高等学校教材

非参数统计方法

吴喜之　王兆军

高等教育出版社

高 等 学 校 教 材

非参数统计方法

吴喜之　　王兆军

高 等 教 育 出 版 社

（京）112号

内容提要

本书主要介绍非参数统计方法的预备知识、对称中心及位置参数的检验、区组设计的分析、秩相关分析、一般的秩检验理论、拟合优度检验、概率密度估计和非参数回归等内容，可供高等学校概率统计专业作为选修课教材使用，也可供实际工作者自学或参考。

图书在版编目（CIP）数据

非参数统计方法/吴喜之，王兆军著．-北京：高等教育出版社，1996
ISBN 7-04-005795-6

Ⅰ．非… Ⅱ．①吴…②王… Ⅲ．非参数统计 Ⅳ．O212.7

中国版本图书馆CIP数据核字（96）第14431号

高等教育出版社出版
北京沙滩后街55号
邮政编码：100009　传真：64014048　电话：64054588
新华书店总店北京发行所发行
北京市顺新印刷厂印装
*
开本 787×1092　1/16　印张 15.5　字数 380 000
1996年9月第1版　1996年12月第1次印刷
印数 0001—1 087
定价 12.00元

凡购买高等教育出版社的图书，如有缺页、倒页、脱页等质量问题者，请与当地图书销售部门联系调换

序

非参数统计是数理统计学的一个分支，它形成于本世纪40年代.第二次世界大战之后，它得到了充分的研究和发展，现已成为一个实用价值很高的方向，为此有条件的高校已开设了非参数统计方法课程.为了满足教学及实际工作者的需要，我们根据1991年国家教委高等学校理科数学与力学教学指导委员会概率论与数理统计教材建设组提出的内容和要求，编写了这本书，供与统计有关各方向的师生使用，也可供实际工作者自学和查阅参考.

本书在实际例子的背景下，叙述了非参数统计的基本理论和方法.选材方面，基于我们6年来教此课的经验并参考了国内外有关的书籍及文献.叙述时，我们力图直观和简单易懂，注重方法产生的思想及实际应用，对某些较繁杂的定理，只给出了结论，而不加证明.如对精深理论感兴趣的读者，可参看每章最后的阅读知识一节.阅读本书的读者仅需要了解概率论与数理统计的基本内容即可.

全书共分7大部分：预备知识；对称中心及位置参数的检验；区组设计的分析；秩相关分析；一般的秩检验理论；拟合优度检验；概率密度估计和非参数回归.为了加深理解所学的内容，每章后都有一定数量的习题，其中有些习题是正文的补充，而且有相当一部分是联系实际的数据题.由于学时所限及实际工作者的需要，本书不讨论收敛速度问题及有关概率密度估计和非参数回归的大样本性质、条件检验和置换检验.

本书收集了众多的非参数统计方法，根据实际工作需要可以直接应用这些方法，而不必理会有关的理论部分.

受国家教委高等教育出版社的委托，中山大学邓永录、杨维权教授主审了此书，参加审稿的还有邓炜材、汤尚勇、高尚华、张润楚、史道济、邓集贤等教授.另外，高等教育出版社的高尚华老师为此书的早日出版费了许多心血，作者在此一并表示衷心的感谢！

由于作者的水平及时间所限，书中错误在所难免，请同行专家和读者批评指正.

作者

1995-07-10

目　录

第一章 引 言

1.1 绪 论

什么是统计？统计就是收集及分析数据，并由此作出推断的科学. 统计要从数据出发建立模型，这叫归纳(induction)；建立模型之后，要用它来进行推断，这叫演绎(deduction). 和以演绎为主并基于公理系统的数学不一样，统计是基于数据的，其数学基础是概率论. 由于现实世界的多样性，在统计中不存在完美的模型，任何一个由数据归纳出来的模型往往要再回到实际中对其检验，并用新的数据对之进行修正. 这种反复的认识及再认识的思想方法是统计的一个突出特点. 数学是一个可以独立存在的逻辑体系. 而对于统计来说，离开了应用，就没有存在的必要.

