sheji
yu
tongji
fenxi

外科科研设计与统计分析

胡良平　毛　玮　主编

中国协和医科大学出版社

外科科研设计与统计分析

[redacted]

外科科研设计与统计分析

胡良平　毛　玮　**主　编**
柳伟伟　王　琪　高　辉　**审　校**

编　者（以姓氏笔画为序）
毛　玮　王　琪　关　雪
吕辰龙　刘惠刚　李子建
李长平　周诗国　柳伟伟
胡良平　胡纯严　郭辰仪
郭　晋　贾元杰　高　辉
陶丽新　鲍晓蕾

中国协和医科大学出版社

图书在版编目（CIP）数据

外科科研设计与统计分析 / 胡良平，毛玮主编. —北京：中国协和医科大学出版社，2012.4
ISBN 978-7-81136-644-0

Ⅰ. ①外… Ⅱ. ①胡… ②毛… Ⅲ. ①外科学-科学研究-研究方法 ②外科学-医学统计-统计分析 Ⅳ. ①R6-3

中国版本图书馆 CIP 数据核字（2012）第 023783 号

外科科研设计与统计分析

主　　编：胡良平　毛　玮
策划编辑：周思龙　韩　鹏
责任编辑：田　奇

出版发行：**中国协和医科大学出版社**
（北京东单三条九号　邮编 100730　电话 65260378）
网　　址：www. pumcp. com
经　　销：新华书店总店北京发行所
印　　刷：北京佳艺恒彩印刷有限公司

开　　本：787×1092　1/16 开
印　　张：19.75
字　　数：400 千字
版　　次：2012 年 7 月第一版　2012 年 7 月第一次印刷
印　　数：1—3000
定　　价：38.00 元

ISBN 978-7-81136-644-0/R · 644

内容简介

本书分为3篇，第1篇为统计学内容概要，全面阐述了外科科研中常用的统计学内容，包括统计表达与描述、试验设计、定量与定性资料的统计分析、简单相关与回归分析、多重线性回归分析与多重logistic回归分析。此外，该篇中还详细讨论了掌握和运用统计学知识的要领和技巧，对具体问题给出了详细的SAS程序和结果解释，方便广大读者学习与使用。第2篇为外科科研中常见统计学错误辨析与释疑，紧紧围绕第1篇中的要点，以外科学（特别是骨科）相关杂志中近几年刊登的论文为主要的资料来源，列举了一些典型的科研案例，并对其在统计学方面出现的各种错误类型和原因进行深入剖析，给出了识别错误的技巧和正确处置的策略，避免其他工作者重蹈覆辙。第3篇为医学统计学要览，以“问题引导”的形式提纲挈领地介绍了“科研设计要览”与“统计分析要览”，为读者正确把握科研设计要领、合理选择统计分析方法和深刻领悟统计学的精髓，奠定了坚实的基础。本书力求通俗易懂、简明扼要，应用性强，富有启发性，注重对读者基础知识的训练和综合应用能力的培养，各章配以丰富的实例，以便读者学习、理解和正确运用。本书可供从事外科学基础和临床研究的科研和临床工作者使用，也可供生物医学领域的其他科研工作者和临床医生、杂志编辑与审稿专家、本科生、研究生参考和借鉴。

前　言

从表面上看，外科（特别是骨科）研究中不像内科和其他学科那样对统计学有极多的需求。事实上，只要做研究，就少不了要考虑多因素影响下多指标的变化规律，因此，外科科研对统计学的需求并不少于其他学科。

因为事物万变不离其宗，只要研究者不被事物的表面现象所迷惑，善于运用辩证唯物主义的思维方法，并以正常人的心态来思考和处置问题，就十分有利于透过现象看清事物的本质，完全有能力驾驭任何科研领域的研究问题。事实上，只要人们掌握了正确的统计学思想，巧妙地运用三型理论（任何事物都存在表现型、原型和标准型，弄清每个具体问题中的这三型，再有的放矢地去解决它，问题也就迎刃而解了），制定出科学完善经济可靠的科研设计方案，在好的方案指导下，注意试验或调查过程中的质量控制，正确地收集科研资料并对其进行合理的分析，所得到的科研成果是经得起时间和实践检验的。为此，笔者总结了外科科研中人们常用的科研设计和统计分析理论与方法，还总结了人们在该科研领域中常犯的统计学错误并对其进行了辨析与释疑，提纲挈领地对前述内容进行了归纳和总结，这正是呈现在读者面前的这本《外科科研设计与统计分析》拙著，但愿它能有利于外科学、特别是骨科学实际工作者针对自己要解决的问题，有的放矢地去学习和运用。

