Join列不同值个数估计

CBase:居兴

2017.11.08

不同值个数

- 不同值个数是指关系表的某一列上**不重复值**的个数
- 不同值个数在cbase里的意义在于:

  1.估计两表join结果的行数(以inner join为例，$nrows = \frac{T1.rows}{T1.diff\_num} * \frac{T2.rows}{T2.diff\_num} * \min(T1.diff\_num, T2.diff\_num))$

  2.确定两表join算子（semi join(left_table.diff_num<100)，hash join）

估计假设

- Join列不同值个数均匀分布，每个不同值重复次数接近
- 不同列的相关性为0，完全独立

不同值个数估计

- 估计公式：$diffNum = T.diffNum * \left(1 - (1 - sel1)^{\frac{T.tuples}{T.dfffNum}}\right) * sel2 * \frac{T.joinedRows}{T.rows}$
- 符号说明：

diffNum是估计的join列不同值个数

T.diffNum是由统计信息得到表T的join列不同值个数

T.tuples是统计信息给出的表T的元组数

T.rows是由where条件估计的选择率sel* T.tuples得到估计的行数

T.joinedRows是T表以及参与join之后中间结果集中来源于T表的行数估计，反应历史信息

sel1是where条件中带join列条件的选择率[0,1]

sel2是where条件中不含join列的条件的选择率[0,1]

公式分析

- $\left(1-(1-sel1)^{\frac{T.tuples}{T.dfffNum}}\right)$：该部分在公式中的含义是join列上的where条件对join列不同值个数的影响。最简单的情况下，join列的不同值个数等于元组数，即该列所有值都是不同值，此时该部分可以化简为sel1，即根据where条件的选择率等比例的选择了相应不同值个数。假设join列的不同值个数只有元组数的一半，即不同值重复率为2，每个不同值出现2次。该部分公式可以化简为sel1*（2-sel1）且大于sel1,即当不同值个数存在重复时，该join列的选择率对不同值个数的选择能力减弱，选择的个数比根据选择率按比例选择出来的不同值个数更多。所以，该部分公式随着$\frac{T.tuples}{T.dfffNum}$越大，$(1-sel1)^{\frac{T.tuples}{T.dfffNum}}$越小，$1-(1-sel1)^{\frac{T.tuples}{T.dfffNum}}$越大越接近1。在最极端情况，不同值只有1个，那么该部分约等于1，join列的where条件的选择率对不同值个数没有任何影响。

公式分析-续

- $T.diffNum$：该项是估计join列不同值个数的基数。
- $sel2$: 该项是评估非join列对join列不同值个数的影响，当前公式认为非join列的选择率对join列的不同值个数是等比例选择的。
- $\frac{T.joinedRows}{T.rows}$：该项是反映表T经过的历史和其他表进行join后对不同值个数的影响。初始时，T.joinedRows等于T.rows，随着表T的join次数越多，joinedRows越小，同样认为$\frac{T.joinedRows}{T.rows}$是对join列的不同值个数是等比例选择的。

案例分析

- 场景1：估小

表T元组数T.tuples=50 2061行，jion列不同值个数T.diffNum=4302，sel1=1，sel2=0.0019（实际0.0038），T.rows=950（1907），$\frac{T.joinedRows}{T.rows}$=1。

估计join列不同值 diffNum=8，决策使用semi join

实际查出来的结果join列的不同值为1900;

原因：违背了假设1，不同值重复率接近，实际join列49 4276行为NULL。其余不同值重复值在1-20间，where条件选择结果集含有8 row join列为NULL。

注：为null值join列inner join会忽略，left join会保留坐标行，右表置null，即使两表join列都为null。

案例分析-续

- 场景2：估大

表T元组数T.tuples=891行，join列的不同值个数T.diffNum=281，sel1=1，

sel2=0.594（实际0.582），T.row=529（519），$\frac{T.joinedRows}{T.rows}$=1。

估计join列不同值 diffNum=167，决策使用hash join

实际查出来的结果join列的不同值为1，为NULL。

原因：违背了假设2，非join列对join列的相关性为0。Join列为null值个数为610，其余279个不同值重复率为1，一个不同值重复率为2，也比较严重地违背了假设1。

公式改进

- 考虑高频值，将高频值与低频值分开，缓解假设1问题，分散sel2的选择能力
- 考虑sel2对join列的不同值的选择是否等比例选择，类似sel1的选择能力
- 考虑$\frac{T.joinedRows}{T.rows}$的选择能力与T.joinedRows的更新规则
- 其他，不同列的相关性?

问题原因在于，
整体的数据分布假设被违背了，
那么我们可以分类讨论，
利用高频值信息将数据进行细化，
各个分类就仍然遵守我们的假设了！

新估计公式

- 子部分: $\mathrm{d}iffNum = T.diffNum * \left(1 - (1 - sel1)^{\frac{T.tuples}{T.dfffNum}}\right) * \left(1 - (1 - sel2)^{\frac{T.tuples}{T.diffNum}}\right) * \frac{T.joinedRows}{T.rows}$

- 完全公式1：

$$diffNum = \left(\sum_{i=1}^{highNum}(1-(1-sel1)^{P_i})*(1-(1-sel2)^{P_i})*\left(1-\left(1-\frac{T.joinedRows}{T.rows}\right)^{P_i}\right)*\sum_{i=1}^{highNum}P_i\right)+(T.diffNum-highNum)*(1-(1-sel1)^{Q})*(1-(1-sel2)^{Q})*\left(1-\left(1-\frac{T.joinedRows}{T.rows}\right)^{Q}\right)*(1-\sum_{i=1}^{highNum}P_i)$$

- 公式说明：

highNum:高频值个数[0,10]

Pi:高频值对应个数；Q：低频值平均频率对应的个数

缺陷：当选择率接近1，且高频值占的比例比较高时，严重估小。

- 完全公式2：

$$\begin{aligned}
diffNum &= \left(\sum_{i=1}^{highNum}\left(1-\left(1-sel1*\sum_{i=1}^{highNum}P_i\right)^{P_i}\right)*\left(1-\left(1-sel2*\sum_{i=1}^{highNum}P_i\right)^{P_i}\right)\right.\\
&\left.*\left(1-\left(1-\frac{T.joinedRows}{T.rows}*\sum_{i=1}^{highNum}P_i\right)^{P_i}\right)\right)+(T.diffNum-highNum)*\left(1-\left(1-sel1*(1-\sum_{i=1}^{highNum}P_i)\right)^{Q}\right)\\
&*\left(1-\left(1-sel2*(1-\sum_{i=1}^{highNum}P_i)\right)^{Q}\right)*\left(1-\left(1-\frac{T.joinedRows}{T.rows}*(1-\sum_{i=1}^{highNum}P_i)\right)^{Q}\right)
\end{aligned}$$

