基于迁徙数据估计武汉感染新型冠状病毒的人员数量

杨 政[1*]，原子霞[2]，贾祖瑶[1]

(1. 电子科技大学经济与管理学院 成都 611731；2. 电子科技大学数学科学学院 成都 611731)

【摘要】根据武汉迁徙数据，该文通过统计分析 2020 年 1 月 29 日至 2 月 9 日全国 50 个城市感染新型冠状病毒的确诊人数比率，估计了武汉市感染病毒的人员数量。研究发现湖北省内 15 个城市的患者确诊比率在均值和中位数上低于省外 35 个城市的均值和中位数。截至 2 月 9 日，利用湖北省内城市确诊比率的均值、中位数和最大值估计，武汉市感染病毒的人数分别是已经确诊人数的 2.1 倍、2 倍和 3.9 倍。利用省外城市确诊比率的均值、中位数和最大值估计，武汉市感染人数分别是已经确诊人数的 3.6 倍、2.6 倍和 8.7 倍。最后利用 Bootstrap 方法对省内外城市的均值和中位数做了稳健性估计。

关 键 词 Bootstrap； 新型冠状病毒； 估计； 迁徙数据； 确诊比率

中图分类号 TP391; C81 **文献标志码** A doi:10.12178/1001-0548.2020030

Estimating the Number of People Infected with COVID-19 in Wuhan Based on Migration Data

YANG Zheng[1*], YUAN Zi-xia[2], and JIA Zu-yao[1]

(1. School of Management and Economics, University Electronic Science and Technology of China Chengdu 611731; 2. School of Mathematical Science, University Electronic Science and Technology of China Chengdu 611731)

Abstract The number of COVID-19 infected persons in Wuhan is estimated by statistically analyzing the ratio of daily confirmed cases number data of COVID-19 in 50 cities in China from January 29, 2020 to February 9, 2020 owing to Wuhan migration. The study finds that the mean and median of diagnosis rate of 15 cities in Hubei Province are lower than that of 35 cities outside Hubei Province. As of February 9, 2020, using the mean, median and maximum values of diagnosis rate in inner cities of Hubei Province, it is estimated that the number of people infected in Wuhan is 2.1 times, 2 times and 3.9 times of those who have been diagnosed, respectively, while using the mean, median and maximum values of the diagnosis rate in cities outside the province, it is estimated that the number of people infected in Wuhan is 3.6 times, 2.6 times and 8.7 times of those who have been diagnosed, respectively. The mean and median of diagnosis rate in cities in and out of Hubei Province are estimated robustly by bootstrap method.

Key words Bootstrap; COVID-19; estimating; migration data; rate of diagnoses

新型冠状病毒肺炎已经成为国际关注的重大紧急公共卫生事件，给人民的生命和生活造成严重危害。因此，阻击病毒传染成了全国人民的共同战役。从 2020 年 1 月 23 日武汉开始“封城”后，各个省市采取多种防控措施。居民按照专家建议减少外出活动，在家隔离以降低被感染的风险。

2020 年 1 月 23 日−2 月 4 日，武汉市确诊人数和疑似病例的数据不断升高。国家在武汉投入更多的力量医治确诊病人，如紧急调拨物资、建立火神山和雷神山医院、派出多批次的支援医疗队等。这些措施给全国人民带来了战胜病毒的信心。此时，明确武汉市感染者的数量对于防控、诊断和治疗有重要意义。那么，武汉市目前受感染的人数有多少？这是本文拟研究的问题。

从 Elsevier 数据库查阅到新型冠状病毒的相关论文大约有 70 余篇，大致分为两类。一类侧重于从医学方面探讨新型冠状病毒的来源[1]、发现和临床诊断[2]、病毒基因分析[3]、公众心理健康[4] 以及

收稿日期：2020 − 02 − 05；修回日期：2020 − 02 − 13；网络首发日期：2020 − 02 − 23
基金项目：教育部人文社会科学研究青年基金 (15YJC790132)
作者简介：杨政 (1978 −)，男，博士生，副教授，主要从事金融计量和非线性时间序列分析等方面的研究. Email: yangzheng@uestc.edu.cn

如何控制病毒流行[5] 等问题。

另一类论文利用大数据、传播动力学模型、统计计算方法等工具对疫情进行了预测分析。文献 [6] 基于包括“易感态-潜伏态-感染态-移除态”的 SEIR 仓室模型，对病毒的基本再生数进行估计。以《人民日报》新型冠状病毒肺炎疫情实时动态数据为基准，估计基本再生数在 2.8～3.3 之间；以国外同行预测的感染人数为基准，基本再生数在 3.2～3.9 之间。文献 [7] 利用传播动力学模型，对新型冠状病毒肺炎传播风险进行了预测分析。该文利用 2020 年 1 月 10 日−1 月 22 日的报告疫情数据，采用动力学模型和统计计算方法预测基本再生数为 6.47(95% 置信区间为 5.71～7.23)，给出了疫情的达峰时间、峰值及最终感染规模，按照 2020 年 1 月 22 日前的控制措施，疫情将在 3 月 10 日左右达到峰值。文献 [8] 分析了分布在全国 31 个省市自治区、552 家医院的 1 099 个确诊病例的临床特征、潜伏期、诊断情况、治疗方式等要素，发现新型冠状病毒感染的平均潜伏期为 3 天。文献 [9] 预测了新型冠状病毒感染者的人数，估计 2020 年 1 月 25 日的感染人数约 7.5 万人。文献 [10] 根据自然增长规律动态提出数据驱动的预测方法，跟踪疫情发展并检测干预措施的有效性。在 2020 年 2 月 5 日预测约 4 天后 (2 月 9 日) 达到峰值，确诊病例总数将在 3.7 万～4.4 万之间。