一般经典的数理统计教科书的主要部分是由估计和检验两大部分组成. 在那里，往往假设产生数据的总体分布的形式是已知的. 所不能确定的是数量有限的一些参数值，而所要做的就是对这些参数进行检验或估计. 但是实践中，在没有足够证据时，去假设一个总体有某种分布形式，并进行参数估计或检验是不负责的，结果是不可靠的，甚至是灾难性的.

非参数统计就是在对总体分布形式不了解时进行推断的统计方法. 这里对于总体分布不作或只作一点诸如对称性之类的简单假设. 虽然不知道分布的形式，我们总可以把数据按大小排队而使每个数据都有自己的“地位”，我们称之为秩(rank). 大小为 n 的样本产生了 n 个秩. 这样，问题就简化为对这些秩的研究了. 幸运的是，这些秩及由其产生的一些统计量的性质和分布是可以得到的，并且与原来的总体分布无关(distribution-free). 除了与秩有关的方法之外，还有其它一些非参数方法. 非参数方法有相当好的稳健性(后面要介绍)，计算简单，处理问题广泛，并且在多数分布未知的情况下比参数方法更有效. 但也应指出：虽然参数方法有局限性，但在总体分布已知时，它比非参数方法利用更多的样本中的信息，因而就更有效.

本章介绍一些为学习后面章节所需要的基本的统计和概率知识. 如已熟悉，可略过不看. 第四节之后的部分最好在用到时再看. 一些概念，如完全估计量和相容估计量等对初学者或非数理统计方向的读者也可略去不看.

1.2 估计和检验

1.2.1 点估计和区间估计

假定我们掷一枚硬币 n 次，得到 S 次正面，需要估计出现正面的概率 p. 由直观，我们可用

$\frac{S}{n}$ 来估计 p. 当然，你可以用任何可以想象的其它方法来估计 p. 这样就产生了一个评价估计好坏的标准问题. 所谓“好”和“坏”，其实只是相对于你的要求而言. 在数理统计课程中已引进了各种不同的标准. 本节仅就本书中要用到的标准作一回顾.

要估计上述的概率 p，绝不能只掷一次硬币. 我们希望在大量试验中，估计量的平均值尽可能地接近所要估计的真值. 这就产生了无偏估计量(unbiased estimator) 的概念. 假设有样本 $X_1, X_2, \cdots, X_n$. 它们的总体分布(函数) 为 $F(x,\theta)$，而 θ 为要估计的参数. 如果我们选定的对 θ 的估计量是 $T(X_1, \cdots, X_n)$(注意，它是样本数据 $X_1, \cdots, X_n$ 的一个函数或统计量，与参数 θ 无关)，在满足

$$E_\theta(T(X_1, \cdots, X_n)) = \theta$$

时，我们称 $T \equiv T(X_1, \cdots, X_n)$ 为 θ 的一个无偏估计量，这里 $E_\theta(\cdot)$ 表示基于 $F(x,\theta)$ 的期望.

我们可以把掷硬币看成是 n 个独立的 Bernoulli 试验，即 S 服从二项分布：$S \sim b(n,p)$. 所以有

$$E\left(\frac{S}{n}\right) = p$$

也就是说，刚才选的对 p 的估计 $\frac{S}{n}$ 是无偏的. 注意，无偏估计可能不唯一，当然和任何其它种类的估计一样，它有它的缺点. 如果有两个统计量 T_1 和 T_2 为参数 θ 的无偏估计，我们自然要选择其方差小的，因为方差越小，统计量的可能值的分散程度越小. 一般来说，我们希望均方误差 $E(T-\theta)^2$ 越小越好. 如果在所有无偏估计中，估计量 T 使均方误差(对无偏估计，这就是方差) 最小，则称 T 为一致最小方差无偏估计(uniformly minimum variance unbiased estimator——UMVUE).

