⊙高等学校统计学类系列教材

非参数统计分析

王静龙　梁小筠　编著

高等教育出版社
HIGHER EDUCATION PRESS

高等学校统计学类系列教材

非参数统计分析

王静龙　梁小筠　编著

高等教育出版社

内容提要

本书共分九章。第一章阐述非参数统计的特点。第二章讲解描述性统计方法。第三、四章分别讲解符号检验法、符号秩和检验法，第五到第八章分别讲解两样本、多样本和区组设计等问题以及相关分析的非参数统计方法。第九章讲解检验的渐近相对效率。本书结合统计软件和案例分析讲解非参数统计的理论与方法，可以作为统计专业的教学用书，也适宜于社会学、心理学、人口学、医学、教育学及生物学等学科的研究和应用的人士阅读，作为这些学科的教学用书。

图书在版编目(CIP)数据

非参数统计分析/王静龙，梁小筠编著. —北京：高等教育出版社，2006.4
ISBN 7-04-018682-9

Ⅰ.非… Ⅱ.①王…②梁… Ⅲ.非参数统计-统计分析(数学)-高等学校-教材 Ⅳ.O212.7

中国版本图书馆 CIP 数据核字(2006)第 014461 号

策划编辑 李 蕊 **责任编辑** 张耀明 **封面设计** 王凌波 **责任绘图** 黄建英
版式设计 王艳红 **责任校对** 金 辉 **责任印制** 孔 源

出版发行	高等教育出版社	**购书热线**	010-58581118
社 址	北京市西城区德外大街4号	**免费咨询**	800-810-0598
邮政编码	100011	**网 址**	http://www.hep.edu.cn
总 机	010-58581000		http://www.hep.com.cn
		网上订购	http://www.landraco.com
经 销	蓝色畅想图书发行有限公司		http://www.landraco.com.cn
印 刷	北京新丰印刷厂	**畅想教育**	http://www.widedu.com
开 本	787×960 1/16	**版 次**	2006年4月第1版
印 张	13.75	**印 次**	2006年4月第1次印刷
字 数	250 000	**定 价**	17.60元

物料号 18682-00

前　言

非参数统计是统计学的一个重要分支，它与总体究竟是怎样的分布几乎没有什么关系，仅需要一些非常一般性（例如连续分布、对称分布等）的假设，进行统计推断时只利用样本观察值中一些非常直观（例如次序）的信息，所以非参数统计分析含有丰富的统计思想并在实践中有着广泛的应用。

1983 年陈希孺院士来我系讲非参数统计。他的讲学使我们领略到非参数统计的魅力，了解了非参数统计产生的背景，它不仅理论丰富、方法巧妙而且应用范围很广。1993 年起系里年年都为本科生开设非参数统计这一门课。我们采用陈希孺院士和柴根象教授所写的《非参数统计教程》与吴喜之和王兆军两位教授所写的《非参数统计方法》作为教学用书。在教学的过程中，我们陆续编写了各个章节的讲义，试图综合这两本书的特点，包含我们在教学过程中处理教材的一些做法，增加一些新近发展的学生不难理解并富有实用价值的内容。本书就是将修改多次的讲义整理加工而成的。

本书共有九章。第一章引言通过实例使得学生了解非参数统计的特点，激发他们学习非参数统计的兴趣。使用非参数统计方法处理数据的第一步工作是描述性统计。第二章将结合 Excel 统计软件讲解描述性统计中的表格法、图形法和数值方法。接下来我们根据数据结构由简单到复杂的原则安排教学内容。第三章的符号检验最为简单直观。由单样本的符号检验到成对数据的符号检验，然后引出符号秩和检验。第四章的符号秩和检验处理成对数据。很自然地，由成对数据到第五章的两样本问题，由两样本到第六章的多样本问题，再由多样本到第七章的区组设计问题。无论是两样本、多样本还是区组设计，都假设观察值相互独立。第八章的相关分析研究数据独立还是相依的问题。随着计算机技术的发展与统计软件的开发，使得复杂的统计运算变得简单可行。本书结合计算机，使用 Minitab、SAS 和 Excel 等统计软件讲解非参数统计方法。渐近相对效率问题是非参数统计的一个理论问题。本书的第九章检验的渐近相对效率试图用学生能理解的方式分别具体计算出 t 检验对符号检验与对秩和检验的渐近相对效率。本书的某一些章节标有星号，初学或若教学时间紧可以跳过去，这并不影响全书的连贯。本书收集、编写了大量的例子，其中有一些案例。它们反映了非参数统计数据分析方法应用的很多方面的问题。本书各章都附有习题，标有星号的习题比较难。

