

大数据时代背景下统计数据质量的评估方法探讨

杨 瑛

深圳市盐田区妇幼保健院, 中国·广东 深圳 518081

【摘要】现如今, 统计工作成为政府掌握国民经济运行信息的重要途径, 在统计领域中应用大数据技术是社会发展中必不可少的环节。本文主要对统计数据质量及其影响要素进行分析, 并分析大数据时代背景下统计数据质量提升的有效策略, 以供参考。

【关键词】大数据时代; 统计数据质量; 评估方法

大数据在现今社会中是重要资源, 其数据量巨大、信息价值高、更新速度快, 使统计学研究、决策获得诸多有价值资料, 并为统计学研究和决策者提供良好发展机遇, 但也存在诸多挑战。单一维度的样本抽查研究方法已经无法保证统计数据准确性, 因此要扩大分析研究的样本数量, 借助多元化评估方法, 保证统计数据客观、真实, 并推动统计事业与经济社会实现整体发展。

1 大数据时代背景下统计数据质量概述

在社会稳步推进背景下, “数据”被广泛认可, 其作为新兴生产要素, 在经济社会发展中起到巨大推动作用。数据质量是指在特定商业环境中, 获得的信息能否满足客户具体需要以及期望^[1]。也就是说, 数据质量是指数据的可靠、可信、可操作、可维护以及可拓展性。随着社会不断进步, 数据质量备受关注。统计学作为复杂学科是通过多样化方法展现事物本质。统计数据是否准确将对其质量产生重要影响, 因此必须要达到一定标准, 如快速、有效、安全、保密等。在大数据时代背景下, 统计数据质量会对政府权威以及公信力产生影响, 统计数据质量不仅有传统数据质量属性, 同时更加系统全面, 使应用需求、客户需要得到满足。当前, 统计数据质量的管理评估已经成为重要时代课题。

2 大数据时代背景下影响统计数据质量的因素

2.1 数据采集

在大数据时代背景下, 统计数据采集形式更为丰富、多样, 导入人工统计报表、传感器采集、系统数据交换等都可以获取数据。但相比较而言, 传统人工统计报表导入形式下, 数据会更加真实准确, 工作效率也更高, 但由于表格规范性、导入系统接口不兼容, 使得数据会出现导入失败的问题。随着科学技术快速发展, 传感器采集在诸多领域有所应用, 如统计数据质量检验、国民经济核算经济流量、动态交通流统计等^[2], 并且获得了显著成就。但传感器是否准确、数据是否安全会受到极大影响。随着社会技

术水平不断提高, 数据采集质量、时效性受到诸多因素影响, 大数据统计效率也面临巨大挑战。此外, 数据来源多样、复杂, 数据间的差异、矛盾也会影响统计结果。

2.2 数据预处理

大数据时代下, 数据来源渠道多, 但有不同, 不论是从何种渠道获取的数据都不能直接进行大数据统计分析, 需做好预处理工作, 保证数据质量与可用性。做好数据清洗工作, 结合特定标准将采集数据中重复、缺失、异常的信息去除。对于数据格式混乱等问题可通过数据转换方式提高数据质量。

2.3 数据储存

在大数据处理中, 分布式储存是主要方式, 可依据储存介质类型、数据分布情况以及储存复杂性对储存技术进行调整^[3], 从而满足大数据处理需要, 让储存更加精准、安全, 并保证其实时性。

2.4 数据处理

大数据技术能够实现数据流转换, 利用数据结构, 时效性、容错性更强, 可将大规模数据转换为超低量秒级数据集^[4], 自动化收集、批量计算, 对大规模数据科学处理。但在对超大规模数据处理时, 其效率会受到影响。此外, 虽大数据技术可以提供更强系统性能, 但对系统软硬件也提出较高要求, 在实际应用中需要更多资源、技术支持, 以便高效、分布式处理数据。当前背景下, 大数据处理计算框架模型广泛应用, 可使大数据分析质量、效率得到提升。在此过程中, 应用聚类、关联分析等技术会让大数据更加精准、可靠, 能够将其价值充分发挥出来。

3 大数据时代背景下统计数据质量提高策略

3.1 提高数据采集质量

在确定数据来源时, 要选择权威、准确的数据来源, 如国家统计局、政府公开数据等。也可以通过对数据提供者的信誉、数据更新频率、样本代表性进行评估, 保证数据来源是真实可靠的。数据采集时, 方法是否科学也十分

重要。应基于统计数据特点等确定合适采集方法,如问卷调查、数据挖掘、网络爬虫等,保证数据完整、准确。此外,随着大数据技术发展,数据采集中出现很多新技术,如通过人工智能技术自动识别问卷调查中的异常值;通过区块链技术保证数据采集透明、可追溯;使用物联网技术实时采集数据等,这些技术手段都可以保证数据采集准确、可信。