- 公式说明：

highNum:高频值个数[0,10]

Pi:高频值对应个数；Q：低频值平均频率对应的个数

案例分析

- 场景1：估小

表T元组数T.tuples=50 2061行，jion列不同值个数T.diffNum=4302，sel1=1，

sel2=0.0019（实际0.0038），T.rows=950（1907），$\frac{T.joinedRows}{T.rows}$=1。

估计join列不同值 diffNum=1.25+0，决策使用semi join

实际查出来的结果join列的不同值为1900；

公式2的还有一个不算严重的缺陷在于当高频值占据比例很小的时候，即重复率和低频值接近时，对高频值的不同值个数估计会很小，为0。

采用公式2，计算仍然估小了，而公式1估计结果更大一点没有决策到semi join。

所以还是假设2被违背。

案例分析-续

- 场景2：估大

表T元组数T.tuples=891行，join列的不同值个数T.diffNum=281，

sel1=1，sel2=0.594（实际0.582），T.row=529（519），

$\frac{T.joinedRows}{T.rows}$=1。

估计join列不同值 diffNum=3.1+4.5，决策使用semi join

实际查出来的结果join列的不同值为1，为NULL。

公式1（估计值89）和公式2都能决策成semi join

模拟分析

根据公式考虑各种变量下的估计公式的表现:
考虑参数：join列选择率sel1，非join列选择率sel2，join列的高、低频比例（用高频重复次数控制），join列高、低频不同值个数（用低频重复次数控制）

模拟环境:
基本设定10000，高频值10个，随机采样20次，模拟选择率的估计结果。
公式1：分开考虑sel1和sel2，即完全公式2。
公式2：公式1基础上将sel1*sel2当做一个整体。
公式3：在公式1基础上乘系数k = 1.4142。
公式4：将高频低频比例在估计公式之后再乘，即完全公式1。

因素：选择率

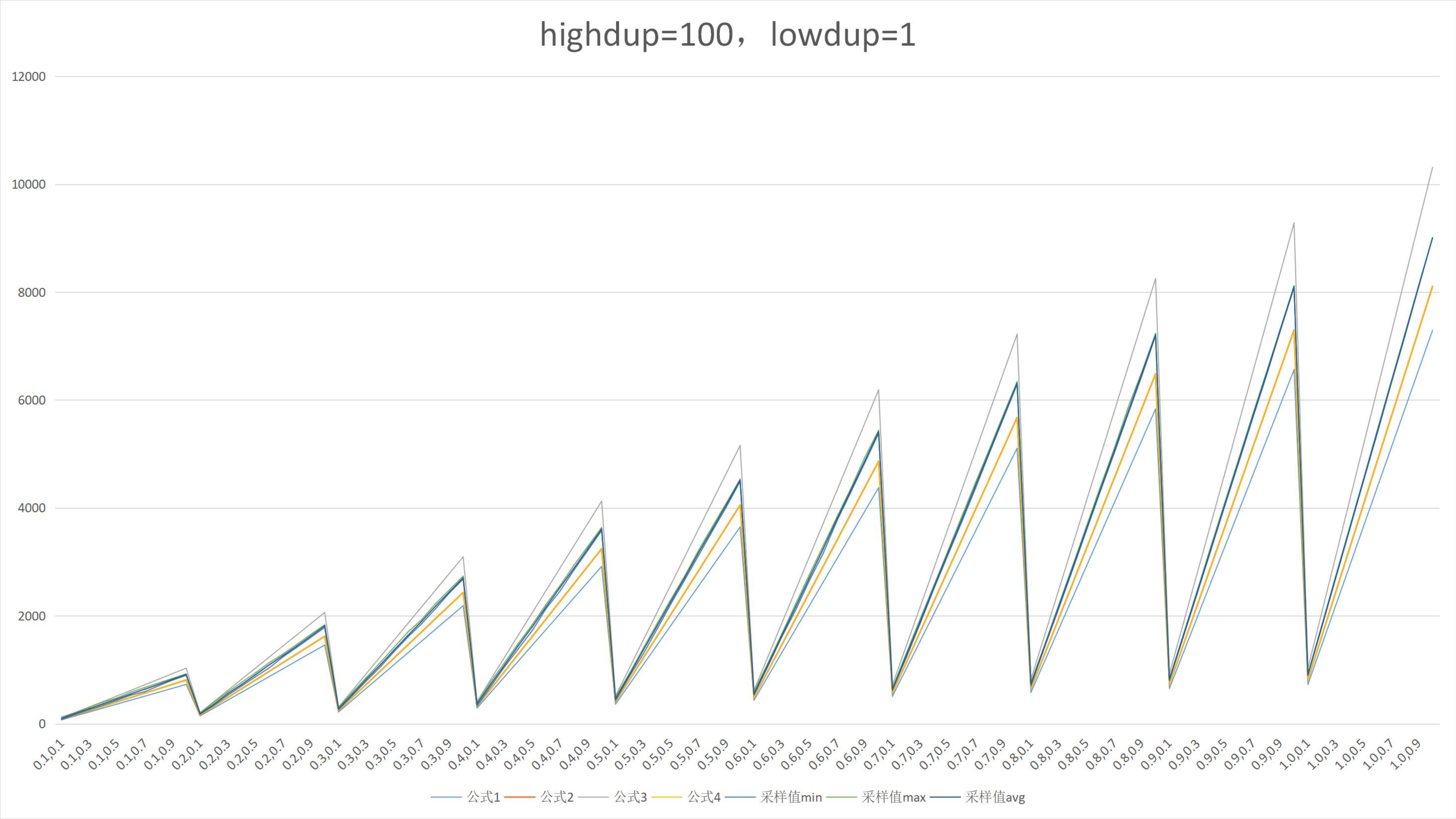
高频值重复次数100，所占总行数比例10%，低频值重复次数1，所占总行数的比例为90%。

结果:

