**lecture1**

What are new value range after normalization by decimal scaling? Justify your

answer.

*Answer: (-1, 1). The values are normalized by moving the decimal point such that the maximal absolute values of the new mapped features is smaller than*

Data Normalization meaning

Because the features have different scales and normalization could make them comparable. Especially, the features with wider scale may dominate the Euclidean distance. TODO: Decimal Scaling -986 to 917, 选 -986的绝对值，然后除以它的最大的位数，比如1000, [-0.986, 0.917]

In real-world data analytic problems, data with missing values for some

attributes is often occurred. Why missing values are common in real-world application?

Describe various methods for handling missing values.

*Answer: Missing data may be due to: (1) Equipment malfunction; (2) Data is not entered; (3) Certain data may not be considered at the time of data collection; (4) Not applicable for some cases Possible methods: (1) Ignore the data sample with missing values; (2) Ignore attributes with missing values; (3) Fill in missing value by a global constant or Attribute mean/Median/mode or using a model to predict the missing value (data imputation)*

Suppose we have the following values for prices (already sorted in increasing

order): [5, 10, 15, 21, 21, 22, 25, 28, 30, 31, 31, 32]. Use smoothing by bin means to smooth the above data using equal-depth binning with bin depth of 4. Illustrate your steps.

*Answer:*

*1. Partition data into bins*

*Bin 1: [5, 10, 15, 21]*

*Bin 2: [21, 22, 25, 28]*

*Bin 3: [30, 31, 31, 32]*

*2. Computing the mean of each bin:*

*Bin1: 12.75*

*Bin2: 24*

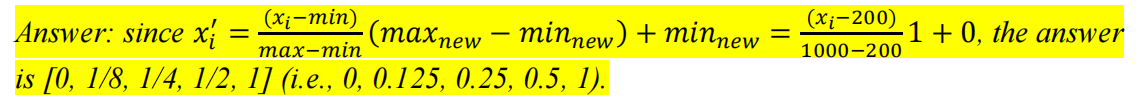
*Bin 3: 31*

*3. Smoothing by mean*

*[12.75, 12.75, 12.75, 12.75, 24, 24, 24, 24, 31, 31, 31, 31]*

Suppose that we have the following data: [200, 300, 400, 600, 1000], normalize

the data by min-max normalization by setting new min to 0 and new max to 1.



Suppose we have the following values for prices (already sorted in increasing

order): [4, 10, 15, 21, 21, 22, 25, 28, 30, 31, 31, 32]. Use smoothing by bin means to smooth the above data using equal-width binning with 4 bins. Illustrate your steps.

*Answer:*

*1. Compute the width of each bin: (32-4)/4 = 7.*

*2. Partition data into bins (equal-width):*

*Bin 1 ( [4, 11) ): [4, 10]*

*Bin 2 ( [11, 18) ): [15]*

*Bin 3 ( [18, 25)): [21, 21, 22]*

*Bin 4 ( [25, 32]): [25, 28, 30, 31, 31, 32]*

*3. Computing the mean of each bin:*

*Bin1: 7*

*Bin2: 15*

*Bin 3: 21.33*

*Bin 4: 29.5*

*4. Smoothing by mean*

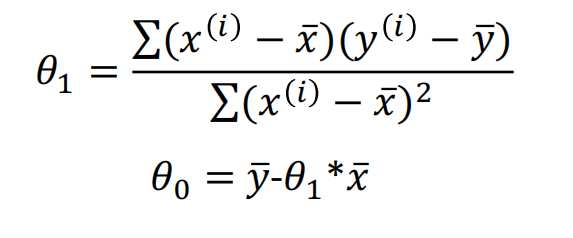
*[7, 7, 15, 21.33, 21.33, 21.33, 29.5, 29.5, 29.5, 29.5, 29.5]*

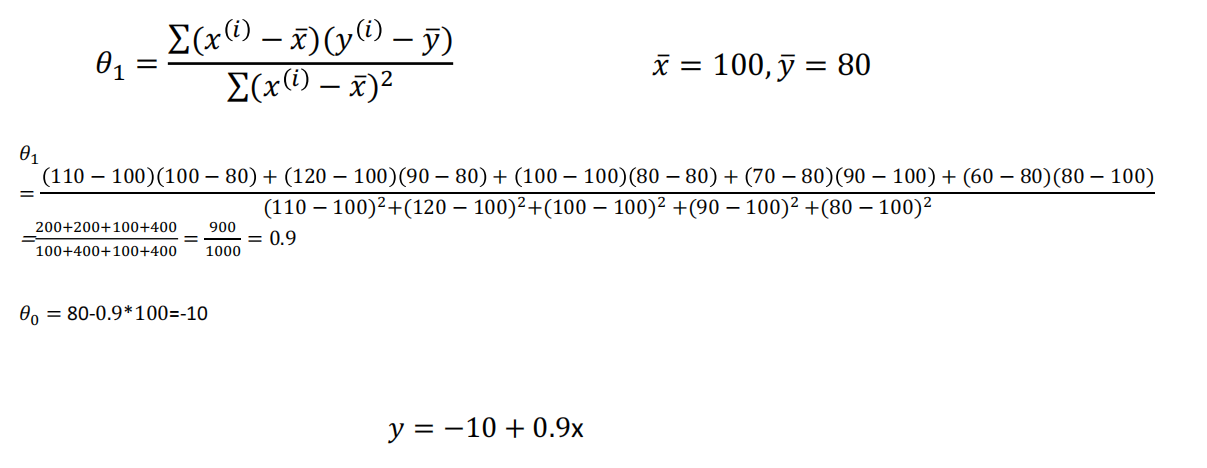
In real-world data analytic problems, data with noisy values for some attributes is often occurred. Why noisy values are common in real-world application? Describe various methods for handling noisy data.

*Answer: Noisy data may due to: (1) Errors in data collection devices; (2) Wrong input; and (3) Technology limitation. Methods for handling noisy data: (1) Binning; (2) Regression; and (3) Clustering*

Regression:



