数理统计与数据分析

课程实践报告

题　目：博士研究生学术水平及就业影响因素分析

姓　名：杜云滔

联系方式：scottdyt@163.com

完成日期：2019 年 1 月 22 日

数理统计与数据分析实验报告

——博士研究生学术水平及就业影响因素分析

学号：10153903105 姓名：杜云滔

摘要：随着高等教育的普及，我国已经成为高等教育规模最大的国家，同时也是学术性博士学位授予最多的国家。博士生作为高等教育的最高学位，其培养模式与本科生培养有很大不同。本文以全国优秀博士论文信息和博士调查问卷为数据，主要探讨了博士研究水平的影响因素，同时对博士就业满意度进行了深刻的数据挖掘。使用方差分析、Softmax 回归等数理统计方法从博士的录取方式、科研环境、读博意愿、学校差异、导师教学水平等多维度进行分析，并对参数进行了显著性检验和有效性论证，最后对高校和在校学生分别给出了对于博士生培养和选择的建议。

关键词： 学术能力、就业满意度、方差分析、Softmax 回归

1 引言

研究生教育规模的扩大已成为全球范围的普遍趋势。在 21 世纪开始的头十年，我国的博士授予数量年均增长率超过 20%[1]。目前，我国是学术性博士学位授予最多的国家，也是高等教育规模最大的国家。然而，博士培养质量问题一直受到社会各界的广泛关注。同时，由于各高校教育水平和硬件条件存在着明显的差异，不同高校培养的博士研究水平差异有多大？如何调整博士培养模式，使得博士生在社会的就业满意度进一步提高？本研究使用全国数百个优秀博士论文的案例以及博士调查问卷对以上问题进行分析，并尝试给出合理的结论和建议。

2 数据与数据处理

2.1 数据来源

2.1.1 国家优秀博士论文信息

该数据记录了全国 285 个 2006 年-2009 年的优秀博士论文获奖者的博士论文标题、涉及的研究方向、影响因子、博士个人信息等属性。

2.1.2 博士问卷调查结果明细表

该数据记录了全国不同高校近 1600 条博士生的调查问卷信息，内容涉及高校满意度、导师评价、学术交流、论文成果等多方面，绝大多数指标使用李克特量表(Likert scale)采集，能较好的反映博士生的满意程度。

2.1.3 高校基本信息

该数据记录了全国近 1300 个高校的基本信息，包括高校代码、所属区域，并加入全国 985 工程、211 工程的共 112 所高校的属性信息。

2.2 数据预处理

2.2.1 优秀博士论文预处理

对于优秀博士论文，本研究更关心其研究时间、影响因子（三大索引）与学校类别之间的关系。因此，首先将 SCI、EI、ISTP 合并为三大索引，并计算出该博士生研究时间（月），最后导出属性表示意图如下所示：

表 1 优秀博士论文信息

博士所在学校代码	发表论文数	SSCI 数	影响因子	发表专利数	研究时长	三大索引
80168	12	0	6.8	2	36	0

2.2.2 博士问卷调查预处理

由于问卷调查存在着较多缺失值，首先需要将缺失数据删去。经数据清理后，其完整数据占总数据的 58.9%。对每一类的评价指标取平均数，并对属性数据使用 one-hot 编码转为数值型数据，同时，针对博士生的综合能力，提出国际论文数、国内论文数、科研活跃度、国际化程度四个指标，其主要评价标准如下表所示：

表 2 博士综合能力评价指标

综合指标	衡量指标
国际论文数	国际学术会议论文篇数、国际期刊论文篇数
国内论文数	国内期刊论文篇数
科研活跃度	参与课题数、参加学术会议次数、专利成果项数
国际化程度	境外学术会议次数、出国经历

对该问卷经过数据预处理、删除不必要的字段后，整合字段如下表所示：

表 3 博士问卷调查表

属性字段	变量取值
性别	男、女
学校名称	如兰州大学
录取方式	1：普通招考 2：硕转博 3：直博生
是否有副导师或导师指导小组	1：是 0：否
是否还会选择读	1：是

