高/等/教/育/体/育/学/教/材

体育多元统计分析

祁国鹰　编著

北京体育大学出版社

体育多元统计分析

祁国鹰　编著

北京体育大学出版社

出 版 人　李　飞
责任编辑　佟　晖
审稿编辑　董英双
责任校对　未　茗
版式设计　佟　晖
责任印制　陈　莎

图书在版编目(CIP)数据

体育多元统计分析 / 祁国鹰编著. -- 北京 : 北京体育大学出版社, 2015.11
ISBN 978-7-5644-2115-1

Ⅰ. ①体… Ⅱ. ①祁… Ⅲ. ①体育统计－多元分析－统计分析 Ⅳ. ①G80-32

中国版本图书馆CIP数据核字(2015)第279929号

体育多元统计分析　　祁国鹰　编著

出　版　北京体育大学出版社
地　址　北京市海淀区信息路48号
邮　编　100084
邮购部　北京体育大学出版社读者服务部 010-62989432
发行部　010-62989320
网　址　http://cbs.bsu.edu.cn
印　刷　北京京华虎彩印刷有限公司
开　本　787×1092毫米　1/16
印　张　25.25

2016年1月第1版第1次印刷
定　价：55.00元

前　言

体育工作者在体育教学、训练和科学研究工作中经常会遇到多元统计分析问题。对于多元统计分析问题，人们过去常常是固定某一研究因素以外的因素来进行分析，结果很难从整体上对问题进行描述和推断。例如，人的体质的强弱与身体形态发育水平、生理机能水平、身体素质及运动能力水平、心理的发育水平以及适应能力等诸多方面的因素均有关系，简单地采用单因素分析的方法，仅能对各个因素的单一影响加以对比研究，而不能揭示出各影响因素的内在联系以及所有研究因素共同对体质的综合影响。多元统计分析是研究多个变量（指标或因素）之间关系的一类统计方法的总称。这类方法的主要目的是从反映事物的相互关联的多种观测记录出发，去寻求事物的简明记述（如分类），对事物进行预测或判断，找寻支配事物的主要因素等等，从而能够综合地对事物的整体进行描述或判断。

中国体育事业迅猛发展的同时迫切要求在体育教学、运动训练、体育管理以及体育科研等方面提高理论水平，要求从经验的、定性的、感性的、直觉的状态下摆脱出来，向科学的、定量的、有理论指导的、能进行分析的、自觉的状态过度。随着计算机统计软件的飞速发展，开辟了多元统计分析在体育领域中应用的广阔前景，愈来愈多的多元统计分析方法已经应用于体育科研领域，并取得了一定的成绩。体育科研人员、体育教师以及教练员能够藉助多元统计分析方法对他们所关心的体育问题进行适当地定量描述，而后运用计算机软件进行统计计算和分析，得到定量的结论。当前，在体育科研领域中，正确地使用多元统计分析解决有关的实际问题，已经成为衡量研究工作水平先进性、科学性的一个重要标志。正因为如此，有必要编写一本适用于体育方面的多元统计分析的书以适应体育科学发展的需要。

笔者曾在1998年出版了《体育用多元分析》一书，该书是从体育工作的实用角度出

发，收集了一些常用的多元统计分析方法，选配了一些体育教学、训练和科研的实例，由浅入深具体地介绍多元统计分析方法。该书的出版受到了广大从事体育教学、训练以及科研工作者的欢迎，其对体育科研科学化水平提高起到了促进作用。该书的部分例题的解题示范使用了SPSS/PC+统计软件。因此，读者在演练或实际运用中需要使用SPSS/PC+的命令进行适当地编程后方能完成统计分析过程。伴随着微型计算机操作系统的升级，视窗下的SPSS软件以及Excel软件界面更加简捷、明了，读者无需编程，只要确实能理解其选择的统计分析过程的目的，仅对相应对话框中的内容进行选择就能完成计算分析，得到满意的结果。于是在《体育用多元分析》的基础上，与时俱进地增补、更新内容，使用SPSS13.0软件和Excel软件作为部分例题的解题示范等进行实质性地提升并重新编写出版一本新书，对于方便体育工作者学习和正确运用多元统计分析方法具有实用价值和现实意义。

