Statistics

21世纪统计学系列教材

Statistics with SPSS

统计学——基于SPSS（第二版）

贾俊平　编著

中国人民大学出版社

Statistics 21世纪统计学系列教材

Statistics with SPSS

统计学——基于SPSS（第二版）

贾俊平　编著

中国人民大学出版社
· 北京 ·

图书在版编目（CIP）数据

统计学：基于SPSS/贾俊平编著. —2版. —北京：中国人民大学出版社，2016.8
21世纪统计学系列教材
ISBN 978-7-300-23175-4

Ⅰ.①统… Ⅱ.①贾… Ⅲ.①统计学-高等学校-教材②统计分析-软件包-高等学校-教材 Ⅳ.①C8

中国版本图书馆CIP数据核字（2016）第179245号

21世纪统计学系列教材
统计学——基于SPSS（第二版）
贾俊平 编著
Tongjixue：Jiyu SPSS

出版发行 中国人民大学出版社
社　　址 北京中关村大街31号　　**邮政编码** 100080
电　　话 010－62511242（总编室）　　010－62511770（质管部）
010－82501766（邮购部）　　010－62514148（门市部）
010－62515195（发行公司）　　010－62515275（盗版举报）
网　　址 http://www.crup.com.cn
http://www.ttrnet.com(人大教研网)
经　　销 新华书店
印　　刷 北京昌联印刷有限公司　　**版　　次** 2014年7月第1版
2016年8月第2版
规　　格 185 mm×260 mm　16开本　　**印　　次** 2016年8月第1次印刷
印　　张 17.5插页1　　**定　　价** 36.00元
字　　数 374 000

第二版前言

第二版在保留第一版内容框架的基础上，增加了一些新内容，并对个别地方做了修订。主要变化如下：

第1章“数据与统计学”的最后增加“本书图解：统计方法分类与本书框架”，展示了本书的结构和各章的逻辑关系。其余各章均增加了“本章图解”，介绍本章的内容结构及逻辑关系。第2章“数据的描述性分析：图表展示”中，箱线图的部分重新进行了编写，详细介绍了箱线图的绘制步骤及其解读，并重新绘制了箱线图的示意图。第4章“随机变量的概率分布”中增加了正态分布和标准正态分布的概率密度函数，4.3节增加了中心极限定理的模拟图示。第8章“方差分析”删除了8.1.3小节“方差分析的基本假定”，8.2.3小节“多重比较”重新编写，增加了多重比较的HSD方法，增加了8.4节，介绍方差分析的假定及其检验方法，包括方差齐性检验的图示方法和检验方法，如Levene方差齐性检验等。第10章“多元线性回归”中增加了对标准化回归系数的解释。

贾俊平

第一版前言

在大数据时代，每天都会产生大量的数据，这些数据需要处理和分析。作为数据分析方法的统计学自然会受到越来越多的人关注，也会越来越广泛地应用于各个领域。难以想象，不使用计算机或统计软件如何处理和分析这些海量数据。

多数人都把统计学当作一门难学的课程来看待，原因之一就是对统计计算望而生畏，对复杂的统计公式望而却步。如果能从繁杂但属于简单劳动的计算中解脱出来，把它交给统计软件来完成，把统计计算统统“秒杀”，从而拿出更多的精力去理解统计方法的思想和原理，就会发现统计学不仅不像想象的那么难学，而且是一门非常有趣、非常有用的科学。

SPSS是最早引入国内的优秀统计分析软件之一，以其视窗操作、易于使用和输出结果直观易懂等特点，被多数人广泛使用。目前，SPSS已有汉化版本，虽然汉化版本中许多术语和表述有不当之处，但还是会方便更多人使用。

本书是一本基于SPSS实现全部计算的统计学教材，书中例题的解答给出了SPSS的详细操作步骤。考虑到多数读者使用上的方便，本书使用的是SPSS 19.0中文版（建议有能力的读者使用英文版）。全书内容共11章，包括数据的描述性分析方法、推断方法以及实际中常用的一些统计方法等。每章均以一个实际问题引入该章要介绍的内容。在写法上完全立足于统计应用，避免统计公式的推导，力求通俗易懂。在形式上，本书给出了例题和练习题的数据文件，读者可通过扫描二维码下载。

本书可作为高等院校经济管理类专业本科生统计学课程的教材使用，也可作为其他文科专业及部分理、工、农、林、医、药专业的教材或参考书使用，对广大实际工作者也极具参考价值。由于作者水平有限，错误难免，希望读者在使用中对本书的不足之处多提宝贵意见，以便进一步修改和完善。

贾俊平

2014年1月于中国人民大学统计学院

目　录

第 1 章

Chapter 1 数据与统计学

问题与思考：怎样理解统计结论?