从图中可以看出各个公式的趋向性与采样结果一直，但是存在一定的误差，其中公式2和公式4的估计结果十分相近，并且贴近采样值。

公式1在选择率比较大的时候出现最明显的估小，并且公式1总是估小。

而公式4乘以系数后，会明显估大。简单调整系数，恐与低频与高频值比例有关。

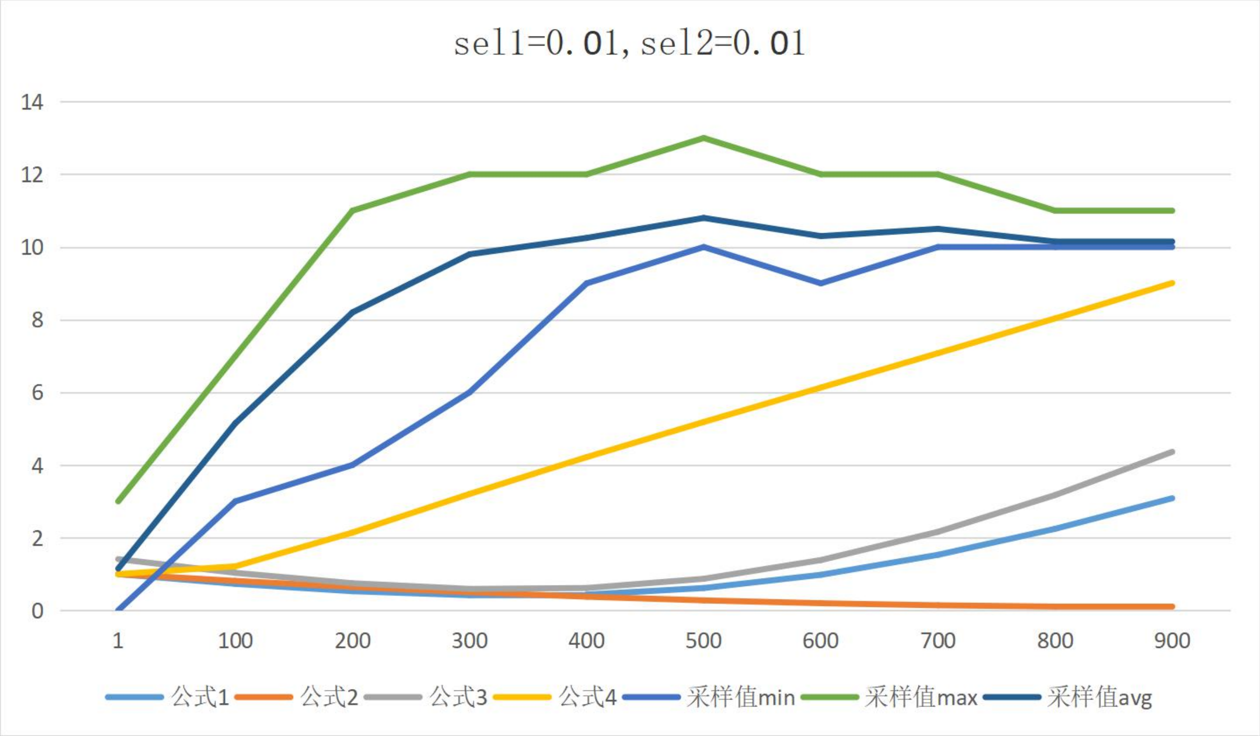


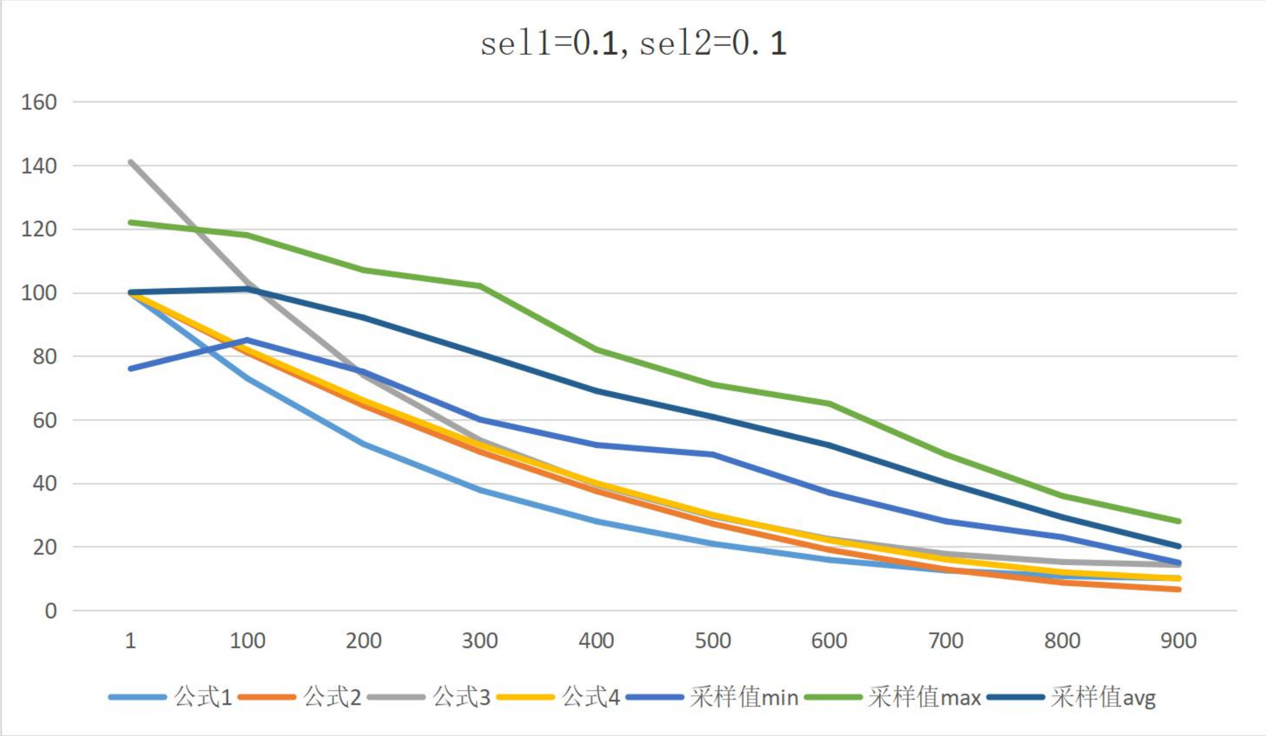
highdup=100，lowdup=1
12000
10000
8000
6000
4000
2000
0
0.1,0.1
0.1,0.3
0.1,0.5
0.1,0.7
0.1,0.9
0.2,0.1
0.2,0.3
0.2,0.5
0.2,0.7
0.2,0.9
0.3,0.1
0.3,0.3
0.3,0.5
0.3,0.7
0.3,0.9
0.4,0.1
0.4,0.3
0.4,0.5
0.4,0.7
0.4,0.9
0.5,0.1
0.5,0.3
0.5,0.5
0.5,0.7
0.5,0.9
0.6,0.1
0.6,0.3
0.6,0.5
0.6,0.7
0.6,0.9
0.7,0.1
0.7,0.3
0.7,0.5
0.7,0.7
0.7,0.9
0.8,0.1
0.8,0.3
0.8,0.5
0.8,0.7
0.8,0.9
0.9,0.1
0.9,0.3
0.9,0.5
0.9,0.7
0.9,0.9
1.0,0.1
1.0,0.3
1.0,0.5
1.0,0.7
1.0,0.9
公式1
公式2
公式3
公式4
采样值min
采样值max
采样值avg