在用一个统计量 $T(X_1, \cdots, X_n)$ 估计参数 θ 时，我们当然要求这个统计量要尽量用到样本中的全部信息，在统计上，称这种统计量为充分的. 确切地说，如果在给定 $T(X_1, \cdots, X_n) = t$ 下，$(X_1, \cdots, X_n)$ 的条件分布与 t 无关，则称 $T(X_1, \cdots, X_n)$ 是分布族 $\{F(x,\theta): \theta \in \Theta\}$ 的充分统计量.

既然 UMVUE 是参数 θ 的一个好的估计，那么它是不是唯一的？为了解决这一问题，又引进了统计上另一个重要的概念 —— 完全统计量. 确切地说，对于分布族 $\{F(x,\theta): \theta \in \Theta\}$，如任给满足

$$\mathrm{E}_\theta g(T) = 0, \quad \forall\, \theta \in \Theta$$

的函数 $g(\cdot)$，都有 $P_\theta(g(T) = 0) = 1$，则称统计量 $T(X_1, \cdots, X_n)$ 的导出分布族是完全的.

在 Bernoulli 试验中，因为

$$\lim_{n\to\infty} \mathrm{Var}\left(\frac{S}{n}\right) = \lim_{n\to\infty} E\left(\frac{S}{n} - p\right)^2 = 0$$

则对任意的 $\varepsilon > 0$ 有

$$\lim_{n\to\infty} P\left(\left|\frac{S}{n} - p\right| > \varepsilon\right) = 0$$

直观上，随着试验次数 n 的增加，估计值 $\frac{S}{n}$ 与实际的参数值 p 应更接近. 一般来说，如对任意 $\varepsilon > 0$，参数 θ 的估计量 $T(X_1, \cdots, X_n)$ 满足

$$\lim_{n\to\infty} P(|T(X_1, \cdots, X_n) - \theta| > \varepsilon) = 0$$

则称 $T(X_1,\cdots,X_n)$ 为 θ 的相容(或相合)估计量(consistent estimator). 注意,相容性是一个大样本性质,在固定的小样本情况,应谨慎对待. 有时,一个相容统计量会没有任何实际意义.

如果取 $T=T(X_1,\cdots,X_n)$ 作为 θ 的一个估计,我们能用它来估计 θ 的一个可能的范围或其可能的上下界,一个常用的范围的形式为 $T(X_1,\cdots,X_n)\pm a$. 当然,因为 T 是个随机变量,所以我们只能说由它导出的区间(置信区间)以某概率(置信度)覆盖参数 θ. 一般地说,如果 $[T_l,T_u]$ 是由一对统计量 $T_l,T_u(T_l\leqslant T_u)$ 所组成的随机区间,如对所有的 θ 有

$$P_\theta(T_l\leqslant\theta\leqslant T_u)=1-\alpha$$

则称 $[T_l,T_u]$ 为 θ 的置信度为 $1-\alpha$ 的置信区间(confidence interval). 这里 $P_\theta(\cdot)$ 表示当 θ 为真实参数值时的概率. 换言之,我们以 $100(1-\alpha)\%$ 的概率或置信度(confidence level)保证 $[T_l,T_u]$ 覆盖 θ.

1.2.2 假设检验

如果在上面掷硬币的试验中,我们怀疑硬币的均匀性,即怀疑是否 $p=\frac{1}{2}$. 我们就要对原假设(null hypothesis)$H_0:p=\frac{1}{2}$ 进行检验. 备择假设(alternative hypothesis)可为 $p\neq\frac{1}{2}$,$p<\frac{1}{2}$ 及 $p>\frac{1}{2}$ 三者之一. 如果备择假设用 $p\neq\frac{1}{2}$,则称检验是双边的. 如备择假设用另外两个之一,则称检验是单边的.

