大数据时代的统计教育

孟生旺　袁 卫

内容提要: 2014 年 11 月,美国统计学会适应大数据时代的要求,发布了统计学本科专业指导性教学纲要。而在 2013 年,我国统计类本科专业刚刚进行了一次较大调整,目前的专业课程设置和教学内容改革还处于探索阶段。美国统计学会发布的这份指导性教学纲要对于推进我国统计类本科专业教育改革具有重要借鉴意义。本文首先概括性地介绍了美国统计学会发布的统计学本科专业指导性教学纲要的核心内容,包括统计专业本科生应该掌握的基本技能和应该修读的主要课程,然后分析了我国统计类本科专业教育存在的问题,并提出了在大数据时代改进我国统计类本科专业教育的几点建议。

关键词: 统计教育; 大数据; 课程体系; 本科专业
中图分类号: C829. 29　　**文献标识码:** A　　**文章编号:** 1002 - 4565(2015) 04 - 0003 - 05

Statistical Education in the Age of Big Data

Meng Shengwang & Yuan Wei

Abstract: In November 2014, American Statistical Association (ASA) published guidelines for undergraduate programs in statistical science to meet the age of big data. In 2013, a significant change was made for statistical undergraduate programs in China. The current curricula and teaching content of statistical undergraduate programs in China are still in the exploratory stage, so the guidelines published by ASA are valuable for statistical education reform in China. The article summarizes the key points of ASA guidelines, including basic skills and main curricula for statistical majors, and then points out the problems of statistical undergraduate programs in China. At the end of the paper, it proposes some suggestions for improving the statistical undergraduate programs in China.

Key words: Statistical Education; Big Data; Curriculum; Undergraduate Program

一、引言

2013 年以前,我国只有一个统计学专业,既可授予经济学学位,也可授予理学学位。从 2013 年开始,这个专业被拆分为统计学、应用统计学和经济统计学三个本科专业。根据教育部高等学校统计类专业教学指导委员会 2013 年 11 月公布的数据,目前全国有 194 所高校开设了统计学专业,156 所高校开设了应用统计学专业,164 所高校开设了经济统计学专业[1]。在美国,统计专业的本科毕业生从 2003 年的 673 人增长到 2013 年的 1656 人,年均增长 9. 42%[2]。毫无疑问,统计类专业的快速发展与当前的大数据时代背景密切相关。

大数据时代对统计专业的学生提出了更高的要求,他们不仅需要掌握统计理论、统计方法和专业统计软件的应用,还要懂得如何提出问题,如何进行数据操作,如何判断数据质量,如何评价模型和方法的有效性,以及如何准确清晰地呈现分析结论。

2014 年 11 月,美国统计学会发布了统计学本科专业指导性教学纲要,这是在大数据背景下对 2000 年首次发布的指导性教学纲要进行的一次全面修订和更新[2-4],主要强调了四个方面的内容: ①数据科学日益重要,统计专业人才不仅需要扎实的数学和统计基础,还要有强大的统计计算和编程能力,可以熟练使用专业统计软件和数据库; ②真实数据是统计专业教育的重要组成部分; ③更加多样化的统计模型和方法; ④通过语言、图表和动画等用户易于理解的方式表达数据分析结论的能力。

美国统计学会适应大数据时代的要求发布的这份统计学本科专业指导性教学纲要代表着美国统计

?1994-2015 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

专业教育在未来的发展方向，对我国目前正在进行的统计教育改革具有重要的参考价值，值得借鉴。

二、统计专业本科生所需的基本技能

在大数据时代，统计专业的本科毕业生所从事的主要工作是实际数据的统计分析。一个完整的统计分析过程由以下环节构成：①问题的公式化表达，即把研究的实际问题转化成一个统计问题；②评价现有数据是否满足需要；③选择合适的统计分析方法；④用可以重复再现的方法进行统计分析；⑤评价统计方法的合理性；⑥得出分析结论，并通过有效的方式呈现给用户。毫无疑问，统计专业的学生应该在统计分析过程的每个环节都受到良好训练。不能指望统计专业的初学者就能达到很高的水平，但要引导他们循序渐进地掌握这些技能。基于上述考虑，美国统计学会把统计专业的本科生应该掌握的知识和技能归纳为下述五个方面[2]。