每天我们都会看到各种统计数字或统计研究的某些结论。下面就是一些有趣的统计结论：

- 吸烟对健康是有害的，吸香烟的男性寿命减少 2 250 天。
- 不结婚的男性寿命会减少 3 500 天，不结婚的女性寿命会减少 1 600 天。
- 身体超重 30%会使寿命减少 1 300 天。
- 每天摄取 500 毫升维生素 C，生命可延长 6 年。
- 身材高的父亲，其子女的身材也较高。
- 一项研究表明，杰出科学家做出重大贡献的最佳年龄在 25～45 岁之间，其最佳峰值年龄和首次贡献的最佳成名年龄随着时代的变化而逐渐增大。
- 学生们在听了 10 分钟莫扎特钢琴曲后做的推理，要比他们听 10 分钟其他娱乐性曲目后做的更好。
- 上课坐在前排的学生平均考试分数比坐在后排的高。
- 中国科学院空间环境研究预报中心的专家称，在神舟七号载人航天飞船飞行期间，遭遇空间碎片的概率在百万分之一以下。

这些结论是怎么得出的？你相信这些结论吗？你相信或不相信的理由是什么？要看懂这些结论似乎并不困难，但要合理解释这些结论就需要具备一定的统计学知识了。统计结论是一种归纳推理，这意味着不能肯定统计结论就一定正确。

在日常生活中，经常会接触到统计数据或一些统计研究结果，比如，在电视、报纸、网络等媒体上会经常看到一些报道使用的统计数据、图表等。作为一门科学的统计学研究什么呢？怎样获得所需要的统计数据呢？这就是本章将要介绍的内容。

1.1 统计学及其应用

每个人都离不开统计，了解一些统计学知识对每个人来说都是必要的。比如，在外出旅游时，你需要关心一段时间内的详细天气预报；在投资股票时，你需要了解股票市场的价格信息，了解某只特定股票的有关财务信息；在观看足球比赛时，除了关心进球数之外，你还要知道各支球队的技术统计，等等。要正确阅读并理解统计数据或统计结论，需要具备一些统计学知识。

1.1.1 什么是统计学

在日常工作或管理中，总会面对各种各样的数据。如果不去分析这些数据，那它们也仅仅是一堆数据而已，没有太多的价值。如何分析这些数据，用什么方法分析数据，并从分析中得出某些结论以帮助我们做出决策，这正是统计学要解决的问题。简言之，**统计学**（statistics）是收集、处理、分析、解释数据并从数据中得出结论的原则和方法。统计学所提供的是一系列有关数据收集、处理和分析的方法。

数据收集就是取得所需要的数据。数据的收集方法可分为两大类：一是观察方法，二是实验方法。观察方法是通过调查或观测获得数据；实验方法是在控制实验对象条件下通过实验获得数据。

数据处理是对所获得的数据进行加工和处理，包括数据的计算机录入、筛选、分类和汇总等，以符合进一步分析的需要。

数据分析是利用统计方法对数据进行分析。数据分析所使用的方法大体上可分为**描述统计**（descriptive statistics）和**推断统计**（inferential statistics）两大类。描述统计主要是利用图表形式对数据进行展示，或通过计算一些简单的统计量（诸如比例、比率、平均数、标准差等）对数据进行分析。推断统计主要研究如何根据样本信息来推断总体的特征，内容包括参数估计和假设检验两大类。参数估计是利用样本信息推断所关心的总体特征，假设检验则是利用样本信息判断关于总体的某个假设是否成立。比如，从一批灯泡中随机抽取少数几个作为样本，测出它们的使用寿命，然后根据样本灯泡的平均使用寿命估计这批灯泡的平均使用寿命，或者检验这批灯泡的使用寿命是否等于某个假定值，这就是推断统计要解决的问题。

