物联网信息处理实验二

实验要求：

使用自己所擅长的程序语言编写水库采样算法(also known as reservoir sampling)。

1：以 stream.txt 文件为自己所写程序的输入，读取文件中数据，设为 $e_1,e_2,..,e_n,.....$(假设每秒到达一个数据)；

2：设定样本集合大小为 s（该参数为程序输入参数），要求在任意 t 大于等于 s 的时刻维持一个采样集合 S，要求对于已经看到过的元素 $e_1,e_2,...,e_t$ 中的每个元素都以相同的概率被选进集合 S。

3：计算采样集合在所有数据读取完毕后的均值，计算过程如下：

假设 $S=\{f_1,f_2,..,f_s\}$，那么均值等于$(f_1+f_2+...f_s)/s$;

4：精确计算整个数据流中元素的均值，计算过程为如下：

假设数据流 delta = $e_1,e_2,...,e_{10000}$(比如 stream.txt 共有 10000 个元素，实际上 stream.txt 的元素个数不是 10000，这里只是举例)，那么均值等于$(e_1+e_2+...e_{10000})/10000$;

实验考察要求：

1：分析并讲解自己所编写程序；

2：设样本集合 s=100，输出最终时刻（所有数据都处理完毕后）采样集合中数据的均值；

3：设样本集合 s=1000，输出最终时刻（所有数据都处理完毕后）采样集合中数据的均值；

4：设样本集合 s=10000，输出最终时刻（所有数据都处理完毕后）

采样集合中数据的均值；

5：分析精确的均值和样本集合上数据均值的差异；

6：不允许组队，如果在考察中，发现两人的程序完全相同，则两人本次的实验成绩为 0。