本书力求结合实际例子讲清楚非参数统计方法的直观意义、来龙去脉、适用于哪一类问题以及其证明的思路。有的证明放在本书的附录中，有的证明请学生参阅有关的参考书目自己完成。

我们建议，教学时数的分配如下：每周3学时（其中1学时为上机实验），教学周20周含复习考试1周。

章次	一	二	三	四	五	六	七	八	九	复习考试
学时	3	3	3	6	12	9	9	9	3	3

本书除了作为大学统计专业的教学用书外，还可以作为从事理论研究和应用的统计工作者、教师和学生的参考用书，此外，本书也适宜于进行社会学、心理学、人口学、医学、教育学及生物学等学科的研究和应用的人士阅读，也可以作为这些学科的教学用书。

感谢陈希孺院士，本书的完稿得益于他1983年来我系的讲学和他写的有关非参数统计的著作，得益于他对我们关心、指导和教诲。谨以此书纪念陈希孺院士。

感谢汪嘉冈、柴根象、孙山泽、吴喜之和王兆军等教授，阅读他们写的书以及与他们的讨论使我们得益匪浅。

感谢茆诗松教授，他在百忙之中审阅了书稿，提出了很多宝贵的意见，并推荐书稿早日出版。我们也要感谢华东师范大学统计系的历届学生，因为有他们的参与，我们在教学中对所涉及的内容越来越有体会，享受着极大的乐趣。如果没有他们的参与，本书难以成稿。最后，要感谢高等教育出版社的李蕊女士。没有他们的关心、支持和辛勤劳动，本书不可能很快出版。

王静龙、梁小筠于华东师范大学统计系

2005年10月

郑重声明

目　录

第一章

引　言

如果根据实践经验，人们知道产生数据的总体具有某种分布（例如正态分布），只是其中的几个参数的值未知（如正态分布的均值或方差未知，或均值和方差都未知），那么在这种类型假设条件下的数据分析方法称为参数型的. 一般的统计教学主要讲授参数数据分析方法及其原理. 参数数据分析方法既利用数据的信息，又利用产生数据的总体的信息，所以它是一个很有效的数据分析方法. 但是在实践中可能发生这样的情况，人们没有足够的实践经验，或者情况比较特殊，难以假设总体具有某种分布. 这时如果仍然使用参数数据分析方法，其统计分析的结果显然是不可信的，甚至有可能是错的. 看下面的两个例子.

例 1.1　供应商供应的产品是否合格？

某工厂产品的零件由某个供应商供应. 合格零件标准长度为 (8.5 ± 0.1) cm. 这也就是说合格零件长度的中心位置为 8.5 cm，允许误差界为 0.1 cm，即长度在 8.4 ~ 8.6 cm 之间的零件是合格的. 为评估近来供应的零件是否合格，随机抽查了 $n = 100$ 个零件. 它们的长度数据 $x_1, x_2, \cdots, x_n$ 如表 1.1.

表 1.1　100 个零件的长度

8.503	8.508	8.498	8.374	8.494	8.500	8.498	8.500	8.502	8.501	8.491	8.504
8.502	8.503	8.501	8.505	8.492	8.497	8.150	8.496	8.501	8.489	8.506	8.497
8.505	8.501	8.500	8.499	8.490	8.493	8.501	8.497	8.501	8.498	8.503	8.505
8.510	8.499	8.489	8.496	8.500	8.503	8.497	8.504	8.503	8.506	8.497	8.507
8.346	8.310	8.489	8.499	8.492	8.497	8.506	8.502	8.505	8.489	8.503	8.492
8.501	8.499	8.804	8.505	8.504	8.499	8.506	8.499	8.493	8.494	8.490	8.505
8.511	8.502	8.505	8.503	8.782	8.502	8.509	8.499	8.498	8.493	8.897	8.504
8.493	8.494	7.780	8.509	8.499	8.503	8.494	8.511	8.501	8.497	8.493	8.501
8.495	8.461	8.504	8.691								

经计算，平均长度为 $\bar{x} = 8.495\ 8$ cm，非常接近所要求的中心位置 8.5 cm. 样本标准差为 $s = \sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2/(n-1)} = 0.104\ 7$ cm. 一般来说，产品的质量指标往

