KEJI BIANJI JICHU ZHISHI

科技编辑基础知识

翁 廉 主编

统计方法的应用与常见错误辨析

TONGJI FANGFA DE YINGYONG YU CHANGJIAN CUOWU BIANXI

中山大学出版社

科技编辑基础知识　翁　廉 主编

统计方法的应用与常见错误辨析

中山大学出版社
·广州·

图书在版编目（CIP）数据

科技编辑基础知识·统计方法的应用与常见错误辨析/翁廉主编. —广州：中山大学出版社，2007. 11
ISBN 978-7-306-02883-9

Ⅰ. 科… Ⅱ. 翁… Ⅲ. 科学技术—编辑工作—基本知识
Ⅳ. G232

中国版本图书馆 CIP 数据核字（2007）第 070352 号

出 版 人：叶侨健
责任编辑：钟永源　李海东
责任校对：中　原　蒴　蔷
封面设计：林　怡
责任技编：黄少伟
出版发行：中山大学出版社
电　　话：编辑部（020）84111996，84113349
　　发行部（020）84111998，84111981，84111160
地　　址：广州市新港西路 135 号
邮　　编：510275　　传真：020-84036565
网　　址：http://www.zsup.com.cn　E_mail.zdcbs@mail.sysu.edu.cn
印 刷 者：广东省农垦总局印刷厂
　　（地址：广州市天河东莞庄路　邮编：510612）
规　　格：850mm×1168mm　1/32
总 印 张：25.75 印张　646 千字
版次印次：2007 年 11 月第 1 版　　2007 年 11 月第 1 次印刷
总 定 价：58.00 元（共 7 册）　　印数：1-1000 套

序

科技期刊是科技信息交流与传播的重要信息载体，在先进生产力的发展中具有智力支持和媒介支持的特殊作用。促进科技期刊出版业的发展，培育更多的精品期刊，满足社会经济日益发展的需要，是我省科技期刊编辑工作者和科技管理工作者共同的责任，也是贯彻落实科学发展观的重要体现。

科学发展观的本质是以人为本。胡锦涛总书记指出："我们要培养大批优秀人才。国家兴盛，人才为本。"所以，为了促进科技期刊出版业的发展，我们既强调提高人的知识水平，完善人的知识结构，以实现人的素质培养为根本目标，一切为了人；又强调依靠人来推动和实现科技期刊业迅速发展，一切依靠人。市场竞争，实质上是人才竞争。在当前市场竞争激烈的形势下，要培育优秀期刊，铸造精品名牌，具有决定性作用的措施是按照科学发展观培养一批具有创新能力的主编和造就一批具有多元知识结构的科技期刊编辑。这是科技期刊出版业持续发展的需要，也是当务之急。

科技编辑知识，是科技期刊编辑人员知识结构的十分重要的组成部分。这部包括七个分册的《科技编辑基础知识》的选题及其内容，是"科技编辑基础知识教育"研究课题组对不同编龄、不同层次的编辑人员进行调查后，根据受访人的知识需求确定的。七个分册所讲的就是科技期刊编辑人员，尤其是青年编辑必须具备的科技编辑的基础知识，对科技编辑人员是有实用价值的。组织编撰这样的一部书，是一项具有创新性的工作，是实践

胡锦涛总书记关于“着力提高人的创新能力，努力造就大批优秀人才”的具体体现。希望这部书对于我省科技期刊编辑队伍的建设能够发挥一些积极的作用，从而促进我省科技期刊出版业持续健康发展。

本书编撰过程中，始终得到广东省科技厅有关领导和广东省科技期刊编辑学会的鼎力支持与帮助。在此，谨向他们表示衷心的感谢。

翁　廉

2007年5月

前　言

用定量分析工具解决实际中的问题是人类文明发展的必然要求，人类的社会实践活动一旦融入了统计的成分便表现出了科学的特征。从原始社会的结绳记事到目前高度发达的信息时代，统计已经有五千年的发展历史。作为一门科学，统计已经广泛应用到自然科学和社会科学的各个领域，尤其在工程技术、生命科学、社会调查、经济管理、农业生产甚至军事理论等方面，统计已经成为不可或缺的手段和工具。

