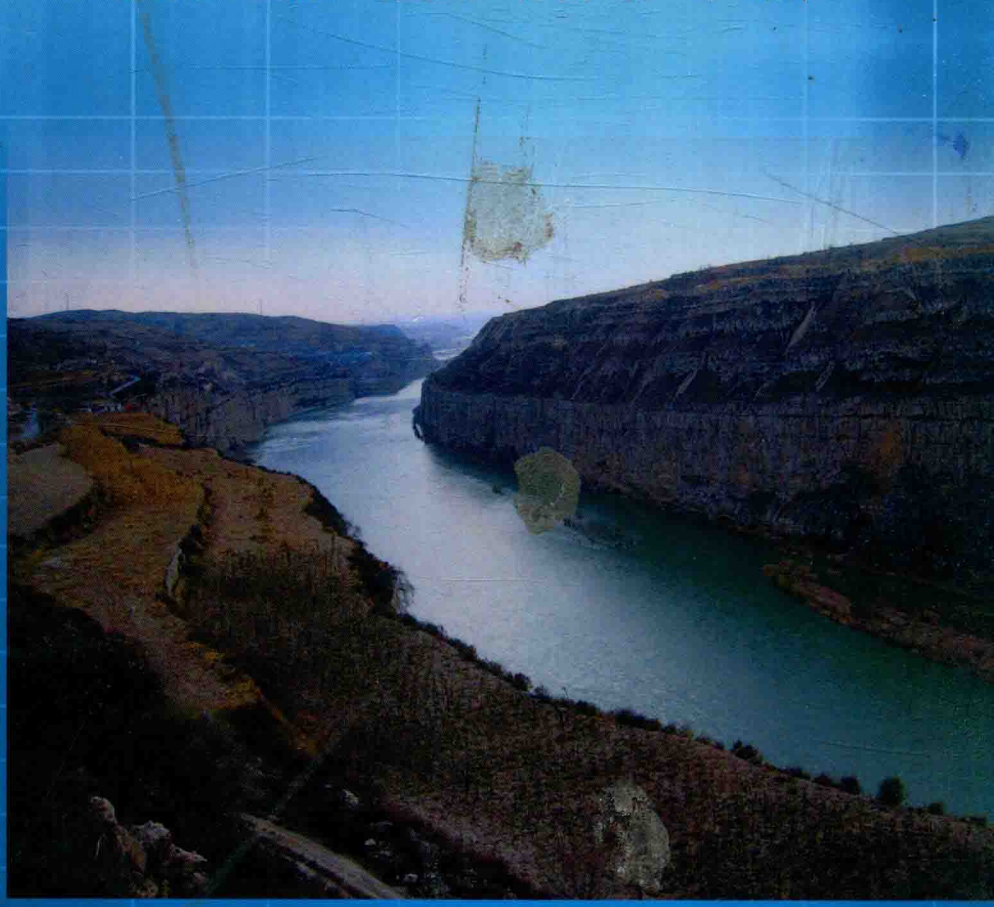
HUANJING SHUJU ZHILIANG PINGJIA TONGJI FANGFA

环境数据质量评价统计方法

孙照东　王任翔　孙晓懿　韦　昊　孙一公　编译



黄河水利出版社

环境数据质量评价统计方法

孙照东　王任翔　孙晓懿
韦　昊　孙一公　编译

黄河水利出版社
·郑州·

内 容 提 要

数据生命周期分为规划(数据质量目标)、实施(数据采集)和评价(数据质量评价)三个阶段。本书描述了数据评价阶段。数据质量评价分为五个步骤,开始于规划文档的评审,结束于对规划阶段提出问题的回答。这些步骤大致与环境统计学家分析数据集合时的行为相应。本书分为五章,详细描述了这五个步骤:评审数据质量目标及抽样设计、进行初步的数据评审、选择统计方法、检验统计方法假设和由数据得出结论。

本书可供资源环境类行政管理、规划、监测、评价专业技术人员和相关专业大中专院校学生、研究生学习参考。数据使用人员,如负责决策和环境评估的项目经理、风险评估师等,会发现本书对于理解和指导其他技术人员生产和分析数据非常有用。数据分析人员会发现本书是对基本评价工具的概要描述。数据生产人员,如分析化验人员和野外样品采集人员等,会发现本书有利于理解他们的工作成果被如何使用并为数据生产效率和有效性的提高奠定基础。

