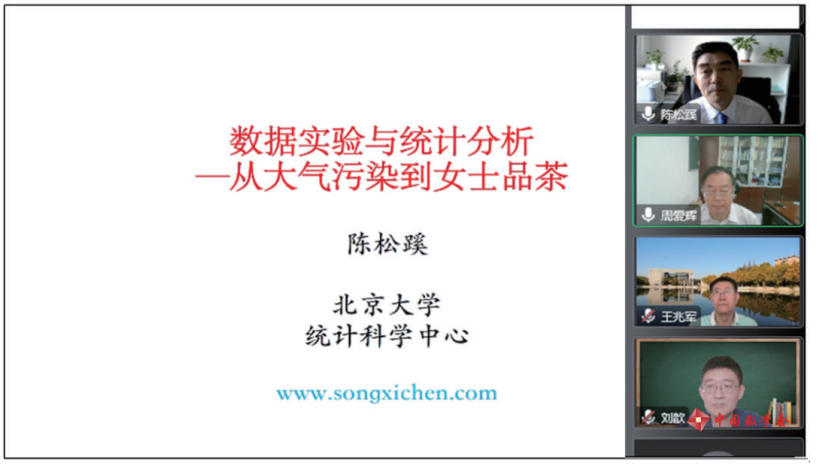
学会科普报告：数据实验与统计分析

当今时代，大数据的获得越来越容易，但其也附带许多混杂因素。要想从大数据中提炼出科学的结果需要使用统计学技术，因此如何使用统计学技术剔除、调整、建模大数据中的混杂因素是数据实验与统计分析中的重要问题。

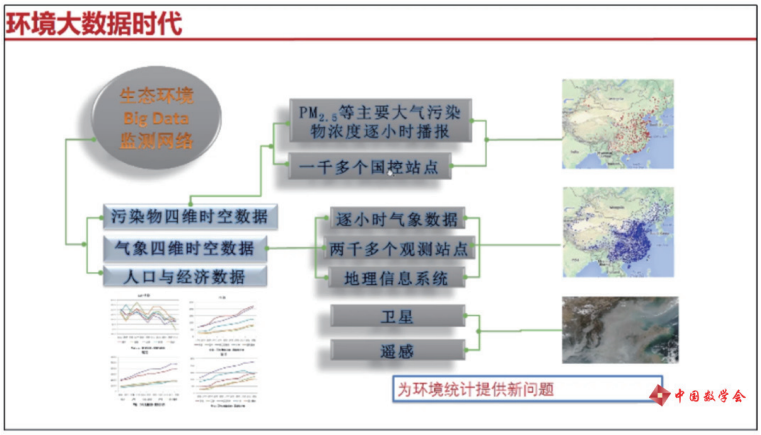
2022 年全国科普日的主题是“喜迎二十大、科普向未来”， 侧重围绕大数据、人工智能等科技发展前沿，让更多公众深刻感知前沿科技魅力。为此，中国数学会联合中国工业与应用数学学会、中国运筹学会和中国现场统计研究会特别邀请北京大学陈松蹊院士，为广大科技工作者和数学爱好者献上了精彩的网络科普报告：“数据实验与统计分析—从大气污染到女士品茶”。中国数学会副理事长周爱辉研究员主持了报告，一起出席的还有中国工业与应用数学学会副理事长王兆军教授、中国运筹学会科普工作委员会主任刘歆研究员。



9 月 18 日上午 9 点，在大家的热切期待中，报告正式开始。陈院士的报告用三个例子说明了为何从大数据中提炼出科学的结果需要使用统计学技术。这三个例子分别是：（1）从大气污染监测网络数据提取污染排放信息，介绍团队八年来分析、追踪北方地区大气污染变化的实证研究，给出大气污染评估的统计学思路和方法；（2）女士品茶及充分随机实验；（3）吸烟对寿命影响的大样本观测研究。

◆ 环境大数据时代 ◆

陈院士从一封与朋友来往的邮件讲起他和大气污染研究的渊源，强调对大气污染的研究关系到人民的生命健康及生活质量。



目前我国已建立包括污染物思维时空数据、气象思维时空数据、人口与经济数据、卫星数据、遥感数据在内的生态环境检测数据，真正进入了环境大数据时代。而如何使用监测大数据度量污染物排放量是大气管理的关键科学问题。陈院士首先对比了用“排放源清单”监测的传统方法和用“环境大数据”监测的新方法，接着详细阐述了用统计学方法剔除气象因素干扰后能更准确地度量污染物排放以及在最理想的情况下用充分随机实验 Treatment Effect 检验方法（t- 检验方法）能很好地解决“如何评判今年的污染低于去年”的问题。最后，陈院士通过对比随机化实验和观测实验，说明充分随机实验的有效性和理想性。