博	0：否
对当前工作满意度	1（非常不满意）- 5（很满意）
工作相关度	1（非常不满意）- 5（很满意）
培养单位评价	1（非常不满意）- 5（很满意）
对导师评价	1（非常不满意）- 5（很满意）
教学评价	1（非常不满意）- 5（很满意）
课题评价	1（非常不满意）- 5（很满意）
学习能力评价	1（非常不满意）- 5（很满意）
读博抉择	1（非常不满意）- 5（很满意）
学校服务评价	1（非常不满意）- 5（很满意）
国际论文数	10
国内论文数	3
科研活跃度	3.5
国际化程度	2

2.2.3 高校基本信息

本课程作业提供了全国 1000 余所高校的基本信息，本研究加入了 985/211 工程的高校信息作为补充，得到高校基本属性如表所示：

表 4 高校基本信息

学校名称	类别	学校代码	所属省	所属区域
上海交通大学	985	10248	上海市	东部

3 基本统计分析

3.1 优秀博士论文分析

3.1.1 论文关键字

全国优秀博士学位论文评选是对博士培养质量进行监督和激励的一项重要举措，对培养和激励在学博士生的创新精神，促进我国博士生培养质量的提高具有积极的作用。那么，这些优秀博士论文主要都集中在什么领域呢？本研究将论文标题作为关键字，使用 jieba 分词对其进行分词处理，并使用词云展示如图：

图 1 论文关键字词云示意图

在研究领域上，很大一部分优秀论文以生物、化学、材料等 21 世纪新学科为主，同时，也有少部分涉及到物理和数学等基础学科。

在研究方法上，大多数论文都是以“研究”和“应用”，只有少部分涉及到“理论”和“方法”的研究，这反映出国内博士生培养主要以应用驱动创新为主。

值得注意的是，不少论文都涉及到“中国”等问题，如《中国草地样带典型生态系统水分利用效率的时空变异特征研究》、《中国人群 SNP 的大规模发掘和慢性乙型肝炎的遗传易感性研究》，这说明当代博士生更加关注身边的实际问题，尝试使用所学领域知识来解决中国实际面临的问题，这是值得提倡的。

3.1.2 博士培养时长

对于高校来说，博士是科研的主力军，那么博士的培养一般需要多少年呢？本研究将入选优秀博士论文的博士培养时间（月）提取出来，并绘制箱型图如下图所示：

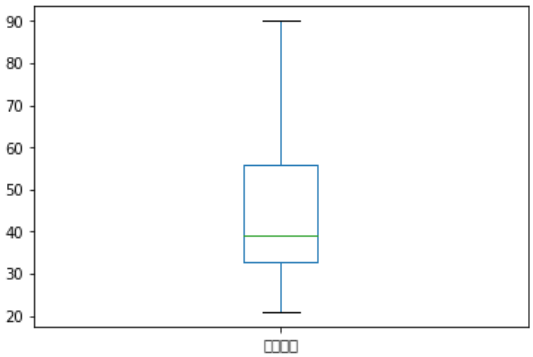

图 2 博士培养时长箱型图

可以发现，大多数博士生培养都在 35 个月（3 年）-56 个月（5 年）左右，同时有个别博士生可以在 2 年内完成学业，也有可能 7 年才拿到博士证书。同时，我们可以看出其分布的中心部分稍微偏向较高的部分。

3.2 博士问卷统计分析

3.2.1 博士区域分布

在收到的完整问卷中，我们发现其调查博士大部分都为 985 院校，只有不到 2%的高校是普通高校，这与实际中普通高校占比远远高于 985/211 的事实不符，因此，本研究只对 985 高校的博士问卷进行分析。

