大数据时代下的统计学

姓名：史君宝 学号：2152118 班级：理科实验班 同和六班

【摘要】：信息技术快速发展，人类进入大数据时代，当前数据的90%是近两年产生的。采用正确的数据分析方法，才能让多元数据真正发挥自身的价值，而这就需要数理科学，尤其是统计学的发展。在大数据时代，统计学在面临巨大的挑战的同时，也有着广阔的应用前景，如何正确平衡两者关系，有助于我们充分利用大数据的优势。

【关键词】：大数据时代；应用统计学；方法创新；噪声与信息；挑战和机遇

【正文】

1. 大数据与统计学

1.1大数据时代的来临

1998年，大数据首次在《大数据的应用程序》一书中被提到，在2003年，大数据仅仅被当作数据参与研究【1】。信息技术快速发展，数据储存和传输的成本降低，通过光纤连接的广大用户之间产生了海量的数据，逐渐成为当代重要的数据发源地，也为大数据时代的来临奠定了基础。

由于互联网收集数据迅速，用户数据实时更新，数据类型多样化、内容多元化，利用这些海量数据，并将结果应用于多领域中，能够产生巨大经济效益。例如在2013年的美国总统选战中，奥巴马利用大数据分析成功连任，2008年，马云就认识到了大数据的重要性，并将其作为阿里巴巴重要的发展战略。进入二十一世纪，信息数据爆炸性增长，近两年数据占比近九成，人类进入大数据时代。

1.2统计学及其发展

中国百科全书对于统计学的定义较为准确：一门研究怎样有效地搜集、整理和分析带有随机性的数据，以对所考察的问题做出推断或预测的学科【2】。从中我们看到，统计学的对象是具有随机性或者不确定性的数据，通过应用科学的统计方法来建立对应模型并进行预测，其中包含着收集数据、分析数据、模型预测等多个过程。

统计学有三百多年的学科历史，统计学发展可分为：古典统计学、近代统计学和现代统计学【3】。回顾历史我们发现统计学的发展都是以社会发展的实际需求为主要的驱动力，人们为了解决科学问题，就创立新的数据收集和处理的方法，统计学也因此建立和发展下来。例如大数理论、误差测定、最小二乘法、正态分布曲线等方法的相继提出，在金融、社会、人口法律的研究中做出了突出贡献。

1. 大数据时代下的统计学面临的挑战

2.1样本和总体概念的改变

传统的数据收集模式是先明确自己的研究目标及其对应的受众情况，并根据已经确定的总体范围，采用抽样调查的方法获得具有代表性的数据，并通过分析样本数据来获得总体情况的信息。而大数据则与之相反，是先有数据，后有总体【4】。具体来说，就是大数据时代的数据不像传统的统计学所面临的问题一样，需要通过确定研究目标和总体来设计问卷调查，将隐藏在人们心中的数据收集起来。大数据时代，数据可以直接方便地获取和储存，网络平台上的海量数据本身就是所有用户的数据集合，我们已经获得了总体数据，所需要做的便是根据研究目标有效挖掘所需数据

这就意味着传统统计学有关数据收集方面的原有理论需要革新来适应大数据时代。比如在大数据时代，抽样调查的功能需要改变，传统抽样是为了抽取有代表性的数据来代替总体进行分析，新型的抽样调查应该转变为在繁杂的总体数据中挖掘出真正需要的有效数据，来服务于之后的工作。

2.2数据来源的局限性

在大数据时代，数据来源具有两个方面的局限性，一是数据来源复杂，数据的承担者的不确定；二是大数据时代实际对于数据来源者加了一个限定，即使用智能手机或电脑的用户，从某种意义上来说对数据来源进行了一个窄化。

数据的来源复杂，数据承担者的不确定使我们难以确定数据的真正意义上的总体。我根据此举个具体的例子来说明，一个无人的百货商店，只通过一台可以识别商品并完成交易行为的机器就可以运行起来，但如果采用一种可以分析顾客所属群体的购物需求并实现智能推送的机器，这家百货商店将具有很大的竞争力。同理在大数据时代，如果我们可以获取某一群体的数据，并可以为其定制服务。但由于很多用户的个人信息不实，我们很难确定信息真正的承担者是谁，就很难确定数据真正意义上的总体。如果解决不了这个实质性的问题，大数据的预测行为难以进行下去，并可能为决策带来巨大的障碍。

大数据时代下的数据来源用户具有一个隐藏的限定，他们都是经过筛选过后的人群。一个具体的例子是波士顿的Street Bump应用程序，为了监测公路的质量情况，他试图从驾驶员的智能手机上取得数据,如果驾车经过路面的坑洼处，智能手机就可以灵敏的反馈所受到的震动【5】。但其实我们从统计学的知识可知，这其实是已经过筛选的样本，因为一部分年老或者贫穷的驾驶者并不会使用智能手机，这种带有隐藏限定的数据反而可能会误导统计学的工作者。

