抽样调查期末论文

数学科学学院
统计与金融数学系2005级
林荟 05231113
张睿琦 05231150

2008 年 1 月 18 日

目 录

§1 前言

随着08年奥运的临近，环境保护成为热点话题，08年奥运北京响亮提出了“绿色奥运”的口号。环境问题的研究的意义越来越大。我们知道，二氧化硫是常见的工业废气及大气污染的成分。有关研究表明，中国每排放一吨二氧化硫造成的经济损失约2万元；中国国民经济和社会发展“十一五”规划纲要明确提出，到2010年，全国二氧化硫排放总量必须控制在2295万吨。2006年国家环保总局与7个排放大省签订了二氧化硫总量控制目标责任书。环保总局还将展开二氧化硫排放指标有偿取得和排污交易试点，充分运用市场机制配置资源的基础作用，以最小的治理成本达到最优的减排效果。2007年国家环保总局得到了关于2006年重点城市工业废气排放及处理情况的调查结果。我们选取06年国家对重点城市企业污染物排放量的普查数据作为分析对象，主要关心的是工业二氧化硫排放量的数据信息。

§2 研究目的

1、对国家重点城市污染物排放情况做一个初步分析，得到一些基本统计量。

2、寻找切实可行的抽样方案避免年年普查耗费大量人力财力。

3、对数据进一步分析，探求各组数据的内在关系，从而得到各城市废气排放的标准。

§3 参数说明

$Z_{\alpha/2}$:正态分布的双侧$\alpha/2$的分位临界限
S_w^2:组内方差
S_i^2:第i层层内方差
n:抽取样本总量
N:全体样本总数
n_i:第i层抽取样本容量
n_{i0}:第i层中每个子总体抽取的样本数
w_i:第i层样本个数占总体的比例
$\bar{y_i}$:第i层样本均值
d_i:第i层的绝对误差上限
d:误差范围
$so2$:二氧化硫排放量
rm:燃煤量
ry:燃油量
qc:二氧化硫去除量
ss:设施数
tl:脱硫数

§4 方案设计总述

对上述问题，我们试图通过运用多种抽样的办法在误差范围较小的情况下,抽取出少量的城市，调查其二氧化硫排放量的数据来估计全国总体二氧化硫的排放情况。通过对已知的2006年重点城市的二氧化硫排放量数据的讨论，我们决定采用分层抽样的方法。将原始数据分层，关键是确定分层标准。我们锁定两种分层的标准：

1.按照工业城市企业的工业总产值划分为大型、中型、小型工业企业城市；

2. 按照二氧化硫排放量依照层内方差小，层间方差大的标准分层。

在**1**的分层方法中，除去每层种二氧化硫排放量的离群点（将它们视为必抽样本），用两种办法进行抽样：

（**1**）简单随机抽样。

（**2**）**Neyman**分配法。

在**2**的分层方法中我们根据在各个子层内按照按比例抽样的方式确定各层的样本容量，然后采用随机抽样的方式确定每一层的样本。在具体实施的过程中在第一次选取的d值下的到的样本个数太多，结果不如意，于是我们对d值进行了调整，第二次的到的样本个数及结果比第一次好很多。

§4.1 方案一

§4.1.1 分层

表 1 统计上大中小型企业划分标准

行业名称	指标名称	计算单位	大型	中型	小型
工业企业	资产总额	万元	40000及以上	4000-40000以下	4000以下

表 2 根据企业规模分层

类型	大型	中型	小型
工业企业资产	40000及以上	4000-40000以下	4000以下
个数	16	96	1
二氧化硫排放量均值	125160	99860	8415
二氧化硫排放量方差	9.8096e+009	7.6954e+009	0

根据国家关于”统计上大中小型企业划分标准”(见表1)计算出全国113个城市企业的工业产值,利用：城市企业的工业产值=工业总产值/汇总工业企业数,得到城市按照工业总产值从大到小的排序(附录1).根据企业规模将总体分为三层.(如表2)

§4.1.2 大型企业

一、必抽样本

表 3 必抽样本

上海	374327
唐山	296038
苏州	228962

考虑到大型企业废气排放量大，所占的比重高，我们决定将该层先单独分析。图1为大型企业二氧化硫排放量的散点图，图2 为箱线图，两个图都明显显示有几个大型工业城市废气排放量显著高于其余城市，考虑到这几个城市的影响力特别大，我们决定将这几个城市列为必抽样本(表3)。