本书包含了15章内容，分列入3篇。第1篇统计学内容概要，介绍了统计表达与描述、试验设计、单因素设计定量资料统计分析、多因素设计定量资料统计分析、单因素设计定性资料统计分析、多因素设计定性资料统计分析、简单相关和回归分析以及多重回归分析，共8章；第2篇外科科研中常见统计学错误辨析与释疑，介绍了统计表达和描述错误与释疑、试验设计错误辨析与释疑、定量资料统计分析错误辨析与释疑、定性资料统计分析错误辨析与释疑、相关和回归分析错误辨析与释疑共5章；第3篇医学统计学要览，以“问题引导”的形式提纲挈领地介绍了“科研设计要览”与“统计分析要览”，为读者正确把握科研设计要领、合理选择统计分析方法和深刻领悟统计学的精髓，奠定了坚实的基础。总之，本书从正反两个视角，全面系统地介绍了外科研究中涉及的试验设计、统计表达与描述、统计分析和SAS实现等方面的内容。

在本书即将出版之际，笔者要衷心感谢我的硕士研究生毛玮。他在本科阶段就读于北京大学医学部，打下了比较坚实的医学基础；又在中国疾病预防控制中心工作了若干年，积累了一些宝贵的实践经验。在硕士阶段，攻读流行病与卫生统计学，潜心研究综合评价方法，同时还完成了本书大部分初稿。在此基础上，笔者结合多年统计教学和咨询经验，补充撰写了第3篇，使本书的理论水平和实用性得到了进一步提升。笔者由衷地感谢本室柳伟伟讲师和在读博士研究生王琪，他们认真负责地对本书初稿进行了通读，纠正了一些差错，并提出了许多宝贵的修改意见。笔者还要感谢本室其他教师和研究生，为本书的质

量提高和校对工作付出了辛勤的劳动。

本书叙述力求通俗易懂、简明扼要，富有启发性，应用性强，便于自学，注重对读者的基础知识的训练和综合应用能力的培养，各章配以丰富的实例，便于读者学习和使用。本书内容不仅适合于从事外科科研工作的人们，也适合于一切从事生物医学、临床各科研究的人们和大学本科以上的学生及学者学习与使用。

由于笔者水平有限，书中难免会出现这样或那样的不妥，甚至错误之处，恳请广大读者不吝赐教，以便再版时修正。

胡良平
于北京军事医学科学院生物医学统计学咨询中心
2012年4月

目　录

第一篇　统计学内容概要

第二篇 外科科研中常见统计学错误辨析与释疑

第三篇 医学统计学要览

第一篇 统计学内容概要

第一章 统计表达与描述

统计表达与描述是指采用统计表、统计图以及以平均水平与变异程度相结合的形式概括统计资料的信息，它是统计推断的前提和基础。统计表达与描述的任务是用恰当的方式高度概括地呈现资料的主要信息，它具有简洁明了、生动形象、使用方便的特点。本章将系统介绍常用的统计描述指标以及编制统计表和绘制统计图的技巧。

第一节 资料类型

一、资料类型划分

统计资料的现代划分方法将资料分为定量资料和定性资料两大类（图 1-1），正确识别资料类型是合理选用统计分析方法的前提条件。

（一）定量资料 其测定值表现为大小不等的数值，一般带有度量衡单位。定量资料分为离散型和连续型两种，分别对应计数资料和计量资料。

计数资料：在定量资料中，如果测定值是只能取零和整数（通常只取正整数）的情况，这种资料属于计数资料，如每分钟脉搏跳动次数、育龄妇女生育子女数等。

计量资料：在定量资料中，若测量值可以取区间内任意值，这种资料称为计量资料，如身高、体重等。

（二）定性资料 通过观测观察单位的属性所得的资料，称为定性资料，又叫做分类资料。定性资料可根据测定指标属性的分类多少，分为二值资料和多值资料。多值资料又可根据测定指标是否有等级关系，分为多值名义资料和多值有序资料。

1. 二值资料：在定性资料中，若观测值只具有相互对立的两种情况，则资料属于二值变量，又称为二分类变量。例如，只有男、女两种情况的性别变量及其观测结果。

2. 多值名义资料：在定性资料中，若观测值的属性无等级之分，称为名义资料。例如，某单位全体员工按 ABO 血型系统可分为 A 型、B 型、AB 型、O 型，显然这些血型之间并没有等级之分。