上述研究并没有直接针对武汉市的感染者人数进行预估。本文在疫情前期把武汉市所有民众看作一个样本总体，离开武汉和留在武汉是由同一个总体分布中抽取的两组随机样本。本文把离开武汉的民众视为实验组样本，把留在武汉的民众视为对照组样本。武汉市受到医护人员不足和医疗物资紧缺的约束，对照组样本的确诊人数低于实际被感染人数。所以，利用实验组样本的确诊数据，分析其统计分布的数字特征，借鉴这些数字特征对武汉市的感染人数进行估计。简单来说，就是利用实验组样本的统计参数，估计对照组样本未受约束时的发展状况。

1　数据和假设

本文利用百度迁徙数据来估计武汉目前受感染的人数。统计从武汉迁入人员数量排名前 50 位城市的人数，迁徙时间从 2020 年 1 月 10 日至 1 月 22 日。表 1 给出了排名前 50 位城市从武汉迁入的人数。

表 1　排名前 50 城市从武汉迁入的人数

省内城市	人数/万	省外城市	人数/万	省外城市	人数/万
孝感市	65.620	信阳市	7.190	广州市	2.660
黄冈市	63.630	郑州市	3.385	深圳市	2.610
荆州市	31.660	南阳市	3.360	南京市	1.845
咸宁市	25.340	驻马店市	3.320	苏州市	1.315
鄂州市	20.525	周口市	2.190	杭州市	1.690
襄阳市	19.240	商丘市	1.655	温州市	1.160
黄石市	18.875	洛阳市	1.120	成都市	2.650
荆门市	15.460	长沙市	5.755	西安市	1.855
随州市	14.980	岳阳市	2.560	南宁市	1.405
仙桃市	14.345	常德市	1.580	昆明市	1.245
宜昌市	13.850	衡阳市	1.215	石家庄市	1.135
天门市	10.165	安庆市	2.310	贵阳市	1.120
恩施市	9.310	合肥市	2.260	厦门市	1.115
十堰市	8.975	阜阳市	1.690	北京市	5.130
潜江市	5.610	六安市	1.155	上海市	3.855
		九江市	2.720	重庆市	6.620
		南昌市	2.490	天津市	1.140
		宜春市	1.455		

为了估算武汉市感染新型冠状病毒的人数，本文做出如下假设。

假设 1：2020 年 1 月 10 日−1 月 22 日离开武汉的迁徙数据是准确的。

由于并没有一手的人员流动数据，故以网络报道的百度迁徙数据为准。百度迁徙数据是根据迁徙人员的手机在不同地点的定位统计的，准确性较高。例如本文根据迁徙数据计算从武汉到信阳的人数是 7.19 万人 (新闻报道的数据是 8.046 万人，包括了 22 日之后的迁徙人数)。因此，在 2020 年 1 月 22 日之前的迁徙数据具有较高的可信度。在 50 个城市中，迁徙人员是确诊新冠病毒肺炎患者的直接来源，至少在潜伏期间迁徙数据是一个最主要的影响因子。因此，50 个城市的确诊病例应该是以武汉迁入人员为主。在潜伏期之后，武汉迁入人员经过病毒潜伏期，确诊比率应该会下降。在病毒的潜伏期内，50 个城市本地居民受到武汉迁入人员传染，确诊比率应该会升高。从武汉迁入 50 个城市的人员数量仍然

是一个基础变量，它持续影响了后续感染患者的数量。

假设 2：从武汉迁徙到 50 个城市的人数在未来一段时间内保持不变。

这个假设是为了对应“封城”后武汉的人数保持不变。计算 50 个城市感染患者人数的基数也应该保持不变。事实上，这个假设对个别城市可能不成立，如温州市，据报道 2020 年 1 月 22 日之后从武汉到温州的人数比之前的人数更多。这对研究结果有一定影响，因此在后面研究中做了一些修正。

假设 3：留在武汉和离开武汉人员感染病毒的概率是同一个分布。

这个假设是估计武汉感染人数的一个重要前提。假设意味着离开和留在武汉的人员都是同一个概率分布的样本。这个假设在 2020 年 1 月 22 日之后几天可能是成立的。由于本文整理数据是 2020 年从 1 月 29 日开始，距离 1 月 22 日已经有 7 天的时间间隔。通过检验湖北省内城市和省外城市的确诊比率，发现两者并不是同一个分布。为此，本文把 50 个城市分成两组。一组是湖北省内的 15 个城市，另一组是省外的 35 个城市。即使这两组样本的分布不同，仍然假设武汉属于这两组样本中的某一种情况。如果武汉不属于这两组样本，那会是一种最差的情况。

假设 4：样本期内留在武汉和离开武汉确诊的人员都是源于相同的病毒感染模式。

这个假设表示武汉市的确诊人员和其他城市的确诊人员是相同的感染模式。比如，其他城市前期的确诊人员都是从武汉迁入被确诊，后期的部分确诊人员是受迁移人员传染而被确诊。武汉市早期的确诊人员和其他城市早期的确诊人员是同一批感染者，后期的部分人员受他人传染被确诊。

除了收集 50 个城市从武汉迁徙来的人员数量，利用百度新型冠状病毒肺炎-疫情实时大数据报告，收集整理这些城市从 2020 年 1 月 29 日至 2 月 9 日的确诊人数，并从万德 (Wind) 数据库收集武汉市的户籍人口数据。

