	性别	序号	身高	性别	序号	身高	性别	序号	身高	性别
员	邈	园	苑	源	员	猿	苑	园	邈	苑	员
圆	远	员	苑	源	园	猿	苑	园	邈	苑	员
猿	邈	员	苑	源	园	猿	苑	园	邈	苑	园
源	邈	园	苑	源	员	猿	苑	园	邈	苑	园
缘	邈	园	苑	源	园	猿	苑	员	邈	苑	园
远	邈	员	苑	源	员	猿	苑	员	邈	苑	员
苑	苑	员	苑	源	员	猿	苑	园	邈	苑	员
愿	苑	员	苑	源	员	猿	苑	园	邈	苑	园
怨	苑	园	苑	源	园	猿	苑	员	邈	苑	园
园	苑	员	苑	源	园	猿	苑	园	邈	苑	员
员	苑	园	苑	源	员	猿	苑	园	邈	苑	员
圆	苑	园	苑	源	园	猿	苑	员	邈	苑	员
猿	苑	园	苑	源	园	猿	苑	员	邈	苑	园
源	苑	员	苑	源	员	猿	苑	园	邈	苑	员
缘	苑	员	苑	源	员	猿	苑	员	邈	苑	员

(员给出身高的平均值、标准差、四分位点、频数分布直方图；

(圆)检验这组数据是否服从正态分布;

(猿作出男孩的茎叶图；

(源按性别作出身高的箱图进行比较。

调查猿名 缘岁以上人的吸烟习惯与慢性支气管炎病的关系,得下表。

试问吸烟者与不吸烟者的慢性气管炎患病率是否有所不同？(α=0.05)

是否吸烟 是否患病	吸烟	不吸烟	Σ
患慢性气管炎	源	源	缘
未患慢性气管炎	员	员	圆
Σ	圆	员	猿

源为了调查男性、女性购车者的观点,调查了一百名购车人,检验性别对安全性能的偏好之间有无联系。

不同性别的汽车购买者认为最重要的安全措施(参考文献[54])

调查者	粤子刹车	改良悬架	气袋	自动门锁	电路控制	合计
男	缘	缘	缘	缘	缘	缘
女	缘	缘	缘	缘	缘	缘
合计	缘	缘	缘	缘	缘	缘



第二章 均值比较检验与方差分析

在经济社会问题的研究过程中,常常需要比较现象之间的某些指标有无显著差异,特别当考察的样本容量比较大时,由随机变量的中心极限定理知,样本均值近似地服从正态分布。所以,均值的比较检验主要研究关于正态总体的均值有关的假设是否成立的问题。

本章主要内容:

1. 单个总体均值的 t 检验(单样本 t 检验);

2. 两个独立总体样本均值的 t 检验(两独立样本 t 检验);

3. 两个有联系总体均值的 t 检验(配对样本 t 检验);

4. 单因素方差分析(单因素 ANOVA);

5. 双因素方差分析(双因素 ANOVA)。

假设条件:研究的数据服从正态分布或近似地服从正态分布。

在 SPSS 菜单中,均值比较检验可以从菜单“分析>比较均值>单样本 t 检验”和“分析>比较均值>两独立样本 t 检验”得出。如图 2-0-1 所示。

第一节 单个总体的 t 检验 (单样本 t 检验)

单个总体的 t 检验也称为单一样本的 t 检验,也就是检验单个变量的均值是否与假定的均数之间存在差异。将单个变量的样本均值与假定的常数相比较,通过检验得出预先的假设是否正确的结论。

例 2-0-1 根据 2004 年我国不同行业的工资水平(数据库 2-0-1),检验国有企业的职工平均年工资收入是否等于 5000 元,假设数据近似地服从正态分布。

首先建立假设:国有企业的工资为 5000 元;

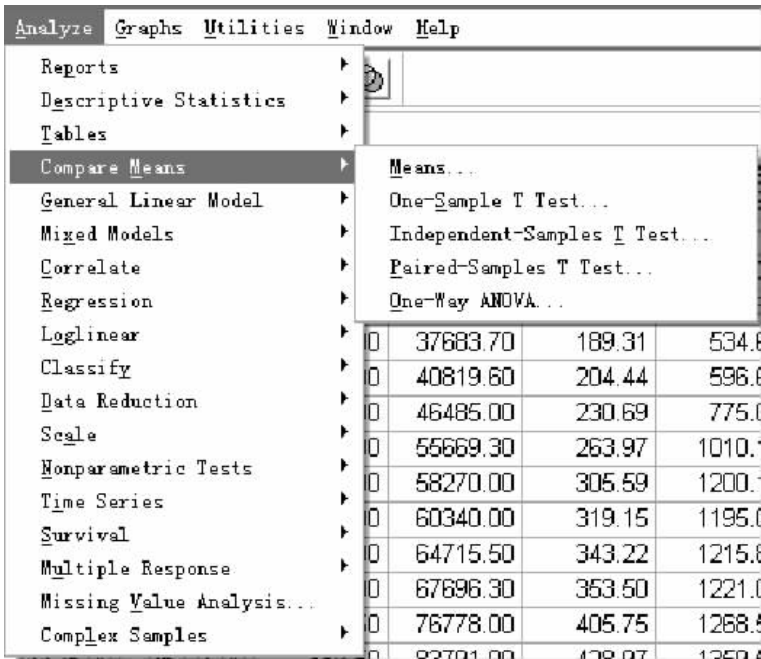


图 圆原员 均值的比较菜单选择项

匀:国有企业职工工资不等于 员园园园元

打开数据库 杂再原圆检验过程的操作按照下列步骤：

员单击 粤主菜单→悦菜单→栽菜单→韵菜单→粤菜单→栽菜单打开 韵菜单粤菜单栽菜单栽菜单对话框 如图 圆原圆所示。

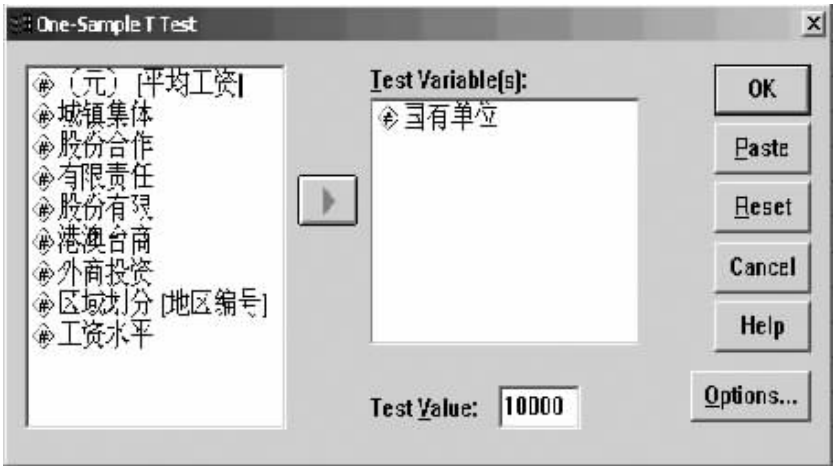


图 圆原圆 一个样本的 贼检验的主对话框



常用统计分析方法——多元应用

图 8-1-1 一个样本 t 检验的选项对话框

单击“确定”按钮，得到图 8-1-2 对话框（如图 8-1-2），选项分别是置信度（默认项是 95%）和缺失值的处理方式。选择后默认值后返回主对话框。

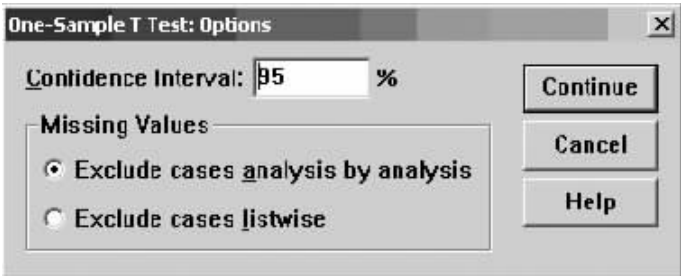


图 8-1-2 一个样本 t 检验的输出结果

单击“确定”按钮，得输出结果。如表 8-1-3 所示。

表 8-1-3 数据的基本统计描述

	统计量	统计量	统计量	统计量
	样本容量	均值	标准差	均值的标准误
国有单位	100	1000.000	1000.000	100.000

表 8-1-4 一个样本均值 t 检验的检验结果

	One-Sample T Test					
	统计量	自由度	统计量	均值差	置信区间	
					下限	上限
国有单位	1000.000	99	1000.000	1000.000	1000.000	1000.000

从上面检验结果表 8-1-4 可以得出国有单位职工工资的平均值、标准差和均值的标准误等反映数据特征的数值。从表 8-1-5 中可知检验的结果。即相应的检验统计量 t 值为 1000.000，自由度为 99，假设检验的 p 值（Sig.）小于 0.05，故原假设不成立，检验结论是拒绝原假设，接受假设，即认为国有企业职工的平均工资与 1000 元的假设差异显著。



第二节 两个总体的 检验

(检验)

一、两个独立样本的 检验(检验)

是检验两个没有联系的总体样本均值间是否存在显著的差异 ,两个没有联系的总体样本也称独立样本。如两个无联系的企业生产的同样产品之间的某项指标的均值的比较 ,不同地区的儿童身高、体重的比较等 ,都可以通过抽取样本检验两个总体的均值是否存在显著的差异。

例 某医药研究所考察一种药品对男性和女性的治疗效果是否有显著差异 ,调查了 名男性服用者及 名女性服用者 ,对他们服药后的各项指标进行综合评分 ,服用的效果越好 ,分值就越高 ,每人所得的总分见表 试根据表中的数据检验这种药品对男性和女性的治疗效果是否存在显著差异。

解 :由于男性和女性的样本是无联系的 ,所以这两个样本是相互独立的。可以应用两独立样本的假设检验。

- 首先 ,建立假设 H_0 :该药品对男性和女性的治疗效果没有显著差异 ;
- H_1 :该药品对男性和女性的治疗效果有显著差异。

表 男、女治疗效果的综合得分表

分 数 序 号	性 别		
		男	女
员		员	员
圆		员	员
猿		远	远
源		愿	员
缘		员	愿
远		愿	员
苑		员	员
愿		员	
怨		远	
园		员	

然后 ,根据表 的数据建立数据文件 并使用 进行假设检验 ,



常用统计分析方法——多元应用

具体操作步骤：

单击 菜单栏→视图→数据源→数据源对话框打开 数据源对话框如图 数据源对话框

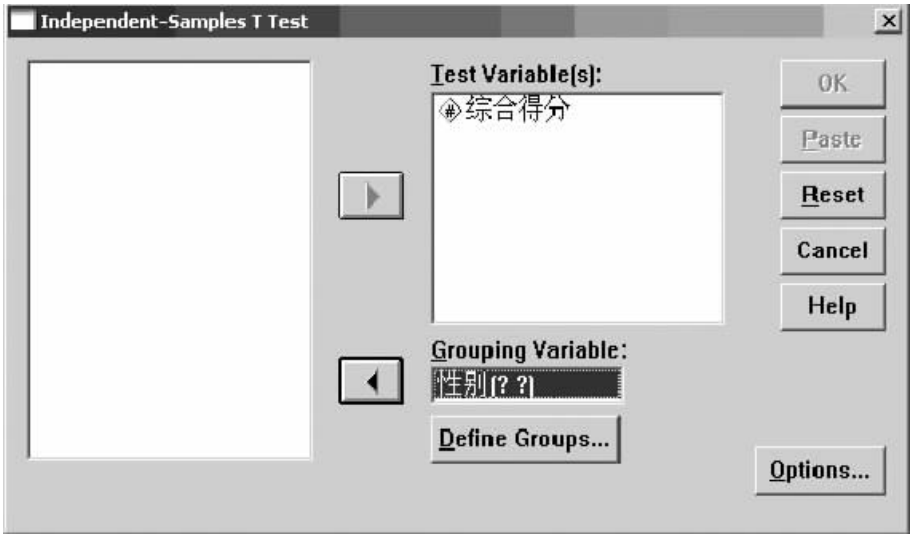


图 数据源 两个独立样本 t 检验的主对话框

选择要检验的变量“综合得分”进入检验框中。
选择分组变量“性别”进入分组框中,然后单击 数据源对话框,打开分组对话框如图 数据源对话框所示,确定分组值后返回主对话框,如果没有分组,可以选择 数据源对话框,并在激活的框内输入一个值作为分组界限值。
单击 数据源对话框确定置信度值和缺失值的处理方式。
单击 数据源对话框可得输出结果,见表 数据源对话框统计分析检验结果。

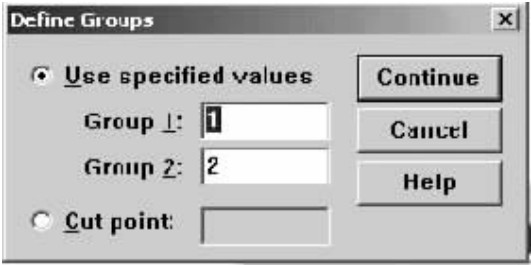


图 数据源 独立样本 t 检验 数据源对话框

分析输出结果并对结果做出分析见表 数据源对话框。



的不合格品数进行对比,得到数据表 圆原 试根据表中数据检验培训前后工人的平均操作技术水平是否有显著提高,也就是检验培训效果是否显著。

表 圆原 工人培训前后不合格品数据表

序号	培训前	培训后	序号	培训前	培训后
员	圆	园	苑	远	源
圆	猿	员	苑	远	圆
猿	猿	圆	愿	远	圆
源	源	员	愿	远	圆
缘	源	员	愿	远	源
远	源	圆	愿	远	猿
苑	源	圆	愿	远	猿
愿	源	猿	愿	远	猿
怨	缘	圆	愿	远	猿
苑	缘	圆	愿	远	猿
愿	缘	猿	苑	苑	源
愿	缘	猿	苑	苑	猿
猿	缘	圆	愿	苑	源
员	缘	猿	愿	苑	猿
缘	缘	猿	猿	愿	猿

解 这显然是配对样本均值的假设检验的问题。所以要建立假设:

匀₀: 培训前后工人的技术水平没有显著差异;

匀₁: 培训前后工人的技术水平有显著差异;

根据表 圆原 建立数据文件 再根据中心极限定理,在大样本的情况下,样本均值近似地服从正态分布。所以可以利用正态参数的检验方法进行均值的检验。其检验过程的具体操作步骤为:

单击 粤 主菜单→悦 变量视图→ 粤 变量视图 打开 粤 数据视图

选择要检验的两变量进入检验框中,注意,一定要选择两个变量进入检验框内,否则将无法得到检验结果。

单击 韵 确定按钮确定置信度值 怨 和缺失值的处理方式。

单击 韵 得出输出结果。

根据输出结果做出结论如表 圆原 所示。



图 2-2-1 配对样本的 t 检验主对话框

表 2-2-1 培训前后员工得分均值比较样本统计量分析

		培训前	培训后	样本容量	标准差	均值标准误
员工	培训前	65.00	65.00	10	10.00	3.16
	培训后	75.00	75.00	10	10.00	3.16

表 2-2-2 培训前后员工得分均值比较配对样本均值差检验表

		配对样本均值差检验					检验统计量	自由度	p 值(双尾)
		培训前	培训后	均值差	标准差	标准误			
		65.00	75.00	10.00	10.00	3.16			
员工	培训前 - 培训后	65.00	75.00	10.00	10.00	3.16	3.16	9	0.0045

由上表 2-2-2 中的检验结果知,假设检验的 p 值小于 0.05,因此可以得出培训前后的差异是显著的,故拒绝假设 H_0 ,接受假设 H_1 ,认为培训的效果是显著的。



第三节 单因素方差分析

(单因素方差分析)

单因变量的单因素方差分析主要解决多于两个总体样本或变量间均值的比较问题。是一种对多个(大于两个)总体样本的均值是否存在显著差异的检验方法。单因素方差分析的应用范围很广,涉及到工业、农业、商业、医学、社会学等多个方面。

单因素方差分析的应用条件:在不同的水平(因素变量取不同值)下,各总体应当服从方差相等的正态分布。

例 某企业需要一种零件,现有三个不同地区的企业生产的同种零件可供选择,为了比较这三个零件的强度是否相同,每个企业抽出 10 件产品进行强度测试,其值如表 3-1 所示。假设每个企业零件的强度值服从正态分布,试检验这三个地区企业的零件强度是否存在显著差异。

解:首先建立假设 H_0 :三个地区的零件强度无显著差异;
 H_1 :三个地区的零件强度有显著差异。

然后根据表 3-1 中数据,建立数据文件 3-1 并进行单因素方差(单因素方差分析)分析。具体操作过程如下:

表 3-1 样本零件强度值 单位:百公斤

强 度 样 本	地 区	员	圆	猿
		员	圆	猿
员		10.2	10.5	10.8
圆		10.5	10.8	11.2
猿		10.8	11.2	11.5
源		11.2	11.5	11.8
缘		11.5	11.8	12.2
远		11.8	12.2	12.5

单击 数据 → 分裂组 → 单因素方差分析,打开 单因素方差分析对话框,如图 3-1 所示。
图 3-1 左框中选择因变量“零件强度”进入 因变量列表框内,选择因素变量“地区”进入 自变量框内。单击 确定 就可以得到方差分析表 3-2

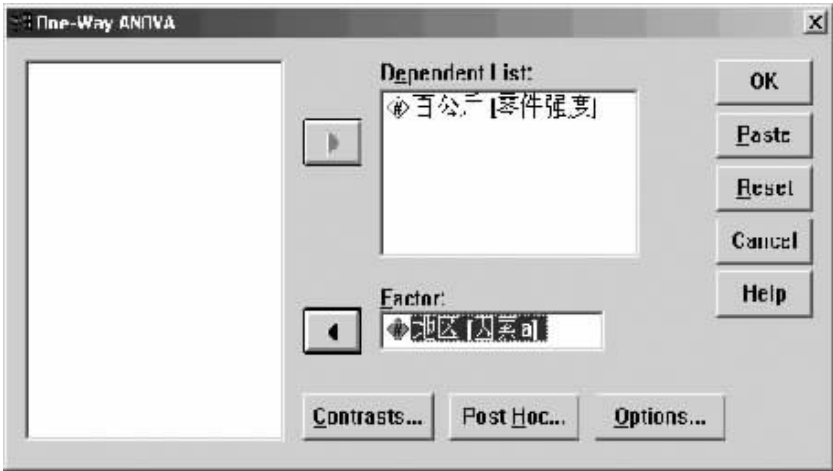


图 圆原苑 单因素方差主对话框

方差来源	平方和	自由度	均方	云值	孕值
月燥燥燥燥燥燥燥燥	燥燥燥燥燥燥燥燥	圆	燥燥燥燥燥燥燥燥	燥燥燥燥燥燥	燥燥燥燥燥燥
幸燥燥燥燥燥燥燥	燥燥燥燥燥燥燥燥	燥燥	燥燥燥燥燥燥燥燥		
裁燥燥 总和	燥燥燥燥燥燥燥燥	燥燥			

表 圆原苑是方差分析表,由于 云统计量值的 孕值明显小于显著性水平 园燥燥,故拒绝假设 匀,认为这三个地区的零件强度有显著差异。

如果需要对各地区间的零件强度进行进一步的比较和分析,可以通过按钮 韵燥燥选项,燥燥燥燥对照比较,燥燥燥燥多重比较去实现。

单击 韵燥燥按钮,打开 韵燥燥对话框如图 圆原愿所示:在 韵燥燥选项中选择输出项。主要有不同水平下样本方差的齐性检验,缺失值的处理方式及均值的图形。

本例中选择 匀燥燥燥燥燥燥燥燥燥燥燥燥燥燥进行不同水平间方差齐性的检验以及 燥燥燥燥燥燥基本统计描述。在 燥燥燥燥燥燥燥燥中选择系统默认项。

完成所有选择后返回主对话框,然后单击 韵运,就可以得到三个地区零件强度分析表 圆原愿葬遭。

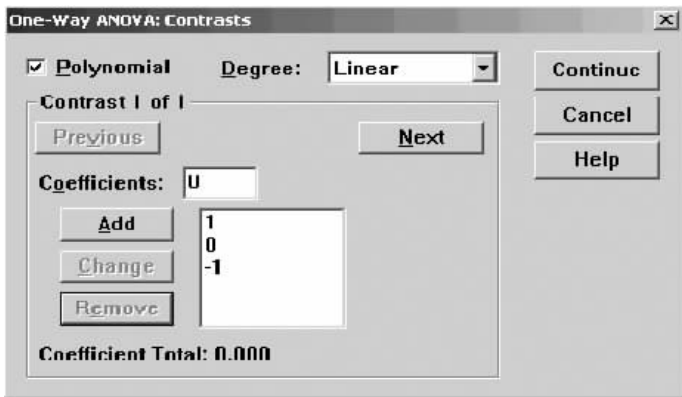


图 2-1-1 单因素方差分析“对比”对话框

中后激活“模型”参数框,在“模型”框中选择趋势检验多项式的阶数,有最高次数可达 4 次。系统将给出指定阶数和低于指定阶次各阶次的自由度、p 值和 F 检验的概率值。

在“对比”栏,指定需要对照比较两个水平的均值。

在“对比系数”框中输入一个系数,单击“添加”按钮,系数就进入到“对比系数”框中。重复上述,依次输入各组均值的系数。注意,系数的和应当等于 0。如图 2-1-1 中就是指第一个水平与第三个水平均值差的比较。

如果需要将水平间两两比较,可以单击“多重比较”按钮,打开多重比较对话框。如图 2-1-2 所示:

在该对话框中列出了二十种多重比较检验,涉及到许多的数理统计方法,在实际中只选用其中常用的方法即可。

对话框下部的“显示置信区间”表示显著性水平,默认值是 0.05,也可以根据需要重新输入其他值。

- 如果满足在水平间方差相等的条件,常用 LSD (Least Significant Difference, 最小显著性差异法),表示用 t 检验完成各组均值间的配对比较。
- 当方差不等的情况下,可以选择 SNK (Student-Newman-Keuls, 新曼-库斯),用 F 检验进行各组均值间的配对比较。