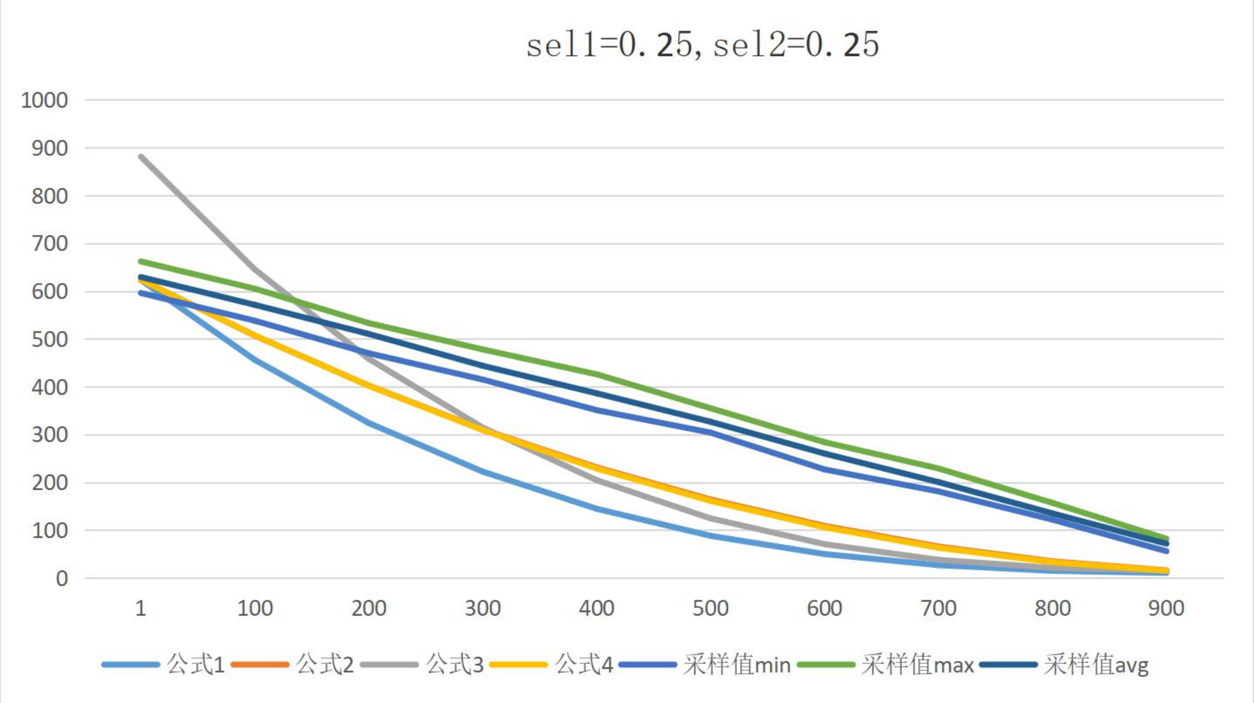
因素：高频比例

控制sel1，sel2,不同highdup值的模拟结果:

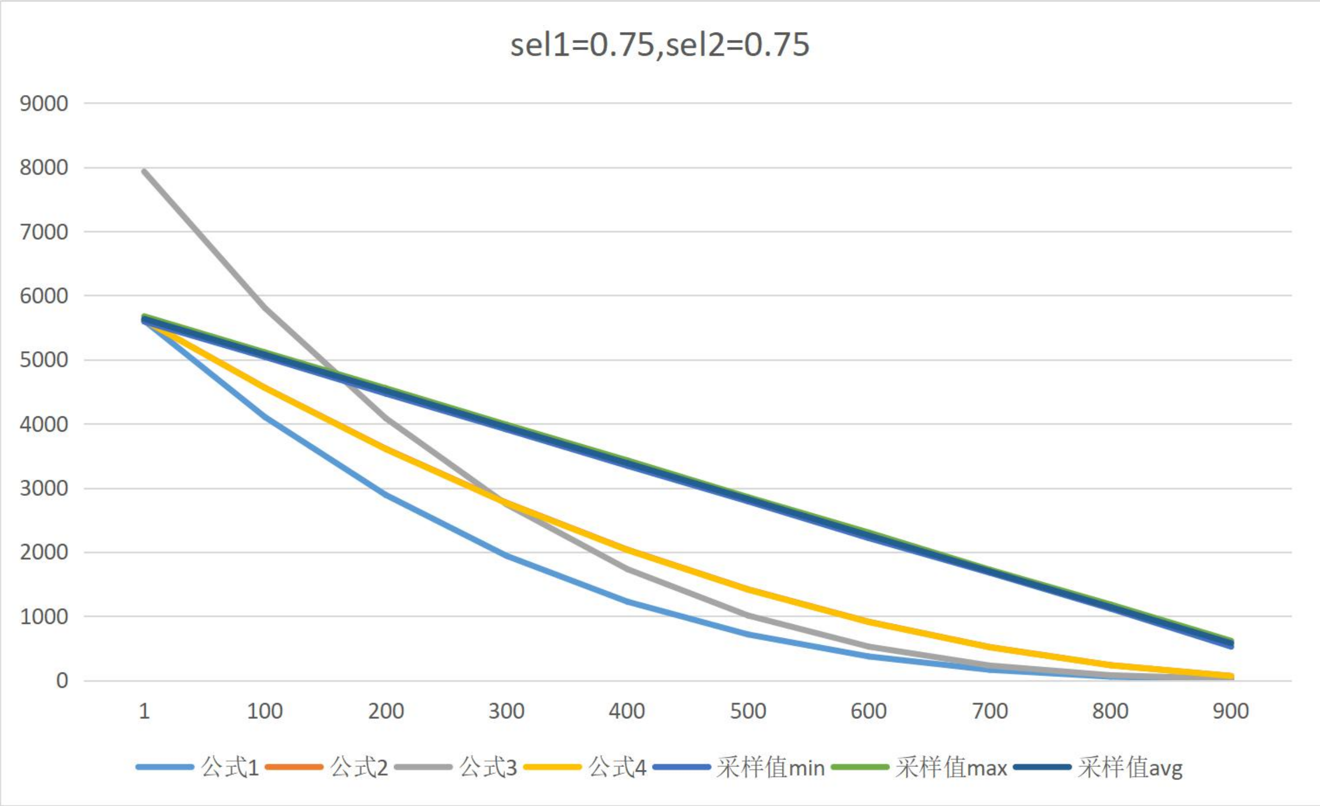
可以观察到，
我们的估计公式随着高频值所占比例加大，下降程度比采样的结果更快，即明显估小，估小对于决策是否走semi join会有严重的影响。
而且，随着选择率增大，即使不同值个数随着高频比例成下降趋势，但是对比低选择率的不同值个数个数明显增多。
或许一个更简单的公式能更好估计不同值个数，diff_number*sel*比例。
高频值重复次数为1,100,200...900时，不同值个数：10000，9010，8010,7010,...,1010

