

第三章

临床检验数据的常规统计方法及应用

统计学是一门通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，包括理论和应用两个方面。

临床检验数据的分析与应用是医学应用统计学的重要方面，同样遵循“目的－数据库－变量类型－变量间关系”模式，以解决实际问题为导向，以建立统计学数据库、分清变量类型为基础，以分析变量与变量间关系为核心，在很多临床工作场景中解决问题。

检验医学是基础生物和自然科学（例如化学、分子生物学、微生物学、免疫学、计量学）与临床医学之间的桥梁和接口。在信息技术、统计学和数据分析领域的新方法引入医学领域时，检验医学一直处于最前沿，随着检验医学的发展，临床检验数据量越来越大，类型越来越丰富，维度越来越繁杂，对于统计学方法需求越来越多。如果在临床工作中能够真正了解临床检验数据的性质和类型，真正理解统计学方法的应用场景，会使我们获取和开辟一些新天地，开创临床实验室工作中的新技术。

第一节 临床检验数据分析前的准备

一、数据的准备

（一）明确数据的使用目的

临床检验数据具有其自身的特点，数据量大，类型丰富，维度繁杂，使用中应该以研究目的为导向，进行合理的研究设计，了解需要解决的实际问题。对于临床检验数据来说，应该明确研究的目的是解决临床检验工作中本身的问题，还是与其他临床

数据相结合以解决其他临床科学问题。

根据不同的研究目的,可以进行不同的研究设计,包括研究类型的选择、设计方案的规划、研究总体、样本量、观察指标、随机化分组、抽样方法的确定,以及数据质量控制和统计分析方法选取等,这些都会影响研究结果的可信度和科学价值。因此,正确的统计学分析一定要建立在明确的研究目的和研究设计的基础之上。

(二) 建立分析用数据库

临床中采集数据的来源有很多,包括患者登记信息、人口学信息、临床试验数据、健康监测数据、电子病历以及问卷调查等,临床检验数据只是其中的一部分。因此采集的数据要想形成分析用数据库,还需要按照一定的规则建立,以确保不同数据类型的准确性、一致性、可信度和及时性。

分析用数据库的格式一般是除第一行属于观察指标外,其余每一行代表一个观察对象的所有观察指标值(即数据),每一列代表某项观察指标所有观察对象的观察值。为了数据库的通用性,文字数据要标化并转换为数字格式,可在适当位置添加批注。

获取的数据通常是零散的、杂乱的,需要进行合适的数据清理、汇总,清理数据也是研究思路整理的过程,清晰地分析用数据库有助于使观察对象的研究指标一目了然。建立好分析用数据库是正确统计分析的前提和基础,甚至决定了研究结果的成败。

另外,数据的四个维度,即种类(variety)、速度(velocity)、数量(volume)和准确性(veracity)将决定数据库规模,其中种类取决于数据的多样性,数据是结构化、半结构化还是非结构化数据;速度取决于数据生成和累积的速度;数量取决于数据的规模;准确性取决于数据的不确定性水平。因此建立的分析用数据库应该从这几个方面进行评估。

(三) 了解数据的性质和类型

由临床检验数据构建分析用数据库,首先,应当明确区分数据库中变量的性质,即因变量与自变量,以及变量的类型,包括数值变量、多项有序分类变量、多项无序分类变量和二项分类变量。同时,还需深入理解数据之间的转化关系等基本信息。这样的做法有助于我们准确选择统计分析方法,从而有效避免混淆变量性质或选择错误的统计分析方法。

在数据库中,变量依据研究目的及其与其他变量的相互作用关系,可被划分为影响变量与结果变量两类。影响变量,又称自变量(X),指的是那些能够自主变动并对结果变量产生影响的因素,而结果变量,又称因变量(Y),则是那些随着影响变量的变动而相应发生变化的量,被视为影响变量作用的结果体现。通常情况下,那些