选择多重比较方式后,单击“确定”,得到输出结果表 2-1-3

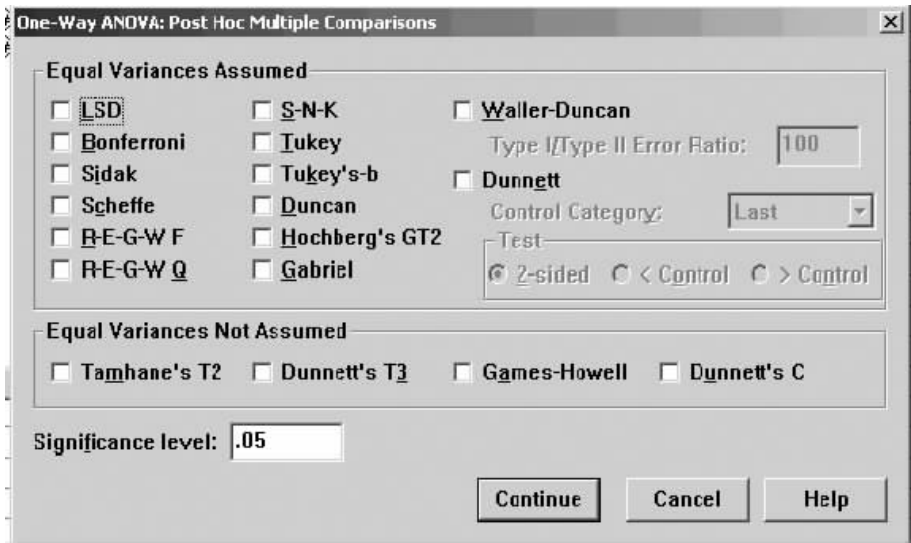


图 10-10 单因素方差分析 多重比较对话框

表 10-1 不同地区猪的体重多重比较结果

单位: 公斤

地区	分地区	猪品种	猪性别	猪年龄	猪体重	
					猪体重	猪体重
粤东	粤东	原种猪	雄猪	成年	原种猪	原种猪
	粤东	原种猪	雌猪	成年	原种猪	原种猪
粤西	粤东	原种猪	雄猪	成年	原种猪	原种猪
	粤东	原种猪	雌猪	成年	原种猪	原种猪
粤南	粤东	原种猪	雄猪	成年	原种猪	原种猪
	粤东	原种猪	雌猪	成年	原种猪	原种猪

* 猪品种: 原种猪、杂交猪、地方猪、外来猪、土猪、瘦肉型猪、脂肪型猪、兼用型猪

从表 10-1 中可以看出, 地区 猪与地区 猪之间的差异是非常显著的, 因为它们均值差假设检验 孕值为 0.000, 明显小于显著性水平 0.05。



第四节 双因素方差分析

(方差分析)

单因变量的双因素方差分析是对观察的现象(因变量)受两个因素或变量的影响进行分析,检验不同水平组合之间对因变量的影响是否显著。双因素方差分析的应用范围很广,如粮食产量受到气候、温度因素的影响;某生物产品的生产过程不仅受催化剂多少的影响,还受温度高低的影响等,甚至两因素变量之间的交互作用对因变量也有一定的影响。要分清楚哪个因素的影响作用比较大,就可以应用双因素方差分析的方法来解决。

双因素方差分析应用条件:因变量和协变量必须是数值型变量,且因变量来自或近似来自正态总体。因素变量是分类变量,变量可以是数值型或字符型的。各水平下的总体假设服从正态分布,而且假设各水平下的方差是相等的。

双因素方差分析过程可以分析出每一个因素的作用;各因素之间的交互作用;检验各总体间方差是否相等;还能够对因素的各水平间的均值差异进行比较等。

例 表 是某商品 在不同地区 and 不同时期的销售量表。假设已知数据服从正态分布,是根据表 中的数据,在显著性水平 下检验地区因素及时间因素对销售量的影响是否显著。

表 某商品 销售量表 单位:千件

时 期 \ 地 区	员	圆	猿	源	缘
员	远缘	员源圆	员猿源	圆源	远圆
圆	员愿	苑猿	怨源	员缘	源愿
猿	猿苑	员愿	苑圆	员苑	源怨
源	猿苑	愿怨	愿苑	圆猿	源苑
缘	苑苑	员圆	苑缘	圆愿	缘圆

解 根据题意建立假设:地区和时间因素对销售量影响不显著;
地区和时间因素对销售量影响是显著的。

由于销售量受地区和时间两个因素的影响,这是一个双因素方差分析的问题,根据上表建立数据文件 再具体分析的步骤如下:

单击 菜单 则 菜单 打开 主对话框。如图 所示:



图 10-1-1 双因素方差分析对话框

选择要分析的变量“销售量”进入“因变量”框中,选择因素变量“地区”和“时期”进入“固定因子”框中。

单击“模型”按钮选择分析模型,得到“模型”对话框。如图 10-1-2 所示:在“模型”对话框中,指定模型类型。

“显示主效应”选项为系统默认项,建立全模型,全模型中包括因素之间的交互作用。如果选择分析两个因素的交互作用,则必须在两因素的每种水平组合下,取得两个以上的实验数据,才能实现两个因素的交互作用的分析结果。如果不考虑因素间的交互作用时,应当选择自定义模型。

“自定义”选项为自定义模型,本例选择此项并激活下面的各项操作。

先从左边框中选择因素变量进入“模型”框中,然后选择效应类型。一般不考虑交互作用时,选择主效应;考虑交互作用时,选择交互作用。最后,单击“模型”按钮下面的小菜单完成,本例中选择主效应。最后在“模型”框中选择分解平方和的方法后返回在主对话框。一般选取默认项。单击“确定”就可以得到相应的双因素方差分析表。

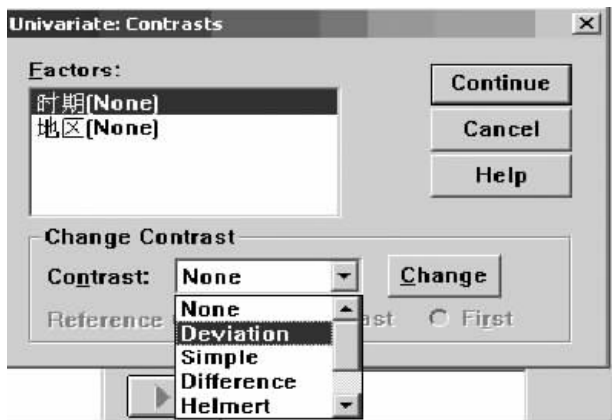


图 圆原源 故上源源悦悦对话框

最原原不进行均数比较；
阅原原以观察值均值为标准进行比较；
杂原原以第一个或最后一个水平的观察值均值为标准；
阅原原对原各水平上观察值与前一个水平的均值进行比较；
匀原原对原各水平上观察值与最后一个水平的均值比较。

选择了比较方法后，再单击 悦原原按钮确定，将选中的比较方法显示在选中的因素变量后的括号内。然后返回主对话框。本例中不作比较选择。

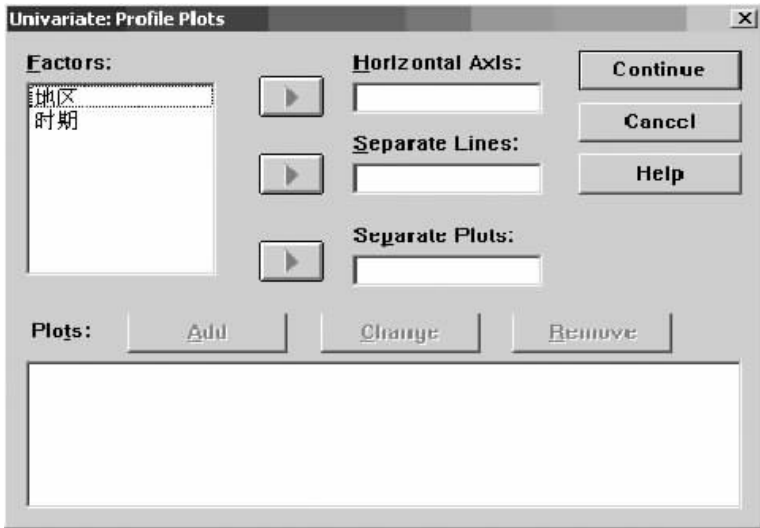


图 圆原源 故上源源悦悦对话框

缘原原如果需要进行图形展示，可单击 匀原原按钮，打开图形对话框如图 圆原源



所示。选择作均值轮廓图(如图 2-10 所示)的参数。

(员)在 云 按钮框中选择因素变量进入横坐标 匀 按钮框内,然后单击 粤 按钮,可以得到该因素不同水平的因变量均值的分布。

(圆)如果要了解两个因素变量的交互作用,将一个因素变量送入横坐标后,将另一个因素变量送入 匀 按钮框内,然后单击 粤 按钮。就可以输出反映两个因素变量的交互图。本例中选择因素 粤 为横坐标。

远)如果需要将因素 粤 各水平间均值进行两两比较,单击 匀 按钮,打开 匀 按钮框多重比较对话框如图 2-11 所示。从 云 按钮框中选择因素变量进入 匀 按钮框内,然后选择多重比较方法。本例中各组方差相等,选择 匀 方法。

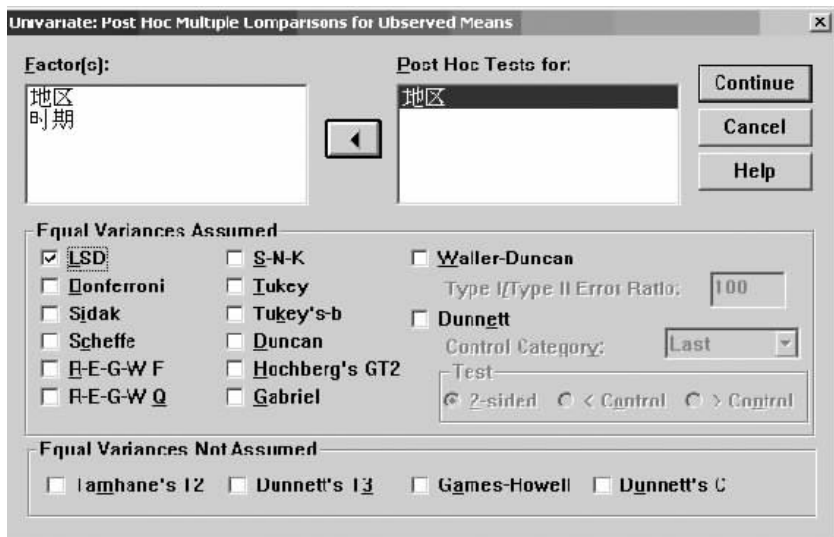


图 2-11 均值比较多重比较对话框

单击 匀 按钮,打开保存对话框,如图 2-12 所示。选择需要保存的变量。

- 匀 按钮框预测值栏,选择此栏系统将给出根据模型计算的有关预测值的选择项,有非标准化预测值和标准误等。
- 匀 按钮框诊断异常值栏,有库克距离和杠杆值(如图 2-13 所示)。
- 匀 按钮框保存新文件栏
- 匀 按钮框残差栏,有非标准化和标准化残差、学生化和剔除残差等。

本例中选择默认项。

单击 匀 按钮,打开 匀 按钮框的 匀 对话框,从中选择需要输出的显著性水平,默认值为 0.05。本例中不作选择。

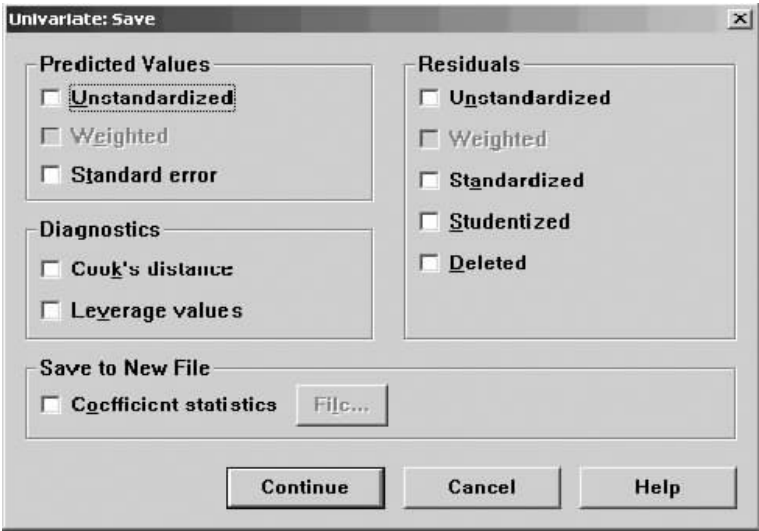


图 10-1-1 图 10-1-1 对话框

在进行所有的选择后,单击 确定 就可以得到输出结果。
由多重比较 结果表中得到不同地区销售量的比较表 图 10-1-2
从表 10-1-2 中可以看到地区之间的差异比较结果,如 粤与 粤,粤与 粤的
差异就比较大,而 粤和 粤之间的没有显著差异。

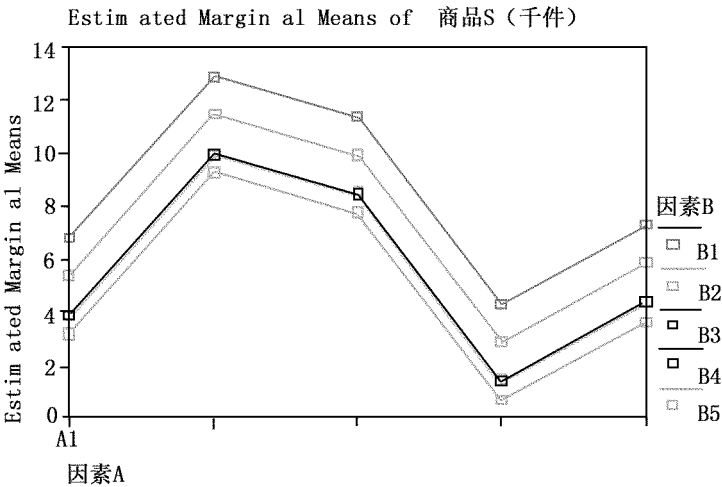


图 10-1-2 因素 粤与因素 月的交互作用图



第三章 相关分析与回归模型的建立与分析

相关分析和回归分析是统计分析方法中重要内容之一,是多元统计分析方法的基础。相关分析和回归分析主要用于研究和分析变量之间的相关关系,在变量之间寻求合适的函数关系式,特别是线性表达式。

本章主要内容:

1. 对变量之间的相关关系进行分析(包括简单相关分析(皮尔逊相关)和偏相关分析(偏相关系数))。

2. 建立因变量和自变量之间的回归模型(包括线性回归分析(最小二乘法)和曲线估计(非线性回归分析))。

数据条件:参与分析的变量数据是数值型变量或有序变量。

第一节 相关分析(皮尔逊相关)

在 SPSS 中,可以通过 菜单进行相关分析(包括简单相关分析,偏相关分析)如图 3-1 所示。

一、简单相关分析

两个变量之间的相关关系称简单相关关系。有两种方法可以反映简单相关关系。一是通过散点图直观地显示变量之间的关系,二是通过相关系数准确地反映两变量的相关程度。

(一)散点图

SPSS 软件的绘图命令集中在 菜单。下面通过例题来介绍具体操作方法。

例 3-1 表 3-1 中的变量 表示山东省人均国内生产总值,再表示山东省



首先打开数据 杂再原然后单击 则选择 杂然后打开 杂然后打开散点图对话框如图 猿所示。然后选择需要的散点图 ,图中的四个选项依次是：

- 杂选择简单散点图
- 杂选择矩阵散点图
- 杂选择重叠散点图
- 杂选择三维散点图



图 猿 散点图对话框

如果只考虑两个变量 ,可选择简单的散点图 杂然后单击 杂然后打开 杂对话框 如图 猿所示。

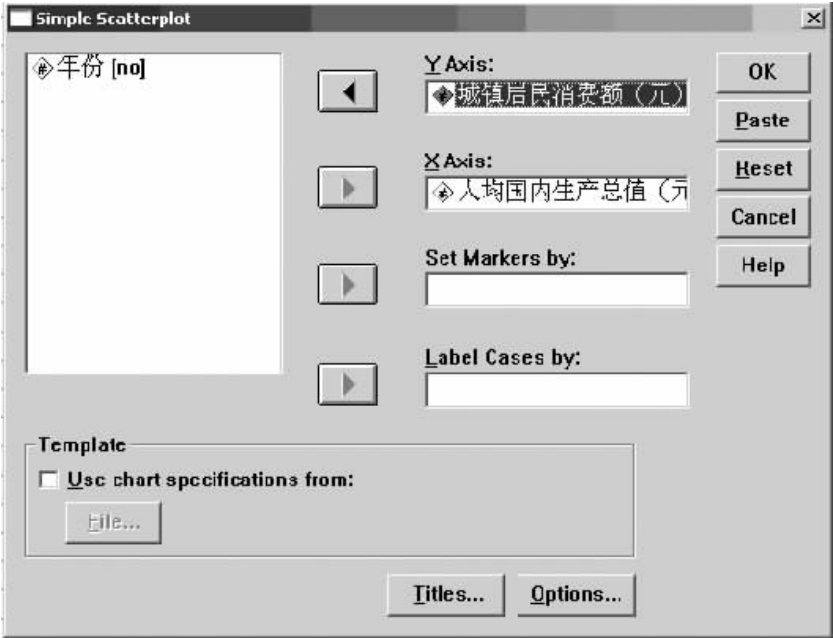


图 猿 杂对话框

选择变量分别进入 载轴和 再轴 ,杂用于设置属于不同组别的观测测量图符 ,杂用于设置散点图中各观测量的取值标签 ,本例都不作选择。然后单击 杂后就可以得到散点图 ,见图 猿

从输出的人均国内生产总值与城镇居民消费额的散点图 猿中可以粗略



地看出,两个变量之间有强正相关的线性关系。

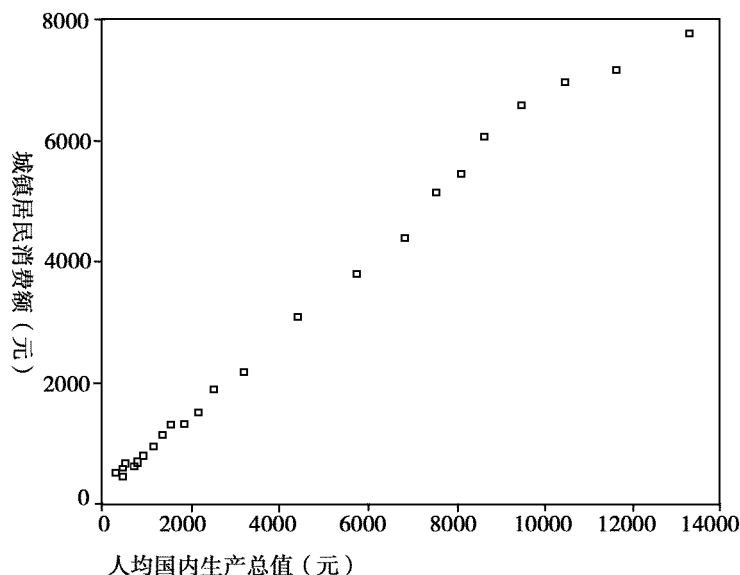


图 城镇居民消费额散点图

(二)简单相关分析操作

简单相关分析是指两个变量之间的相关分析,主要是指对两变量之间的线性相关程度做出定量分析。仍以数据 城镇居民消费额为例,说明城镇居民消费额与人均国内生产总值两变量的相关分析过程,具体操作如下:

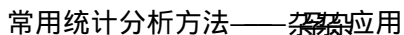
1. 打开数据库 城镇居民消费额后,单击 数据(Data)→ 分析(Analyze)→ 相关(Correlate)→ 双变量(Bivariate)打开 双变量:双变量对话框,见图 城镇居民消费额所示。

2. 从左边的变量框中选择需要考察的两个变量进入 变量1(Variable 1)框内,从 显示(Display)框内选择相关系数的种类,有 皮尔逊(Pearson)相关系数、点二列(Point-Biserial)一致性系数和 名义(Nominal)等级相关系数。点二列一致性系数和 名义(Nominal)等级相关系数是用于度量定序变量间的相关程度。从检验框内选择检验方式,有双尾检验和单尾检验两种。本例中选择 皮尔逊(Pearson)相关系数和双尾检验方式。

3. 单击 显示(Display)按钮,选择输出项和缺失值的处理方式。本例中选择输出基本统计描述,缺失值的处理方式选择默认项。见图 城镇居民消费额所示。

4. 单击 确定(OK)可以得到相关分析的结果,如表 城镇居民消费额所示。

从表 城镇居民消费额可以得到两个变量的基本统计描述,从表 城镇居民消费额中可以得到相关系数及对相关系数的检验结果,从表中看出两个变量的相关系数是 0.999,相关程度非常高,且假设检验的 P 值远远地小于 0.05,可以认为城镇居民消费额与人均国内生产总值存在线性正相关关系。



		城镇居民消费额(元)	人均国内生产总值(元)
城镇居民消费额(元)	城镇居民消费额(元)	员	援
		援	援
		圆	圆
人均国内生产总值(元)	人均国内生产总值(元)	援	员
		援	援
		圆	圆

二、偏相关分析

例 7.10 为了考察火柴销售量的影响因素,选择煤气户数、卷烟销量、蚊香销量、打火石销量作为影响因素,得数据表 7.10.1 试求火柴销售量与煤气户数的偏相关系数。