全书共分十一章：第一章列联表分析；第二章方差分析；第三章回归分析；第四章判别分析；第五章聚类分析；第六章主成分分析；第七章因子分析；第八章数量化方法；第九章多指标多方案综合评价方法；第十章单因素统计设计方法；第十一章对应分析。各章的内容相对独立，读者可以结合实际工作和科研的需要而选读。

在本书的编写过程中，参阅了有关的书籍和文献，并引用了其中的一些材料和统计图表等，特在此谨向各书的编著者和出版者表示深切地感谢。夏晓硕士研究生参与了本书的编写工作，在资料的录入、统计软件的更新以及计算结果的复核等繁杂工作中极其认真负责，故在此一并表示感谢。

由于作者的水平有限，加之时间仓促，书中错误或不妥之处恳请专家和读者批评指正。

祁国鹰

2013年1月 于北京体育大学

内容提要

本书结合体育教学、运动训练以及体育科研的实例介绍了多元统计分析的一些常用方法，其重点在于方法的实用。本书的特点是“淡化系统，突出实用”。部分例题的解题使用了SPSS13.0软件和Excel软件，并对输出的结果进行了较详细地解释或说明。全书主要内容有：列联表分析，方差分析，回归分析，判别分析，聚类分析，主成分分析，因子分析，数量化方法，多指标多方案综合评价方法，单因素统计设计方法，对应分析。

本书的读者对象是体育统计工作者、体育科研人员、体育教师或教练员、体育院校高年级学生以及各体育专业的研究生等。本书亦可作为体育院校体育多元统计分析相应课程的教学参考书。

目录 CONTENTS

第一章
列联表分析

教学提示

对某一总体中的个体可用多种不同方式进行分类。例如，人能分为男性和女性、已婚和单身、有选举权和无选举权等等，这是二级分类的一些例子。多级分类的例子如：人的气质类型可分为多血质、胆汁质、粘液质、抑郁质四种；人的能力评定可分为优、良、中、差、劣五等；人对某项体育运动的态度可分为十分喜欢、一般喜欢、不喜欢三类等等。当个体属性分类确定后，总体即可以被分为若干类，我们便可以对每类中个体的数目计数。这些计数值或频数，通常被称为定性数据。

本章主要介绍以交叉分类或以列联表形式出现的定性数据的分析方法。

第一节　列联表和卡方检验

一、列联表

列联表是一种多重分类法，即是在一张表上安排两类指标因子的表格，一个指标因子（定性变量B）安排在纵列上，另一指标因子（定性变量A）安排在横行上。如对某一事物的N次观测所组成的样本，按照A和B两个定性变量分类，A分有1,2，…，i，…，r个互不相容的类；B分有1,2，…,j，…，c个互不相容的类。由它们形成的频数分布表，称为r×c列联表。这种表主要是用以检验两类指标因子之间是否存在相互独立关系。为此，先假设它们是相互独立的，然后构造统计量，对此假设进行检验和判定。仅含有两个变量的列联表，称为二维列联表，见表1-1所示。

表1-1　二维列联表的一般型式

		列变量（B）						总计
		1	2	……	j	……	c	
行变量（A）	1	n_{11}	n_{12}	……	n_{1j}	……	n_{1c}	$n_{1\cdot}$
	2	n_{21}	n_{22}	……	n_{2j}	……	n_{2c}	$n_{2\cdot}$
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	i	n_{i1}	n_{i2}	……	n_{ij}	……	n_{ic}	$n_{i\cdot}$
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	r	n_{r1}	n_{r2}	……	n_{rj}	……	n_{rc}	$n_{r\cdot}$
总计		$n_{\cdot 1}$	$n_{\cdot 2}$	……	$n_{\cdot j}$	……	$n_{\cdot c}$	$n_{\cdot\cdot}=N$