图书在版编目(CIP)数据

环境数据质量评价统计方法/孙照东等编译. —郑州:黄河水利出版社,2014.5

ISBN 978-7-5509-0708-9

Ⅰ.①环… Ⅱ.①孙… Ⅲ.①环境影响-评价-统计方法 Ⅳ.①X820.3

中国版本图书馆 CIP 数据核字(2014)第 001785 号

出 版 社:黄河水利出版社

地址:河南省郑州市顺河路黄委会综合楼 14 层 邮政编码:450003

发行单位:黄河水利出版社

发行部电话:0371-66026940、66020550、66028024、66022620(传真)

E-mail:hhslcbs@126.com

承印单位:黄河水利委员会印刷厂

开本:787 mm×1 092 mm 1/16

印张:10.75

字数:248 千字 印数:1—2 000

版次:2014 年 5 月第 1 版 印次:2014 年 5 月第 1 次印刷

定价:30.00 元

编译者序

编译者在开展“黄河下游水量稀释调度方案初步研究”过程中，发现特殊时段水利系统氨氮常规监测数据与环境保护系统的自动监测数据差别很大；在编制“入河排污口设置论证”报告过程中，发现我国水利系统和环境保护系统的部分水质参数监测数据使用困难，并存在现行监测方法与 GB 3838—2002《地表水环境质量标准》不匹配的情况。怎样找到监测数据的规律性、如何计算可能最大值、如何看待出自不同部门的监测数据、如何处理离群值、如何处理未检出是实际工作中需要解决的重要问题，这些问题无不与环境监测数据的质量密切相关。数据的质量，直接影响调度方案的编制质量及实施效果，直接影响到入河污染物影响范围的确定，直接影响到污染物限排方案的提出，并会影响到主管部门的决策质量。经过检索 Internet 网络资源，在美国环境保护局网页上找到了有关数据质量评估的系列重要文献。

本书以其中的一本为基础，其英文书名为“EPA QA/G - 9S Data Quality Assessment: Statistical Methods for Practitioners”。该系列中另外三本重要文献为《数据质量目标程序指南》(EPA QA/G - 4 Guidance on Systematic Planning Using the Data Quality Objectives Process)、《环境数据采集抽样设计选择指南》(EPA QA/G - 5S Guidance on Choosing a Sampling Design for Environmental Data Collection)、《数据质量评价：评审人员指南》(EPA QA/G - 9R Data Quality Assessment: A Reviewer's Guide)。它们可从美国环境保护局网页 http://www.epa.gov/quality/qa_docs.html 下载。

EPA QA/G - 9S Data Quality Assessment: Statistical Methods for Practitioners，文字通俗易懂，是美国环境保护局不同项目办公室在环境数据设计统计分析方面的经验总结，融入了以往的指导性文件、统计学知识和科学规划，在数据质量评价标准和设计任务说明书方面提供一般性指南。该书编辑风格独具特色，与一般教科书和其他指南性文件的不同之处在于：未采用线性的和连续性的阅读方式，而是设计成数据质量评价方面有用技术的“工具框”，其总体结构能够使分析人员使用系统的方法论调查许多问题，每一种统计技术均采用系列步骤的形式进行演示，并用实例按照描述的步骤进行说明。该书是一本环境监测数据质量评价方面难得的工具书，建议作为环境、资源类技术及管理人员的案头书使用。英文原版书已于 2012 年由美国 Bibliogov 出版社正式出版，可以从 Amazon 等网站购买。

在本书编译过程中，编译人员力争反映原书的语言风格，力争准确表达原意，专业术语依据 GB/T 3358.1—2009《统计学词汇及符号 第 1 部分：一般统计术语与用于概率的术语》和 GB/T 3358.2—2009《统计学词汇及符号 第 2 部分：应用统计》确定。编译人员在编译过程中，发现了一些小的印刷错误，本书定稿时进行了修正。

本书由黄河水资源保护科学研究院孙照东教授级高级工程师组织编译、审稿和定稿。参与本书编译的人员有黄河水资源保护科学研究院孙照东(副译审、教授级高级工程

师)、王任翔(高级工程师)、孙晓懿(博士、工程师)、韦昊(助理工程师)和河海大学孙一公。

在本书编译过程中,编译者得到了黄河水资源保护科学研究院彭勃院长的大力支持,在此表示感谢!