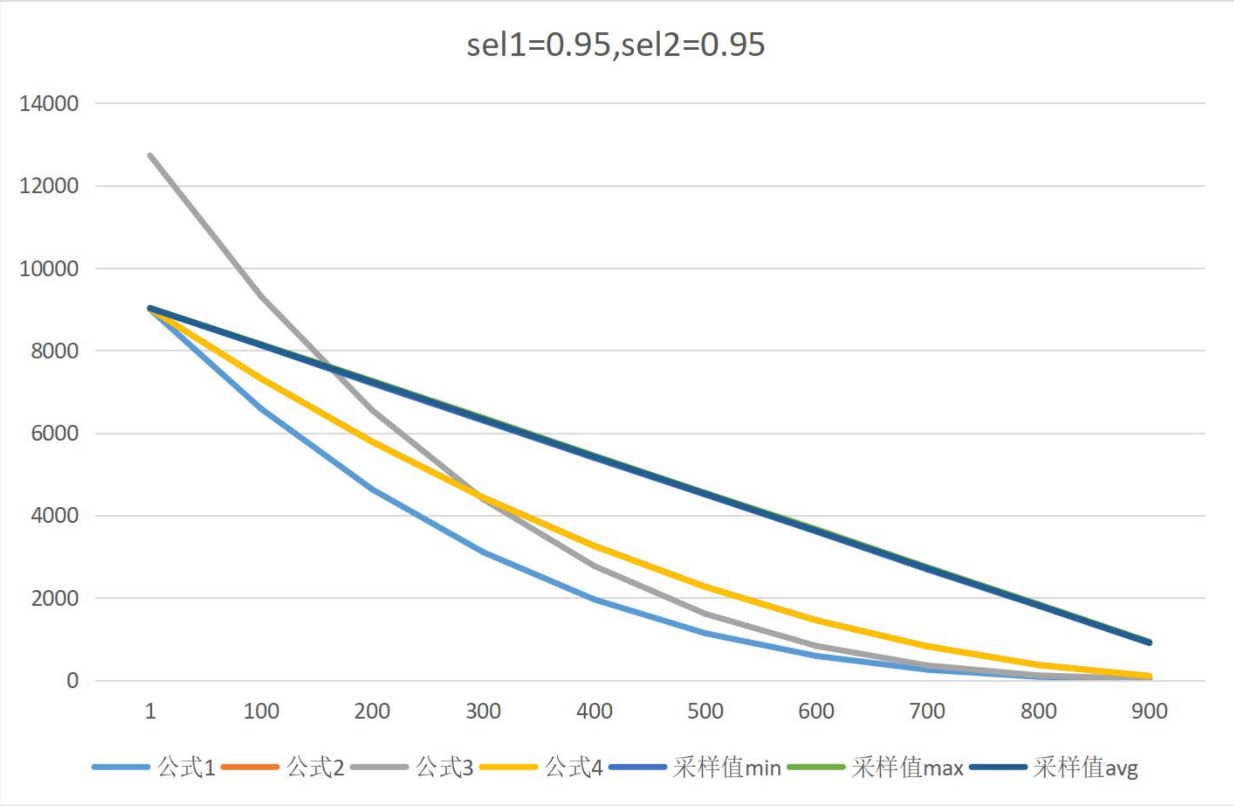

sel1=0.01, sel2=0.01
14
12
10
8
6
4
2
0
1
100
200
300
400
500
600
700
800
900
公式1
公式2
公式3
公式4
采样值min
采样值max
采样值avg


sel1=0.1, sel2=0.1
160
140
120
100
80
60
40
20
0
1
100
200
300
400
500
600
700
800
900
公式1
公式2
公式3
公式4
采样值min
采样值max
采样值avg


sel1=0.25, sel2=0.25
1000
900
800
700
600
500
400
300
200
100
0
1
100
200
300
400
500
600
700
800
900
公式1
公式2
公式3
公式4
采样值min
采样值max
采样值avg


sel1=0.5,sel2=0.5
4000
3500
3000
2500
2000
1500
1000
500
0
1
100
200
300
400
500
600
700
800
900
公式1
公式2
公式3
公式4
采样值min
采样值max
采样值avg


sel1=0.75,sel2=0.75
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
1
100
200
300
400
500
600
700
800
900
公式1
公式2
公式3
公式4
采样值min
采样值max
采样值avg