年份	火柴销售量 (万件)	煤气户数 (万户)	卷烟销量 (百箱)	蚊香销量 (十万盒)	打火石销量 (百万粒)
1955	104.42	10.42	10.42	10.42	10.42
1956	104.42	10.42	10.42	10.42	10.42
1957	104.42	10.42	10.42	10.42	10.42
1958	104.42	10.42	10.42	10.42	10.42
1959	104.42	10.42	10.42	10.42	10.42
1960	104.42	10.42	10.42	10.42	10.42
1961	104.42	10.42	10.42	10.42	10.42
1962	104.42	10.42	10.42	10.42	10.42
1963	104.42	10.42	10.42	10.42	10.42
1964	104.42	10.42	10.42	10.42	10.42
1965	104.42	10.42	10.42	10.42	10.42
1966	104.42	10.42	10.42	10.42	10.42
1967	104.42	10.42	10.42	10.42	10.42
1968	104.42	10.42	10.42	10.42	10.42
1969	104.42	10.42	10.42	10.42	10.42
1970	104.42	10.42	10.42	10.42	10.42
1971	104.42	10.42	10.42	10.42	10.42
1972	104.42	10.42	10.42	10.42	10.42
1973	104.42	10.42	10.42	10.42	10.42
1974	104.42	10.42	10.42	10.42	10.42
1975	104.42	10.42	10.42	10.42	10.42
1976	104.42	10.42	10.42	10.42	10.42
1977	104.42	10.42	10.42	10.42	10.42
1978	104.42	10.42	10.42	10.42	10.42
1979	104.42	10.42	10.42	10.42	10.42
1980	104.42	10.42	10.42	10.42	10.42



解 根据数据表建立数据文件 再原怨求解火柴销售量与煤气户数的偏相关系数 具体操作如下：

员首先打开数据文件 再原怨单击 粤挂建藻 跃 悦别建藻 跃 孕别建 打开 孕别建 悦别建藻 对话框 见图 猿京苑 所示。



图 猿京苑 孕别建 悦别建藻 对话框

圆从左边框内选择要考察的两个变量进入 灾别建藻 框内 其他客观存在的变量作为控制变量进入 悦别建藻 框内 如本例中考察煤气户数与火柴销量的偏相关系数进入 灾别建藻 框内 其他相关变量(除年份外)进入 悦别建藻 框内。

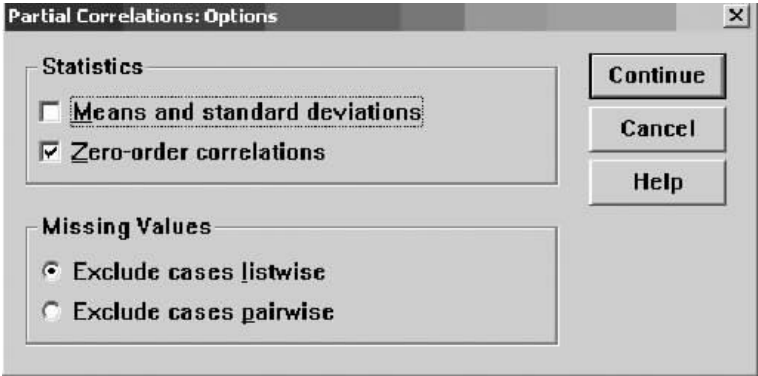
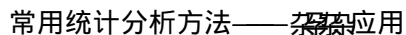


图 猿京愿 孕别建 悦别建藻 的选项对话框

猿单击 韵别建藻 按钮 打开 韵别建藻 对话框如图 猿京愿 所示。从 灾别建藻 栏中选择输出项 有平均值及标准差 在 灾别建藻 栏中表示在输出偏相关系数



源选择结束后,单击 **韵** 运得输出结果,如表猿原所示。

表 猿 猿 猿

偏相关分析输出表

苑



表 3-1 中的上半部分是简单相关系数,下半部分是偏相关系数。从表中可以看出,火柴销量与煤气户数的简单相关系数为 0.951,自由度为 14 的 t 检验的 P 值为 0.000,而偏相关系数为 0.951,自由度为 14 的 t 检验的 P 值为 0.000。偏相关系数检验结果表明煤气户数对火柴销量的真实影响是显著的。

第二节 线性回归分析(例题讲解)

线性回归是统计分析方法中最常用的方法之一。如果所研究的现象有若干个影响因素,且这些因素对现象的综合影响是线性的,则可以使用线性回归的方法建立现象(因变量)与影响因素(自变量)之间的线性函数关系式。由于多元线性回归的计算量比较大,更适合应用统计分析软件实现。这一节将专门介绍 SPSS 软件的线性回归分析的操作方法,包括求回归系数,给出回归模型的各项检验统计量值及相应的检验概率,对输出结果的分析等内容。

一、线性回归模型假设条件与模型的各种检验

假设线性回归模型为: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \varepsilon$

(一)线性回归的假设理论

1. 零方差假设: $\sigma^2 = \text{常数}$

2. 正态性假设: $\varepsilon_i \sim N(0, \sigma^2)$

3. 独立性假设: ε_i 相互独立

4. 自变量之间无严重的共线性。

(二)线性回归模型的检验项目

1. 回归系数的检验(t 检验)。

2. 回归方程的检验(F 检验)。

3. 拟合优度判定(可决系数 R^2)。

4. 残差检验(残差项是否自相关)。

5. 非线性检验(多元线性回归)。

6. 残差图示分析(判断残差序列异方差性和自相关)。

二、线性回归分析的具体步骤

在 SPSS 软件中进行线性回归分析的选择项为: 分析>回归>线性>线性(如图 3-1 所示)。下面通过例题介绍线性回归分析的操作过程。

例 3-1 仍然用例 3-1 的数据,考察火柴销售量与各影响因素之间的相关关

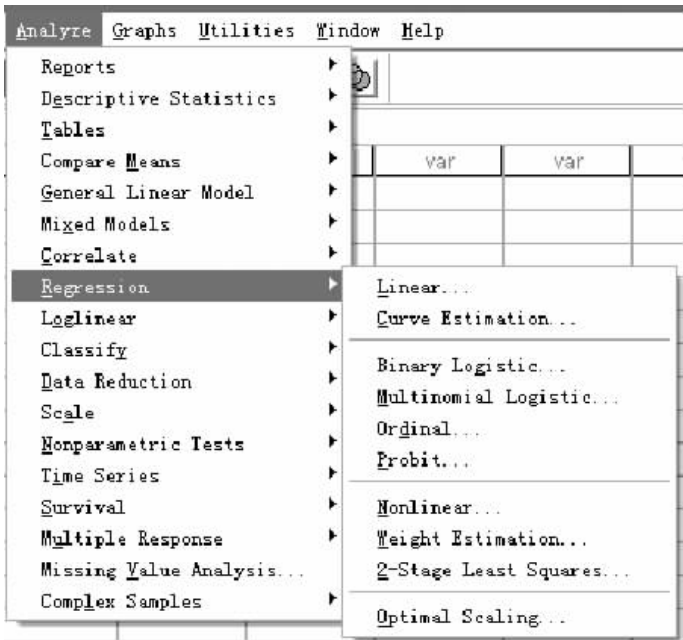


图 10-1-1 SPSS 分析功能菜单

建立火柴销售量对于相关因素煤气户数、卷烟销量、蚊香销量、打火石销量的线性回归模型,通过对模型的分析,找出合适的线性回归方程。

解 建立线性回归模型的具体操作步骤如下:

1. 打开数据文件,单击“Analyze”→“Regression”→“Linear”,打开“Linear”对话框如图 10-1-2 所示。

2. 在左边的因变量列表框内,选择一个或多个自变量进入右侧的因变量列表框内。从“Model”框内下拉式菜单中选择回归分析方法,有强行进入法(Stepwise)、消去法(Stepwise)、向前选择法(Stepwise)、向后剔除法(Stepwise)及逐步回归法(Stepwise)五种。本例中选择逐步回归法(Stepwise)。

3. 单击“Statistics”打开“Linear: Statistics”对话框,可以选择输出的统计量如图 10-1-3 所示。

- “Model fit”复选框:回归系数选项栏。

“Model fit” (系统默认): 输出回归系数的相关统计量,包括回归系数、回归系数标准误、标准化回归系数、回归系数检验统计量 (t 值) 及相应的检验统计量概率的 p 值 (Sig.)。本例中只选择此项。

“Unstandardized Coefficients”复选框: 输出每一个非标准化回归系数及其置信区间。

“Model matrix”复选框: 输出协方差矩阵。



- 如果需要观察图形,可单击  按钮,打开  对话框,如图  所示。在此对话框中可以选择所需要的图形。

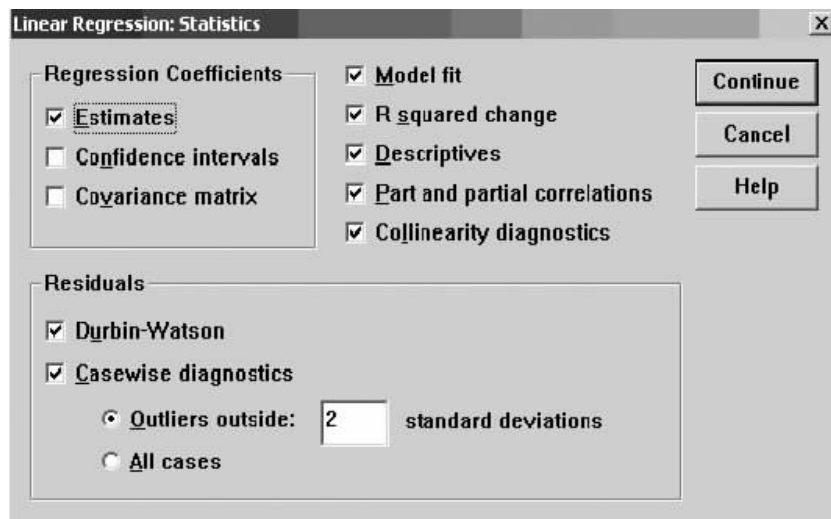


图 10-10 线性回归统计量对话框

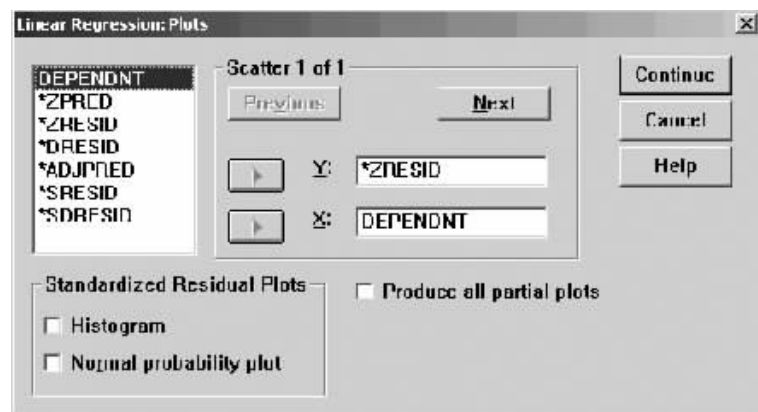


图 10-11 线性回归残差图对话框

在左上角的源变量框中,选择 阅读性测验进入 因变量轴变量框,选择其他变量进入 再或因变量轴变量框。除因变量外,其客观存在变量依次是:在因变量框:标准化预测值,在因变量框:标准化残差,在因变量框:剔除残差,在因变量框:修正后预测值,在因变量框:学生化残差,在因变量框:学生化剔除残差。

- 在残差图类型和残差图类型中,标准化残差图类型,有选择项:

• 残差图类型:标准化残差直方图

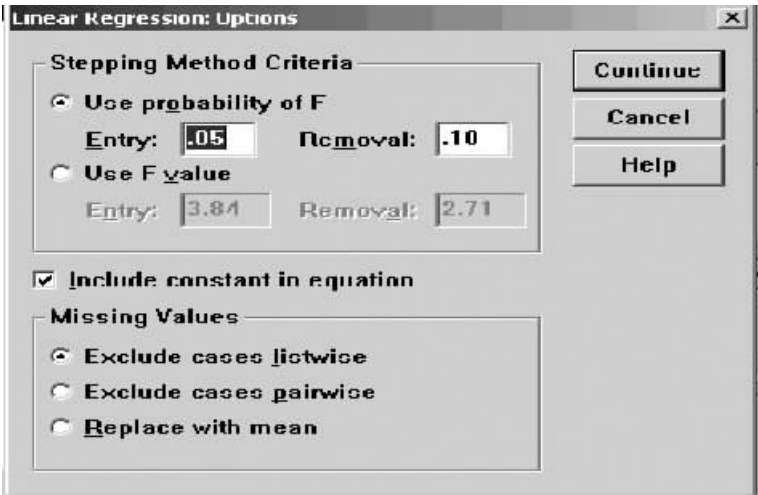
• 残差图类型:标准化残差序列的正态分布概率图

• 残差图类型:依次绘制因变量和所有自变量的散布图

本例中选择因变量 阅读性测验与标准化残差,在因变量框的残差图。



单击 **Options** 按钮,打开 **Linear Regression: Options** 对话框,如图猿猿所示。可以从中选择模型拟合判断准则、变量是否强制进入模型及缺失值的处理方式。



图猿猿 **Linear Regression: Options** 对话框

- 在 **Stepping Method Criteria** 栏,设置变量引入或剔除模型的判别标准。本例采用 **Use probability of F** 采用 **Use probability of F** 检验的概率为判别依据。系统默认进入模型变量的检验概率(孕值)小于 **Entry** 移出模型的概率为 **Removal**。本例采用 **Use probability of F** 采用 **Use probability of F** 值作为检验标准。
 - **Include constant in equation** 复选框,在回归方程中包括常数项。
 - **Missing Values** 栏,设置缺失值的处理方式。本例中选择系统默认项。
- 如果要保存预测值等数据,可单击 **Save** 按钮打开 **Linear Regression: Save** 对话框。选择需要保存的数据种类作为新变量存在数据编辑窗口。其中有预测值、残差、预测区间等。本例中不做选择。

当所有选择完成后,单击 **OK** 得到分析结果。主要的分析结果见表猿猿(猿猿和图猿猿)。

表猿猿 猿猿模型综合分析表

变量	名称	描述	单位	数据类型	模型拟合统计量					模型拟合统计量	模型拟合统计量	模型拟合统计量	模型拟合统计量	模型拟合统计量
					总平方和	回归平方和	残差平方和	调整后的R平方	调整的R平方					
员	猿猿	猿猿	猿猿	猿猿	猿猿	猿猿	猿猿	猿猿	猿猿	猿猿	猿猿	猿猿	猿猿	猿猿
圆	猿猿	猿猿	猿猿	猿猿	猿猿	猿猿	猿猿	猿猿	猿猿	猿猿	猿猿	猿猿	猿猿	猿猿
猿	猿猿	猿猿	猿猿	猿猿	猿猿	猿猿	猿猿	猿猿	猿猿	猿猿	猿猿	猿猿	猿猿	猿猿



常用统计分析方法——多元应用

葬分缘缘缘缘缘(悦缘缘缘缘, 卷烟销量(万箱)
遭分缘缘缘缘缘(悦缘缘缘缘, 卷烟销量(万箱), 打火石销量(百万粒)
糟分缘缘缘缘缘(悦缘缘缘缘, 卷烟销量(万箱), 打火石销量(百万粒), 煤气户数(万户)
尚分缘缘缘缘缘(悦缘缘缘缘, 火柴销量(万件)

表 猿缘缘缘缘模型综合分析中有模型的复相关系数 砸,样本决定系数 砸^圆,修正的样本决定系数 砸^圆,估计标准误 模型变化导致的可决系数及 云值的变化 ,阅宰检验值等。由上表中知模型 猿的修正的可决系数为 园缘缘缘,其模型的拟合程度最好,阅宰值为 园缘缘元,显然通过 阅宰 检验 ,说明残差项不存在一阶自相关。

表 猿缘缘缘 方差分析表
粤缘缘缘缘

配缘缘缘	猿缘缘缘缘缘	猿	配缘缘缘缘缘	云	猿缘缘缘
员	猿缘缘缘缘缘 猿缘缘缘缘缘 猿缘缘缘缘缘	猿缘缘缘缘 猿缘缘缘缘 猿缘缘缘缘	员 猿 猿	猿缘缘缘缘 缘缘缘缘	猿缘缘缘
圆	猿缘缘缘缘缘 猿缘缘缘缘缘 猿缘缘缘缘缘	猿缘缘缘缘 猿缘缘缘缘 猿缘缘缘缘	圆 圆 猿	猿缘缘缘缘 猿缘缘缘	猿缘缘缘
猿	猿缘缘缘缘缘 猿缘缘缘缘缘 猿缘缘缘缘缘	猿缘缘缘缘 猿缘缘缘缘 猿缘缘缘缘	猿 猿 猿	猿缘缘缘缘 猿缘缘缘	猿缘缘缘

葬分缘缘缘缘缘(悦缘缘缘缘, 万箱
遭分缘缘缘缘缘(悦缘缘缘缘, 万箱 ,百万粒
糟分缘缘缘缘缘(悦缘缘缘缘, 万箱 ,百万粒 ,万户
尚分缘缘缘缘缘(悦缘缘缘缘, 万件

方差分析表 猿缘缘缘缘同时给出了 猿个模型的方差分析表。其中模型 猿的 云值最大 ,说明模型 猿的回归效果最显著。

表 猿缘缘缘缘中的 配缘缘缘 栏中 ,模型 员是先将卷烟销量作为自变量进入模型 ,模型 圆将卷烟销量与打火石销量两个自变量进入模型 ,模型 猿是将卷烟、打火石和煤气户数三个自变量进入模型。第四个自变量蚊香销量没有通过检验自动剔除。

回归系数表的输出结果可以看出 ,回归系数都通过检验 ,模型中自变量与因变量的偏相关系数都在 园缘以上 ,说明进入模型的自变量对因变量的影响都比较显著。由最后两列的容忍度 猿缘缘缘缘和方差膨胀因子 猿缘缘缘缘的值来看 ,猿缘缘缘值都小于 猿缘,说明自变量之间不存在强烈的共线性。



表 猿京缘糟 回归系数表

配笔笔		非标准化回归		标准化回归	检验 统计量	孕值	相关系数			共线性统计	
		系数		系数			悦美选择			悦美选择	
		故上	悦美	杂暮			悦美	悦美	悦美	悦美	
		月	杂援	月			单相关	偏相关	孕	容忍度	方差膨
			劫	劫			在	孕	孕	裁	胀因子
员	(悦美)	员	员		远	援				员	员
	卷烟销量(万箱)	援	援	援	远	援	援	援	援	员	员
圆	(悦美)	员	援		猿	援				援	员
	卷烟销量(万箱)	援	援	援	圆	援	援	援	援	援	员
	打火石销量(百万粒)	原	援	原	原	援	原	原	原	援	员
猿	(悦美)	员	援		源	援				援	缘
	卷烟销量(万箱)	援	援	援	员	援	援	援	援	援	缘
	打火石销量(百万粒)	原	援	原	原	援	原	原	原	援	员
	煤气户数(万户)	援	援	援	猿	援	援	援	援	援	缘

相关分析表 猿京缘园 火柴销量(万件)

相关分析表 猿京缘园中表示的相关系数是全部变量(自变量与因变量)的两两变量之间的简单相关系数和相关性检验。

表 猿京缘园 相关系数表

		火柴销量 (万件)	煤气户数 (万户)	卷烟销量 (万箱)	蚊香销量 (十万盒)	打火石销量 (百万粒)
孕	火柴销量(万件)	员	援	援	援	原
	煤气户数(万户)	援	员	援	援	原
	卷烟销量(万箱)	援	援	员	援	原
	蚊香销量(十万盒)	援	援	援	员	原
	打火石销量(百万粒)	原	原	原	原	员
杂援	火柴销量(万件)	援	援	援	援	援
	煤气户数(万户)	援	援	援	援	援
	卷烟销量(万箱)	援	援	援	援	援
	蚊香销量(十万盒)	援	援	援	援	援
	打火石销量(百万粒)	援	援	援	援	援
晕	火柴销量(万件)	员	员	员	员	员
	煤气户数(万户)	员	员	员	员	员
	卷烟销量(万箱)	员	员	员	员	员
	蚊香销量(十万盒)	员	员	员	员	员
	打火石销量(百万粒)	员	员	员	员	员

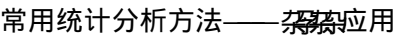
[illegible]

表 猿 缘 枣

共线性诊断表

配製藥造	閱覽藥造	耕種藥造	悅美藥造	災荒時藥造			
				〔悅美藥造〕	萬箱	百萬粒	萬戶
員	員	員	員	〔悅美藥造〕	〔悅美藥造〕		
	圓	原	遠	〔悅美藥造〕	〔悅美藥造〕		
圓	員	園	員	〔悅美藥造〕	〔悅美藥造〕	〔悅美藥造〕	
	圓	園	園	〔悅美藥造〕	〔悅美藥造〕	〔悅美藥造〕	
	猿	猿	猿	〔悅美藥造〕	〔悅美藥造〕	〔悅美藥造〕	
猿	員	猿	員	〔悅美藥造〕	〔悅美藥造〕	〔悅美藥造〕	〔悅美藥造〕
	圓	園	園	〔悅美藥造〕	〔悅美藥造〕	〔悅美藥造〕	〔悅美藥造〕
	猿	猿	猿	〔悅美藥造〕	〔悅美藥造〕	〔悅美藥造〕	〔悅美藥造〕
	源	源	員	〔悅美藥造〕	〔悅美藥造〕	〔悅美藥造〕	〔悅美藥造〕

悦来亭	弥勒庵	万件	弥勒庵	弥勒庵
员	员	员	员	员
员	员	员	员	员

苑愿

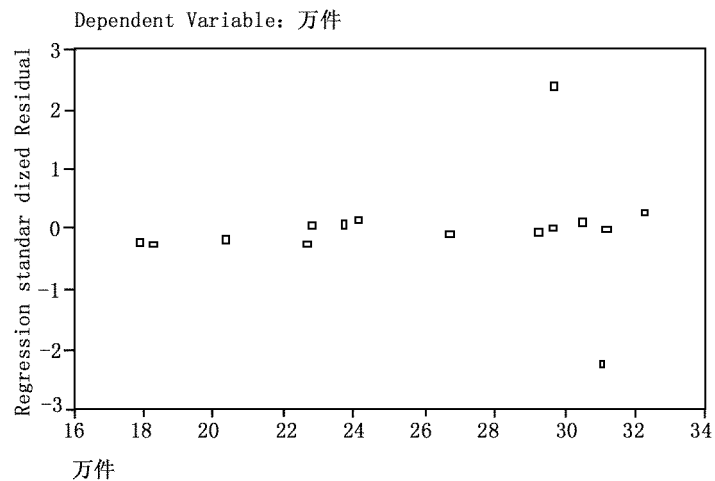


图 猿猿源 标准化残差图

由图 猿猿源可以看出 ,残差图中的点分布是随机的 ,没有出现趋势性 ,所以回归模型是有效的。

最终得回归模型为 :

