非参数统计讲义（总14页）

--本页仅作为文档封面，使用时请直接删除即可--
--内页可以根据需求调整合适字体及大小--

第一章绪论

本章主要内容：1. 非参数方法介绍

2. 预备知识

第一节非参数方法介绍

一. 非参数方法的概念和实例 复习参数方法**定义：**设总体X的分布函数的形式是已知的，而未知的仅仅是分 布函数具体的参数值，用样本对这些未知参数进行估计或进行某种形式的假设 检验，这类推断方法称为参数方法。

先来看两个实例。

例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（土）cm。这 也就是说合格零件长度的中心位置为，允许误差界为，即长度在-之间的零件 是合格的。为评估近年来供应的零件是否合格，随机抽查了门二100个零件，它 们的长度数据X见第一章附表。

解答：

根据我们已学过的参数统计的方法，如何根据数据来判断这批零件合格否

用参数数据分析方法，在参数统计中，运用得最多的是正态分布，所以考 虑假设供应商供应的零件长度X服从正态分布，即

$$X \sim N \quad (\text{呻}\ 2)$$

其中两个参数均未知，但可用样本均值估计R，样本方差估计。2。

由已知的数据计算可得：零件的平均长度，即样本均值为*X*二，样本标准差 为s二。

则零件合格的可能性近似等于

$$P(8.4 < X < 8.6) = \text{中}((8.6\ \text{f})/\text{。}) - ①((8.4 - \text{四})/\text{。})$$

2 ((8.6 − 8.9458)/0.1047)—中((8.4 − 8.4958)/0.1047)

*牝*66%

这个说明：约有三分之一的零件不合格，该工厂需要换另一个供销商了。

但这个结论与实际数据符不符合呢这是我们要思考的问题。

我们可以对数据做一个描述性分析，先对这100个样本数据做一个频率分 布。

观察到：在这100个零件中有91个零件的长度在˜之间，所以零件合格的 比例为91%,超过66%很多！

统计分析的结论与数据**不吻合**的！这是什么原因呢

我们可以作出数据的直方图来分析数据的分布情况。由图知，该数据的总 体不是近似服从正态分布的！所以我们对于数据的总体分布的假设错了！问题 就出在假设总体是正态分布上！继续看直方图，能否很容易就观察出来它大概 是什么分布呢答案是不易看出，所以试图先确定数据的分布函数，再利用参数 的方法来分析是不太容易的。

例 哪一个企业职工的工资高

这里有22名职工的工资情况，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)如附表。

仅从数据来看，显然企业1职工的工资较高。

根据我们已学过的参数统计的方法，这个问题用什么方法来解决呢（提 问）

采用参数数据分析方法，假设企业1和企业2职工的工资分别服从正态分 布Ng_2）和N（b,o_2），则该问题转化为假设检验问题：

$$H_0{:}a = b,\quad {}_{\%:}a > b$$

即两样本t检验。

计算可得，检验统计量的值t二。

若取a二，其临界值为％$_5$（20）=1.725，不能拒绝原假设，即认为二者没有区 别；

若取a二，其临界值为％（20）=1.325，仍不能拒绝原假设！计算p值得到的 结论也一样。

这个统计分析的结论显然与数据不吻合！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， 并且有一部分人的工资比较高，所以分布的右边有较长的尾巴。

对于以上的这样的问题，若想用参数数据的分析方法，就不能再假设总体 服从正态分布，必须给它们赋一个较合理的分布函数，做到这点对于很多实际 问题上是难度比较大的。除了这个办法之外，我们还可以用另外的处理办法， 例如，非参数统计、参数和非参数方法相结合等等。这门课，我们主要讨论非 参数方法。

二. 非参数统计方法特点

1. 非参数统计方法通常称为“分布自由”的方法，即非参数数据分析方法对 产生数据的总体的分布不做假设，或者仅给出很一般的假设，例如连续型分 布、对称分布等一些简单的假设，结果一般有较好的稳定性。所以适用范围非 常宽泛。