相对固定、不易改变的特性或是易于人为操控的处理因素，常被视作影响变量或影响因素。相对的，那些易于波动、较难精确界定的观察效应或最终状态，则被视为结果变量，作为最终观测结果的体现。值得注意的是，影响变量与结果变量的界定并非绝对，其划分需依据具体的研究目的及实际情境而定。在某些情况下，两者之间的界限甚至可能变得模糊，甚至无须进行明确的区分。

从数据库和数据分析的专业视角出发，变量作为能够体现数据库中数据内在数量关系的要素，在统计计算中扮演着重要角色。不同的研究目的会引导我们构建不同的数据库，进而决定了数据库中所包含的各类变量。在应用统计学中，根据统计分析方法的选择，临床检验数据可被精准地划分为数值变量、多项有序分类变量、多项无序分类变量以及二项分类变量，见表 3-1-1。

表 3-1-1 临床检验数据的不同类型与举例

数据类别	数据类型	等级次序	举例
计数资料	二项分类数据	无或有	隐血（-、+），性别（男、女），革兰染色（-、+）
	多项无序分类数据	无	血型（A、B、AB、O）
等级资料	多项有序分类数据	有	尿蛋白（-、+、++、+++）
计量资料	数值变量	-	各类检测的定量结果

不同变量可遵循以下顺序转化：从数值变量，到多项有序分类变量，到多项无序分类变量，再到二项分类变量，称为降级转化，但这种转化过程会不断丧失蕴藏的数据信息。

在数据分析过程中，根据研究需要在有关专业理论指导下，各类数据资料间可以相互转化，以满足不同统计分析方法的要求。计量资料在专业理论的支持下，不难转化为计数资料或等级资料。因此，在获取临床检验数据时，应尽可能地获取定量数据，而对于那些原本为计数或等级的资料，在分析过程中，为满足某些统计分析方法的要求，在有关理论和实践的指导下，可以转化为计量资料（指标数据化）。

二、软件的准备

“工欲善其事，必先利其器”，数据分析与统计的软件也是一种工具，称心适手的工具可以带来更高的分析效率。在学习和工作中，大家接触和听说过很多数据分析与统计软件，这些软件有商业软件、共享软件，也有免费软件。不同的软件各有特点，可以选择一种或几种软件进行数据分析与统计，或在数据分析的不同阶段采用不同的软件。但是因为学习不同软件需要花费大量的时间，因此如何选择适合自己喜好和需求的软件，是一种谨慎的策略。

（一）Excel

Excel 是美国微软公司开发的办公软件 Microsoft office system 中的电子表格系统，是目前应用最为广泛的办公室表格处理软件之一。Excel 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，是建立数据库并进行常用统计分析的好工具。尤其是在数据收集、初步清理、趋势观察等方面非常有效，是与其他专业统计软件配合使用的好帮手。

（二）社会科学统计软件包（SPSS）

社会科学统计软件包（statistical package for the social science, SPSS）是 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 总部推出的专业统计软件。SPSS 软件最大的特点是采用菜单操作，方法齐全，图形与表格绘制较为方便，结果输出直观易懂，但是其在统计分析功能方面略显逊色，特别是难以同时分析处理多个数据文件，不如编程语言式统计软件灵活自由，对于大数据分析能力有限。

（三）统计分析系统（SAS）

统计分析系统（statistics analysis system, SAS）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 针对专业用户进行设计，采用编程语言，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。

（四）Stata

Stata 统计软件由美国计算机资源中心 1985 年研发，是一款集数据分析、数据管理以及绘制专业图表等多种功能于一体的统计分析软件，它提供了数据科学和推理，数据处理、探索、可视化、统计、报告和可重复性所需的一切，适用于各种领域的研究人员，包括社会科学、生物统计学、健康科学、流行病学、心理学和经济学。

（五）R 语言

“R”语言（<http://www.r-project.org/>）由新西兰奥克兰大学的 Ross Ihaka 和 Robert Gentleman（R&R）在 1993 年开发的一种编程语言，是一种集数据、统计、图形分析为一体的工具。R 是现今最受欢迎的数据分析和可视化平台之一。R 是一个开源项目，在很多操作系统上都可以免费使用。基本的安装就提供了数以百计的数据管