然而，随着科学技术的高速发展对统计实践和理论要求的不断深入，以及大量的统计方法在实践中的广泛应用，许多问题也随之出现，尤其是统计方法的错误使用所导致的数据或结论与事实的严重背离已经成为后续工作正常展开的直接障碍和危险因素，甚至会成为整个工作计划失败的开端和根源，进而必然会对人们的生产实践活动造成损失。综观统计学发展的历史，建立在数据分析基础之上的严密的数学理论是统计科学得以正确发展的直接保障和依据。但对统计理论的理解与掌握并非一促而就，需要一个努力的过程，单凭主观上的热情和愿望并无助于问题的解决，过度的急功近利必然会导致错误。所以在对待用统计方法解决实际问题上要持有严谨的态度。遗憾的是，许多研究者及论文的作者为了片面追求“学术成果”，忽视统计科学的规律和要求，在并未充分掌握基本理论的前提下，照本宣科，盲目甚至滥用统计分析方法，结果得到的是荒谬或与事实相反的结论。更有甚者，为使实验结果符合“预期的结论”，置科研要求中的科学

性和严谨性于不顾，毫无根据地选择统计方法或修改实验数据，致使数理统计方法的可信度降低，科学的权威性遭到破坏。结合现在许多科技期刊中存在的统计学错误，这不能不说已经成为一种严重的社会问题。基于此，本书从基本的统计理论出发，系统地介绍了完整的统计过程所用到的基本原理和方法，并根据科技论文中常见的统计学错误的表现形式，给出了发现、判断以及避免、排除相应错误的方法。

本书共分四章，先简要介绍统计方法，然后按统计学的大类，从统计描述、统计推断和统计预测三个方面分别阐述统计方法的原理和应用中的注意事项及常见错误辨识。限于篇幅，有些理论本书只作了列举并未给出完整的推证，而且由于不同专业的特殊要求，本书无法进行全面介绍，建议有兴趣的读者参阅相关书籍。

由于时间仓促及编者的水平有限，本书中难免存在一些纰漏和错误，恳请读者进行批评指正。

刘淑华　明宗峰

2007 年 8 月

目　录

第一章　统计方法简介

第一节　统计的含义

统计作为一种社会实践活动是随着人类社会经济的发展，适应于国家治理和处置各种社会事务的需要而产生的，是伴随人类文明的进步一同发展起来的。现在已经发展成为一门重要的科学。

“统计”最基本的含义是对客观事物进行数量方面的计量和分析，是对所关心的对象获取量的信息的实践活动。统计本身可能无法直接得到解决问题的方法，但根据实际需要所做的信息的整理与收集却是解决问题、制订计划或预测事件发展规律的有效途径。一般情况下，通过统计方法所获取的信息可以为下一步工作计划的展开提供方法上的指导，有效的统计甚至可能直接成为解决实际问题的工具。作为一门科学，它主要通过客观对象数量上的特征或者表征客观对象特定属性间的数量的相关性来反映事件发展的规律，而这种规律本身可能是诸多非数量特征的客观原因所导致的必然结果。比如，股市的价格浮动可能是银行利息调整、国家经济政策宏观调控、企业经营状况以及有雄厚实力的投资者的人为操纵等原因的直接表现，但单纯任何一个单一的因素都不足以准确描述股市价格变动的规律。由于影响股市因素的多样性和不确定性，考察全部的影响因素显然是不现实的，因此基于影响股价的因素而建立的数学模型，无论多么复杂，都不可能对股市的走向