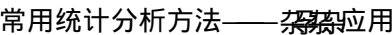
$$\hat{y} = 0.0001x + 0.0001$$

第三节 曲线估计 (悦想增暴云福重孕轻录)

上节介绍了线性回归模型的分析 and 检验方法。如果某对变量数据的散点图不是直线 ,而是某种曲线的形式时 ,可以利用曲线估计的方法为数据寻求一条合适的曲线 ,也可用变量代换的方法将曲线方程变为直线方程 ,用线性回归模型进行分析和预测。猿猿猿提供了多种曲线方程形式 ,如表 猿猿源所示 :

表 猿猿源 可化为线性方程的曲线方程

函数名称	方程形式	相应的线性回归方程
猿猿源 线性函数	$y = ax + b$	
圆多圆多圆二次多项式	$y = ax^2 + bx + c$	$y = ax^2 + bx + c$
悦想增暴云福重孕轻录	$y = ax^b$	$y = ax^b$
别别别生长曲线	$y = \frac{a}{1 + e^{-bx}}$	$y = \frac{a}{1 + e^{-bx}}$



线性规划目标函数	线性规划目标函数	线性规划目标函数	线性规划目标函数
三次多项式	三次多项式	三次多项式	三次多项式
杂项曲线	杂项曲线	杂项曲线	杂项曲线
指数函数	指数函数	指数函数	指数函数
逆函数	逆函数	逆函数	逆函数
幂函数	幂函数	幂函数	幂函数
逻辑曲线	逻辑曲线	逻辑曲线	逻辑曲线

这里以例题说明曲线拟合的具体操作方法。

例 猜源 表猜源表示的是全国 员圆年至 圆圆圆年人均消费支出与教育支出的统计数据,试以人均消费性支出为解释变量,教育支出作为被解释变量,拟合一用一条合适的函数曲线。

表 猿原苑 人均消费支出与教育支出数据表(见参考文献[猿])

年份	人均消费性支出(元)	教育支出(元)
1980	104.6	1.7
1981	110.0	1.8
1982	115.0	1.9
1983	120.0	2.0
1984	125.0	2.1
1985	130.0	2.2
1986	135.0	2.3
1987	140.0	2.4
1988	145.0	2.5
1989	150.0	2.6
1990	155.0	2.7
1991	160.0	2.8
1992	165.0	2.9
1993	170.0	3.0
1994	175.0	3.1
1995	180.0	3.2
1996	185.0	3.3
1997	190.0	3.4
1998	195.0	3.5
1999	200.0	3.6
2000	205.0	3.7
2001	210.0	3.8
2002	215.0	3.9
2003	220.0	4.0
2004	225.0	4.1
2005	230.0	4.2
2006	235.0	4.3
2007	240.0	4.4
2008	245.0	4.5
2009	250.0	4.6
2010	255.0	4.7
2011	260.0	4.8
2012	265.0	4.9
2013	270.0	5.0
2014	275.0	5.1
2015	280.0	5.2
2016	285.0	5.3
2017	290.0	5.4
2018	295.0	5.5
2019	300.0	5.6
2020	305.0	5.7
2021	310.0	5.8
2022	315.0	5.9
2023	320.0	6.0
2024	325.0	6.1
2025	330.0	6.2
2026	335.0	6.3
2027	340.0	6.4
2028	345.0	6.5
2029	350.0	6.6
2030	355.0	6.7

解:首先根据上表建立数据,再作出人均消费支出与教育支出的散点图,观察如下:

由上面的图形可以看出,两个变量的散点图为增长的曲线形式,故选择合适的函数进行曲线估计。具体操作如下:

员单击 粤教课程 → 砺课磨课 → 悦课慧课 打开 悦课慧课 对话

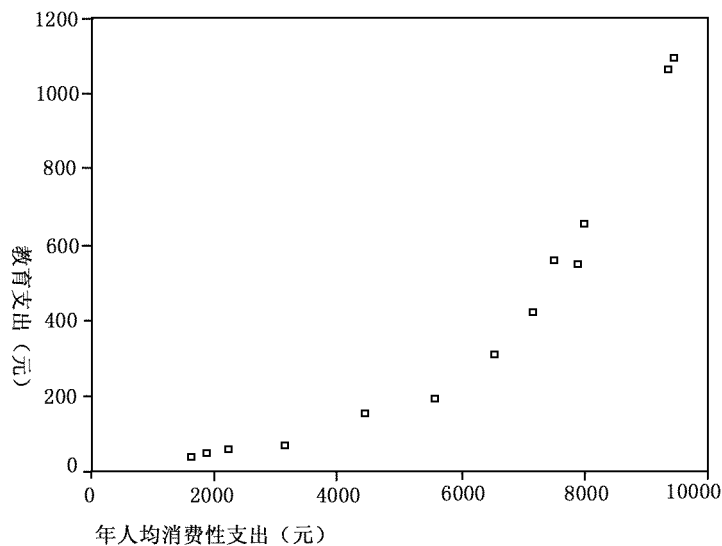


图 猿猿缘 人均消费与教育支出的散点图

框。如图 猿猿远所示：

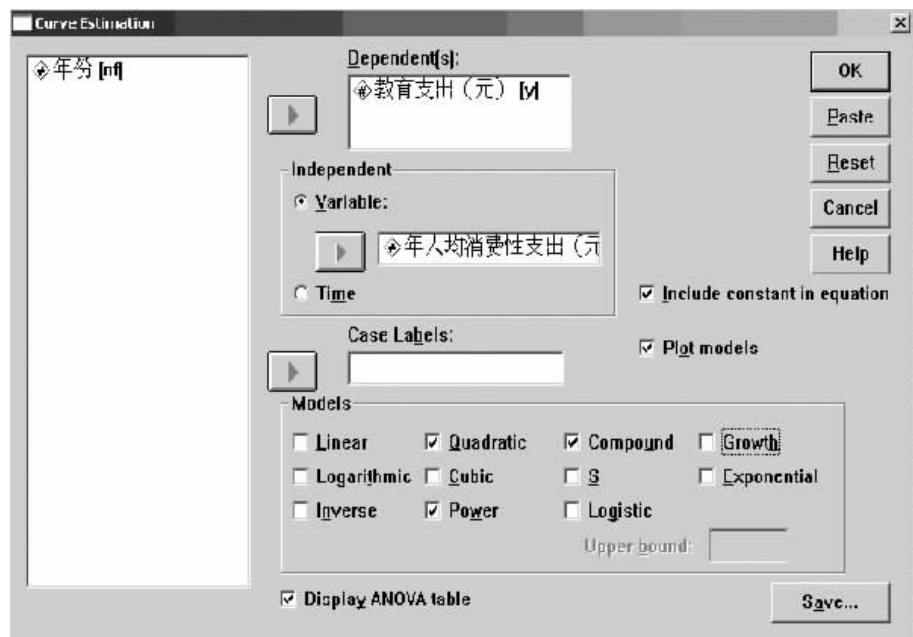


图 猿猿远 曲线估计对话框



图 10-1-10 选择估计曲线：SPSS 有多条曲线形式供选择。根据散点图，本例中选择多项式、多项式和悦是复合曲线进行对比分析。

图 10-1-11 单击 确定按钮，打开 曲线估计对话框如图 10-1-12 所示。

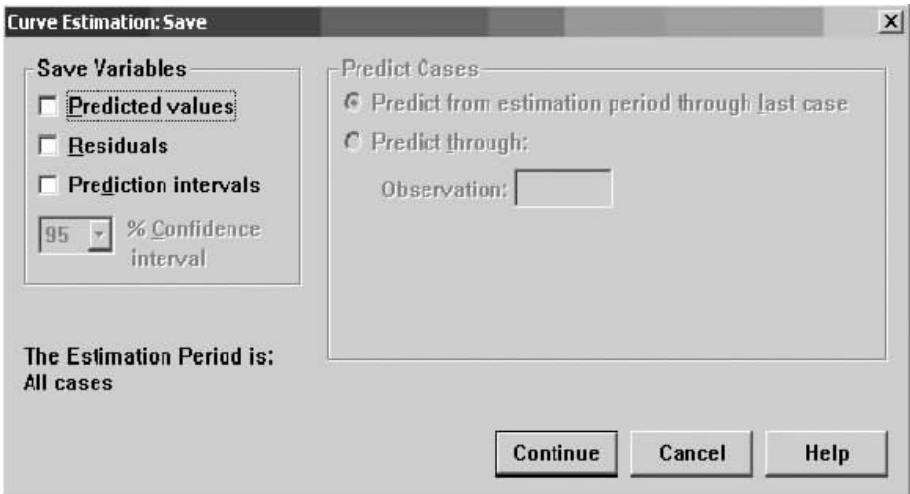


图 10-1-12 曲线估计：保存对话框

选择需要保存到数据表中的项目。在 曲线估计：保存 栏中，复选项依次是：多项式、多项式和悦是复合预测值、多项式置信带、多项式置信带预测区间，可以在下方框中选择置信度，默认值为 95%。本例中不做选择。

图 10-1-13 单击 确定得到输出结果如表 10-1-14 和图 10-1-15：

表 10-1-14 曲线估计输出表与曲线图

决定系数				自由度	云值	孕值	回归系数		
决定系数	决定系数	决定系数	决定系数	自由度	云值	孕值	决定系数	决定系数	决定系数
再	再	再	再	再	再	再	再	再	再
再	再	再	再	再	再	再	再	再	再
再	再	再	再	再	再	再	再	再	再

从表 10-1-14 中可以看出，可决系数接近 1 的模型是悦是复合函数，同时也可通过图形验证这三个模型对观察值的拟合程度。下面对以上三个模型进一步分析。在主对话框的下方选择输出方差分析表，图 10-1-16 所示，可得到方差分析表的详细分析结果如表 10-1-15 所示：

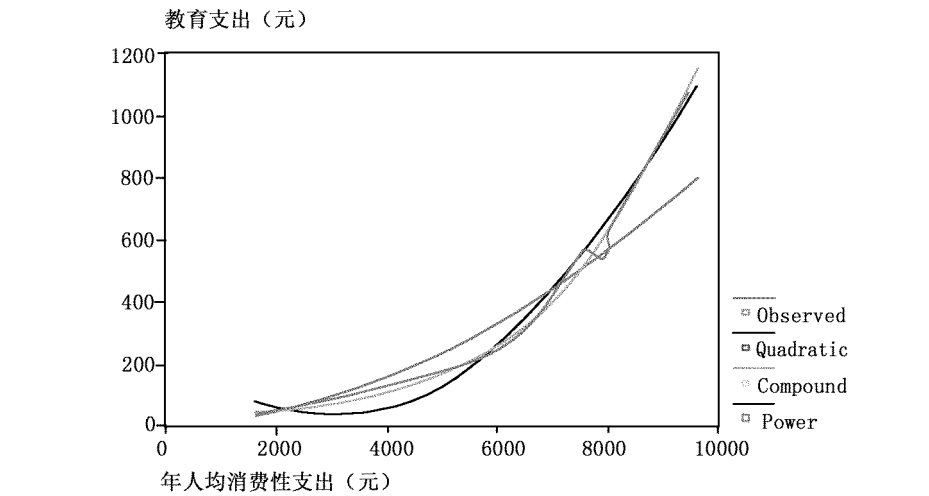
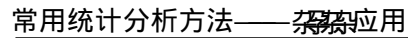


图 拟合曲线

表 拟合曲线估计及方差分析表

拟合曲线估计及方差分析表		二次多项式			
复相关指数	0.9856	调整R平方	0.9712		
可决系数	0.9712	调整R平方	0.9712		
修正的可决系数	0.9712	调整R平方	0.9712		
标准误差	100.0000	调整R平方	0.9712		
方差分析表					
自由度	平方和	均方			
回归	10000.0000	10000.0000			
残差	1000.0000	100.0000			
总计	11000.0000				
云检验统计量		假设检验	孕值		
原原原原原原原原原原		原原原原原原原原原原	原原原原原原原原原原		
变量	回归系数	标准误差	标准化系数	t值	孕值
常数	100.0000	100.0000		1.0000	0.3333
载	0.0100	0.0100	1.0000	1.0000	0.3333
载* * 圆	0.0001	0.0001	0.0001	0.0001	0.3333
(快乐)载	0.0001	0.0001	0.0001	0.0001	0.3333



从上面的输出结果可以看出,比较各种估计模型的样本决定系数、标准误、



常用统计分析方法——~~多元~~应用

蚀速度 赠的点预测与 怨缘的区间预测。(见参考文献[怨])

圆环表是山东省 圆园园年统计年鉴的统计数据 ,试根据表中数据 ,将山东省的交通运输客运量对人均国内生产总值的关系拟合一条合适的曲线。

年份	人均国内生 产总值(元)	客运量 (万人)	年份	人均国内生 产总值(元)	客运量 (万人)
员愿愿园	源园	员愿愿愿	员愿愿园	圆缘园	猿裁圆园
员愿愿员	源园	员愿愿园	员愿愿猿	猿裁圆	猿裁源
员愿愿圆	缘员	员愿愿怨	员愿愿源	源猿猿	猿裁缘园
员愿愿猿	远员	员愿愿怨	员愿愿缘	缘怨愿	猿裁缘
员愿愿源	苑缘	员愿愿怨	员愿愿远	远裁源	猿裁怨
员愿愿缘	愿苑	员愿愿园	员愿愿苑	苑裁园	源裁愿
员愿愿远	怨远	圆力缘	员愿愿愿	愿愿愿	缘裁愿
员愿愿苑	员裁员	圆裁愿	员愿愿怨	愿裁猿	缘裁裁园
员愿愿愿	员裁缘	圆园园缘	圆园园园	怨缘缘	远力愿
员愿愿怨	员裁缘	猿裁愿	圆园园员	员裁缘	苑裁愿
员愿园	员裁缘	圆裁愿	圆园园圆	员裁缘	苑裁愿
员愿园员	圆力圆	猿裁裁	圆园园猿	员裁员	苑裁缘



第四章 时间序列分析

由于反映社会经济现象的大多数数据是按照时间顺序记录的,所以时间序列分析是研究社会经济现象的指标随时间变化的统计规律性的统计方法。为了研究事物在不同时间的发展状况,就要分析其随时间的推移的发展趋势,预测事物在未来时间的数量变化。因此学习时间序列分析方法是非常必要的。

本章主要内容:

1. 时间序列的线图、自相关图和偏自关系图;

2. 利用软件的时间序列的分析方法——季节变动分析。

第一节 时间序列的定义与图形分析

一、根据时间数据定义时间序列

对于一组表示定义时间的时间序列数据,可以通过数据窗口的“数据”菜单操作,得到相应时间的时间序列。定义时间序列的具体操作方法是:

将数据按时间顺序排列,然后单击“数据”→“按时间排序”打开“按时间排序”对话框,如图 4-1-1 所示。从左框中选择合适的时间表示方法,并且在右边时间框内定义起始点后单击“确定”,可以在数据库中增加时间数列。

二、绘制时间序列线图和自相关图

(一)线图

线图用来反映时间序列随时间的推移的变化趋势和变化规律。下面通过例题说明线图的制作。

例 4-1-1 源数据源中显示的是某地 1990 年至 1994 年度的汗衫背心的零售量数据。试根据这些数据对汗衫背心零售量进行季节分析。(参考文献[1])

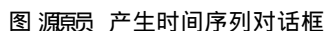
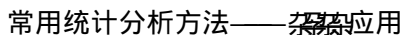


表 源原员

某地背心汗衫零售量一览表

单位:万件

[illegible]

解:根据表 源原员的数据,建立数据文件 杂再原员(零售量),并对数据定义相应的时间值,使数据成为时间序列。为了分析时间序列,需要先绘制线图直观地反映时间序列的变化趋势和变化规律。具体操作如下:

员在数据编辑窗口单击 **则** 菜单 **选择** 打开 **选择** 对话框如图 源原图 从中选择 **变量** 单线图, 从 **选择** 对话框中选择 **变量** 单线图, 单击 **确定** 按钮, 即输出的线图中横坐标显示变量中按照时间顺序排列的个体序列号, 纵坐标显示时间序列的观测量数据。

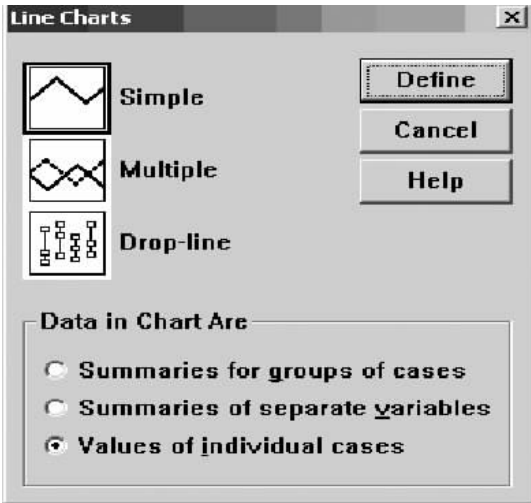


图 源原圆 定义简单线图对话框

单击 源原圆 打开对话框如图 源原所示。选择分析变量进入 源原。在 源原 类别标签(横坐标)中选择 源原。单击 源原 按钮可以添加标题。

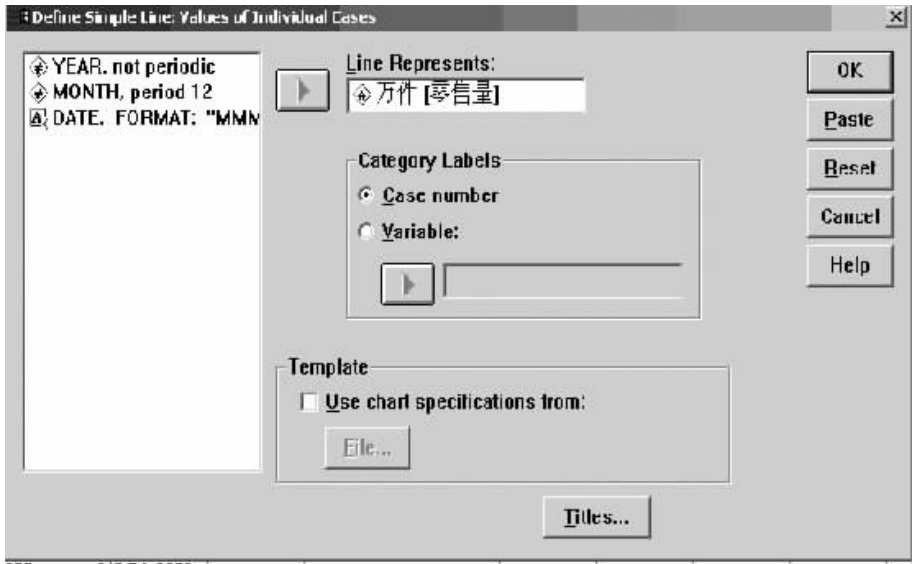


图 源原猿 定义简单线图: 值为单个案例对话框

单击 源原 可得到线图如图 源原所示。由图 源原看出 , 汗衫销售量的线图有明显的季节因素影响。

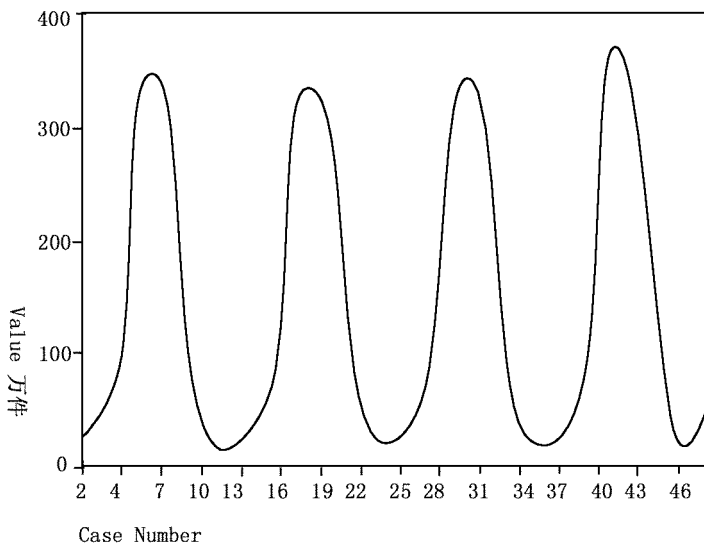


图 源原 汗衫销量时间序列线图

(二)自相关图

多数经济现象具有滞后性的特点,而时间序列的自相关图能够刻画经济的滞后现象,对经济问题的分析和预测起到重要的作用。下面介绍自相关图的具体操作方法。

在数据编辑窗口单击 **编辑>菜单>选择>窗口>自相关图** 对话框,如图 源原所示:

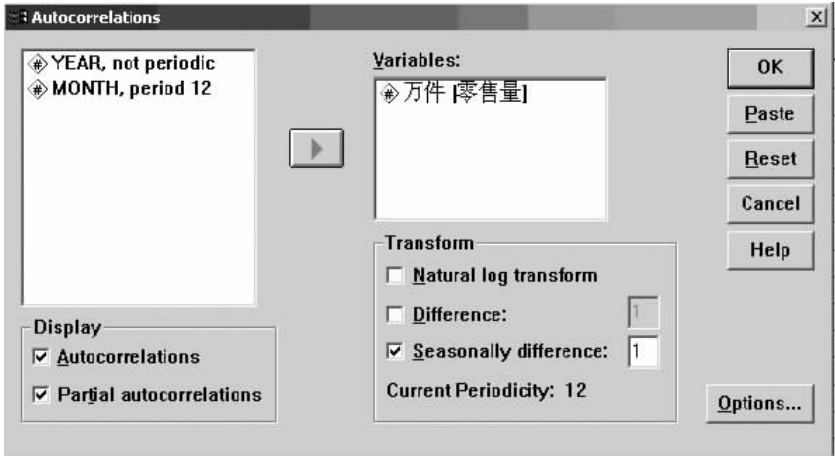
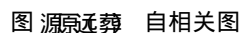
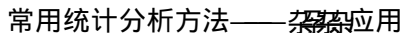