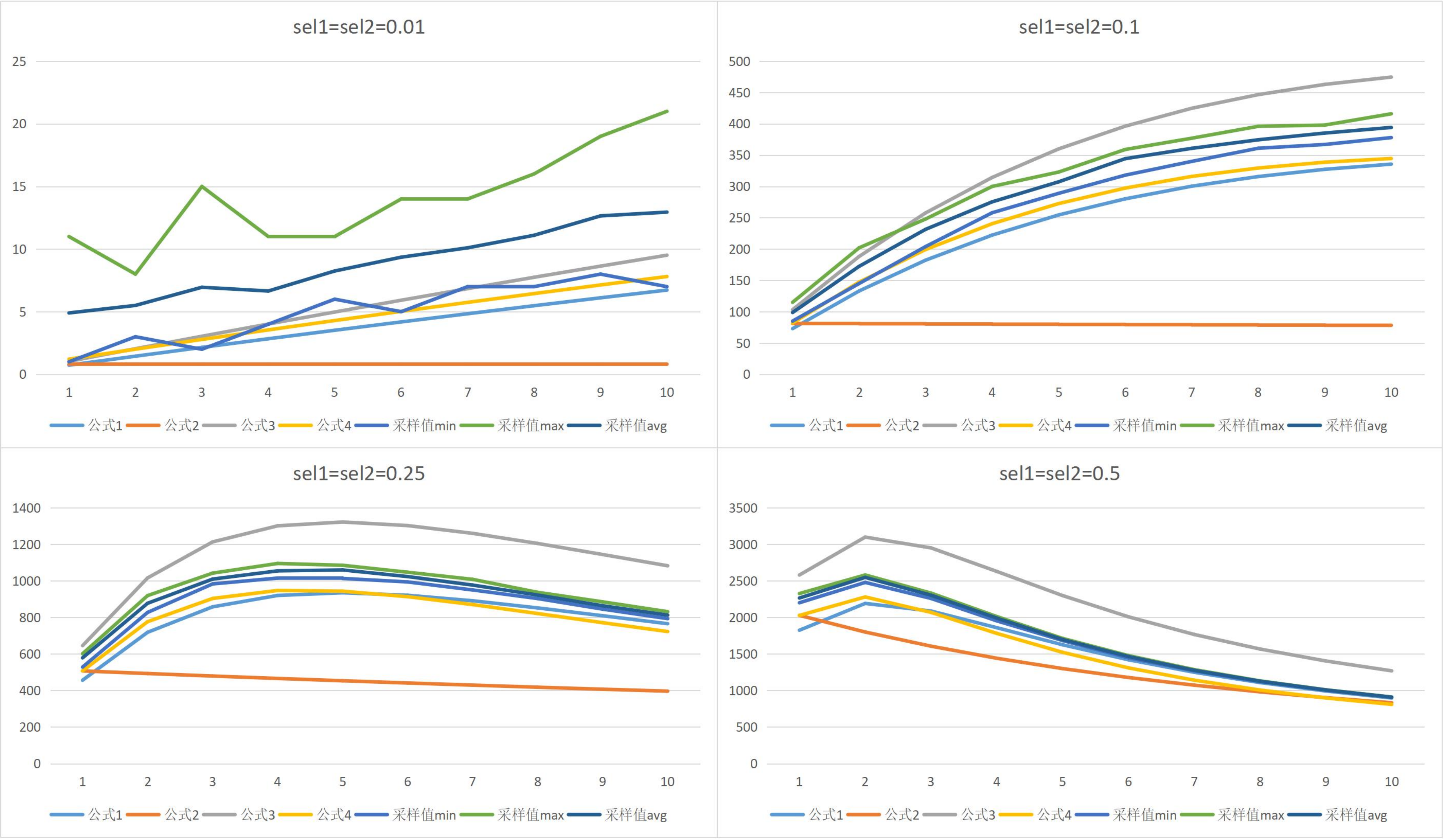

sel1=0.95,sel2=0.95
14000
12000
10000
8000
6000
4000
2000
0
1
100
200
300
400
500
600
700
800
900
公式1
公式2
公式3
公式4
采样值min
采样值max
采样值avg

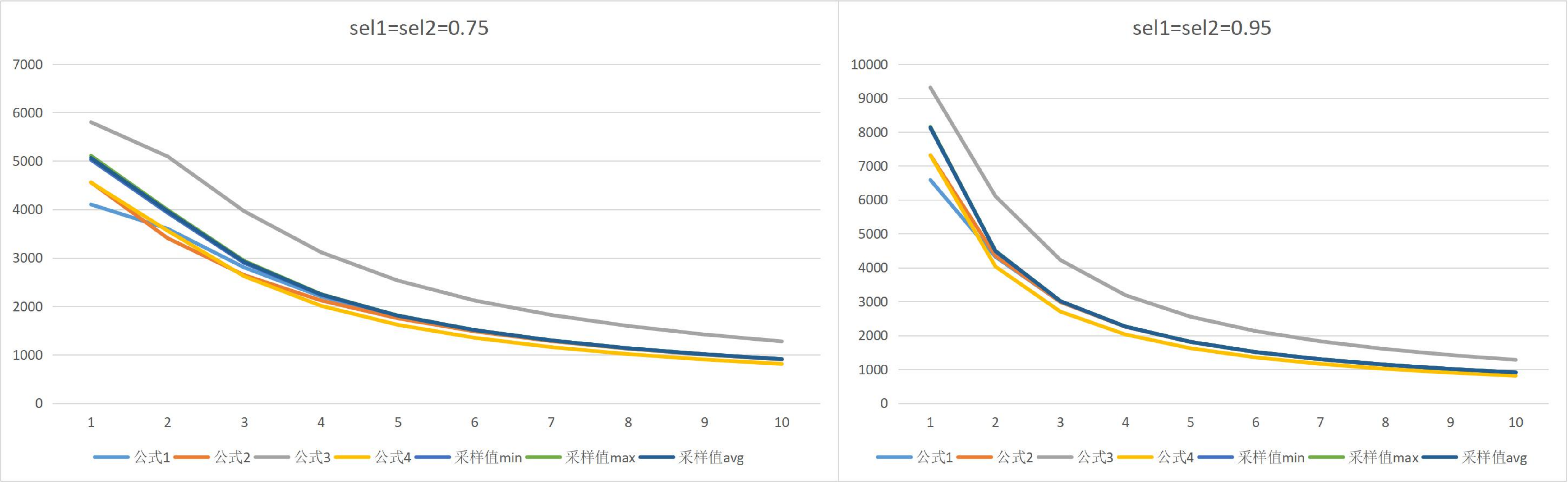
因素：不同值个数

固定高频值的重复率为100，占比10%。

低频值重复个数为1,2,...,10时，不同值个数：
9010,4510,3010,2260,1810,1510,1296,1135,1010,910

结果，在这个情况下，公式1在后期随着低频重复次数增大，估计的更准，而公式4在低频重复次数前期（小于5）估计的更准，误差越来越接近高频值所占比例。
总体上看，公式4估计虽然在后期估计的稍弱与公式1，但是不是十分明显。


sel1=sel2=0.01
0
5
10
15
20
25
1
2
3
4
5
6
7
8
9
10
公式1
公式2
公式3
公式4
采样值min
采样值max
采样值avg
sel1=sel2=0.1
0
50
100
150
200
250
300
350
400
450
500
1
2
3
4
5
6
7
8
9
10
公式1
公式2
公式3
公式4
采样值min
采样值max
采样值avg
sel1=sel2=0.25
0
200
400
600
800
1000
1200
1400
1
2
3
4
5
6
7
8
9
10
公式1
公式2
公式3
公式4
采样值min
采样值max
采样值avg
sel1=sel2=0.5
0
500
1000
1500
2000
2500
3000
3500
1
2
3
4
5
6
7
8
9
10
公式1
公式2
公式3
公式4
采样值min
采样值max
采样值avg