因编译者水平有限,本书编译文本肯定存在许多不足之处,希望读者将斧正意见发往sunzd2003@ vip. sina. com,以便不断改进。

孙照东

2014 年 3 月于郑州

前　言

数据质量评价(DQA)是对环境数据的科学及统计评估,以确定它们是否符合项目规划目标,确定环境数据是否以正确的类型、质量和数量支持其预期用途。数据质量评价建立的一个基本前提:仅当与数据的预期用途相联系时数据的质量才有意义。本书旨在描述环境数据集评估中数据质量评价的技术要点。对于数据质量评价过程概念性的介绍见《数据质量评价:评审人员指南》(EPA QA/G-9R)(U.S. EPA,2004)。

通过使用数据质量评价,评审人员可以回答下述四个重要问题:

(1)给定的数据质量能否按照期望的确定性水平进行决策(或估计)?

(2)抽样设计进行得有多好?

(3)如果相同的取样设计策略再次应用于类似研究,能否期望这些数据以所需要的确定性水平支持相同的预定应用?

(4)如果影响确实存在的话,能否采集到足够的样本使得评审人员能看到这一影响?

第一个问题处理评审人员的直接需求。例如,如果这些数据正在被用于决策,并提供了有力证据支持作出正确选择,那么决策者可以继续进行决策,直至决策得到正确的数据支持。但是,如果此数据没有显示出足够有力的证据倾向于一项选择,那么数据分析结果将提醒决策者注意这种不确定性。现在,决策者就处于如何继续决策,作出明智选择的境地(譬如,在决策前收集更多的各种数据,或者继续进行决策,而不在乎得出错误结论的可能性相对较高)。

第二个问题处理在稍有改变的条件下该抽样设计的鲁棒性如何。如果此设计对于潜在的干扰影响非常敏感,那么对于结果的解释可能很困难。通过回答第二个问题,评审人员提防由独特环境下产生的虚假结果的可能性。

第三个问题处理是否将此作为独特情形,即此数据质量评价结果仅仅应用于该情形而不能外推到其他情形。同时,也处理将该数据收集方案再用于未来项目的适用性问题。例如,如果想将某抽样设计用于不同地点,而在该地点又是第一次采用该抽样设计,就应该基于此次采样活动的结果和环境条件与原活动不同之假定来决定能够期望该方案表现得多好。由于环境条件会从一地一时到另外的一地一时变化,抽样设计的充分性应当在可能的结果和条件的更广泛的范围内进行评估。

最后一个问题处理研究中是否使用了足够的资源。譬如,在流行病学调查中,是否能够基于实际获得的有限的样本数可靠地观察到所感兴趣的影响。

数据生命周期包括三个步骤:规划、实施和评价。在规划阶段,数据质量目标(DQO)过程(或任何其他系统规划程序)用来定义标准以确定样本采集的数量、位置和时间安排等,以便得到具有期望的确定性水平的结果。这与采样方法、分析程序,以及相应的质量保证(QA)与质量控制(QC)程序等一起编入质量保证项目计划中。实施阶段是根据质量保证项目计划规定采集数据。在评价阶段开始时,对数据进行核实验证,以确保遵守质量

保证项目计划中规定的采样与分析协议，测量体系执行质量保证项目计划规定的准则。然后，数据质量评价，通过确定数据质量目标的规划目标规定的性能和验收标准是否得到实现，再由数据得出结论，结束数据的生命周期。

数据质量评价涉及五个步骤，开始于规划文档的评审，结束于对规划阶段提出问题的回答。这些步骤与环境统计员分析数据集的活动相似。这五个步骤将在本书下述的章节中详细描述：

1. 评审数据质量目标及抽样设计；
2. 进行初步的数据评审；
3. 选择统计方法；
4. 检验统计方法假设；
5. 由数据得出结论。

上述五个步骤以线性序列呈现，但数据质量评价实际上是一个迭代过程。例如，如果数据初步评审显示数据集的模式或异常现象与该项目的目标不一致，那么该研究规划中的某些方面需要在步骤1中重新审议。同样，如果数据不支持统计方法的基本假设，那么可能必须重新审视数据质量评价前面的步骤。