对原假设进行检验的结果只能是下列两个决策之一:1. 拒绝原假设 H_0;2. 不能拒绝原假设 H_0. 有些作者用"接受备择假设"来代替第 2 个决策,这是不对的. 因为在检验中,我们一直在原假设条件下进行概率运算,在原假设不对时,没有任何理由来"接受"备择假设. 我们尊重他人基于历史原因的选词,但为了科学的准确性及避免逻辑混乱,我们不主张用"接受备择假设"的说法.

上面的原假设只包含一个点,称为简单假设(simple hypothesis). 一般地,假定 Θ 为所有可能的参数值 θ 的集合. 原假设为 $\theta\in\Theta_0$,备择假设为 $\theta\in\Theta_1$. 而 $\Theta_0\subset\Theta$,$\Theta_1\subset\Theta$ 及 $\Theta_0\cap\Theta_1=\varnothing$. 当 Θ_0 包含多于一个点时,称检验为复合假设(composite hypothesis). 注意,在简单假设下,分布被唯一确定. 而在复合假设情况则不尽然.

在检验中,我们需要选择一个检验统计量(test statistic):$T\equiv T(X_1,\cdots,X_n)$. 因为检验统计量完全确定了检验的性质,所以,检验统计量也称为检验. 在原假设成立时,它的可能值只以很小的概率属于某个范围,比如集合 W. 如果事件 $(T\in W)$ 的确发生了,它在原假设下是一个小概率事件. 换句话说,原假设有问题,应该拒绝. 这时,W 称为拒绝域(rejection region 或 critical region). 如果事件 $(T\notin W)$ 发生了,则我们没有理由拒绝原假设. 当 W 是诸如 $(-\infty,c]$ 或 $[c,\infty)$ 一类的区间时,$T\in W$ 等价于 $T\leqslant c$ 或 $T\geqslant c$. 这时称 c 为临界值(critical value). 在决策中,我们可能会犯两种错误. 一种是原假设对,我们拒绝了它,这是所谓的第 I 类错误;另一种是原假设不对,但没有拒绝,即所谓的第 II 类错误. 犯这两类错误的概率分别为 $P_\theta(T\in W|\theta\in\Theta_0)$ 和 $P_\theta(T\notin W|\theta\in\Theta_1)$. 人们自然会希望这两个概率越小越好,但在样本给定之后不可能两全其美. 通常是先限制第 I 类错误概率不大于预先给定的概率 $0<\alpha<1$,它被称为显著性水平(level of significance)或检验水平(size of test). 即对任意的 $\theta\in\Theta_0$,

$$P_\theta(T\in W)\leqslant\alpha.$$

在此条件下，选择合适的检验统计量使犯第 II 类错误的概率尽可能地小，即使 $P_\theta(T \in W|\theta \in \Theta_1)$ 尽可能地大. 我们称 θ 的函数 $\beta(\theta) \equiv P_\theta(T \in W)$ 为势(函数)(power function). 显然当 $\theta \in \Theta_0$ 时，$\beta(\theta)$ 是犯第 I 类错误的概率. 而当 $\theta \in \Theta_1$ 时，$1-\beta(\theta)$ 是犯第 II 类错误的概率. 上面的限制第 I 类错误概率条件可写成

$$\sup_{\theta \in \Theta_0} \beta(\theta) \leqslant \alpha$$

注意，势函数实际上也依赖于检验 T 的选择，我们可记它为 $\beta(\theta,T)$. 如果一个水平 α 的检验 T^* 使得对于所有的水平 α 的检验 T 及所有的 $\theta \in \Theta_1$ 有

$$\beta(\theta,T^*) \geqslant \beta(\theta,T)$$

则称检验 T^* 是一致最优势的(uniformly most powerful——UMP). 因为人们总希望在水平 α 尽量小的时候拒绝原假设. 举例说，如果我们可以在 $\alpha=0.01$ 拒绝，当然也可以在 $\alpha=0.05$ 拒绝；但总是选小的 α 以证明我们拒绝得有道理. 因此，在实践及各种计算机软件中，人们并不预先指定水平的值，而是很方便地利用由数据产生的下面定义的 p 值. 在取得了 $X_1,\cdots,X_n$ 的观察值 $x_1,\cdots,x_n$ 之后，我们称概率