2 湖北省内外城市确诊比率的计算

用 $i=1,2,\cdots,50$ 表示 50 个城市中的第 i 个城市，用 $x_{t,i}$ 表示第 i 个城市在第 t 天的累积确诊人数，用 y_i 表示从武汉迁出到第 i 个城市的人数。计算第 i 个城市每天的确诊比率 $p_{t,i}$ 为：

$$p_{t,i}=\frac{x_{t,i}}{y_i}\times 1\,000 \tag{1}$$

这样得到 50 个城市累计确诊比率的时间序列数据。

接下来，用 R 软件对各城市每天的数据进行描述性统计分析。为了避免个别城市的特殊值影响整体分析，除了最大值和最小值，还在描述性统计中增加了第二大值和第二小值。表 2 给出湖北省内和省外城市在确诊比率方面的描述性统计。由于确诊人数是累计值而不是每天增量值。因此均值、中位数、最大值、最小值和标准差等随着时间增加而变大。对比省内外城市，省外城市在均值，中位数、标准差、最大值及第二大值这些统计量的数值上都大于当日的省内城市。省外城市确诊比率的最小值和第二小值在 5 日和 6 日之前都小于省内城市，之后都大于省内城市。原因是省内城市早期有确诊人员，初始值较大，但是受限于医务人员的不足，确诊比率增长较慢。省外城市的初始值小，随着省外城市的医疗资源充足，潜在患者被迅速确诊，确诊比率的最小值和第二小值迅速超过省内城市。

表 2 中雅克贝拉 (Bera–Jarque, JB) 检验统计量用于检验分布是否属于正态分布。对于每日确诊比率，检验原假设 $H_0: p_t$，服从标准正态分布。根据文献 [11]，定义 JB 统计量为：

$$\mathrm{JB}=N\left[\frac{b_1^2}{6}+\frac{(b_2-3)^2}{24}\right] \tag{2}$$

式中，偏度 $b_1=\dfrac{E\left[p_{t,\cdot}^3\right]}{(\sigma^2)^{3/2}}$；峰度 $b_2=\dfrac{E\left[p_{t,\cdot}^4\right]}{(\sigma^2)^2}$；方差 $\sigma^2=E(p_{t,\cdot}-\bar{p}_{t,\cdot})^2$ 以及均值 $\bar{p}_{t,\cdot}=Ep_{t,i}$；$N$ 为样本数。从雅克贝拉 (JB) 检验 p 值来看，在 10% 的显著性水平上，省内城市并不拒绝每日的确诊比率服从正态分布。省外城市的 p 值都小于 1%，则拒绝每日确诊比率是正态分布。根据表 2 找到湖北省内和省外出现极值的城市，见表 3。湖北省内城市的感染情况更严重，不同城市出现在最大值和第二大值。特别是随州市，虽然在确诊的绝对人数上没有孝感市和黄冈市高，但是确诊比率在省内城市中连续 7 日保持第一，表明随州市的疫情非常严峻。

表 2　湖北省内和省外城市确诊比率的描述性统计

城市	时间	均值	中位数	最大值	第二大值	第二小值	最小值	标准差	JB值	p值
省内城市	1月29日	0.507	0.455	0.981	0.919	0.223	0.143	0.239	1.052	0.591
	1月30日	0.699	0.608	1.326	1.235	0.383	0.178	0.311	0.724	0.696
	1月31日	0.970	0.890	1.671	1.522	0.627	0.214	0.413	0.538	0.764
	2月1日	1.223	1.106	2.029	1.993	0.676	0.481	0.520	1.444	0.486
	2月2日	1.535	1.335	2.563	2.549	0.971	0.624	0.660	1.605	0.448
	2月3日	1.821	1.576	3.057	2.852	1.131	0.624	0.772	1.177	0.555
	2月4日	2.149	1.936	4.279	3.285	1.151	0.784	0.994	1.216	0.545
	2月5日	2.453	2.252	4.713	3.820	1.259	0.963	1.079	0.977	0.613
	2月6日	2.781	2.840	5.567	4.090	1.358	1.141	1.251	0.941	0.625
	2月7日	3.069	2.981	6.108	4.404	1.604	1.319	1.343	1.008	0.604
	2月8日	3.435	3.525	6.362	4.880	1.719	1.426	1.341	0.420	0.811
	2月9日	3.568	3.365	6.569	5.203	1.837	1.462	1.496	0.691	0.708
省外城市	1月29日	1.476	1.016	9.828	3.018	0.179	0.089	1.634	552.498	0.000
	1月30日	1.946	1.359	14.828	4.083	0.179	0.089	2.420	783.319	0.000
	1月31日	2.474	1.787	19.569	5.030	0.268	0.089	3.177	859.612	0.000
	2月1日	2.872	1.993	20.776	6.513	0.446	0.357	3.413	693.042	0.000
	2月2日	3.132	2.131	22.845	7.510	0.625	0.536	3.796	637.518	0.000
	2月3日	3.336	2.469	25.086	8.659	0.625	0.625	4.126	724.137	0.000
	2月4日	4.280	2.946	29.310	10.307	1.071	0.893	4.795	667.981	0.000
	2月5日	4.670	3.299	31.379	11.073	1.518	1.071	5.124	658.624	0.000
	2月6日	5.066	3.633	34.138	12.031	1.875	1.339	5.556	681.783	0.000
	2月7日	5.472	3.849	36.293	12.797	2.063	1.607	5.907	665.006	0.000
	2月8日	5.734	4.177	37.759	13.448	2.063	1.607	6.142	661.369	0.000
	2月9日	6.025	4.492	38.621	14.023	2.115	1.964	6.274	641.141	0.000