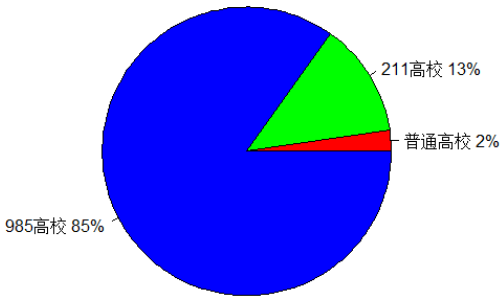

图 3 有效问卷中的高校占比

在 985 高校中，48%的博士生来自中部地区，34%来自西部，其余 18%来自东部。但其样本量在统计意义上都是“大样本”，因此可以进行接下来的数据分析工作。

3.2.2 博士录取方式分析

我们发现，100%的博士都认为“如果再次选择，依然会读博士”，可见高校博士生培养策略还是比较成功的。下表列出了他们攻读博士学位的录取方式：

表 5 博士录取方式

录取方式	数量	读博抉择
普通招考	365（48.8%）	4.456164
硕转博	293（39.2%）	4.088737
直博	89（12%）	4.230337

可以发现，大部分博士生是通过普通招考的方式进入博士生涯的学习，也有一部分是在硕士期间对研究方向感兴趣，觉得有必要继续深造从而攻读博士学位，而直博的比例相对较少。然而，普通招考的学生攻读该博士学位的决心更大，这可能是由于其普通招考的选择面更大，因此对报考的专业和学校更加认同。

3.2.3 论文发表地域差异

本研究统计了不同区域内发表国际、国内论文的数量，如下图所示：

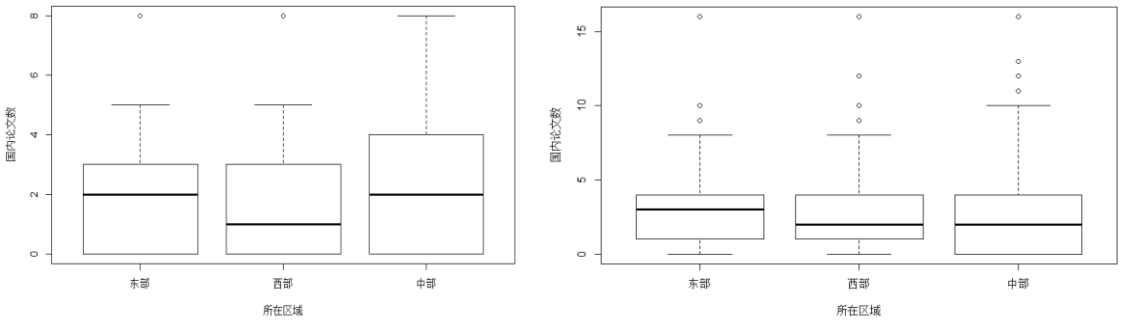

图 4 国际、国内论文发表数量箱型图

一个很有趣的现象是：东部博士更倾向于发表国际论文，而中部地区的博士更倾向于发表国内论文。一般而言，国际论文的研究水平高于国内论文，因此可看出，不同地域的博士生教育水平存在着明显差异。

4 分析方法

4.1 方差分析

4.1.1 优秀博士研究水平地域差异

由于各地区的高校教学水平存在较大的差异，其培养的博士生水平也参差不齐。但对于优秀博士论文的获奖者，其区域差异是否还存在呢？本研究使用影响因子作为衡量学术水平的指标，将其与学校所属类别和区域差异作为影响因子进行方差分析可得下表：

表 6 研究水平方差分析表

来源	Df	$Sum\ Sq$	$Mean\ Sq$	$F\ value$	$P\ value$
学校类别	2	298	149.19	8.293	0.003 ***
所属区域	3	112	37.17	2.066	0.104
交互	6	72	12.06	0.67	0.67