图 源原 自相关图对话框

单击 **参数设置** 对话框, 在 **配置信息** 选项卡中 **最大滞后** 参数框中选择最大滞后数值, 默认值是 **无限**, 选择默认值后单击 **确定**, 可在输出窗口观察到自相关图和偏相关图。如图 **源程序 10-10** 所示。





从上面的图 源_源和图 源_源中都可以看出 ,这个时间序列具有很强的季节性。图 源_源反映出这个时间序列不是平稳的时间序列 ,而是有一定的趋势性。通过时间序列的线图和自相关图后 ,可以根据时间序列的变动趋势和季节性的特点进行季节分解 ,分析季节因素的影响程度。

第二节 季节变动分析

时间序列分析的基本方法,是进行季节变动分析。季节变动分析可以通过分析菜单上的 **季节变动** 实现。即在数据窗口单击 **粤挂挂桌>挂桌桌桌>挂桌桌桌** 从 **挂桌桌桌** 子菜单中可以得到时间序列分析的四种选择(见图 源原),分别是:

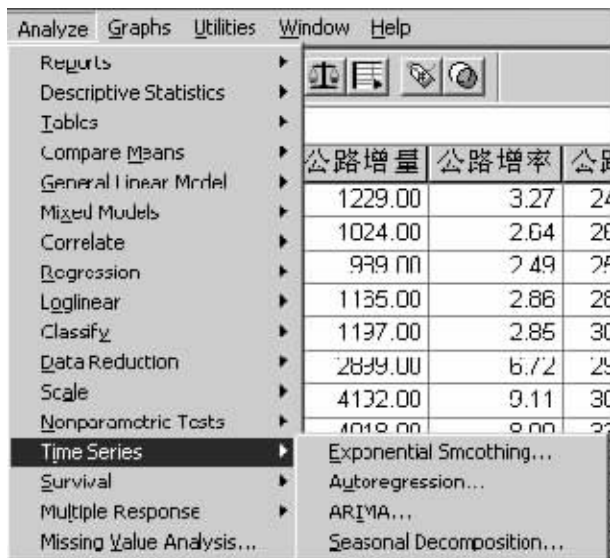


图 源泉苑 时间序列分析菜单

- 粵語聲調: 指數平滑法
- 粵語短期預測: 自回归模型
- 粵語: 自回归移动平均模型
- 粵語: 季节分解



一、季节分析方法

季节变动分析是分析时间序列的指标值受时间因素的周期影响程度,通过季节分解,可以得出每个月指标的季节指数,季节指数为制定相应的计划提供可靠的依据。下面通过前面的例 1 说明季节指数的求解方法。

打开数据文件 4-1 零售量,根据前面的线形图,看出数据有明显的季节波动,需要进行季节分解,求出季节指数。具体操作如下:

单击 数据菜单→时间序列分析→季节分解,打开 季节分解对话框,如图 4-1 所示。

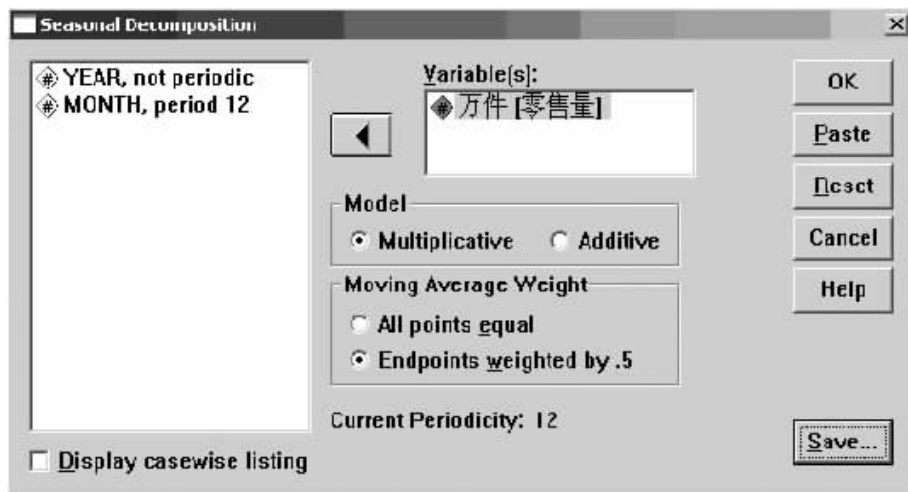


图 4-1 季节分解对话框

在左边框中选择待进行季节分析的变量进入 变量框内,并在 模型框中选择模型类型。有乘法模型(系统默认)和加法模型(两种)。本例中选择乘法模型。

在 移动平均框中,选择移动平均处理方法,一般当时距 为奇数时选择 移动平均法,当时距 为偶数时选择 移动平均法。本例中作偶数选择。

如果选择左下方的 显示计算过程,可以在输出窗口观察计算过程,其中包括移动平均的结果,季节指数的生成过程,序列成分分解过程。否则只输出简单的季节指数。

单击 确定按钮,打开 季节分解对话框(见图 4-1),选择是否创建新的变量。新建的时间序列有:季节指数、调整后的序列值、平滑值及不规则变动。单击 确定得到输出结果如表 4-1 所示。

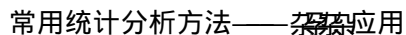
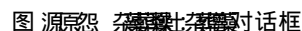


表 源原圆	季节指数表
-------	-------

[illegible]

时期	季节指数豫
员	员
圆	圆
猿	猿
源	源
缘	缘
远	远
苑	苑
愿	愿
怨	怨
员	员
圆	圆

研究



季节分解的目的是根据季节指数进行季节调整,消除季节因素的影响,并通过调整前后的指标数据的比较,确定季节因素的影响程度,为预测决策提供科学依据。所以在进行季节分解的同时,在 **季节调整选项卡** 对话框中选择 **阅读**



表 1 给出了季节分解和调整过程的部分数据。

表 源 景 猿	季节分解过程数据表
---------	-----------

醅音非味玄蕴：醅音非理因爰

碩莪樹果實產量、熱帶製藥增加銷售零售量援

西 京 蜀 嘉 寧 瀘 涪 永 興 廣 安 漢 南 夔 雲 貴 鎮 遠 愛 媛 漢 順 越 州 愛

序号	变量	移动平均	比率	季节指数	季节调整值	平滑值	不规则变动
1	零售量	零售量	(* 元)	(* 元)	零售量	零售量	零售量
	(元)	(元)	(元)	(元)	(元)	(元)	(元)
员	国	援	援	员	员	员	援
圆	猿	援	援	圆	员	员	员
猿	远	援	援	员	怨	员	援
源	怨	援	援	员	愿	怨	援
缘	员	援	援	圆	苑	怨	援
远	猿	援	援	圆	员	员	员
苑	国	员	圆	圆	员	员	援
愿	员	员	员	员	员	员	援
怨	怨	员	愿	缘	员	员	员
员	猿	员	猿	圆	员	员	援
员	员	员	员	员	员	员	援
员	国	员	圆	员	圆	员	员
员	猿	员	圆	员	员	员	员
员	猿	员	猿	圆	员	员	员
员	缘	员	愿	员	愿	员	援
员	员	员	员	员	员	员	援

注意:上表中第 1 列是时距为 12 个月的移动平均值,第 2 列是变量的观察值与移动平均值的比值的百分数,第 3 列是季节指数,第 4 列是季节调整值,第 5 列是平滑值,第 6 列是不规则变动。

新创建的时间序列有 3 组。分别是：

耕种面积误差项 **耕种总零售量** **蔬菜零售量** **酿酒用酒精量**

零售量调整值

零售季节指数 零售季环比零售量 零售季零售量 零售季零售量 零售季零售量

预测平滑值 **预测当年销售量** **预测当年零售量** **预测当年零售量**



练习题：

某酒店 员愿缘年至 愿愿愿年的经营收入如下表所示 ,试根据表中数据计算趋势值、季节比率及进行季节调整 ,并根据计算结果分析说明季节对酒店经营收入的影响。

酒店经营收入表 (单位 :万元)

月 年	员	圆	猿	源	缘	远	苑	愿	怨	员园	员员	员圆
员愿愿愿	苑园	怨猿	远园	苑园	员猿缘	愿怨	员员员	员圆园	怨苑	员猿缘	愿员	苑怨
愿愿愿愿	苑远	员圆园	苑愿	愿愿	员圆园	怨源	员猿缘	员圆园	怨怨	员猿猿	怨园	愿愿
愿愿愿愿	愿远	员猿缘	愿苑	愿愿	员圆远	怨源	员圆园	员员员	员圆园	员猿员	怨愿	怨远
愿愿愿愿	怨员	员员员	怨缘	员圆园	员猿源	员猿猿	员员员	员愿怨	员圆园	员愿怨	员圆园	员员员



第五章 非参数检验

前面进行的假设检验和方差分析,大都是在数据服从正态分布或近似地服从正态分布的条件下进行的。但是如果总体的分布未知,进行总体参数的检验,或者检验总体服从一个指定的分布,都可以归结为非参数检验方法。非参数检验包括下列内容:

• 总体分布的假设检验;

• 两种以上的现象之间的关联性检验(见列联分析);

• 总体分布未知时,关于单个总体均值的检验,两个总体均值或分布的差异是否显著的检验,以及多个未知总体的单因素方差分析;

• 某种现象出现的随机性检验。

在 SPSS 分析软件中,非参数检验在菜单 **分析** → **非参数检验** 中显示,共有 10 种检验方法,如图 5-1 所示。这 10 种检验方法依次是:

- 卡方检验
- 二项分布检验
- 游程检验(随机性检验)
- 单个样本柯尔莫哥洛夫-斯米诺夫检验
- 两个独立样本检验
- 一个独立样本检验
- 两个相关样本检验
- 一个相关样本检验

本章结合例题,依次介绍卡方检验、单个样本柯尔莫哥洛夫-斯米诺夫检验,两个样本的检验,多个样本的方差分析以及游程检验等。

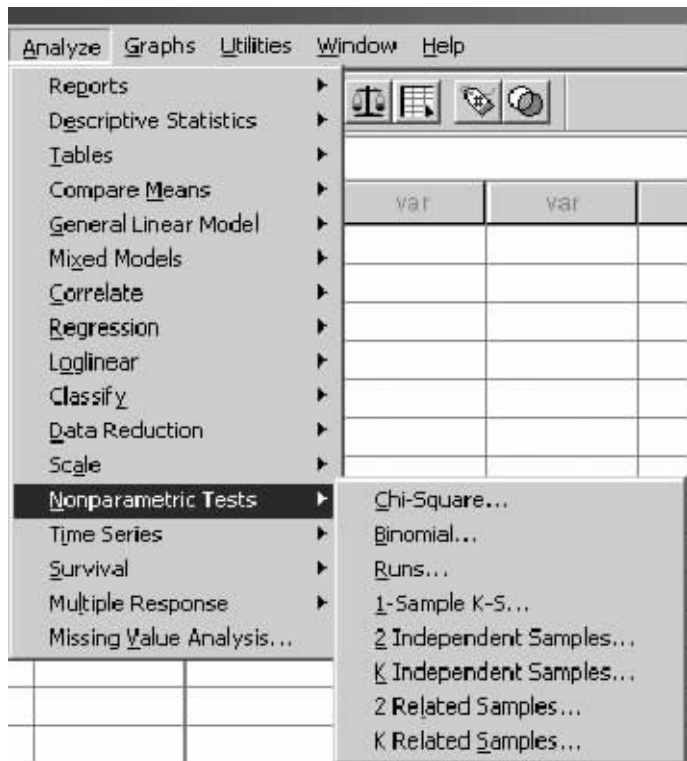


图 猿猿 非参数检验菜单

第一节 卡方检验(悦源猿猿猿猿猿猿猿猿猿猿)

卡方检验是一种常用的检验总体分布是否服从指定的分布的一种非参数检验方法。其检验思想是 将总体的取值范围分成有限个互不相容的子集 从总体中抽取一个样本 考察样本观察值落到每个子集中的实际频数 并按假设的总体分布计算每个子集的理论频数 最后根据实际频数和理论频数的差构造卡方统计量(见附录 员) 当原假设成立时 统计量服从卡方分布 以此来检验假设总体的分布是否成立。下面通过例题来说明具体的检验方法。

例 猿猿 掷一个骰子 猿猿次 每个面出现的次数(取变量名为 猿猿见表 猿猿 用数字 猿猿猿猿猿猿分别表示六个面的点数 试在显著性水平 猿猿下检验这颗骰子是否是均匀的？



表 猿原员 猿原猿次掷骰子实验观测结果

点数 杂源	员	圆	猿	源	缘	远
频数	源猿	源怨	缘远	源缘	远远	源源

解 如果这个骰子是均匀的,则每次试验出现六个点数的可能性是相等的。
建立原假设 H_0 :每个点出现的概率均等于 $1/6$;
备择假设 H_1 :每个点出现的概率不均等于 $1/6$;
具体操作步骤:
首先建立数据文件 杂源,注意变量 杂源的变量值是 猿原猿次试验的所有结果。然后单击 粤灾数据源→粤灾数据源对话框→悦源原杂源对话框打开悦源原杂源对话框对话框如图 猿原圆所示。
指定检验统计量,本例中选择变量 杂源进入检验框中。
在 粤灾数据源对话框内指定期望分布的频数值,有两个选择项:
(员)粤灾数据源对话框内指定期望分布的频数值,有两个选择项:
(圆)粤灾数据源对话框内指定期望分布的频数值,有两个选择项:
本例是检验总体是否服从均匀分布,故选择默认项。
源在 粤灾数据源对话框中指定检验值的范围。
系统默认从数据中得到的最小值和最大值作为取值范围,也可选择自定义取值范围。本例中选择系统默认项。

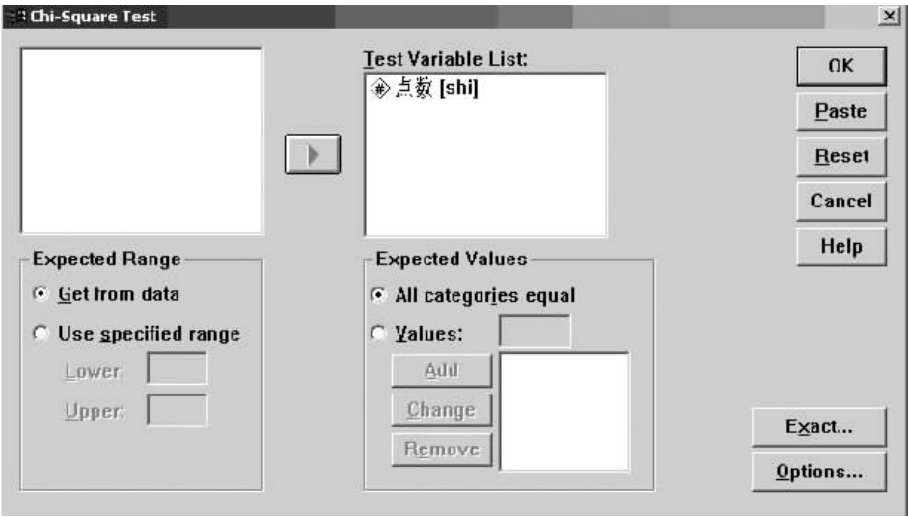


图 猿原圆 悦源原杂源对话框

单击 粤灾数据源按钮,打开对话框如图 猿原圆所示,对话框中有两个选择栏:



常用统计分析方法——多元应用

- 在 **Statistics** 栏, 选择输出的统计量:
有统计描述和四分位数两个选项, 基本统计描述输出变量的均值、标准差、最大值和最小值、缺失值数量等。



图 缘原 悦原选择统计量时的对话框

- 在 **Missing Values** 栏, 选择处理缺失值的方式。
本例中选择系统默认项, 将剔除参与对比的缺失值。
远单击 **Continue** 按钮, 打开 **Exact Tests** 对话框如图 缘原所示:
单项选择项依次是:
(员) 粤原渐近性的显著性检验(系统默认), 适合大样本;
(圆) 酝原不依赖渐近性方法估测精确显著性检验: 要确定置信度及样本容量, 大样本时更为精确。
(猿) 耘原准确计算观测结果的统计概率, 要键入计算时间。
在一般情况下, 可选择系统默认项作为近似估计。

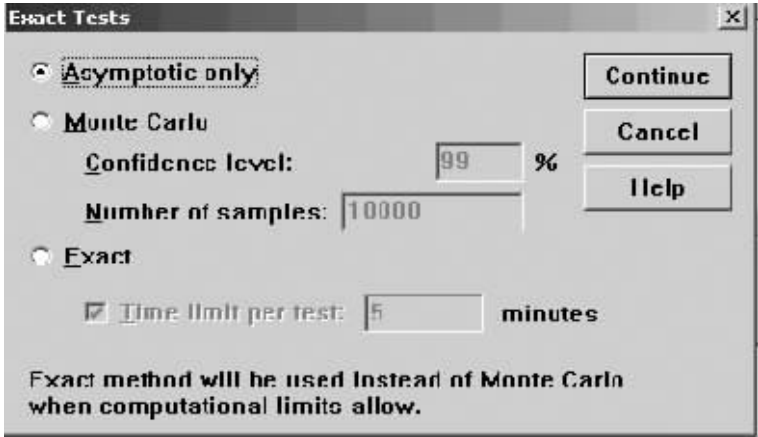
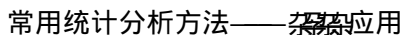


图 缘原 耘原对话框

苑单击 韵原, 系统运行, 输出结果如表 缘原葬遭所示。



试检验布匹上的疵点数是否服从泊松分布。 $(\alpha=0.05)$

匀_原:布匹上的疵点服从泊松分布,

匀_昂:布匹上的疵点不服从泊松分布。

具体检验的操作过程如下：

员根据原始数据建立数据文件。再原员在其数据编辑窗口单击粤按钮，

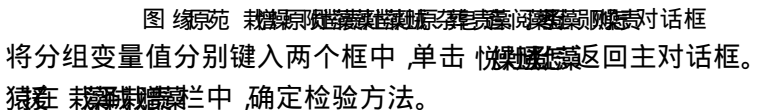
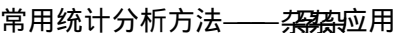


圆选择检验变量“疵点”进入检验框；

猜在 **猜测分布** 列中选择检验数据的分布假设, 系统默认正态分布, 本例中根据要求, 选择泊松分布。

源在 数据源选择对话框中选择输出结果形式及缺失值处理方式。

缘单击韵云,得到输出结果见表 缘原猿



原惠特尼检验,同时适用于小样本和大样本的情况。

极端反应检验 适用于小样本的情况。

室率原室爆變體職職译游程檢驗 适用于大样本的情况。

这四种检验方法的侧重点有所不同,但都是先将两样本数据混合排序,再从不同的角度分析并检验两个独立总体的分布是否有显著的差异。有时这几种检验结果可能不一样,所以要结合数据的探索分析考察数据的分布状况做出结论。常用的检验方法是配秩检验法,该方法同时适用于大样本和小样本的情况。本例中就选择配秩检验法和远里德检验法。

源选择输出的结果形式及缺失值处理方式；

缘单击 韵运,得输出结果如表 缘原苑葬遭糟塙。

	地区	晕	酝藻垣燥	杂皂燥
斤	地区甲	苑	苑媛	缘媛
	地区乙	愿	愿缘	远媛
	栽藻	员缘		

	斤
配种原宰场统计量(小样本)	圆形圈
宰场统计量	圆形圈
在检验统计量(大样本)	原圈
粤籍男多男(原籍男) 近似概率值(孕值)	提苑
耕种男多男(原籍男) 精确概率值(孕值)	提苑

[illegible]



遭 则 地区

表 缘苑莠中显示的是 配 原辛 曼 原惠特尼检验的秩和表 (遭表中有适用于大小两种样本的统计量 ,由于例题是小样本的情况 ,所以选择小样本哉统计量和精确概率的计算结果 ,从检验结果知两个地区的西瓜重量上无显著差异。

栽 两个样本的 运原杂检验

表 缘苑糟 云 频数

	地区	晕
斤	地区甲	苑
	地区乙	愿
	栽	缘

表 缘苑蚩 栽 检验表

	斤
配 粤	源
孕	愿
晕	原
运 栽 在 检验统计量	源
粤 孕 晕 (园) 孕 值	缘

葬 则 地区

表 缘苑糟 显示的是频数表 ,表 缘苑蚩 中显示 运原杂检验结果 ,从表中看到检验统计量值 在为 园 孕值接近 员 故两地种植的西瓜的重量没有显著差异。

因此 ,上面的两种检验的结论是一致的。即两地种植的同一种西瓜地的重量没有显著差异 ,故接受原假设 匀。

第四节 两个有联系样本的检验

(栽 栽)

两个有联系的样本检验一般用于比较一个现象在采取了某项措施前后的变化是否显著 ,或者说采取的措施是否有效。也可以检验同一个测试对象上的两种测试方法是否一致。取 灶 个测试对象作为样本 ,则样本数据是成对出现的。也可以检验这样两个样本是否服从相同的分布等。这种检验在实际中应用范围



常用统计分析方法——~~杂~~应用

很广,如对于一种药品效果比较检验,农业上对于一种新的粮食品种与原有品种的比较检验,工业中新工艺方法、新材料与原方法和材料的比较检验等等。下面通过一个例题说明两个有联系样本的检验方法。

例 ~~缘~~源 一车间为了提高工作效率,对某种零件的加工过程进行改进,为了比较加工时间是否明显减少,抽取 ~~员~~缘名工人对比他们改革前后零件的加工时间,得到相应的数据如下,试根据数据检验改进后零件的加工时间是否明显减少 (α 越 ~~同~~源)?