$$P_\theta(T(x_1,\cdots,x_n) \in W|\theta \in \Theta_0)$$

为该检验的 p 值(p-value) 或观察水平(observed size) 或显著概率(significance probability). 对于任何大于 p 值的水平，人们可以拒绝原假设，但不能在任何小于它的水平下拒绝原假设. p 值是使人们可以拒绝原假设的最小水平.

例 1.1　假设在 $n=10$ 次掷硬币的试验中，共出现正面 $S=3$ 次. 要检验该硬币是否均匀. 令 θ 为出现正面的概率. 原假设为 $H_0:\theta=0.5$，而备择假设为 $H_1:\theta<0.5$. p 值为

$$P_{0.5}(S \leqslant 3) = 2^{-10}\sum_{k=0}^{3} C_{10}^k = 0.1719$$

因此，对于所有小于 0.1719 的水平，我们不能拒绝原假设.

1.2.3　稳健性及稳健统计

我们知道，统计就是要使所建立的模型和其所反映的现实世界尽可能地一致. 但是，不存在完美的模型，也不存在不含误差的数据. 只能希望我们的方法或模型对于有危险的误差不至于太敏感. 这就是稳健性的概念(robustness). 稳健概念实际上是针对统计中的假设过分理想化而产生的. 稳健性是非参数统计的基本特点. 但是稳健统计是介于非参数统计和经典的(参数)统计之间的一些理论的集合，它是近似半参数模型的统计. 稳键统计的目的主要有以下几条：

1. 描述出适合于大多数数据的结构；2. 找出离群值(outliers)，如果需要的话，改变我们已有的结构；3. 在不平衡的数据结构中(如在回归分析中)，发现高度有影响的数据点(leverage points)，并给出警告；4. 对假定的诸如独立性等的相关结构进行审查并改进.

实际上，对于一个不太熟悉的数据结构，很难说清哪些影响点是真正满足我们要找的模型还是纯属误差的产物. 这就要对问题的背景有所了解. 纯数学式的思维方式是行不通的.

下面给出一个例子对稳健性进行说明.

例 1.2　设 $F(x)$ 为一关于 μ 对称的连续分布函数. $X_1,\cdots,X_n$ 是服从该分布的一个样本. 我们来比较两个 μ 的估计量. 一个是样本均值 $\overline{X}=\frac{1}{n}\sum_{i=1}^{n}x_i$, 另一个是样本中位数 X_{med}, 定义为顺序统计量的中间值, 即当 n 为偶数时它取 $\frac{\left(X_{\left(\frac{n}{2}\right)}+X_{\left(\frac{n}{2}+1\right)}\right)}{2}$, 而当 n 为奇数时它为 $X_{\left(\frac{n+1}{2}\right)}$. 这里顺序统计量 $X_{(1)}\leqslant\cdots\leqslant X_{(n)}$ 是按自小到大次序重新排列的 $X_1,\cdots,X_n$. 显然如果 $X_{(n)}$ 趋于无穷大, 则 $\overline{X}$ 也趋于无穷. 这说明 $\overline{X}$ 对个别数据的不寻常值很敏感. 而 X_{med} 则不因 $X_{(n)}$ 的异常变化而改变, 即 X_{med} 是 μ 的一个稳健估计. 我们还可看出, 虽然样本中位数具有稳健性, 但样本均值包含了更多的样本所具有的信息. 因此, 在不存在异常点时, 样本均值是更常用的.

1.3　数据初步分析

在拿到一个新的数据之后, 首先要有对该数据的直观了解. 本节介绍一些简单的数据分析, 使我们对数据的特点、大概的分布形状等有个粗略的了解, 为以后的进一步统计推断作好准备.

