# Interpretation of clusters by comparing centers with grand means+Bootstrap+iK-means

Contents:

- Interpretation of clusters via centers
- Center: DA perspective
- Center: CS perspective
  - Validation of center using bootstrap
  - Comparing centers using bootstrap
- K-Means: Complementary criterion and anomalous cluster
- Home work 2 & 3

# Interpretation of clusters by comparing centers with grand means+Bootstrap+iK-means

Contents:

- **Interpretation of clusters via centers**
- Center: DA perspective
- Center: CS perspective
  - Validation of center using bootstrap
  - Comparing centers using bootstrap
- K-Means: Complementary criterion and anomalous cluster
- Home work 2 & 3

# Interpretation of clusters, I

- Iris data 150x4:

| | w1 | w2 | w3 | w4 |
|---|---|---|---|---|
| 1 | 5.1 | 3.5 | 1.4 | 0.3 |
| 2 | 4.4 | 3.2 | 1.3 | 0.2 |
| 3 | 4.4 | 3.0 | 1.3 | 0.2 |
| 4 | 5.0 | 3.5 | 1.6 | 0.6 |
| 5 | 5.1 | 3.8 | 1.6 | 0.2 |
| ..... | | | | |
| 150 | 6.5 | 3.2 | 5.1 | 2.0 |

# Interpretation of clusters, 11

- Iris data: 150 x 4
- Taxon 1: 1:50
- Taxon 2: 51:100
- Taxon 3: 101:150

# Interpretation of clusters, III

- Means

| | w1 | w2 | w3 | w4 |
|---|---|---|---|---|
| T1 | 5.0060 | 3.4280 | 1.4620 | 0.2460 |
| T2 | 5.9360 | 2.7700 | 4.2600 | 1.3260 |
| T3 | 6.5880 | 2.9740 | 5.5520 | 2.0260 |
| G | 5.8433 | 3.0573 | 3.7580 | 1.1993 |

# Interpretation of clusters, IV

- Relative Difference: 100*(CMean –GMean)/GMean

| | w1 | w2 | w3 | w4 |
|---|---|---|---|---|
| T1 | -14.3297 | 12.1239 | **-61.0963** | **-79.4886** |
| T2 | 1.5859 | -9.3982 | 13.3582 | 10.5614 |
| T3 | 12.7439 | -2.7257 | **47.7382** | **68.9272** |
| G | 0 | 0 | 0 | 0 |

# Interpretation of clusters, V

Relative Difference: $\Delta = 100*(CMean - GMean)/GMean$

| | w1 | w2 | w3 | w4 |
|---|---|---|---|---|
| T1 | -14.3297 | 12.1239 | **-61.0963** | **-79.4886** |
| | | **small petal** | | |
| T2 | 1.5859 | -9.3982 | 13.3582 | 10.5614 |
| | | **pretty average** | | |
| T3 | 12.7439 | -2.7257 | **47.7382** | **68.9272** |
| | | **big petal** | | |

**Look at large** $\Delta$ or $-\Delta$!!!

# Interpretation of clusters, VI

- Relative Difference: 100*(CMean –GMean)/GMean
- | | w1 | w2 | w3 | w4 |
  |---|---|---|---|---|
  | T1 | | small petal | | |
  | T2 | | pretty average | | |
  | T3 | | big petal | | |

**Why is interpretation important?**

In the very end, clustering is accepted if related to existing domain knowledge

No sepal in interpretation – probably should be removed from clustering process

# Interpretation of clusters, VII

- Relative Difference: 100*(CMean –GMean)/GMean

| | |
|---|---|
| T1 | small petal |
| T2 | pretty average |
| T3 | big petal |

# Interpretation of clusters by comparing centers with grand means+Bootstrap+iK-means

Contents:

- Interpretation of clusters via centers
- **Center: DA perspective**
- Center: CS perspective
  - Validation of center using bootstrap
  - Comparing centers using bootstrap
- K-Means: Complementary criterion and anomalous cluster
- **Home work 2 & 3**

# What is **center,** 1: **Data analysis view**

Consider a feature over *N* entities (transposed)