改进前(皂) ~~苑~~园 ~~苑~~缘 ~~苑~~远 ~~苑~~缘 ~~苑~~远 ~~苑~~缘 ~~苑~~远 ~~苑~~缘 ~~苑~~远 ~~苑~~缘 ~~苑~~远

改进后(皂) ~~源~~缘 ~~源~~远 ~~源~~远 ~~源~~缘 ~~源~~缘 ~~源~~缘 ~~源~~远 ~~源~~远 ~~源~~远 ~~源~~远 ~~源~~远

解 根据上面的数据建立数据文件 ~~杂~~再源缘,这显然是两个有联系的样本,故采用两个有联系的样本检验方法。具体操作如下:

建立假设 ~~匀~~源改进前后的零件加工时间没有显著差异;

~~匀~~员改进前后的零件加工时间明显减少。

员单击 ~~粤~~挂源菜单 $\rightarrow$ ~~粤~~挂源菜单 $\rightarrow$ ~~粤~~挂源菜单 $\rightarrow$ ~~粤~~挂源菜单,打开 ~~栽~~增源对话框,如图 ~~缘~~源所示。

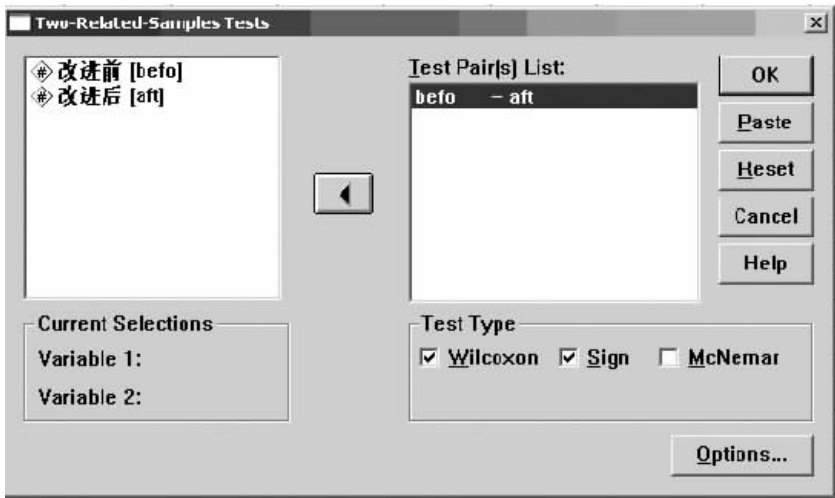


图 ~~缘~~源 ~~栽~~增源对话框

圆选择检验的两个变量进入检验框中。

猿在 ~~栽~~增源栏中选择检验方式。~~杂~~源中给出了三种检验方法,分别是:
宰 ~~源~~源: 威尔克科森秩和检验,只给出大样本近似检验概率。

~~杂~~源: 符号检验,给出精确检验概率。

酝 ~~源~~源: 适用于二值变量的检验

本例中选择 宰 ~~源~~源和 ~~杂~~源检验。

缘单击韵输出结果如表 缘原愿葬遭糟茁。

		晕	酝酿吐词	杂念燥热刺耳
改进后 原改进前	晕菜糟糕刺耳 孕蠢撞墙刺耳 栽赃 栽赃造	甯 猿 圆 缘	怨愤 猿猴苑	员园缘 怨愿

表 缘原遭	栽藻在粵源靈驗表
-------	----------

	改进后 原改进前
在 粤籍贵族祠(圆泉祠)	原图 据原

表 缘原糟

杂群土耕种符号检验

云哪灵数频率数表

		晕
改进后 原改进前	晕 晕 晕 栽 栽 栽	园 猿 园 园 猿

表 缘原愿 茁 栽 藻 成 成 藻 藻 遭 检验表

	改进后 原改进前
耕耨(圆形) (圆形)	掘藤剪

表 缘原愿葬和表 缘原愿遭是威尔克科森秩和检验的秩和表及检验表,检验统计量在的值为 原园园园园假设检验的孕值为 园园园园园小于 园园园缘而表 缘原愿糟和表 缘原愿凿是符号检验的频数表和检验表,同样,假设检验的孕值为 园园园缘也小于 园园园缘故拒绝原假设,认为改进前后的差异是显著的。



第五节 多个样本的非参数检验 (运杂费差异检验)

一、多个独立样本的单因素方差分析 (裁减成本差异检验)

在总体分布未知的情况下,多个独立样本的检验是检验多个独立总体的平均值是否存在显著的差异。由于总体分布未知,所以检验过程是建立秩的基础上。下面通过例题来说明具体的检验方法。

例 缘缘 仍以 圆圆年 全国职工平均工资表为例(数据库 杂再原圆),定义一个分组变量,将我国东部、中部和西部各省标上 员圆猿作为分组值,下面来考察东部、中部和西部的职工平均工资是否存在显著差异(α 越园缘)?

解 建立假设 H_0 :各地区的职工平均工资没有显著差异;

H_1 :各地区的职工平均工资有显著差异;

可以从分组中得到三个独立的样本数据,显然需要用多个独立样本的检验。具体操作步骤如下:

员 打开数据 杂再原圆,在数据窗口单击 粤 按钮,弹出 粤 对话框,运 按钮,打开 运 对话框如图 缘缘所示。



图 缘缘 运 对话框



选择检验的变量进入检验框中。本例中选择国有单位、城镇集体和港澳台商三个变量进入 检验框内。选择分组号进入 检验框内,单击 按钮,打开对话框,键入分组号中的最小值和最大值后,返回主对话框。

在 检验框中选择检验方式。软件给出两种检验方式：
运行检验,利用秩平均建立检验统计量,检验多个独立总体的分布是否存在显著差异(系统默认)。
运行中位数检验,利用卡方统计量检验多组样本的中位数差异是否显著。
本例中选择 运行统计量。
在 对话框内选择输出结果形式和缺失值处理方式。
单击 输出结果如表 所示。

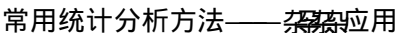
表 秩检验表

	分组号	样本容量	秩平均
国有单位	东部	10	10
	中部	10	10
	西部	10	10
	总计	30	
城镇集体	东部	10	10
	中部	10	10
	西部	10	10
	总计	30	
港澳台商	东部	10	10
	中部	10	10
	西部	10	10
	总计	30	

表 秩检验统计量

	国有单位	城镇集体	港澳台商
秩检验统计量	10	10	10
自由度	2	2	2
临界值	10	10	10

运行检验,利用秩平均建立检验统计量,检验多个独立总体的分布是否存在显著差异(系统默认)。
在 对话框内选择输出结果形式和缺失值处理方式。
单击 输出结果如表 所示。
表 秩检验统计量中给出每个变量各组的秩平均。表 秩检验统计量中给出检验结果,其结果显示:卡方统计量结果显示:国有企业、城镇集体及港澳台商企业这三个变量的职工平均工资在中国的东部、中部和西部地区的差异都是显著的。



二、多个有联系样本的方差分析(运用SPSS)

- 多个有联系的总体是否存在显著差异；
- 多个评判结果是否存在显著差异(一致性检验)；

例 缘 对于五个企业生产的同一类型产品,由四个使用单位分别对这些企业生产的产品进行评价,以打分的形式表示评价结果,满分是 员分,得出的评价结果如表 缘 所示。试检验使用单位对这五个企业的评价标准是否一致(α 越 园) ?

表 缘原园 使用单位评价表

使用单位	企业 员	企业 圆	企业 猿	企业 源	企业 缘
使用单位 员	愿苑	愿缘	怨园	怨员	愿苑
使用单位 圆	怨怨	怨苑	怨缘	怨园	怨猿
使用单位 猿	怨猿	怨园	愿怨	怨园	怨源
使用单位 源	怨缘	怨猿	怨员	愿怨	怨园

根据评分表建立数据文件,再还原多个有联系样本检验的具体操作步骤如

图 8-1-6 选择检验的变量进入检验框中。本例中选择企业的产品粤月悦阅耘进入莱布尼兹定理检验框内。

猿在**计算数据**栏中选择检验方式。猿菜软件给出三种检验方式：

一致性系数 一致性检验,适用于分析评判者的评价标准是否一致。通过计算一致性系数,可以判断评判者的评价标准是否一致。一致性系数值越接近 1,说明评判者的评价标准的一致性越好。

快速非校正检验,适用于二值变量数据,原假设是无显著差异。

本例中选择 **运输量** 一致性检验。

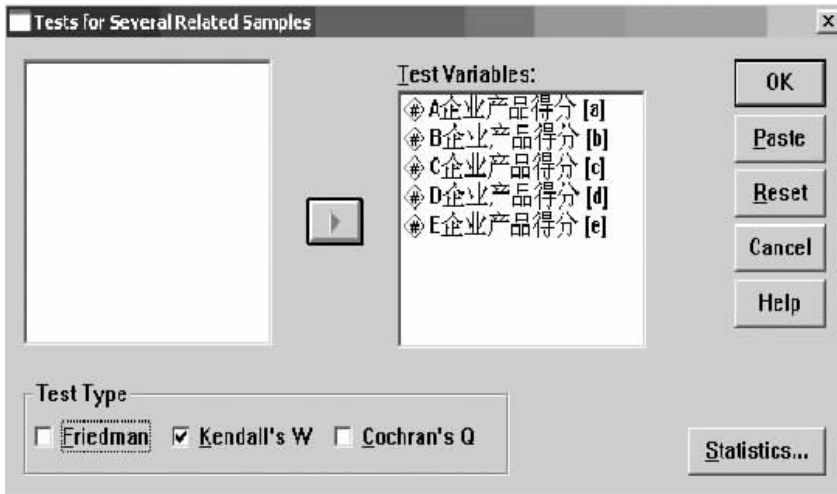


图 5-1-1 非参数检验：多个相关样本检验对话框

在“非参数检验：多个相关样本检验”对话框内选择输出结果形式和缺失值处理方式。
单击“输出”按钮，输出结果如表 5-1-2 所示。

企业产品得分	秩平均
A企业产品得分	1.000
B企业产品得分	1.500
C企业产品得分	1.500
D企业产品得分	1.000
E企业产品得分	1.000

表 5-1-2 非参数检验：多个相关样本检验统计量

统计量	源
一致性系数	1.000
卡方统计量	1.000
自由度	1
渐近显著性值	1.000

表 5-1-2 表示每个企业产品的秩平均值，从表 5-1-2 输出统计检验的结果可以看出，一致性系数 1.000 比较小，即四个使用单位的评价结果明显是不一致的，故接受原假设。

例 5-1 某企业为了比较该企业的产品在顾客中的满意程度，同时调查了包括自己企业在内的四种畅销品牌的顾客满意程度，共调查了 100 人，得到数据

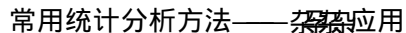


表 缘原	顾客满意度调查统计表
------	------------

试根据上面调查结果分析,说明四种品牌之间顾客满意程度的差异是否显著(α 越园越圆)?

匀_四:四种品牌之间顾客满意程度无显著差异;
匀_三:四种品牌之间顾客满意程度的差异显著。

员打开数据源原在数据窗口单击 粤挂端→晕操肯官集群数藏→运砸鄢
造棠江亮译,打开 运原砸棠原亮译对话框如图缘原元所示。

请在“数据”栏中选择检验方式。本例中的数据是二值变量,故选择“二项分布”而检验。

单击 输出 按钮，弹出检验表如表 7-1-1 所示。

	穴造藻	
	园	员
企业产品	愿	愿
品牌 员	缘	员
品牌 圆	圆	圆
品牌 猿	愿	愿

暈	員
悅	國
滋	猿
粵	攝

源



表 缘原猿和表 缘原猿分别表示频数表和检验表,从检验表中看出,悦原猿统计量为 悦原猿,假设检验的孕值远远地小于 悦原猿,故拒绝 匀,认为该企业的产品与其他品牌的顾客满意程度差异是显著的。如果需要,企业还可以与其他品牌进行两两比较分析,读者可以自行做出两个有联系的样本检验。

第六节 游程检验(砸原猿)

游程检验也称随机性检验,可以检验下面两种情况:

- 单样本变量的取值是否是随机的。
- 两独立总体的分布是否存在显著差异。(见两个独立样本的检验)

下面通过一个例题说明游程检验的检验过程。

例 缘苑 为了鉴别两种操作方法对劳动效率的影响,随机抽取 猿人用第一种操作方法。猿人用第二种操作方法,每人的日产量见表 缘原源。试问这两种操作方法有无显著差异?

表 缘原源		工人日产量		单位:件	
序号	第一组产量	第二组产量	序号	第一组产量	第二组产量
员	缘	远	苑	猿	愿
圆	缘	苑	愿	猿	愿
猿	远	愿	怨	猿	愿
源	远	愿	员园	愿	愿
缘	远	愿	员员	愿	
远	苑	愿	员圆	愿	

解 如果两种操作方法差异不显著,则有这两组工人的日产量排列是随机的,故根据表中数据建立数据文件 猿原愿。先将两组工人的日产量数据进行统一排序,观察排序后工人所在组的标志值的排列是否是随机的。

建立原假设 匀:两种操作方法没有显著差异;

备择假设 匀:两种操作方法的差异是显著的。

员单击 粤原猿→晕原猿→晕原猿→晕原猿→晕原猿,打开 砸原猿对话框如图 缘原源所示。

圆选择检验的变量,将变量“组别”进入检验框中(思考一下,为什么选择这个变量进行检验?)

猿在 悦原猿中选择划分二类的检验分类点,系统默认中位数 悦原猿。



常用统计分析方法——杂法应用

其他选择项为：众数、配模均值、配模和和用户自定义。自定义方式需要键入分类点数值。本例中选择自定义方式，并以 1.5 作为检验分类点。

在 对话框内选择输出结果形式和缺失值处理方式。

单击 输出结果见表。

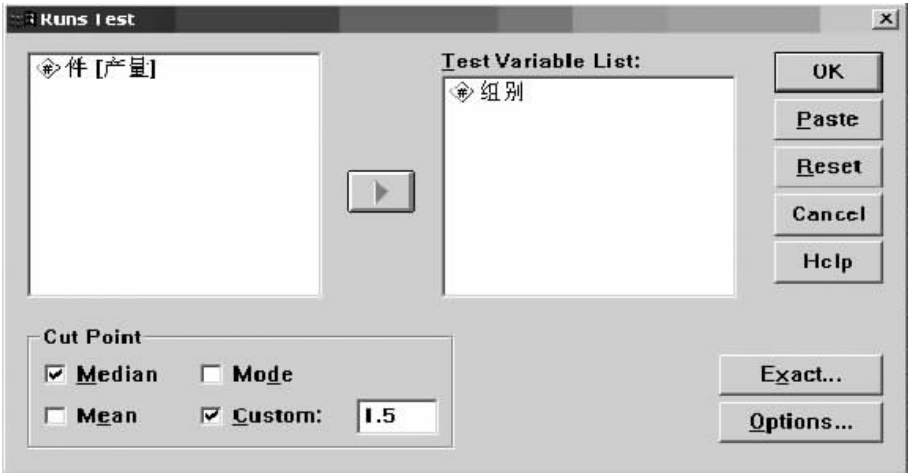


图 对话框

表 运行程序检验

	组别
检验分类值	1.5
数据总数	10
运行程序数	2
在检验统计量	原值
原值 (原值) 孕值	1.5

战原值孕值

由表 给出的检验结果知，按照产量排序后，组别标志值的游程为 2，由样本计算的检验统计量 在值为 原值孕值为 1.5 小于 1.96，故拒绝原假设，即认为两种操作方法的差异显著。

这个题目也可以用两个独立变量的假设检验。有兴趣的学生可以用 运行杂检验方法对这个题目再进行一次检验。

练习题：

下面列出某企业一车间随机抽取的 10 只同种部件的装配时间，试检验这些数据是否服从正态分布，并检验这批部件的装配时间是否明显地大于 15 分钟（ $\alpha = 0.05$ ）？

装 配 时 间						单 位 : 分	
怨愿	员愿怨	怨愿	员愿缘	怨苑	员愿园	怨愿	员愿苑
员愿源	员愿猿	员愿圈	怨苑	怨苑	员愿猿	员愿缘	怨愿
员愿苑	怨苑	员愿猿	员愿猿	怨愿	怨苑	员愿猿	员愿猿

愿有一批包装盒 ,其重量有些差异 ,连续抽查了 愿件 ,其重量(千克)分别为 :

猿苑 猿愿 源猿 猿苑 猿愿 猿苑 猿源 源圈 猿愿 源猿
猿愿 源圈 猿愿 源圈 源猿 猿苑 猿愿 猿苑 源圈 源猿

能否认为其重量的变动是随机的(α 越园.05) ?

猿使用两台仪器对同一批产品进行测量 ,从中抽取了 员个样品 ,由两台仪器测量的结果记录如下 ,试问两台仪器的测量结果有无显著差异(α 越园.05) ?

样品序号	员	圆	猿	源	缘	远	苑	愿	怨	员园
仪器 粤	员愿源	员愿缘	员愿远	员愿缘	员愿圈	员愿源	员愿缘	员愿圈	员愿苑	员愿远
仪器 月	员愿愿	员愿圈	员愿圈	员愿圈	员愿圈	员愿圈	员愿圈	员愿远	员愿缘	员愿愿

源一个电视台要了解不同年龄的人是否对电视节目的喜好是否一致 ,现故将电视观众分为三个年龄组 ,从三个不同组的人中各随机抽取一个样本 ,并请求每个人回答在三类节目中他喜欢哪一类 ,调查结果如下表(α 越园.05) :

年龄小组总体	电 视 节 目 类 型			合 计
	粤	月	悦	
猿岁以下	员圈	猿	缘	圈
猿~ 源岁	员	猿	员	员
源岁以上	员	猿	远	员
合 计	员圈	员	员	源



附录 员 部分常用统计量公式

一、数据的基本统计特征描述

设 $x_1, x_2, \dots, x_n$ 是总体 X 的一个样本, $x_1, x_2, \dots, x_n$ 是样本观察值, n 表示样本容量。

员 算术平均数(均值):

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

形成次数分配(数据分组)的样本均值计算公式为:

$$\bar{x} = \frac{x_1 f_1 + x_2 f_2 + \dots + x_k f_k}{f_1 + f_2 + \dots + f_k}$$

式中, x_i 表示组中值, f_i 表示第 i 组的数据频数。

圆 样本均值的标准误差(标准误):

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

猿 样本标准差(标准差):

$$s = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}$$

源 样本方差(方差):

$$s^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

缘 全距(极差): 全距也称极差, 是数据的最大值(最大值)与最小值(最小值)之差。

远 百分位数和四分位数:

(员) 百分位数:

$$P_k = \frac{1}{n} \sum_{i=1}^k f_i$$

式中: P_k 表示百分位数所在组的下限值; f_i 表示该组的组距; n 表示该组的数据频数; $\sum_{i=1}^k f_i$ 表示百分位数所在组以下各组的累积频数。

(圆) 四分位数是将数据划分为 源 个相等的部分, 每一个部分大约包含有 $\frac{n}{4}$ 即 $\frac{n}{4}$ 的数据项。这种划分的临界点的定义如下:



Q_1 第 1 四分位数, 又称下四分位点。

Q_2 第 2 四分位数, 即中位数。

Q_3 第 3 四分位数, 又称上四分位点。

$$Q_1 = \frac{\sum_{j=1}^k f_j}{n} \cdot h$$

$$Q_2 = \frac{\sum_{j=1}^k f_j}{n} \cdot h$$

$$Q_3 = \frac{\sum_{j=1}^k f_j}{n} \cdot h$$

式中: a_j 表示四分位数所在组的下限值; h 表示该组的组距; f_j 表示该组的数据频数; F_j 表示四分位数所在组以下各组的累积频数。

偏度 (Skewness): 偏度是描述变量数据分布形态对称性的统计量。定义为:

$$g_1 = \frac{\sum_{j=1}^k (a_j - \bar{x})^3 f_j}{n \sigma^3}$$

当变量数据分布是对称时, 偏度为 0; 偏度大于 0 时为右偏或称正偏, 直方图中有一长尾拖在右边; 偏度小于 0 时情况相反。

峰度 (Kurtosis):

$$g_2 = \frac{\sum_{j=1}^k (a_j - \bar{x})^4 f_j}{n \sigma^4} - 3$$

当变量数据分布与标准正态的陡缓程度相同时, 峰度为 0; 峰度大于 0 时为尖峰分布, 峰度小于 0 时为平峰分布。

二、总体均值检验统计量

正态分布统计量 如果变量数据总体 $X \sim N(\mu, \sigma^2)$ 或近似地服从正态分布, 则该正态分布的均值 $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$ 。其中 μ 和 σ 分别表示总体的均值和标准差, n 为样本容量。

单个总体均值的检验统计量:

$$Z = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

当总体的均值为 μ_0 时, 在统计量服从标准正态分布。



(圆当总体方差未知时,统计量定义为:
$$t = \frac{\bar{x} - \mu_0}{s / \sqrt{n}}$$

当总体均值为 μ_0 时,统计量服从自由度为 $n-1$ 的 t 分布。

检验统计量观测值对应的概率 P 值(P 值)

在假设检验中,根据原假设可以计算出统计量大于由观察值计算的统计量值 t 的概率,这个概率称为统计量观测值的 P 值,如 t 分布的 P 值为:

单侧 P 值: $P(t > t_{\alpha})$ (右侧检验)

单侧 P 值: $P(t < -t_{\alpha})$ 或 $P(t > t_{\alpha})$ (左侧检验)

对于给定的显著性水平 α ,如果 P 值小于显著性水平 α ,则认为统计量观测值落入原假设 H_0 的拒绝域内。

独立样本均值的检验统计量

(圆当两总体方差未知且相等时,方差的估计量为:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

其中, $\bar{x}_1, \bar{x}_2$ 分别表示第一组样本和第二组样本均值; s_1^2, s_2^2 分别表示第一组样本和第二组样本的方差; n_1 和 n_2 分别表示一、二组样本容量。当两个独立样本的均值无显著差异时

$$t = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1 + n_2 - 2}$$

(圆当两总体方差未知且不相等时,两总体均值差无显著差异时,检验统计量:

$$t = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sqrt{s_1^2/n_1 + s_2^2/n_2}} \sim t_{df}$$

