第五次作业

姓名：刘文晨

学号：202222280328

习题4.2.1

4.2.5 Exercises for Section 4.2

Exercise 4.2.1: Suppose we have a stream of tuples with the schema

Grades(university, courseID, studentID, grade)

Assume universities are unique, but a courseID is unique only within a university (i.e., different universities may have different courses with the same ID, e.g., "CS101") and likewise, studentID's are unique only within a university (different universities may assign the same ID to different students). Suppose we want to answer certain queries approximately from a 1/20th sample of the data. For each of the queries below, indicate how you would construct the sample. That is, tell what the key attributes should be.

(a) For each university, estimate the average number of students in a course.

(b) Estimate the fraction of students who have a GPA of 3.5 or more.

(c) Estimate the fraction of courses where at least half the students got "A."

(a)

对每个元组产生一个随机整数，范围是0~19，并当且仅当随机数为0是才存储该元组。

对存储下来的元组，以university属性分类，对于同一university属性的元组，统计不同的courseID数量，将元组数/courseID数，得到的就是每所大学在一个课程中的平均学生数目。

伪代码：

```
SELECT university, COUNT(*)/COUNT(DISTINCT courseID) AS avg_student
FROM Grades_hash_0
GROUP BY university
```

(b)

对每个元组产生一个随机整数，范围是0~19，并当且仅当随机数为0是才存储该元组。

对存储下来的元组，统计grade>=3.5的数量，除以元组数就得到GPA不低于3.5分的学生所占的比例。

伪代码：

```
SELECT COUNT(*) AS num
FROM Grades_hash_0
WHERE grade>=3.5

SELECT COUNT(*) AS sum
FROM Grades_hash_0

ratio=num/sum
```

(c)

对每个元组产生一个随机整数，范围是0~19，并当且仅当随机数为0是才存储该元组。

设置一个计数器count，初始化为0。

对存储下来的元组，以courseID属性分类，统计课程数量sum。对于同一courseID属性的元组，统计greade为A的数量和元组数量，若两者相除结果≥0.5，计数器count加1。

最终结果=count/sum。

伪代码：

```
count=0

for courseid in courseID:

    SELECT COUNT(*) AS num_A
    FROM Grades_hash_0
    WHERE grade='A'
    GROUP BY courseID

    SELECT COUNT(*) AS num
    FROM Grades_hash_0
    GROUP BY courseID

    if(num_A/num>=0.5)
        count++

SELECT COUNT(DISTINCT courseID) AS sum
FROM Grades_hash_0

ratio=count/sum
```

习题4.2.1

4.3.4 Exercises for Section 4.3

Exercise 4.3.1: For the situation of our running example (8 billion bits, 1 billion members of the set S), calculate the false-positive rate if we use three hash functions? What if we use four hash functions?

如果使用3个哈希函数，相当于往80亿个靶位上投30亿支飞镖，某个位为0的概率为$e^{-3/8}$。一个非S中的元素若要成为伪正例的话，那么就必须在3个哈希函数的作用下都映射为1，而该概率为$(1-e^{-3/8})^3 \approx 0.0306$。

如果使用4个哈希函数，相当于往80亿个靶位上投40亿支飞镖，某个位为0的概率为$e^{-1/2}$。一个非S中的元素若要成为伪正例的话，那么就必须在4个哈希函数的作用下都映射为1，而该概率为$(1-e^{-1/2})^4 \approx 0.0240$。

习题4.5.1

4.5.6 Exercises for Section 4.5

Exercise 4.5.1: Compute the surprise number (second moment) for the stream 3, 1, 4, 1, 3, 4, 2, 1, 2. What is the third moment of this stream?

在这个流中，有3个元素出现了2次，有个1元素出现了3次。

二阶矩$=3\times2^2+1\times3^2=21$

三阶矩$=3\times2^3+1\times3^3=51$

习题4.6.1

4.6.8 Exercises for Section 4.6

Exercise 4.6.1: Suppose the window is as shown in Fig. 4.2. Estimate the number of 1's the the last k positions, for $k =$ (a) 5 (b) 15. In each case, how far off the correct value is your estimate?

当k=5时，估计值为3，包括2个大小为1的桶的大小，1个大小为2的桶的大小的一半，真实值为3，此时，估计值与真实值的差值为0。

当k=15时，估计值为10，包括2个大小为1的桶的大小，1个大小为2的桶的大小，1个大小为4的桶的大小以及1个大小为4的桶的大小的一半，真实值为9，此时，估计值与真实值的差值为1。