

“二十大报告”齐普夫定律的拟合分析

张周¹ 徐晓岭² 顾蓓青² 张霞妮³ 胡高娥⁴

(1. 上海对外经贸大学国际商务外语学院, 上海 201620;

2. 上海对外经贸大学统计与信息学院, 上海 201620;

3. 上海对外经贸大学国际经贸学院, 上海 201620;

4. 上海对外经贸大学工商管理学院, 上海 201620)

【摘要】本文分别采用随机法、并列法、最小值法、平均法与最大值法,对“二十大报告”文本语料库的词频词序进行统计和分析,并依据散点分布图绘制出了齐普夫对数分布曲线,尤其是最小值法、平均法和最大值法其呈现较好的线性关系,符合齐普夫定律的设定,验证了齐普夫定律对“二十大报告”文本的适用性。

【关键词】二十大报告;齐普夫定律;最小二乘法

一、齐普夫定律简介

1932年,齐普夫(G. Zipf)提出不同单词分布的经验规律即齐普夫定律。齐普夫定律是文献计量学中的重要定律之一,并应用于多个领域。如文献计量学、文本特征选择、词典编撰、机器翻译和关键词抽取等均建立在齐普夫定律基础之上。

美国哈佛大学语言学教授齐普夫在前人研究的基础上,收集了大量的文本语料,并进行了系统的分析,正式创立了词频分布定律,验证下面的公式:若把一篇较长的文章中每个词出现的频次从高到低进行递减排列,某个词在文中出现的频率次数 f (词频)与它的排列序号数 r (词序按照词频高低排列)的乘积为一个常数 C ,即所谓的齐普夫第一定律(经典的齐普夫定律):

$$f = \frac{C}{r} \quad (1)$$

后来又引入参数 α ,推广了(1)式,即如下更一般的(2)式:

$$f(r) = \frac{C}{r^\alpha} \quad (2)$$

将(2)式两边取对数,得到:

$$\ln f = \ln C - \alpha \ln r \quad (3)$$

即在双对数坐标系下, $-\alpha$ 即直线的斜率, $\ln C$ 即为直线的截距。

二、分词原则与词语等级

分词原则是指切分词汇单位的基本原则,而在词语频次的统计中主要采用随机法、并列法、平

均值法、最小值法和最大值法这五种方法确定词语等级,除随机法外,后四种方法把同频词均放在一个等级上。具体分词原则与词语等级划分可见文献^{[1]-[4]}。

下面选用“二十大报告”作为研究分析的来源资料,此文本共有32749个书写字符,累积总词数11476个。筛选与统计的过程中,参考《现代汉语词典》,共计有2050词汇。遵循上述原则,“二十大报告”切分与统计词汇(词素)结果如表1所示,并分别列出了“二十大报告”词汇分布及随机法、并列法以及最小值法、平均值法和最大值法确定的词语等级结果。

三、统计分析

令 $y = \ln f, x = \ln r, \beta = \ln C$, 并记 $l_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2$,

$$l_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}),$$

则有: $y = \beta - \alpha x$, 于是可利用最小二乘法分别得到 α, β 的估计值 $\hat{\alpha}, \hat{\beta}$:

$$\hat{\alpha} = -\frac{l_{xy}}{l_{xx}}, \hat{\beta} = \bar{y} + \alpha \bar{x}$$