表 3　每天确诊比率极值对应的城市

城市	时间	最大值	第二大值	第二小值	最小值
省内城市	1月29日	十堰市	荆门市	仙桃市	潜江市
	1月30日	十堰市	荆门市	仙桃市	潜江市
	1月31日	十堰市	随州市	仙桃市	潜江市
	2月1日	随州市	宜昌市	仙桃市	潜江市
	2月2日	随州市	宜昌市	咸宁市	潜江市
	2月3日	随州市	十堰市	天门市	潜江市
	2月4日	随州市	襄阳市	天门市	潜江市
	2月5日	随州市	襄阳市	天门市	潜江市
	2月6日	随州市	襄阳市	天门市	潜江市
	2月7日	随州市	宜昌市	天门市	潜江市
	2月8日	随州市	十堰市	恩施市	潜江市
	2月9日	随州市	十堰市	恩施市	潜江市
省外城市	1月29日	温州市	杭州市	洛阳市	贵阳市
	1月30日	温州市	杭州市	洛阳市	贵阳市
	1月31日	温州市	杭州市	洛阳市	贵阳市
	2月1日	温州市	深圳市	洛阳市	贵阳市
	2月2日	温州市	深圳市	洛阳市	贵阳市
	2月3日	温州市	深圳市	洛阳市	贵阳市
	2月4日	温州市	深圳市	洛阳市	贵阳市
	2月5日	温州市	深圳市	洛阳市	贵阳市
	2月6日	温州市	深圳市	洛阳市	贵阳市
	2月7日	温州市	深圳市	厦门市	贵阳市
	2月8日	温州市	深圳市	厦门市	贵阳市
	2月9日	温州市	深圳市	石家庄市	贵阳市

省外城市在 2020 年 1 月 29 日−2 月 9 日期间的变化不大。前 3 日确诊比率最高的是浙江温州和杭州。随着浙江采取严格的防控措施，2 月 1 日之后确诊比率第二大值出现在深圳市。确诊比率最小值一直由贵阳市保持。第二小值在洛阳市、厦门市和石家庄市中变换。

对比湖北省内城市和省外城市的表现。考虑到省内和省外城市的样本数和分布相同，采用文献 [12] 提出的 F 检验做省内外的均值相等性检验，采用文献 [13] 的χ^2检验做中位数相等性检验。应用 R 软件检验均值和中位数的相等性，检验结果见表 4。

从均值检验结果看，表 4 显示省内外均值相等的原假设在 5% 水平被显著拒绝，说明湖北省内城市和省外城市的均值差异明显。从中位数检验看，2020 年 1 月 30 日−2 月 1 日，中位数相等的原假设在 1% 水平上被显著拒绝。2 月 2 日的中位数检验在 10% 水平上并不显著。在 2 月 3 日−6 日的结果出现反转，显示省内中位数持续低于省外的中位数。一个好的信号出现在 2 月 7 日，p 值在 10% 水平上不拒绝省内外中位数相等的原假设。这说明湖北省内和省外的确诊状况暂时进入一个新阶段。2 月 8 日和 9 日的中位数检验结果强化了这一结论。

3　武汉市感染病毒人数的估计

根据式 (1)，估计每日武汉感染人数为:

$$\hat{x}_{t,j}=\frac{p_{t,j}*y}{1\,000} \tag{3}$$

表 4 均值和中位数检验

时间	均值检验H_0: $M_{hubeicity} = M_{othercity}$		中位数检验H_0: $Med_{hubeicity} = Med_{othercity}$	
	Welch F	p-value	Chi-square	p-value
1月29日	11.705	0.002***	11.524	0.001***
1月30日	8.948	0.005***	11.524	0.001***
1月31日	7.548	0.009***	7.714	0.006***
2月1日	7.750	0.008***	4.667	0.031**
2月2日	5.783	0.021**	2.381	0.123
2月3日	6.692	0.014**	4.667	0.031**
2月4日	6.282	0.016**	4.667	0.031**
2月5日	5.937	0.019**	4.667	0.031**
2月6日	5.293	0.027**	4.667	0.031**
2月7日	5.167	0.028**	2.381	0.123
2月8日	4.415	0.042**	2.381	0.123
2月9日	4.741	0.035**	0.857	0.355

注：***，**和*表示在1%、5%和10%置信水平上显著

式中，$y = 883.73$万是武汉市的户籍人数；$p_{t,j}$是表 2 的比率，表示第t日第j种情况下的数值；j分别表示均值、中位数、最大值、第二大值、第二小值和最小值这 6 种情况。常见的统计估计应该包括某些置信水平下的区间估计，比如估计武汉感染人数的 95% 区间，在本文中并没有做区间估计。由于武汉市的情况很特殊，也可能不属于省内和省外的两种分布，汇报区间估计的意义不大。直接采用 4 种极值比率来估计，这样能够看到极端情况下武汉市感染人数的估计值。

表 5 给出了 6 种情况估计的武汉感染人数。表 5 的最后一列是每日公布的武汉市确诊人数。为了更好地理解估计结果，把估计值除以每日公布的确诊人数，得到估计值和公布确诊人数的比值，结果见表 6。表 5 和 6 的结果总结为以下 3 点。