3. 多值有序资料：在定性资料中，若观测值的属性有等级之分，称为多值有序资料。例如，某病患者治疗后的疗效可划分为治愈、显效、好转、无效、死亡 5 个等级，这时属性变量不同取值之间是有好坏程度之分的。

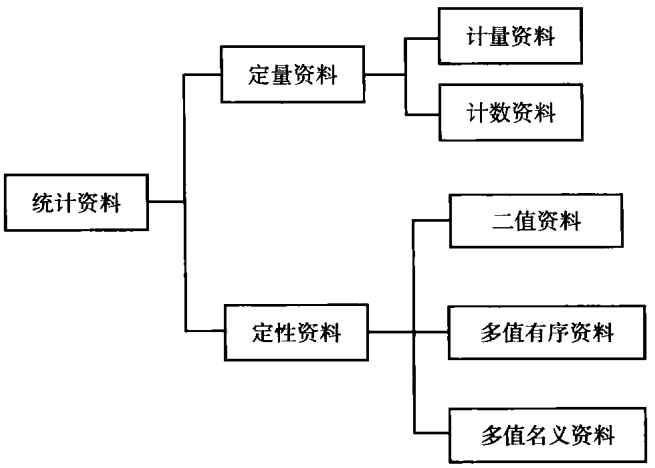

图 1-1 资料类型的现代划分方法

二、资料类型的本质

资料类型是从资料性质的角度来划分的，也可以说资料性质决定了资料类型。有时资料的性质很难识别，因此还应给出必要的描述信息，例如，当我们看到一个数值“1”时，不能确定这个“1”代表什么含义，它可以表示 1 个人、1 厘米、甚至是等级 1。

在统计软件中，不可能通过大量文字来描述变量信息，通常有固定的方式规定变量的属性（数值型或字符型）和性质（定量变量还是定性变量）。例如，若未作特殊说明，SAS 软件中的数值型变量一般视作定量变量，这时的变量值 1，可能是带有量纲的 1 厘米、1 小时等。当使用“freq”和“class”等语句指明把当前变量看作“频数变量”或“分类变量”，这时的变量值 1，就可能是频数为 1，或者等级为 1。在 SPSS 软件中新建变量时，通常需要指明变量性质，说明是度量变量、有序变量还是名义变量。

三、资料类型的转换

根据研究需要，有时可以进行定量资料和定性资料的相互转换。一般来说，定量资料包含的信息较定性资料丰富，而定性资料较定量资料的描述更简洁。定量资料转换为定性资料时通常比较方便，只需要明确划分标准即可；而定性资料转换为定量资料时，通常因为信息不全使得转换起来比较困难。

例如，假设 10 名患者的年龄（岁）分别为：19、23、26、31、42、44、53、54、58、67，显然这是一个定量资料。如果把年龄按照一定的标准（如 <30、$30\sim50$、$\geqslant50$）来划分，这时年龄这个指标可以从低到高由“1”“2”“3”三个级别来划分，该资料便由定量

资料转化为有序资料，每个级别患者的人数如表 1-1 所示。

表 1-1 10 名患者年龄的分级情况

年龄分级	例数
1（<30）	3
2（30～50）	3
3（≥50）	4

第二节 定量资料的统计描述

对定量资料的统计描述，常从平均水平和离散程度两个方面进行。平均水平和离散程度是资料描述的两大特征，平均水平不言而喻，常用的描述统计量有均数、中位数等。离散程度反映资料的波动范围，常用的描述统计量有标准差、方差和变异系数等。

此外，我们常对资料是否服从某种分布来做假设检验，例如，对正态性作假设检验可能会涉及若干统计量，如偏度系数、峰度系数等，它们可以反映定量资料与正态分布的符合程度。常用的统计描述指标如表 1-2 所示。

表 1-2 定量资料常用的统计描述指标及适用场合

描述内容	指标	意 义	应用场合
集中趋势	算术均数	所有观测值的平均值	对称分布定量资料
	几何均数	基于对数变换的平均值	对数正态分布定量资料
	中位数	全部数据按大小排序后，位次居中的数值	各种分布的定量资料
	众数	出现频率最高的观测值	各种分布的定量资料
	调和均数	基于倒数变换的平均值	正偏态分布定量资料
离散程度	全距	观测值最大值与最小值之差	各种分布的定量资料
	标准差（方差）	观测值离开算术均数的变异程度	对称分布定量资料
	四分位间距	中间半数观测值的全距	各种分布的定量资料
	变异系数	标准差与均数的比值	各种分布的定量资料