数据解释是对分析结果进行的说明，包括结果的含义、从分析中得出的结论等。

统计学是一门关于数据的科学，它研究的是来自各领域的数据，提供的是一套通用于所有学科领域的获取数据、分析数据并从数据中得出结论的原则和方法。统计方法是通用于所有学科领域的，而不是为某个特定的问题或领域构造的。当然，统计方法和技术并不是一成不变的，使用者在给定的情况下必须根据所掌握的专业知识选择使用这些方法，而且如有需要还要进行必要的修正。

正如有的学者所指出的那样："统计学基本上是寄生的，靠研究其他领域内的工作而生存。这不是对统计学的轻视，这是因为对很多寄主来说，如果没有寄生虫就会死。对有的动物来说，如果没有寄生虫就不能消化它们的食物。因此，人类奋斗的很多领域，如果没有统计学，虽然不会死亡，但一定会变得很弱。"① 看上去统计似乎被边缘化了，但实际上正说明了统计在各学科领域的独特地位和作用，也表明了统计作为一门独立学科而具有的特点。

1.1.2　统计学的应用

说出哪些领域要应用统计，这很困难，因为几乎所有的领域都用统计；说出哪些领域不用统计，同样也很困难，因为几乎找不到一个不用统计的领域。可以说，统计是适用于所有学科领域的通用数据分析方法，是一种通用的数据分析语言。只要有数据的地方就会用到统计方法。

1. 统计学的应用领域

统计学广泛应用于各个学科领域，为各学科的发展做出了重要贡献。这里，我们不想列举统计学的应用领域，只想通过几个简单的例子说明统计学的应用。

例 1—1

用统计识别作者。1787—1788 年，三位作者亚历山大·汉密尔顿（Alexander Hamilton）、约翰·杰伊（John Jay）和詹姆斯·麦迪逊（James Madison）为了说服纽约人认可宪法，匿名发表了著名的 85 篇论文。这些论文中的大多数作者已经得到了确认，但是，其中的 12 篇论文的作者身份引起了争议。通过对这些论文不同单词的频数进行统计分析，得出的结论是，詹姆斯·麦迪逊最有可能是这 12 篇论文的作者。现在，对于这些存在争议的论文，认为詹姆斯·麦迪逊是原创作者的说法占主导地位，而且几乎可以肯定这种说法是正确的。

例 1—2

用简单的描述统计量得到一个重要发现。费希尔（R. A. Fisher）在 1952 年的一篇文章中举了一个例子，说明如何由基本的描述统计量知识引出一个重要的发现。20 世纪早期，哥本哈根卡尔堡实验室的施密特（J. Schmidt）发现在不同地区捕获的同种鱼类的脊椎骨和鳃线的数量有很大不同，甚至在同一海湾内不同地点所捕获的同种鱼类也有这样的倾向。然而，鳗鱼的脊椎骨数量变化不大。施密特在从欧洲各地、冰岛、亚速尔群岛以及尼罗河等几乎分离的海域里所捕获的鳗鱼的样本中，计算发现了几乎一样的均值和标准偏差值。由此，施密特推断所有各个不同海域内的鳗鱼是由海洋中某公共场所繁殖的。后来名为"戴纳"（Dana）的科学考察船在一次远征中发现

① C. R. 劳：《统计与真理——怎样运用偶然性》，北京，科学出版社，2004。

了这个场所。

例 1—3

挑战者号航天飞机失事预测。1986 年 1 月 28 日清晨，载有 7 名航天员的挑战者号进入发射状态。发射几分钟后，航天飞机发生爆炸，机上的航天员全部遇难。在此次失事前，该航天飞机 24 次发射成功。将航天飞机送入太空的两个固体燃料推进器由 6 个 O 型项圈密封，在几次飞行中，曾发生过 O 型项圈被腐蚀或气体泄漏事故。这类事故与气温是否有关系呢？本次发射时天气预报气温为－0.56℃。下面的表 1—1 是 23 次飞行中 O 型项圈因腐蚀或泄漏事故损坏的个数（因变量 y）及发射时火箭连接处的温度（自变量 x）数据。

