DOI: 10.16661/j.cnki.1672-3791.2005-5042-5940

# 国家开放大学毕业生学习时间的Pareto拟合研究

孙煜 宋丽哲
(国家开放大学信息化部 北京 100039)

**摘 要:** 远程开放教育是当今社会中最重要的教育方式之一,而远程学习者往往十分关心自己完成学业的时间,该文以学期为单位,使用$\chi^2$检验法对国家开放大学2007—2018年共12年的毕业生学习的学期数进行检验,结果表明,国家开放大学毕业生学习的学期数显著服从Pareto分布。另外,进一步使用最小二乘法对Pareto分布的参数进行了估计,为后续研究打下了基础。

**关键词:** 学习时间分布 Pareto分布 $\chi^2$检验法 拟合研究
**中图分类号:** G434 **文献标识码:** A **文章编号:** 1672-3791(2020)12(b)-0191-05

# Pareto Fitting Study of Study Time for Graduates of the Open University of China

SUN Yu SONG Lizhe
(Ministry of Information Technology, The Open University of China, Beijing, 100039 China)

**Abstract:** Distance open education is one of the most important forms of education in today's society, and distance learners are often very concerned about the time they take to complete their studies, the article examines the number of semesters studied by graduates of the Open University of China in a total of 12 years from 2007—2018 in terms of semesters using the chi square test, and the results show that the number of semesters studied by graduates of the Open University of China significantly follows the Pareto distribution. In addition, the parameters of the Pareto distribution were further estimated using the least squares method, which provided the basis for subsequent studies.

**Key Words:** Learning time distribution; Pareto distribution; Chi square test; Fitting research

现代信息化技术的快速发展和广泛应用为高等教育注入了强大的动力,远程开放教育成为学习型社会最重要的教育方式之一。远程开放教育主要通过先进信息技术和传统教育紧密结合的手段来构筑知识经济时代的终身学习体系。而参加远程教育的学生一般要经过多长的学习时间才能毕业?以往学生的毕业率情况如何?这些问题不仅为学生所关注,对于远程教育本身也具有重要意义。

国家开放大学原名中央广播电视大学,同地方广播电视大学一起,组建成为一个完整的教学和管理体系。目前,国家开放大学由总部、分部、地方学院、学习中心和行业、企业学院共同组成完整的办学组织体系,在籍学生约为453万(数据来源于国家开放大学最新统计数据)。因此,对其进行学生毕业所用学期数的相关研究具有重要意义。

过去,出于辍学现象的普遍性及给远程教育带来不利影响的严重性,大量研究人员选择“辍学”作为研究选题,例如张凤来、王文婷[1-2]在其研究中都指出了远程开放教育中,辍学率的研究一直是一个重要的课题。而对于毕业的研究相对于辍学来说,正如一个硬币的两面,研究毕业现象以提升毕业率,同样可以降低辍学率。国内也有少部分学者针对毕业率进行研究,例如宿红艳(2015)[3],徐辉、梁晓琦(2018)[4]以某所远程教育机构为个案,采用描述性统计方法研究不同