往服从正态分布.因此,工厂使用参数数据分析方法来评估零件是否合格.假设供应商供应的零件长度 X 服从正态分布:$X \sim N(\mu,\sigma^2)$,则零件合格的可能性近似等于

$$\begin{aligned}P(8.4 \leqslant X \leqslant 8.6) &= \Phi((8.6-\mu)/\sigma) - \Phi((8.4-\mu)/\sigma)\\ &\approx \Phi((8.6-8.495\,8)/0.104\,7) - \Phi((8.4-8.495\,8)/0.104\,7)\\ &\approx 66\%,\end{aligned}$$

所以约有三分之一的零件不合格.由于样本标准差 0.104 7 cm 非常接近允许误差界 0.1 cm,我们也可以根据正态分布 3σ 原则:

$$P(8.4 \leqslant X \leqslant 8.6) \approx P(\mu-\sigma \leqslant X \leqslant \mu+\sigma) = 68.26\%,$$

知道大约只有三分之二的零件合格.看来该工厂需要换另一个供应商了.

这个统计分析的结论和数据不相吻合.事实上,在这 100 个样本零件中有 91 个零件的长度在 8.4 ~ 8.6 cm 之间,所以零件合格的比例为 91%,超过三分之二很多.考虑到这家供应商的价格不高,它的近 10% 的不合格率是允许的,该工厂不需要换供应商.上面的统计分析的结论之所以和数据不相吻合,就是由于假设供应商供应的零件长度 X 服从正态分布.在使用参数数据分析方法之前,最好对数据进行描述性分析,比如列出频率分布表和画直方图.100 个样本数据的频率分布表和直方图分别如表 1.2 和图 1.1.

表 1.2　100 个零件长度的分布表

长度(cm)	频率(%)
~8.40	5
8.40 ~ 8.46	0
8.46 ~ 8.48	1
8.48 ~ 8.50	45
8.50 ~ 8.52	45
8.52 ~ 8.60	0
8.60 ~	4
合计	100

频率分布表和直方图告诉我们,将供应商供应的零件长度 X 的分布假设为正态分布是不合理的.由描述性分析可以看到,90% 的零件长度在 (8.5 ± 0.02) cm 之间,误差只有 0.02 cm,它们是相当不错的.有 9% 的零件不合格,而且它们的长度离标准长度的中心位置很远,这很可能是因为少数几名职工操作不当,或者由于生产工具失灵的缘故.这些问题工厂应该向供应商指出.看来该工厂不需要

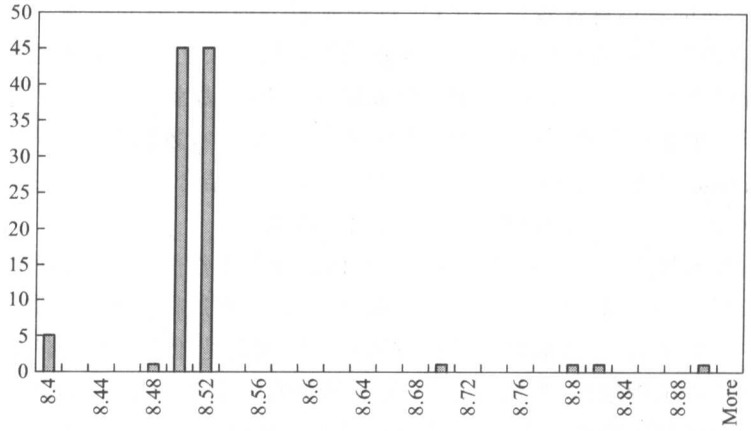

图 1.1　100 个零件长度的直方图

换另一个供应商.

注:本书的频率(或频数)分布表的每一个分组都是左开右闭,它们都只包含右端的值,而不包含左端的值. 这里的 ~8.40,8.40 ~ 8.46,…,8.60 ~ 分别表示 $(-\infty, 8.40]$,$(8.40, 8.46]$,…,$(8.60, +\infty)$.

例 1.2　哪一个企业职工的工资高?

这里有 22 名职工,其中的 12 名职工来自企业 1,另外的 10 名职工来自企业 2. 他们的工资(单位:千元)如表 1.3.