二、其余样本分层后简单随机抽样

其余的13个城市由于层内方差大，分两层进行简单随机抽样。分层结果如表4。

$$
\begin{aligned}
d_1 &= 300000 \\
d_2 &= 200000 \\
n_{01} &= \frac{Z_{\alpha/2}^2 S^2}{d_1^2} = \frac{1.96^2 * 3.2118 * 10^8}{30000^2} = 1.7439 \\
n_1 &= \frac{n_0}{1 + n_0/N} = 1.3512 \approx 1 \\
n_{02} &= \frac{Z_{\alpha/2}^2 S^2}{d_2^2} = \frac{1.96^2 * 7.1313 * 10^8}{20000^2} = 6.8489 \\
n_2 &= \frac{n_{02}}{1 + n_{02}/N} = 3.4618 = 3
\end{aligned}
$$

最终结果如表5。

表 4 其余13个样本分层

层1	南京	145627
	无锡	141992
	武汉	132572
	呼和浩特	121131
	金昌	106401
	鞍山	103237
层2	北京	93755
	厦门	68969
	长春	52727
	马鞍山	51969
	芜湖	46867
	克拉玛依	24544
	延安	13448

表 5 大型企业分层结果

标准	总体	均值	方差	样本个数
>100000	6	120670	$4.0855*10^8$	1
$<=100000$	7	50326	$7.1313*10^8$	3
必抽	3	299780	$5.2932*10^9$	3
总计	16	475266	$6.3275*10^9$	7

§4.1.3 中型企业

一、必抽样本

为了直观反映数据分布，我们首先画出了中型企业二氧化硫排放量的点图，箱线图(图3,图4)。与分析大型企业城市相类似的方法，我们将4个离群点(表6)视为必抽样本。

二、其余92个元素

画出剩余的92个中型企业城市二氧化硫的排放量散点图(图5) 分析图可知

表 6 中型企业城市必抽样本

重庆	711537
渭南	338956
洛阳	295548
天津	232282

1、数据分布较为均匀，不存在明显的分层现象。
2、拉萨、海口的二氧化硫排放量分别为总体均值的0.5%，0.2%，可以不纳入考虑范围。
我们将剩余的中型企业平均分为三层(表**7**)，对每层进行简单随机抽样。

令

$Z_{\alpha/2}=1.96$， 置信度为95%

$$\begin{aligned}
d_1 &= 20000\\
d_2 &= 8000\\
d_3 &= 8000\\
n_{01} &= \frac{Z_{\alpha/2}^2 S_1^2}{d_1^2}=\frac{1.96^2*1.2023*10^9}{20000^2}=11.5469\\
n_1 &= \frac{n_0}{1+n_0/N_1}=8.3377\approx 8\\
n_{02} &= \frac{Z_{\alpha/2}^2 S_2^2}{d_2^2}=\frac{1.96^2*1.1559*10^8}{8000^2}=6.9383\\
n_2 &= \frac{n_{02}}{1+n_{02}/N_2}=5.6350\approx 5\\
n_{03} &= \frac{Z_{\alpha/2}^2 S_3^2}{d_3^2}=\frac{1.96^2*2.4023*10^8}{8000^2}=9.7388\\
n_3 &= \frac{n_{03}}{1+n_{03}/N_3}=7.3521\approx 7
\end{aligned}$$

计算结果如表**9**。

§4.1.4 小型企业

只有张家界一个城市,为必抽样本。

§4.1.5 方案一最终抽样结果

按上述的方法进行抽样，数据结果如表8。由抽样结果可见,样本与总体均值的差在误差允许范围之内,抽样结果比较令人满意. 这种将城市按照企业工业总产值划分3层,在每层种采用简单随机抽样的办法不能保证各个层内的样本具有比较相似的性质，则这样的性质导致抽取到的少量样本不能较好的反应每层的信息,这就需要增加抽取样本的容量,增加了抽样的费用和负担并且不能达到较高的精度要求.