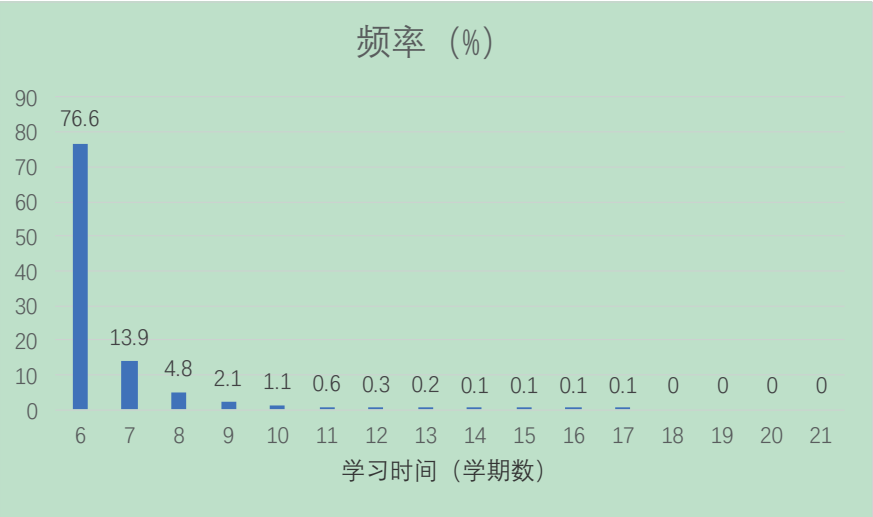


图1 学习时间频率直方图

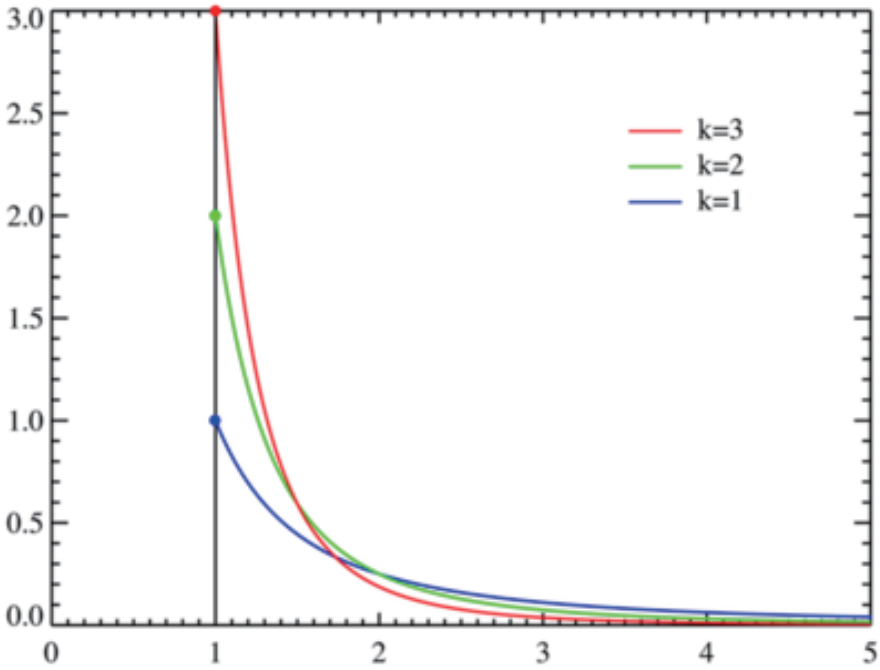


图2 Pareto分布（$x_{\min}=1$）

专业、不同性别学生的毕业率。

然而由于学生的个体差异较大，学习方式复杂，给统计和研究工作带来很大的困难，至今少有关于学生毕业学期数这方面的报告。

为了能够更好地分析影响学生学习学期数的因素，首要的就是要研究清楚学生学习学期数的分布情况。

由此，该文通过对国家开放大学2007—2018年12年的教务数据对毕业生学习学期数进行分析，根据其频率直方图以及Pareto分布的概率密度函数图，判断其服从Pareto分布，并采用$\chi^2$检验的方法来进行检验。在通过检验的基础上，建立模型估计其Pareto分布的参数，为进一步研究影响学生学习学期数的参数分析等提供一定的参考。文章第一部分介绍该文的研究意义以及方向；第二部分对数据的分布做初步拟合；第三部分介绍模型构建及估计的过程，并给出分布参数的估计；第四部分对文章做全面总结，给出结论。

## 1 学生学习时间分布初步拟合

该文选取国家开放大学2007—2018年共12年的教务数据，数据量达到了900万条，经过数据筛选，去除空数据条目，删除无效数据等，剩余数据依然有390万之多，对其进行初步统计得出，频数统计见表1。

学习时间基础统计见表2。

表1 学习时间频数统计表

| 有效学习时间 | 频数 | 百分比 | 有效百分比 | 累积百分比 |
|---|---|---|---|---|
| 6 | 2 988 879 | 76.6 | 76.6 | 76.6 |
| 7 | 542 833 | 13.9 | 13.9 | 90.6 |
| 8 | 185 391 | 4.8 | 4.8 | 95.3 |
| 9 | 80 915 | 2.1 | 2.1 | 97.4 |
| 10 | 42 470 | 1.1 | 1.1 | 98.5 |
| 11 | 23 348 | 0.6 | 0.6 | 99.1 |
| 12 | 12 543 | 0.3 | 0.3 | 99.4 |
| 13 | 6 680 | 0.2 | 0.2 | 99.6 |
| 14 | 5 505 | 0.1 | 0.1 | 99.7 |
| 15 | 4 125 | 0.1 | 0.1 | 99.8 |
| 16 | 3 620 | 0.1 | 0.1 | 99.9 |
| 17 | 2 912 | 0.1 | 0.1 | 100 |
| 18 | 673 | 0 | 0 | 100 |
| 19 | 85 | 0 | 0 | 100 |
| 20 | 20 | 0 | 0 | 100 |
| 21 | 1 | 0 | 0 | 100 |
| 总计 | 3 900 000 | 100 | 100 | |