$\mathbf{x} = (x_1, x_2, \ldots, x_N)$

**Def.** **Center of x is a value c satisfying equations**

$$x_i = c + e_i, \text{ for all } i = 1,2,\ldots, N$$

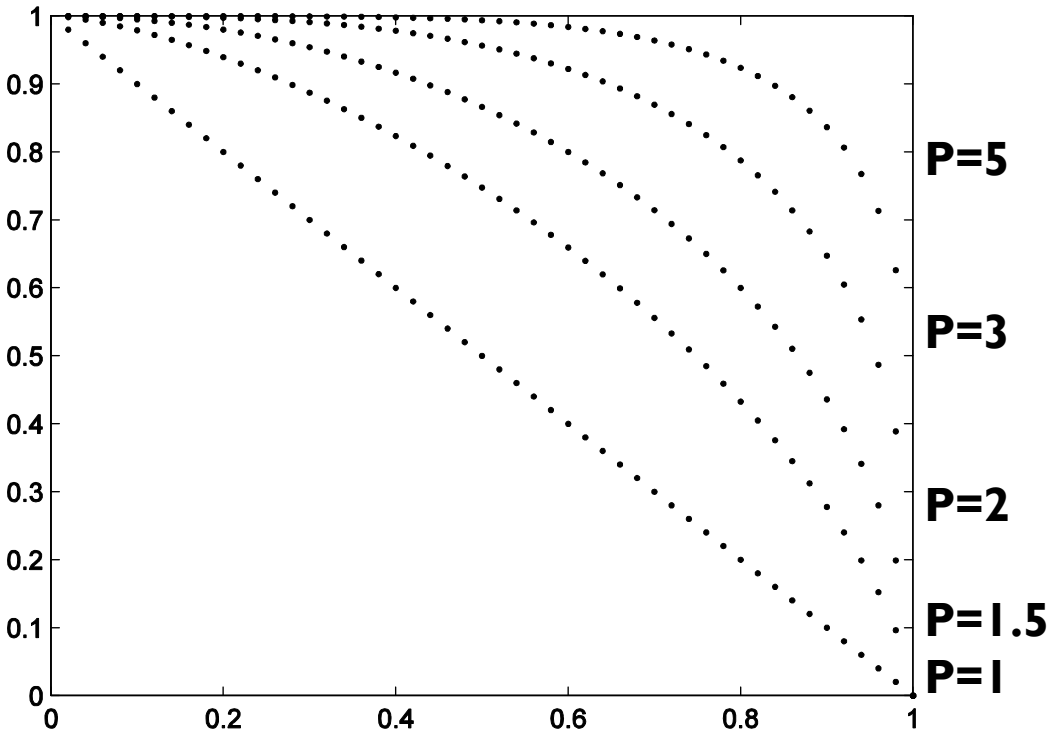
***at as small residuals*** $e_i$ ***as possible***

**Def.**

$$L_p = [|e_1|^p + |e_2|^p + \ldots + |e_N|^p\ ]/N$$

**Minkowski criterion: min** $L_p$

# Minkowski distance: curve $x^p+y^p=1$ at different p



# What is **center, 4**

## Data analysis view: Minkowski p-center ($p \geq 1$)

*Minimize* $L_p=[|c-x_1|^p+|c-x_2|^p+\ldots+|c-x_N|^p\ ]/N$

with respect to all possible c

**At different p, different solutions!**

**$L_p$ is a measure of spread of the feature around center**

# What is **center, 5**

**Data analysis view: Minkowski p-center (p ≥ 1)**

*Minimize* $L_p=[|c-x_1|^p+|c-x_2|^p+\ldots+|c-x_N|^p\ ]/N$

with respect to all possible c

**Take p=2.** Then $L_p$ is quadratic. First-order minimum condition can be applied, it leads to optimal

**c=Mean(x)!**

**At this c,**

$L_2$ **is the square of the standard deviation!**

(The minimum $L_2$ is referred to as the variance, and its square root, as the standard deviation.)