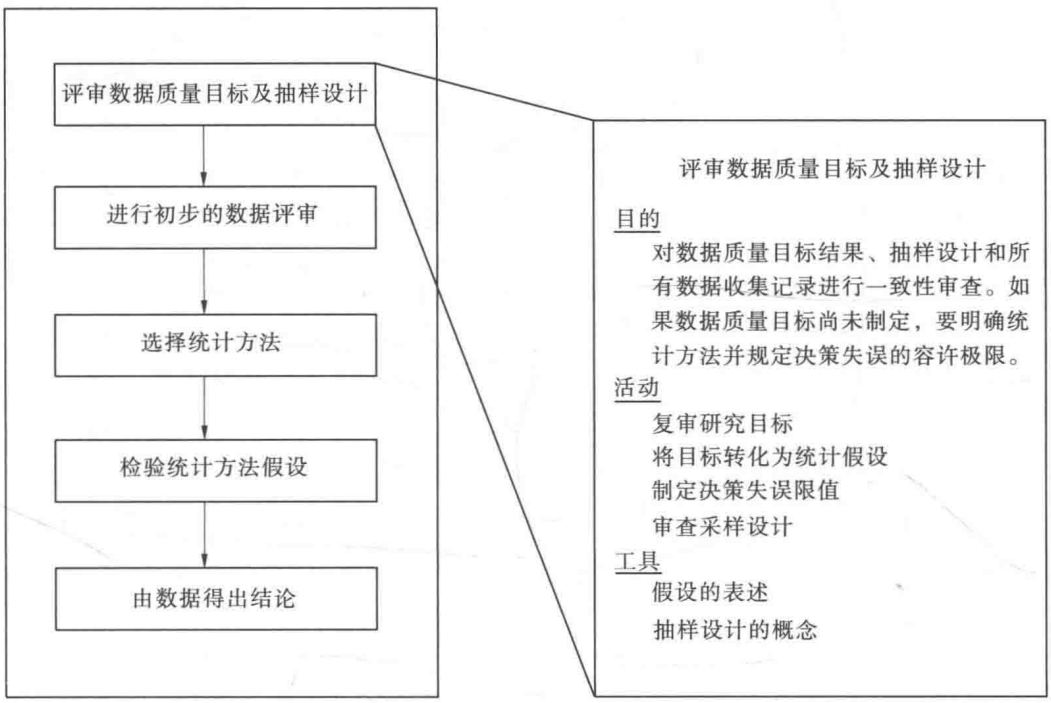
本书是为潜在的数据用户、数据分析人员及数据生产者等广大读者群而编写的。数据用户（如负责决策或对环境特性做出估计的项目经理或风险评估人员）应该会发现本书有助于理解数据生产者和分析人员的技术工作。数据分析人员会发现本书是对基本评价工具的概要描述。数据生产者（如负责环境样本采集和分析并报告结果数据的分析化学家或野外现场采样专业人员）将会发现本书对他们理解自己的工作成果被如何应用大有裨益，从而为提高数据的生产效率和有效性奠定了基础。

本书介绍了用于执行数据质量评价的背景信息和统计工具；非统计学方面的数据质量评价问题讨论可参见配套文档《数据质量评价：评审人员指南》（EPA QA/G－9R）（U.S. EPA，2004）。每章对应于数据质量评价中的一个步骤，从完成该步骤要进行的各项活动的概述开始。第1章～第4章，概述之后，描述了专门的图形或统计工具，运用实例提供了具体步骤。第5章就统计检验的诠释给出了一些建议。附录里包含了有关统计表，参考文献中列出了对深入的统计分析有用的出版物。

目 录

第 1 章　步骤 1:评审数据质量目标及抽样设计

数据质量评价过程



步骤 1:评审数据质量目标及抽样设计

- 评审研究目标
  - 如数据质量目标已制定,则评审数据质量目标程序的结果。
  - 如数据质量目标尚未制定,则弄清楚这些目标应是什么。
- 将数据用户目标转化为对主要统计假设的叙述
  - 如数据质量目标已制定,则将其转化为对初步统计假设的陈述。
  - 如数据质量目标仍未制定,则弄清楚作了什么假设或估计。
- 将用户数据目标转换为Ⅰ型或Ⅱ型决策误差极限值
  - 如数据质量目标仍未制定,记录数据用户对决策误差、灰色区域的宽度和估计初值的可能容许极限。
  - 如数据质量目标已制定,则确认此决策误差极限值。
- 评审抽样设计并标注任何特性和潜在问题
  - 就任何潜在的严重偏差,评审抽样设计。