行变量第i类，列变量第j类中的观测频数或实计数，即表中第ij格中的频数，用n_{ij}表

示。行变量第 i 类中的总观测数用 $n_{i\cdot}$ 表示，列变量第 j 类中的总观测数用 $n_{\cdot j}$ 表示，称它们是边缘总计。由表中频数给出：

$$n_{i\cdot} = n_{i1} + n_{i2} + \cdots + n_{ic} = \sum_{j=1}^{c} n_{ij} \tag{1-1}$$

$$n_{\cdot j} = n_{1j} + n_{2j} + \cdots + n_{rj} = \sum_{i=1}^{r} n_{ij} \tag{1-2}$$

$$n_{\cdot\cdot} = \sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij} \tag{1-3}$$

$$= \sum_{i=1}^{r} n_{i\cdot} = \sum_{j=1}^{c} n_{\cdot j} = N \tag{1-4}$$

$n_{\cdot\cdot}$ 表示样本的总观测数，常用 N 表示。这种表示法常称为点表示，各点表明对特定下标求和。

若在二维列联表中的两个定性变量各有两个分类，如 A 有两个分类，B 也有两个分类，则构成了 2×2 列联表，其一般型式见表 1-2 所示。

表 1-2　2×2 列联表的一般型式

		列变量（B）		总计
		类 1	类 2	
行变量（A）	类 1	n_{11}	n_{12}	$n_{11}+n_{12}$
	类 2	n_{21}	n_{22}	$n_{21}+n_{22}$
总计		$n_{11}+n_{21}$	$n_{12}+n_{22}$	$n_{11}+n_{12}+n_{21}+n_{22}$

2×2 列联表又称为“四格表”。例如表 1-3 中展示了 100 个吸烟者的年龄调查情况，这是一张按照两个定性变量，吸烟量和年龄各取两个分类的四格表。

表 1-3　吸烟量与年龄关系调查结果

组别	60 岁以上	60 岁以下	总计
20 支以上/日	50	15	65
20 支以下/日	10	25	35
总计	60	40	100

二、独立分类与连带

列联表分析的基本问题是研究构成列联表的两类特征和属性之间是否为相互独立的。由表 1-3 的数据可知，如果吸烟量与年龄无关，则理论上讲，60 岁以上年龄的人群中的日吸烟 20 支以上的人所占该人群的比例数，应和年龄在 60 岁以下的人群中的日吸烟 20 支以上的人所占该人群的比例数相等。如果这些比例数不同，则表明吸烟量与年龄有着更为密切的连带关系。

2×2 列联表中独立性意味着两个比例数相等。在一般的 r×c 列联表中这个概念的含义是：总体的一个个体落在一个变量的分类中的概率不受它属于另一个变量分类的影响，反之亦然。假定在总体中，属于行变量第 i 类、列变量第 j 类的一次观测的概率用 P_{ij} 表示；因此，同样本的 N 个个体所得的二维列联表中第 ij 格的理论频数 E_{ij} 用下式给出：

$$E_{ij} = NP_{ij} \tag{1-5}$$

若用 $P_{i\cdot}$ 表示总体中一次观测是属于行变量第 i 类的概率，用 $P_{\cdot j}$ 表示列变量第 j 类的相应概率。不难证明，概率 $P_{i\cdot}$ 和 $P_{\cdot j}$ 的估计量 $\hat{P}_{i\cdot}$ 和 $\hat{P}_{\cdot j}$ 在观测值的相应边缘总计不变的条件下，即有：

$$\hat{P}_{i\cdot} = \frac{n_{i\cdot}}{N} \text{ 和 } \hat{P}_{\cdot j} = \frac{n_{\cdot j}}{N} \tag{1-6}$$

由概率乘法定律知总体中两定性变量相互独立的条件是：

$$P_{ij} = P_{i\cdot}P_{\cdot j} \tag{1-7}$$

若列联表中的理论频数 E_{ij} 的估计值为 $\hat{E}_{ij}$，则独立性意味着

$$\hat{E}_{ij} = N\hat{P}_{i\cdot}\hat{P}_{\cdot j} \tag{1-8}$$

即
$$\hat{E}_{ij} = \frac{n_{i\cdot}n_{\cdot j}}{N} \tag{1-9}$$

当两个定性变量独立时，用公式（1-9）所估计的理论频数与观测频数间仅仅相差一个可归因于偶然因素的量；然而如果两个变量之间存在着某种连带关系，则理论频数与观测频数间能预期会有一个大的差异出现。因此可见，利用 $\hat{E}_{ij} - n_{ij}$ 所提供的信息可对两定性变量独立性作出判断。

为便于计算，将上述理论列成表 1-4。表中 n_{ij} 为观测频数。$\hat{P}_{i\cdot}$ 和 $\hat{P}_{\cdot j}$ 为 $P_{i\cdot}$、$P_{\cdot j}$ 的估计量。$\hat{E}_{ij}$ 为理论频数 E_{ij} 的估计量。