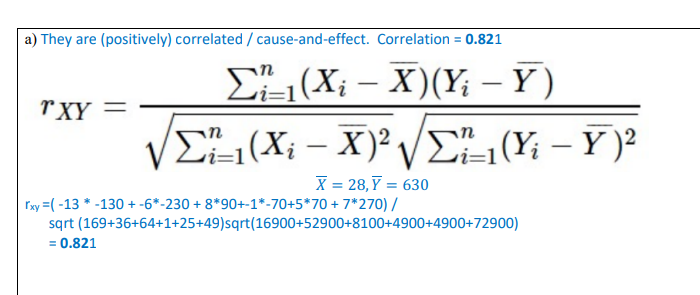




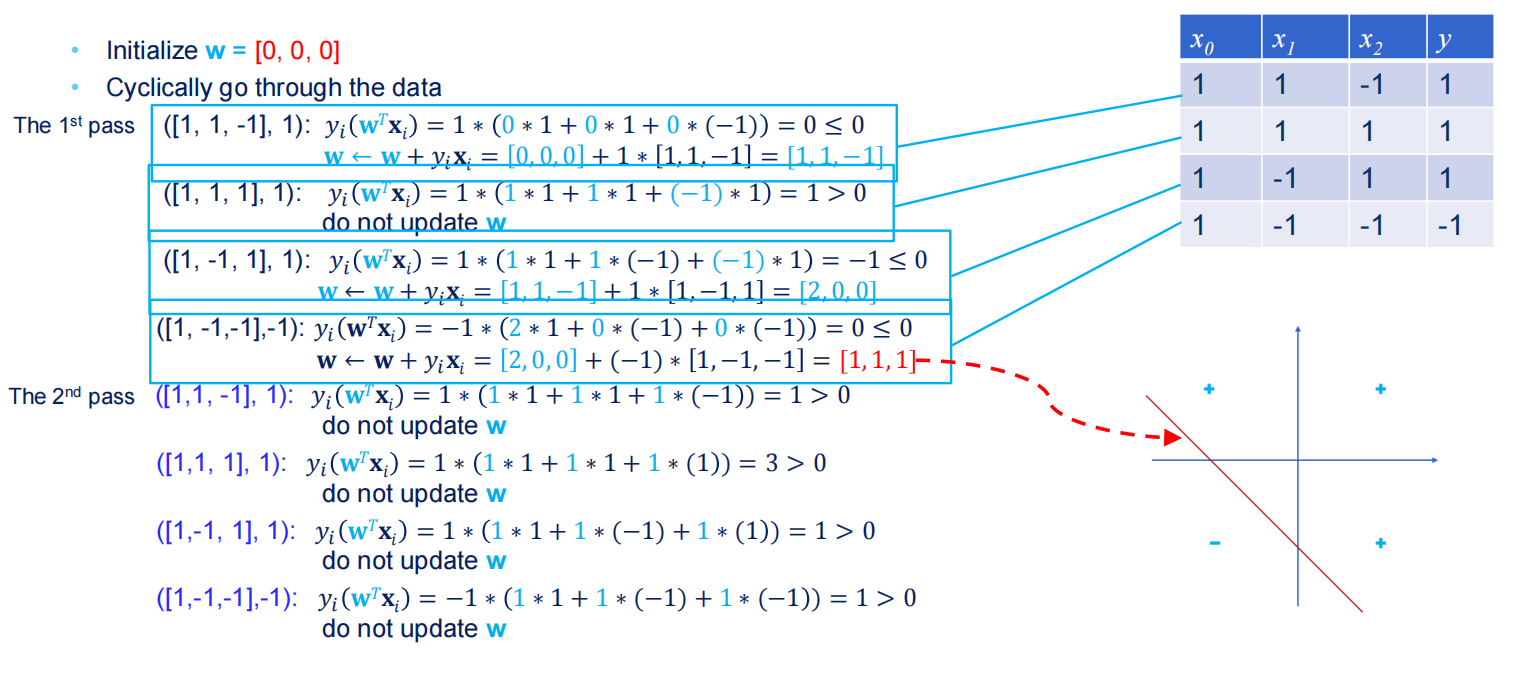
Residual

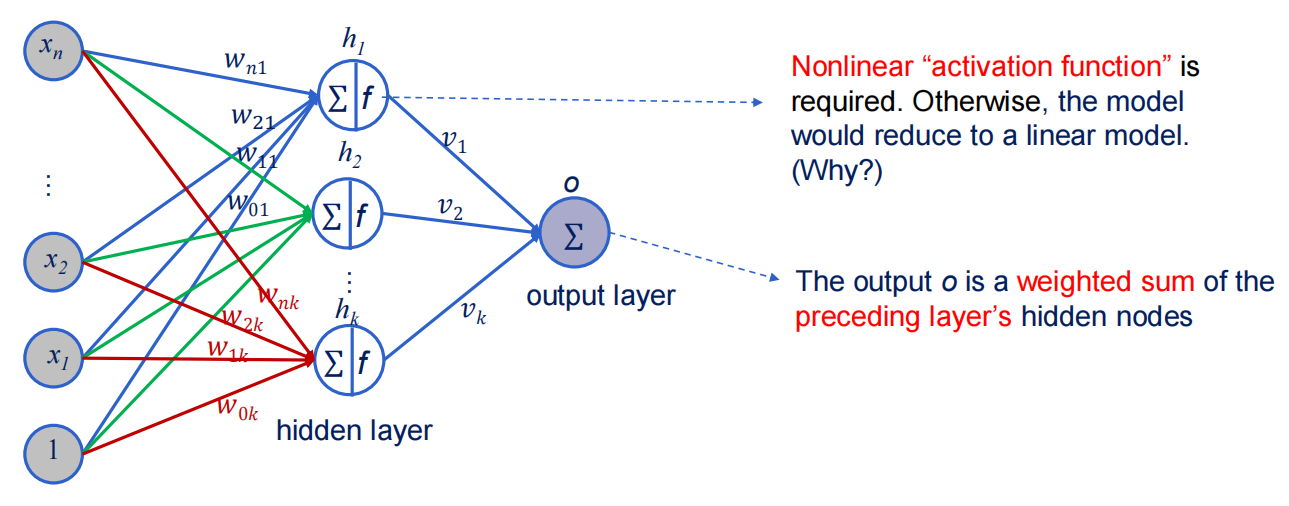
R = |Y - Ypredicted|

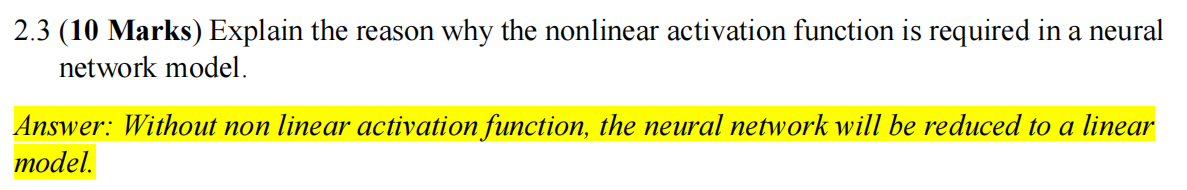
Correlation

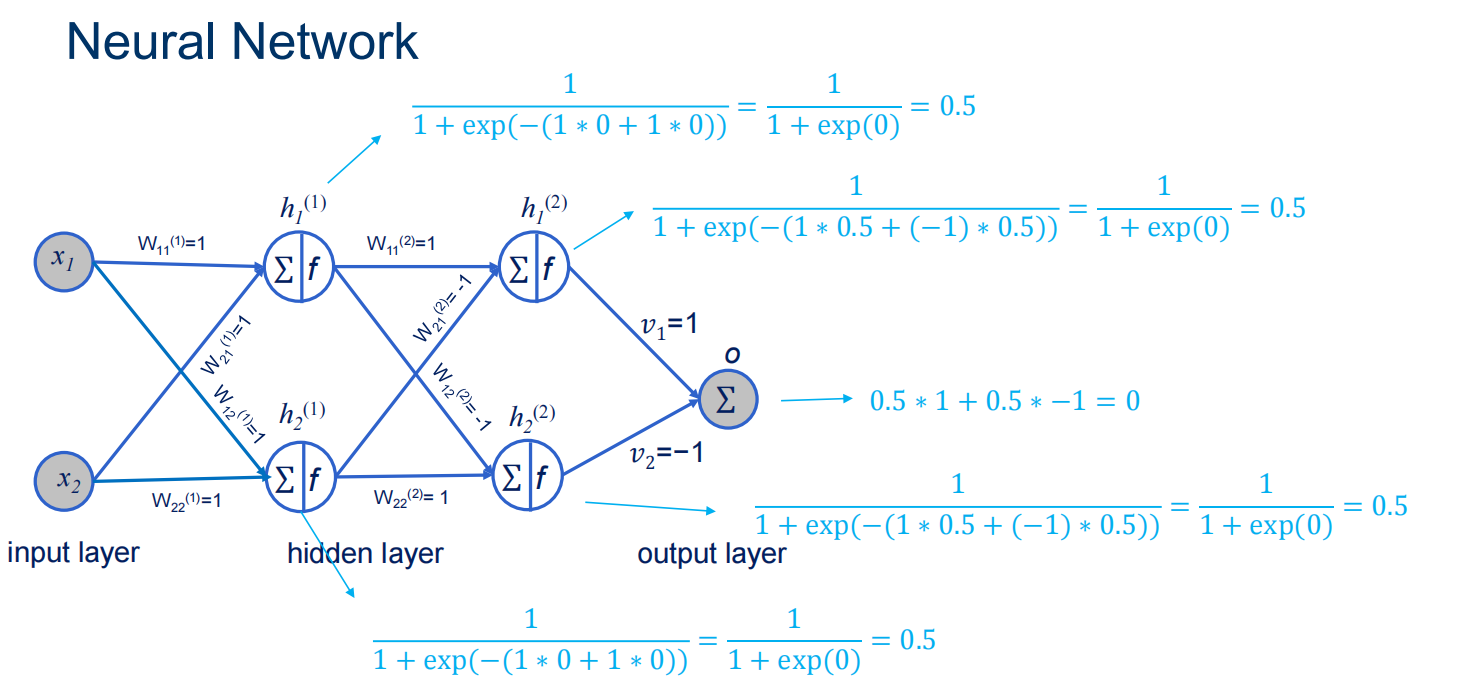


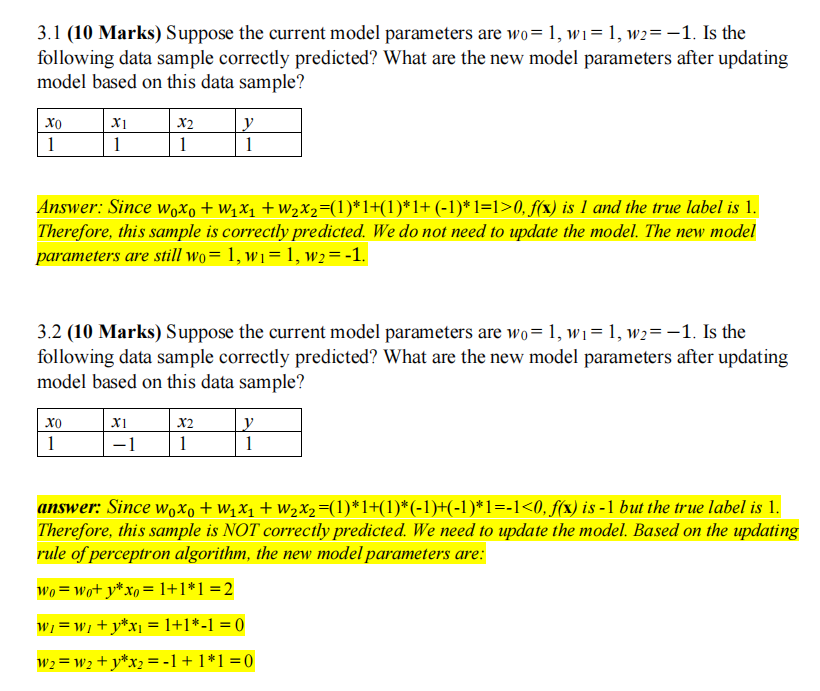
**Lecture2**



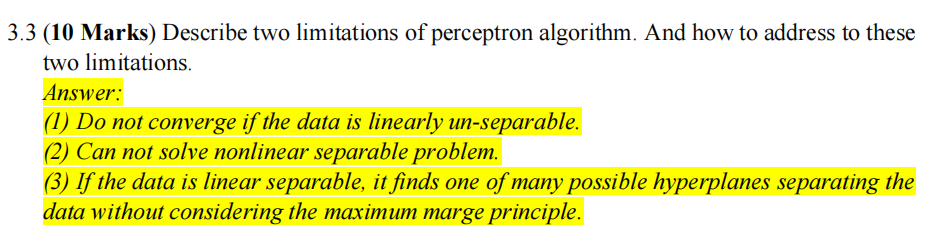








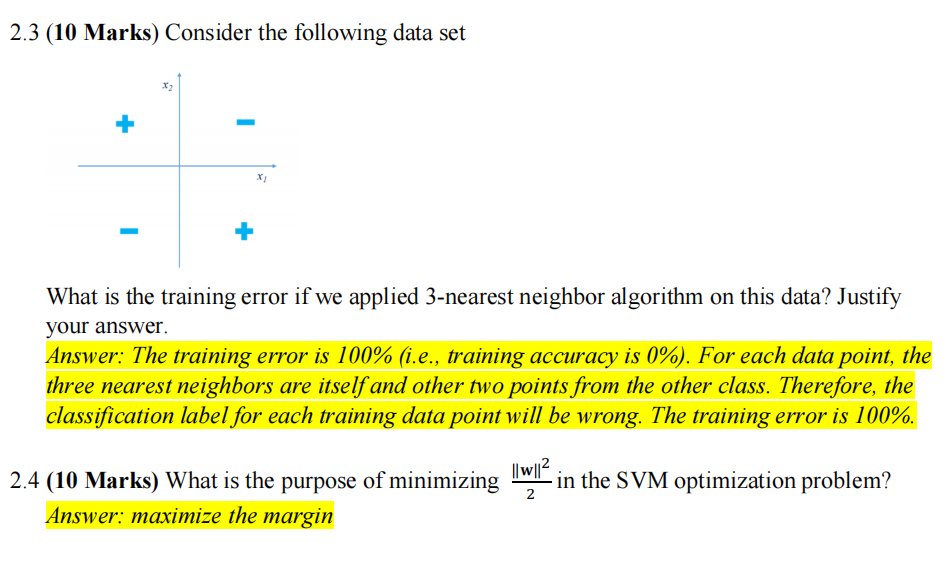
**If Yi(Wt\*Xi) <=0 then W = W+YiXi**

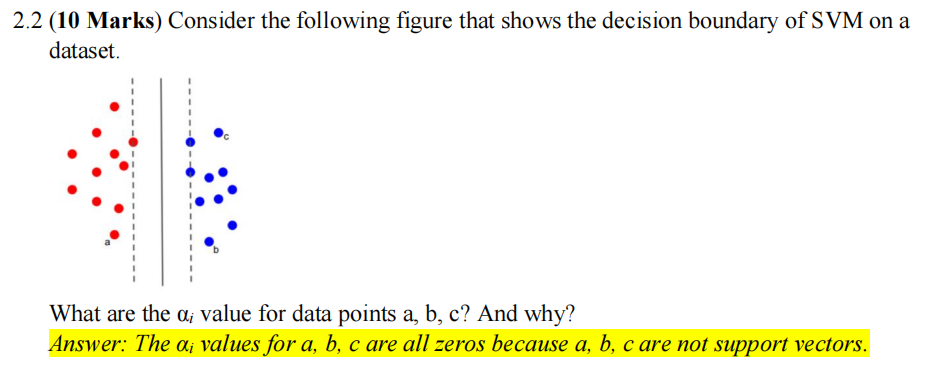


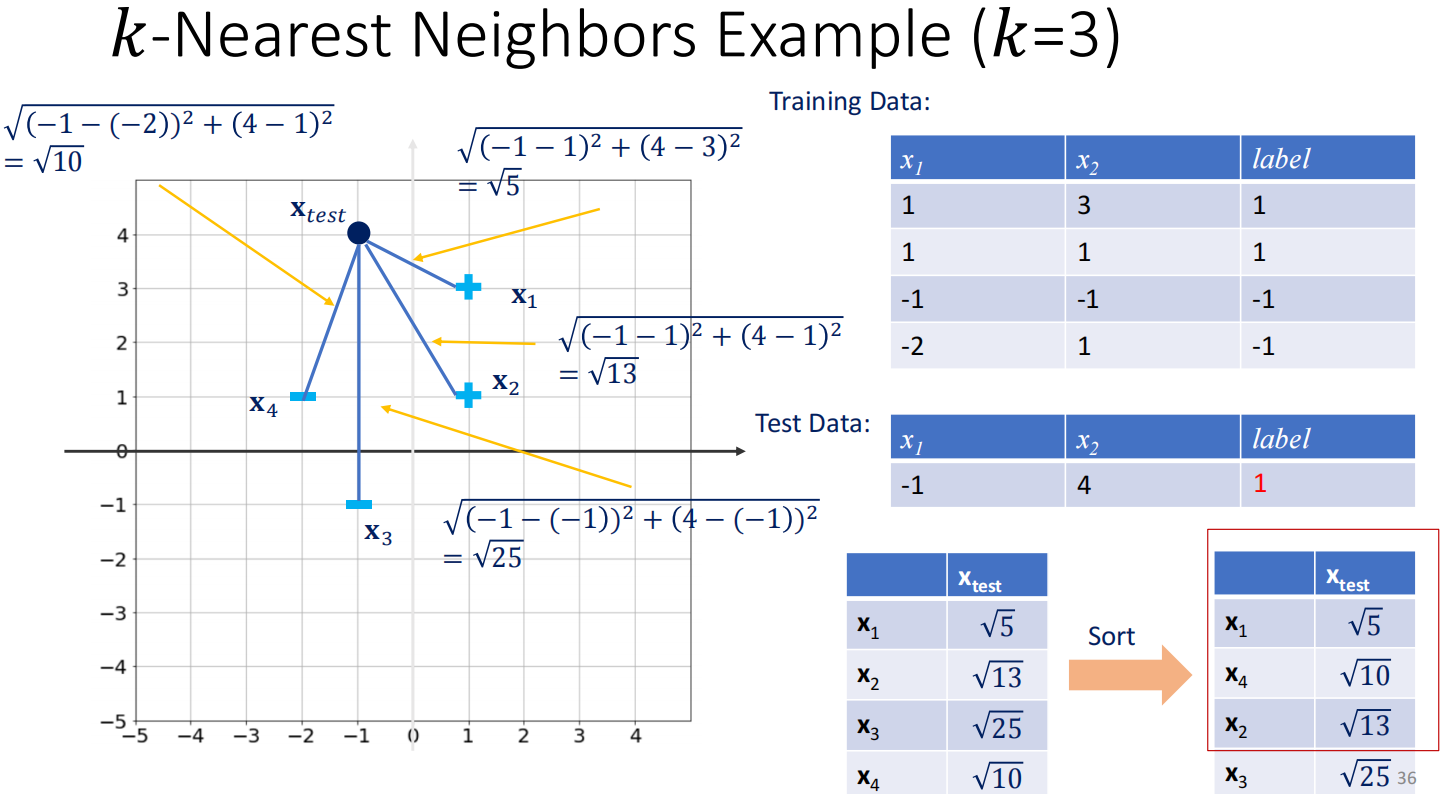
Multiple layer MLP to draw two lines to solve the problem which the data is linearly un-separable