数据质量评价，开始于对数据生命周期中规划阶段的关键性成果，如对数据质量目标、质量评价项目计划和所有关联文档等进行评审。研究目标，对理解数据采集工作目的提供相关的背景情况，为评价拟采用数据的质量设立定性的及定量的依据。抽样设计（在质量评价项目计划中形成的）提供关于如何解释这些数据的重要信息。通过研究抽样设计，分析人员可以了解抽样设计制定时的假设，了解这些假设与目标的关系。

对于在研究规划阶段还没有制定项目目标的情况，有必要在实施数据质量评价之前形成数据用户目标陈述。在数据分析前陈述用户目标的主要目的是为评价与用户预期用途有关的数据的质量制定适当的评价标准。不熟悉数据质量目标程序的分析人员可参考《数据质量目标程序指南》（EPA QA/G－4）（U. S. EPA，2000）、有关统计规划与分析的书籍，或咨询统计学专家。数据质量目标程序的七个步骤如图 1-1 所示。

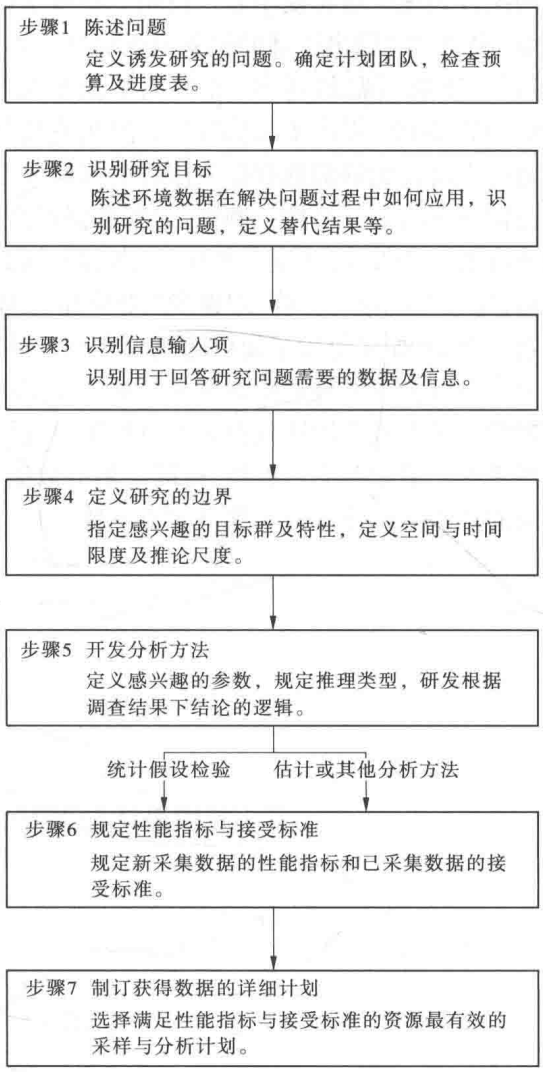


图 1-1　数据质量目标程序

如果项目已经设计为一个假设检验，其不确定度范围可以表达为用户对犯误拒绝（Ⅰ型或假阳性）或者误接受（Ⅱ型或假阴性）决策错误的容忍程度。当原假设被拒绝而事实上为真时就发生了误拒绝错误。当原假设未被拒绝而事实上为假时就发生了误接受错误。其他相关术语包括“显著水平”和“功效”，“显著水平”等于Ⅰ型错误的概率，“功效”等于 1 减去Ⅱ型错误的概率。统计功效实际上是一个函数，是描述Ⅱ型错误范围内的“功效曲线”。功效曲线的特性对选择适当的统计检验非常重要。关于如何确定误拒绝和误接受决策的错误率的详细信息见《数据质量目标程序指南》（EPA QA/G－4）（U. S. EPA，2000）。

如果该项目已经依据置信区间设计，其不确定性可以表达为两个相互关联术语的组合：置信区间的宽度（区间越小对应的不确定度越小）或者相关参数真值处于该区间之内的置信水平（较高的置信水平代表较小的不确定度）。