2.3噪声与信号

除了对于数据的来源进行确定外，这种从虚拟世界所获得的数据质量也值得我们去探讨。例如有媒体报道说，许多的推特账号实际上是机器人自动程序或“半机器人”系统（即得到机器人程序辅助的人工控制账号），以及部分的虚假账号【6】。我们对于收集的数据要抱有客观态度，收集的数据不可能都是正确的，真正的数据其实是非常嘈杂的。数据就在那里，信息与数据同时空存在，统计学家纳特·西尔弗说：“只要能将信号和噪声区分开来，我们就能获得所需要的任何信息。”这位统计学家所说的确实是一个真理，信号与噪声难以区分，二者随使用者的变化而变化【7】。从我们的实际经验来看，大部分的数据对于使用者来说都是噪声，有用的数据反而会被噪声所淹没，将它们从噪声中搜集出来十分困难。当我们利用网络数据进行社会的分析或者决策时，这些由后台的自动化算法产生的数据是最让人头痛的，这种数据本身就存在很大的错误，同时相对于人工数据产生的更加快捷，如果不能将它们剔除，这种错误数据的循环利用，将会将原来存在的噪声或者误差进一步放大，导致相关行业承受巨大的经济损失。

2.4数据结构类型多样

随着大数据的进一步发展，数据不再像传统统计学中可以计算，可以用二维表格进行表示的结构化数据了。视频、音频、图片、图像、文档等不同数据表现形式产生，涉及到了非结构化数据、半结构化数据、异构数据。这些数据形式的变化给统计学工作者带来了巨大的挑战，作为数据科学的统计学急需解决这一问题。近年来一些统计学家也提出了一些解决这些数据的方法，但是值得敲响警钟的是，在大数据时代与统计学相辅相成的计算机科学也在这方面发起了挑战。比如在超高维数据的处理上，SIS方法可以实现数据的降维，然而随着计算机技术的发展，这种方法的价值反而在减小【8】。基于分布式的大规模数据软件平台，通过并行处理可以成倍地提高计算机性能，经过并行计算的优化，理论上在GPU上的运算速度会比CPU快50-100倍，不借助SIS方法就可以直接处理超高维数据【9】。我们知道统计学发源于对数据处理的实际需要，如果未来对复杂数据的处理，我们依赖于计算机技术的提升而非统计学理论的创新，那么这种实际需求将不再能推动统计学的发展，统计学也将开始萎缩。

1. 大数据时代下统计学的机遇

3.1统计成本降低，时效性更强

在当今大数据时代，数据收集过程在过去需要人工参与，例如设计问卷、走访调查，查询文献。大数据时代的人工成本下降甚至消失，可以轻松快捷地获得大量的数据。同时，传统统计学中，数据收集中如果出现了数据失效的情况，需要重新进行调查，而在大数据时代就可以轻松解决这个问题，大大降低了统计的成本

同时数据的时效性也能得到保证，比如，我们利用元素周期表的原子的相关信息，我们可以充分相信前辈的测量数据，但是对于中国社会的经济发展状况我们就必须采用实时数据。以居民消费价格指数CPI为例，该指数广泛应用于实时反映居民的生活水平，它作为国民性的指标，数据收集计算十分麻烦，同时受到市场价格的波动，该数据也是在实时变化的，对于它的测算，利用大数据的优势将会大大简化，时效性增强。

3.2统计学体系大大延伸

在大数据时代，应该用动态、辩证的眼光去看待统计学的发展，构建新的学科体系，将大数据时代的新的思想、方法、知识迁移到学科建设中。我们知道统计学是以需求出发，在大数据时代，数据类型多样化，数据处理过程也浮现了很多的问题，如果我们积极把握这个巨大变革的时期，将会大大丰富统计学的内容。

3.3数据挖掘更加深入

在大数据时代，由于计算机技术的发展，在保证数据质量的基础上，对于数据内容的挖掘将会更加深入。具体来说，在传统的统计学框架下，我们寻求两个事物之间的关系时其实是从我们认知的角度出发，之后我们会对其展开调查分析，我们更加专注于通过因果关系去分析问题。而在大数据时代，我们将会更加注重事物的相关关系。大数据时代下，需要从泥沙俱下的大数据中提炼出有价值的知识和信息，这些有价值的知识和信息显然是非预期的【10】。这意味着我们会在数据分析的过程中，挖掘出两个看似不相关事物之间的相关关系，收获预期之外的经济效益。比如通过谷歌趋势服务，预测股市的涨跌，发现Twitter用户的情绪有助于预测股市【11】。如果将类似的相关分析应用到多领域的研究中，将实现数据向价值的直接转化，带来巨大的社会经济效益。

【总结】：

大数据时代，赛博世界为我们提供了大量的数据，我们需要与其相对应的数理科学来解决。统计学正是一门根据具体需求发展而来的数据科学，在面对多样化的数据类型，颠覆的统计模式，更高的统计需求时，统计学现有理论远远不够，需要不断革新自身理论，来适应大数据时代。将统计学与计算机科学相结合，创造新型的统计模式，深入挖掘数据资源，并把实践成果应用于其他领域中，必将带给多学科明媚的春天。

【参考文献】

【1】《应用统计学在大数据背景下的应用与创新》 高文滨 杨映瑶

《江苏科技信息》 2019年7月第19期

【2】《大数据时代对传统统计学变革的思考》 朱建平 张悦涵

《统计研究》 2016年2月 第33卷第2期

【3】【8】【9】【11】《大数据时代统计学发展的若干问题》

“大数据中的统计方法”课题组 《统计研究》 2017年1月 第34卷第1期

【4】《基于大数据思维的统计学若干理论问题》 李金昌

《统计研究》 2016年11月 第33卷第11期

【5】【6】【7】《大数据时代对统计学的挑战》 邱东

《统计研究》 2014年1月 第31卷第1期

【10】《大数据时代统计学的重构与创新》 《统计研究》 2015年2月第2期

中国人民大学“大数据与应用统计”研究组