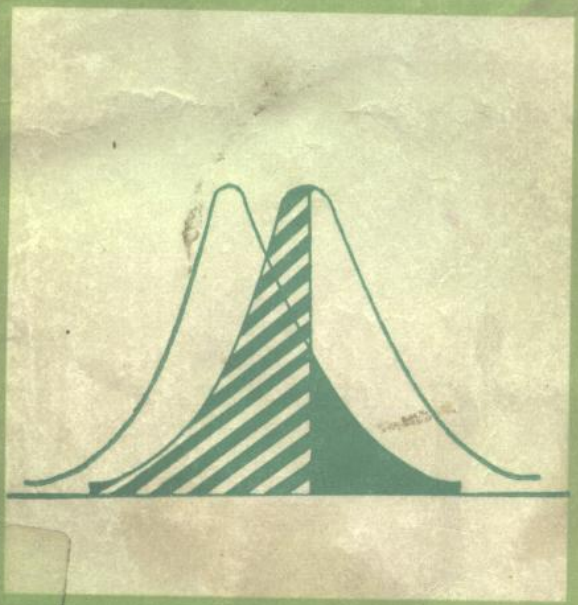
医学统计学基础

汤旦林　编著



卫生统计应用丛书

医学统计学基础

汤旦林　编著

田凤调　审

人民卫生出版社

医学统计学基础

汤旦林　编著

人民卫生出版社出版
（北京市崇文区天坛西里10号）
河北省遵化县印刷厂印刷
新华书店北京发行所发行

787×1092毫米32开本　7⅛印张　152千字
1989年10月第1版　1989年10月第1版第1次印刷
印数：00,001—10,000
ISBN 7-117-01093-2/R·1094　定价：2.70元
〔科技新书目200—147〕

卫生统计应用丛书编写说明

为了提高我国卫生统计知识水平，促进卫生统计工作的发展，更好地适应我国四化建设的需要，经过较长时间的同行酝酿，并经与人民卫生出版社协商，决定编写出版这套丛书。

本套丛书以介绍卫生统计基础知识、基本方法为主，注意实用性、科学性。既照顾到读者实际接受的可能性，又要求反映出时代的特点，介绍新的内容。

主要读者对象是：卫生统计专业工作者；医务人员和卫生防疫人员；医学界有关专业的科研与教学工作者；也可作为医学院校学生与研究生的参考书。

卫生统计应用丛书的选题包括下列几个方面：医学统计方法，居民健康统计，卫生资源统计，卫生业务统计，计算机应用技术，卫生统计工作改革等。每册一般为10～15万字，分批出版。

卫生统计应用丛书编委会

1988年12月

序

统计学的原理与方法，对医药卫生的科学研究与卫生事业管理水平的提高，都有重要的作用。无论是调查或实验研究的设计、数据信息的处理或电子计算机的医学应用，如能融会贯通地结合运用统计学的观点与方法，就能事半功倍。凡了解这一点的医药卫生科研、管理人员都有学好统计学的愿望。只是深入浅出、引人入胜的书还不多。

1982年汤旦林同志曾被邀请举办过“医用生物统计”系列讲座，随后在大家的鼓励下又进一步将讲稿整理出来，陆续发表在《中华医学杂志》等刊物上。我读过其中的一些单行本，觉得很有特色。现在他将这些讲稿整理成册，相信会对读者有学习、参考的价值。故乐以为序。

1988年7月29日

目　录

原书缺页

原书缺页

假设检验 常用于新药鉴定、病因分析、理化检验方法或技术水平的考核等等。如不善于运用统计学的显著性检验方法来把关，往往不易作出正确的判断。已往有不少的反面教训，“卤碰治百病”就是记忆犹新的一例。

抽样方法和分布理论 是统计学的重要基础。在制定正常范围、研究患者（或病原体或中间宿主）在空间或时间上的分布规律方面亦常用到。

相关与回归 用于病因学研究、发育或生理功能评价、各种预测、趋势分析和鉴别诊断等等。

方差分析 用于多组间的比较和多因素分析，可帮助我们明确各因素的主次和因素之间的交互影响。

概率单位方法 常用于药物、毒物以及物理因素等对机体的作用强度的鉴定与比较，以及它们的联合作用的分析。

多元分析 可用于疾病的诊断、鉴别、辨证、分型、预后、病因分析、新药筛选、治疗决策、动态观察以及诊疗自动化等。它和电子计算机的医学应用相辅相成。是很有发展和应用前途的一类方法。

不论分布方法 绝大多数经典的统计方法都建立在“数据为正态分布”这一基本假定之上，而医学数据多半并非正态分布，采用不论分布的方法就可以绕过这一困难。这类方法还有计算简便等优点。

寿命表方法 用于估计和比较不同处理后的生存率。

统计模型 用于复杂问题，包括多因素、多变量的深入细致的分析。例如：对数线性模型，用于多维列联表的分析；*logit*模型，用于多个致病因子的分析；*Cox*模型，用于生存曲线的描绘与比较等。