# What is **center**, 6

**At Minkowski p=2, Given x = $(x_1, x_2, \ldots, x_N)$,**

**Spread** **Standard deviation *std***

**Center** **Mean** $\bar{x} = \sum_{i=1}^{N} x_i / N$

**Consider definition**

$$std^2 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N} = \frac{\sum_{i=1}^{N} x_i{}^2 - N\bar{x}^2}{N}$$

Reformulate

$$\sum_{i=1}^{N} x_i^2 = N(\bar{x}^2 + std^2) \qquad (*)$$

# What is **center, 7**

**At Minkowski p=2,** **Given x = $(x_1, x_2, \ldots, x_N)$,**

$$std^2 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N} = \frac{\sum_{i=1}^{N} x_i^2 - N\bar{x}^2}{N}$$

**Data scatter$\equiv\sum_{i=1}^{N} x_i^2$ decomposed in a Pythagorean way**

**Data scatter$= N(\bar{x}^2 + std^2)$ (*)**

$N\bar{x}^2$ ***Explained part (by the model $x_i = c + e_i$ )***

$std^2$ ***Unexplained part***

*The greater the mean, the greater the explained part*

*Similar decompositions hold at multivariate summarizations*

# What is **center,** 8: Minkowski's *p*

| *p* | Center | Comment |
|---|---|---|
| 2 | Mean | Intuitive;<br>Gaussian<br>Sensitive to removal/addition of outliers |
| 1 | Median | Stable over removal/addition of outliers |
| ∞ | Midrange | Does not depend on the distribution shape<br>Sensitive to change of range boundary points |

Other values of *p* can be beneficial too, but we know very little of this

# Interpretation of clusters by comparing centers with grand means+Bootstrap+iK-means

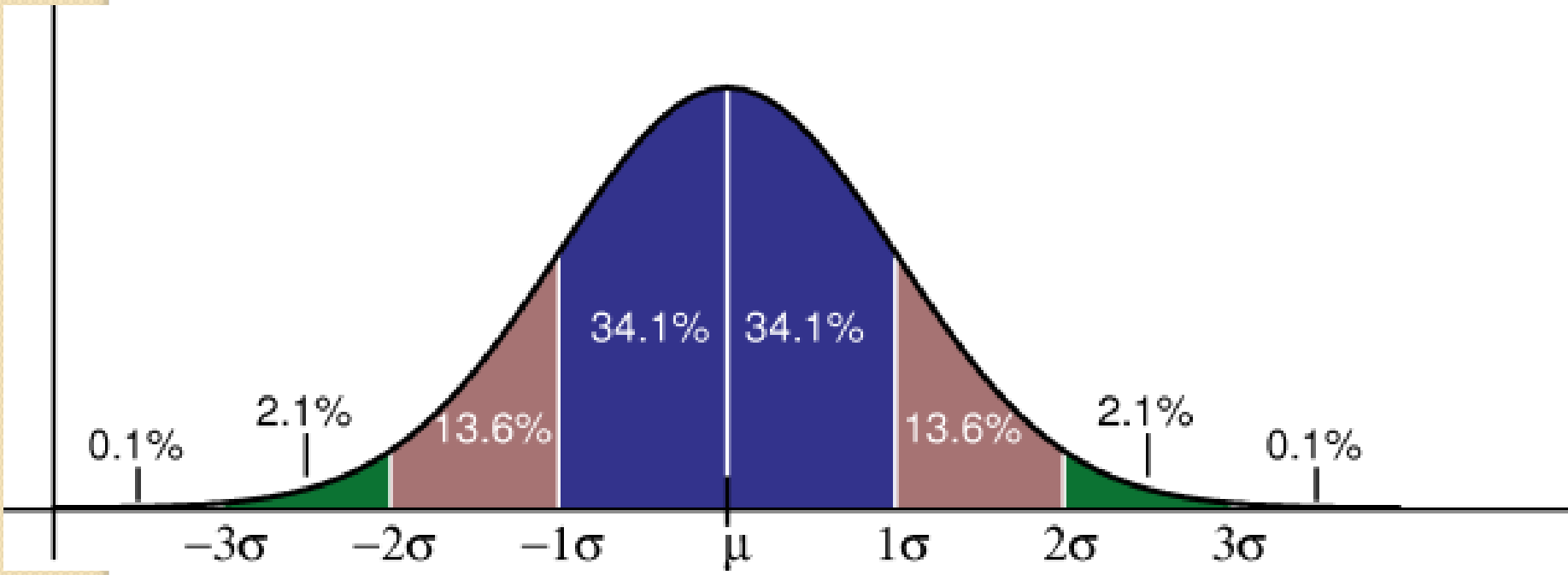
Contents:

