# 医学应用统计分析

## （SAS、SPSS版）

主编　陈青山

人民卫生出版社

# 医学应用统计分析

（SAS、SPSS 版）

主　编　陈青山

副主编　尹　平　吴泰顺　程锦泉

人民卫生出版社

**图书在版编目（CIP）数据**

医学应用统计分析/陈青山主编.—北京：人民卫生出版社，2015

ISBN 978-7-117-20876-5

Ⅰ.①医… Ⅱ.①陈… Ⅲ.①医学统计-统计分析-软件包-教材 Ⅳ.①R195.1-39

中国版本图书馆 CIP 数据核字(2015)第 129123 号



医学应用统计分析

主　　编：陈青山
出版发行：人民卫生出版社（中继线 010-59780011）
地　　址：北京市朝阳区潘家园南里 19 号
邮　　编：100021
E - mail：pmph @ pmph.com
购书热线：010-59787592　010-59787584　010-65264830
印　　刷：北京盛通印刷股份有限公司
经　　销：新华书店
开　　本：787×1092　1/16　　印张：13
字　　数：316 千字
版　　次：2015 年 6 月第 1 版　2015 年 6 月第 1 版第 1 次印刷
标准书号：ISBN 978-7-117-20876-5/R · 20877
定　　价：45.00 元
打击盗版举报电话：010-59787491　E-mail：WQ @ pmph.com
（凡属印装质量问题请与本社市场营销中心联系退换）

# 编　委

（以姓氏笔画为序）

# 序

我们正在进入一个信息时代。

人们所指的信息时代是信息传输速度快和信息量大的时代。信息除了文字和声像外还包括数据，数据赋予信息更具体简洁的内涵。统计学是收集、分析、表述和解释数据的科学，作为数据分析的一种有效工具，统计方法已广泛应用于医学科学各个领域，是医学科学工作者和科学研究者不可或缺的知识和技能。

在计算机日益普及的时代，教材编委结合长期教学和实践经验，潜心研究出一套实用型统计学教学方法以及由一个基本配置和若干可选的常用统计部件构成的统计计算软件包，使用者只要把收集到的数据按要求建好数据库，分清变量类型，正确选用所需要的统计学方法，录入数据就能够获得所需的统计结果。把异常复杂和难于操作的统计学方法综括为“明确分析目的-建好数据库-分清变量类型-正确选择统计方法”，是本书的指导思想和编写特色，是淳朴的科研假设经过艰苦的研究转化为便捷结果的方法。上百次的尝试和一而再的失败与成功，从二项分类变量、多项无序分类变量、多项有序分类变量到数值变量，逐步深入浅出，最终囊括了科学研究中常用的统计学方法，满足了科学研究数据的常用统计分析需求，达到了简单、快捷、实用的效果。不论是阳春白雪或是下里巴人，只要通读一遍本书，按图索骥进行学习与操作，就能够驾轻就熟、如愿以偿地获得所需要的统计结果。

实践过这种统计学方法的人都异口同声地说，医学应用统计分析竟然如此简捷方便和不必求人，于是令我想起“前人种树后人乘凉”的成语。编委们穷其全力、锲而不舍，创建了“目的-数据库-变量类型-变量间关系分析”的应用统计学教学方法，为统计学应用者找到了一条便捷实用的学习途径。这种不弃不舍、孜孜以求的科学态度和甘为人梯的专业责任心令人肃然起敬。

《医学应用统计分析》就像一叶来自于远方的帆船，给人带来便捷的希望。两岸青山相对出，孤帆一片日边来。

国家教学名师 **王声湧**

2015 年元月于暨南园

# 前　言

统计学是一门理论科学，也是一门应用科学，在实际工作中有着广泛的用途。但统计学难学，统计学难用，究其原因或许各种各样，不是老师没有教好，不是教材没有编好，更不是学生没有学好。主要的客观因素在于众多的学习者，如绝大多数非数学类专业（医学、商学、管理学等）学生，只是学过或没有学过高等数学、只具备一些基础数学知识，而没有更多的概率论和数理统计等相关课程的背景知识，所以学起来吃力，用起来费劲。