3.2 做好数据预处理工作

数据预处理也就是在正式开始分析数据前开展的处理工作,涉及数据清洗、转换以及规范化等。首先,对于数据清洗,这是数据预处理中十分重要的环节。由于数据来源比较复杂,数据量大,因此不乏各种错误和异常值,如重复值、错误值等,通过数据清洗能够有效识别和处理这些问题,保证数据真实准确。其次,数据转换。其中主要涉及数据类型转换、格式转换以及单位转换^[5]。在数据转换下,保证数据满足分析工具需要,能够使数据分析更加准确、高效。最后,数据规范化,这是将数据转换为统一单位、格式,为比较分析提供便利。此外还涉及数据标准化、归一化处理,避免尺度、量纲影响分析结果。

3.3 加强数据储存

为避免采集到的数据出现丢失、损坏,影响统计结果质量,应做好数据储存工作,完善数据储存系统,做好数据备份,同时使用科学可靠的数据储存技术,如云储存、分布式储存等,维护数据安全性。同时,对数据统一整体归类,完善数据目录以及索引系统,保证数据是可用的,易访问的,为数据分析利用提供便利。此外,还应优化数据维护管理制度,定期对数据进行检查、更新,规范化监督数据利用,保证数据储存、利用过程中是一致、准确的。

3.4 采用科学数据质量评估方法

大数据时代下,统计数据质量的评估方法越来越受到重视,统计数据质量评价方法有很多种,如逻辑规则检验法、核算数据重估、计量模型分析、统计分布验证、调查偏差评估、多维评估延伸以及基于机器学习和人工智能的数据质量评估方法等方法,下面就其中几个方法进行简要分析。

首先,计量模型。其是统计数据检验的重要方法,这一方法对于总量数据进行评估的方法,也就是在无法确定统计数据具体逻辑关系时,形成模型评估,查看是否有差异。在这种方法中,建立模型是关键,找到异常点,操作时需先依据统计理论与目标建立计量模型,然后对样本数据估算检查,反复调整,多次代入评估,最终获得结论。计量模型主要包括线性回归模型、时间序列模型、面板数据模型、Logit模型、Probit模型等^[6]。这些模型可以检验数据质量各方面,如数据的集中趋势、离散程度、分布形态、精确性、相关性、影响程度、稳定性、周期性、趋势性、不确定性以及模糊性等。在实际应用中,要根据具体

问题和数据特点选择合适的计量模型来评估数据质量。

其次,多维评估延伸。大数据时代背景下,统计数据内容多样且多维化,评估方法也需实现多维延伸。多维评估法的评估路径主要有四种,即多级量度、指标检测、用户需求、统筹综合。虽然长时间以来对数据统计分析的侧重使得多维评估还处于发展阶段,并不成熟,但随着统计数据非结构化特点日渐明显,人们对统计数据质量要求不断提高,多维化评估必将受到更多关注和重视,其应用也会日渐完善。

最后,基于机器学习和人工智能的数据质量评估方法。这主要分为基于监督学习、无监督学习以及深度学习的评估方法。基于监督和无监督学习评估方法适用于已知和未知数据质量问题类型的情况。而基于深度学习的方法在大数据时代下有巨大发展潜力,其中深度神经网络可以自动学习数据中的复杂特征,让数据质量评估更加准确。循环神经网络和长短时记忆网络等序列模型可以评估时序数据质量。

结束语:

总而言之,如今信息技术快速发展,社会进入到数字化、大数据时代,在此背景下,统计工作也面临着诸多挑战,但其依旧通过对数据收集、整理、计算、分析来展现事物特点规律,对社会发展以及经济进步有着重要影响。同时要认识到,在大数据时代背景下,统计数据质量保证是极具系统性、挑战性的任务,统计数据质量依旧受到诸多因素影响,数据评估方法也有很大不足,会对统计数据发展、经济社会进步产生影响。为此,政府部门、统计机构需提高重视程度,加强对统计数据质量评估方法研究分析,顺应大数据时代发展需要,保证统计数据结果真实准确,推进统计事业发展,让统计工作朝着更加现代化、数字化的方向发展,助力统计研究新发展、新变革。

参考文献:

- [1]李瑞青.大数据背景下统计数据质量控制方法研究[J].中文科技期刊数据库(全文版)经济管理,2023(3):0004-0007.
- [2]王洪峰.大数据背景下乡镇基层统计数据质量优化分析[J].中文科技期刊数据库(全文版)经济管理,2023(3):0009-0012.
- [3]高华东.统计数据质量的影响因素分析及改进方法思考[J].中文科技期刊数据库(全文版)经济管理,2023(7):0005-0008.
- [4]曹琦.基层统计数据质量管理的优化方式探究[J].时代人物,2023(3):0138-0140.
- [5]王伟.基于SAM的宏观统计数据质量评估方法分析[J].大众商务,2022(1):0320-0321.
- [6]张海亮.大数据背景下统计数据质量控制方法研究[J].生产力研究,2022(5):156-160.