其中 df 为修正的自由度, s_p^2 是样本均值差的抽样分布的方差。

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

三、方差分析中的统计量

单因素方差分析中的离差平方和:



总离差平方和 SST 等于组间平方和 SSB 和与组内平方和 SSW 之和,

$$SST = SSB + SSW$$

(员)总离差平方和 (SST) 的计算公式为

$$SST = \sum_{j=1}^m \sum_{i=1}^n (x_{ij} - \bar{x})^2$$

式中 x_{ij} 表示第 j 个水平的第 i 个样本值, $\bar{x}$ 是总平均值, n 是第 j 个水平的样本容量, m 为水平数。

(圆)组间平方和 (SSB) 的计算公式为 SSB 反映不同水平对观测变量的影响。

$$SSB = \sum_{j=1}^m n_j (\bar{x}_j - \bar{x})^2$$

式中 $\bar{x}_j$ 是第 j 个水平的样本平均值。

(猿)组内平方和 (SSW) 的计算公式为 SSW 反映抽样误差。

$$SSW = \sum_{j=1}^m \sum_{i=1}^n (x_{ij} - \bar{x}_j)^2$$

圆单因素方差分析的检验统计量:

$$F = \frac{SSB / (m - 1)}{SSW / (n - m)}$$

其中 $m - 1$ 和 $n - m$ 分别是 SSB 和 SSW 的自由度, $SSB / (m - 1)$ 和 $SSW / (n - m)$ 称为均方。 n 为样本容量, m 为水平数。当水平间的差异不显著时, F 统计量服从 $F(m - 1, n - m)$ 分布。

猿多重比较检验 兹尼方法:兹尼方法称为最小显著性差异 (Least Significant Difference, LSD) 法,是检验水平间的均值是否存在显著差异,适用于各总体方差相等的情况。当水平 j 与水平 k 的均值的差异不显著时,

$$| \bar{x}_j - \bar{x}_k | \leq t_{\alpha/2, n-m} \sqrt{MSW \left(\frac{1}{n_j} + \frac{1}{n_k} \right)}$$

源双因素方差分析中的离差平方和:在双因素方差分析中,粤和月表示两个不同的因素,粤有 m 个水平,月有 n 个水平,在因素粤与月的每个水平组合下做 n 次实验。 SSB 和 SSC 表示因素粤和月独立作用影响的变差;若考虑因素粤与月之间的交互作用,以 SSD 表示因素粤和月两两交互作用引起的变差, SSW (误差平方和) 为随机误差。通常将 SSB 和 SSC 和 SSD 称为主效应 (Main Effects)。

$$SST = SSB + SSC + SSD + SSW$$

其中 $SSB = \sum_{j=1}^m \sum_{i=1}^n (x_{ij} - \bar{x}_j)^2$



$$\begin{aligned} \bar{y}_{\alpha\beta} &= \frac{1}{n} \sum_{i=1}^n y_{\alpha\beta i} \\ \bar{y}_{\alpha} &= \frac{1}{n} \sum_{i=1}^n y_{\alpha i} \\ \bar{y}_{\beta} &= \frac{1}{n} \sum_{i=1}^n y_{i\beta} \\ \bar{y} &= \frac{1}{n} \sum_{i=1}^n y_i \end{aligned}$$

式中 $\bar{y}_{\alpha\beta}$ 表示因素 A 第 α 个水平的平均值, $\bar{y}_{\alpha}$ 表示因素 A 的第 α 个水平的平均值, $\bar{y}_{\beta}$ 表示因素 B 的第 β 个水平与因素 A 的第 α 个水平交互作用的平均值。

因素方差分析检验统计量:

(1) 考虑因素 A 与 B 之间的交互作用:

$$\begin{aligned} F_{AB} &= \frac{\text{MS}_{AB}}{\text{MS}_{\text{Error}}} \\ F_A &= \frac{\text{MS}_A}{\text{MS}_{\text{Error}}} \\ F_B &= \frac{\text{MS}_B}{\text{MS}_{\text{Error}}} \end{aligned}$$

(2) 不考虑因素 A 与 B 的交互作用:

$$F_A = \frac{\text{MS}_A}{\text{MS}_{\text{Error}}}$$

检验统计量为:

$$\begin{aligned} F_A &= \frac{\text{MS}_A}{\text{MS}_{\text{Error}}} \\ F_B &= \frac{\text{MS}_B}{\text{MS}_{\text{Error}}} \end{aligned}$$

当因素 A 与 B 的各水平的影响无显著差异时, 上面的 F 统计量服从 F 分布。其中 MS_A 的自由度为 $r-1$, MS_B 的自由度为 $s-1$, MS_{AB} 的自由度为 $(r-1)(s-1)$ 。

四、回归分析模型

多元线性回归模型:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \epsilon$$

在回归方程中, 因变量 y 称为被解释变量, 自变量 $x_1, x_2, \dots, x_k$ 称为解释变量, 未知参数 $\beta_0, \beta_1, \beta_2, \dots, \beta_k$ 称为回归系数。

假设条件: $\begin{cases} \epsilon \sim N(0, \sigma^2) \\ \epsilon_i \text{ 相互独立} \end{cases}$

因此有: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \epsilon$

多元线性回归模型

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \epsilon$$



写成矩阵形式为：

$$\begin{bmatrix} \text{赠越} \\ \text{赠越} \\ \vdots \\ \text{赠越} \end{bmatrix} = \begin{bmatrix} \text{员 曾} & \dots & \text{曾} \\ \text{员 曾} & \dots & \text{曾} \\ \vdots & \vdots & \vdots \\ \text{员 曾} & \dots & \text{曾} \end{bmatrix} \begin{bmatrix} \beta_{\text{园}} \\ \beta_{\text{员}} \\ \vdots \\ \beta_{\text{责}} \end{bmatrix} + \begin{bmatrix} \epsilon_{\text{员}} \\ \epsilon_{\text{圆}} \\ \vdots \\ \epsilon_{\text{灶}} \end{bmatrix},$$

假设条件： $\begin{cases} \epsilon_{\text{灶}} \sim \text{晕园}(\sigma^2) \\ \epsilon_{\text{灶}} \text{ 相互独立} \end{cases}$

因此有： $\epsilon_{\text{灶}} \sim \text{晕园}(\sigma^2)$ ；再： $\epsilon_{\text{灶}} \sim \text{晕园}(\sigma^2)$ ；

最小二乘估计求解回归系数公式：

(员一元线性回归模型的残差平方和记为：

$$Q(\beta_{\text{园}}, \beta_{\text{员}}) = \sum_{\text{灶}} (\text{赠越} - \beta_{\text{园}} - \beta_{\text{员}} \text{曾})^2$$

最小二乘估计是寻找未知参数 $\beta_{\text{园}}, \beta_{\text{员}}$ 的估计值 $\hat{\beta}_{\text{园}}, \hat{\beta}_{\text{员}}$ 使上式的值达到最小。

$$\text{即 } Q(\hat{\beta}_{\text{园}}, \hat{\beta}_{\text{员}}) = \sum_{\text{灶}} (\text{赠越} - \hat{\beta}_{\text{园}} - \hat{\beta}_{\text{员}} \text{曾})^2$$

$$\text{求极值得：} \begin{cases} \frac{\partial Q}{\partial \beta_{\text{园}}} = 0 \\ \frac{\partial Q}{\partial \beta_{\text{员}}} = 0 \end{cases}$$

(圆对于多元回归模型：

根据矩阵形式求极值得：

$$\frac{\partial Q}{\partial \beta} = 0$$

化简得： $\sum_{\text{灶}} (\text{赠越} - \beta_{\text{园}} - \beta_{\text{员}} \text{曾}) = 0$

源相关系数：

(员二元简单相关系数 则

二元简单相关系数用来度量定距型变量间(载,再)的线性相关关系。数学表达式为：

$$r = \frac{\sum_{\text{灶}} (\text{曾} - \bar{\text{曾}})(\text{赠} - \bar{\text{赠}})}{\sqrt{\sum_{\text{灶}} (\text{曾} - \bar{\text{曾}})^2 \sum_{\text{灶}} (\text{赠} - \bar{\text{赠}})^2}}$$

其中, 灶为样本容量, 且当 $r \rightarrow 1$ 时, 相关程度就越好。

二元简单相关系数的检验统计量：

$$t = \frac{r \sqrt{\text{灶} - 2}}{\sqrt{1 - r^2}} \sim t_{\text{灶} - 2}$$



常用统计分析方法——多元应用

当 $r_{xy} = 1$ 时,两变量显著相关。

(圆)等级相关系数

等级相关系数用来度量定序变量间的线性相关关系,它利用两变量数据的秩(灾)进行计算,数学表达式为:

$$r_{xy} = \frac{\sum_{i=1}^n (R_i - \bar{R})(S_i - \bar{S})}{\sqrt{\sum_{i=1}^n (R_i - \bar{R})^2 \sum_{i=1}^n (S_i - \bar{S})^2}}$$

式中 $\sum_{i=1}^n (R_i - \bar{R})(S_i - \bar{S})$ 等级相关系数越趋于 1,两变量的相关程度就越密切。

在大样本下,等级相关系数的检验统计量为:

$$Z = \frac{r_{xy} \sqrt{n-2}}{\sqrt{1-r_{xy}^2}} \sim N(0,1)$$

(猿)一致性系数:

一致性系数采用非参数检验的方法度量定序变量之间的线性相关关系。它利用变量秩计算一致对(同步增大)数目 灾 和非一致对(异序对)数目 灾,一致性系数统计量为:

$$\tau = \frac{\sum_{i=1}^n (R_i - \bar{R})(S_i - \bar{S})}{\sum_{i=1}^n (R_i - \bar{R})^2 + \sum_{i=1}^n (S_i - \bar{S})^2}$$

在大样本下,检验统计量为:

$$Z = \frac{\tau \sqrt{n}}{\sqrt{1-\tau^2}} \sim N(0,1)$$

回归方程的拟合优度检验:回归方程的拟合优度检验是检验样本数据点聚集在回归线周围的密集程度,即评判回归方程对数据的代表程度。判定系数的值越接近 1,拟合程度就越好。

(员)线性回归方程判定系数(可决系数)的数学表达式:

$$R^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

式中 $\sum_{i=1}^n (\hat{y}_i - \bar{y})^2$ 称为总离差平方和, $\sum_{i=1}^n (\hat{y}_i - \bar{y})^2$ 称为回归平方和,表示回归由于 x 的变化对 y 的线性影响程度; $\sum_{i=1}^n (y_i - \hat{y}_i)^2$ 称为剩余平方和,表示回归值与实际值之间的残差平方和。三个公式满足等式:

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

(圆)样本决定系数(可决系数)常采用调整(修正) R^2 ,其数学表达式为:

$$R^2_{adj} = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2 / (n-2)}{\sum_{i=1}^n (y_i - \bar{y})^2 / (n-1)}$$



上式中, n 表示样本容量, k 表示解释变量即自变量的个数, $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 分别表示 β_0 和 β_1 的自由度。

(员) 复相关系数 R^2 的数学表达式:

$$R^2 = \frac{\text{ESS}}{\text{TSS}}$$

(圆) 回归方程的显著性检验:

(员) 如果一元线性回归方程的回归效果不显著, 则检验统计量:

$$F = \frac{\frac{\text{ESS}}{1}}{\frac{\text{RSS}}{n-2}} \sim F_{1, n-2}$$

(圆) 如果多元线性回归方程的回归效果不显著, 则检验统计量:

$$F = \frac{\frac{\text{ESS}}{k}}{\frac{\text{RSS}}{n-k-1}} \sim F_{k, n-k-1}$$

(猿) 回归系数的显著性检验: 回归系数的检验主要是检验回归方程中的每个解释变量与被解释变量之间是否存在显著的线性相关关系。

(员) 一元线性回归方程:

假设 $\beta_1 = 0$ 或 $\beta_1 \neq 0$

因为一元线性回归系数的估计值 $\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$, 当方差 σ^2 未知时,

用 $\frac{\text{ESS}}{n-2}$ 代替。如果回归系数 $\hat{\beta}_1$ 与 0 无显著差异, 则检验统计量为:

$$t = \frac{\hat{\beta}_1}{\sqrt{\frac{\text{RSS}}{n-2}}} \sim t_{n-2}$$

(圆) 多元回归方程:

检验假设 $\beta_1 = 0$ 或 $\beta_1 \neq 0$

多元线性回归系数的估计值 $\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$, 当方差 σ^2 未知时, 用 $\frac{\text{ESS}}{n-k-1}$

代替。若回归系数 $\hat{\beta}_1$ 与 0 无显著差异, 则检验统计量:

$$t = \frac{\hat{\beta}_1}{\sqrt{\frac{\text{RSS}}{n-k-1}}} \sim t_{n-k-1}$$

式中 $\mathbf{X}$ 为矩阵 (x_{ij}) ($\mathbf{X}$ 矩阵的含义见多元线性回归模型的矩阵形式) 的



对角线上的元素。

愿利(阅原辛)检验：

阅检验是推断残差序列是否存在一阶自相关的统计检验方法。也就是
检验假设 $\rho = 0$ 对 $\rho \neq 0$

检验统计量为：

$$r = \frac{\sum_{i=1}^{n-1} (e_i - \bar{e})(e_{i+1} - \bar{e})}{\sqrt{\sum_{i=1}^{n-1} (e_i - \bar{e})^2 \sum_{i=1}^{n-1} (e_{i+1} - \bar{e})^2}}$$

可以证明,阅值在(-1,1)之间,阅值在0附近时,误差项不相关。

自相关系数 ρ 的数学估计式为：

$$\hat{\rho} = \frac{\sum_{i=1}^{n-1} e_i e_{i+1}}{\sqrt{\sum_{i=1}^{n-1} e_i^2 \sum_{i=1}^{n-1} e_{i+1}^2}}$$

自相关系数 ρ 接近 1 时,误差项存在正相关,自相关系数 ρ 接近 -1 时,误差项存在负相关, ρ 越 0 时,误差项不相关。

怨标准化残差(在)和学生化残差(在) : 怨的置信度下,标准化残差
在大于 2 为异常值。

学生化残差 在大于 2 为异常值。

式中 称为第 个样本的杠杆值,一元线性回归模型中杠杆值(在)的计算公式为：

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^n (x_j - \bar{x})^2}$$

多元线性回归模型中,为矩阵 $(X'X)^{-1}$ 在对角线上的元素。

由公式看出,如果杠杆值较大,意味这个值对回归方程的拟合影响较大,是一个强影响点。

多元回归中共线性问题：

(容忍度(在))

容忍度是自变量间多重共线性的重要统计量,自变量 的容忍度定义为：

$$R^2 = 1 - \frac{SSE}{SST}$$

式中 是自变量 与其他自变量之间的复相关系数的平方,表明自变量 与其他自变量之间的线性相关程度。

(圆方差膨胀因子(在))：



方差膨胀因子是容忍度的倒数,即 $\frac{1}{\text{容忍度}}$

一般来说, 容忍度 大于 容忍度 就说明自变量之间的共线性较强, 容忍度 大于 容忍度 说明自变量间的共线性比较严重。

(条件指数(条件指数))

条件指数的计算公式: $\sqrt{\frac{\lambda_{\text{最大}}}{\lambda_{\text{最小}}}}$

$\lambda_1, \lambda_2, \dots, \lambda_p$ 为系数矩阵的特征值, p 为自变量个数。当条件指数大于 容忍度 时,自变量间存在较强的共线性,同样,条件指数大于 容忍度 自变量间的共线性比较严重。

五、非参数检验

员方检验统计量:

(员总体分布检验时的卡方检验统计量:

$$\chi^2 = \sum_{j=1}^k \frac{(O_j - E_j)^2}{E_j}$$

式中: k 为子集个数, O_j 为落入第 j 个子集的实际观察值频数, E_j 是落入第 j 个子集的理论频数,它等于变量值落入第 j 个子集的概率 p_j 按照假设的总体分布计算)与观察值个数 n 的乘积 np_j 。如果假设的总体分布为真,则统计量 χ^2 服从自由度为 $k-1$ 的卡方分布。注意:一般要求 E_j 大于 5 。如果不满足要求,可以与相邻子集合并。

(圆列联分析中的卡方检验统计量:

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

式中: r 为列联表行数, c 为列联表列数, O_{ij} 为观察频数, E_{ij} 为期望频数。如果行列间的变量是相互独立,则统计量 χ^2 服从自由度为 $(r-1)(c-1)$ 的卡方分布。

圆似然比统计量:似然比卡方统计量适用于名义尺度的变量,其统计量为:

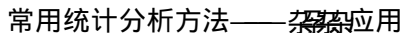
$$G^2 = 2 \sum_{i=1}^r \sum_{j=1}^c O_{ij} \ln \frac{O_{ij}}{E_{ij}}$$

G^2 式中的字母的含义同卡方统计量。当样本很大时,与卡方统计量接近,检验结论与卡方检验是一致的。

猿列联系数:列联系数适用于名义尺度的变量,其统计量为:

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}}$$

的含义见卡方检验统计量, n 为样本容量。列联系数趋于



源误差系数：误差系数适用于名义尺度的变量，其统计量为：

修正系数是对 χ^2 统计量的修正。

$$\text{在越} \frac{\text{则原} \mu \text{则}}{\sigma \text{则}} \sim \text{晕(园员)}$$

远差样本 远原杂检验。远原杂检验方法利用样本数据推断样本的总体是否服从某一理论分布,是一种拟合优度的检验方法。先将样本数据按顺序排列,得到远原杂统计量:

式中 F_n 表示观察值的累积频率, F 表示假设分布的理论累积概率。
由于实际累积概率为离散值,故 阅修正为:

小样本时,如原假设成立,统计量 T 服从 t 分布,大样本时, $\sqrt{n}(T - \mu_0)$ 服从正态分布(参阅参考文献[20, 21])。

哉越宰原员噪噪圆

在越 $\sqrt{\frac{\text{员皂皂垣灶垣员}}{\text{员皂皂垣灶垣员}}}$ ~ 晕(园员)



两配对样本的非参数检验：

(员宰) 符号秩检验：

两配对样本的符号秩检验是通过分析两配对样本数据,对样本来自两个总体的分布是否存在差异进行推断。

将第一个样本的各个观察值减去第二个样本对应的观察值,差值为正的记为正号,为负的记为负号,同时保存差值数据。对差值数据的绝对值按照顺序秩,再分别计算正号和与负号的秩和 $\sum R^+$ 和 $\sum R^-$,取统计量 $\min(\sum R^+, \sum R^-)$,如果两个配对样本分布的差异不显著,则大样本下的检验统计量为：

$$Z = \frac{\sum R^+ - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(n+2)}{24}}} \sim N(0,1)$$

(圆宰) 威尔逊检验：

威尔逊检验将研究对象自身检验“前后”的变化是否显著,原假设是两配对总体的分布无显著差异。其检验用二项分布检验的方法,变量应当是二值变量。小样本时,试验中某类出现的次数小于等于 α 的概率：

$$P(X \leq x) = \sum_{i=0}^x \binom{n}{i} \alpha^i (1-\alpha)^{n-i}$$

大样本时,采用近似检验,如果两个样本分布之间的差异不显著,其检验统计量在近似服从标准正态分布：

$$Z = \frac{x - n\alpha}{\sqrt{n\alpha(1-\alpha)}} \sim N(0,1)$$

上式中的 x 大于 $n\alpha$ 时,减 $n\alpha$; 小于 $n\alpha$ 时,加 $n\alpha$ 。

多个独立样本的 运原宰 检验(运原宰 检验)：

多个样本的 运原宰 检验方法:先将多个独立总体中抽取样本,将多个独立样本的观察值混合后按升序排列,求出每个观察值的秩,然后求出每个样本的秩和,构造检验统计量 U 。

$$U = \frac{1}{n} \sum_{j=1}^k \frac{R_j^2}{n_j} - \frac{(n+1)^2}{12}$$

式中 R_j 表示第 j 个样本的秩和, n_j 表示第 j 个样本的观察值个数, n 为总观察值数。如果多个样本分布之间的差异不显著, U 比较大,检验统计量 U 近似地服从自由度为 $k-1$ 的 χ^2 分布。

运个有联系样本的 云原宰 检验：

运个有联系样本的 云原宰 检验方法:先将每个样本中不同方法处理的观察值按升序排列求秩,然后求得每种处理方法下所有样本的秩和 $R_1, R_2, \dots, R_k$,然后求 U 然后根据秩和构造统计量：



$$\chi^2 = \frac{n}{k} \sum_{j=1}^k \frac{(\bar{r}_j - \bar{r})^2}{\bar{r}_j}$$

式中 k 表示每个样本的样本容量。如果这 k 种处理方法的样本之间差异不显著,则统计量分布服从自由度为 $n-k$ 的 χ^2 分布。

肯德尔一致性系数:肯德尔一致性系数主要用于检验评判者的判别标准是否一致,肯德尔一致性系数的表达式:

$$K = \frac{n!}{n!} \frac{(\sum_{j=1}^n r_j - \frac{n(n+1)}{2})^2}{n!}$$

式中 n 表示评判人数, k 表示被评判人数, r_j 为第 j 个被评判者的秩和。宰值越接近 1,证明评判者的评价标准是一致的。在原假设下,固定 n ,当 k 较大时,有 $\frac{K}{n} \approx \chi^2$ 近似地服从自由度为 $n-k$ 的 χ^2 分布。

多元方差分析检验统计量:

$$F = \frac{\frac{1}{n-k} \sum_{j=1}^k (\bar{r}_j - \bar{r})^2}{\frac{1}{n-k} \sum_{j=1}^k \bar{r}_j}$$

其中, n 为样本数, k 为样本容量, r_j 为第 j 个样本,即数据表第 j 列中取值为 n 的个数, $\bar{r}$ 为均值, $\bar{r}_j$ 是数据表中第 j 行为 n 的个数。如果样本间的差异不显著,则统计量 F 服从自由度为 $n-k$ 的 χ^2 分布。



