高 等 学 校 教 材

SPSS

统计分析高级教程

（第3版）

张文彤 董 伟 编著

高等教育出版社

高 等 学 校 教 材

SPSS

统计分析高级教程

（第3版）

张文彤 董 伟 编著

藏言 图书馆

高等教育出版社·北京

内容简介

本书全面、系统地介绍了各种多变量统计模型、多元统计分析模型、智能统计分析方法的原理和软件实现，是一本使用 SPSS 进行高级统计分析的实用性很强的指导书和参考书。本书共分 4 个部分，分别是一般线性模型、线性混合模型与广义线性模型，回归模型，多元统计分析方法，以及其他统计分析方法。本书基于 IBM SPSS Statistics 24 中文版，并结合作者多年的统计分析实战经验和 SPSS 行业应用经验，侧重对统计新方法、新观点的讲解，在保证统计理论严谨权威的同时注重叙述的浅显易懂，使本书更加易学易用。

本书可作为高等学校本科生和研究生统计学相关课程教材，也可作为市场营销、金融、财务、人力资源管理等行业中需要做数据分析的人士，或从事咨询、研究、分析等工作的人士的参考书。

图书在版编目（CIP）数据

SPSS 统计分析高级教程 / 张文彤，董伟编著. --3 版. --北京：高等教育出版社，2018.1

ISBN 978-7-04-049033-6

Ⅰ. ①S… Ⅱ. ①张… ②董… Ⅲ. ①统计分析-统计程序-高等学校-教材 Ⅳ. ①C819

中国版本图书馆 CIP 数据核字(2017)第 302226 号

SPSS Tongji Fenxi Gaoji Jiaocheng

策划编辑	刘 艳	责任编辑	刘 艳	封面设计	于文燕	版式设计	马敬茹
插图绘制	于 博	责任校对	刘 莉	责任印制	赵义民		

出版发行	高等教育出版社	网 址	http://www.hep.edu.cn
社 址	北京市西城区德外大街 4 号		http://www.hep.com.cn
邮政编码	100120	网上订购	http://www.hepmall.com.cn
印 刷	北京市联华印刷厂		http://www.hepmall.com
开 本	787 mm× 1092 mm 1/16		http://www.hepmall.cn
印 张	33.75	版 次	2004 年 9 月第 1 版
字 数	830 千字		2018 年 1 月第 3 版
购书热线	010-58581118	印 次	2018 年 1 月第 1 次印刷
咨询电话	400-810-0598	定 价	65.00 元

本书如有缺页、倒页、脱页等质量问题，请到所购图书销售部门联系调换

版权所有 侵权必究

物 料 号 49033-00

前　言

本书自2004年第1版出版以来，受到了广大读者的欢迎，被国内数百所高校选为本科生或研究生相关课程教材，在此感谢广大读者的支持与厚爱。

本书第2版出版于2013年，SPSS在这几年间已经升级了4个版本，而且最新的版本易用性更强，软件功能更丰富，因此需要对全书内容进行有针对性的修订。在第2版的基础上，结合SPSS版本的更新和读者的反馈，本版以IBM SPSS Statistics 24中文版为准，对内容做了如下调整。

1. 内容进一步拓展

本书第2版全面覆盖了SPSS自身提供的各种高级统计分析功能，但SPSS提供的主要是成熟且常用的统计模型，许多独特的新模型并未提供。实际上，SPSS通过Python插件和R插件的方式提供了这些模型。为进一步拓展读者的知识面，除介绍SPSS新版本直接提供的新模型外，本书还对一些比较重要的用插件方式提供的模型进行了介绍，包括分位数回归、弗斯Logistic回归、潜类别分析、支持向量机、随机森林、项响应模型等，并在附录中介绍了相应插件的安装方法，以帮助读者及时跟进统计分析领域的最新进展，并在工作中充分发挥SPSS的作用。

2. 更加浅显易懂

本书涉及的统计模型都比较复杂，为降低学习难度，确保读者能够掌握相应的统计分析方法，本书在第2版的基础上进一步减少了案例数量。通过对同一案例在不同方法框架下的分析结果进行比较，再辅以对统计理论深入浅出的讲解，降低了初学者的入门难度，最大限度地优化了学习效果，有利于读者学以致用。