进而得参数 C 的估计: $\hat{C} = e^{\hat{\beta}}$ 。

下面分别利用随机法、并列法、最小值法、平均值法与最大值法得到的词语等级利用最小二乘法来估计参数 α, β, C 。

由随机法所得到的图1所显示的齐普夫对数分布的线性性可以看出,采用随机法分析“二十大报告”文本中的词频与词级,能够满足齐普夫分布定律,但

表 1: 党的“二十大报告”词汇（词素）的统计一览表

词（词素）	出现频率	词（词素）	出现频率	词（词素）	出现频率
的	583	战略，推动	61	战，向…	29
和	376	经济	60	新，心，从	28
发展	239	文化	58	奋斗，体制…	27
化	198	主义，问题	57	团结，统筹…	26
社会	193	世界，领导	56	理，反，所…	25
党	189	全，创新，实现	55	区，五…	24
中国	178	实，好	54	发，导…	23
人民	177	高	53	保，复兴…	22
坚持	173	国，特色，健全	52	会，质量…	21
建设	166	改革，加快	51	才，稳定…	20
是	141	工作，治理…	50	强，核心…	19
国家	137	中，民主	49	行，强化…	18
体系	119	历史，保障	48	现，创造…	17
社会主义	114	要，能力，基本…	47	加，结构…	16
性	112	上，重大	46	主，区域…	15
推进	110	更，国际	45	领，集中…	14
全面	108	十，维，群…	44	生，经，正…	13
不	103	精神	42	民，也…	12
现代	98	等，层…	41	系，信…	11
为	95	思想，促进…	40	本，族…	10
制度，加强	94	管，路，伟大…	39	法，技术…	9
完全	91	同，机制，必须	38	善，改进…	8
一，我们	87	观，进，护…	37	们，特，法律…	7
治，力，以	83	强国，人才…	36	展，渠道…	6
在	82	命	35	委，征程…	5
大	78	把，维护，保护	34	修，时刻…	4
人，完善	77	用，实施…	33	题，篇章…	3
有	72	建，力量，群众…	32	届，逃，巍…	2
各	66	到，马克思	31	蹕，毅，希…	1
政治	64	最，文明，深入…	30		

拟合程度一般，而 R^2 、方差分析以及回归系数估计和检验，回归方程与各系数都显著。此时

$$\hat{\alpha} = 1.064, \hat{\beta} = 8.570, C = e^{\hat{\beta}} = 5271.12979.$$

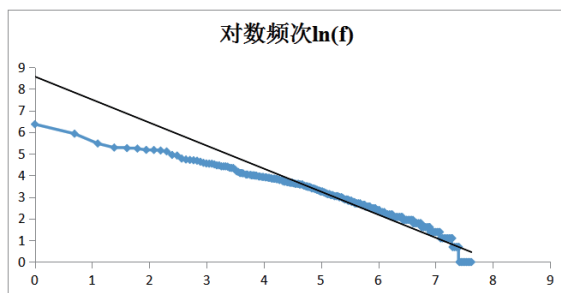


图 1: 随机法对数分布及线性回归

由并列法所得到的图 2 所显示的齐普夫对数分布的线性也可以看出，采用并列法分析“二十大报告”

文本中的词频与词级，能够满足齐普夫分布定律，但拟合程度较差，而 R^2 、方差分析以及回归系数估计和检验，回归方程与各系数都显著。此时

$$\hat{\alpha} = 1.055, \hat{\beta} = 7.428, C = e^{\hat{\beta}} = 1682.4393.$$

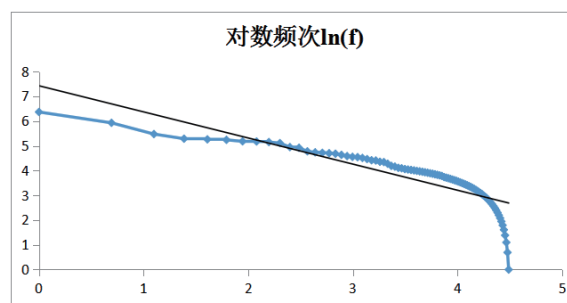


图 2: 并列法对数分布及线性回归

由最小值法所得到的图3所显示的齐普夫对数分布的线性性可以看出,采用最小值法分析“二十大报告”文本中的词频与词级,很好地满足齐普夫分布定律,而 R^2 、方差分析以及回归系数估计和检验,回归方程与各系数都显著。此时

$$\hat{\alpha} = 0.733, \hat{\beta} = 6.760, C = e^{\hat{\beta}} = 862.6422.$$

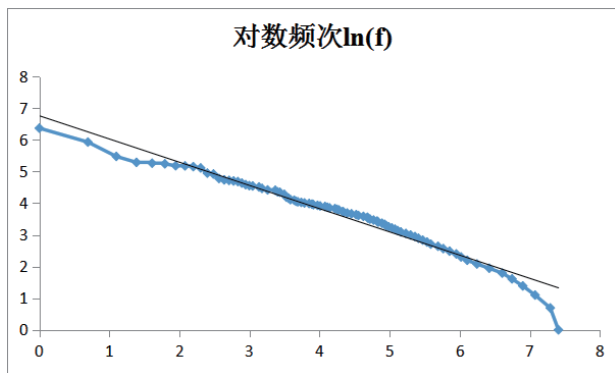


图3: 最小值法对数分布及线性回归

由平均值法所得到的图4所显示的齐普夫对数分布的线性性可以看出,采用平均值法“二十大报告”文本中的词频与词级亦很好地满足齐普夫分布定律,而 R^2 、方差分析以及回归系数估计和检验,回归方程与各系数都显著。此时