在经典的统计框架下，正态分布一直是最引人注目的，但是对总体的分布 不是随便做出来的，如以上两例，盲目地做出正态分布的假设有时候是起反作 用的。

当数据的分布不是很明确，特别当样本含量不大，几乎无法对分布作推断的 时候，此时使用参数方法就有一定的风险，我们就可以考虑用非参数的方法。

但要**注意，非参数方法是与总体分布无关，而不是与所有分布无关！**

2. 非参数统计可以处理所有类型的数据。我们知道，统计数据按照数据类 型可以分为两大类：定性数据和定量数据。一般地，参数统计是处理定量数据，如果所收集到的数据不符合参数模型的假定，比如：数据只有顺序，没有 大小，则很多参数模型无能为力，此时只能尝试非参数方法。例如：研究急性 白血病患儿血液中血小板数与出血症状之间的关系。血小板数可用数据衡量， 但出血症状则只能分为：明显、较明显、有出血点和无这4类。类似于这样的 “等级资料”，参数方法没辙，可用非参数方法中的Spearman等级相关方法来 做。

3. 在不知道总体分布的情况下，如何利用数据所包含的信息呢一组数据最 基本的信息就是次序。非参数统计就是利用这个最基本的信息。如果把数据点 按从小到大的次序排队，每一个具体数目都有它在整个数据中的位置，这称为 该数据的秩（rank）。非参数统计的一个**基本思想:用数据的秩代替数据，构 造统计量进行统计推断。**数据有多少个观察值，就有多少个秩。在一定的假设 条件下，这些和由它们构成的统计量的分布是求得出来的，而且和原来的总体 分布无关。就可以进行所需要的统计推断了。所以说，非参数统计只是和总体 的分布无关，但和秩以及它们统计量的分布是密切相关的！

另外，其它与总体分布无关的统计方法也属于非参数统计。

4. 在考虑非参数统计量的分布时，我们较多考虑这些统计量的渐近分布， 由于利用到一些大样本方面的定理，得出来的渐近分布都服从正态分布或是由 正态分布导出的分布，较容易计算和处理。

5. 非参数方法与参数方法

通过刚才上面的解说，不要产生错觉，认为非参数方法总比参数方法有 效！

非参数方法不是总比参数方法有效！

毕竟非参数方法利用到的数据信息非常有限。如果人们对总体有充分的了 解且足以确定其分布类型，则非参数方法比参数方法效率低。

例如在总体分布族已知的情况下，非参数统计一般不如参数统计结果精确！ 另外，在总体分布是均匀分布时，正态的参数方法又比非参数方法好！这点可 以通过计算渐近相对效率来说明。

三. 非参数统计的历史

相对参数统计而言，非参数统计起步较晚，但有后来者居上的趋势。

非参数统计的形成主要归功于20世纪40年代~ 50年代化学家F. Wilcoxon等人的工作。Wilcoxon于1945年提出两样本秩和检验。1947年Mann和Whitney两人将结果推广到两组样本量不等的一般情况。之后，相继涌现出大量论文。Savage 1962年统计的非参数论文就有3000多项。

Pitman于1948年回答了非参数统计方法相对于参数方法来说的相对效率方面的问题。

1956年，J.L. Hodges和E.L. Lehmann则发现了一个令人惊讶的结果，与正态模型中t检验相比较，秩检验能经受住有效性的较小损失。而对于 重尾分布所产生的数据，秩检验可能更为有效。第一本论述非参数应用的书于 1956年由出版，有人记载从1956年到1972年，该书被引用了 1824次。这也 说明非参数统计在这一时期的发展是相当活跃的。