做出长期准确的预测。现在的经济学家更倾向于用统计的方法来揭示这一规律，原因在于事物数量上的相关特性必然使得其在演变过程中留下数字变化特征的痕迹。统计方法完全抛开各种具体的影响因素而纯粹从事件发展过程中所表现出的数量特征来捕捉其演变规律，简单说来就是从客观对象数量的变化关系中寻找量与量之间的函数关系。用数字的相关性所总结或归纳出来的演变规律其实是包含了影响事件发展的全部因素，从这个意义上讲，统计作为一种方法论或一门科学有着其他物理手段不可替代的作用，其广泛的应用领域和科学严密的数学手段必然使其具有极其深刻而十分丰富的内容。

人们的实践活动是复杂的、具体的，统计作为获取信息的手段，其方法也是多样的。而且在为实现同一目的所作的统计中，统计的内容也各不相同。比如，要获取一个单位的职工数量，最简单的办法就是直接找到该单位的人事部门或其主管部门询问。有时这种方法可能不太现实，我们则可以从对该单位的食堂规模、职工宿舍的建筑面积、停车场的大小以及单位时间内出入单位的平均人数等信息的统计中推断出一个大致的人数范围。这里需要强调的一点就是，如果不是对所考察的对象作直接的统计分析，统计的对象和内容必须要与所考察的对象有数量上的关联，而且这种关联具有事先的确定性。例如，统计了一个学校预订教材的数量就可以大体知道该校的学生人数，因为每个学生只能订一份教材是事先确知的，而且一般情况下所有的学生都必须拥有教材。

由此看来，统计是我们获取信息的最直接的手段，科学的统计方法和全面的统计资料有助于我们对未知事件作出合理的预测，从而有助于我们制定完善的计划或对解决实际问题提出有效的方法。春秋时期，秦国著名政治家商鞅在《商君书·去疆》中强调："强国知十三数，竟内仓、口之数，壮男、壮女之数，老、弱之数，官、士之数，以言说取食者之数，利民之数，马、牛、刍稿之数。欲强国，

不知国十三数，地虽利，民虽众，国愈弱至削。”意思是要实现国家的强盛必须先对国情国力有个基本的统计，并以此来制定正确的治国方针，否则即使地广人丰，国家也会因情况不明而导致决策失误，从而使国家渐趋衰落，甚至灭亡。《孙子兵法》中亦有言，“知彼知己，百战不殆”。其实这个“知”的过程必然是一个统计的过程，是为“战”服务的，卓越的军事家总是善于运用统计的方法以获取用于指导战争的翔实的资料，他们能够将各种统计方法和资料应用得出神入化。这些都是统计资料在预测和决策中所处地位的重要性的具体体现。人们可能比较关心的问题是，统计作为获取信息的工具是否只是解决问题的辅助手段？统计能否成为解决问题的直接的方法？对这个问题的正确回答可能会有助于我们加深对统计这一科学的真正理解。一般情况下，统计活动的结果作为一种信息或资料在人们处理问题的过程中确实是一种重要的辅助，但随着科技的发展，尤其是计算机技术的广泛运用，再加上统计方法本身的不断完善，统计成为一种直接解决实际问题的手段已是不争的事实。统计自产生之日起就赋有服务于人类实践需要的属性，只是由于时代不同和应用水平的局限，其作用的体现有所侧重而已。

我们所处的时代是文明高度发达的时代，是信息化的时代，在这样的历史条件下，统计比以往更具有实际的意义。自然科学的飞速发展和技术的不断进步已经为统计赋予了新的历史使命。可以用马寅初先生的一段话来总结其重要性：“人类社会，日臻繁复，耳目有所未周，则不能无赖于统计焉。盖个人动作，在与社会有关，倘于社会事实，未尽了了，则闭门造车，难期合辙。自然界现象，变化万端，亦非一二人所能穷，则综合统计又为必要。是故学者不能离统计而研学，政治家不能离统计而施政，事业家不能离统计而执业也。”总之，统计作为一门科学，其应用范围已经渗透到自然科学和社会科学的各个领域。其作用和价值可以概括为以下几点：