三、几个重要的基本概念

1. 个体差异 前已指出，各种观测指标都是(或可以转成）随机变量。对一个观察单位（例如一个7岁男孩，或一位孕妇）进行观测所获得的某种指标（如身高，或血压）的一个实际的数值（如身高116.1cm，或血压110/70mmHg）称为一个个体。同一来源的各个个体彼此并不相同，它们之间的这种差异称为个体差异。注意：这里定义的"个体"，指的是观测值，而不是观测对象。计数指标的个体用1或0表示，如阳性者为1，阴性者则为0。又如记男为1，则女为0。

人体（或人群）的各种观察指标的具体数值均有较大的个体差异，单个的观测值往往意义不大，所以有必要对相当数量的个体进行统计分析，个体差异愈大，进行统计分析的必要性也愈大。

2. 总体、样本、误差 来源相同的全部个体(例如所有7岁男孩的身高值，或所有孕妇的血压）称为一个总体。实验或调查研究的目的，都是从观测结果（一部份个体）去估计总体的情况。为了估计总体的情况而抽查的那部份个体就称为样本。样本对总体的代表性取决于抽样的方法、样本大小（含量）和个体差异的程度。

总体中的个体数目一般是较大的，往往是无限的。样本中的个体数目总是有限的，相对来说是较小的。由于存在着个体差异，用一小部份个体组成的样本去估计总体的情况难免会有些误差，这种误差称为抽样误差。如果选取对象或观测对象的方法或条件有偏向，也会产生误差，称为系统误差。前者是不可避免的，但我们可以估计出它的大小；后者是应

当也必须千方百计地加以防止的。防止系统误差是科研设计所要解决的主要问题之一。

3. 平均数与百分数 计算定量指标X的样本平均数$\bar{X}$的目的，是为了估计总体的平均数μ（它是一个未知常数）。在计算平均数之前，应检查一下原始数据的同质性，不可将不同质的数据（即来自不同总体的个体）混在一起算均数，这样会歪曲事物的本质。样本含量较小时，可以按下式直接算出均数。

$$\bar{X}=\Sigma X/n \qquad (1\text{-}1)$$

其中Σ是求和的符号，ΣX就是观察值X的合计，n是样本含量。当样本含量较大时，用上式就不方便，可先将数据按大小分组（表1－1），再用下式计算（无论是否利用计算器都远较上式方便）。

表1-1 161名7岁男孩身高值的频数分布

身高(cm) X	组中值 z	用划记法分组*	组频数 f	zf
104—	106	正正	10	1060
108—	110	正正正正正正一	31	3410
112—	114	正正正正正正正正正一	51	5814
116—	118	正正正正正正正正下	43	5074
120—	122	正正正正丁	22	2684
124—	126	正	4	504
合计			161	18546

* 若每人一张卡片，则用分卡法更方便。

$$\bar{X}=\Sigma zf/n \qquad (1\text{-}2)$$

其中z表示各组数据的代表值，又称组中值，它等于该组指

标的上、下限的平均值；f表示属于该组的个体数，又称组频数；所以Σzf相当于观察值的合计ΣX（误差很小，可以忽略）。

将表中数据$n=161$及$\Sigma zf=18546$代入（1－2）式，得平均身高为$\bar{X}=18546/161=115.2$(cm)。

如果数据的分布很不对称，例如尿铅、发砷、粉尘的直径或浓度等等，往往有少数的特大值，此时用平均数并不能恰当地代表数据的一般水平，可以采用中位数——即将数据按大小次序排列后位于正中的数。

平均数是从样本算出的统计量，所以用它来估计总体均数时还应考虑到抽样误差。关于这个问题，我们将在以后作进一步讨论。

对于计数的定性指标，常用阳性率（如生存率、患病率、治愈率等等）来描述或估计总体的情况。其实，阳性率也就是计数数据的平均数。例如：10人中有3人阳性，则阳性率为$P=\frac{3}{10}=0.3$（或30%）；若把阳性者的观测值X记为1，而将阴性者的观测值X记作0，则$\Sigma X=3$，$n=10$，平均数$\bar{X}=\Sigma X/n=3/10=0.3$，和前面算得的阳性率$P=0.3$相同。

阳性率常用百分数来表示。医学中常用的百分数有两类：反映机会大小的百分率和反映比重大小的百分比（又称构成比）。在使用或阅读时应避免混淆，否则容易导致错误。例如：在某省的肿瘤科研协作经验交流会上，一位县医院的医生说农民的患癌率最高；而另一位市医院的医生则说工人的患癌率最高。其实他们所依据的都是各自医院里的癌症患者职业构成比，而不是患癌率。正如临床病例分析中的性别比例、年龄构成等数据，也都是构成比，它们主要与医院的服