表 1.3　两个企业职工的工资

企业 1	11	12	13	14	15	16	17	18	19	20	40	60
企业 2	3	4	5	6	7	8	9	10	30	50		

显然,企业 1 职工的工资高. 倘若假设企业 1 和企业 2 职工的工资分别服从正态分布 $N(a,\sigma^2)$ 和 $N(b,\sigma^2)$,则这两个企业职工工资的比较问题就转化为一个参数的假设检验问题,其原假设为 $H_0: a = b$,备择假设为 $H_1: a > b$. 由于两样本 t 检验统计量的值 $t = 1.282$,显著性水平 $\alpha = 0.05$ 的临界值为 $t_{0.95}(20) = 1.725$,所以我们不能拒绝原假设,因而认为这两个企业职工的工资没有差异. 即使取显著性水平 $\alpha = 0.10$,由于临界值 $t_{0.90}(20) = 1.325$,我们仍然不能拒绝原假设. 我们也可以通过计算 p 值完成检验的程序. p 值等于自由度为 20 的 t 分布随机变量大于等于检验统计量的值 1.282 的概率,所以检验的 p 值为 $P(t(20) \geqslant 1.282) = 0.107\,3$. 由于 p 值不小,因而我们仍不能拒绝原假设. 这个统计分析的结论显然和数据不相吻合. 之所以有问题,就是因为假设职工的工资服从正态分布的缘故. 一般来说,工资、收入等的分布是不对称的,并且由于有一部分人的收入比较

高，所以分布的右边有比较长的尾巴. 将企业职工的工资的分布假设为正态分布，并且在正态分布假设条件下使用参数数据分析方法是不恰当的. 这时通常有以下两种处理方法. 一是仍想使用参数数据分析方法，那就必须对企业职工工资分布给予一个非正态的、其他类型分布的假设. 二是非参数数据分析方法. 当然，还有另外的处理方法，例如将参数方法和非参数方法结合起来，这就是所谓的半参数数据分析方法. 半参数数据分析方法本书从略.

非参数数据分析方法对产生数据的总体的分布不作假设，或仅给出很一般，例如连续型分布、对称分布等这样一些简单的假设. 在这样条件下的数据分析方法称为非参数型的. 非参数数据分析方法与总体究竟是什么分布没有什么关系，所以它的应用范围很广. 当然，如果关于总体分布有很多的信息，以至于能够给出总体分布的具体数学形式，只是其中的几个参数值未知，基于这种类型分布的参数数据分析方法，一般来说较非参数数据分析方法好. 人们很可能会就此认为，既然非参数数据分析方法应用范围广，它的效率势必不高. 其实不然. 本书第九章将正态总体假设下的 t 检验方法与相应的非参数方法进行比较. 从中大家可以看到，即使在正态总体时使用 t 检验方法的效率也仅比相应的非参数方法略微高一些，而在非正态总体假设时使用 t 检验方法的效率有的还比相应的非参数方法低. 从这个意义上讲，非参数数据分析方法犹如所谓的广谱类抗生素. 广谱是指这类抗生素对很多类病菌均有作用. 阿莫西林克拉维酸钾片（商品名：君尔清）是广谱类抗生素. 下面摘录它的说明书《药理毒理》栏目中的一段话.

本品为阿莫西林和克拉维酸钾的复方制剂. 阿莫西林为广谱青霉素类抗生素，克拉维酸钾本身只有微弱的抗菌活性，但具有强大的广谱 β - 内酰胺酶抑制作用，两者合用，可保护阿莫西林免遭 β - 内酰胺酶水解. 本品的抗菌谱与阿莫西林相同，且有所扩大. 对产酶金黄色葡萄球菌、表皮葡萄球菌、凝固酶阴性葡萄球菌及肠球菌均具良好作用，对某些产 β - 内酰胺酶的肠杆菌科细菌、流感嗜血杆菌、卡他莫拉菌、脆弱拟杆菌等也有较好的抗菌活性.

阿莫西林克拉维酸钾片用于很多种感染的治疗. 引起感染例如呼吸道感染的病菌有好几个类别. 要检验出究竟是哪一类病菌引起感染通常不是一件轻而易举立即就可解决的事. 在不能确切肯定引起感染是什么病菌的情况下，为了及时治好感染，医生往往倾向于选用广谱类抗生素例如阿莫西林克拉维酸钾片等. 广谱类抗生素应用范围广，且其治疗效果也不错. 在某种程度上，用非参数统计方法去分析数据犹如用广谱类抗生素去治病.