基于该现实，改变传统统计学教材的“方法-原理-公式-应用”编写模式，不以复杂的原理方法、公式推导为重点，以解决实际问题为目的，内容简单、方法明确、易于理解、便于操作，按“目的-数据库-变量类型-变量间关系分析”新型教学模式编写应用统计学教材是现实的需要，也是一种大胆的创新。

本书遵循简单、实用的原则，力避复杂的数学原理和公式推导，以明确分析目的，建好统计数据库，分清变量类型为基础，以分析变量与变量之间关系为主线阐述统计学的分析方法。本书编写中力求思维清晰明确、内容实用简单、操作便捷可行、安排合理有序，使之成为一本学术思想先进、实用价值较高、适用对象更加广泛的统计学教材。

本书的主要内容包括基础部分：统计学基本概念、思维方法，建数据库，分清楚变量类型，选择统计方法；分析部分：单一变量分析，两变量关系分析，多变量关系分析。其中两变量关系分析是教材的主要内容。

该书的编写目的是冀望提高广大学习者实际应用统计学的能力，但限于水平，不一定遂愿，诚恳期待广大读者的批评指正并不断完善。

主　编　**陈青山**

2015 年元月

# 目 录

# 第一章

# 绪　论

统计学是一门透过同质事物的变异性、揭示内在事物规律性和实质性的科学。确切地讲，是一门关于客观数据分析的科学，研究数据的收集、整理和分析，在实际工作中，有着广泛的用途。

应用统计分析是应用者围绕应用分析的目的，根据数据或数据库中变量的特征和类型，以及变量与变量间关系所实施的数据分析。学好应用统计分析需要掌握一些统计学的基本概念、基本特征以及基本学习方法。

## 第一节　统计学的几组基本概念

应用统计分析有几组较为重要的基本概念，掌握它们可以全面地理解什么是统计学。

### 一、指标与变量

指标（index）即观察指标，是由研究目的确定的观察对象的内在属性特征或其相关的影响因素。例如，需要研究某地区饮用和不饮用早餐奶等对小学生身体生长发育（如身高、体重等）的影响，那么身高、体重反映了小学生身体生长发育的特征，分别称为研究的身高指标、体重指标，影响身高体重的性别、年龄等因素，称为研究的性别指标、年龄指标等。

变量（variable）即观察变量，也称变化的量，实际上就是观察指标，一般特指用于数学、统计或软件计算的分析指标。例如，反映小学生身体生长发育的身高、体重指标，在统计计算时，分别称为身高变量、体重变量。

某一变量的观察值或测量结果即为变量值，如测得某个小学生的身高 1.20m、体重 30kg，可分别称该小学生身高的变量值为 1.20m、体重的变量值为 30kg。

笼统地讲，统计学是一门关于变量（实际上是变量值）分析的科学。

### 二、影响变量与结果变量

变量按是否影响其他变量，或是否受到其他变量的影响分为影响变量和结果变量。影响变量（affect variable），也称自变量（independent variable），是指自身变化并影响结果变量变化的量；结果变量（outcome variable）又称因变量（dependent variable）或反应变量

(response variable)，是指受到影响变量的影响而变化的量，看作影响变量变化的结果。如果分析某地小学生体重依赖于年龄的变化规律，那么年龄可看作是影响变量，体重则为结果变量；如果分析不同性别之间身高是否存在统计学差异，那么性别是影响变量，身高是结果变量。

分清楚变量特征，即分清楚结果变量与影响变量，是选择统计分析方法的重要步骤。一般而言，那些相对固有的、不易改变的指标（如性别、籍贯等），或易于被人控制的处理因素（如实验分组、疫苗接种与否等）作为影响变量或影响因素；而那些容易变化，或较难确定的观察效应或结局指标（如疗效、患病与否等）作为结果变量，看成是最后观察或反应的结果。但影响变量和结果变量的划分是相对的，视研究目的和具体情况而定，有时甚至不加区分。

