hyperloglog原理理解

举一个我们最熟悉的抛硬币例子，出现正反面的概率都是1/2，一直抛硬币直到出现正面，记录下投掷次数k，将这种抛硬币多次直到出现正面的过程记为一次伯努利过程，对于n次伯努利过程，我们会得到n个出现正面的投掷次数值$k_1,k_2.....k_n$，其中最大值记为k_{max}，那么可以得到下面结论：

1. n次伯努利过程的投掷次数都不大于k_{max}
2. n次伯努利过程，至少有一次投掷次数等于k_{max}

对于第一个结论，n次伯努利过程的抛掷次数都不大于k_{max}的概率用数学公式表示为：
$P_n(X \le k_{max}) = (1 - 1/2^{k_{max}})^n$

第二个结论至少有一次等于k_{max}的概率用数学公式表示为：
$P_n(X \ge k_{max}) = 1 - (1 - 1/2^{k_{max}-1})^n$

当$n \ll 2^{k_{max}}$时，$P_n(X \ge k_{max}) \approx 0$，即当$n$远小于$2^{k_{max}}$时，上述第一条结论不成立；
当$n \gg 2^{k_{max}}$时，$P_n(X \le k_{max}) \approx 0$，即当$n$远大于$2^{k_{max}}$时，上述第二条结论不成立。因此，我们似乎就可以用$2^{k_{max}}$的值来估计$n$的大小。

以上结论可以总结为：进行了n次进行抛硬币实验，每次分别记录下第一次抛到正面的抛掷次数k，那么可以用n次实验中最大的抛掷次数k_{max}来预估实验组数量n：$\hat{n} = 2^{k_{max}}$

此处的证明有误

当$n \ll 2^{k_{max}}$时，$P_n(X \ge k_{max}) \approx 0$，即当$n$远小于$2^{k_{max}}$时，上述第一条结论不成立；
当$n \gg 2^{k_{max}}$时，$P_n(X \le k_{max}) \approx 0$，即当$n$远大于$2^{k_{max}}$时，上述第二条结论不成立。

这里的上述证明写反了，应该是第一行证明第二条结论不成立，同时第二行证明第一条结论不成立

分析:
第一行，说 Pn 约等于零，对于第二条结论说
n 次伯努利过程，至少有一次投掷次数等于 Kmax
此时 Pn 的概率主要考虑等于 Kmax 的情况约等于 0，我觉得没有违背结论

第二行 ，说 Pn 约等于零，对于第一条结论来说
N 次伯努利过程的投掷次数都不大于 Kmax
此时的概率应该约为 1

再说为什么 2 的 k 次方能够估计 n 的量
此时主要考虑 n 与 2 的 k 次方是否处在同一个量级之中

比如说$a^{\frac{x}{y}}$这个公式

如果 xy 在同一个量级，那么整体不会归为 0 或者无穷大
而 x>>y,那么整体趋于无穷大
X<<y 那么整体趋于无穷小