表 1—1　　挑战者号航天飞机 23 次飞行中损坏的 O 型项圈个数和发射时的温度

飞行频数	O 型项圈的损坏个数	温度（℃）	飞行频数	O 型项圈的损坏个数	温度（℃）
1	2	11.7	13	1	21.1
2	1	13.9	14	1	21.1
3	1	14.4	15	0	22.2
4	1	17.2	16	0	22.8
5	0	18.9	17	0	23.9
6	0	19.4	18	2	23.9
7	0	19.4	19	0	24.4
8	0	19.4	20	0	25.6
9	0	20.0	21	0	26.1
10	0	20.6	22	0	27.2
11	0	21.1	23	0	24.4
12	0	21.1			

根据表 1—1 的数据进行线性回归得到的回归方程为 $\hat{y}=2.177\,1-0.085\,6x$。由此得到当温度为－0.56℃时，O 型项圈发生事故的预计个数为 2.225 个。结果显示连接处的温度与 O 型项圈事故之间有一定的相关性。如果当时管理者看到了回归的预测结果，选择延迟发射也许会成为最佳选择。

前两个是统计得以应用并取得成效的例子，后一个是统计结果未被采纳而酿成惨剧的例子。不管怎样，它们都表明统计在许多领域都有广泛的应用。

2. 统计的误用与滥用

大约在一个世纪以前，政治家本杰明·迪斯雷利（Benjamin Disraeli）曾有一个著名的论断："谎言有三种：谎言、糟透的谎言和统计。"统计常常被人们有意或无意地滥用，比如错误的统计定义、错误的图表展示、不合理的样本、数据的篡改或造假，等等。这些误用有些是常识性的，有些是技术性的，有些则是故意的。作为从数据中寻找事实的统计，却被有些人变成了歪曲事实的工具。你也许常常看到这样的产品质检报告：某某产品的抽样合格率是 80%。乍看上去还可以，但如果实际上只抽查了 5 件产品，有 4 件合格，这样的合格率能说明什么问题呢？在马路上随便采访几个

人，他们的看法能代表大多数人的观点吗？“调查结果表明……”调查了多少个人？是随机调查的吗？样本是怎样选取的？这看上去是在用事实说话，实际上成了统计陷阱。

在管理领域，统计也往往被作为两个极端使用：一个极端是复杂问题简单化，一些不懂或不太懂统计的人认为统计没什么用，他们因为不懂统计而看不起统计，他们不用或几乎不用统计方法分析数据，即使做些统计分析，也往往是表面上的。走入这一极端的人，他们决策的依据就是自己大脑中一些杂乱无章的信息组合出的某种直觉。如果他们的决策是正确的，他们会更加自信，更加感到不用统计也挺好；如果他们的决策出了毛病，则会找出一大堆推脱的理由：市场难测、环境突变、竞争激烈、需求疲软、价格下跌、管理不善、成本上升、出口下降……另一个极端是把简单问题复杂化，特别是在管理领域。一些管理者把本来可以用简单方法解决的问题故意复杂化，他们不用简单的分析方法，而用复杂的分析方法；他们为证明管理的科学性，建立一个别人看不懂的模型，编一大堆程序，输出一大堆数字和符号；他们得出用统计语言陈述的结论，提出一些似是而非的建议……这样的分析往往既脱离了管理问题，对实际决策也未必有用。在管理中，这两个极端都是不可取的。管理决策中不用统计几乎不可想象，但把简单问题复杂化对管理决策未必有用。从统计的实际应用来看，简单的方法不一定没用，复杂的方法也不一定有用。统计应该被恰当地应用到它能起作用的地方。不能把统计神秘化，更不能歪曲统计，把统计作为掩盖事实的陷阱。

曲解统计是一种常见现象。在有些人看来，使用统计就是寻找支持：在他们的心目中可能早已有了某种“结论”性的东西，或者说他们希望看到符合他们需要的某种结论，便去找些数据来支持他们的结论。如果数据分析的结果与他们预期的结论一致，他们就会声张自己是用科学方法得到的结论；如果与预期的不一致，他们要么会篡改数据，要么对统计弃而不用。这恰恰背离了数据分析的本质。数据分析的真正目的是从数据中找出结论，从数据中寻找启发，而不是寻找支持。真正的数据分析事先是没有结论的，通过对数据的分析才得出结论。