可以发现，尽管都是优秀博士，学校类别对于博士生的学术水平有明显的影响，而所属区域的影响不显著。

4.1.2 国际论文数量影响因素

国际论文一般意味着更高的研究质量和更好的研究水平，因此，衡量其影响因素对于博士生的培养方案制定具有重要意义。本研究使用方差分析讨论国际论文数量的的影响因素如下：

表 7 国际论文方差分析表

来源	Df	$Sum\ Sq$	$Mean\ Sq$	$F\ value$	$P\ value$
所属区域	2	80	40.2	5.625	0.0037 **
录取方式	2	326	163.03	22.899	2.25e-10 ***
导师教学	4	94	23.45	3.22	0.01 *

观察 P 值可知，国际论文的发表数量不仅与所属区域有关（通常意味着国际化程度），与学生的录取方式、导师负责程度也密切相关。通过对比箱型图可知，“硕转博”的学生发表国际论文的数量普遍比“普通报考”学生高 2 篇左右。

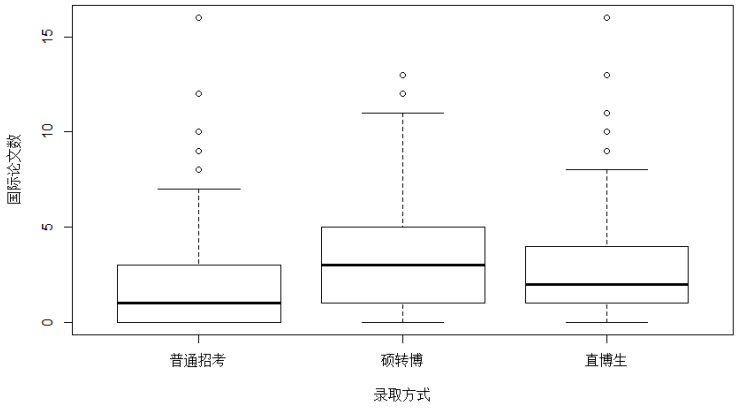

图 5 国际论文发表数量箱型图

4.1.3 副导师制度对科研效果的影响

由于世界一流大学多采用多元指导导师制度[3]，因此近些年来，我国部分高校也开始实施导师团队制度。那么，这种制度是否真的能带来博士生的科研水平提高呢？本研究以博士的国际、国内论文数量和科研活跃度作为评价指标，分析其显著水平。

表 8 副导师制度方差分析表

来源	目标	Df	$Sum\ Sq$	$Mean\ Sq$	$F\ value$	$P\ value$
是否有副导师制度	论文数量	1	4	4	0.357	0.55
是否有副导师制度	科研活跃度	1	147	147	9.323	0.002 ***

可以发现，是否有副导师制度与最终的论文产出没有明显关系，但却能较大的促进学术的科研活跃度。

4.2 softmax Regression

4.2.1 模型构建

本研究想探讨博士对工作的满意度的影响因子。由于在调查问卷中满意度分为 5 类，因此可以使用 Multinomial Logistic Regression（softmax regression）来构造模型。

多类别 Logistic 回归其实就是 Logistic 回归的扩展，主要有两种处理方法[4]：

1. 直接将多类别回归看成是 k 个二元 Logistic 回归；

2. 将 sigmoid 函数代替为 softmax 函数，构造 Softmax Regression

在本研究中，由于每个类别都是互斥的，更适用于构造 Softmax 回归。对于标记有 k 个类别的分类问题，其分类函数为：

$$h_\theta(x^{(i)}) = \begin{bmatrix} p(y^{(i)}=1|x^{(i)};\theta) \\ p(y^{(i)}=2|x^{(i)};\theta) \\ \vdots \\ p(y^{(i)}=k|x^{(i)};\theta) \end{bmatrix} = \frac{1}{\sum_{j=1}^{k} e^{\theta_j^T x^{(i)}}} \begin{bmatrix} e^{\theta_1^T x^{(i)}} \\ e^{\theta_2^T x^{(i)}} \\ \vdots \\ e^{\theta_k^T x^{(i)}} \end{bmatrix}$$

softmax 回归算法的代价函数如下所示：

$$J(\theta) = -\frac{1}{m}\left[\sum_{i=1}^{m}\sum_{j=1}^{k} 1\left\{y^{(i)}=j\right\} \log \frac{e^{\theta_j^T x^{(i)}}}{\sum_{l=1}^{k} e^{\theta_l^T x^{(i)}}}\right]$$