表 1-4　二维列联表的一般计算格式表

		列变量（B）				总计（行）
		1	2	j	c	
行变量（A）	1	n_{11}	n_{12}	n_{1j}	n_{1c}	
		$P_{11} = \hat{p}_{1\cdot}\hat{p}_{\cdot 1}$	$P_{12} = \hat{p}_{1\cdot}\hat{p}_{\cdot 2}$	$P_{1j} = \hat{p}_{1\cdot}\hat{p}_{\cdot j}$	$P_{1c} = \hat{p}_{1\cdot}\hat{p}_{\cdot c}$	$n_{1\cdot}$
		$\hat{E}_{11} = NP_{11} = \frac{n_{1\cdot}n_{\cdot 1}}{N}$	$\hat{E}_{12} = NP_{12} = \frac{n_{1\cdot}n_{\cdot 2}}{N}$	$\hat{E}_{1j} = NP_{1j} = \frac{n_{1\cdot}n_{\cdot j}}{N}$	$\hat{E}_{1c} = NP_{1c} = \frac{n_{1\cdot}n_{\cdot c}}{N}$	$\hat{p}_{1\cdot} = \frac{n_{1\cdot}}{N}$
		n_{21}	n_{22}	n_{2j}	n_{2c}	
		$P_{21} = \hat{p}_{2\cdot}\hat{p}_{\cdot 1}$	$P_{22} = \hat{p}_{2\cdot}\hat{p}_{\cdot 2}$	$P_{2j} = \hat{p}_{2\cdot}\hat{p}_{\cdot j}$	$P_{2c} = \hat{p}_{2\cdot}\hat{p}_{\cdot c}$	$n_{2\cdot}$

	2	$\hat{E}_{21}=NP_{21}=\frac{n_{2\cdot}n_{\cdot1}}{N}$	$\hat{E}_{22}=NP_{22}=\frac{n_{2\cdot}n_{\cdot2}}{N}$	$\hat{E}_{2j}=NP_{2j}=\frac{n_{2\cdot}n_{\cdot j}}{N}$	$\hat{E}_{2c}=NP_{2c}=\frac{n_{2\cdot}n_{\cdot c}}{N}$	$\hat{p}_{2\cdot}=\frac{n_{2\cdot}}{N}$
	i	n_{i1}	n_{i2}	n_{ij}	n_{ic}	
		$P_{i1}=\hat{p}_{i\cdot}\hat{p}_{\cdot1}$	$P_{i2}=\hat{p}_{i\cdot}\hat{p}_{\cdot2}$	$P_{ij}=\hat{p}_{i\cdot}\hat{p}_{\cdot j}$	$P_{ic}=\hat{p}_{i\cdot}\hat{p}_{\cdot c}$	$n_{i\cdot}$
		$\hat{E}_{i1}=NP_{i1}=\frac{n_{i\cdot}n_{\cdot1}}{N}$	$\hat{E}_{i2}=NP_{i2}=\frac{n_{i\cdot}n_{\cdot2}}{N}$	$\hat{E}_{ij}=NP_{ij}=\frac{n_{i\cdot}n_{\cdot j}}{N}$	$\hat{E}_{ic}=NP_{ic}=\frac{n_{i\cdot}n_{\cdot c}}{N}$	$\hat{p}_{i\cdot}=\frac{n_{i\cdot}}{N}$
	r	n_{r1}	n_{r2}	n_{rj}	n_{rc}	
		$P_{r1}=\hat{p}_{r\cdot}\hat{p}_{\cdot1}$	$P_{r2}=\hat{p}_{r\cdot}\hat{p}_{\cdot2}$	$P_{rj}=\hat{p}_{r\cdot}\hat{p}_{\cdot j}$	$P_{rc}=\hat{p}_{r\cdot}\hat{p}_{\cdot c}$	$n_{r\cdot}$
		$\hat{E}_{r1}=NP_{r1}=\frac{n_{r\cdot}n_{\cdot1}}{N}$	$\hat{E}_{r2}=NP_{r2}=\frac{n_{r\cdot}n_{\cdot2}}{N}$	$\hat{E}_{rj}=NP_{rj}=\frac{n_{r\cdot}n_{\cdot j}}{N}$	$\hat{E}_{rc}=NP_{rc}=\frac{n_{r\cdot}n_{\cdot c}}{N}$	$\hat{p}_{r\cdot}=\frac{n_{r\cdot}}{N}$
总计（列）		$n_{\cdot1}$	$n_{\cdot2}$	$n_{\cdot j}$	$n_{\cdot c}$	$n_{\cdot\cdot}=N$
		$\hat{p}_{\cdot1}=\frac{n_{\cdot1}}{N}$	$\hat{p}_{\cdot2}=\frac{n_{\cdot2}}{N}$	$\hat{p}_{\cdot j}=\frac{n_{\cdot j}}{N}$	$\hat{p}_{\cdot c}=\frac{n_{\cdot c}}{N}$	