一、对称分布定量资料的统计描述

对于呈对称分布定量资料，常用算术均数描述其平均水平，用标准差（或方差）描述其变异程度。

（一）算术均数（arithmetic mean） 简称均数（mean），它描述一组服从对称分布定

量资料的平均水平。

均数把一组性质相同的观测数据转变为一个代表性的数值，高度提取了所有观测数据所反映的信息。因此，用均数来概括资料的平均水平，包含资料的主要信息。而另一方面，均数将观测数据之间的差异性掩盖起来，并且受极端值的影响明显。所以，它适用于表达一组服从或近似服从正态分布的定量资料。

通常用 $\bar{X}$ 表示样本均数，用 μ 表示总体均数。均数的计算方法有直接法和加权法两种。直接法的计算公式为：

$$\bar{X} = \frac{X_1 + X_2 + \cdots X_n}{n} = \frac{\sum X}{n} \qquad (1\text{-}1)$$

上式中 $X_1, X_2, \cdots, X_n$ 为各观测值。

【例 1-1】 现测得某地 8 只大鼠的骨密度（g/cm^2）如下：0.202，0.166，0.184，0.163，0.173，0.186，0.197，0.174，求这 8 只大鼠的骨密度均数。

因本例数据个数不多，目测数据变异程度不大，故以直接法求其算术均数。

$\bar{X}$ =（0.202 +0.166 +0.184 +0.163 +0.173 +0.186 +0.197 +0.174）/8 =0.181（g/cm^2）

当样本量较大时（如 n >30）时，常将原始数据整理成如表 1-3 频数表的形式，使用加权法来计算。

表 1-3 某地 100 名正常成年人非蛋白氮（mg/100ml）频数分布

非蛋白氮	组中值 X_i	频数	累积频数
22 ~	23	2	2
24 ~	25	5	7
26 ~	27	11	18
28 ~	29	18	36
30 ~	31	26	62
32 ~	33	21	83
34 ~	35	10	93
36 ~	37	6	99
38 ~40	39	1	100
合 计	-	100	-

注：频数表“频数”和“累积频数”两栏分别除以总频数，可得到各组段的“频率”和“累积频率”，从而形成频率表，加权法中也可以使用“频率”来作运算。

加权法的计算公式为：

$$\bar{X} = \frac{f_1X_1 + f_2X_2 + \cdots + f_kX_k}{f_1 + f_2 + \cdots + f_k} = \frac{\sum fX}{\sum f} \qquad (1\text{-}2)$$

上式中 $X_1, X_2, \cdots X_i$ 分别为各组段的组中值，$f_1, f_2, \cdots, f_i$ 分别为各组段的频数，相当于各组段组中值的权重。

【例 1-2】 基于表 1-3 资料求某地 100 名正常成年人非蛋白氮（mg/100ml）均数。

因本例数据以频数表形式呈现，从频数分布来观察，数据大致上符合对称分布，故以加权求其算术平均数。

$$\bar{X} = \frac{2 \times 23 + 5 \times 25 + \cdots + 1 \times 39}{2 + 5 + \cdots + 1} = \frac{3100}{100} = 31.0 \text{（mg/100ml）}$$

这 100 名正常成年人非蛋白氮的算术平均值为 31.0mg/100ml。

（二）方差（variance）和标准差（standard deviation） 标准差和方差均用于反映一组对称分布的观测值在数量上的变异程度。

个体偏离总体平均水平的程度为“$X - \mu$”，就是所谓的离均差（deviation from average），但是“$X - \mu$”的平均水平不能反映总体中个体值的变异程度，这是因为“$X - \mu$”有正有负，总和为 0，使用离均差的绝对值又不方便。而离均差平方可以消除正、负值的影响，为此，人们将离均差平方和的平均值作为总体中个体值偏离平均水平的概括性指标，称作总体方差（population variance），记为 σ^2。

$$\sigma^2 = \frac{\sum (X - \mu)^2}{n} \tag{1-3}$$

方差的量纲是原始数据量纲的平方，为了用原量纲表示变异程度，把总体方差开平方，将其算术平方根称为总体标准差（population standard deviation），记为 σ。