§4.2 方案二——用NEYMAN分配

将中型企业平均分为三层,用**NEYMAN**分配法进行抽样:

$$
\begin{aligned}
w_i &= \frac{N_i}{N} \\
S_w^2 &= \sum_{i=1}^{n} w_i S_i^2 \\
\bar{S} &= \sum_{i=1}^{n} w_i S_i \\
n_0 &= Z_{\alpha/2}^2 N^2 S_w^2 / d^2 \\
n_0' &= Z_{\alpha/2}^2 N^2 (\bar{S})^2 / d^2 \\
n &= \frac{n_0'}{1 + n_0/N} \\
n_i &= n * \frac{w_i S_i}{\sum_{i=1}^{n} w_i S_i} \\
w_1 &= \frac{16}{113} = 0.1416 \quad w_2 = \frac{32}{113} = 0.2832 \\
w_3 &= \frac{32}{113} = 0.2832 \quad w_4 \\
&= \frac{32}{113} = 0.2832 \quad w_5 = \frac{1}{113} = 0.0088 \\
S_1 &= 9.8096 * 10^9 \ , \ s_1 = 9.904 * 10^4 \\
S_2 &= 4.9175 * 10^9 \ , \ s_2 = 7.0125 * 10^4 \\
S_3 &= 1.5299 * 10^9 \ , \ s_3 = 3.9114 * 10^4 \\
S_4 &= 1.6755 * 10^1 0 \ , \ s_4 = 1.2944 * 10^5 \\
S_5 &= s_5 = 0 \\
S_w^2 &= 7.9596 * 10^9 \\
\bar{S} &= 8.1614 * 10^4 \quad \bar{S^2} = 6.6609 * 10^9
\end{aligned}
$$

$$
\begin{aligned}
n_0 &= 9.7611 * 10^3 \quad n_0' = 8.9423 * 10^3 \quad n = 102.3364 \approx 102 \\
\sum_{i=1}^{5} w_i s_i &= 16/113 * 9.9044 * 10^4 + 32/113 * (7.0125 * 10^4 + 3.9114 * 10^4 + 1.2944 * 10^5) + 1/113 * 0 \\
&= 8.1614 * 10^4 \\
n_1 &= 20 \\
n_2 &= 24 \\
n_3 &= 13 \\
n_4 &= 45 \\
n_5 &= 0
\end{aligned}
$$

可以看出即使是用**NEYMAN** 分配效果也是非常不好的，这是因为虽然按照总产值来分层，比较简单明显，但是工业企业的规模大小与二氧化硫的排放量之间没有很强的相关性，我们尝试用更加显著的因素做为依据重新进行抽样。方案三我们以二氧化硫排放量为标准分层。

§4.3 方案三——以二氧化硫排放为标准

§4.3.1 对数据初步分析

为了直观的反映二氧化硫排放量数据的统计性质，我们计算出数据的各统计量如表**10**所示，可以看出城市间二氧化硫排放量差别很大。进一步画出二氧化硫排放量的箱线图**(**图**6)**。由箱线图可以看出有**5**个离群值，且都是较大离群值，考虑到这些地区二氧化硫排放量占的比重较大，我们决定将这**5**个地区单独选出作为必抽地区**(**表**11)**。

§4.3.2 分层抽样

除去五个必抽样本后，我们对其余的**108**个城市采用分层抽样

一、子层的划分

准则：各子层方差的和最小，层间方差最大。具体方法如下：

1、确定分的层数**k**。
2、对除掉**5**个离群值后的**108**个数据进行排序。
3、对排序后的**108**个数据通过编程得到所有的分为**k**层的方案，每种方案算出各层方差的和，取方差和最小的那个。

通过**MATLAB**编程计算，我们将总体分为**4**层，对于按二氧化硫排放量

排序后的样本序列第**1 9**为第一层，**10 34**为第二层，**35 70**为第三层，**71 108**为第四层。通过计算，得到各层方差分别为：$1*10^9*(3.00323.05812.75702.5406)$

§4.3.3 子层内部按比例抽样

由上我们可以看出各子层的方差差别不是很显著，因此我们在各个子层采用按比例抽样得到n_i。

一、给定置信区间为95% ，$d=5000$

$$\begin{aligned} S_w^2 &= \sum_{i=1}^{4} w_i S_i^2 = 2.5487*10^8 \\ S_w &= 15965 \\ \bar{S} &= \sum_{i=1}^{4} w_i S_i = 15782 \\ \bar{S^2} &= 2.4908 \\ n_0 &= \frac{Z_{\alpha/2}^2 * S_w^2}{d^2} = 39.6178 \\ n &= \frac{n_0}{1+\frac{n_0}{N}} = 28.7435 \approx 29 \end{aligned}$$