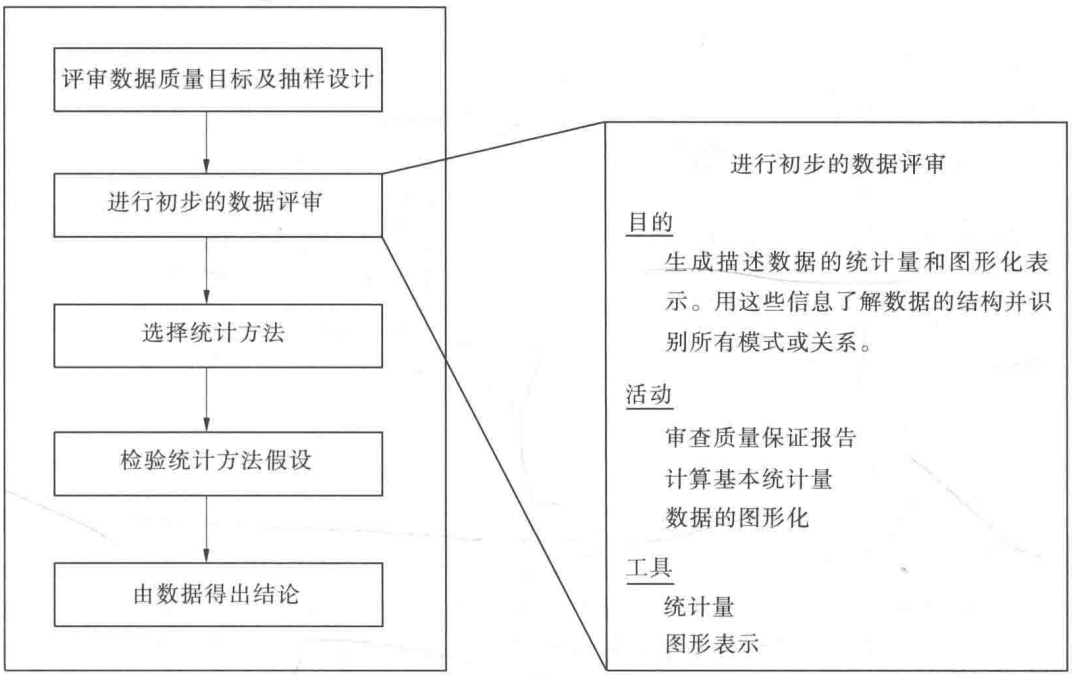
就抽样设计的评审而言，记得抽样设计中的关键差异在于判断性或权威性抽样（其中，采样数量及地点依据专家意见而定）与概率抽样（其中，样本数和地点随机选择，并且

目标群的每一元素包含在样本中的概率是已知的)之间的差异。判断性采样应仅在调查目标不具有统计特性的情况下予以考虑(例如,当研究目的是甄别污染物泄漏的具体位置或研究只关注采样地点本身时)。通常,权威性抽样样本得到的结论仅适用于个体样本,因缺少代表性,聚集作用可能引起严重偏差,从而导致错误的结论。判断性抽样样本仅用于原有目的,不能用于其他目的。如果使用判断性抽样样本数据,在解释所有与结论相关的统计表述时应特别小心。使用一些概率性陈述和判断性抽样样本是不正确的,应该避免,因为它将给出正确的假象但正确性并不存在。判断性抽样距真正的随机样本越远,得到的结论问题就越多。《环境数据采集抽样设计选择指南》(EPA QA/G - 5S)(U. S. EPA,2002)提供了与抽样设计问题及其数据判读的含义有关的广泛信息。

分析人员应能够根据数据使用者的目标评审抽样设计文档。寻找支持这些目标或与这些目标矛盾的设计特性。例如,如果数据用户所关注的是基于污水中污染物随时间的平均浓度进行决策,那么混合抽样是适宜的。如果数据用户正在寻找危险废物场地的污染热点,应当慎用混合抽样,以避免"平均地远离"热点。并且查找抽样设计实施中的潜在问题。例如,如果使用了简单随机采样,对于简单随机抽样设计,验证空间(或时间)上每一个点具有相同的被选择的概率。抽样方案的小偏差可能对由数据集得出的结论产生极小影响。重大或实质性偏差应进行标注,其潜在影响在整个数据质量评价过程中都应当仔细考虑。最重要的一点是,验证采集的数据与质量评价项目计划、抽样与分析计划,或研究的总体目标是如何与规定的相一致。