实际工作中总体均数 μ 往往未知，只能用样本均数 $\bar{X}$ 来估计 μ。若用 $\bar{X}$ 代替 μ，样本中的个体偏离 $\bar{X}$ 的程度比其偏离 μ 的程度缩小一些，以致离均差平方的平均值也缩小一些。英国统计学家 Gosset W S 提出用“$n - 1$”代替 n 来计算样本中离均差平方的平均水平，以纠正上述低估现象，于是样本方差（sample variance）计算公式如下：

$$S^2 = \frac{\sum (X - \bar{X})^2}{n - 1} \tag{1-4}$$

样本标准差（sample standard deviation）的公式如下：

$$S = \sqrt{\frac{\sum (X - \bar{X})^2}{n - 1}} \tag{1-5}$$

式中“$n - 1$”称为自由度（degree of freedom，df），记为 ν，有时也用 df 表示。

由于 $\bar{X} = \sum X/n$，有时使用式（1-5）经简单代数运算演化成的式（1-6）计算样本标准差更简便：

$$S = \sqrt{\frac{\sum X^2 - \frac{(\sum X)^2}{n}}{n - 1}} \tag{1-6}$$

自由度是统计学术语，其意义是随机变量能“自由”取值的个数。例如，一个 $n = 4$ 的样本，若已知 $\bar{X} = 5$，三个数据是“自由”取值的，一旦三个数据确定了，受到 $\bar{X} = 5$

这个条件的限制，第四个数也就确定了，这时自由度 $\nu = n - 1 = 4 - 1 = 3$。

标准差和方差越大，说明观测值偏离其算术均数的变异程度越大，这时均数对个体值的代表性就越差；反之，标准差和方差越小，则观测值偏离其算术均数的变异程度越小，均数的代表性也越好。

【例 1-3】 求例 1-1 中 8 只正常大鼠的骨密度的标准差。

此资料为一次小样本抽样所得定量资料，故采用式（1-6）计算标本标准差，由资料可得，$\sum X^2 = 0.262$，$(\sum X)^2 = 2.088$，

$$S = \sqrt{\frac{0.262 - \frac{2.088}{8}}{8 - 1}} = 0.012 \ (g/cm^2)$$

计算大样本资料的标准差和方差时，常将原始资料整理成如表 1-3 频数表的形式，然后根据加权法按式（1-7）计算。

$$S = \sqrt{\frac{\sum fX^2 - (\sum fX)^2 / \sum f}{\sum f - 1}} \tag{1-7}$$

上式中 X 为各组段组中值，f 为相应组段频数。

【例 1-4】 利用加权法计算表 1-3 资料中 100 名正常成年人非蛋白氮（mg/100ml）的标准差及方差。

有表中资料可算得，$\sum fX^2 = 97180$，$\sum fX = 3100$，根据公式（1-7）：

$$S = \sqrt{\frac{97180 - (3100)^2/100}{100 - 1}} = 3.3 \ (mg/100ml)$$

$$S^2 = \frac{97180 - (3100)^2/100}{100 - 1} = 10.9 \ (mg/100ml)^2$$

该地 100 名正常成年人非蛋白氮的标准差为 3.3mg/100ml，方差为 10.9 $(mg/100ml)^2$。

二、非对称分布定量资料的统计描述

对于非对称分布定量资料，常用中位数描述其平均水平，用四分位间距描述其变异程度。

（一）中位数（median） 中位数是将一组观测值从小到大按顺序排列，位次居中的那个数值，用 M 表示。在全部观测值中，小于和大于中位数的个体数相等。中位数可以应用于任何分布类型定量资料。

中位数是位置平均数，不受极端值影响，在具有极端值的数据中，中位数比算术均数更具有代表性，因此通常多用于偏态定量资料中，甚至是一端或两端无确切值的情况。例如，测量值超出仪器或试剂的测量范围，而无法获得测量结果。然而，用中位数描述定量资料会损失很多信息，而样本量较小时中位数不太稳定。

小样本中位数的计算方法：当样本量较小（如 $n < 30$）时，先将观测值按由小到大顺序排序，再按下式计算，

n 为奇数时 $M = X_{(\frac{n+1}{2})}$ (1-8)

n 为偶数时 $M = \frac{X_{(\frac{n}{2})} + X_{(\frac{n}{2}+1)}}{2}$ (1-9)