表 5 6 种情况下武汉感染人数的估计

城市	时间	均值	中位数	最大值	第二大值	第二小值	最小值	公布的确诊人数
省内城市	1月29日	4 483	4 020	8 665	8 117	1 971	1 260	1 905
	1月30日	6 178	5 373	11 717	10 918	3 388	1 575	2 261
	1月31日	8 572	7 866	14 770	13 451	5 544	1 890	2 639
	2月1日	10 811	9 774	17 934	17 611	5 976	4 253	3 215
	2月2日	13 565	11 799	22 654	22 524	8 579	5 513	4 109
	2月3日	16 088	13 929	27 019	25 207	9 998	5 513	5 142
	2月4日	18 993	17 111	37 815	29 029	10 172	6 931	6 384
	2月5日	21 680	19 902	41 650	33 760	11 128	8 506	8 351
	2月6日	24 575	25 097	49 201	36 148	11 998	10 082	10 117
	2月7日	27 126	26 347	53 980	38 922	14 171	11 657	11 618
	2月8日	30 352	31 150	56 221	43 128	15 188	12 602	13 603
	2月9日	31 531	29 735	58 050	45 983	16 232	12 917	14 982
省外城市	1月29日	13 040	8 975	86 849	26 669	1 578	789	1 905
	1月30日	17 197	12 009	131 036	36 081	1 578	789	2 261
	1月31日	21 862	15 791	172 937	44 448	2 368	789	2 639
	2月1日	25 383	17 614	183 603	57 561	3 945	3 156	3 215
	2月2日	27 675	18 829	201 886	66 365	4 734	5 523	4 109
	2月3日	29 480	21 820	221 694	76 522	5 523	5 523	5 142
	2月4日	37 823	26 038	259 024	91 082	9 468	7 891	6 384
	2月5日	41 271	29 154	277 308	97 854	13 414	9 469	8 351
	2月6日	44 765	32 104	301 687	106 318	16 570	11 836	10 117
	2月7日	48 356	34 015	320 733	113 090	18 229	14 203	11 618
	2月8日	50 674	36 915	333 684	118 846	18 229	14 203	13 603
	2月9日	53 248	39 699	341 303	123 925	18 687	17 359	14 982

1) 按照确诊比率的最小值 (即最小比率) 来估计。基于省内城市确诊比率最小值 (即潜江市)，估计武汉市受感染人数。除了前 3 日的估计人数低于确诊人数外，后面 5 日的估计人数都高于确诊人数。自 2020 年 2 月 8 日开始，武汉市确诊人数开始大于估计的感染人数。再按第二小值的比率估计，1 月 29 日的估计值是 1 971 人，和确诊的 1 905 人较为接近。2 月 9 日，根据第二小值 (即恩施市) 的确诊比率估计武汉市感染人数，估计值大约是确诊人数的 1.1 倍。

从省外城市来看，按照确诊比率的最小值 (即贵阳市) 估计 2020 年 2 月 1 日武汉市的受感染人

数，估计值和确诊人数持平。用第二小值 (即洛阳市) 估计，在 2 月 1 日的估计值已经开始大于确诊人数。按照省外城市确诊比率的最小值，即最乐观的估计，2 月 2 日之后武汉市感染人数的估计值全部大于确诊人数。

2) 按照省内城市的平均值估计，表 5 显示武汉在 2020 年 1 月 29 日−2 月 9 日的感染人数分别是 4 483 人和 31 531 人。表 6 展示了比值的动态变化，由 1 月 29 日的 2.4 倍开始增加，到 2 月 1 日到达峰值即 3.4 倍，之后比值开始持续减少，2 月 9 日的比值是 2.1 倍。

在按照省外城市的平均值估计，从 2020 年 1 月 29 日−2 月 9 日的感染人数分别是 1.304 0 万人和 5.324 8 万人。表 6 的比值随时间变化的动态特征和省内情况类似，从 1 月 29 日的 6.8 倍到 1 月 31 日达到峰值即 8.3 倍，比值从 2 月 1 日开始持续下降，到 2 月 9 日估计的感染人数是确诊人数的 3.6 倍。

表 6　武汉感染人数的估计和公布确诊人数的比值

城市	时间	均值	中位数	最大值	第二大值	第二小值	最小值	公布的确诊人数
省内城市	1月29日	2.4	2.1	4.5	4.3	1.0	0.7	1.0
	1月30日	2.7	2.4	5.2	4.8	1.5	0.7	1.0
	1月31日	3.2	3.0	5.6	5.1	2.1	0.7	1.0
	2月1日	3.4	3.0	5.6	5.5	1.9	1.3	1.0
	2月2日	3.3	2.9	5.5	5.5	2.1	1.3	1.0
	2月3日	3.1	2.7	5.3	4.9	1.9	1.1	1.0
	2月4日	3.0	2.7	5.9	4.5	1.6	1.1	1.0
	2月5日	2.6	2.4	5.0	4.0	1.3	1.0	1.0
	2月6日	2.4	2.5	4.9	3.6	1.2	1.0	1.0
	2月7日	2.3	2.3	4.6	3.4	1.2	1.0	1.0
	2月8日	2.2	2.3	4.1	3.2	1.1	0.9	1.0
	2月9日	2.1	2.0	3.9	3.1	1.1	0.9	1.0
省外城市	1月29日	6.8	4.7	45.6	14.0	0.8	0.4	1.0
	1月30日	7.6	5.3	58.0	16.0	0.7	0.3	1.0
	1月31日	8.3	6.0	65.5	16.8	0.9	0.3	1.0
	2月1日	7.9	5.5	57.1	17.9	1.2	1.0	1.0
	2月2日	6.7	4.6	49.1	16.2	1.3	1.0	1.0
	2月3日	5.7	4.2	43.1	13.6	1.1	1.0	1.0
	2月4日	5.9	4.1	40.6	14.3	1.5	1.2	1.0
	2月5日	4.9	3.5	33.2	11.7	1.6	1.1	1.0
	2月6日	4.4	3.2	29.8	10.5	1.6	1.2	1.0
	2月7日	4.2	2.9	27.6	9.7	1.6	1.2	1.0
	2月8日	3.7	2.7	24.5	8.7	1.3	1.0	1.0
	2月9日	3.6	2.6	22.8	8.3	1.2	1.2	1.0