很明显，上述公式是 logistic 回归损失函数的推广。

4.2.2 参数显著性检验

首先需要对问卷调查中的所有参数进行显著性检验，这里还是使用方差分析的方法，得到其参数的方差分析表为：

表 9 参数显著性检验

变量	F value	P value
是否有副导师制度	3.358	0.067
录取方式	0.094	0.75
培养单位评价	43.063	9.99e-11 ***
导师评价	5.874	0.0156 *
课题评价	13.172	0.0003 ***
教学评价	8.839	0.003 **
读博抉择	11.824	0.0006 ***
科研活跃度	2.761	0.096
国际化程度	3.711	0.054

选定显著性水平 0.1，可以看出，“培养单位”、“导师水平”、“课题喜爱程度”、“教学水平”和“读博抉择”这 5 个变量是显著的。因此，去除不显著的变量，对这些变量进行参数估计。

4.2.3 参数估计

本研究使用牛顿法迭代求解，根据以上提到的目标函数进行建模，得到其参数估计为下表所示：

表 10 影响工作满意度的 softmax 回归结果

变量	非常不满意	比较不满意	一般	比较不满意	非常满意
培养单位	0.45878586	0.29298017	0.39645906	-0.28466718	-0.29178425
导师水平	-0.48125492	-0.65389009	-0.70738256	0.02159774	0.77351132
课题喜爱程度	0.03064116	0.2024683	-0.13997582	-0.09224358	-0.51308147
教学水平	-0.05859511	-0.01933294,	-0.03913871	0.1028815	-0.17446187
读博抉择	0.05042302	0.17777457	0.49003803	0.25243152	0.20581627
Intercept	-6.2984561	2.88437025,	3.55430944	3.43318641	-3.57341001

5 分析结果

5.1 博士研究水平影响因素分析

由方差分析结果可知，博士研究水平受到以下几方面的影响：

1. 对于优秀博士生而言，不同类型的高校研究水平有较大差异。

本研究分析了全国优秀博士论文获奖者的基本信息，发现处于不同类型的学校（985/211/普通高校），其顶尖博士生的水平仍然存在着较大差距。这可能是由于不同高校的教育资源和人才培养能力、资金有较大差异导致的。但同时，对于优秀博士来说，其所处学校的地域因素影响不大。

2. 不同录取方式的博士生，其学术能力有较大差异。

本研究使用国际论文来衡量其学生的学术能力。研究发现，不同录取方式的博士生，其学术能力也有着较大差异。例如，“硕转博”的学生发表国际论文的数量普遍比“普通报考”学生高 2 篇左右。

这种现象可能是由于：“硕转博”和“直博生”一方面比“普通报考”的学术更加优秀，同时其也较早受到科研锻炼的机会，因此更有机会在国际会议和期刊上发表论文。

3. 副导师制度对研究成果没有显著影响。

近些年来，为了进一步减小师生比，国内部分高校推出了副导师制度。但据本研究发现，副导师制度并不能有效提高学生的学术水平。这可能是因为老师的时间和精力有限，副导师或导师团体制度并不能增加老师对学术的学术指导时间，因此其效果并不显著。

但另一方面，副导师制度能很好的激发学生的科研兴趣。一部分原因可能是提供副导师制度的高校一般教学条件比较好，能提供更多的学术交流机会。同时，副导师制度也意味着学生能参与更多的课题项目，进一步提高了科研参与度。