第 2 章　步骤 2:进行初步的数据评审

数据质量评价过程



步骤 2:进行初步的数据评审

- 审查质量保证报告
  - 查找执行样本采集和分析方法过程中出现的问题或异常现象。
  - 仔细检查质量控制数据,取得用于检验隐含数据质量目标、采样与分析计划、质量评价项目计划的假设的信息。
- 计算统计量
  - 考虑计算适当的百分位数(见 2.2.1)。
  - 选择集中趋势(见 2.2.2)和离散的度量(见 2.2.3)。
  - 如果数据包含两个变量,计算其相关系数(见 2.2.4)。
- 采用图示法显示数据
  - 选择解释数据集结构并突出隐含数据质量目标、采样与分析计划及 QA 项目计划的假设的图示法(见 2.3)。
  - 采用多种图示法检查数据集的不同特性。

2.1 概述及活动

在数据质量评价的这一步骤中，分析人员实施对数组的初步评价，计算一些基本的统计量值，用图示法检查数据。实施初步数据评审的第一项活动是评审所有有关的质量保证报告，对能够用于检查数据质量目标过程中确定的假设的信息应予以特别注意。记录数据中明显的离群值、缺失值、与标准操作方法的偏差以及非标准数据采集方法的使用等极为重要。

图示法可以用于识别模式和趋势，快速确认或驳斥假定，发现新的现象，识别潜在问题，提出补救措施。由于单个图示不能反映数据集的全貌，分析人员应选择不同的图示方法来说明该组数据集各种不同的特性。2.3 节对于常用图示法进行了描述并给出了应用实例。

关于本步骤概况和活动更广泛的讨论见《数据质量评价：评审人员指南》(EPA QA/G-9R)(U.S. EPA,2004)。

2.2 统计量

2.2.1 相对位置度量

有时候，分析人员感兴趣于了解一项或几项观察值相对于所有观察结果的相对位置。百分位数或四分位数是相对位置的度量，对概括数据有用。百分位数是大于或等于给定百分数的数据值中的那个数据值。以数学术语表述，p^{th}百分位数是大于或等于 $p\%$ 的数据值的数据值以及小于或等于$(100-p)\%$的数据值的数据值。因此，如果"x"是p^{th}百分位数，那么数据集中 $p\%$ 的数据值要小于或等于 x，并且有$(100-p)\%$的数据值大于 x。一个样本百分位数可能处于一对观察值之间。例如，一个有 10 个观察值的数据集的75^{th}百分位数不是唯一确定的。因此，有多种方法用于计算样本百分位数，其中最常用的方法如方框 2-1 所描述。

方框 2-1 计算百分位数的说明与实例

设 $X_1, X_2, \cdots, X_n$ 代表 n 个数据点。为了计算 p^{th} 百分位数 $y(p)$，首先从小到大对这些数据点排序并将这些点标记为 $X_{(1)}, X_{(2)}, \cdots, X_{(n)}$。则 p^{th} 百分位数为：

$$y(p) = (1-f) \cdot X_{(i)} + f \cdot X_{(i+1)}$$

其中，$r = (n-1)p + 1$，$i = \text{floor}(r)$，$f = r - i$ 且 $X_{(n+1)} = X_{(n)}$。注意：floor(r)意思是计算 r，然后舍弃所有小数。

实例：计算下述 10 个数据点(从小到大排序)的90^{th}和95^{th}百分位数：4,4,4,5,5,6,7,7,8 和 10 ppb。

对于95^{th}百分位数，$r = (10-1) \times 0.95 + 1 = 9.55$，$i = \text{floor}(9.55) = 9$，$f = 9.55 - 9 = 0.55$。因此，$95^{th}$百分位数是 $y(0.95) = 0.45 \times 8 + 0.55 \times 10 = 9.1$。

对于90th百分位数，$r=(10-1)\times0.9+1=9.1$，$i=\text{floor}(9.1)=9$，$f=9.1-9=0.1$。因此，90th百分位数是$y(0.9)=0.9\times8+0.1\times10=8.2$。

通常评审的重要的百分位数是数据的四分位数，即25th、50th、75th百分位数。50th百分位数也被称作样本的中位数（见2.2.2），而25th和75th百分位数用于估计数据组的离散情况（见2.2.3）。环境数据中90th、95th和99th百分位数也非常重要，决策者希望90%、95%和99%的污染水平肯定低于某一设定的风险水平。