按中位数估计，省内城市的比值从 1 月 31 日的 3 倍减少到 2 月 9 日的 2 倍。利用省外城市估计的感染人数和确诊人数的比值，从 1 月 29 日的 4.7 倍增加到 1 月 31 日的 6 倍，再逐步减少到 2 月 9 日的 2.6 倍。

3) 从感染确诊比率的最大值 (即最大比率) 来估计。按照省内城市的最大值 (即十堰市和随州市) 估计，从 2020 年 1 月 29 日−2 月 9 日，估计的武汉感染人数是当日确诊人数的 4.5 倍和 3.9 倍，期间比值在 2 月 4 日达到最大的 5.9 倍。按照省内城市的第二大值 (分别是荆门市和十堰市) 来估计，1 月 29 日和 2 月 9 日的估计值分别是 0.811 7 万人和 4.598 3 万人。武汉市估计的感染人数是确诊人数的 4.3 倍和 3.1 倍。

按照省外城市的最大值 (即温州市) 估计，从 2020 年 1 月 29 日−2 月 9 日，估计的感染人数和确诊人数比值从 45.6 倍 (1 月 29 日) 增加到 65.5 倍 (1 月 31 日)，再逐步减少到 22.8 倍 (2 月 9 日)。1 月 29 日的估计值是 8.685 万人，2 月 9 日的估计值是 34.130 万人。用省外城市的第二大值来估计，1 月 29 日 (杭州市) 和 2 月 9 日 (深圳市) 的估计值分别是 2.666 9 万人和 12.392 5 万人，估计的感染人数分别是当日确诊人数的 14 倍和 8.3 倍。

用省外城市的最大比率 (即温州市) 估计出武汉市的感染人数在 2 月 9 日是 34.130 万人，这个结果令人吃惊。追查从武汉回到温州的人数，温州

市副市长在 2020 年 1 月 29 日采访中提到：“武汉‘封城’后，1 月 23 日至 27 日 5 天，仍然有 1.88 万人从湖北特别是武汉到达温州，平均每天有 3 600 多人”。因此从武汉回到温州是 3.04 万人，大于百度迁徙数据计算的 2020 年 1 月 10 日至 22 日的 1.16 万人，回到温州的实际人数增加了 1.6 倍。假设温州的累计确诊人数是从 3.04 万人中得到的，那么估算武汉市感染人数大约为 13.127 万人 (34.130 万人/2.6)，仍然高于由第二大值 (即深圳市) 确诊比率估计的 12.393 万人。调整后的估计感染人数与当日确诊人数的比值是 8.7 倍，高于第二大值的 8.3 倍。从表 6 看到，省外城市感染比率第二大的城市在 1 月 29 日是杭州。从 2 月 1 日之后就是深圳市。显然，从确诊率高的温州市和深圳市估计武汉市的感染人数，结果较为一致。

4 省内外城市均值和中位数的 Bootstrap 估计

由于研究样本较少，尤其是省内的 15 个样本属于小样本情形。对均值和中位数的估计可能会有一些影响。为此，采用 Bootstrap 方法[14] 重新估计每日的均值和中位数。具体步骤为：

1) 在原始样本$p_{t,i}$中有放回的抽样，得到 N 个样本$p^*_{t,i}$，其中省内样本$N=15$，省外样本$N=35$。

2) 利用抽取的 Bootstrap 样本$p^*_{t,i}$，计算 Bootstrap 抽样下的均值统计量：

$$\text{Mean}^* = \frac{1}{N}\sum_{i=1}^{N} p^*_{t,i} \tag{4}$$

和中位数统计量：

$$\text{Med}^* = p^*_{t,k}, k = [N/2]+1 \tag{5}$$

其中$[a]$表示对a取整。

3) 重复第 1) 步和第 2) 步共 B 次，得到均值$\text{Mean}^*_1, \text{Mean}^*_2, \cdots, \text{Mean}^*_B$和中位数$\text{Med}^*_1, \text{Med}^*_2, \cdots, \text{Med}^*_B$。

4) 计算 Bootstrap 均值

$$\overline{\text{Mean}}^* = \frac{1}{B}\sum_{b=1}^{B} \text{Mean}^*_b \tag{6}$$

和 Bootstrap 中位数的平均：

$$\overline{\text{Med}}^* = \frac{1}{B}\sum_{b=1}^{B} \text{Med}^*_b \tag{7}$$

对湖北省内外城市的每日样本进行 Bootstrap 抽样 B=10 000 次，计算得到 Bootstrap 均值和中位数，结果见表 7。图 1 给出从 2020 年 1 月 29 日到 2 月 9 日每天的直方图。

表 7 基于 bootstrap 抽样计算的省内外城市的均值和中位数

时间	省内城市		省外城市	
	均值	中位数	均值	中位数
1月29日	0.508	0.461	1.473	1.054
1月30日	0.698	0.655	1.950	1.383
1月31日	0.970	0.892	2.476	1.808
2月1日	1.225	1.093	2.864	2.035
2月2日	1.535	1.353	3.260	2.280
2月3日	1.821	1.661	3.701	2.668
2月4日	2.149	1.957	4.287	3.017
2月5日	2.449	2.308	4.690	3.341
2月6日	2.780	2.678	5.125	3.596
2月7日	3.070	2.953	5.499	3.874
2月8日	3.432	3.397	5.748	4.137
2月9日	3.570	3.462	6.029	4.408

把表 7 的 Bootstrap 均值和中位数与表 2 的均值和中位数对比，二者数据非常接近，说明均值和中位数的结果是鲁棒的。

图 1 的直方图反映了湖北省内城市和省外城市确诊比率在均值上的差异。直方图是经验分布的直观表现。图 1 表明省内外均值的差异是多方面的。省内确诊比率的均值始终小于省外均值，峰值高表明确诊比率在均值周围的频次非常高，表明省内城市的确诊比率非常集中，方差小同样表明确诊比率在均值周围变化小。省外城市确诊比率的特点是方差更大。