1. 统计方法与统计理论。统计专业的本科生应该能够设计研究方案，通过统计图进行探索性数据分析，建立统计模型并对模型的输出结果进行评价，熟悉统计推断，能够从数据分析中得出恰当的结论。为此，要帮助学生逐渐积累应用统计方法的经验，让他们可以基于实际问题对各种统计方法的适用性做出评价，并能有效呈现和表达他们的分析结论。当然，为了提高数据分析结果的可靠性，统计专业的本科生还需具备扎实的统计理论基础。

2. 数据操作和统计计算。统计专业的学生应该能够熟练使用一款专业统计软件进行探索性数据分析，发现和清洗数据中的错误记录，具有编程能力和算法思维，可以进行各种数据操作，譬如把不同来源和不同格式的数据进行合并处理以满足当前研究的需要。统计专业的学生还应该掌握统计计算技术，能够进行模拟研究。通过随机模拟来验证解析方法已经得出的结论，是一种非常有效的学习方法。

3. 数学基础。统计专业的本科生应该掌握微积分、线性代数、概率论和数理统计的基础知识。计划攻读统计学博士学位的学生还需修读更加高级的数学课程，如数学分析或高等微积分，也可以修读随机过程、图论、微分方程、优化方法、组合数学、代数统计学等课程。

4. 实践训练和表达能力。统计专业的学生应该具有良好的表达和交流能力，善于通过图示和动画等听众易于理解的方式展示分析结论，具有团队合作精神和项目领导能力。表达能力与统计技术的培养应协调一致，通过很有说服力的方式把一个错误的分析结论传输给听众会产生更加糟糕的后果。

5. 特定领域的知识。统计学提供了数据分析的基本方法和工具，只有与其他领域的知识结合，才能真正体现其应用价值，因此，统计专业的学生必须掌握特定应用领域的知识，并用统计学特有的思维方法来分析和解决该领域的实际问题，具体表现为：首先把特定领域的实际问题转化为统计问题，然后搜集数据并进行统计分析，最后把分析结论通过听众易于理解的方式表达出来。统计专业的学生如果能在其他领域副修一个学位或选修一系列课程，可以有效提高统计专业人才的培养质量。

三、统计专业的课程设置和教学内容

根据美国统计学会公布的统计学本科专业指导性教学纲要，统计专业的课程设置应该涵盖统计方法与统计理论、数据操作与统计计算、数学基础、实践训练[2]。

(一) 统计方法与统计理论类课程

应用统计方法解决问题的一般思路是，从特定领域的实际问题出发，通过搜集和分析数据，寻找解决问题的答案。统计专业的本科生必须深入理解统计学的基本概念，掌握多种统计方法。统计方法是不断发展更新的，不能指望本科毕业生熟悉所有的统计方法，但至少应该掌握下述内容：

1. 统计理论。包括随机变量及其分布，似然理论，点估计与区间估计，假设检验，决策理论，贝叶斯统计和重抽样(包括自助法和置换检验)。

2. 探索性和可视化数据分析方法。包括高级可视化技术，核密度估计和地图应用等。可视化技术在识别数据中的错漏和异常值方面也有重要价值。

3. 研究方案设计及其相关主题。如随机分配，随机抽样，数据搜集，区组和分层，自适应设计，以及偏差、因果推断和共线性等。

4. 统计模型。如线性回归模型，广义线性回归模型，广义可加模型，时间序列模型，混合模型，生存分析，空间统计，回归树，以及模型的选择、诊断和交叉验证等。此外，还应包括多元统计分析和机器学习的内容。

?1994-2015 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

(二) 数据操作与统计计算类课程

具体教学内容包括:

1. 一种或多种专业统计软件,如 R/Rstudio、SAS 或 Python,并辅以 shell scripts 和 knitr[5,6]。

2. 对各种类型和格式的数据(如 CSV、JSON、XML、数据库和文本数据)进行检索、合并和重组等操作的技术,以及缺失数据的处理方法。

3. 基本的编程概念和技术,如把一个问题分解为若干个模块的方法,基于算法的思维方式,以及结构化编程、程序调试和提高算法效率的方法等。

4. 计算密集型的统计方法,如迭代算法、优化方法、重抽样和蒙特卡洛随机模拟等。这些方法对本科生尤为重要,因为通过随机模拟可以验证解析结果的正确性并形成直观感性的认识。