5.2 博士就业满意度影响因素分析

由 Softmax 回归结果可知，影响博士就业满意度的因素主要有“培养单位”、“导师水平”、“课题喜爱程度”、“教学水平”和“读博抉择”这 5 个变量。结合拟合结果，可以得到如下结论：

1. 培养单位的环境在一定程度上可以影响学生的就业。

若培养单位环境特别恶劣（学术氛围不浓厚、学科整体水平不高），则在很大程度上会影响学生的研究水平，进而影响其就业。但对于大部分培养单位而言，学生的就业与培养单位的硬件条件没有太大关系。

2. 导师的学术水平呈两级分化的趋势。

一方面，若老师的学术能力太差，不会对学生的就业造成太大影响。这可能是因为博士阶段以自我学习为主，老师的帮助有限。但另一方面，若老师的学术能力很强且教学方法得当，会在很大程度上促进学生的学术水平，提高就业满意度。

3. 选择学术感兴趣的课题设计能提高学生的就业满意度。

若学生的科研课题是自己感兴趣、且投入了大量时间和精力，能有效调动学术的学术热情和积极性，进而提高在就业市场的竞争力。

4. 教学水平对学生的影响不明显。

不同于本科生，对于博士生来说，更多的是需要自己发现问题和解决问题的能力。因此，高校的教学水平对博士的学术和就业能力的影响有限。

5. 读博意愿对自我能力提升的影响较大。

大部分学生认识到博士阶段的学习对于自己学术能力提升重要性。同样，这种重要性也很大程度上影响了学生的就业水平和就业能力。

6 结论与政策建议

通过以上的统计分析，本研究分别对高校和学生提出如下的建议：

对于高校：

1. 建议增加直博生、硕转博学生的比例。

由统计分析可以发现，直博生、硕转博学生的学术能力普遍高于普通招考的学生。然而，目前高校的博士生入学途径主要还是依靠普通招考，这不利于学校整体科研能力的提高。

2. 慎重选择副导师制度。

虽然近些年来副导师制度在国内比较流行，但通过统计分析，并没有发现其制度能明显的提高博士生的科研水平。对于教师资源有限的普通高校而言，没有必要施行这种制度，增加教师的负担。但对于师资力量雄厚的 985 高校，可以尝试施行这种制度，也许在提高学生科研参与度的同时，也能增强学生的整体学术能力。

3. 增强高校的整体学术氛围和能力，引进优秀教师。

良好的科研环境能一定程度上影响博士生的就业，同时，优秀教师在博士生的培养中起到了重要的作用。而相比之下，更好的学术资源存在边际效应递减的情况。

对于高校学生：

1. 对于优秀的学生而言，进入一个好的高校比地理位置更重要。

本研究发现，对于优秀的博士生，决定其学术水平的主要影响因素是所在高校的水平，而不依赖于绝对的地理位置。例如，在博士生学校选择方面，一个中部地区的 985 高校也许比东部 211 高校更合适。

2. 教学水平不一定决定其学术能力，自我学习能力更为重要。

对于博士生而言，其教学水平的高低并不能决定最终的科研产出。在博士生阶段，自我管理和学习的能力、读博意愿比外界的影响更大。

3. 尽早决定是否读博。

本研究发现，直博生、硕转博学生在学术研究上的成果明显优于招考生。因此，若自己有读博意愿，应该尽早决定方向，尽早接触科研训练。

参考文献

[1]赵世奎，沈文钦．中美博士教育规模扩张的比较分析——基于 2 0 世纪 6 0 年代以来博士教育发展的数据分析[J]．教育研究，2014(1)：138－149.

[2] 王铮,许敏.电影票房的影响因素分析——基于 Logit 模型的研究[J].经济问题探索,2013(11):96-102.

[3]徐玉珍.多元指导制度下导师在博士生培养中的职责——基于香港大学与加州大学洛杉矶分校的相关规定[J].现代教育科学,2019(01):120-124+129.

[4] Softmax Regression