图 1 的另一个重要特征是随着时间变化，省内城市的直方图和省外城市的直方图产生了交集，当时间增加，交集重合的部分越来越多。从表 7 看到，省内城市的均值以更快的速度增加，1 月 29 日是 0.508，2 月 9 日是 3.570，大约增加了 6 倍。省外城市的均值在 1 月 29 日是 1.473，2 月 9 日是 6.029，大约增加了 3 倍。从表 7 还可以看到从 1 月 29 日到 2 月 9 日，省内城市的中位数增加了 6.5 倍，而省外城市的中位数增加了 3.2 倍。这表明随着湖北省内城市医疗条件的改善，确诊比率提高得越来越快，逐渐跟上省外城市的确诊趋势。只有当省内外的均值和中位数在统计检验上不再有显著差异时，省内外的疫情达到了相同的水平。

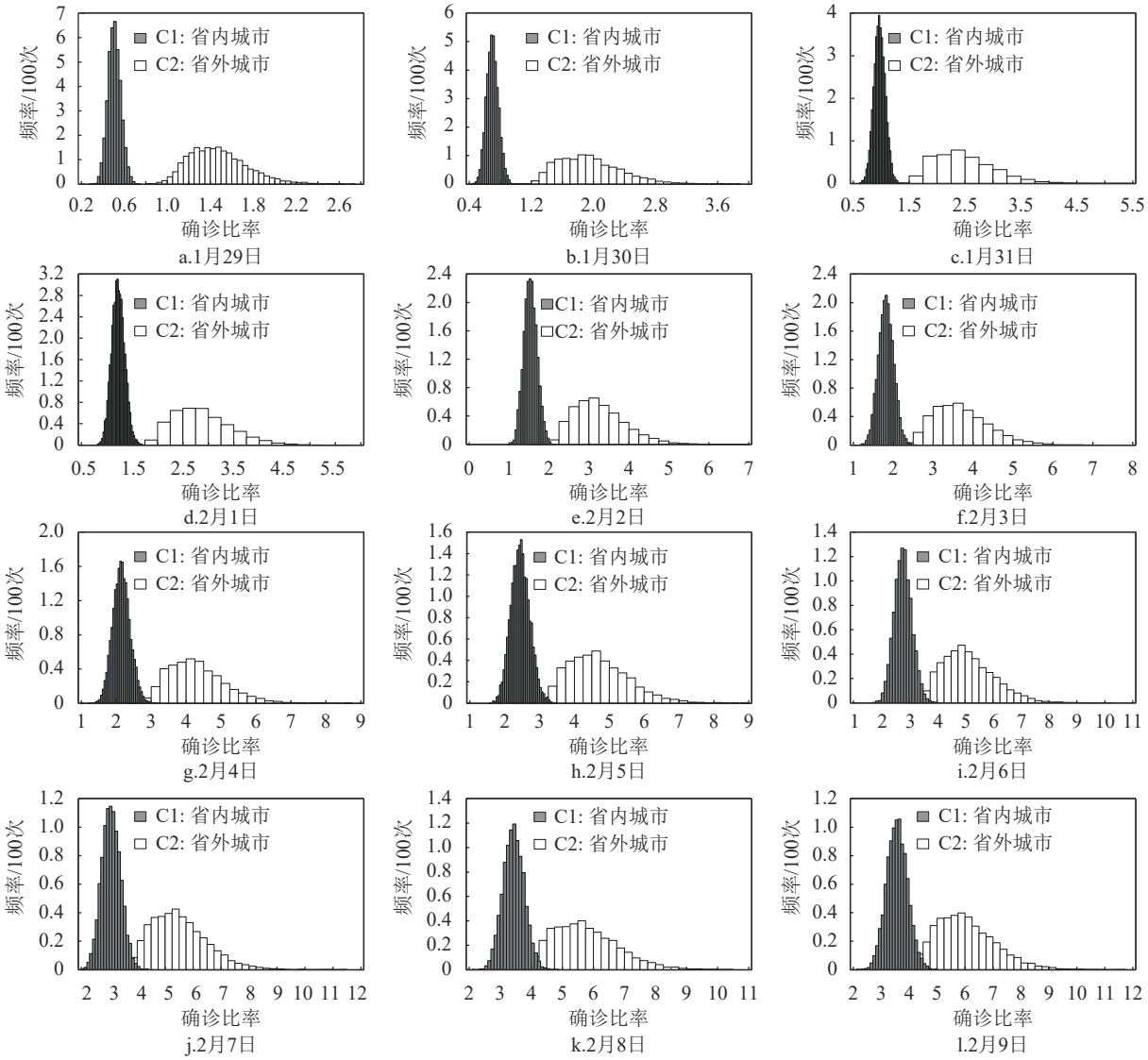

图 1　基于 10 000 次 Bootstrap 抽样估计均值的直方图

5　结 束 语

5.1　研究结论

利用 2020 年 1 月 10 日−1 月 22 日的百度迁徙数据，本文统计从武汉市到全国 50 个城市的迁移人数。同时，收集 2020 年 1 月 29 日−2 月 9 日这 12 天内 50 个城市感染新型冠状病毒确诊人数的数据。首先，利用统计方法计算了感染新型冠状病毒人数占迁移人数的比率。其次，对省内外每天的确诊比率进行描述性统计，以及均值和中位数相等性检验。接下来，根据省内外的统计结果，对武汉市的感染人数进行估计。最后对均值和中位数进行了 Bootstrap 抽样计算，均值和中位数结果具有稳健性。本文研究得到以下结论。