表2 学习时间基础统计表

| 个案数 | | 平均值 | 标准偏差 | 方差 | 总和 |
|---|---|---|---|---|---|
| 有效 | 缺失 | 6.44 | 1.11 | 1.233 | 25 123 871 |
| 3 900 000 | 0 | | | | |

而后，根据频数统计表，画出频率直方图，具体见图1。

由图1可以看出，学生学习时间的分布属于截尾分布，魏顺平（2011）[5]中曾使用生存分析法、cox回归模型来对毕业影响因素进行分析，生存分析法以及cox回归模型都是针对于截尾分布的分析模型，其在不要求估计资料的分布类型的情况下，以生存结局和生存时间为应变量，能够同时分析众多因素对生存期的影响。由于这些优点，所以，在分析中很受研究人员的欢迎，但也正是因为此类模型通过半参数拟合，规避了分布类型，所以导致往往只能够关注一个终点事件，例如，只关心学生8年内能够毕业的毕业率，而不能对多个终点进行预测，而往往人们想知道的并不仅仅是一个结果，例如毕业所用学习时间的期望等。

出于这种目的，该文针对学习所用学期数进行分布拟合。

帕累托分布（Pareto distributions）[6]是以意大利经济学家维弗雷多·帕雷托命名的，是维弗雷多·帕雷托在大量真实世界的现象中，发现的幂次定律分布。这个分布在经济学以外的领域，也被称为布拉德福分布。一个多世纪以来,它在不同的领域范围内，广泛应用，也来越收到科研人员的重视。由于Pareto分布具有递减的失效率函数,经常用来描述诸如个人收入(收入越高,获得更高收入的能力就会增加)、某种药理过程后病人的存活时间(存活时间越长,能够继续存活更长时间的可能性就越高)等模型。

在帕累托分布中，如果$X$是一个随机变量，则$X$的概率分布如公式（1）所示：

$$P(X>x)=(\frac{x}{x_{\min}})^{-k} \quad (1)$$

式中，$x$是任何一个大于$x_{\min}$的数，$x_{\min}$是$x$最小的可能值（正数），$k$是为正的参数。可以看出，Pareto曲线族是由两个数量参数化的，即$x_{\min}$和$k$。

图2分别给出了在$k$=1,2,3时的Pareto分布曲线。

结合图1和图2，可以看出，学生毕业所用学期数的频率图，与Pareto分布曲线图中的$k$=3的情况大致已知，加之Pareto的广泛应用，具有良好的分析特性、丰富的参考资料。故该文考虑用Pareto分布拟合学习时间的分布。

## 2 Pareto分布的$\chi^2$检验法

前文中，根据频率图与Pareto分布曲线考虑使用

表3 分布检验对于学习时间变换表

| $i$ | $t_i$ | $x_i=\ln t_i$ | $\hat{F}(t_i)=\frac{i}{n+1}$ | $y_i=\ln\frac{1}{1-\hat{F}(t_i)}$ |
|---|---|---|---|---|
| 1 | 6 | 1.791 759 469 | 0.076 923 077 | 0.080 042 708 |
| 2 | 7 | 1.945 910 149 | 0.142 857 143 | 0.154 150 68 |
| 3 | 8 | 2.079 441 542 | 0.2 | 0.223 143 551 |
| 4 | 9 | 2.197 224 577 | 0.25 | 0.287 682 072 |
| 5 | 10 | 2.302 585 093 | 0.294 117 647 | 0.348 306 694 |
| 6 | 11 | 2.397 895 273 | 0.333 333 333 | 0.405 465 108 |
| 7 | 12 | 2.484 906 65 | 0.368 421 053 | 0.459 532 329 |
| 8 | 13 | 2.564 949 357 | 0.4 | 0.510 825 624 |
| 9 | 14 | 2.639 057 33 | 0.428 571 429 | 0.559 615 788 |
| 10 | 15 | 2.708 050 201 | 0.454 545 455 | 0.606 135 804 |
| 11 | 16 | 2.772 588 722 | 0.478 260 87 | 0.650 587 566 |
| 12 | 17 | 2.833 213 344 | 0.5 | 0.693 147 181 |