理、统计和图形函数。但是，R 的学习曲线较为陡峭，因为它的功能非常丰富，相应的帮助文件也很多，很多功能都是由独立贡献者编写的可选模块提供，要掌握 R 的所有功能，可以说是具有挑战性的。因此，可以根据需要学习相应的功能。

（六）Python

Python 是由荷兰数学和计算机科学研究学会的 Guido van Rossum 于 20 世纪 90 年代初设计，提供了高效的高级数据结构，还能简单有效地面向对象编程。Python 语法和动态类型，以及解释型语言的本质，使它成为多数平台上写脚本和快速开发应用的编程语言，随着版本的不断更新和语言新功能的添加，逐渐被用于独立的、大型项目的开发。众多开源的科学计算软件包都提供了 Python 的调用接口，例如著名的计算机视觉库 OpenCV、三维可视化库 VTK、医学图像处理库 ITK。而 Python 专用的科学计算扩展库就更多了，它们为 Python 提供了快速数组处理、数值运算以及绘图功能。

（七）Graphpad Prism

Graphpad Prism 是一款集生物统计和科技绘图于一体的数据处理软件。它可以直接输入原始数据，自动进行基本的生物统计，产生高质量的科学图表。虽然在数据统计分析方面的功能不如 SAS 和 SPSS 强大，但是其具有的功能非常实用。具体应用及软件可以在 GraphPad Prism 中国官网（www.graphpad-prism.cn）得到。

（八）Origin

Origin 是由 OriginLab 公司开发的一个科学绘图、数据分析软件，其支持各种 2D/3D 图形。Origin 中的数据分析功能包括统计、信号处理、曲线拟合以及峰值分析。Origin 的绘图功能是基于模板的，其提供了几十种 2D/3D 绘图模板，并且允许用户自己定制模板，可以与各种数据库软件、办公软件、图像处理软件等方便地连接。

（九）MedCalc

MedCalc 软件是一个专门为医学工作者设计的医学计算器，功能齐全，可以帮助医生快速做出普通的医学计算，从而对症下药。提供超过 76 种常用的规则和方法，包括患者数据、单位参数、费用计算等。易于学习，快速且可靠，并且包括 220 多种统计测试、过程和图形。

（李 江）

第二节 临床检验数据常规统计方法

统计方法是指有关收集、整理、分析和解释统计数据，并对其所反映的问题做出一定结论的方法，统计方法可以通过已知的观测数据去分析随机现象的数据规律，可以帮助我们分析数据的不确定性。统计方法包括了两大部分：描述统计（descriptive statistics）与统计推断（statistical inference）（图 3-2-1）。描述统计可以帮助我们读懂数据，提供数据的基本信息。统计推断可以对数据进行推论，并进一步给出推理性结论，其中参数估计可以用“样本数值”估计“总体数值”，假设检验可以验证一个想法或者假设的研究问题。统计方法的选择如图 3-2-2 所示。