假定我们有三个班的 97 个学生的考试成绩表(表 1.1).

表 1.1　考试成绩

一班			二班			三班		
82	45	89	99	87	72	58	46	72
82	67	72	81	88	82	84	74	48
64	89	93	66	71	88	116	91	69
78	87	75	58	84	68	53	65	109
115	57	86	86	70	88	91	69	69
73	86	85	91	77	108	86	45	48
82	90	104	109	73	81	61	70	84
64	83	77	96	60	92	96	63	90
83	78	81	85	104	98			
96	62	77	104	57	25			
53	113	67	96	74	74			
103	39		72	96	88			
			84	62				

表中成绩是按学生姓氏笔画排列的, 人们从中并不容易一眼看出该数据的特征. 下面将对它进行初步的分析.

1.3.1 直方图

最常用的一个表现数据的方法是直方图 (histogram). 它通常把数据的值域分成若干相等区间，于是数据就按区间分成若干组，每组作成一个矩形，其高和该组中数据的多少成比例，其底为所属区间. 这些矩形就是直方图，它给数据的分布一个直观的形象. 图 1.1 就是表 1.1 的数据的直方图. 这里数据被分成 10 个区间，并形成 10 个矩形. 比如分数 40—49 有 5 个人，相应地形成高为 5(至多乘一常数)，宽为 10 的位于该区间的矩形.

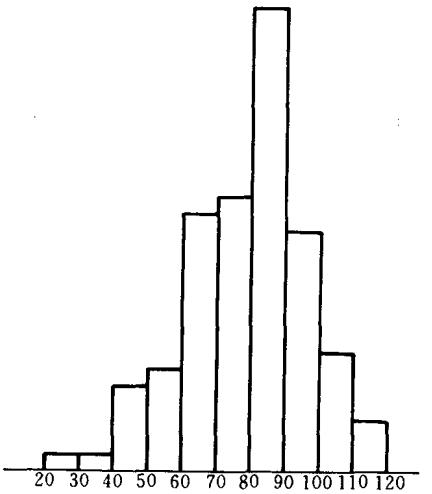

图 1.1　表 1.1 数据的直方图

一般说来，对于观测数据 $X_1,\cdots,X_n$，选择两个适当的常数 X_0 和 $h(>0)$，把 $(-\infty,+\infty)$ 分成一些小区间 $\Delta_i=[X_0+(i-1)h, X_0+ih)$，$i=0,\pm1,\pm2,\cdots$，并以 n_i 记 $X_1,\cdots,X_n$ 落在 Δ_i 的个数. 我们以 Δ_i 为底，$\frac{n_i}{nh}$ 为高做一矩形. 对 $i=0,\pm1,\pm2,\cdots$ 而得的许多矩形就是一个直方图. 直方图的形状依赖于区间的选择. 数据的特点及画图者的观点都对此有影响.

1.3.2 茎叶图

一个茎叶图(stem-and-leaf display)和直方图类似，只不过用数据代替矩形. 具体地说，把数据按除了最后一位数之外的前面数字的异同来分组；相同的分为一组(或若干组，依具体数据情况而定). 每一组数占一行，以前面的数字作为该行的标记，放在行头；并把这些数按由小到大的顺序从上往下排，这就形成了一个“茎”. 每一行则是该组的所有数据的最后一位数字的排列(通常按由小到大的顺序从左至右排列)，这就是“叶子”. 一组中，数据越多“叶子”越长. 这既直观，又显示了具体数据.

我们把表 1.1 中的得分作出若干茎叶图：图 1.2 是三班成绩的茎叶图(没有按大小排“叶子”). 图 1.3 是所有学生的成绩的茎叶图(每行按大小排列). 图 1.4 是二班和三班成绩的背靠背茎叶图(back to back stem and leaf display)，它使这两个班的成绩共用一个茎，但两个班的“叶子”分别向上下两边排列. 从该图可看出两个班成绩的不同分布特点. 这些图的茎中的值是

不同的，但如前面提到的，若干行可共有相同的茎值. 图 1.5 和图 1.6 就是对一个班 30 个学生成绩的因不同分组而作成的两个茎叶图，其中一个茎值相应于若干行.