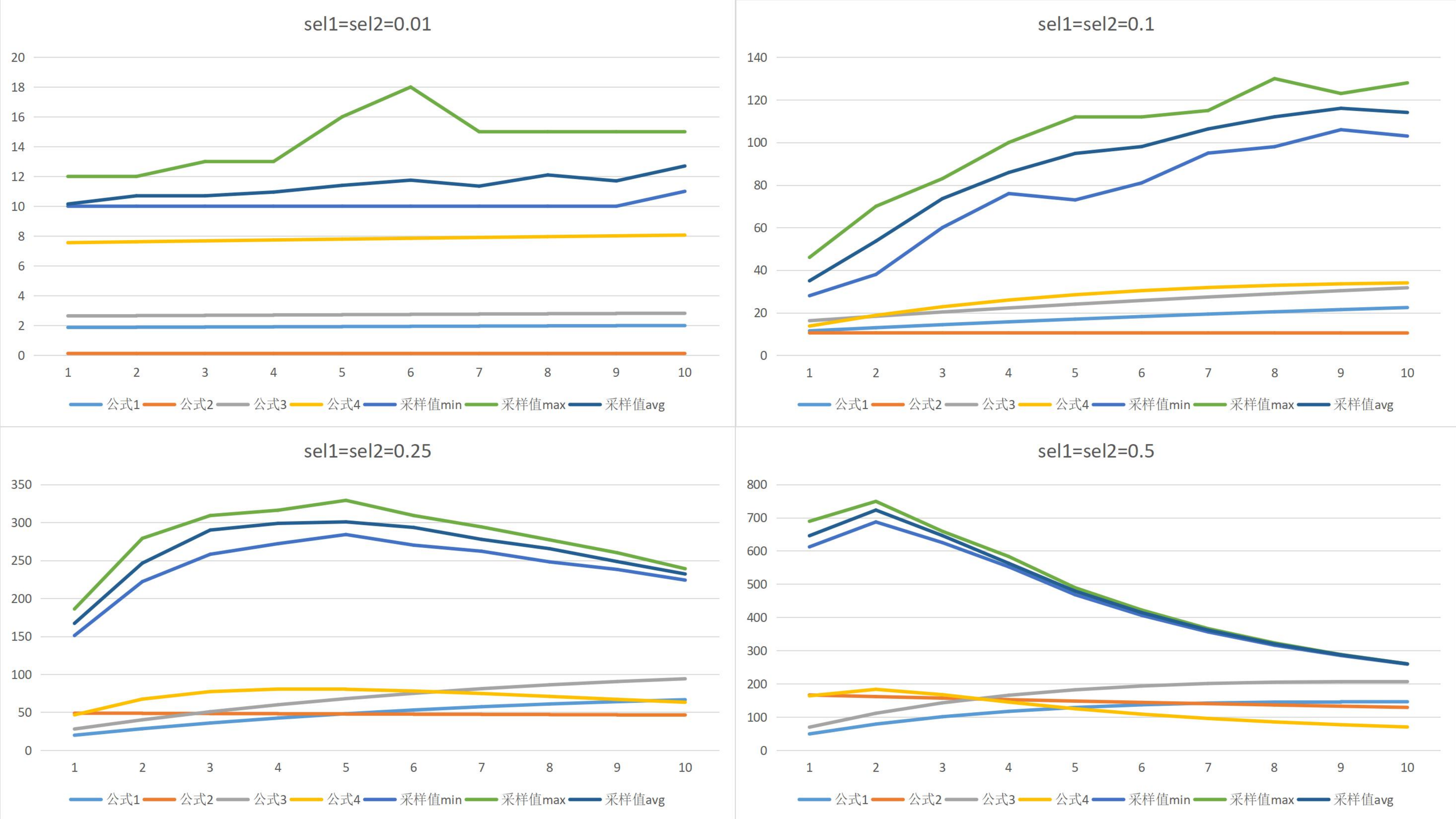
sel1=sel2=0.75
7000
6000
5000
4000
3000
2000
1000
0
1
2
3
4
5
6
7
8
9
10
公式1
公式2
公式3
公式4
采样值min
采样值max
采样值avg
sel1=sel2=0.95
10000
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
1
2
3
4
5
6
7
8
9
10
公式1
公式2
公式3
公式4
采样值min
采样值max
采样值avg

