1.在Bloom Filter 中，出现false positive 的概率近似为

$$(1-e^{-kn/m})^k$$

给定n,m，请问k应该取值多少比较合适？(25分)

2. 定义一个随机变量X，使得

$$Pr[|X-EX|\geq k\sigma]\geq \frac{1}{10k^2}$$

(25分)

3. 假设S由n个数组成，中位数是M。我们均匀随机采样(可以重复)k次，得到$X_1,X_2,...,X_k$。证明:当$k\geq O(1/\epsilon^2)$，$\tilde{M}=median(X_1,...,X_k)$是$M$的一个好的近似：

$$Pr[S\text{有}\frac{1}{2}-\epsilon\text{部分大于}\tilde{M},\text{且S有}\frac{1}{2}-\epsilon\text{部分小于}\tilde{M}]>0.9$$

(25分)

4. 密歇根州有540万选票，假设一张选票统计错误的概率是1%,那么统计错误的选票数超过1.4%的概率是多少？(25分)

1 解:

原问题可以等价为求

$$\arg min\Big(1-exp\{-\frac{kn}{m}\}\Big)^k=\arg min k\log\Big(1-exp\{-\frac{kn}{m}\}\Big)$$

令

$$f(k)=k\log\Big(1-exp\{-\frac{kn}{m}\}\Big)$$

对k求偏导，有

$$\frac{\partial f(k)}{\partial k}=\log\Big(1-exp\{-\frac{kn}{m}\}\Big)+\frac{kn}{m}\frac{exp\{-\frac{kn}{m}\}}{1-exp\{-\frac{kn}{m}\}}$$

记

$$t=1-exp\{-\frac{kn}{m}\}\in(0,1)$$

令$\frac{\partial f(k)}{\partial k}=0$，有

$$\log t-\log(1-t)\frac{1-t}{t}=0$$

解得当$t=\frac{1}{2}$时，原式取得最小值。此时$k=\ln 2\cdot\frac{m}{n}$。

2 解:

构造随机变量X，使得其满足

$$\begin{cases} f(x)=-f(x) \\ f(x)=\begin{cases} m, & |x|<\sigma \\ g_k(x), & k\sigma<|x|<(k+1)\sigma \end{cases} \end{cases}$$

其中$\sigma=\mathrm{Var}[X]$，该随机变量满足E[X]=0。令

$$\begin{aligned} Pr[|X-E[X]|\geq k\sigma] &= Pr[|X|\geq k\sigma] \\ &\geq Pr[(k+1)\sigma\geq|X|\geq k\sigma] \\ &= \int_{k\sigma}^{(k+1)\sigma} g_k(x)dx \\ &= \frac{1}{10k^2} \end{aligned}$$

又

$$\int_{-\infty}^{+\infty}f(x)dx=1 \iff m+\frac{1}{10}(1+\frac{1}{2^2}+\cdots)=\frac{1}{2} \iff m=\frac{1}{2}-\frac{\pi}{60}$$

因此m可以显式求出。又

$$\int_{-\infty}^{+\infty}x^2f(x)dx=\sigma^2$$

是一个关于σ的方程，存在这样的函数族$g_k(x),k=1,2,\ldots$满足上述条件，构造完毕。

3 解:

记事件A为S里小于$\tilde{M}$的数不足$(\frac{1}{2}-\epsilon)n$，即重复k次取数，取得S中较小的$(\frac{1}{2}-\epsilon)$个的次数大于$\frac{k}{2}$。原题等价为证明

$$\begin{aligned} Pr[A]\leq 0.05 &\iff \sum_{i=\frac{1}{2}k}^{\infty}C_k^i(\frac{1}{2}-\epsilon)^i(\frac{1}{2}+\epsilon)^{k-i}\leq 0.05 \\ &\iff F\Big(\mathcal{F}_{distribution}(\frac{1-2\epsilon}{1+2\epsilon}\cdot\frac{k+2}{k};k,k+2)\Big)\leq 0.05 \\ &\iff F\Big(\mathcal{F}_{distribution}(\frac{1-2\epsilon}{1+2\epsilon}\cdot(1+2\epsilon^2);\frac{1}{\epsilon^2},\frac{1}{\epsilon^2}+2)\Big)\leq 0.05 \end{aligned}$$

由算法求最优值，得到当$\epsilon=0.179480$时，原函数取最大值$0.028386\leq 0.05$。原问题得证。

4 解：由二项分布

$$f(k) = C_n^k p^k (1-p)^{n-k}$$

有

$$Pr[X \leq K] = F\Big(\mathcal{F}_{distribution}\Big(x = \frac{1-p}{p} \cdot \frac{k+1}{n-k}; d_1 = 2(n-k), d_2 = 2(k+1)\Big)\Big)$$

带入$t = 75600, n = 5400000, p = 0.01$，有

$$Pr[X \geq t] = 1 - Pr[X \leq t-1] \simeq 1.1 \times 10^{-16}$$

因此选票错误数大于1.4%的概率约为1.1×10^{-16}。