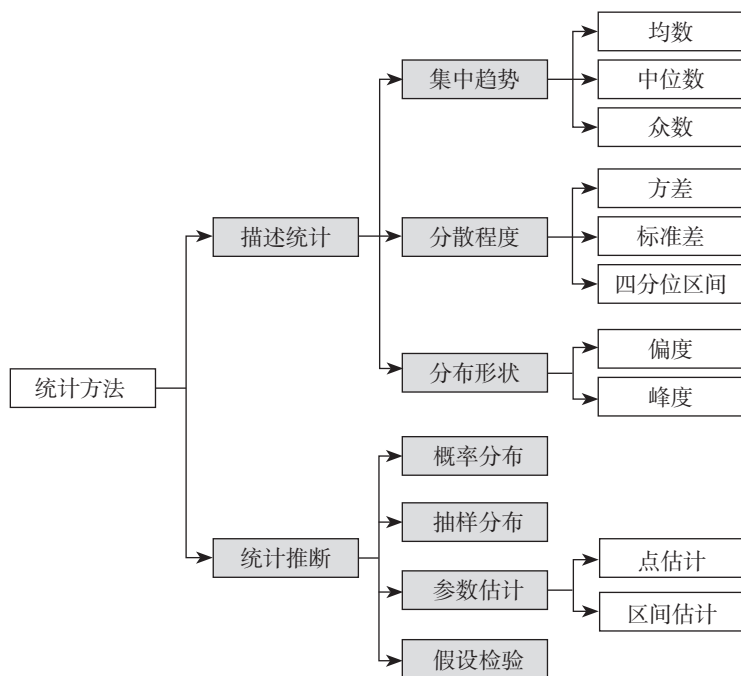


图 3-2-1 临床检验数据常规统计方法的构成

一、描述统计

描述统计是指用统计指标和适当的统计图表来描述数据资料的分布规律及其数量特征，即利用收集的数据，通过一些统计方法来反映数据的特点。

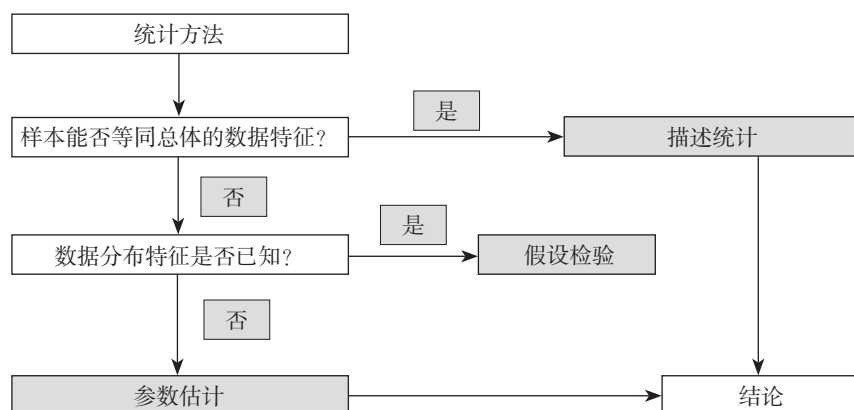


图 3-2-2 临床检验数据常规统计方法的选择

在分类变量的情况下，可以描述离散值的简单计数（如一组数据中男性和女性的数量和百分比）；在连续变量时，可以使用集中趋势和分散程度等统计量来描述（如不同性别组的年龄均值和标准差）。在临床检验工作中，可以利用临床检验质控数据的集中趋势、分散程度等特征来制定质控规则，并通过图表对收集的数据进行必要的可视化，进一步综合概括和分析得出数据的客观规律，如绘制质控图；利用重复测量的检验数据评估检测系统的精密度和正确度；利用参考人群的数据分布特征建立参考范围等。

（一）数值法

数值法进行统计描述就是用不同的统计量从不同的维度描述数据的特征，如描述集中趋势可以用平均数、中位数、众数；描述离散趋势可以用四分位区间、方差、标准差；还可以进行相关关系的描述，如正相关关系、负相关关系等。

根据统计量的产生途径，可以分为参数法和非参数法。参数法统计量是基于已定义公式的推导和计算，例如算术平均值、几何平均值、标准差等。非参数法统计量是从数据中直接计数得到的，例如第 50 百分位的中位数、第 25 百分位与第 75 百分位的四分位区间等。

根据数据的不同类型和特征，在进行数据分析时，需要采用正确的描述统计方法，对于计数资料可以采用率或构成比描述；对于符合正态分布的计量资料可以采用参数法的均数和标准差表示；对于不符合正态分布的计量资料可以采用非参数法的中位数和四分位数间距表示。

（二）图表法

图表法也称为描述统计的可视化，就是利用不同的图形和表格从不同的维度描述

数据特征,进行可视化显示。例如描述数据的构成,可以用饼图、漏斗图、金字塔图;描述数据的联系,可以用散点图、气泡图、雷达图;描述数据的趋势,可以用折线图;描述数据的分布,可以用柱状图、箱式图、小提琴图等。