4	6	8	5	8			
5	8	3					
6	9	5	9	9	1	3	
7	2	4	0				
8	4	6	4				
9	1	1	6	0			
10	9						
11	6						

图 1.2　三班成绩的茎叶图

					9				
					9				
					8				
					8				
					8				
					8				
					7				
					7				
					6				
				8	6				
				8	6				
				7	6				
			9	7	5				
			9	5	5	9			
			9	4	4	8			
			8	4	4	6			
			7	4	4	6			
			7	3	4	6			
			6	3	3	6			
			5	2	3	6			
			4	2	2	3	9		
		8	4	2	2	2	9		
		8	8	3	2	2	1	8	
		8	7	2	1	2	1	4	
		6	7	2	1	1	1	4	6
		5	3	1	0	1	0	4	5
5	9	5	3	0	0	1	0	3	3
2	3	4	5	6	7	8	9	10	11

图 1.3　表 1.1 的茎叶图

						8				
						8				
						8				
						8				
					4	7				
					4	6	9			
					3	5	8			
					2	4	6			
				8	2	4	6	9		
				6	1	2	6	8		
			8	2	1	1	2	4		
5			7	0	0	1	1	4		
2	3	4	5	6	7	8	9	10	11	
		5	3	1	0	4	0	9	6	
		6	8	3	2	4	1			
		8		5	4	6	1			
		8		9			1			
				9						
				9						

图 1.4　背靠背茎叶图

5	9											
6	1	3	4	4	4	4						
6	6	6	7	7	7	7	8	8	8	9	9	9
7	0	0	1	1	1	2	3	3	4			
7	7	8										

图 1.5　分两组的茎叶图

5	9						
6	1						
6	3						
6	4	4	4	4			
6	6	6	7	7	7	7	
6	8	8	8	9	9	9	
7	0	0	1	1	1		
7	2	3	3				
7	4						
7	7						
7	8						

图 1.6　分五组的茎叶图

1.3.3　五数概括

直方图和茎叶图包含了大量的样本信息，但没有作任何加工或简化. 我们有时需要用少数

几个统计量来对大量的原始数据进行概括. 下面引进所谓的五数概括(five-number summaries).

有了一组数据之后,我们首先感兴趣的可能是数据的"中心". 通常人们首先想到的"中心"的度量是样本均值. 样本均值的确用得很多,但正如前面所说,样本中位数也是一个可取的关于数据"中心"的度量,它具有某种稳健性. 这一节我们就用它来度量数据的"中心".

我们引入层(depth)的概念. 如果把数据按大小次序排列(假定有 n 个数据),则最外面的两个,即最大和最小的两个数称为第一层,然后依次往里称为第二层,第三层等等. 当 n 为奇数时,最后一层(第$\frac{n+1}{2}$层)只剩一个数,即中位数;而当 n 为偶数时,最后一层(第$\frac{n}{2}$层)剩两个数,它们的平均是中位数. 我们用 μ 表示中位数,其层数用 $d(\mu)$ 表示.

例 1.3

有 8 个数:	46	48	58	72	74	84	91	116
层:	1	2	3	4	4	3	2	1

在茎叶图中,定义茎中每一数的层为该茎所对应的叶中数据的层的最大值. 中位数所在的茎的层为该叶所具有的数据数目,并用(·)表示.

例 1.4 图 1.7 为表 1.1 数据的茎叶图.