表4 参数估计对应学习时间数据变换表

| $i$ | $t_i$ | $x_i=\ln t_i$ | $f(t_i)$ | $y_i=\ln f(t_i)$ |
|---|---|---|---|---|
| 1 | 6 | 1.791 759 469 | 0.766 379 231 | −0.266 078 152 |
| 2 | 7 | 1.945 910 149 | 0.139 187 949 | −1.971 930 11 |
| 3 | 8 | 2.079 441 542 | 0.047 536 154 | −3.046 264 724 |
| 4 | 9 | 2.197 224 577 | 0.020 747 436 | −3.875 332 611 |
| 5 | 10 | 2.302 585 093 | 0.010 889 744 | −4.519 933 888 |
| 6 | 11 | 2.397 895 273 | 0.005 986 667 | −5.118 220 505 |
| 7 | 12 | 2.484 906 65 | 0.003 216 154 | −5.739 569 091 |
| 8 | 13 | 2.564 949 357 | 0.001 712 821 | −6.369 613 845 |
| 9 | 14 | 2.639 057 33 | 0.001 411 538 | −6.563 075 062 |
| 10 | 15 | 2.708 050 201 | 0.001 057 692 | −6.851 665 812 |
| 11 | 16 | 2.772 588 722 | 0.000 928 205 | −6.982 257 806 |
| 12 | 17 | 2.833 213 344 | 0.000 746 667 | −7.199 891 702 |

Pareto分布进行拟合，在拟合之前，需要对分布进行假设检验。

由于Pareto分布与双参数分布有直接的关系，在数据分析的时候，可以采用对数变换，然后利用一种针对双参数指数分布的$\chi^2$检验方法来进行分布的检验[7]，具体检验过程如下。

对一容量为$n$，截尾数为$r$的样本$t_1\leqslant\cdots\leqslant t_r$做如下变换(1)，令：

$$u_i=\frac{\sum_{j=2}^{i}y_i}{\sum_{j=2}^{r}y_i},\ i=2,\ldots\ r-1 \tag{2}$$

检验统计量为：

$$\chi^2=-2\sum_{i=2}^{r-1}\ln u_i \tag{3}$$

可以证明，统计量(2)服从自由度为$2(r-2)$的$x^2$分布。当Pareto分布假设成立时，统计量的$\chi^2$取值在给定显著水平$\alpha$后，取：

$\{\chi^2\leq\chi^2_{\frac{\alpha}{2}}(2(r-2))$ 或 $\chi^2\geq\chi^2_{1-\frac{\alpha}{2}}(2(r-2))\}$为其检验的拒绝域。

一般的，我们取显著水平为$\alpha=0.05$，学生学习数据

(下转197页)

[2] 陆杰华,沙迪.新时代农村养老服务体系面临的突出问题、主要矛盾与战略路径[J].新疆师范大学学报:哲学社会科学版,2019,40(2):78-87.
[3] 贺键雨.衡南县新型农村养老保险制度实施效果研究[D].湖南农业大学,2015.
[4] 王增文.新型城镇化背景下城乡养老保险制度及服务整合路径研究[J].华中科技大学学报:社会科学版,2019,33(2):124-130,137.
[5] 席悦.浅析我国农村养老保险制度存在的问题及对策[J].经济研究导刊,2019(1):26-27.
[6] 张玉帅.我国农村养老保险制度存在问题与解决对策研究[J].现代经济信息,2016(24):87.
[7] 李维.我国新型农村养老保险存在的问题和对策研究[J].商,2016(6):76.