数据的可视化,重要的是理解图形含义,正确表达数据特征。图形不是越炫酷越好、越复杂越好,而是适当地结合不同的图形表达数据的特征。

二、统计推断

统计推断是从总体中随机抽取一个或几个样本,通过样本信息了解总体特征,是在对样本数据进行统计描述的基础上,对统计总体的未知数据特征做出以概率形式表述的推断。统计推断的结果通常是为了得到下一步的分析策略。

参数估计(parameter estimation)和假设检验(hypothesis testing)是统计推断的两个组成部分,也是统计学的核心内容。它们都是利用样本对总体进行某种推断,但推断的角度不同,使得它们在观察性研究、实验性研究以及预测评价方面有着截然不同的应用。

在临床检验工作中,经常会遇到类似的问题,例如,一组研究对象的某个标志物水平是否与另一组研究对象的标志物水平不同?某个新的检测方法和传统检测方法是否一致?诊断效能是否优越?某种标志物与疾病状态是否相关?是否可以预测或诊断疾病的发生、发展及预后?我们该如何判别和证明这些信息和想法的真伪呢?

这一类以“判断”为输出结果的问题,可以统一归纳为假设检验问题,按照假设检验的流程进行处理,可以从数据的差异与一致性、相关与回归、时间因素、重复测量等对数据的影响等方面进行统计分析,实现不同的研究目的。

探索数据之间的关系,是一切研究的基础,抛开研究目的分析数据是毫无意义的,但不关注数据的特征,设计研究目的也是盲目的。如何理解数据,找出数据之间的关联,根据数据特征选择合适的统计方法,是分析数据间关系的关键。不同类型数据间关系的分析应该选择适合的统计方法。

在临床检验工作中,参数估计和假设检验也有许多不同的应用场景,例如检验结果报告对于疾病的诊断就是利用了参数估计中点估计的原理;质量管理工作中利用区间估计建立检验项目的参考值区间、室内质控规则以及评估检验结果的不确定度;实验室认可工作中利用假设检验进行不同检测平台和检测方法的一致性评价;临床诊疗工作中利用假设检验构建诊断和预测模型等。

（一）参数估计

参数估计指用样本指标值（统计量）推断总体指标值（参数）。参数估计包括点估计（point estimation）和区间估计（interval estimation）两种情况。

点估计是用样本统计量的某个取值直接作为总体参数的估计值。区间估计是在点估计的基础上，按预先给定的概率（ $1-\alpha$ ）所确定的包含未知总体参数的一个范围，通常是由样本统计量加减估计误差得到，该范围称为参数的可信区间或置信区间（confidence interval, CI）。

（二）假设检验

假设检验是一种基本的统计推断形式，用来判断样本与样本，样本与总体的差异是由抽样误差引起还是本质差别造成的。

可以根据研究问题的需要对所研究的总体作某种假设，记作原假设（ H_0 ）；选取合适的统计量，这个统计量的选取要使得在假设 H_0 成立时，其分布为已知；由实测的样本，计算出统计量的值，并根据预先给定的显著性水平进行检验，作出拒绝或接受假设 H_0 的判断。

假设检验可以简单地看作是一门关于变量与变量之间关系分析的科学。常用的假设检验方法非常丰富，根据变量类型的不同，变量特征的差异，需要选择不同的统计方法。如果首选统计方法的条件不合适，一般可以通过降级转化选择“低”一级的统计方法或其他统计方法。

变量与变量之间的关系可以从不同的方面去阐述，如数据间的比较、数据间的趋势以及数据间的关系。例如数据之间的一致性和相关性。一致性和相关性是临床实践中广泛应用的两个概念，用以评估变量之间关系的存在和强度。数据之间的相关性与一致性有密切的关系，但统计学本质上是不同的概念，需要采用不同的统计量进行描述（图 3-2-3）。