- Interpretation of clusters via centers
- Center: DA perspective
- **Center: CS perspective**
  - Validation of center using bootstrap
  - Comparing centers using bootstrap
- K-Means: Complementary criterion and anomalous cluster
- **Home work 2 & 3**

# Week2. What is **center:** Probabilistic perspective

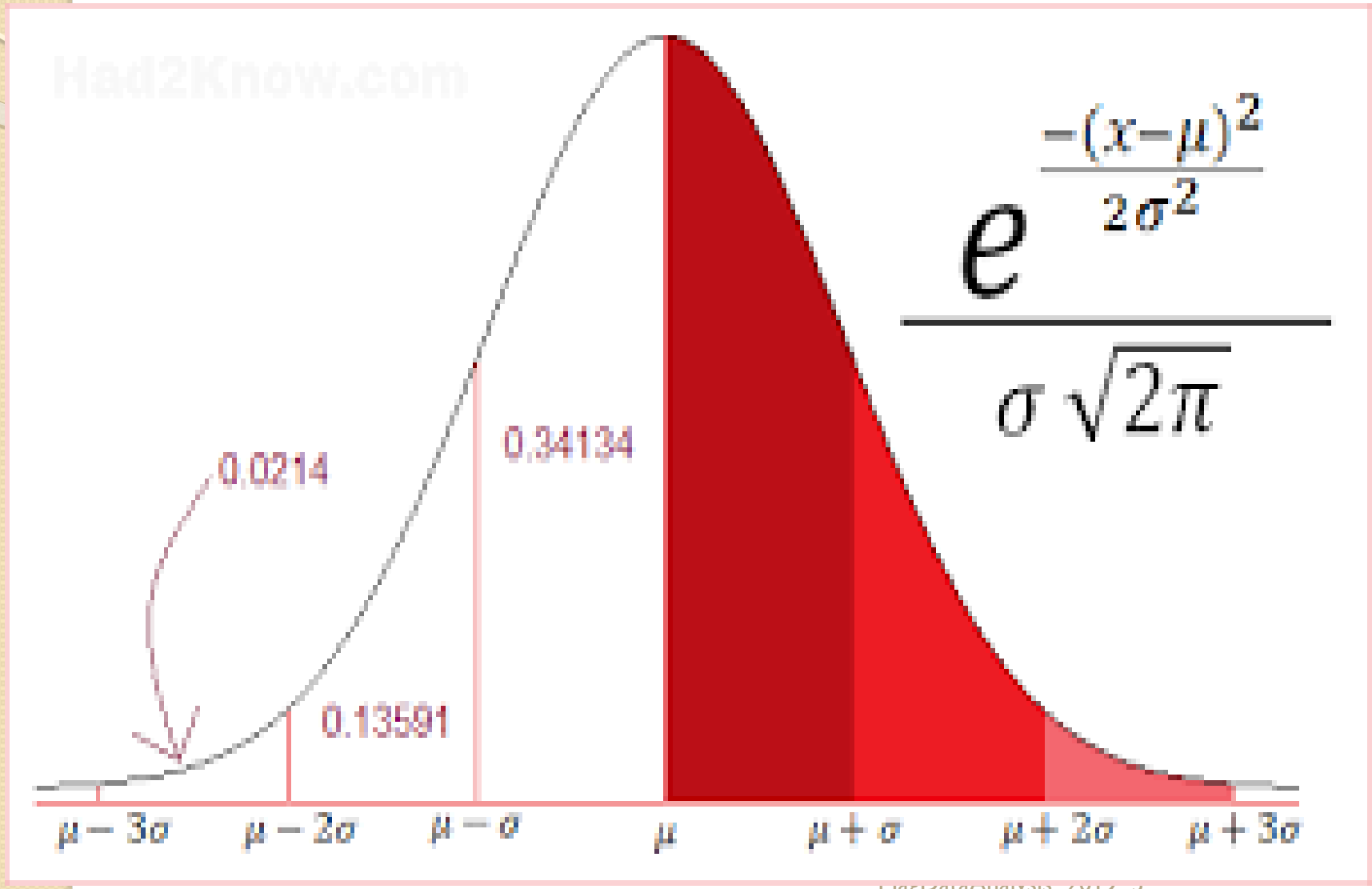
## Gaussian density function

$$p(x)=C\ exp[-(x-a)^2/2\sigma^2]$$



# Week2. What is **center:** Probabilistic perspective

## Gaussian density function



# What is **center:** Probabilistic perspective

**Estimates of parameters in the Gaussian density**

$$p(x)=Cexp[-(x-a)^2/2\sigma^2]$$

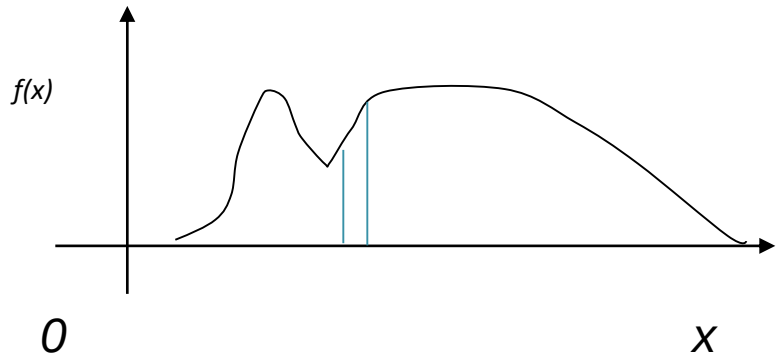