$$\hat{\alpha} = 0.724, \hat{\beta} = 6.738, C = e^{\hat{\beta}} = 843.8713.$$

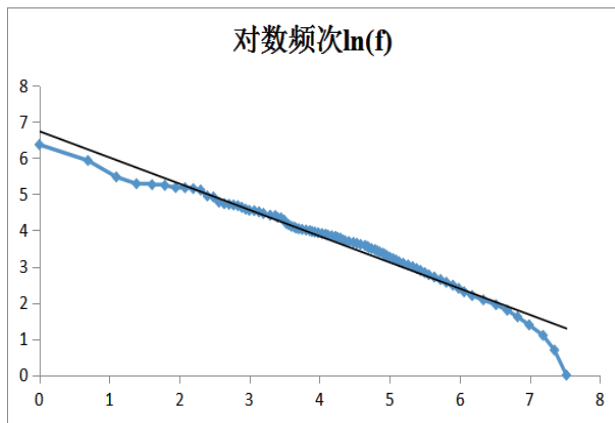


图4: 平均值法对数分布及线性回归

由最大值法所得到的图5所显示的齐普夫对数分布的线性性可以看出,采用最大值法分析“二十大报告”文本中的词频与词级,也很好地满足齐普夫分布定律,而 R^2 、方差分析以及回归系数估计和检验,回归方程与各系数都显著。此时

$$\hat{\alpha} = 0.716, \hat{\beta} = 6.717, C = e^{\hat{\beta}} = 826.33478.$$

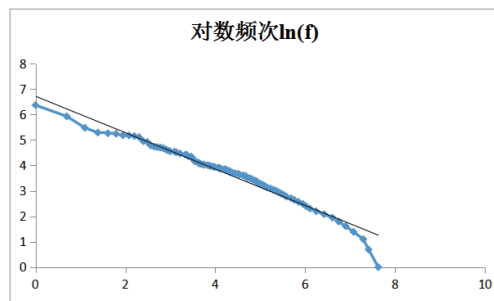


图5: 最大值法对数分布及线性回归

四、总结

综合比较各方法调整的 R^2 的值,发现用最大值法拟合最佳,其次是平均值法、最小值法、随机法,而并列法最差。而各斜率 $-\alpha$ 的值,随机法与并列法都接近于设定值-1,符合齐普夫第一定律,即满足(1)式,最小值法、平均法和最大值法相近,但都与设定值-1有较大差距,即满足(2)式。在“二十大报告”(中),出现频次很高的有:发展、党、中国、人民、坚持、建设、国家、体系、社会主义、全面、现代等,此外,经济、文化、创新、特色、治理、民主等频次也比较高,这较好地体现了在党中央的坚强领导下,中国过去五年的发展状况以及未来五年的发展趋势。

以上主要基于对“二十大报告”文本中语料库的词频词序的统计和分析,对齐普夫定律在汉语中的适用性进行了研究和验证,分别采用随机法、并列法、最小值法、平均法与最大值法,对“二十大报告”文本语料库进行了数据的统计和分析,并依据散点分布图绘制出了齐普夫对数分布曲线,尤其是最小值法、平均法和最大值法呈现出较好的线性关系,符合齐普夫定律的设定。可见,统计结果中的词频分布呈现出较为明显的齐普夫分布规律。

参考文献:

- [1] 王崇德, 来玲. 汉语文集的齐夫分布[J]. 情报科学, 1989, 10(2): 1-8, 42.
- [2] 孙清兰. 齐夫定律若干理论问题探讨与发展[J]. 情报学报, 1992, 11(2): 128-137.
- [3] 韩普, 路高飞, 王东波. 基于最大似然估计方法的齐普夫定律验证[J]. 情报理论与实践, 2012, 35(11): 6-11.
- [4] 钱学森. 科技情报工作的科学技术[J]. 科技情报工作, 1983, (10): 1-9.

基金项目: 2023年度“大学生创新创业训练计划”项目