1. 比较分析

例如前面我们提到统计方法经常提出的一个假设问题是：“一组研究对象在某些特征上是否与另一组研究对象不同？”这个问题可以归结为两个群体的数据集中趋势的比较，以及每个群体在中心值上的分散程度分析。

当数据满足假设条件时，如足够的数据量、相同的数据分布（方差齐）、独立的数据构成，两组计量资料最常用的组间比较方法是学生 t 检验；但是当数据的假设无效时，则可以使用非参数检验来进行组间比较，例如 Wilcoxon 秩和检验；当数据为离散数据，不能用分布来表示，可以采用四格表描述，使用卡方检验、Fisher 确切检

验等方法进行组间比较。

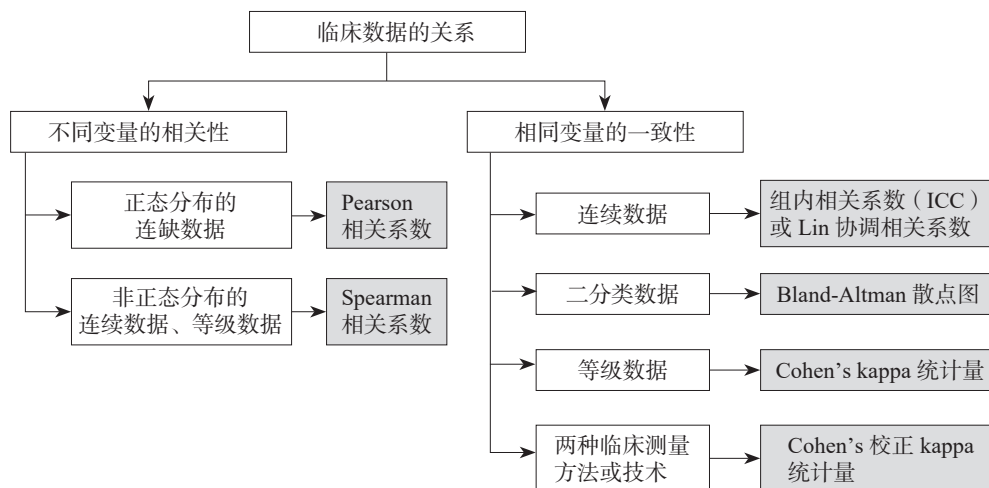


图 3-2-3 临床数据关系假设检验方法的选择

当要比较两个以上不同组的数据时，可以采用方差分析（analysis of variance, ANOVA），这种情况下的假设 H_0 是所有组的平均值相同，假设 H_1 是并非所有的平均值都是相等的（有些可能相同，但有些可能不同）。这时数据也应该满足随机抽样，相互独立，正态分布，方差齐等条件，当不满足这些理想条件时，可以使用其他方法进行比较。例如曼-惠特尼检验（the Mann-Whitney test）、Wilcoxon 秩和检验和 Kruskal-Wallis 秩和检验等非参数检验方法。

2. 趋势分析

在临床检验工作中，经常会遇到判别两组数据之间是否存在某种关联，并估计这种关系强度的问题，例如生化检验项目制订标准曲线时所需的吸光度与浓度关系，不同生物标志物之间的关系，以及生物标志物与临床症状和结局的关系等。

最广义的相关性就是衡量数据之间的关联。在相关性数据分析中，一个数据变量幅度的变化与另一个数据变量幅度的变化相关联，要么是在相同方向（正相关），要么是相反方向（负相关）。

在进行相关性分析前，首先需要根据数据类型和分布特征，选择合适的方法（表 3-2-1）。例如，Pearson 相关性是两个正态分布随机变量之间线性关联的度量；Spearman 秩相关则描述了两个变量之间的单调关系，适用于非正态分布的连续数据或有序数据，对异常值相对稳健。

进行相关性分析时，应注意以下几点：应避免利用相关性推断因果关系；相关性分析不适合进行一致性评价；应辨别相关与回归的区别和联系，相关用于说明两变量之间的关系方向和密切程度，没有主次之分，回归更进一步用于定量刻画两变量在数