*Mean, of a:* $$m=\frac{\sum_{i=1}^{N} x_i}{N}$$

*Variance $\sigma^2$ (Standard deviation squared)*

$$s^2=\frac{1}{N}\sum_{i=1}^{N}(x_i-a)^2 \quad \textbf{or}$$

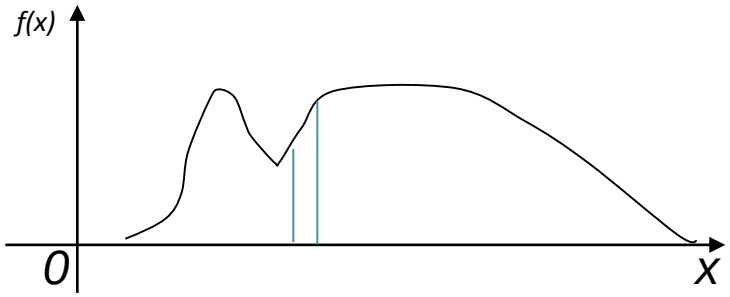
$$s^2=\frac{1}{N-1}\sum_{i=1}^{N}(x_i-m)^2$$

# Classic Statistics: Center, 1




**Probabilistic view**: observed *M* values of feature $x=(x_i)$ – set of *M* independent random variables with the same density function *f(x)*.

# Classic Statistics: Center, 2



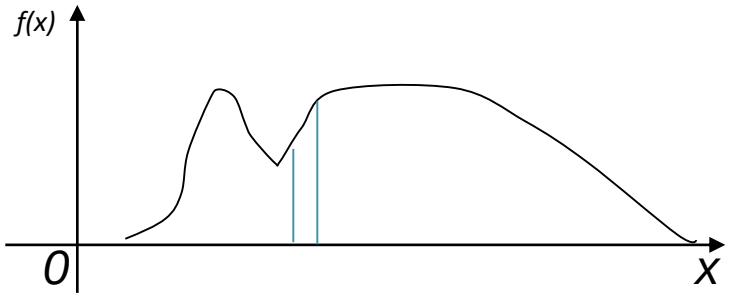

Mathematical Expectation of a random variable

$$ME(f) = \int_{-\infty}^{+\infty} \boldsymbol{x f(x) dx};$$

Variance $\sigma^2$, $ME([f(x)-ME(f)]^2)$.