层	茎	叶													
1	2	5													
2	3	9													
7	4	5	5	6	8	8									
13	5	3	3	7	7	8	8								
28	6	0	1	2	2	3	4	4	5	6	⋯				
46	7	0	0	1	1	2	2	2	2	3	3	4	⋯		
(27)	8	1	1	1	2	2	2	2	3	3	4	4	4	4	⋯
24	9	0	0	1	1	1	2	3	6	6	6	⋯			
10	10	3	4	4	4	8	9	9							
3	11	3	5	6											

图 1.7 层及中位数

从该图亦可找到中位数. 因中位数的层数为 $d(\mu)=49$,而图中茎值为 7 的层为 46,所以,茎值为 8 的叶中第三小的数为中位数($\mu=81$).

除了中位数之外,我们还对数据的分散程度感兴趣. 这里我们不考虑样本方差,而考虑极大值、极小值、上四分位数(upper quantile)Q_U 及下四分位数(lower quantile)Q_L. 四分位数的层定义为 $d(Q)=\frac{n+2}{4}$ 或 $d(Q)=\frac{n+3}{4}$ 依 n 为偶数或奇数而定. 得到了层数也就有了上下两个四分位数. 从例 1.3 的数据,易得 $Q_L=53$ 和 $Q_U=87.5$. 中位数、极值和四分位数就是我们所谓的五数概括. 数据落在 Q_L 和 Q_U 之间的概率为 0.5. 在它们之外太远的数则有可能为异常值. 记 $H=Q_U-Q_L$. 人们认为在区间$(Q_L-1.5H, Q_U+1.5H)$之外的数据可看作是异常值. 如表 1.1 中的 25 可为一例.

1.3.4 盒子图

五数概括并不直观，现把这五个数画在一个图上：在 Q_L 与 Q_U 之间画一矩形盒子，在极大值与 Q_U 之间，极小值与 Q_L 之间画两线段，并在中位数处画一竖线就成了我们的盒子图(box-plot). 图 1.8 为表 1.1 数据的盒子图.

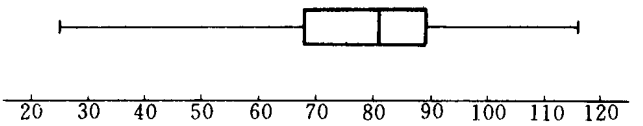

图 1.8 表 1.1 中数据的盒子图

图中矩形描述了中间的 50% 的数据；左右的水平线段代表了上下 25% 的数据的分布情况. 图 1.8 显示出高低两部分数据(各占 25% 的数据)并不对称. 数据在中位数 81 附近还是集中的(盒子短).

以上这节所作的数据分析虽然很初等，但是简单明了，直观性强. 它不要求数据符合任何统计模型，是获得数据之后的一种处理方法.

1.4 顺序统计量的基本性质

非参数统计的一大特点就是利用样本数据的大小关系来进行研究. 因此，对在第一节已涉及的顺序统计量(order statistics)的研究构成了非参数统计的基础. 设

$$X_{(1)} \leqslant X_{(2)} \leqslant \cdots \leqslant X_{(n)}$$

为样本 $X_1, \cdots, X_n$ 的顺序统计量. $X_{(i)}$ 为第 i 个顺序统计量；$X_{(1)}$ 和 $X_{(n)}$ 分别称为极小值和极大值；$X_{(n)} - X_{(1)}$ 称为极差. 在第一节中，我们已用顺序统计量来表示了样本中位数. 实际上，前面所讲的“五数”概括中的五数都是 p 分位数(p-quantile) 的特例. p 分位数 定义为

$$m_p = X_{([np])} + (n+1)\left(p - \frac{[np]}{n+1}\right)(X_{([np]+1)} - X_{([np])})$$

其中 $[x]$ 表示不大于 x 的最大整数.

本节介绍一些有关顺序统计量的基本知识.

1.4.1 顺序统计量的精确分布

本节始终考虑独立同分布 (iid) 的样本 $X_1, \cdots, X_n$. 假设它们的总体分布函数为 $F(x)$，而其顺序统计量 $X_{(r)}$ 的分布函数为 $F_r(x)$，分布密度函数为 $f_r(x)$. 我们有

$$F_r(x) = P(X_{(r)} < x)$$