**Lecture3**

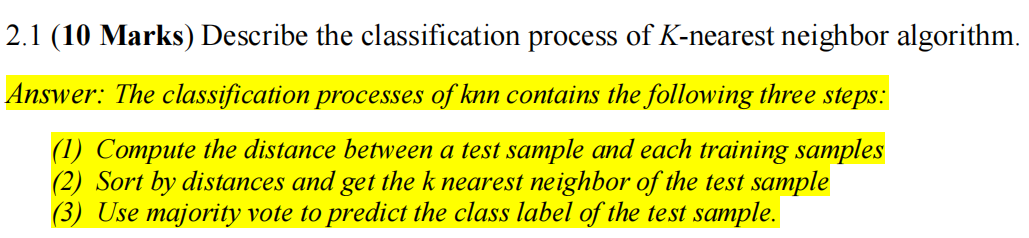








K近邻是解决分类问题



K-means 步骤：

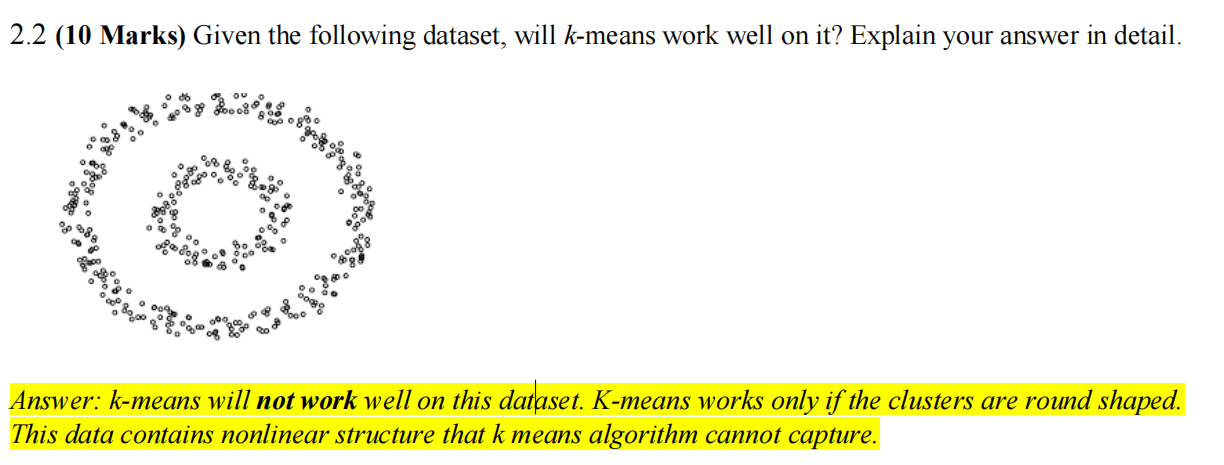
1. 初始化initialize start point

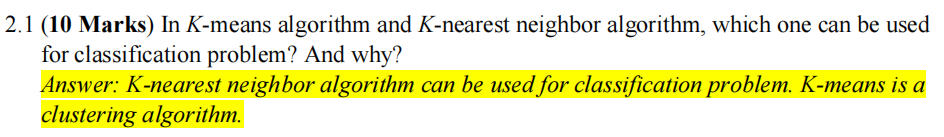
2. iterate the sample, xi to cluster center

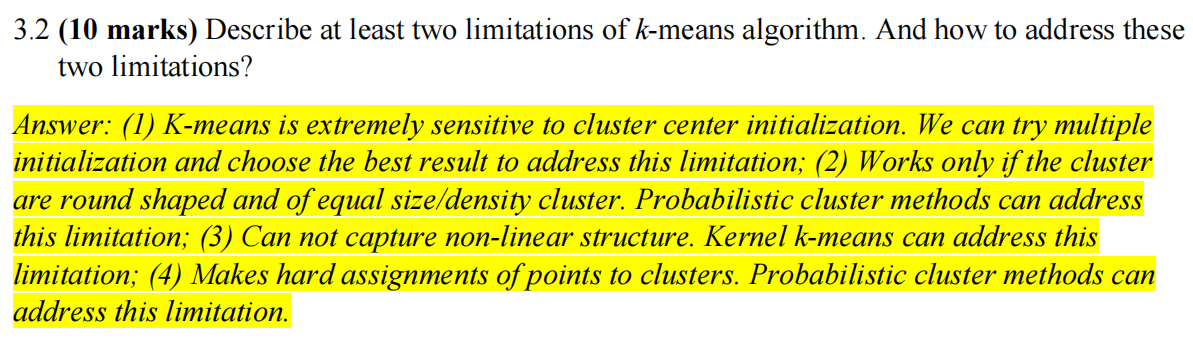
3. Coverage criteria

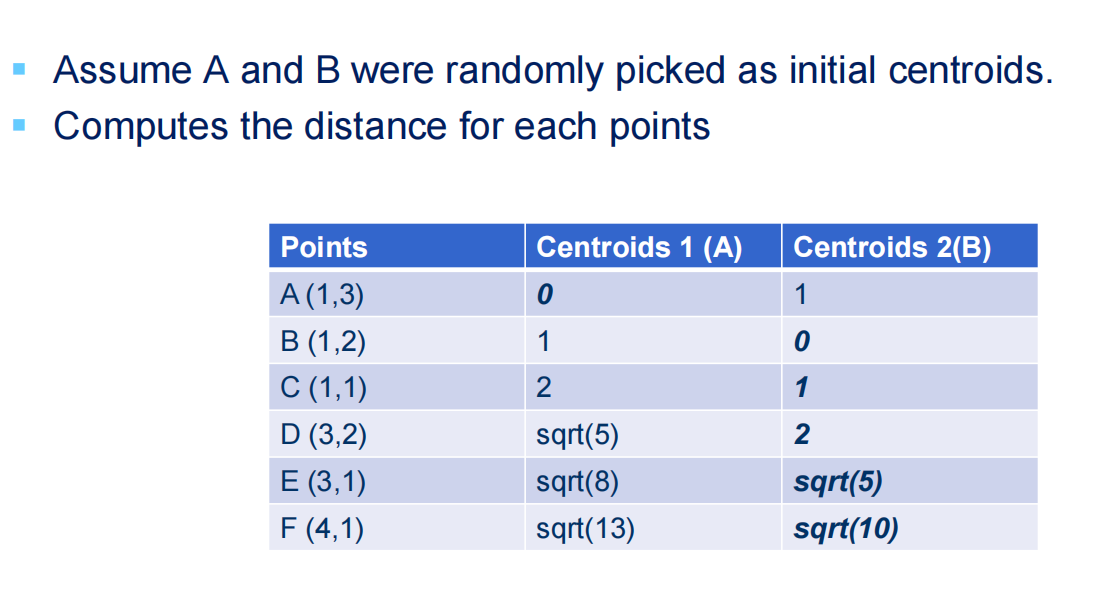
TODO: Hierarchical clustering example

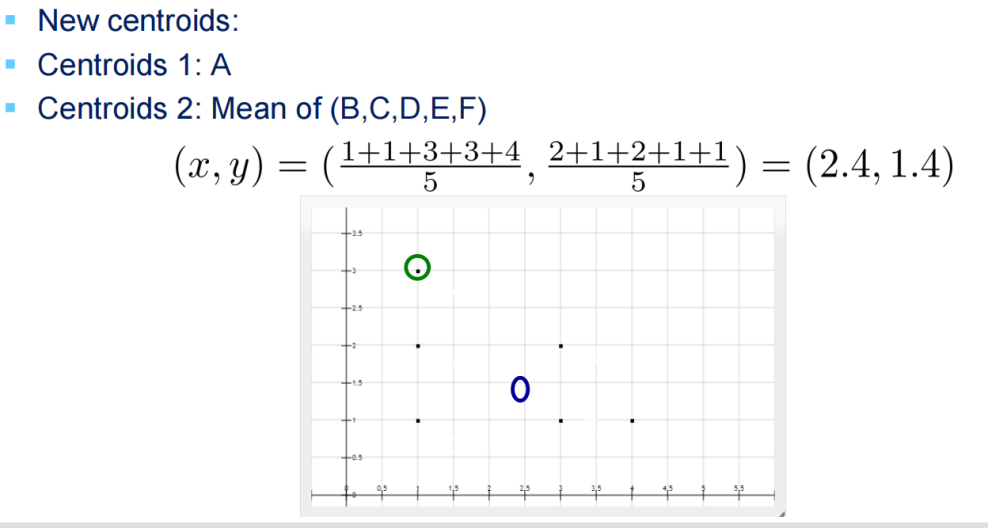
**Lecture4**









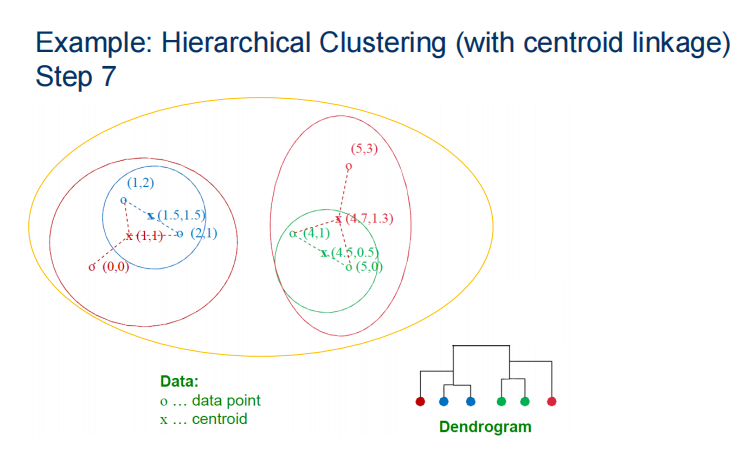


求新的cluster，就是把这个Cluster中所有的X加起来除以N，Y加起来除以N,得到一个的 (X, Y) 坐标。新的cluster是根据每个点到cluster的距离，离cluster的近就是那个cluster了。