1.2 数据及其来源

统计分析离不开数据，没有数据统计方法就成了无米之炊。数据是什么？怎样获得所需的数据？这就是本节将要介绍的内容。

1.2.1 变量与数据

观察一个企业的销售额，你会发现这个月和上个月有所不同；观察股票市场上涨股票的家数，今天与昨天的数量不一样；观察一个班学生的生活费支出，一个人和另一个

人不一样；投掷一枚骰子观察其出现的点数，这次投掷的结果和下一次也不一样。这里的“企业销售额”、“上涨股票的家数”、“生活费支出”、“投掷一枚骰子出现的点数”等就是变量。简言之，**变量**（variable）是描述观察对象某种特征的概念，其特点是从一次观察到下一次观察可能会出现不同结果。变量的观测结果就是**数据**（data）。

根据观测结果的特征，变量可以分为类别变量和数值变量两种。

类别变量（categorical variable）是取值为事物属性或类别以及区间值的变量，也称**分类变量**（classified variable）或**定性变量**（qualitative variable）。比如，观察人的性别、公司所属的行业、用户对商品的评价时，得到的结果就不是数字，而是事物的属性。例如，观测性别的结果是“男”或“女”，公司所属的行业为“建造业”、“零售业”、“旅游业”等；用户对商品的评价为“很好”、“好”、“一般”、“差”、“很差”。人的性别、公司所属的行业、用户对商品的评价等作为变量取的值不是数值，而是事物的属性或事物的类别。此外，考虑学生月生活费支出的档次可能分为1 000 元以下、1 000～1 500元、1 500～2 000 元、2 000 元以上 4 档，变量“月生活费支出档次”的这 4 档取值也不是普通的数值，而是数值区间，因而变量也称为区间值类别变量。人的性别、公司所属的行业、用户对商品的评价、学生月生活费支出的档次等都是类别变量。

类别变量根据取值是否有序通常分为两种：**名义**（nominal）**值类别变量**和**顺序**（ordinal）**值类别变量**。名义值类别变量也称**无序类别变量**，其取值是不可以排序的。比如“公司所属的行业”这一变量的取值为“建造业”、“零售业”、“旅游业”等，这些取值之间不存在顺序关系。又如“商品的产地”这一变量的取值为甲、乙、丙、丁，这些取值之间也不存在顺序关系。顺序值类别变量也称**有序类别变量**，其取值可以排序。例如“对商品的评价”这一变量的取值为“很好”、“好”、“一般”、“差”、“很差”，这 5 个值之间是有顺序的。取区间值的变量当然是有序的类别变量。当类别变量只取两个值时也称**二值**（binary）**类别变量**，例如“性别”这一变量的取值为“男”和“女”。二值变量可以看做名义变量，也可以看做有序变量。

类别变量的观测结果称为**类别数据**（categorical data）。类别数据也称**分类数据**或**定性数据**。与类别变量相对应，类别数据相应分为名义值类别数据和顺序值类别数据两种。其中只取两个值的类别数据也称**二值类别数据**。

数值变量（metric variable）是取值为数字的变量，也称**定量变量**（quantitative variable）。例如“企业销售额”、“上涨股票的家数”、“生活费支出”、“投掷一枚骰子出现的点数”等这些变量的取值可以用数字来表示，都属于数值变量。数值变量的观察结果称为**数值数据**（metric data）或**定量数据**。

数值变量根据其取值的不同，可以分为**离散变量**（discrete variable）和**连续变量**（continuous variable）。离散变量是只能取有限个值的变量，而且其取值可以一一列举，如“企业数”、“产品数量”等就是离散变量。连续变量是可以在一个或多个区间中取任意值的变量，它的取值是连续不断的，不能一一列举，比如“年龄”、“温度”、“零件尺寸的误差”等都是连续变量。当离散变量的取值很多时，也可以将离散变量当作连续变量来处理。