60年代，J.L. Hodges和E.L. Lehmann从秩检验统计量出发，导出了

若干估计量和置信区间。这些方法为后来非参数方法成功应用于试验设计数据 开启了一道大门。之后，非参数统计的应用和研究获得了巨大的成功。

上世纪六十年代中后期，Cox和Ferguson最早将非参数方法应用于生存 分析。上世纪70年代到80年代，非参数统计借助计算机技术和大量计算获得 了更稳健的估计和预测，以P. J. Huber和F. Hampel为代表的统计学家从计算 技术的实现角度，为衡量估计量的稳定性提出了新准则。

上世纪90年代有关非参数统计的应用和研究主要集中在非参数回归和非参 数密度领域，其中较有代表性的人物是Silverman和. Fan。

四. 非参数统计主要内容

非参数统计可以分成两个范畴，一个是比较经典的基于秩的，以检验为主 的非参数统计推断，而另一部分是近二三十年来发展的非参数回归、非参数密 度估计、自助法以及小波方法等现代非参数统计方法。这两者均不对总体分布 做较为确定的假设，但除此之外，这两部分内容在方法和概念上均没有多少共 同点。我们首先介绍经典地基于秩的，以检验为主的非参数统计推断，这也是 我们的主要内容，然后介绍现代非参数统计的部分内容。

第二节预备知识

一、秩统计量

1 . 定义：设*Z.Z*是来自连续分布*F*(*z*)的简单随机样本，*Z*(1) < …< *Z*(〃)为其 次序统计量。定义随机变量

$$R = r，当\ Z = Z_{()}，\ i = 1,2,\cdots,n。$$

当是唯一确定时，称样本观测值*Z*. 有秩*R* , *i* = 1,2,...,*n*。（由于*F(z)*连续，因而*R*不唯一确定的概率为0。）$_i$

即*R*. 是第i个样本单元*Z*在样本次序统计量(*Z*),..•%)中的位置。

例1：已知一组数据，请写出它们相应的秩。

(1) 20, 10, 30。

解:先将该组数据从小到大排列如下：10, 20, 30。所以10对应的秩为1, 20 对应的秩为2, 30对应的秩为3。

(2) 200,100,300。

解:先将该组数据从小到大排列如下：100, 200, 300。所以100对应的秩为1, 200对应的秩为2，300对应的秩为3。

注意：这两组数据显然区别较大，但他们对应的秩却都是1，2，3。没有差别II

2. 性质。

定理1记$R = (R_1, \ldots, R_n)$, 集合

M = {(尸，…,r)|(r，…,r)是(1,…,n)的一个排列}，
1 n 1 n

则R在沉上均匀分布。

证明：易知R仅在沉上取值。对任意一个$r =$(个…, r) 6沉,

$$P\{R = r\} = P\{(R, \ldots, R) = (r, \ldots, r)\}$$
1 n 1 n

$$= P\{(Z_1, \ldots, Z_n) = (Z_{(r1)}, \ldots, Z_{(rn)})\}$$

$$= P\{(Z1L'' = (Z(1), \ldots, Z(n))\}$$

$$= P\{Zd1 < \bullet\bullet\bullet < Zdn\},$$

其中$q = k$，当 ⌈i时，即$d, (i = 1,...,n)$是i在排列r中的位置。又由于

$$(Z_1, Z, \ldots, Z_n) d (Z_{d1}, Z, \ldots, Z_{dn}),$$

所以$P\{R = r\} = P\{Z_1 < \bullet\bullet\bullet < Z_n\}$对任意r e沮，上式均成立，所以对任意$r$，这个概率均相等。而全部这样的事件互不相容且它们的和是必然事件，故对任意r 6 沮，有 $P\{R = r\} = 1/n!$。

定理2 $R = (R$ ⌈气, …, $R_n)$的边缘分布也是均匀分布，特别一维边缘分布有

$$P(R=r)=\begin{cases}\text{当}\ r=1,2,\sim,\text{时},\\ 0,\text{其他}。\end{cases}$$

以上内容仅为本文档的试下载部分，为可阅读页数的一半内容。如要下载或阅读全文，请访问：https://d.book118.com/308000101016006057