（1）统计是人们揭示自然规律，认识世界的重要手段；

(2) 统计是制定计划、处理事务的有力保障；

(3) 统计是科学研究的有效工具。我们作为社会的一员，一切工作的进一步展开，都有赖于一定的统计知识，因此学好用好统计，也是我们的必然选择。

第二节　统计学中常用的基本概念和术语

鉴于编写本分册的初衷是介绍统计方法的常用知识和应用技巧，以实现在科技论文的编辑过程中及时准确地辨识一些常见的统计错误，而且由于统计方法在科技论文中的应用具有相对的独立性，以及限于篇幅，本分册不准备过多地讲述概率论方面的相关知识，尽管有些理论是统计学的基础。如果读者对于统计上涉及概率论的某些概念或理论感觉模糊或抽象，可以查阅概率论方面的有关书籍。以下只对统计中经常用到的一些概率中的基本概念和术语作简单的陈述。

一、统计中有关概率方面的基本概念

1. 随机试验

满足如下条件的试验称为随机试验：

(1) 试验可以在相同条件下重复进行；

(2) 每次试验的可能结果不止一个，但所有可能的结果都是事先已知的；

(3) 进行一次试验前不能确定会出现哪个结果，但试验结果必是已知结果中的一个。

2. 样本空间

随机试验的所有可能结果组成的集合称为样本空间，常用 S 来表示。S 中的元素称为样本点。S 中包含的样本点可以是有限个，也可能是无穷多个，或在区间及整个数轴上取值。

3. 随机事件

样本空间的子集称为随机事件，简称事件。它在每次试验中可能发生，也可能不发生，带有一定的随机性。常用大写字母 $A,B,C,\cdots$ 来表示。每次试验中，当且仅当这一子集合中的一个样本点出现，则称这一事件发生。比如，若 $e \in A$，每次试验中，如果样本点 e 出现，则称事件 A 发生；反之，如果事件 A 发生了，则这一子集合中必有某一个样本点 e 出现。以下是随机事件的特殊情形：

基本事件：随机试验 E 的每一个可能结果，称为基本事件，随机事件就是由若干基本事件组成的集合。

必然事件：随机试验中一定发生的事件，称为必然事件。样本空间 S 包含所有的样本点，它是 S 自身的集合，在每次试验中它必然发生，所以称为必然事件。

不可能事件：每次试验中不可能发生的事件，称为不可能事件，常记为 $\varnothing$。它不包含任何样本点，作为样本空间的子集，它在每次试验中都不发生，所以称为不可能事件。

4. 频率与概率

频率：在相同条件下进行 n 次试验，其中事件 A 发生的次数 n_A 称为事件 A 发生的频数，比值 $\frac{n_A}{n}$ 称为事件 A 发生的频率，记为 $f_n(A) = \frac{n_A}{n}$。当 n 较小时，用频率来表达事件发生的可能性的大小是不恰当的。但随着 n 的逐渐增大，频率 $f_n(A)$ 会趋于某个常数 p，这个常数 p 就是度量事件发生可能性大小的概率，记为常数 $P(A) = p$。这是概率的统计定义。

概率的公理化定义：记所有随机事件的集合为 F，即 $F = \{A \mid A \in S; A$ 为随机事件$\}$。对于每个事件 A，取一个实数与之对应，记为 $P(A)$，$P(A)$ 是定义在 F 上的集合函数。如果 $P(A)$ 满足下列条件：①非负性：$P(A) \geqslant 0$；②规范性：$P(S) = 1$；③可列可加性：设

$A_1, A_2, \cdots, A_n, \cdots$ 是两两互不相容的事件，则有 $P(A_1 \cup A_2 \cup \cdots) = \sum_{i=1}^{\infty} P(A_i)$。则称 $P(A)$ 为事件 A 发生的概率。

