

# CSCI262 – SYSTEM SECURITY

---

## Statistical Inference

14 October 2022

# Statistical inference

Name	Gender	Hobby	Marital Status	Age
Andrew	Female	Reading	Divorced	35
Biggles	Female	Running	Single	30
Cross	Female	Running	Single	30
Dupont	Male	Reading	Divorced	35
Evans	Female	Reading	Single	19
French	Female	Walking	Single	22
Green	Male	Reading	Married	29
Hall	Male	Running	Single	24
Ivanov	Male	Reading	Single	27
Jones	Male	Walking	Married	27
King	Male	Running	Single	19
Lee	Male	Walking	Single	24

# Statistical inference

- Assuming that we have only a statistical interface, and hence only aggregate queries can be executed. We also know Evans is **female**, **single** and had **reading** as hobby. In addition she is also attractive 😊😊😊 (of course the last information is redundant, ha..ha..).
- We want to know her age and we can only infer because we only have statistical interface and face with various query size restrictions.

# Statistical inference

- Assume there is **no** limit on the query size, that is, the statistical interface does not restrict on the minimum number of records to respond to the query, give a sequence of two queries that will identify Evans' age.

# Statistical inference

- Before we can infer, we need to have some information with respect to the number of records satisfying the criteria – female, single and reading. So we executed the following query:

```
SELECT    count(*)           // Assuming the table
FROM      Particular         // name is Particular
WHERE     Gender = 'Female'
AND       Hobby = 'Reading'
AND       MaritalStatus = 'Single';
```

This query will  
return a value 1.

# Statistical inference

- Of course, if there is no requirement that we need to use two queries (as specified in the question), we can get the age as follow:

```
SELECT sum(age)
FROM    Particular
WHERE   Gender = 'Female'
AND     Hobby = 'Reading'
AND     MaritalStatus = 'Single';
```

... But unfortunately we cannot.

# Statistical inference

- We need to find another statistic information based on a subset conditions such as female and reading as hobby, or female and single, or single and reading as hobby.
- Assuming, we tried all the possible subset conditions 😊

```
SELECT count(*)  
FROM Particular  
WHERE Gender = 'Female'  
AND Hobby = 'Reading';
```

This query will return a value 2.  
(The result of counting the records Andrews and Evans.)

# Statistical inference

```
SELECT count(*)  
FROM    Particular  
WHERE   Gender = 'Female'  
AND     MaritalStatus = 'Single';
```

This query will return a value 4. (The result of counting the records Biggles, Cross, Evans, and French.)



# Statistical inference

```
SELECT count(*)  
FROM    Particular  
WHERE   Hobby = 'Reading'  
AND     MaritalStatus = 'Single';
```

This query will return a value 2. (The result of counting the records Evans and Ivanov.)

# Statistical inference

- From these three queries, the first and the third queries can be used, but not the second query. Why? 😊
- The different in number of records between the second query and our query with triplet conditions earlier is 2.

```
SELECT    sum(age)
FROM      Particular
WHERE     Gender = 'Female'
AND       Hobby = 'Reading'
AND       MaritalStatus = 'Single';
```

# Statistical inference

- With these information, we can now construct the following two queries to infer Evans' age:

```
SELECT    sum(age)
FROM      Particular
WHERE     Gender = 'Female'
AND       Hobby = 'Reading'
```

```
MINUS
SELECT    sum(age)
FROM      Particular
WHERE     Gender = 'Female'
AND       Hobby = 'Reading'
AND       MaritalStatus <> 'Single';
```

The result returned by the queries is 19, that is,  $54 - 35 = 19$ . Hence we now know Evans is 19 years old 😊.

# Statistical inference

- With a similar deduction, we can use the following queries to infer Evans' age:

```
SELECT    sum(age)
FROM      Particular
WHERE     MaritalStatus = 'Single'
AND       Hobby = 'Reading'
```

```
MINUS
SELECT    sum(age)
FROM      Particular
WHERE     Gender != 'Female'
AND       Hobby = 'Reading'
AND       MaritalStatus = 'Single';
```

The result returned by the queries is 19, that is,  $46 - 27 = 19$ . Hence we now know Evans is 19 years old 😊.

# Statistical inference

- Now, suppose there is a **limit on the query limit of 2**, that is, the statistical interface will not respond to query if the minimum number of records is less than 2, give a sequence of two queries that will identify Evans' age.

# Statistical inference

- Similar as (a), before we can infer, we need to have some information with respect to the number of records satisfying the criteria – female, single and reading. So we executed the following query:

```
SELECT    count(*)           // Assuming the
FROM      Particular         // table name is Particular
WHERE     Gender = 'Female'
AND       Hobby = 'Reading'
AND       MaritalStatus = 'Single';
```

This query will not return any result because the total number of record satisfying this query is 1, which is less than the limit set.

# Statistical inference

- We try another query, and we have:

```
SELECT    count(*)           // Assuming the table
FROM      Particular         // name is Particular
WHERE     Gender = 'Female'
AND       MaritalStatus = 'Single';
```

This query will return a value of 4. (The records satisfying the conditions are Biggles, Cross, Evans, and French.)

# Statistical inference

- Assuming that we are lucky, in our first try of subset condition we obtained the second query that we need, that is,

```
SELECT    count(*)  
FROM      Particular  
WHERE     Gender = 'Female'  
AND       MaritalStauts = 'Single'  
AND       Hobby != 'Reading';
```

This query will return a value 3. (The records satisfying the above condition are Biggles, Cross, and French.)



# Statistical inference

- Now, we can infer Evans's age by using the following two queries:

```
SELECT    sum(age)
FROM      Particular
WHERE     Gender = 'Female'
AND       MaritalStatus = 'Single'
```

```
MINUS
SELECT    sum(age)
FROM      Particular
WHERE     Gender = 'Female'
AND       MaritalStauts = 'Single'
AND       Hobby != 'Reading';
```

This query will return  
a value 19, that is,  
 $101 - 82 = 19$ . 😊

# Statistical inference

*Hope the example help!* 😊 😊 😊