本书由张文彤和董伟共同编写，可作为高等学校各专业本科生和研究生的统计学相关课程教材，也可作为市场营销、金融、财务、人力资源管理等行业中需要进行数据分析的人士，或从事咨询、研究、分析等工作的人士的参考书。学习本书的读者需要具备统计分析及SPSS操作的基础知识，需要补充这部分知识的读者可以先学习《SPSS统计分析基础教程》(第3版)。对于希望进一步提升统计分析和数据挖掘实战能力的读者，则可以在学习完本书后继续阅读作者的实战案例精粹系列书籍，以进一步提高实战经验。为便于读者交流和使用这套书，读者可以关注微信公众号：统计之星。本书的案例数据文件、拓展资料等可到本书配套的数字课程网站和“医学统计之星”网站上下载。

希望广大读者一如既往地踊跃提出自己的宝贵意见和建议，使得本书再次改版时能够更上一层楼，更好地满足大家的学习和工作需求。

作　者

2017年10月

目　录

第一部分　一般线性模型、线性混合模型与广义线性模型

试读结束:需要全本请在线购买:　www.ertongbook.com

试读结束：需要全本请在线购买：　www.ertongbook.com

1

第一部分

一般线性模型、线性混合模型与广义线性模型

第1章　方差分析模型

通过对《SPSS 统计分析基础教程》(第 3 版)(以下简称基础教程)的学习,读者已经全面掌握了 IBM SPSS Statistics 的基本操作,以及图表绘制、描述统计技术和单因素统计分析方法。但是,真实的世界复杂多变,各种变量间的联系错综复杂,仅仅依靠描述统计或者简单的统计推断方法往往无法满足分析需求,必须依靠更强大的统计模型来解决问题。本书将进一步介绍各种高级统计模型,而本章将要介绍的方差分析模型就是其中最基础和最常用的一种。

一见模型二字,很多读者就会觉得继续学习下去的心理压力很大。实际上,模型无非是对复杂现实世界的一种简化描述而已,一个出色的模型必然比现实世界更简明、易懂,而模型也不一定是枯燥的公式表达,比如说当其以一种赏心悦目的姿态出现在 T 台上时,中文就将其翻译为"模特"。怎么样,这样解释之后,大家对模型的感觉是不是变好了一点?

1.1　模型简介

在实际项目中,研究者在分析数据时往往需要同时考察多个因素对因变量的影响,如要研究性别对身高的影响,就应当控制年龄、遗传、营养状况等因素的影响。对此单因素统计分析方法是无能为力的,而以方差分析模型为代表的多因素统计分析方法可以在控制其他因素影响的同时研究两个因素之间的关系,分析的效率更高,适用范围更广。

此外,许多时候各个因素之间还存在交互作用。例如,在研究催化剂对化学反应的催化能力时,如果该催化剂只在某个温度范围内效果最佳,则单独研究该催化剂的催化作用并无实际意义,此时催化剂和温度之间的交互作用也应成为研究的重点,即必须要研究在什么温度下该催化剂的催化能力最佳。对交互作用的分析是方差分析模型的特长之一。

1.1.1　模型入门

1. 单因素方差分析模型的结构

首先从一个简单的统计模型开始,假设现在希望描述某个人群的月收入状况,那么根据统计学知识,如果月收入服从正态分布(请读者注意这个前提假设),则平均值能够表示集中趋势,标准差能够表示离散趋势,任何一位受访者 i 的月收入 Y_i 可以被表达为如下形式:

$$Y_i=\mu+\varepsilon_i$$

其中,Y_i 代表第 i 位受访者的月收入。显然,此时总体平均值 μ 是 Y_i 的最佳估计值,而 ε_i 则表示因各种原因导致的第 i 位受访者实际月收入和平均值之差,或者说反映了抽样中的随机误差。由于已经假定月收入服从正态分布,因此模型可以设定 ε_i 服从平均值为 0、标准差为某个定值的正态分布 $N(0,\sigma^2)$。