(三) 数学基础类课程

统计专业的本科生应该深刻理解统计方法为什么和在什么情况下有用,以及如何用数学语言来表达统计思想,并能够解释数学推导与统计应用之间的相互关系。

统计专业的本科生需要学习的数学基础类课程应该包括: 一元和多元微积分; 线性代数,如矩阵运算、线性变换、欧式空间投影、特征值、特征向量和矩阵分解等; 概率论,如一元和多元随机变量的性质、离散和连续型随机变量的分布、以及马尔科夫链等。在讲授这些基础数学课程时,应该强调它们在统计中的应用,如 delta 方法。

(四) 实践训练类课程

统计专业的学生不仅需要掌握扎实的统计专业知识,同时还需要具有良好的交流和表达能力,从而可以有效地将统计分析结论用易于理解的方式呈现给听众。这些技能包括: 口头和书面表达能力、演示和分析结论的可视化技巧、团队合作能力、以及与用户的交流和沟通能力。

统计专业的本科生应该了解统计学家和数据科学家所从事的所有工作,养成重视数据质量、科学选择统计方法、客观报告分析结论的职业习惯。对统计专业本科生的训练还可以包括实习、论文写作、参与课题研究和提供咨询建议等。

四、存在的问题和改革建议

从 2013 年开始,我国的统计类本科专业包括统计学、应用统计学和经济统计学。

以美国统计学会 2014 年 11 月公布的统计学本科专业指导性教学纲要为参考依据,对比我国目前的统计类本科专业教育,无论是在培养目标和课程设置方面,还是在教学内容和教学方法方面,都还存在不小差距。

(一) 学习目标不易量化观测和评价

在人才培养方案中,首先需要明确学习目标。

学习目标的设定应该简洁清晰,能够转化为可以观测和量化的行为。我国统计类专业的培养方案关于学习目标的表述大多比较抽象,不易进行量化观测,很难评价毕业生是否达到了设定的最终学习目标。相应地,我们在人才培养过程中比较注重课程层面上的评价,比较轻视专业层面上的总体性评价,缺乏对毕业生专业综合技能的反馈机制。在专业层面上对学生的总体性评价要远远复杂于在课程层面上对学生个体的评价,每门课程的良好效果并不能保证所有课程组合在一起就必然可以实现预定的学习目标,因为某些不同的课程可能仅仅培养了一种相同的技能。

作为参考范例,下面引用加州大学伯克利分校为统计学本科专业所制定的学习目标[7],包括: 概率论基本原理; 统计思维和统计推断方法; 统计计算方法; 统计建模方法及其局限性; 通过图示等手段对数据进行描述和解释,并进行探索数据分析的方法; 良好的书面和口头表达能力。

我国现在设有三个统计类专业,如何明确设定它们各自的学习目标,既体现统计类专业的共性,又能个性鲜明,避免专业边界模糊不清,是一个亟待研究解决的问题。

(二) 课程设置存在一定随意性

一个合理的课程体系是为了实现培养方案所制定的最终学习目标而设定的一系列课程,这些课程之间应该相互配合,逻辑关系清晰,避免内容上的重复和遗漏,同时还应该足够灵活,为学生的自主发展预留下宽松的空间。

我国许多高校的统计类专业培养方案由于学习目标不够明确,因此在课程设置上缺乏对总体性学习目标的充分考虑,存在一定的随意性。目前流行的课程体系是按照统计学的各个研究主题设置的,强调每门课程在教学内容上的系统性和完整性。这种课程设置方式历史较长,已经积累了丰富的经验,但也难免存在一些缺陷,如不同课程之间的教学内

?1994-2015 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

容容易出现交叉和重复，每门课程的教学内容与培养方案所设定的总体学习目标不易完全对接。

统计类专业的课程设置必须体现其应用型学科的特点，这就要求开设统计类专业的高校能够提供不同应用领域的课程群供学生进行选修，此外，还应该设有一个数学类课程群，主要供那些计划攻读统计学硕士学位或博士学位的学生进行选修。