可以讲，统计学是一门关于结果变量与影响变量（简称变量与变量）间关系分析的科学。

## 三、总体与样本

总体（population）是根据研究目的确定的同质观察单位的全体，更确切地说，是同质的所有观察对象某变量值的集合。笼统地讲，总体可以是一个社区、一个特定的人群、一组血样、一群细胞等；具体而言，总体是所有观察对象的某个观察指标（即变量）的全部观察值。例如，在饮用和不饮用早餐奶对某地区小学生身体生长发育影响的研究中，该地区符合条件的所有小学生常常被认为是该研究的总体，实际上还要具体区分不同指标的总体，该研究的身高总体是所有研究对象的身高值，该研究的体重总体是所有研究对象的体重值。研究的总体中，有的研究对象（或变量值）的个数是可数的，称为有限总体，有的是不可数的，称为无限总体。

在实际应用中，由于往往无法或者没有必要得到总体中每个变量的值，所以常常应用随机抽样的方法研究其中的某一部分。所谓随机抽样，就是一种从总体中随机抽取具有代表性的部分个体进行统计分析并用来研究总体的方法。从总体中随机抽样获得的部分观察对象的变量值称为样本（sample），样本中变量值的个数称为样本含量（sample size）。

已经证明，一定样本含量的样本信息可以推断其总体的相关特征。从这个意义上讲，统计学是一门研究样本，推论总体的科学。

## 四、同质与变异

同质（homogeneity）是指研究对象具有相同或相近的性质、条件或影响因素。在上述早餐奶对某地区小学生身高体重影响的研究中，该地区全体小学生可认为是同质的，因为这些研究对象具有相同的地域、相同的身份、相近的年龄……许多研究中常常给出筛选对象的诊断标准、纳入标准和排除标准，目的就是为了保证研究对象的同质性。

同质研究对象的某些研究特征又具有差异性，这种现象称为变异（variation）。在早餐奶的研究中，该地区全体小学生具有同质性，但他们的身高有高有矮、体重有轻有重……表现为变异。

同质总体中个体间的变异是绝对的，这是统计学赖以存在的基础。从这个角度来看，统计学是一门研究变异的科学。

## 五、参数和统计量

参数（parameter）是描述研究总体特征的指标。用希腊字母代表，如：总体均数$\mu$、总体率$\pi$、总体标准差$\sigma$等。

统计量（statistic）是根据样本的变量值计算的、描述样本特征的指标。用拉丁字母代表，如：样本均数$\bar{x}$、样本率$p$、样本标准差$S$等。

在总体参数未知时，常常通过样本的统计量对总体参数进行估计或假设检验。所以，统计学是一门研究样本统计量估计总体参数的科学。

## 六、本质差异和抽样误差

不同样本的统计量或分布存在不同程度的差异，常有两个原因：一是本质差异，二是抽样误差。

本质差异（essential difference）是指不同的研究因素影响或作用于不同的研究总体，导致不同总体参数之间或相应样本统计量之间存在的差异。例如，饮用和不饮用早餐奶可引起两组身高体重的不同，视为研究因素导致的本质差异。

抽样误差（sampling error）是指由于随机抽样的原因引起的样本统计量与总体参数或不同样本统计量之间的差异。例如，饮用同量早餐奶的全部小学生平均身高1.20m，随机抽取了其中10名小学生的平均身高为1.19m，这两个平均身高不等视为抽样误差。又如，饮用同量早餐奶的小学生如果用随机分组方法分成两组，一般来讲两组的平均体重不完全相同，也可看作是抽样误差。

引起抽样误差的直接原因是随机抽样，内在原因是总体中个体间的变异。因为个体变异的绝对性，所以抽样误差不可避免，但抽样误差的大小可用统计学方法予以估算。从此意义来理解，统计学则是一门研究抽样误差的科学。

## 七、正态分布与偏态分布

正态分布（normal distribution），又称为高斯分布（Gaussian distribution），是一种常见的、具有以均数为中心、左右两侧基本对称、钟形、两头低中间高等特征的连续型分布。统计学上把以均数为$\mu$、方差为$\sigma^2$的正态分布记作$N(\mu, \sigma^2)$，其中$\mu=0$，$\sigma^2=1$的正态分布称为标准正态分布，记作$N(0, 1)$。大多数医学数据呈正态分布或近似正态分布，有的数据尽管不呈正态分布，但经适当的变量变换，可使变换后的数据服从正态分布或近似正态分布。

偏态分布（skewed distribution），是一种较为常见的、没有或缺少正态分布曲线特征的连续型分布，表现为分布曲线的峰值与平均值不相等，即不以均数为中心，左右两侧明显不对称。根据曲线峰值小于或大于平均值可分为正偏态分布或负偏态分布。