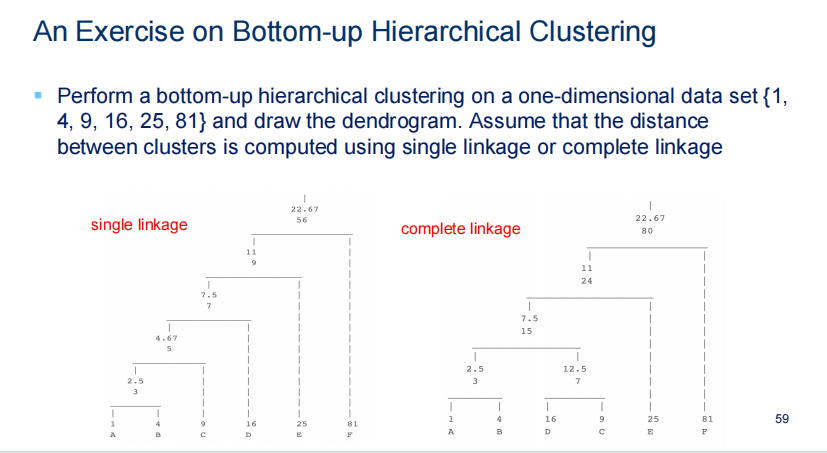
比如A(1,3) 与ClusterA距离0，ClusterB距离为1，那么A(1,3)就是ClusterA了。

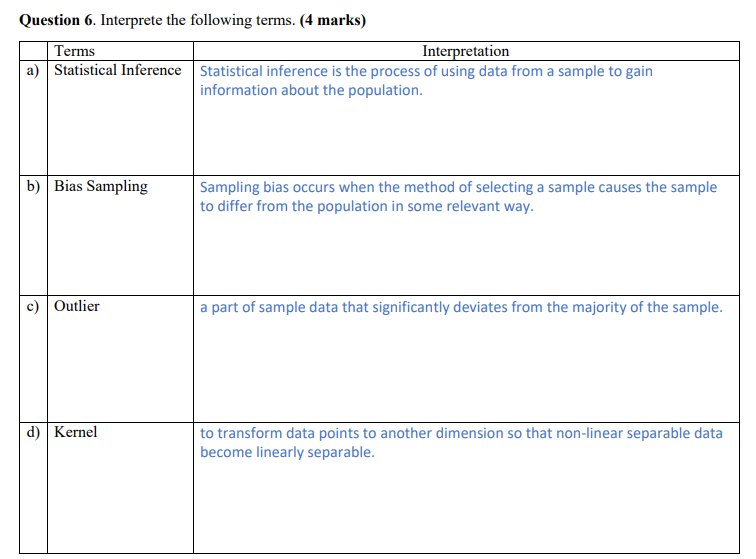
然后根据属于这个cluster的点的坐标的平均值，在算出新的cluster的坐标。重复这个步骤到每个点的cluster不变为止。

Hierarchical Clustering (with Centroid Linkage)

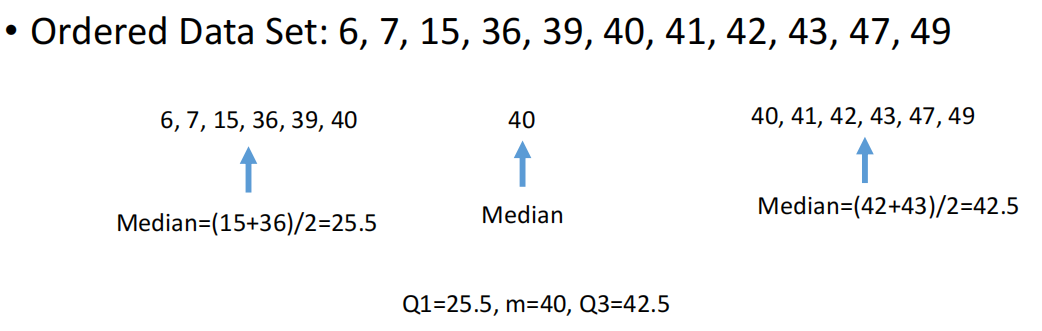


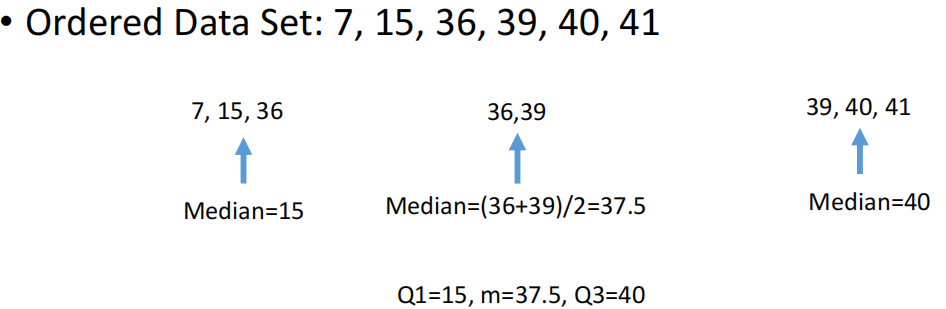
1. 先算每个点的欧拉距离，看哪两个点的距离小， 所以选择选择(1,2) (2,1)连起来, (4,1)(5,0)的欧拉距离也是一样的。然后是 (0,0) -> (1.5,1.5)的欧拉距离最近。



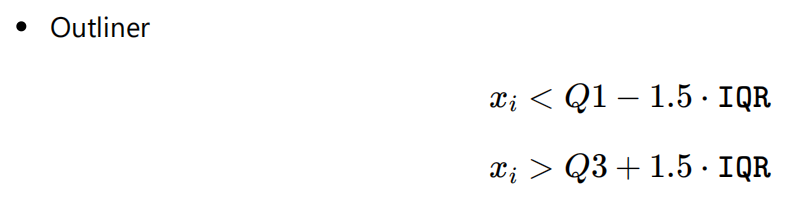


**Lecture5**



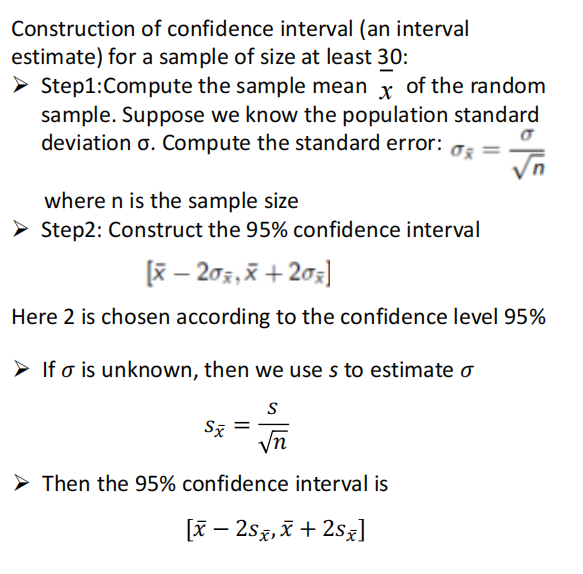


Range: max-min, IQR = Q3-Q1

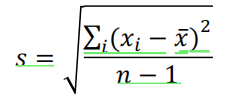


Parametric Tests: assume the distribution of sample data (i.e. normality)

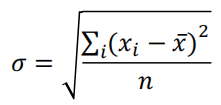
Non-Parametric Tests: do not assume data are drawn from any particular distribution