1) 通过对比发现，湖北省内城市确诊人数的均值和中位数都低于省外城市的均值和中位数。原因是疫情初期湖北省内的医疗资源不足，许多感染病人还未得到有效的诊断和治疗。潜在病人尚未被发现，这需要特别重视。随州、十堰、襄阳、宜昌市和荆门等城市，在样本期内的确诊比率处于 15 个省内城市的前两位。从均值检验来看，湖北省内城市和省外城市的差异是显著的。从中位数检验来看，湖北省内城市和省外城市的差异在 2 月 7 日发生了改变，不拒绝在 10% 水平上中位数相等的假设。这表明湖北省内城市感染者的确诊逐渐赶上省外城市确诊的速度。当省内和省外城市在均值检验也无显著差异时，才能认为省内外疫情状态达到同一个水平，疫情防控进入一个新的阶段。

2) 从最近一天 (2 月 9 日) 的情况来看，利用省内城市确诊比率的均值和中位数估计武汉市的感染

人数，估计值是确诊人数的 2.1 倍和 2 倍。利用最大值和第二大值估计，感染人数是确诊人数的 3.9 倍和 3.1 倍。利用省外城市的均值和中位数估计武汉市的感染人数，估计值是确诊人数的 3.6 倍和 2.6 倍。用最大值和第二大值估计的武汉市感染人数，是确诊人数的 8.7 倍 (修正后) 和 8.3 倍。这些结果无不说明武汉市内有很多潜在的感染患者尚未得到诊断。根据通报信息，武汉市前期已经征用和开辟了 9 000 张床位，雷神山医院的 1 000～1 500 张床位，火神山医院的 700～1 000 张床位。2 月 4 日武汉市征用 11 家场馆改造成“方舱医院”，改造完成后，可提供万余张床位。这些床位数量加在一起仍然小于估计的感染人数。

3) 利用 Bootstrap 方法重新估计湖北省内外城市确诊比率的均值和中位数。稳健性的结论进一步支持对武汉市感染患者的预测结果。

5.2 进一步的讨论

首先，由于作者不具备医学方面的专业知识，无法从传染病模型、病毒潜伏期、基本再生数及感染传播机制等方面进行分析。本文的假设也忽略了病毒二代传播在不同地方的差异性，这些差异性对估计结果会有一定的影响，使得估计值和实际感染人数有一定偏差。

其次，基础数据的准确性会影响估计结果。由于研究数据是根据网络上百度迁移数据整理得到，而实际情况更复杂，整理的数据与实际数据有差异。50 个城市迁徙人员的基数变小使得计算的确诊率偏高，导致武汉感染人数的估计值也偏高。

最后，文中没有考虑 50 个城市每日增加确诊人数的动态特征，利用面板分析方法研究动态数据会得到新的启示，比如判断疫情拐点的出现。利用更多的数据信息，更复杂的统计和大数据研究方法，研究结论将会更丰富。

参 考 文 献

[1] DONALD R J, SINGER B M D, FBPHARMS F. A new pandemic out of China: The Wuhan coronavirus syndrome[J]. Health Policy and Technology, 2020, 9(1): 1-2.

[2] PHAN T. Novel coronavirus: From discovery to clinical diagnostics[J]. Infection, Genetics and Evolution, 2020, 79: 104211.

[3] PARASKEVIS D, KOSTAKI E G, MAGIORKINIS G, et al. Full-genome evolutionary analysis of the novel corona virus (2019-nCoV) rejects the hypothesis of emergence as a result of a recent recombination event[J]. Infection, Genetics and Evolution, 2020, 79: 104212.

[4] BAO Yan-ping, SUN Yan-kun, MENG Shi-qiu, et al. 2019-nCoV epidemic: Address mental health care to empower society[J]. The Lancet, 2020, 395(10224): E37-E38.

[5] WANG Fu-sheng, ZHANG Chao. What to do next to control the 2019-nCoV epidemic?[J]. The Lancet, 2020, DOI: 10.1016/S0140-6736(20)30300-7.

[6] ZHOU Tao, LIU Quan-hui, YANG Zi-mo, et al. Preliminary prediction of the basic reproduction number of the novel coronavirus 2019-nCoV[EB/OL]. [2020-02-03]. http//arxiv.org/abs/2001.10530.

[7] TANG Biao, WANG Xia, LI Qian, et al. Estimation of the transmission risk of 2019-nCov and its implication for public health interventions[DB/OL]. (2020-01-27). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3525558.

[8] GUAN Wei-jie, NI Zheng-yi, HU Yu, et al. Clinical characteristics of 2019 novel coronavirus infection in China [EB/OL]. (2020-02-06). https://www.medrxiv.org/content/10.1101/2020.02.06.20020974v1.

[9] WU J T, LEUNG K, LEUNG G M. Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: A modelling study[J]. The Lancet, 2020, 395(10225): 689-697.

[10] HUANG N E, QIAO Fang-li. A data driven time-dependent transmission rate for tracking an epidemic: A case study of 2019-nCoV[J]. Science Bulletin, 2020, 65(6): 425-427.

[11] BROOKS C. Introductory econometrics for finance[M]. 3rd Ed. Cambridge: Cambridge University Press, 2014.

[12] WELCH B L. On the comparison of several mean values: An alternative approach[J]. Biometrika, 1951, 38(3-4): 330-336.

[13] CONOVER W J. Rank tests for one sample, two samples, and k samples without the assumption of a continuous distribution function[J]. The Annals of Statistics, 1973, 1(6): 1105-1125.

[14] EFRON B. Bootstrap methods: Another look at the Jackknife[J]. The Annals of Statistics, 1992, 7: 1-26.

编 辑 蒋 晓