因素：不同值个数-续

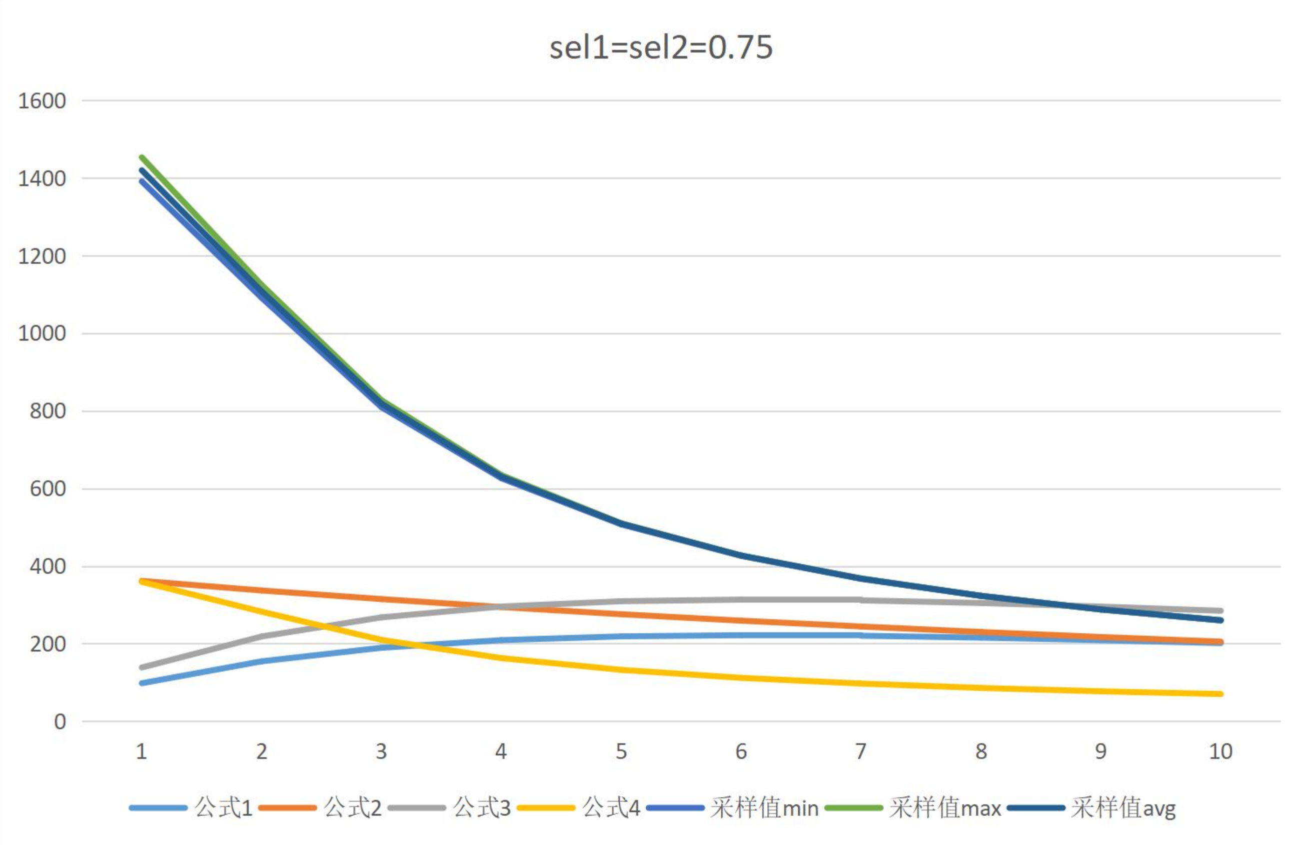
固定高频值的重复率为750，占比75%。

低频值重复个数为1,2,...,10时，不同值个数：2510,1260,844,635,510，427,368,323,288,260,260

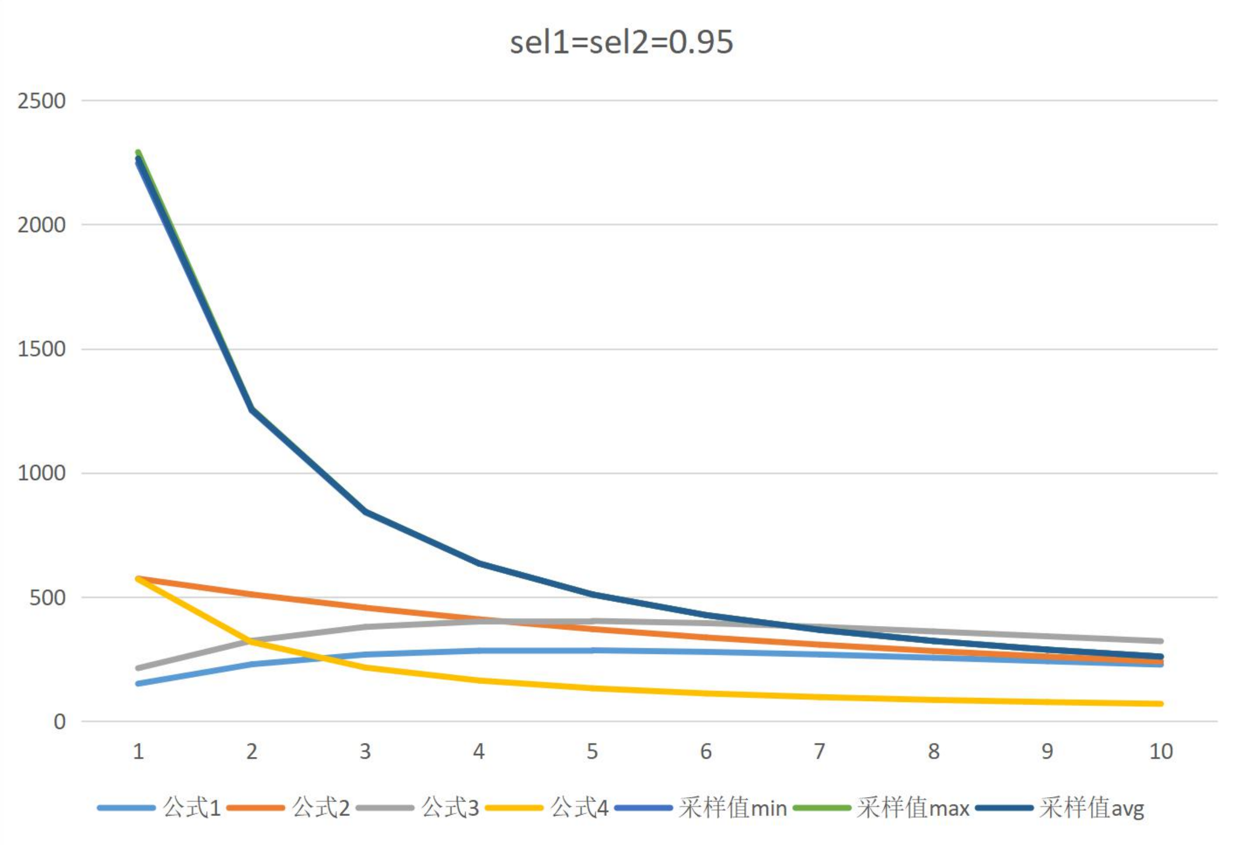
结果，可以看出随着选择率的增加，各个公式估小的程度更加明显。



sel1=sel2=0.01
公式1 公式2 公式3 公式4 采样值min 采样值max 采样值avg
sel1=sel2=0.1
公式1 公式2 公式3 公式4 采样值min 采样值max 采样值avg
sel1=sel2=0.25
公式1 公式2 公式3 公式4 采样值min 采样值max 采样值avg
sel1=sel2=0.5
公式1 公式2 公式3 公式4 采样值min 采样值max 采样值avg



sel1=sel2=0.75
1600
1400
1200
1000
800
600
400
200
0
1 2 3 4 5 6 7 8 9 10
公式1 公式2 公式3 公式4 采样值min 采样值max 采样值avg



sel1=sel2=0.95
2500
2000
1500
1000
500
0
1 2 3 4 5 6 7 8 9 10
公式1 公式2 公式3 公式4 采样值min 采样值max 采样值avg

总结

综合比较选择率，高频值比例和低频值重复次数这几个变量情况下的各个公式的表现，可以看出公式4在选择率，高频值比例的变化情况下最优，在低频值重复次数较小情况下也表现较优。但是公式4的不足之处，在于处理高频比例很大，而选择率较高情况下，仍然会明显估小

所以，得出结论使用公式4，能在多数情况下对不同值个数的估计更准。而对于处理高频部分所占比例极大情况下，需要有新的子公式进行平衡。

END,THANKS