（三）教学内容有待更新

大数据时代是以数据为中心的时代，统计类专业的教学内容必须适应这个时代的要求，对传统教学内容进行及时调整和更新。

1. 突出数据的重要性。无论是统计学的初级课程还是高级课程，都有必要引入真实数据的分析，让学生接触来自现实问题中的各种原始数据，引导他们把实际问题转化为统计问题并进行数据分析，培养学生用数据思考的能力。这是一项极具挑战性的教学内容改革，但对提高统计人才的培养质量意义重大。我国统计教育的现状是，大多数学生很难接触到来自实际问题的原始数据，通常使用教科书提供的二手数据进行建模分析。二手数据经过了加工整理，数据背后的实际问题已经被淡化，这种培养模式在一定程度上割裂了实际问题与数据分析之间的联系，不易培养学生的数据思维能力和解决复杂问题的能力，当他们遇到大规模的非结构化数据时可能束手无策。

2. 使用专业统计软件。从理论上讲，只要一款统计软件足够灵活和强大，就可以在教学中使用，但从实践的角度看，推荐使用 R 或 SAS。R 是免费开源软件，统计建模、统计计算和可视化功能强大，新用户增长迅速，也是最新统计方法发布的主要平台，非常有利于培养学生的编程能力和知识更新能力。SAS 软件被很多公司用于数据管理和数据分析，在实际应用领域具有长期而深远的影响，是数据分析不可或缺的专业统计软件。SAS9.3 及其以后的版本还可以自由调用 R，实现了 SAS 与 R 的完美结合，这必将进一步拓展 R 和 SAS 的应用领域。当然，教学中也可以尝试使用其他专业统计软件，如 Python 和 Stata 等。但无论如何，基于 EXCEL 讲授统计方法的时代应该尽早结束。

3. 增加统计模型类课程。统计模型是进行数据分析的主要工具。从目前的情况看，大多数高校为本科生开设的统计模型类课程主要是线性回归模型、时间序列分析和多元统计分析等，远远不能满足数据分析的实际需要，因此在教学内容中很有必要增加一些新的统计模型，如广义线性模型、广义可加模型和线性混合模型等。在学时一定的条件下，增加新内容必然涉及传统内容的删减或压缩问题，需要慎重研究。例如，可以尝试把广义线性模型的部分内容纳入线性回归模型课程中进行讲授。事实上，如果简化了模型的数学推导过程，把注意力集中在模型的思想和应用层面，讲授统计模型所需要的课时完全可以压缩。对于大多数需要直接进入就业市场的本科生而言，统计模型的学习应该更加强调应用层面的问题，如模型设定、模型检验和评价、模型输出结果的解释等。

4. 重视统计计算。传统的统计教育比较重视统计思想和统计理论，推崇数学推导过程，而对统计计算的重要性认识不足。在许多情况下，用直观的数值模拟代替复杂的数学推导，不仅不会降低精度，而且结论更加容易理解，这对统计专业的本科生而言无疑是非常有效的学习方法。譬如，贝叶斯统计具有理论完善和结论直观的优点，缺点是计算量大，但在计算技术飞速发展的今天，贝叶斯统计过去曾经面临的计算瓶颈正在逐渐消失，基于 MCMC 的统计方法在数据分析中的强大威力日益显现。因此，在大数据时代，强调统计计算的重要性应是大势所趋[8]。

5. 整合传统教学内容。许多传统的统计方法是基于小样本数据建立的，并不适用于大数据分析的需要。在大数据时代，这些传统教学内容的相对重要性也会发生变化。譬如，在数据量很大的情况下，统计显著性检验的重要性就会明显降低，而如何克服数据中偏差所带来的影响，以及如何发现高维数据中的结构模式和相关关系将越来越重要。对传统教学内容进行整合的必要性毋庸置疑，但具体实施的难度很大。美国统计学会在最新公布的统计学本科专业指导性教学纲要中也没有明确指出哪些统计方法应该是必修内容，哪些是选修内容，哪些可以删减或压缩。这是一个极具争议的话题。

（四）对非技术性素质的培养重视不够

非技术性素质主要指学生对统计分析结果的表达能力，以及统计分析过程应该遵守的职业道德。我国目前的统计类专业培养方案对这两种素质重视不够，对统计职业道德的培养几乎处于空缺状态。

?1994-2015 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

统计分析过程的一个重要环节是把数据分析结论有效地传达给用户，潜在的用户可能是统计专业人士，也可能是非统计专业人士。这就要求毕业生具有良好的口头和书面表达能力，擅长使用图表和动画等工具把数据分析的结论准确直观地呈现给用户。这种能力可以通过课堂讨论和撰写研究报告等方式进行培养。