2.2.2 集中趋势度量

集中趋势度量描述数据集的中心特性。最常用的三个估计值是平均值、中位数及众数。计算这些统计量的说明见方框2-2，实例见方框2-3。

方框2-2 计算集中趋势度量的说明

设$X_1,X_2,\cdots,X_n$代表n个数据点。

样本均值：样本均值$\overline{X}$，是数据的和除以样本量n：

$$\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$$

样本中位数：样本中位数$\widetilde{X}$，是有序数据集的中心数。为了计算样本中位数，将数据由小到大排序，并将数据点标记为$X_{(1)},X_{(2)},\cdots,X_{(n)}$。则

$$\widetilde{X}=\begin{cases}\dfrac{1}{2}[X_{(n/2)}+X_{(n/2+1)}], & n\text{为偶数}\\ X_{[(n+1)/2]}, & n\text{为奇数}\end{cases}$$

样本众数：样本众数是样本中出现频率最大的那个数。样本众数可能不存在或不唯一。对每一个值出现的次数计数，样本众数就是出现最频繁的那个数。

方框2-3 集中趋势度量的计算实例

以下是对10个数据点：4,4,7,7,4,10,4,3,7和8计算样本均值、中位数、众数的例子。

样本均值：

$$\overline{X}=\frac{4+4+7+7+4+10+4+3+7+8}{10}=\frac{58}{10}=5.8$$

样本中位数：排序后的数据为3,4,4,4,4,7,7,7,8和10。因$n=10$为偶数，所以中位数为：

$$\frac{1}{2}[X_{(10/2)}+X_{(10/2+1)}]=\frac{1}{2}[X_{(5)}+X_{(6)}]=\frac{1}{2}\times(4+7)=5.5$$

样本众数：计算每一值出现的次数，得到：

3出现1次，4出现4次，7出现3次，8出现1次，10出现1次。

由于4最常出现，它就是该数据集的样本众数。

最常使用的数据集的中心度量是样本平均值，以$\overline{X}$表示。样本平均值可认为是数据集的“重心”。样本平均值是简单抽样设计的算术平均值；不过，对于复杂抽样设计，如分层法，样本平均值则是加权算术平均值。样本平均值受极值（极大或极小）和未检出处理的影响（详见4.7）。

样本中位数是较常用的数据集的中心度量。该值直接落在排序数组的中间。这意味着：有1/2的数据小于样本中位数，1/2的数据大于样本中位数。样本中位数是50^{th}百分位数的另外一个名称（见2.2.1）。样本中位数不受极值影响，且可以在未检出存在的情况下较容易地使用。

另一个度量数据中心的方法是样本众数。样本众数是以最大频数出现的值。由于样本众数可能不存在或不唯一，因此它是最不常用的数据集的中心度量。不过，众数对于定性数据很有用。

2.2.3 离散度量

如有关于中心的数值分散度量相伴，集中趋势度量更有意义。数据集的离散度量包括极差、方差、样本标准差、变异系数和四分位数间距。计算这些度量的说明见方框2-4，实例见方框2-5。

方框2-4 计算离散度量的说明

设$X_1, X_2, \cdots, X_n$代表n个数据点。

样本极差：样本极差w，是指数据集中最大值与最小值的差值，即$w = \max(X_i) - \min(X_i)$。

样本方差：样本方差s^2按下式计算：

$$s^2 = \frac{\sum_{i=1}^{n} X_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} X_i\right)^2}{n-1}$$

样本标准差：样本标准差s，是样本方差的平方根，即$s = \sqrt{s^2}$。

变异系数：变异系数（C_v）是样本标准差除以样本均值（见2.2.2），即$C_v = s/\overline{X}$。

四分位数间距：四分位数间距（IQR）为75^{th}和25^{th}百分位数的差，即$IQR = y(75) - y(25)$。

就计算而言，离散最简单的度量是样本极差。对于小样本来说，极差易于解释并可能较充分地表示该组数据的离散。而对于大样本，极差的信息量不足，因为仅考虑了极值，所以它会深受离群值的影响。

一般而言，样本方差度量了数据点与样本均值之间的均方差。样本方差较大意味着数据未向均值附近集聚。样本方差较小（相对于均值来说），则意味着绝大多数数据都在均值附近。样本方差受极值和大量的未检出的影响。样本标准差是样本方差的平方根，与数据度量单位一致。