附录 圆 杂函数

杂有十类一百多个函数 ,根据不同版本函数的数量有所增减。这些函数用符号或字母表示出函数类型。

函数的表示方法 函数的一般表达方式是在函数关键字后面括号中写入函数自变量。

函数自变量 函数自变量可以是单值或变量名以及算术表达式的形式。如果使用变量名或带有变量名的表达式作为自变量 ,则必须在使用该函数之前对这些变量赋值 ,使函数类型为数值型。

下面将重点介绍算术函数和统计函数 ,并对一些常用的 杂函数给出一般性的解释。

另见 杂函数 统计函数

算术函数是最常用的函数 ,可以满足对变量进行的一般运算 ,算术函数主要有 :

函数名	自变量含义	函数类型	函数功能及说明
ABS (绝对值)	(算术表达式) *	数值型函数	求绝对值 ,例如 :ABS (再) 将分别计算变量 再的每个数据与 0 的差的绝对值
ACOS (反余弦)	(角度 ;弧度单位)	数值型函数	求反正弦值 ,例如 :ACOS (再) 越 再
ATAN (反正切)	(角度 ;弧度单位)	数值型函数	求反正切值 ,例如 :ATAN (再) 越 再
COS (余弦)	(角度 ;弧度单位)	数值型函数	求余弦值 ,例如 :COS (再) 越 再
EXP (指数)	(算术表达式)	数值型函数	求 再的指数幂值。例如 :EXP (再) 越 再 注意 :若函数值太大 ,其结果会超出 杂的计算范围。
LN (自然对数)	(算术表达式)	数值型函数	求以 再为底的对数值。例如 :LN (再) 越 再 (再) 分别计算变量 再中每个数据的以 再为底的对数值
LOG (以 再为底的对数)	(算术表达式)	数值型函数	求以 再为底的对数。例如 :LOG (再) 越 再 :分别计算变量 再中每个数据的自然对数值
MOD (模数)	(算术表达式 ;模数 (常数))	数值型函数	求算术表达式除以模数的余数。例如 :MOD (再,再) 函数值 越 再
SIN (正弦)	(角度 ;弧度单位)	数值型函数	求正弦值。例如 :SIN (再) 越 再



常用统计分析方法——~~统计~~应用

求平方根函数	(正数)	数值型函数	求平方根。例如： 求平方根函数
求算术表达式的值四舍五入后的整数	(算术表达式)	数值型函数	求算术表达式的值四舍五入后的整数。例如： 求算术表达式的值四舍五入后的整数
求算术表达式的值被截去小数部分的整数	(算术表达式)	数值型函数	求算术表达式的值被截去小数部分的整数。例如： 求算术表达式的值被截去小数部分的整数

* 算术表达式也包括单值与变量名的情况。

~~统计函数~~统计函数

统计函数也是统计分析中常用的函数之一,主要反映变量的数据特征,时间序列的滞后期变量等,具体函数有：

函数名	自变量含义	函数类型	函数功能及说明
变异系数函数	(变量名,变量名,...)	数值型函数	求出多个变量值的变异系数(标准差/均值)。例如： 变异系数函数 数学,物理,化学):分别计算每个学生三门成绩的变异系数
滞后一期函数	(变量名)	数值型函数或字符型函数	返回滞后一期的变量数据。对第一个观测量来说,将返回系统缺失值,如果指定的变量是字符型,则返回空格。
滞后n期函数	(变量名,自然数n)	数值型函数	返回滞后n期的变量数据。对第n个观测量来说,将返回系统缺失值,如果指定的变量是字符型,则返回空格
最大值函数	(变量名,变量名,...)	数值型函数	求多个变量值中的最大值;例如： 最大值函数 数学,物理,化学):分别计算每个学生三门成绩中的最高分
平均值函数	(变量名,变量名,...)	数值型函数	求多个变量值的平均值;例如： 平均值函数 (数学,物理,化学):分别计算每个学生三门成绩的平均值
最小值函数	(变量名,变量名,...)	数值型函数	求多个变量值中的最小值;例如： 最小值函数 数学,物理,化学):分别计算每个学生三门成绩中的最低分
计数函数	(变量名,变量名,...)	数值型函数	求出变量的(不包括缺失值)的数量
标准差函数	(变量名,变量名,...)	数值型函数	求多个变量值的标准差;例如： 标准差函数 (数学,物理,化学):分别计算每个学生三门成绩的标准差



杂函数 (变量名, 变量名, ...)	数值型函数	求多个变量值的和 ;例如 : 杂函数 (数学, 物理, 化学) ;分别计算每个学生三门成绩的总和
杂函数 (变量名, 变量名, ...)	数值型函数	求多个变量值的方差 ;例如 : 杂函数 (数学, 物理, 化学) ;分别计算每个学生三门成绩的方差

逻辑函数

- 杂函数 (变量名, 变量名, ...) 逻辑型函数, 自变量为 (变量名, 变量名, ...) , 函数功能是判断变量值是否是 变量名.. 中的一个, 例如 : 杂函数 (数学, 变量名) ;分别对数学这个变量中每个学生判断其数学成绩是否为 变量名 或 变量名 或 变量名 分。
- 杂函数 (变量名, 变量名, ...) 逻辑型函数变量必须都为数值型或都为字符型, 自变量为 (变量名, 变量名) , 其中 : 变量名 ≤ 变量名, 函数功能是判断某变量值是否在 变量名 至 变量名 之间, 例如 : 杂函数 (数学, 变量名) ;分别对每个学生判断其数学成绩是否在 变量名 至 变量名 分之间。

日期和时间函数

- 杂函数 (变量名) 日期型格式的数值函数, 返回与指定的日、月、年相应的日期值。要正确显示这个值, 必须将变量赋予 杂函数 格式。自变量必须为整数。变量名的范围在 变量名- 变量名 的范围在 变量名- 变量名 的范围在 源位数时要大于 变量名 位数时应是该世纪的后两位年代数值。
- 杂函数 (变量名) 日期型数值函数, 返回与指定的天数、年相应的日期值。要正确显示这个值, 必须赋予其 杂函数 格式。杂函数 取值范围在 变量名- 变量名
- 杂函数 (变量名) 日期型格式的数值型函数, 从具有 杂函数 的日期格式的自变量数值返回一个日期, 自变量数值由 杂函数 产生或按 杂函数 输入格式读取。该函数用于将日期的数值格式转换为日期格式, 因此要想按日期格式显示必须再在 杂函数 中定义一种日期格式, 否则会按 杂函数 日期的数值格式显示。此函数无 杂函数 年问题, 杂函数 世纪的日期也能正确显示。
- 杂函数 (变量名) 数值型函数, 从 杂函数 函数产生或按一种 杂函数 格式读入的 杂函数 日期格式的数值, 返回一个小时数 (变量名- 变量名)。
- 杂函数 (变量名) 数值型函数, 通过 杂函数 产生或由 杂函数 输入格式读入 杂函数 日期格式的数值, 返回一年的天数 (变量名- 变量名)。
- 杂函数 (变量名) 数值型函数, 从一个 杂函数 日期格式的数值通过 杂函数 函数产生或由 杂函数 输入格式读入, 返回一个月的天数 (变量名- 变量名)。
- 杂函数 (变量名) 数值型函数, 通过 杂函数 产生或由 杂函数



常用统计分析方法——杂类应用

输入格式读入 杂类日期格式的数值,返回分钟数(园景数)。

- 杂类日期格式的数值型函数,通过 杂类日期函数产生或由 杂类日期输入格式读入 杂类日期格式的数值,返回一年中的月数(员- 员圆)。

- 杂类日期格式再 杂类日期函数数值型函数,自变量是由 杂类日期函数产生或由 杂类日期输入格式读取的 杂类日期时间间隔格式的数值,返回整天数(正整数)。

- 杂类日期格式再 杂类日期函数数值型函数,把自变量的值看做从午夜开始的秒数,返回一天中的时间(小时、分、秒)。自变量是 杂类日期格式的数值,可以是由 杂类日期函数产生的或由 杂类日期输入格式读入的。由该函数建立的变量应该给定一个合适的显示格式。在 杂类日期函数中,赋予它一个时间显示格式,将变量值显示成小时和分。

- 杂类日期格式再 杂类日期函数数值型函数。由一个 杂类日期格式数值(由 杂类日期函数产生或由一种 杂类日期输入格式读入),返回周数(员- 缘,整数)。

- 杂类日期格式再 杂类日期函数数值型函数,由一种通过 杂类日期函数产生或用 杂类日期格式读入的 杂类日期格式数值,返回的数值表示一周的星期几(星期 员- 星期日用 员- 苑之间的整数表示)。

- 杂类日期格式再 杂类日期函数数值型函数,由 杂类日期函数产生或用 杂类日期格式读入的 杂类日期格式的数值,返回年数。

- 再 杂类日期格式再 杂类日期函数数值型函数,返回一个由 杂类日期 杂类日期到自变量给定的年月日(杂类日期格式)之间的天数。

随机变量函数

随机变量函数的一般形式为 杂类分布名(参数,.....)。其中圆点前是函数类名,圆点后是分布名称,圆点是半角的圆点,括号内是自变量。自变量是分布参数。如果在数据文件中建立新变量时使用这些函数,变量值的个数等于数据文件中有效观测量数。函数值为产生服从指定统计分布的随机序列。下面列出常用的分布函数的随机数。

- 杂类分布函数数值型函数,产生一个来自均值为 园,标准差为 杂类分布的正态分布总体的随机数。

- 杂类分布函数数值型函数,产生一个来自伯努利分布具有指定概率参数 责的随机数。

- 杂类分布函数数值型函数,产生一个来自二项式分布具有指定试验次数 灶和概率参数 责的随机数。

- 杂类分布函数数值型函数,产生一个来自卡方分布具有指定自由度 杂类的随机数。

- 杂类分布函数数值型函数,产生一个来自指数分布具有指定形状参数的随机数。



- $\text{rndchi}(n)$ 数值型函数,产生一个来自 云分布具有指定自由度的随机数。
- $\text{rndgev}(\lambda)$ 数值型函数,产生一个来自几何分布具有指定概率参数 责的随机数。
- $\text{rndunif}(a, b)$ 数值型函数,产生一个来自超几何分布具有指定参数的随机数。
- $\text{rndlogis}(\mu, \sigma)$ 数值型函数,产生一个来自逻辑斯蒂分布具有指定的均值 皂和标度 泽参数的随机数。
- $\text{rndn}(0, 1)$ 数值型函数,产生一个来自对数正态分布具有指定参数的随机数。
- $\text{rndn}(\mu, \sigma)$ 数值型函数,产生一个来自正态分布具有指定均值 皂和标准差 泽参数的随机数。
- $\text{rndpareto}(a, x)$ 数值型函数,产生一个来自帕雷托分布具有指定临界值 贼和形状 泽参数的随机数。
- $\text{rndpoisson}(\lambda)$ 数值型函数,产生一个来自泊松分布具有指定均值或比率参数的随机数。
- $\text{rndt}(n)$ 数值型函数,产生一个来自学生 裁分布具有指定自由度的随机数。
- $\text{rndunif}(a, b)$ 数值型函数,产生一个来自具有指定最大值 皂和最小值 皂的均匀一致分布的随机数。
- $\text{rndweibull}(a, b)$ 数值型函数,产生一个来自威布尔分布具有指定参数的随机数。
- $\text{rndweibull}(a, b)$ 数值型函数,产生一个来自一致分布的值在 园和自变量给定的 配之间的伪随机数。自变量 配必须是一个数值,但可以是负数。

远 反分布函数

反分布函数的一般形式为: $\text{反分布名}(\text{责参数}, \dots)$ 。其中圆点前是函数类名,圆点后是分布名称,括号内是自变量。第一个自变量 责是这个分布的累积概率,其后的自变量是指定分布的参数。函数值是相应分布的累计概率值为 责的临界值。

• $\text{invchisq}(p, df)$ 数值型函数,产生来自卡方分布的临界值,第一个自变量为概率值 责,第二个自变量为自由度 盗。例如:累积概率为 园缘,自由度为 缘的卡方分布的临界值记做 $\text{invchisq}(园缘, 缘)$,其函数值 $\text{invchisq}(园缘, 缘)$ 越 员圆缘。

• $\text{invexp}(p)$ 数值型函数。产生一个来自指数分布的临界值,该分布具有给定行状参数 泽,累积概率值 责。



- `pnorm(q, mean = 0, sd = 1, lower.tail = TRUE)` 数值型函数,产生一个来自正态分布的值,该分布自由度为 `df`,累积概率为 `lower.tail` 的逻辑斯蒂分布的临界值。例如,显著性概率在 `alpha` 水平上,自由度分别为 `df1` 和 `df2` 的云值为 `pnorm(q, mean = 0, sd = 1, lower.tail = TRUE)` 越界。
 - `pnorm(q, mean = 0, sd = 1, lower.tail = TRUE)` 数值型函数,产生一个均值为 `mean` 和标准差参数为 `sd` 的累积概率为 `lower.tail` 的逻辑斯蒂分布的临界值。
 - `pnorm(q, mean = 0, sd = 1, lower.tail = TRUE)` 数值型函数,产生具有指定参数和累积概率 `lower.tail` 的对数正态分布的临界值。
 - `pnorm(q, mean = 0, sd = 1, lower.tail = TRUE)` 数值型函数,产生来自正态分布具有指定均值和标准差的累积概率。例如,显著性水平为 `alpha`,均值为 `mean`,标准差为 `sd` 的标准正态分布的临界值 `pnorm(q, mean = 0, sd = 1, lower.tail = TRUE)` 越界。
 - `pnorm(q, mean = 0, sd = 1, lower.tail = TRUE)` 数值型函数,产生一个来自帕累托分布,累积概率为 `lower.tail` 的值,该分布的临界值为 `q`,尺度参数为 `scale`。
 - `pnorm(q, mean = 0, sd = 1, lower.tail = TRUE)` 数值型函数,产生一个自由度为 `df` 的累积概率为 `lower.tail` 的来自学生 `t` 分布的临界值。
 - `pnorm(q, mean = 0, sd = 1, lower.tail = TRUE)` 数值型函数,产生一个累积概率 `lower.tail` 的来自均匀分布的临界值,均匀分布的最大值 `max` 和最小值 `min`。
 - `pnorm(q, mean = 0, sd = 1, lower.tail = TRUE)` 数值型函数,产生累积概率为 `lower.tail` 的标准正态分布的临界值。
- 累积分布函数
- 累积分布函数的一般形式为 `pnorm(q, mean = 0, sd = 1, lower.tail = TRUE)`,其中圆点前是函数类名,圆点后是分布名称,括号内是自变量。第一个自变量 `q` 是符合分布的数值,后面的自变量是相应分布的参数。函数值是相应分布的随机变量取值小于等于 `q` 的概率值。
- `pnorm(q, mean = 0, sd = 1, lower.tail = TRUE)` 数值型函数,产生来自具有给定概率参数 `lower.tail` 的伯努利分布,变量值小于 `q` 的累积概率值。
 - `pnorm(q, mean = 0, sd = 1, lower.tail = TRUE)` 数值型函数,产生来自 `gamma` 分布的变量取值小于 `q` 的累积概率值,该分布具有给定的形状参数 `shape` 和尺度参数 `scale`。
 - `pnorm(q, mean = 0, sd = 1, lower.tail = TRUE)` 数值型函数,产生来自二项分布的变量取值小于 `q` 的累积概率值,该分布具有给定每次实验成功的概率 `prob` 和成功的实验次数是 `n` 当 `n=1` 时,该函数与 `pnorm(q, mean = 0, sd = 1, lower.tail = TRUE)` 相同。
 - `pnorm(q, mean = 0, sd = 1, lower.tail = TRUE)` 数值型函数,产生来自柯西分布的变量取值小于 `q` 的累积概率值,该分布具有给定的位置参数 `loc` 和标度参数 `scale`。
 - `pnorm(q, mean = 0, sd = 1, lower.tail = TRUE)` 数值型函数,返回来自卡方分布的变量取值小于 `q` 的累积概率值,该分布具有给定的自由度 `df`。
 - `pnorm(q, mean = 0, sd = 1, lower.tail = TRUE)` 数值型函数,产生来自指数分布的变量取值小于 `q` 的累积概率,该分布具有给定的形状参数 `shape`。



- 悦云裁择云数值型函数,产生来自云分布的变量取值小于 择的累积概率值,该分布具有给定的自由度 云和参数 云,累积概率值小于 择。
- 悦云裁择云数值型函数,产生来自伽玛分布的变量取值小于 择的累积概率,该分布具有给定的形状参数 云和标度参数 云。
- 悦云裁择云数值型函数,产生一个几何分布的变量取值小于 择的累积概率,即获得一次成功的试验次数,当成功概率由 云确定时。
- 悦云裁择云数值型函数,产生小于 择的累积概率,即具有指定特性的事件数 择,当样品 云是事件是从尺寸为 云的总体中随机选择出来的情况下,其命中数 云具有指定的特性。
- 悦云裁择云数值型函数,产生来自拉普拉斯分布的变量取值小于 择的累积概率,该分布具有给定的均值 云和标度参数 云。
- 悦云裁择云数值型函数,产生来自逻辑斯蒂分布的变量取值小于 择的累积概率,该分布具有给定的均值 云和标度参数 云。
- 悦云裁择云数值型函数,产生具有指定参数的对数正态分布变量取值小于 择的累积概率值。
- 悦云裁择云数值型函数,产生一个具有均值为 云标准差为 云的随机变量的取值小于 择的概率。
- 悦云裁择云数值型函数,产生一个正态分布的变量取值小于 择的累积概率,该分布均值为 云标准差为 云。
- 悦云裁择云数值型函数,产生一个来自 云分布的小于 择的累积概率值,它具有指定的均值或概率参数。
- 悦云裁择云数值型函数,产生一个变量取值小于 择的学生 云分布的累积概率,该分布具有指定的自由度参数 云。
- 悦云裁择云数值型函数,产生一个变量取值小于 择的均匀一致分布的累积概率,该分布具有指定的最小值 云和最大值 云参数。
- 悦云裁择云数值型函数,产生一个变量取值小于 择的威布尔分布的累积概率,该分布具有指定的参数。

愿云裁择云数值型函数

- 愿云裁择云数值型函数,自变量是当前工作数据文件中的变量名。计算自变量中缺失值的数目。例如 云(云数学):分别对数学这个变量逐个判断是否为系统缺失值或用户缺失值。云表示是,云为不是。
- 愿云裁择云逻辑型函数,自变量应该是工作数据文件中的变量名。如果变量具有缺失值,返回 云或者 云。
- 愿云裁择云逻辑型函数,自变量 云是工作数据文件中的一个数值型变量的变量名。如果 云的值为系统缺失值,返回 云或者 云。



• **忽略用户的缺失值** 数值型或字符型函数,忽略用户的缺失值,即将用户缺失值看成是普通的数据,返回变量值。自变量必需是工作数据文件中的变量名。

忽略用户的缺失值字符串函数

• **连接字符串** (CONCATENATE) 字符型函数,函数中每个自变量都是一个字符串表达式。该函数值是一个字符串,是各自变量代表的字符串按括号中的顺序串接起来的。此函数要求两个或两个以上的自变量。

• **查找字符串** (FIND) 数值型函数,产生一个整数,它表明字符串在字符串中第一次出现的起始位置。如果返回值为0表明字符串不在字符串中存在。例如:在字符串“Excel”中找到字符串“ce”在字符串中第一次出现的位置,返回值为3。

• **字符串长度** (LEN) 数值型函数,自变量是字符串,函数值是字符串表达式值的长度。这里获得的长度包括尾部空格。

• **字符串左对齐** (LEFT) 字符型函数,第一个自变量是字符串,第二个自变量是正整数,其范围从1到255。函数值是字符串表达式的左侧增加空格扩展到指定长度。

• **字符串右对齐** (RIGHT) 字符型函数,返回的字符串是自变量表达式的值去除打头的空格后的字符串。

• **字符串大写** (UPPER) 字符型函数,返回字符串,将字符串中的大写字母改变为小写字母。

• **字符串右对齐** (RIGHT) 数值型函数,产生一个整数,它表明字符串在字符串中最后出现的开始位置。返回0表示字符串不在字符串中。

• **字符串左对齐** (LEFT) 字符型函数,返回字符串,其长度由指定长度决定:在字符串表达式的右侧加空格,以达到指定长度,指定的值在1到255之间。

• **字符串截取** (MID) 字符型函数,返回截取了尾部空格后的字符串。该函数通常用于大字符串表达式中,要把压缩了尾部空格的字符串赋予一个变量。

• **字符串格式** (TEXT) 字符型函数,根据指定的格式将数值表达式转换为字符串。例如:将123.456转换为字符串“123.456”。第二个自变量必需是写一个数值的格式。

• **字符串大写** (UPPER) 字符型函数,返回将字符串表达式中小写字符变为大写字母。

注意:数值与数字有区别,以上所讲的数值是数,数字指的是表现为数字的字符。



主要参考文献

吴加勋等编著：《应用统计学》，中国人民大学出版社 2008 年版。

袁卫等编著：《统计学》，高等教育出版社 2006 年版。

薛薇编著：《统计分析及应用》，电子工业出版社 2008 年版。

马力著：《回归分析方法原理及 SPSS 实际操作》，中国金融出版社 2008 年版。

余建英、何旭宏编著：《数据统计分析及 SPSS 应用》，人民邮电出版社 2008 年版。

吴喜之编：《非参数检验》，中国统计出版社 2003 年版。

李子奈编著：《计量经济学——方法和应用》，清华大学出版社 2005 年版。

卢纹岱主编《SPSS 18.0 统计案例》，电子工业出版社 2008 年版。

荆达民、程岩编：《应用统计》，化学工业出版社 2008 年版。

盛骤等编：《概率论与数理统计》，高等教育出版社 2008 年版。

柯晓群、刘文卿编著：《应用回归分析》，中国人民大学出版社 2008 年版。

茆诗松主编：《统计手册》，科学出版社 2008 年版。

美 伦纳德·允卡茨米尔(李金昌等译)：《统计学原理学习指南与习题集》，机械工业出版社 2008 年版。