参数和非参数数据分析方法是针对不同情况提出来的两种统计分析方法，它们各有其优缺点，是互为补充的. 本书着重介绍非参数数据分析方法.

有关例 1.2 这个问题的非参数数据分析方法的讨论见第五章.

习 题 一

由26个被调查者组成的样本提供了每周看电视和读书的时间(单位:小时),数据列于下表.取水平0.05,检验原假设:看电视和读书的时间一样多,备择假设:看电视的时间比读书的时间多.这是成对数据的比较问题,应算出看电视的时间与读书的时间的差值.差值也列于下表内.

被调查者	看电视	读书	差值	被调查者	看电视	读书	差值
1	10	8	2	14	19	16	3
2	14	10	4	15	10	8	2
3	4	17	-13	16	17	2	15
4	6	7	-1	17	10	6	4
5	12	14	-2	18	12	4	8
6	13	12	1	19	7	10	-3
7	14	10	4	20	19	3	16
8	13	11	2	21	12	11	1
9	10	5	5	22	11	7	4
10	14	9	5	23	2	25	-23
11	9	9	0	24	9	10	-1
12	12	8	4	25	8	6	2
13	4	18	-14	26	16	5	11

试用单样本 t 检验法检验上述假设.你认为该检验结果可信吗?

第二章

描述性统计

描述性统计是在对产生数据的总体的分布不作任何假设的情况下，整理数据、显示数据和分析数据，将数据中有用的信息提取出来的统计方法. 常用的描述性统计方法有表格法、图形法和数值方法.

§2.1 表格法和图形法

表格法主要有列频数分布表和频率分布表. 看下面的例子.

例 2.1 某电子公司测试新灯丝的燃烧寿命，表 2.1 列出了 200 个灯泡样本的可使用小时数.

表 2.1 灯丝寿命数据

107	73	68	97	76	79	94	59	98	57	73	81	54	65	71	80	84	88	62	61
79	98	63	65	66	62	79	86	68	74	61	82	65	98	63	71	62	116	65	88
64	79	78	79	77	86	89	76	74	85	73	80	68	78	89	72	58	69	82	72
92	78	88	77	103	88	63	68	88	81	64	73	75	90	62	89	71	71	74	70
74	70	85	61	65	81	75	62	94	71	85	84	83	63	92	68	81	62	79	83
93	61	65	62	92	65	64	66	83	70	70	81	77	72	84	67	59	58	73	83
78	66	66	94	77	63	66	75	68	76	73	76	90	78	71	101	78	43	59	67
61	71	77	91	96	75	64	76	72	77	74	65	82	86	79	74	66	86	96	89
81	71	85	99	59	92	94	62	68	72	77	60	87	84	75	77	51	45	63	102
85	67	87	80	84	93	69	76	89	75	59	77	83	68	72	67	92	89	82	96

这 200 个数杂乱无章，不经过整理难以发现其内在的规律. 首先构造它的频数分布表和频率分布表，步骤如下：

① 找出最小值 43 和最大值 116；

② 将数据分组. 一般取组数:5 ~ 20,组距 ≈ (最大值 - 最小值)/组数. 200 个灯泡的灯丝寿命数据分为 16 组,组距为 5；

③ 手工或使用软件(例如 Excel,使用 Excel 作分布表和直方图的方法将在下面介绍)列出 200 个灯泡的灯丝寿命数据的频数分布表和频率分布表(如表 2.2).

表 2.2 灯丝寿命数据频率分布表

灯丝寿命(小时)	个数	频率(%)
40—44	1	0.5
45—49	1	0.5
50—54	2	1.0
55—59	8	4.0
60—64	24	12.0
65—69	28	14.0
70—74	30	15.0
75—79	34	17.0
80—84	23	11.5
85—89	22	11.0
90—94	14	7.0
95—99	8	4.0
100—104	3	1.5
105—109	1	0.5
110—114	0	0.0
115—119	1	0.5
总和	200	100.0

由分布表可以看到,灯丝寿命很长和很短的都是少数,多数是中间状态,从中间往两头看,寿命长和短的比例基本相等. 这种趋势在直方图中更明显地显示出来. 200 个灯泡的灯丝寿命数据的直方图见图 2.1.

其纵轴的单位是“个”. 直方图钟形对称. 所谓钟形对称,就是直方图“中间高,两头低,左右近似对称”.

下面介绍利用 Excel 作频数分布表和直方图的步骤.