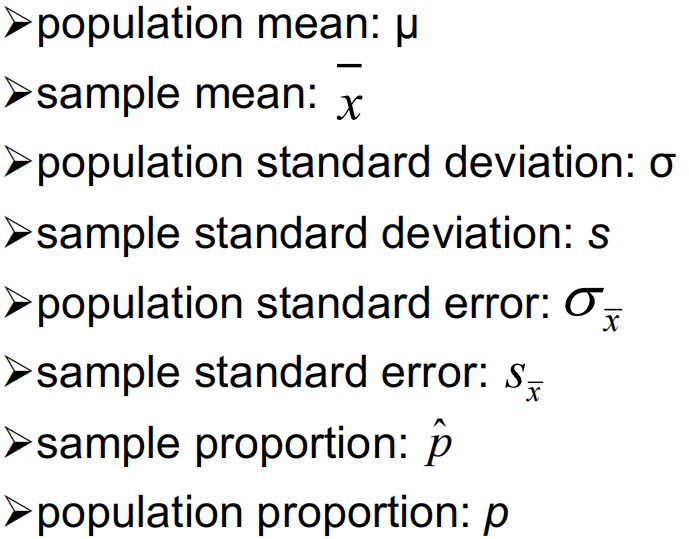
Sample Standard Deviation



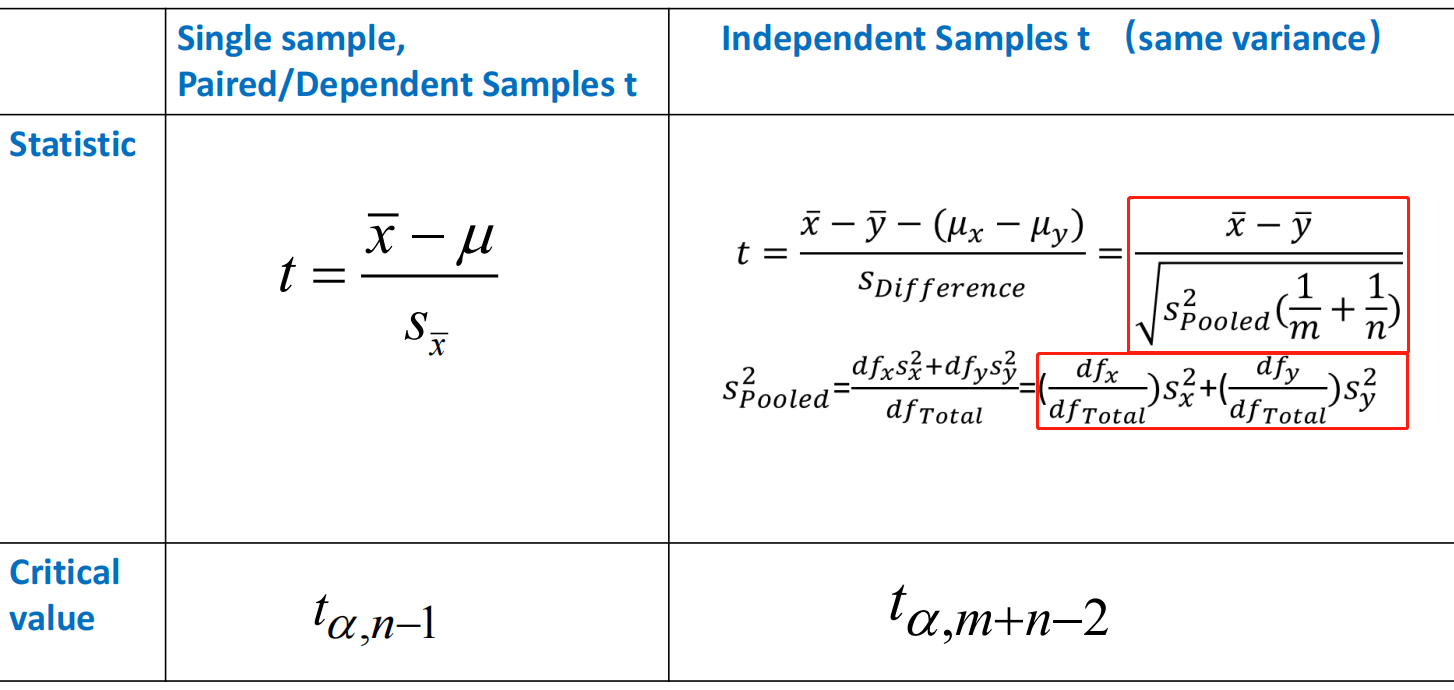
Population Standard Deviation



**Lecture6**



Sx sample standard error 样品的标准差 = s / 根号 n

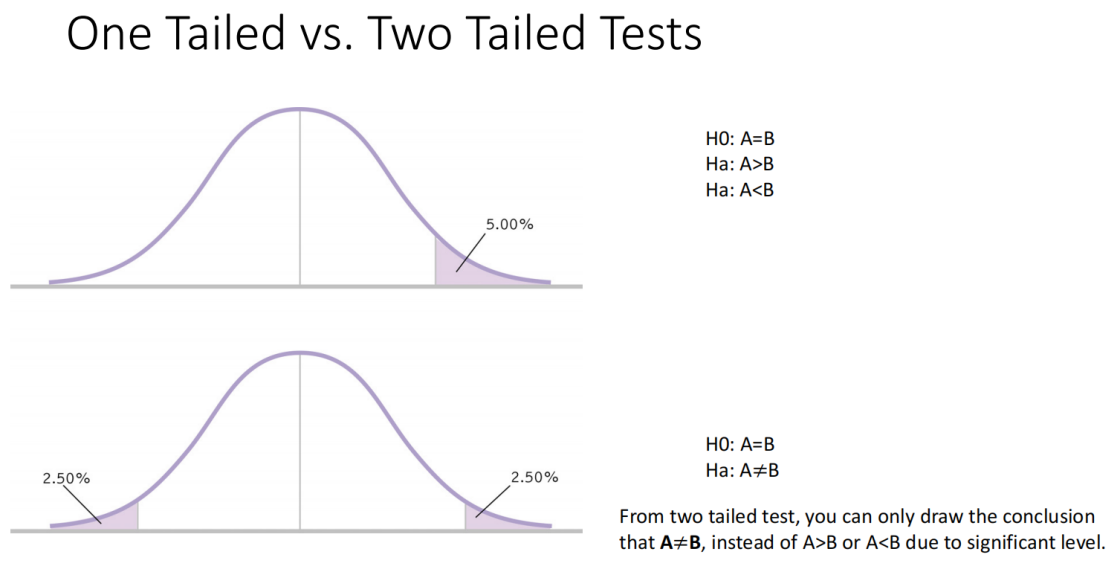


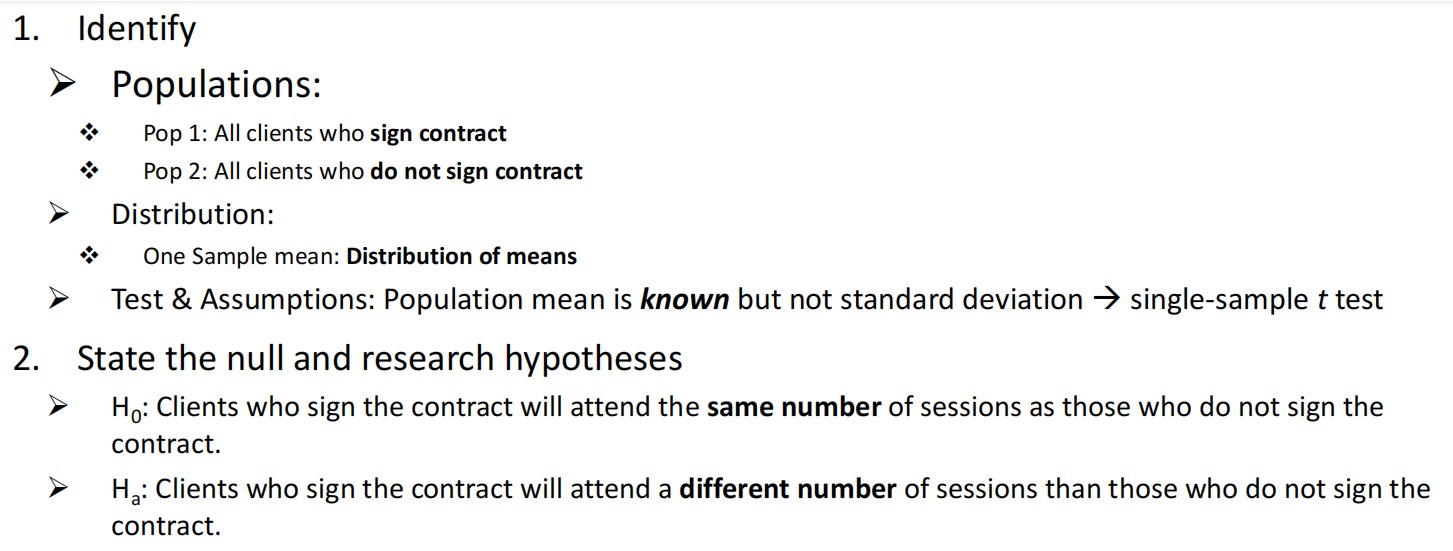
1. Identify 2. State the hypotheses

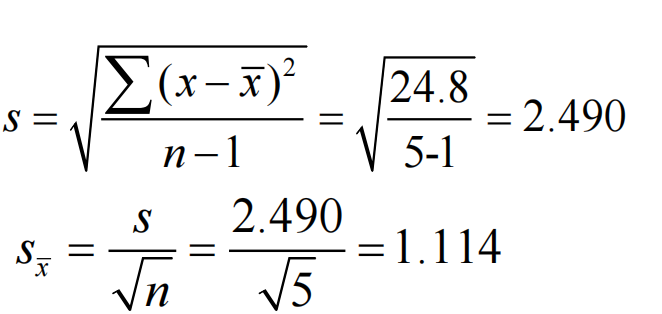
3. Characteristics of the comparison distribution计算

