- (25分)在$Misra-Greis$算法中，用k表示计算器的个数，n是数据流的长度，n'表示算法结束时计数器中的数值之和。对于任意元素，算法返回的频率$\hat{f}_i$和真实的频率f_i的差距满足

$$f_i - \frac{n-n'}{k+1} \le \hat{f}_i \le f_i$$

- (25分)考虑$Misra-Greis$算法的并行情况，一台机器处理数据流σ_1，一台机器处理数据流σ_2，分别得到k个计数器。我们将两组计数器合并：对于相同元素，让对应的计数器数值相加；如果最后有超过k个计数器，那么按照数值从大到小排序，记第$k+1$个计数器的数值为c，我们只保留前k个计数器，并将数值全部减去c。证明：对于连在一起的数据流$\sigma_1 \cdot \sigma_2$:对于任意元素，算法返回的频率$\hat{f}_i$和真实的频率f_i的差距满足

$$f_i - \frac{n-n'}{k+1} \le \hat{f}_i \le f_i$$

- (25分) 假设有一个长度为m的数据流，每次读取一个元素，我们保存它的概率是p，忽视它的概率为$1-p$。令f_i'是保存元素i的次数，f_i是元素i在数据流中出现的次数。令$p = min(1, \frac{400}{m\epsilon^2})$。证明：

$$Pr[\forall i, |f_i'/p - f_i| \le (\epsilon/2)m] \ge 99/100$$

（注：本方法可以用计算majority的频率，改进空间复杂度，感兴趣可以思考一下）

1 解：

k表示计算器的个数，n是数据流的长度，n'表示算法结束时计数器中的数值之和，算法返回的频率$\hat{f}_i$和真实的频率f_i。

在算法中，总共减去了$n-n'$个计数，每次都将所有的k个计数器减1，并且舍弃了当前元素，则每次减去的计数为$k+1$，减少的次数即减少的元素个数为

$$\frac{n-n'}{k+1}$$

那么可以得到真实的计数范围为

$$\hat{f}_i \leq f_i \leq \hat{f}_i + \frac{n-n'}{k+1}$$

即有

$$f_i - \frac{n-n'}{k+1} \leq \hat{f}_i \leq f_i$$

2 解：

设连在一起的数据流$\sigma_1 \cdot \sigma_2$的总元素个数为n，最后保留的k个计数器的总和为n'，数据流σ_1，σ_2分别的计数器数值为n_1'和n_2'。首先有

$$f_i - \frac{n-n_1'-n_2'}{k+1} - c \leq \hat{f}_i \leq f_i$$

由于只保留前k个计数器，并将数值全部减去c，那么至少有$k+1$个计数器减去了c，即有

$$n_1' + n_2' \geq n' + (k+1)c$$

由此可以得到

$$c \leq \frac{n_1'+n_2'-n'}{k+1}$$

那么

$$\begin{aligned}
\hat{f}_i &\geq f_i - \frac{n-n_1'-n_2'}{k+1} - c \\
&\geq f_i - \frac{n-n_1'-n_2'}{k+1} - \frac{n_1'+n_2'-n'}{k+1} \\
&= f_i - \frac{n-n'}{k+1}
\end{aligned}$$

则有$f_i - \frac{n-n'}{k+1} \leq \hat{f}_i \leq f_i$成立.

3 解：

$$Pr\Big(\big|\frac{f_i'}{p} - f_i\big| \leq \frac{\epsilon m}{2}\Big) = Pr\Big(\big|f_i' - pf_i\big| \leq \frac{\epsilon mp}{2}\Big) \tag{1}$$

由于数据流可以看做二项分布，那么有

$$E[f_i'] = pf_i$$

$$Var[f_i'] = p(1-p)f_i$$

由Chebyshev不等式可得

$$
\begin{aligned}
(1)\text{式} &\geq 1-\frac{p(1-p)f_i}{(\frac{\epsilon mp}{2})^2}\\
&=1-\frac{4(1-p)f_i}{\epsilon^2m^2p}\\
&=1-\frac{(1-p)f_i}{100m}\\
&\geq 1-\frac{1-p}{100}\\
&\geq 1-\frac{1}{100}=\frac{99}{100}
\end{aligned}
$$