由

$$n_i = n \cdot w_i$$

得

$$n_1 = 2,\ n_2 = 7,\ n_3 = n_4 = 10\ (\sum_{i=1}^{4} n_i = 29)$$

用计算机模拟简单随机抽样，编写函数**fun1(x,y)**–从**x**个样本中选出**y**个。对每个子层随机抽样。抽样结果见表**13**。

通过对抽出的样本数据计算得到各层估计统计量：

$$\begin{aligned} S_1^2 &= 1.2247*10^9 \\ S_2^2 &= 3.6312*10^8 \\ S_3^2 &= 1.9727*10^8 \\ S_4^2 &= 2.4705*10^8 \\ \bar{y_1} &= 2.0422*10^5 \end{aligned}$$

$$\begin{aligned} \bar{y_2} &= 1.3052 * 10^5 \\ \bar{y_3} &= 81197 \\ \bar{y_4} &= 45032 \end{aligned}$$

由

$$\begin{aligned} \bar{y} &= \sum_{i=1}^{4} w_i \bar{y_i} \\ S^2 &= \sum_{i=1}^{n} w_i S_i^2 = 3.388 * 10^8 \end{aligned}$$

得总体均值为：93475

由

$$Y = 108 * \bar{y}$$

求得总排放量的估计值为：12111658，总方差的估计值为：$3.388e+008$，与真实值11597565 的残差为-514093，估计值比真实值大。

分析：用以上方法必须抽出34个城市，超出当前调查的能力，所以要进一步改进使得所抽样本更小，尽量控制在15个城市以内。基于以上原因我们对抽样方法进行优化，适当增大误差范围。

定置信区间给定置信区间95% $d = 10000$

以相同方法按比例抽样得到各层内的样本结果如下：$n1 = 1$ ，$n2 = 2$ ，$n3 = 3$ ，$n4 = 3$，具体样本见表**12**

得到各层估计统计量：

$$\begin{aligned} \bar{y_1} &= 232282 \\ \bar{y_2} &= 141510 \\ \bar{y_3} &= 76700 \\ \bar{y_4} &= 32672 \end{aligned}$$

$$\begin{aligned} S_1^2 &= 0 \\ S_2^2 &= 3.3834 * 10^7 \\ S_3^2 &= 9.2554 * 10^7 \\ S_4^2 &= 2.9064 * 10^8 \end{aligned}$$

估计的总体均值为：89176。
总排放量的估计值为：11647430。
总方差的估计值为：$1.4094 * 10^8$。

总排放量的估计与真实值11597565 的残差为-49865，估计值比真实值大。

分析：改变d后需要抽出的样本只有**14**个，估计的值也比之前的方式更接近真实值，估计的效果很好。

§5 多元线性回归分析

参考有关资料，我们假定二氧化硫排放量可能与燃料煤消费量（＋）、燃料油消费量（＋）、工业二氧化硫去除量（-）、废气治理设施数、费用（-）、脱硫数（-）有回归关系，进行拟合,拟合结果如下：

$$so2 = 15148.6621 + 53.4202 * rm + 0.2008 * qc - 265.5428 * ry + 41.0048 * ss + 0.1599 * tl$$

用**MATLAB**做出拟合点图与原数据点比较图，如图**7**。

为了分析残差**r**我们做出了它对正态分布的**QQ**图(图**8**),可以看出,除了尾部的异常点之外，残差近似服从正态分布。这说明二氧化硫的排放量与这些因素之间有一定的线性关系。由于时间有限，我们没有过多的分析这些量之间精准的关系，猜想如果减少或是增加一些因素，对原始数据进行数据变换，可能会得到更好的回归关系。这些还有待进一步研究。