三、卡方检验

由前面分析可知，当两定性变量相互独立时，$\hat{E}_{ij}-n_{ij}$ 应为随机误差；当两定性变量具有连带关系时，$\hat{E}_{ij}-n_{ij}$ 将会有一个大的差异出现。因此 $\left(\hat{E}_{ij}-n_{ij}\right)^2$ 中包含这两定性变量的

独立与连带的信息。

K·Pearson（1904）首先构造了检验独立性的统计量：

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{\left(\hat{E}_{ij} - n_{ij}\right)^2}{\hat{E}_{ij}} \tag{1-10}$$

当理论频数都不太小且N充分大时，统计量χ^2在两定性变量相互独立的假设下近似服从自由度$n' = (r-1)(c-1)$的卡方分布。

两个定性变量A、B的独立性检验方法如下：

H_0：$P_{ij} = P_{i\cdot}P_{\cdot j}$（即A与B相互独立）

选择检验统计量：

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{\left(n_{ij} - \frac{n_{i\cdot}n_{\cdot j}}{N}\right)^2}{\frac{n_{i\cdot}n_{\cdot j}}{N}} \tag{1-11}$$

依据自由度n'和预先给定的显著性水平α，查卡方分布表找出临界值$\chi^2_\alpha(n')$。

若$\chi^2 > \chi^2_\alpha(n')$，则否定$H_0$，即在$\alpha$水平上可信A与B有显著地连带关系；若$\chi^2 < \chi^2_\alpha(n')$，则接受$H_0$，即在$\alpha$水平上可信A与B相互独立。

以表1-3中的数据作为本节的数值例子。所需检验的假设H_0：吸烟量与年龄无关。

计算χ^2的第一步是用公式（1-9）计算理论频数。例如$\hat{E}_{11}$由下式算出：

$$\hat{E}_{11} = \frac{65 \times 60}{100} = 39$$

将其余计算程式编入表1-5。表中最末一行的和便是χ^2值。

表 1-5　χ^2 值计算表

n_{ij}	50	15	10	25	
$\hat{E}_{ij}$	39	26	21	14	
$(n_{ij}-\hat{E}_{ij})^2$	121	121	121	121	χ^2
$(n_{ij}-\hat{E}_{ij})^2/\hat{E}_{ij}$	3.10	4.65	5.76	8.64	22.15

本例子中 $r=c=2$，则自由度 $n'=(2-1)(2-1)=1$。如果显著性水平取 0.05，即 $\alpha=0.05$。从卡方分布表查得：$\chi^2_{0.05}(1)=3.84$。由于计算出的 χ^2 值远大于这个值，于是可以得出结论：在 0.05 水平上可信吸烟量与年龄有显著的某种连带关系。事实上，表 1-3 中反映出的每日吸烟 20 支以上 60 岁以上人的相应比例数（50/60=0.833）明显不同于 60 岁以下人的相应比例数（15/40=0.375）。

应该指出，借助于卡方检验所发现的显著连带，未必就意味着所涉及的变量间存在着任何因果关系；然而，它又确实提示这种连带的原委是值得研究的。

第二节　2×2 列联表

由两个二值变量所产生的列联表是最简单的，称为 2×2 列联表，已在前面介绍过。本节将详细研究这种列联表。

2×2 列联表，在社会科学、体育教学、医学研究等领域应用的都比较广泛。表中的数据可由不同方式产生，如从某总体中抽取 N 个受试对象，按照两个二值变量对个体进行分类，即可以产生这样的数据。前述用以研究吸烟量与年龄关系的表 1-3 中的数据，就是按这一方