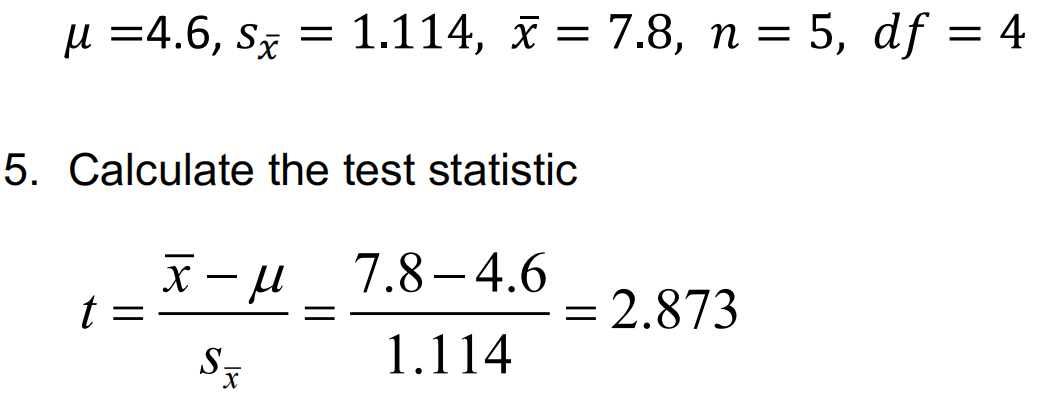
4. Critical values

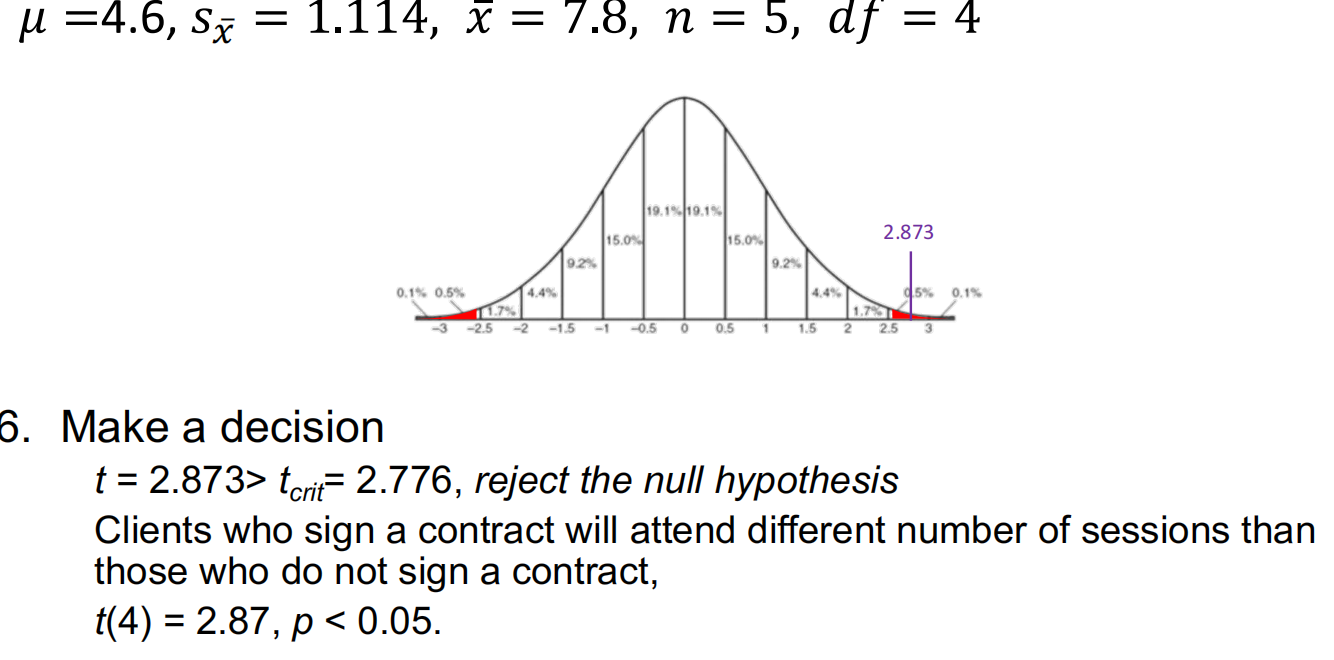
5. Calculate 6. Decide



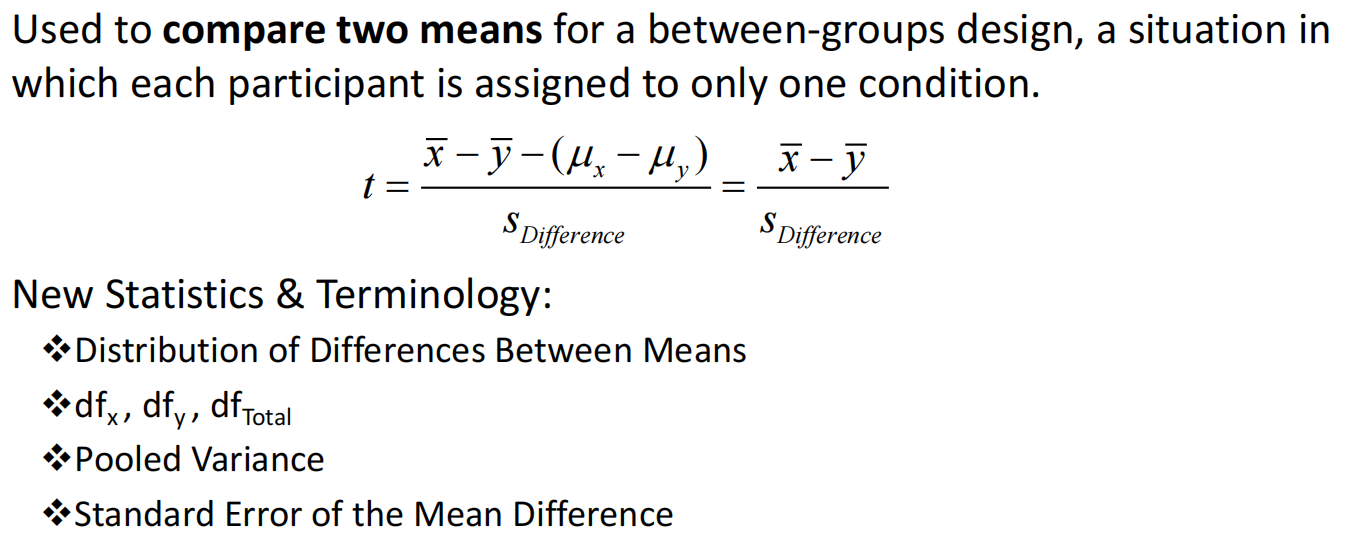


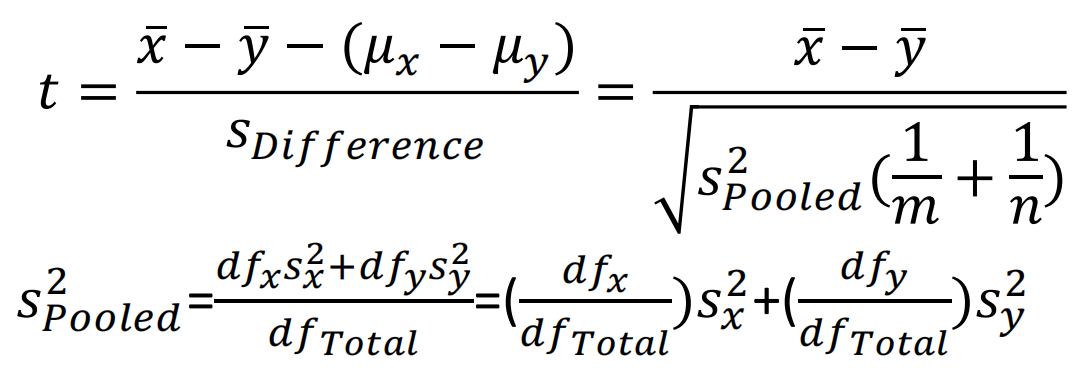


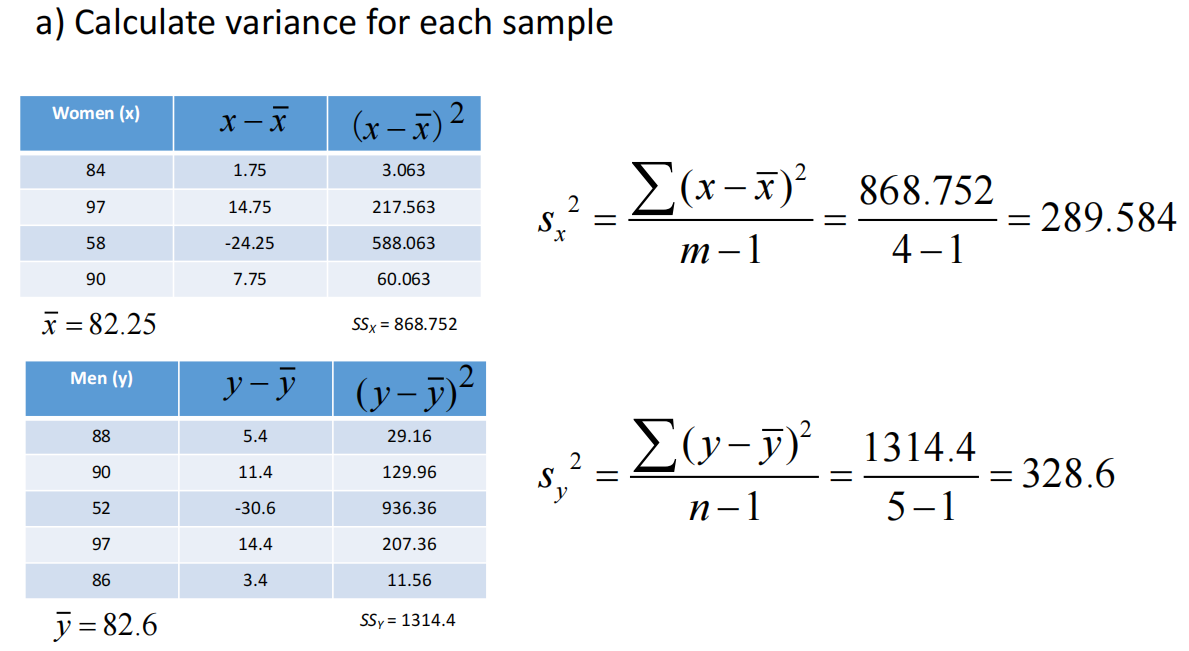


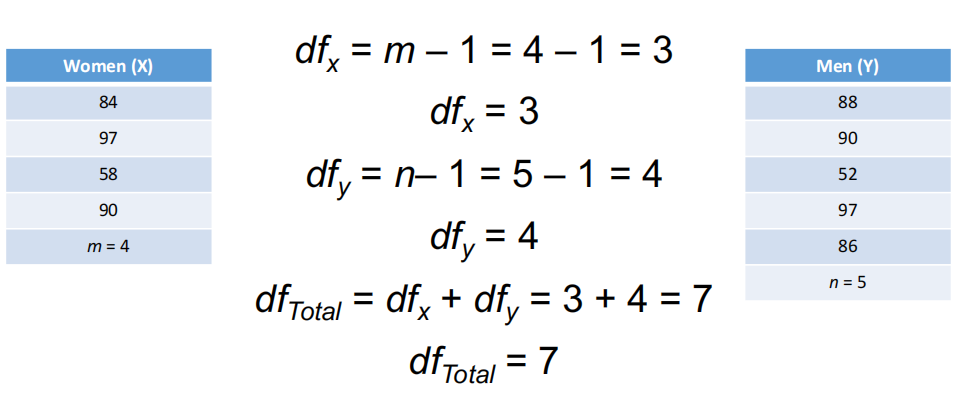


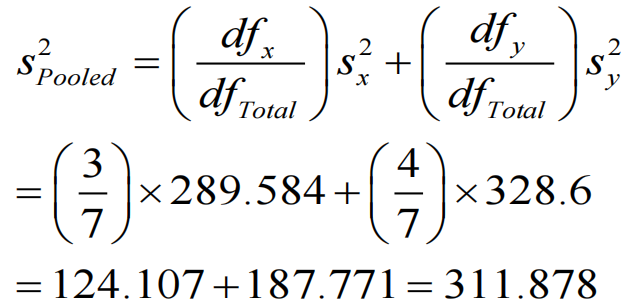
Independent Samples *t* Test

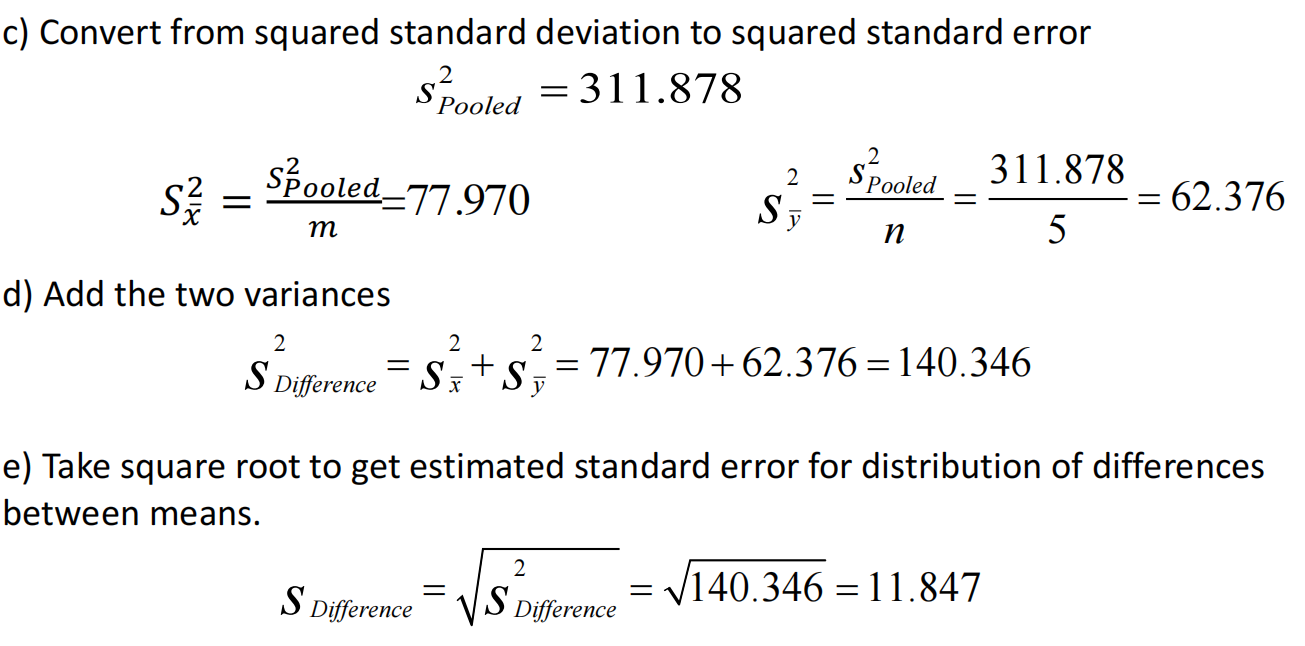


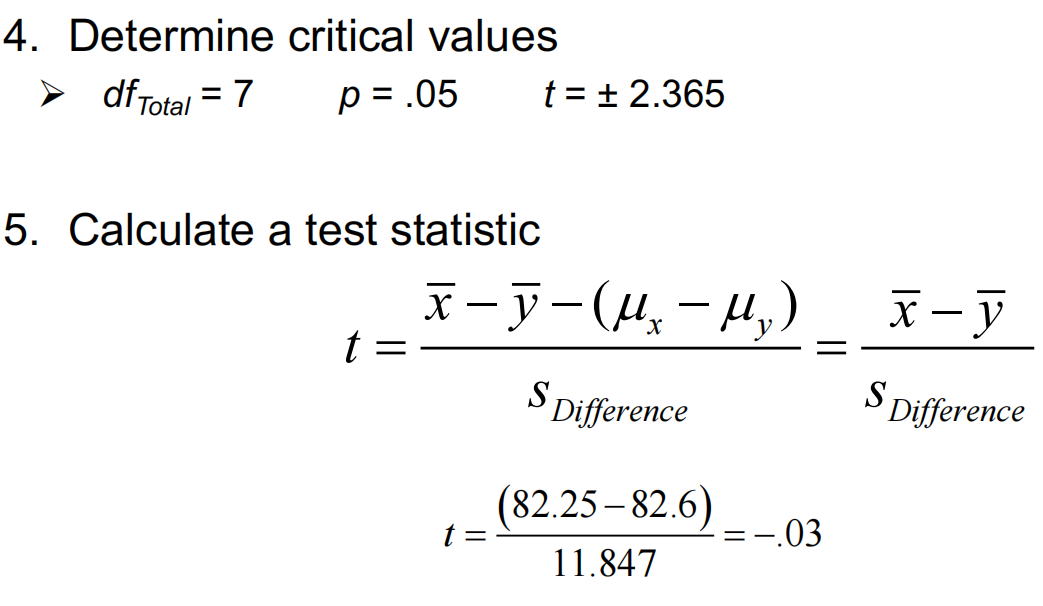




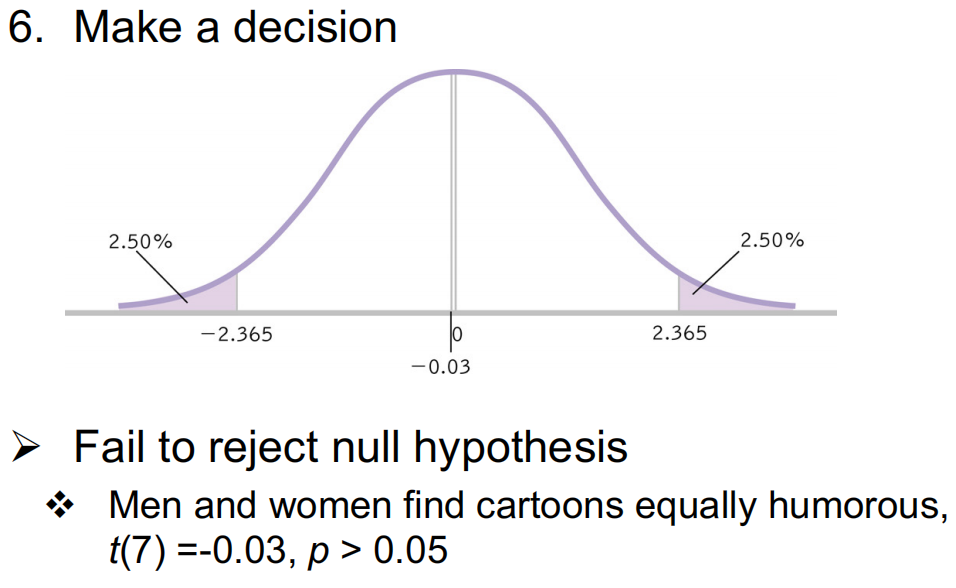


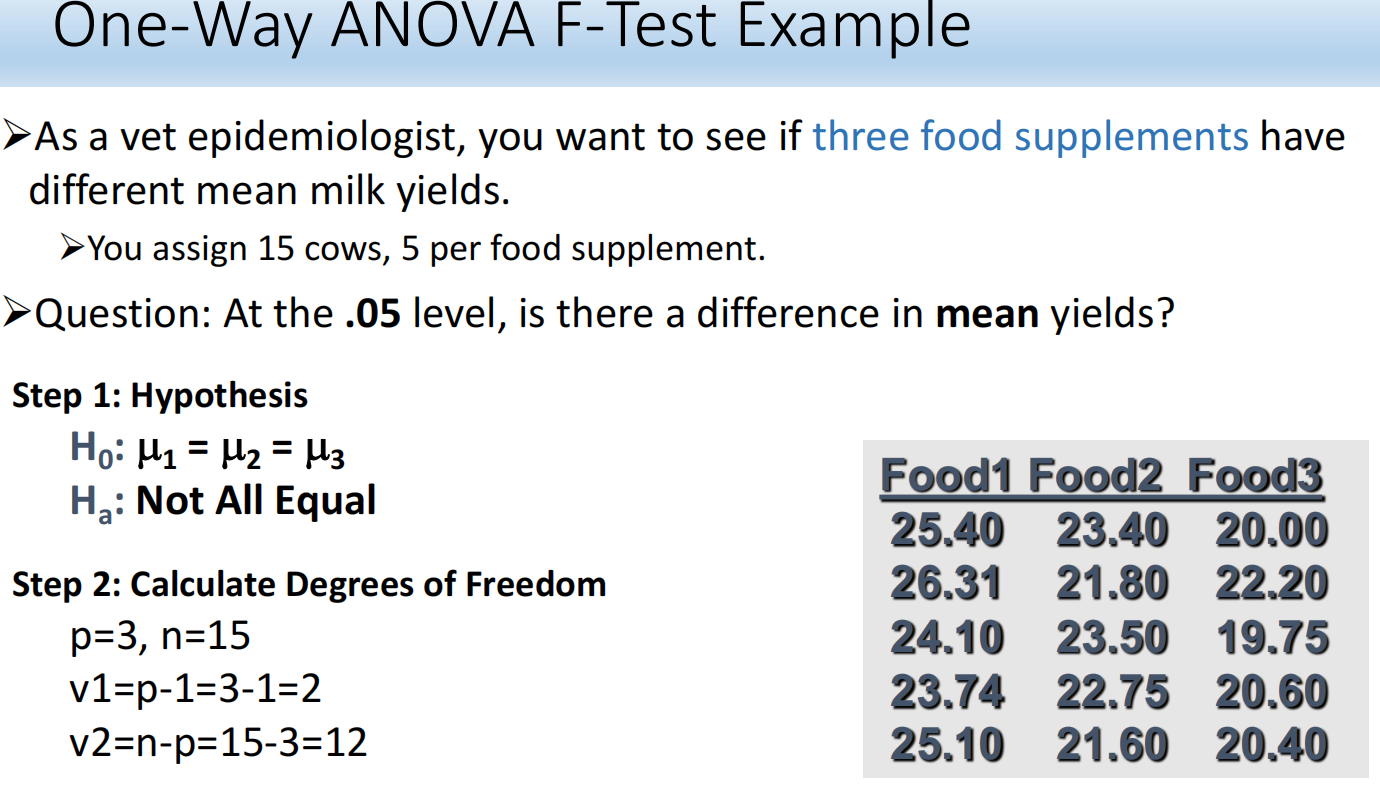


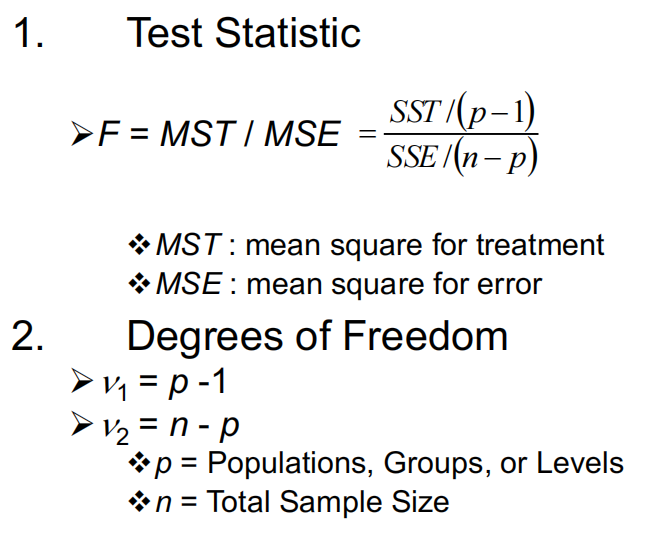


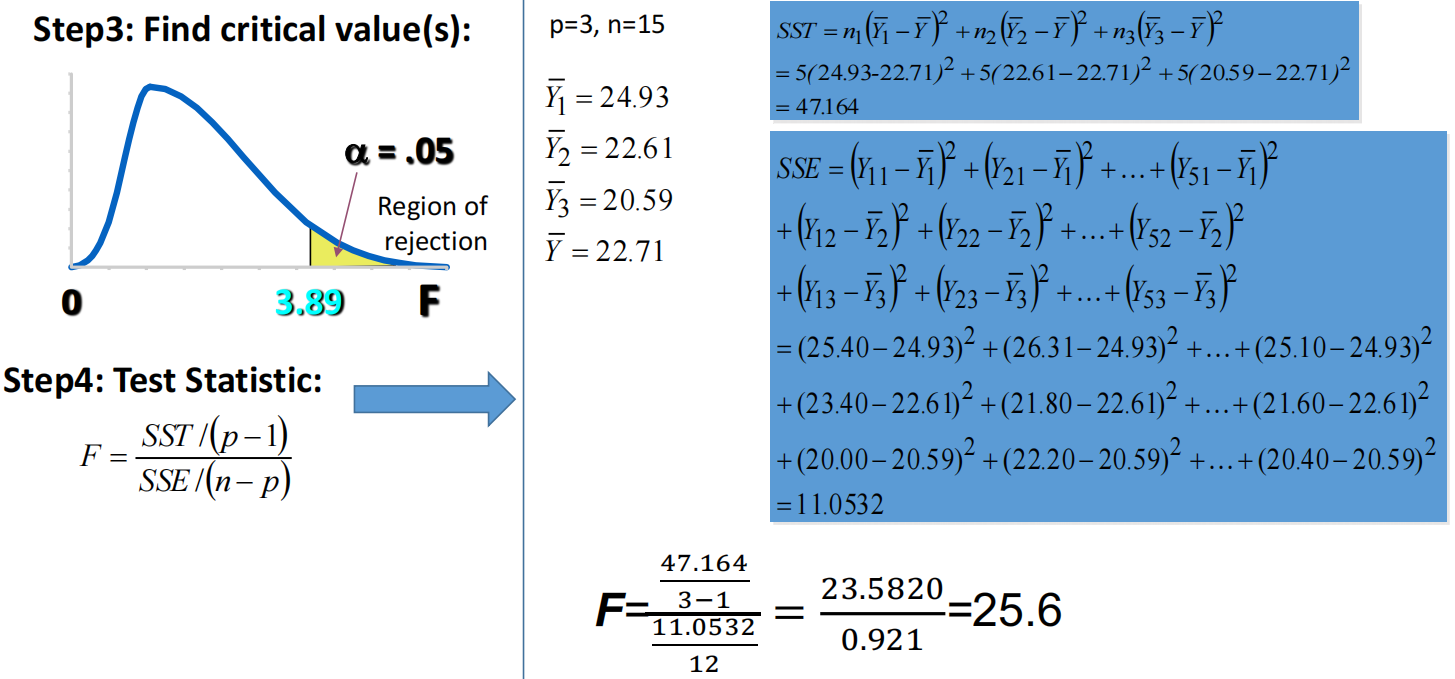


dftotal=7, 查表

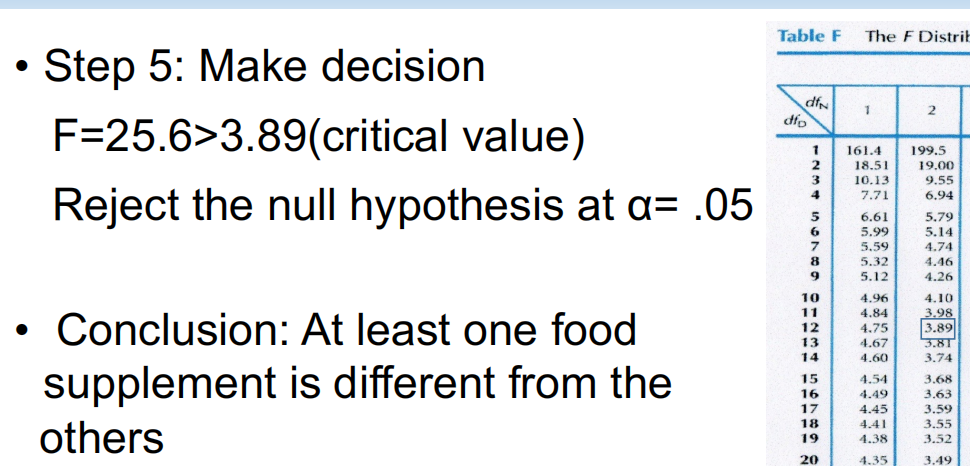