§6 附录

表 7 中型企业城市分层

层1		层2	层3		
淄博	213895	福州	101500	哈尔滨	60751
宁波	211263	阳泉	100105	韶关	59845
石家庄	209879	烟台	99049	常德	58123
邯郸	198324	南通	97498	湖州	57187
赤峰	191173	西安	97240	日照	56331
贵阳	187689	九江	93099	桂林	54908
包头	179470	枣庄	92107	吉林	53102
徐州	171221	大连	89447	秦皇岛	52802
三门峡	165108	扬州	89442	齐齐哈尔	51836
宜宾	156722	株洲	88368	泉州	50977
郑州	154297	宝鸡	87474	湛江	49296
遵义	137401	绵阳	87464	长沙	48340
平顶山	133929	镇江	86405	牡丹江	44823
本溪	133061	常州	83847	深圳	42380
石嘴山	130225	柳州	83793	北海	42216
咸阳	130157	抚顺	83063	宜昌	40512
太原	126495	临汾	79149	泸州	40010
济宁	125091	湘潭	79030	连云港	39507
广州	124765	济南	77833	荆州	36281
杭州	121189	泰安	76752	自贡	35291
成都	121143	曲靖	76269	南昌	32795
乌鲁木齐	119762	兰州	73657	珠海	29509
长治	118730	西宁	72797	合肥	27211
青岛	117978	绍兴	71156	汕头	25736
大同	115785	温州	70834	开封	22052
焦作	115400	保定	70186	德阳	21706
攀枝花	113652	沈阳	70005	银川	16289
潍坊	109961	岳阳	69993	铜川	13184
安阳	107097	锦州	66064	玉溪	11754
昆明	103273	南宁	65963	南充	9168

表 8 方案一抽样结果

分类	样本		样本均值	总体均值	总体-样本
大型企业	上海	374327	$2.9978 * 10^5$	$2.9978 * 10^5$	0
	唐山	296038			
	苏州	228962			
	金昌	106401	106401	125160	18759
	北京	93755	$5.5055 * 10^4$	$5.0326 * 10^4$	$-4.7298 * 10^3$
	芜湖	46867			
	克拉玛依	24544			
中型企业	淄博	213895	$1.6639 * 10^5$	$1.4480 * 10^5$	$-2.1586 * 10^4$
	宁波	211263			
	石家庄	209879			
	贵阳	187689			
	遵义	137401			
	济宁	125091			
	广州	124765			
	成都	121143			
	阳泉	100105	$8.3066 * 10^4$	$8.2653 * 10^4$	-412.6333
	绵阳	87464			
	济南	77833			
	曲靖	76269			
	兰州	73657			
	常德	58123	$4.4379 * 10^4$	$3.9464 * 10^4$	$-4.9145 * 10^3$
	日照	56331			
	湛江	49296			
	深圳	42380			
	北海	42216			
	南昌	32795			
	珠海	29509			
小型企业	张家界	8415	8415	8415	0

表 9 中型企业城市分层

抽样方法	总体个数	均值	方差	样本个数
离群值	4	394580	$4.6568 * 10^10$	4
简单抽样	30	144800	$1.2023 * 10^9$	8
	30	82653	$1.1559 * 10^8$	5
	30	39464	$2.4023 * 10^8$	7
合计	96	709032	$5.9891 * 10^10$	24

表 10 二氧化硫排放量基本统计量表

Mean	102633.3
Median	86405.00
Maximum	711537.0
Minimum	175.0000
Std. Dev.	89440.03
Skewness	3.401995
Kurtosis	21.28088

表 11 二氧化硫排放量显著大的城市

重庆	711537
上海	374327
渭南	338956
唐山	296038
洛阳	295548

表 12 改变d后的各层样本

层一	苏州	232282
层二	遵义	137401
	南京	145627
层三	兰州	73657
	宝鸡	87474
	厦门	68969
层四	牡丹江	44823
	铜川	13184
	泸州	40010

表 13 简单随机抽样结果

层一	包头	179470
	苏州	228962
层二	太原	126495
	焦作	115400
	乌鲁木齐	119762
	石嘴山	130225
	徐州	171221
	青岛	117978
	武汉	132572
层三	绍兴	71156
	厦门	68969
	曲靖	76269
	阳泉	100105
	扬州	89442
	岳阳	69993
	沈阳	70005
	大连	89447
	保定	70186
	金昌	106401
层四	哈尔滨	60751
	银川	16289
	吉林	59845
	马鞍山	53102
	韶关	52727
	长春	51969
	齐齐哈尔	51836
	湛江	49296
	南昌	32795
	德阳	21706