5. 随机变量

设随机试验 E 的样本空间为 S，如果对于每一个样本点 $e \in S$，变量 X 都有一个确定的实数值与之对应，则 X 是定义在 S 上的实值函数，即 $X = X(e)$，称这样的变量 X 为随机变量。显然随机变量是样本空间到实数集的单值映射，如果映射的范围为有限个或可列个，则称随机变量是离散型随机变量；若映射的范围为某个实数区间，则称随机变量是非离散型随机变量。

6. 分布

一个随机现象的规律通常通过随机事件及其概率来描述。一个随机试验的所有结局事件与对应的概率的排列称为分布。对应于样本数量值的分布称其为频率分布；对应于总体数量值的分布称为概率分布。对于离散型随机变量，有以下几种重要的概率分布：

(1) 伯努力分布：设随机变量 X 的所有可能的取值为 0 和 1，它的分布为

$$P(X = k) = p^k(1-p)^{1-k},\ k = 0,1,\ 0 < p < 1$$

称 X 服从伯努力分布，又称 $(0,1)$ 分布，简记为 $X \sim B(1,p)$。

(2) 超几何分布：设随机变量 X 的概率分布为

$$P(X = k) = \frac{C_M^k C_{N-M}^{n-k}}{C_N^M},\ \max\{0, n-N+M\} \leqslant k \leqslant \min\{n, M\}$$

其中 n, M, N 都是正整数，$M \leqslant N$，则称随机变量 X 服从超几何分布，简记为 $X \sim H(n,M,N)$，其中 n, M, N 是分布的参数。

(3) 二项分布：设随机变量 X 的概率分布为

$$P(X = k) = C_n^k p^k q^{n-k},\ k = 0,1,2,\cdots,n$$

其中 n 为正整数，$0 < p < 1$，$p + q = 1$，则称随机变量 X 服从参数

为(n,p)的二项分布，简记为$X \sim B(n,p)$。它满足$P(X=k) \geqslant 0$，$\sum_{k=0}^{n} C_n^k p^k q^{n-k} = (p+q)^n = 1$。对于二项分布，当$k$增加时，概率$P(X=k)$随之增加，直至达到最大值，然后随着$k$的继续增加，概率单调减少。

(4) 泊松分布：设随机变量X的概率分布为

$$P(X=k) = \frac{e^{-\lambda}\lambda^k}{k!}, k = 0,1,2,\cdots$$

其中$\lambda > 0$，则称随机变量X服从参数为λ的泊松分布，简记为$X \sim P(\lambda)$，它满足$P(X=k) \geqslant 0$，$\sum_{k=0}^{n} P(X=k) = \sum_{k=0}^{n} \frac{e^{-\lambda}\lambda^k}{k!} = e^{-\lambda}e^{\lambda} = 1$。

(5) 几何分布：在n重伯努力试验中，每次试验事件A发生的概率都为p，直到事件A发生为止，所进行的试验次数X服从几何分布。X的概率分布为

$$P(X=k) = pq^{k-1}, k = 1,2,\cdots$$

对于非离散型随机变量，由于所有可能的结果不能一一列举，因此不能用分布率来表示。我们所关心的是这种随机变量落在某一范围内的概率，而不是它取某个值的概率。如等车时常关心的是多长时间等到车的概率，而不是某一时刻等到车的概率；考察产品的使用寿命，关心的是产品的寿命大于某个值的概率，而不是等于某值的概率等等。为此，引入随机变量分布函数的概念：设X是一个随机变量，对于任意实数x，称$F(x) = P(X \leqslant x)$为随机变量X的分布函数。

不论随机变量是离散型随机变量或非离散型随机变量，分布函数$F(x)$全面地描述了随机变量X的统计规律性。

7. 概率密度

如果对于随机变量X的分布函数$F(x)$，存在非负函数$f(x)$，使对于任意实数x，有$F(x) = \int_{-\infty}^{x} f(t)\mathrm{d}t$，则称$X$是连续型随机变