数据分析包含一定的经验成分在内，难免受分析者个人主观判断的影响，这就必然涉及数据分析的职业道德问题。统计专业的毕业生在实际工作中可能会面临许多职业道德问题，譬如，有意忽略客观环境对采集数据的影响，只报告具有统计显著性的结果，漠视数据分析的假设前提，混淆因果关系与相关关系，隐藏数据分析方法可能存在的缺陷，把统计显著性解释为实际显著性，根据主观需要选择变量和数据等[9] [10]。统计职业道德的培养，一方面可以在专业课程的教学中潜移默化地让学生逐渐养成关心数据质量、科学选择统计方法、客观呈现分析结论的良好习惯，另一方面也可以采用专题讲座的形式，集中讨论数据分析过程应该遵守的职业道德问题。

五、结语

大数据时代是一个以数据为中心的时代，统计专业人才必须学会用数据进行思考。2013 年，我国统计类本科专业刚刚进行了一次较大调整，把原来的一个统计学专业拆分为目前的统计学、应用统计学和经济统计学三个本科专业。在这种情况下，如何适应大数据时代的要求，对我国统计类专业进行改革，是一个亟待认真研究和解决的问题。

统计学是方法论学科，我国统计类本科专业的设置与美国有所不同，但在学习目标、课程体系和教学内容上的差异不应很大，因此，美国统计学会近期公布的统计学本科专业指导性教学纲要对于我国的统计教育改革应该具有重要的参考价值和借鉴意义。

与美国统计学会公布的指导性教学纲要进行对比，我国统计类专业在教学目标、课程体系和教学内容等方面都还存在一定差距。本文对我国统计专业教育现存问题的剖析未必全面和准确，提出的改革建议尚不成熟，但在大数据时代开展统计教育改革的重要性不言而喻，期待能引起统计教育界的足够重视。

参考文献

[1] 教育部高等学校统计类专业教学指导委员会. 统计学专业教学单位. http://statstsc.org/category/信息公开/教学单位，2013 - 11 - 15.

[2] American Statistical Association. 2014 Curriculum Guidelines for Undergraduate Programs in Statistical Science [EB/OL]. http://www.amstat.org/education/curriculumguidelines.cfm，2014 - 11 - 15.

[3] American Statistical Association. 2000 Curriculum Guidelines for Undergraduate Programs in Statistical Science [EB/OL]. http://www.amstat.org/education/2000curriculumguidelines.cfm，2014 - 11 - 15.

[4] G. Rex Bryce，Robert Gould，William I. Notz，Roxy L. Peck. Curriculum guidelines for Bachelor of Science degrees in statistical science [J]. The American Statistician，2001：7 - 13.

[5] R Core Team. R：A Language and Environment for Statistical Computing [CP]. http://www.R-project.org，2014 - 10 - 31.

[6] YihuiXie. Dynamic Documents with R and knitr [M]. New York：CRC Press，2013.

[7] Beth Chance. From Curriculum Guidelines to Learning Objectives：A Survey of Five Statistics Programs. [EB/OL]. http://www.amstat.org/education/curriculumguidelines.cfm，2014 - 11 - 15.

[8] Tim Hesterberg. What Teachers Should Know about the Bootstrap：Resampling in the Undergraduate Statistics [EB/OL]. http://www.amstat.org/education/curriculumguidelines.cfm，2014 - 11 - 19.

[9] Steve Cohen. Ethics for Undergraduates [EB/OL]. http://www.amstat.org/education/curriculumguidelines.cfm，2014 - 11 - 15.

[10] Steve Cohen. Some Thoughts on the Importance of Internships as Part of an Undergraduate Program [EB/OL]. http://www.amstat.org/education/curriculumguidelines.cfm，2014 - 11 - 15.

作者简介

孟生旺，男，甘肃秦安人，1998 年毕业于中国人民大学统计学系，获经济学博士学位，现为中国人民大学统计学院教授，中国人民大学应用统计科学研究中心研究员，博士生导师。研究方向为应用统计、风险管理与精算。

袁卫，男，天津人，1988 年毕业于中国人民大学统计学系，获经济学博士学位，现为中国人民大学统计学院教授，中国人民大学应用统计科学研究中心研究员，中国调查与数据中心主任，博士生导师。研究方向为应用统计、统计教育与统计史。

（责任编辑：方　原）

?1994-2015 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net