在某种程度上来讲，统计学是一门研究数据分布的科学。

## 八、频率与概率

一枚硬币，投掷10次，如果观察出现正面的次数，可能为1次、2次、3次……10次

或0次，计算这10次投掷中出现正面次数与总投掷次数之比，就是计算投掷10次硬币出现正面的频率。一般认为，频率（frequency）是在有限少量次数如几次或几十次试验中，某现象出现的次数与总试验次数的比值。

当投掷硬币的次数不断增加，正反面出现的次数与总次数的比值将逐渐接近50%。可以设想的是，当投掷无限多次时，正面或反面出现的频率就是50%，此即为投掷硬币出现正面或反面的概率。

可见，概率（probability）是在无限多次试验中，某现象出现的次数与总试验次数的比值，或者说是频率的极限值。它反映某一事件发生的可能性大小，常以符号 $P$ 表示，$P$ 越接近1表示该事件发生的可能性越大，$P$ 越接近0表示该事件发生的可能性越小。其取值范围在0到1之间，可以用小数或百分数表示。

所以，统计学也是一门研究概率大小的科学。

## 第二节 应用统计分析的实质和基本特征

站在不同的角度，对统计学有不同的理解和认识，但统计学的实质内容就是数据分析，包括理论和应用两部分。

理论统计学是研究数据分析的原理、方法、条件和公式等。

应用统计学则是应用现代计算机技术（包括软件技术）和理论统计学的成果，围绕分析目的，分析实际数据中变量与变量间的关系。

从这个意义上来讲，应用统计分析属于应用统计学的范畴，是从解决实际问题的角度阐述如何应用统计学方法，因此具有应用统计学的一些基本特征。

1. 实用性 解决实际数据的统计分析问题，不涉及或尽量少涉及统计理论、公式推导等内容，甚至不太多的考虑其计算公式或中间的计算过程等。

2. 目的性 有明确的实际应用目的。一堆杂乱无章、没有任何分析目的的数据是没有价值的，尽管理论上有很好的分析方法。

3. 数据性 某种意义上，统计分析就是数据分析，因此收集的数据要按照数据间关系、数据库的要求进行整理、呈现，建立的数据库能被统计软件调用，并按目的要求进行分析。

4. 借用性 借用理论统计学的研究成果和现代计算机的科学技术（包括软件技术）解决实际问题，主要强调如何应用、如何得出结果。

## 第三节 学好应用统计分析的方法

如何学好应用统计分析，不能一言以蔽之，但在学习中，需要掌握以下几点。

### 一、明确分析目的

研究目的是统计分析的目标和方向，决定了研究设计、研究对象、研究指标等，而研究的设计方案、分析指标是选择不同统计分析方法的决定因素。因此，正确的统计学分析一定要建立在明确的研究目的基础上，那些没有目的的统计分析，或者事先没有研究设

计，事后找来一堆数据的统计分析都是不可取的。

## 二、建好分析数据库

一般来讲，统计分析需要借助统计分析软件计算，而统计分析软件都要有完整、符合要求的数据或数据库，所以建好分析数据库是统计分析的必要条件。此外，建好分析数据库还可以理清分析思路。在试验或调查研究中获取的数据有时多而零散，如果不能进行科学的整理汇总，就会杂乱无章，理不清头绪，抓不住要点，甚至无所适从，最后可能束之高阁、弃之不用，造成数据的极大浪费。相反，建好数据库，可以使观察对象的研究指标一目了然，使研究思路清晰明确。因此，建好数据库是正确统计分析的前提和基础。

## 三、分清楚变量类型

数据库中各个研究对象的每项观察指标都可以看作是一个分析的变量，变量的类型是统计分析中选择不同统计方法的依据，分清楚变量的类型是正确选择统计方法的基础和关键。变量分为数值变量和分类变量两类，其中分类变量按是否有序以及项数的多少，又分为二项无序、多项无序、二项有序、多项有序几种类型。实际应用中，常常将二项无序分类变量和二项有序分类变量合并为二项分类变量，详见第二章。