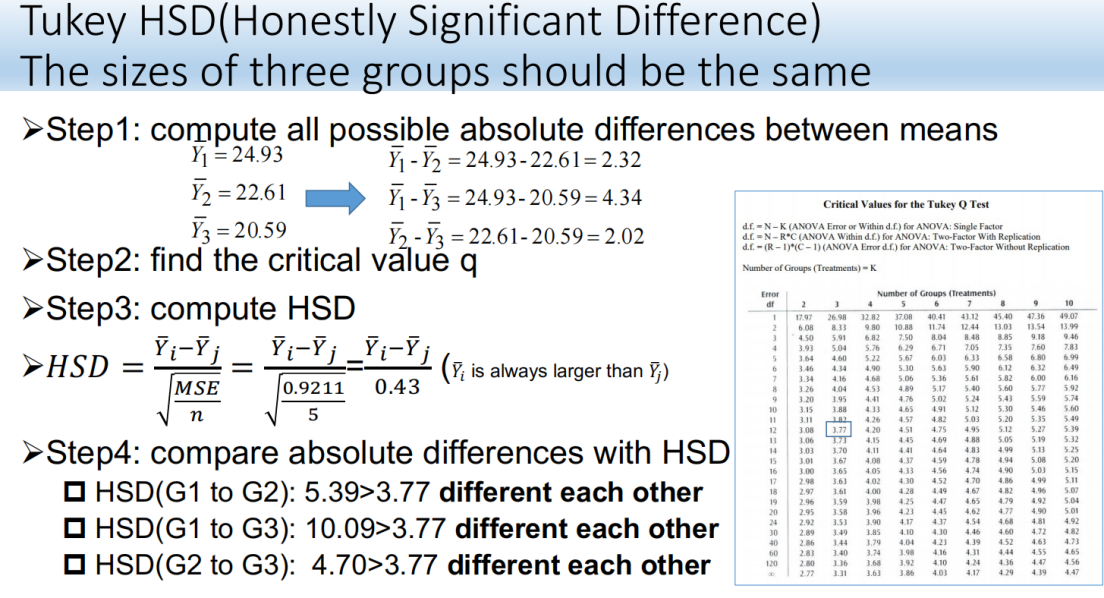




根据 V1, V2查表



**Ybar是整体的平均值，Y1bar 是 第一列的平均值， n1,n2,n3 表示这一列的数据数量。 Y12 表示第一行第二列，Y21表示第二行第一列。SSE计算的每一列的数据减去这一列的平均值然后平方相加。查表就是p-1列，n-p行**



**然后把每一列的平均值相减求绝对值，比较tukey表，如果大于就是different each other**

MST=SST/p-1

MSE=SSE/(*n*-p*)*

*F*=MST/MSE

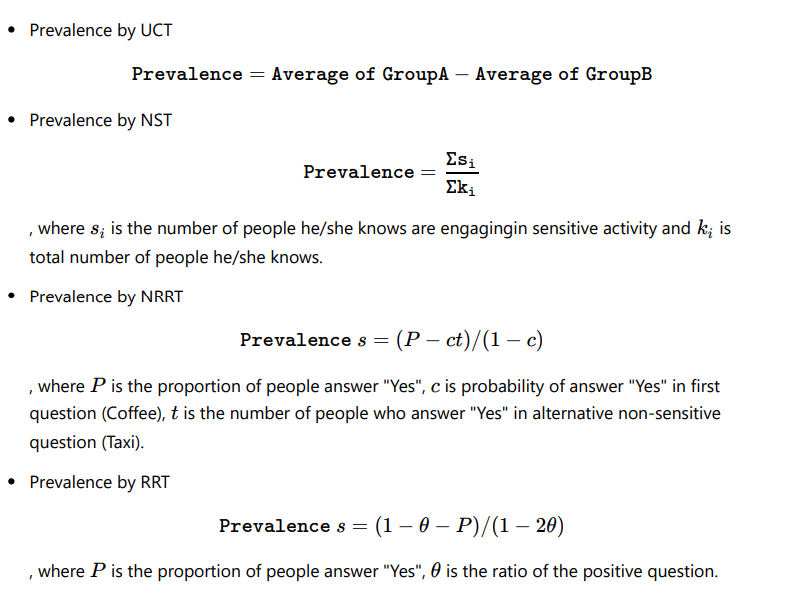
MSE = SSE / (n-p)

算出来每两列的HSD，然后查表

12行,3列 = n-p 行, p列

**Lecture 7**

DB Join, select group by having



Encryption is a process of transforming a plaintext into a ciphertext.

A public key is used to encrypt a plaintext to a ciphertext

But the ciphertext cannot be decrypted using a public key.

It needs a private key to decrypt the message