(上接194页)

中,共研究其学习分布点16个,即:$n=r=12$,列出研究点的时间表见表3(单位:学期)。

根据式(1)计算$u_i$,并将学习数据带入式(2)中,计算得:

$$\chi^2 = 27.3076$$

根据$\chi^2$分布的分位数表可知:$\chi^2_{0.025}(24)=12.401$,$\chi^2_{0.975}(28)=39.364$,显然,未落入拒绝域(3)中,接受原假设,即:学习时间符合Pareto分布。

## 3 学习时间分布参数的最小二乘估计

由第三节结论,学习时间的分布服从Pareto分布:

首先,其的分布函数为:

$$F(t)=1-(\frac{\theta}{t})^{\alpha}\ ,\ t\geq\theta>0,\ \alpha>0 \tag{4}$$

其概率密度函数为:

$$f(t)=\alpha\theta^{\alpha}t^{-1-\alpha} \tag{5}$$

式中,$\alpha$是形状参数,$\theta$称为尺度参数。

根据式(4),对(4)式两侧同时取对数,可以得出:

$$\ln f(t)=\ln\alpha+\alpha\ln\theta-(1+\alpha)\ln t \tag{6}$$

由样本给出$f(t_i)$,设$x_i=\ln t_i$,$y_i=\ln f(t_i)$,$i=1,2,\ldots,n$,可以构建如下模型:

$$y=\ln\alpha+\alpha\ln\theta-(1+\alpha)x+\varepsilon^2 \tag{7}$$

其中,$\varepsilon\sim N(0,1)$,作为随机误差项存在。另外,令:

$$Y=\begin{pmatrix}y_1\\ \vdots\\ y_n\end{pmatrix},\ X=\begin{pmatrix}1 & x_1\\ \vdots & \vdots\\ 1 & x_n\end{pmatrix},\ \beta=\begin{pmatrix}\ln\alpha+\alpha\ln\theta\\ -1-\alpha\end{pmatrix}$$

由最小二乘法(OLS)可得:

$$\hat{\beta}=(X'X)^{-1}X'Y \tag{8}$$

计算整理可得参数$\alpha,\theta$的估计值为:

$$\begin{aligned}\hat{\alpha}&=\frac{\sum x_i\sum y_i-n\sum x_iy_i}{n\sum x_i^2-(\sum x_i)^2}-1\\ \hat{\theta}&=\exp\left(\frac{\sum x_i^2\sum y_i-\sum x_i\sum x_iy_i}{\hat{\alpha}n\sum x_i^2-\alpha(\sum x_i)^2}-\hat{\alpha}\ln\alpha\right)\end{aligned} \tag{9}$$

对学习时间数据做相应变换,得到表4。

将相应数据带入式(8)中可得参数估计约为:

$$\begin{aligned}\hat{\alpha}&=5.497\\ \hat{\theta}&=5.112\end{aligned} \tag{10}$$

即,国家开放大学12个学期的学生毕业所用学期数所服从的Pareto分布,形状参数α的值为5.497,尺度参数$\theta$的值为5.112。

## 4 结语

该文基于国家开放大学一共12个学期的学生毕业所用的学期数进行研究分析,假定其服从Pareto分布,并使用卡方检验方法对假定进行了假设检验,根据假设检验的结果,确定其服从Pareto分布。在此基础之上,根据Pareto的分布函数,通过最小二乘估计方法给出了分布的参数估计值。

远程教育学生毕业所用学期数的研究,对于分析其影响因素有着基础性的作用,在确定了所用学期数的分布情况之后,才可以更加准确地研究影响毕业所用学期数的因素,从而为提高毕业率、降低辍学率提供方向。

## 参考文献

[1] 张凤来.湖南远程开放教育辍学问题探析[J].创新创业理论研究与实践,2018,1(15):41-43.
[2] 王文婷.开放大学学生辍学影响因素量表的编制——基于教育功能论的开放大学学生辍学管理实践研究[J].内蒙古电大学刊,2017(3):78-82.
[3] 宿红艳.上海交通大学继续教育学院网络教育学生毕业率的研究[J].成人教育,2015,35(6):74-80.
[4] 徐辉,梁晓琦.影响开放教育续修生毕业率的相关因素研究[J].海南广播电视大学学报,2018,19(3):144-148.
[5] 魏顺平.网络高等教育学生毕业时间预测研究[J].中国远程教育,2011(10):18-27,49,95.
[6] 李海芬.Pareto分布的统计分析[D].华东师范大学,2004.
[7] 茆诗松,王静龙,濮晓龙.高等数理统计[M].北京:高等教育出版社,2006.
[8] 崔媛媛.步加试验下Pareto分布的统计分析[D].温州大学,2016.
[9] 郑丹丹.多维视角下远程开放教育学习者辍学问题的研究[J].科教导刊,2019(6):191-192.