下面开始扩展模型框架。假设现在比较三种职业,即医生、律师和软件工程师的月收入,并

判断其有无差异，那么最简单的做法就是对每种职业的人群都进行随机抽样，得到三组受访者，收集他们的月收入数据，然后进行检验。在此问题中，每一位受访者月收入的平均估计值 Y_{ij} 可以被表达为如下形式：

$$Y_{ij}=\mu_i+\varepsilon_{ij}$$

其中，Y_{ij} 代表第 i 组中第 j 位受访者的月收入。显然，在此表达式中 μ_i 表示某一组月收入的平均值，i 的取值范围为 1~3，分别代表三种职业之一；而 ε_{ij} 表示第 i 组的第 j 位受访者的月收入相对于本组月收入平均值的随机误差。

下面来看模型中对 ε_{ij} 的设定，在原始的模型框架中，显然各组的 ε_{ij} 可以服从各自的正态分布，但这类研究往往更关心平均值的差异，因此为了简化模型架构，一般会进一步假设各组的 ε_{ij} 服从同一个正态分布，即无论 i 取值是多少，ε_{ij} 均服从同一个平均值为 0、标准差为某个定值的正态分布 $N(0,\sigma^2)$。这样一来，如果三种职业的月收入无差异，那么它就应当等于总体平均值（平均水平）再加上一个随机误差项，实际上就变成了同一个变量的正态分布 $N(\mu,\sigma^2)$。为了能够对收入水平进行预测，人们又规定 $E(Y)=\mu_i$，即第 i 组个体的月收入估计值等于该组月收入的平均水平，结合模型结构，这应当不难理解。实际上，如果对应的是样本数据，该估计值就是各组的样本平均值。

为了统计推断的需要，以上模型往往被改写成如下形式：

$$Y_{ij}=\mu+\alpha_i+\varepsilon_{ij}$$

其中，μ 表示不考虑职业时月收入的总平均水平；α_i 表示职业 i 的主效应，即职业 i 月收入平均水平的改变情况。例如，$\alpha_i=1\ 000$，表明职业 i 月收入的平均水平要比总平均水平高 1 000 元。如果职业 1 和职业 3 的平均月收入不相等，则有 $\alpha_1\neq\alpha_3$。反之，如果三种职业的平均月收入无差异，则因各种职业均不存在主效应而有 $\alpha_1=\alpha_2=\alpha_3=0$。因此，检验职业种类是否对月收入有影响，就是检验如下假设：

$$H_0:\text{对任意的 } i \text{ 取值，都有 } \alpha_i=0;\ H_1:\text{至少有一个 } \alpha_i\neq 0$$

μ、α_i 等显然是一个相对的大小，可以有无限多组取值方式。例如，职业 1 比职业 3 的平均月收入高 1 000 元，则当 α_3 为 500 时，α_1 就应当是 1 500；当 α_3 为 100 时，α_1 就应当是 1 100，总之加上 1 000 即可。为了能够在实际问题中得到 μ、α_i 的具体估计值，模型拟合中又会对它们有一些附加的设定，这被称为模型拟合时的约束条件，相关的详细介绍见后。

在基础教程中，大家已经学习了方差分析的基本思想是变异分解。例如，在单因素方差分析中总变异被分解为如下两部分：

$$\text{总变异}=\text{处理因素导致的变异}+\text{随机变异}$$

对照前面的模型表达式，就会发现实际上 α_i 对应了处理因素导致的变异，而 ε_{ij} 对应了相应的随机变异。在多因素方差分析模型中，这一原理没有任何变化，只是模型中考虑的因素更多而已。

2. 两因素方差分析模型的结构

下面开始对单因素模型进行扩展。同样是上面的问题，有研究者提出：性别也对收入水平有影响，也许正是因为其中一组中男性的比例高于其他组，才导致该组的平均月收入高于后者。为什么不考虑控制性别的影响？如果要同时考虑性别和职业对收入的影响，则模型扩展为

$$Y_{ijk}=\mu+\alpha_i+\beta_j+\varepsilon_{ijk}$$

该模型对应如下变异分解式：

试读结束：需要全本请在线购买： www.ertongbook.com