Anyone can encryption, only the receiver can decrypt (not even the sender!)

The length of the private key is recommended to have 256 bits more

**Data security**

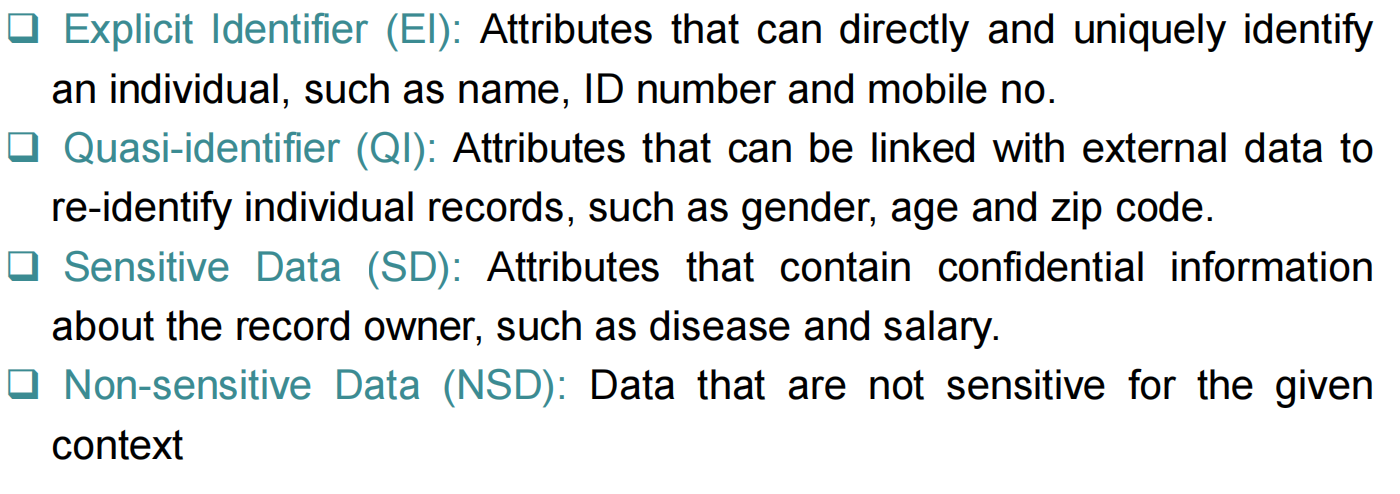
Name the advantages/disadvantages for each method(UCT, NST, NRRT and RRT) which one is the most effective.

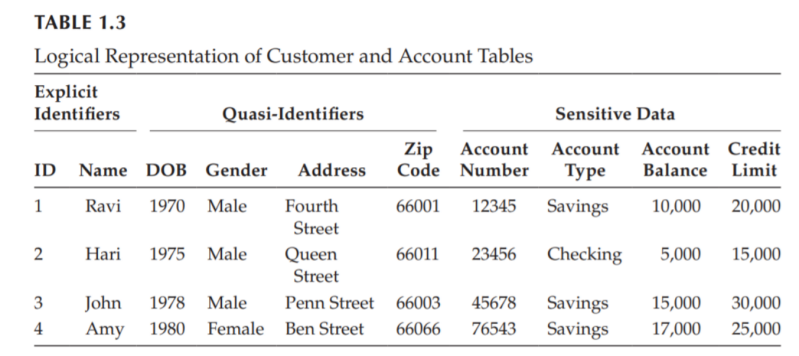
Indirect questioning can improve the data quality by increasing the perceived privacy of the respondents

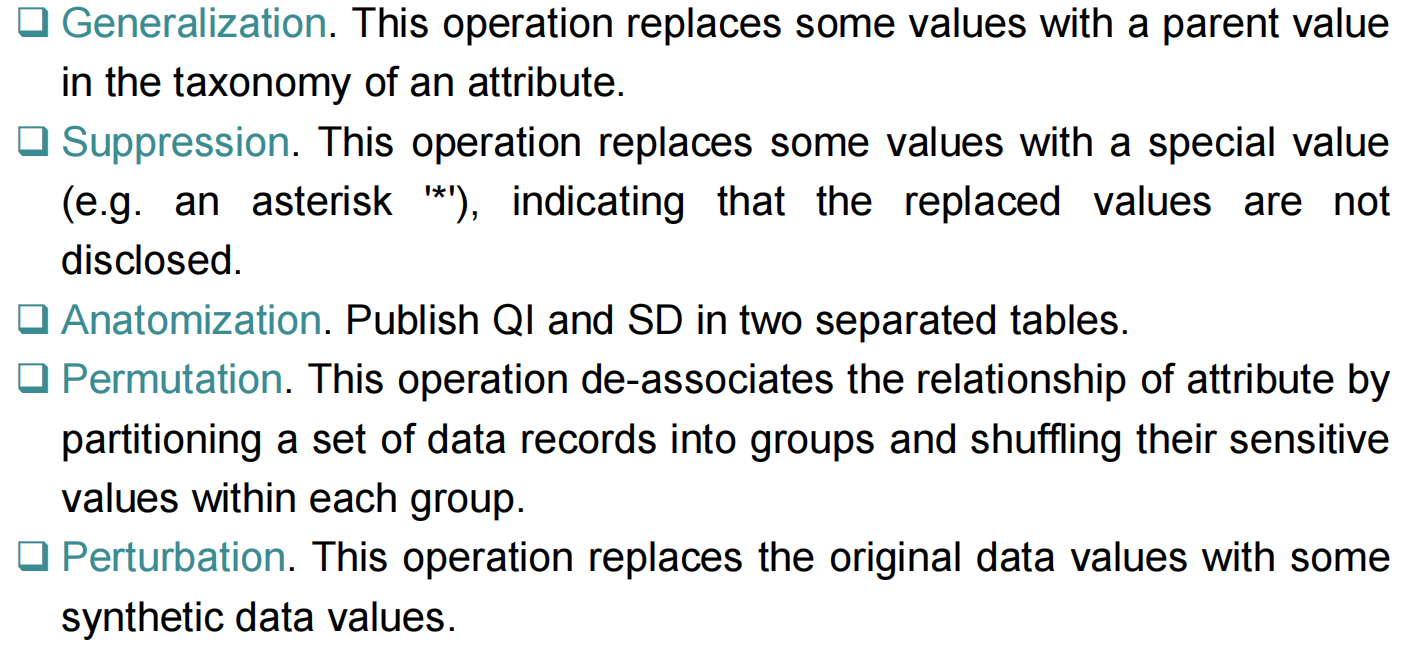
Limitations:

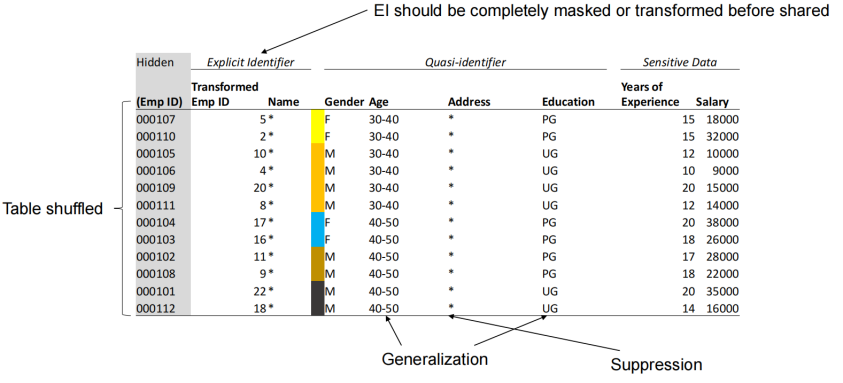
1. Hard to define you know someone in the way you know their hidden side.

2.People’s social networks may not represent the population









Sql Injection, prepareStatement