## 四、正确选用统计学方法

统计学分析可看作是变量与变量间的关系分析，当研究目的和设计方案确定以后，不同特征类型的变量组合决定了不同统计方法的选择。如，二项分类变量与二项分类变量关系的分析选用$\chi^2$检验，数值变量与二项分类变量关系的分析选用 $t$ 检验，数值变量与多项无序分类变量关系的分析选用 $F$ 检验，数值变量与数值变量关系的分析选用直线相关回归分析……。详见第三章以后各章节。

## 五、熟悉常用的统计分析软件

统计分析软件是统计分析的必备工具，目前有许多种（套）。常用的国际公认的统计分析软件有：统计分析系统（SAS）、社会学统计程序包（SPSS）。微软公司的电子表格系统 Microsoft Office Excel 也有广泛应用。

### （一）统计分析系统（SAS）

SAS（Statistics Analysis System）是统计分析系统的英文缩写，最早由北卡罗来纳大学的两位生物统计学研究生编制，1976 年由 SAS 软件研究所正式推出。SAS 完全针对专业用户进行设计，以编程为主。其最大特点是分析模块调用，功能强大，深浅皆宜，简短编程即可同时对多个数据文件进行分析。但对一般用户而言，人机界面不太友好，初学者编写、使用程序会存在各种难度。本书介绍的是 SAS 9.2 版本的程序。

### （二）社会学统计程序包（SPSS）

SPSS（Statistical Package for the Social Science）是社会学统计程序包的英文缩称，20 世纪 60 年代末由美国斯坦福大学的三位研究生研制，1975 年由芝加哥 SPSS 总部推出。SPSS 系统的最大特点是菜单操作，方法齐全，绘制图形、表格较为方便，输出结果比较直观。但其统计分析功能略显逊色，特别是难以同时分析处理多个数据文件。本书介绍的

是 SPSS 13.0 版本的程序。

**（三）微软公司的电子表格系统**（Microsoft Office Excel）

Microsoft Office Excel（简称 Excel）是美国微软公司开发的电子表格系统，是目前应用最为广泛的办公室表格处理软件之一。Excel 作为 Office 软件的一员被众多用户所熟知，具有数据处理、函数运算、数据库、图表制作等功能，进行统计分析时具有易得、快速、直观、简单、运算可视等优点，也是建立数据库、进行常用统计分析的好工具。

不同软件各有利弊、互有长短，用户可根据需要和使用习惯，选择一种或几种软件进行数据分析。本书仅介绍 SAS、SPSS 的计算方法和程序，Excel 统计分析另书出版，其他不作阐述。

# 第二章

# 变量、数据和数据库

统计数据是实施统计分析的前提和基础，常常以数据库的方式呈现。数据库由不同观察对象的观察指标（变量）及其相应的数据值组成。掌握变量、数据和数据库的基本知识，正确区分变量、数据类型有利于正确选择统计方法并实施统计分析。

## 第一节 变 量

从数据库、数据分析的角度来看，变量是指能反映数据库数据的内在数量关系、可用于统计计算（包括软件计算）的指标。一般而言，不同的研究目的决定了不同的数据库，实际上决定了组成数据库的不同变量。

弄清楚变量类型及其转换关系是应用统计分析的重要内容。

### 一、变量的类型

变量分为分类变量和数值变量两种类型。

#### （一）分类变量

分类变量（categorical variable），又称定性变量（qualitative variable），是指用定性方法确定的、说明观察单位某项属性特征或类别的指标。

在有关分类变量的统计分析中，由于选择的统计方法与分类变量的分类个数、分类类别之间是否存在等级或程度差异等有关，因此根据分类变量的分类项数和各项数间有无等级程度差异分为二项分类变量（包括二项无序分类变量和二项有序分类变量）、多项无序分类变量和多项有序分类变量，见表2-1。

**表2-1 不同类别的分类变量**

| 类别 | 项数 | 等级次序 | 举例 |
|---|---|---|---|
| 二项分类变量 | 二项 | 无或有 | 性别（男、女）、结果（阴性、阳性） |
| 多项无序分类变量 | 多项 | 无 | 血型（A、B、AB、O） |
| 多项有序分类变量 | 多项 | 有 | 营养状况（优、良、中、差） |

#### （二）数值变量

数值变量（numerical variable），又称定量变量（quantitative variable），是指用定量方