[Alas, density of the summary variable $x_1 + x_2$ is not *2f(x), but rather more complex* ***convolution*** *of f(x)* with itself]

# Classic Statistics: Center, 3




Mathematical Expectation of a random variable

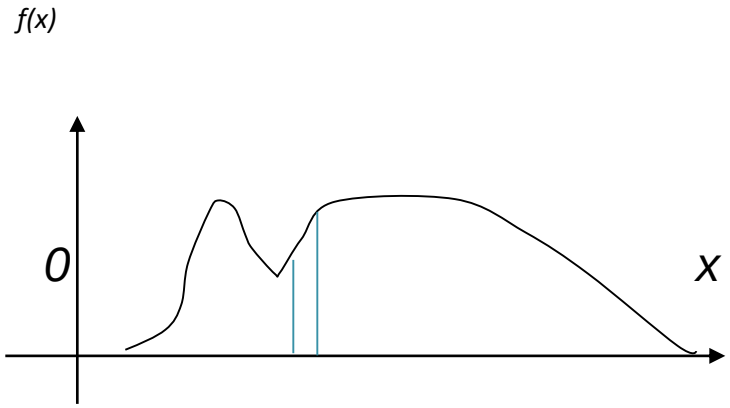
$$ME(f) = \int_{-\infty}^{+\infty} \boldsymbol{x f(x) dx};$$

Variance $\sigma^2$, $ME([f(x)-ME(f)]^2)$.

Relation to DA: average

$\bar{x} = \sum_{i=1}^{M} x_i / M$ is unbiased estimate of $\mu$

# Classic Statistics: Center, 4




Unidimensional data $x=(x_i)$ $(i=1,2,\ldots,M)$: $M$ independent random variables with the same density $f(x)$.

**Central Limit Theorem:**

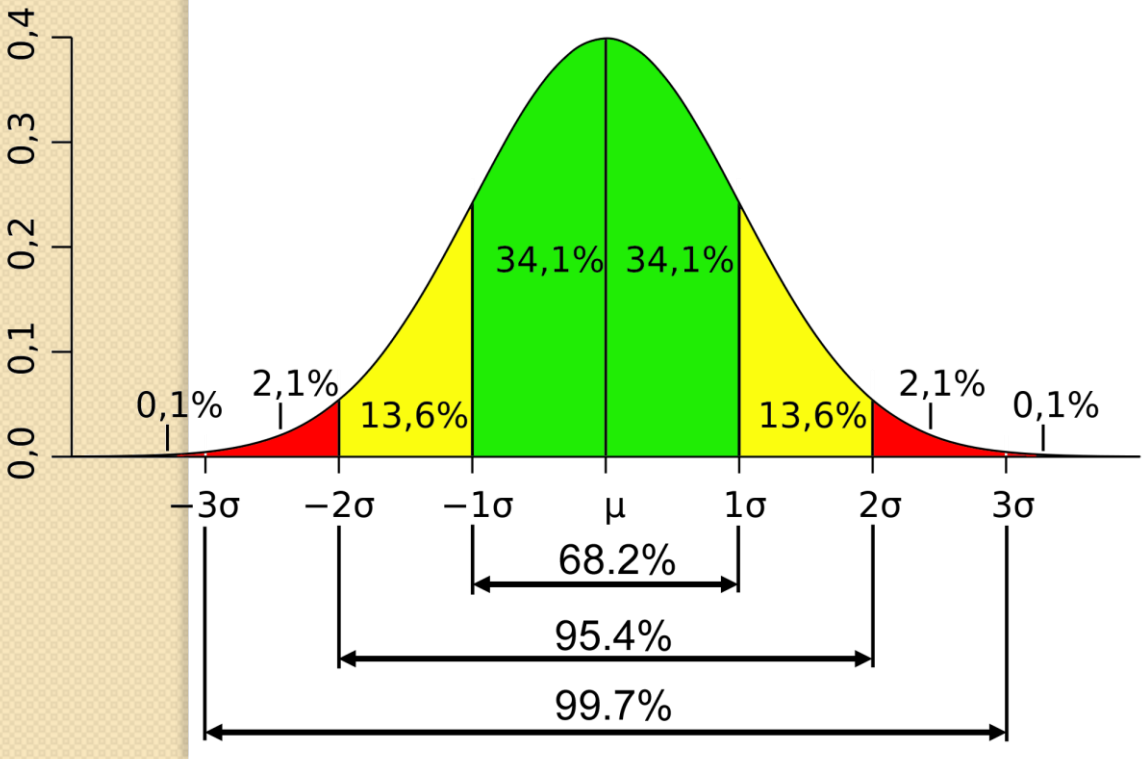
Density function of the sum $x_1 + x_2 + \ldots + x_M$ converges to the Gaussian density function with mathematical expectation $\mu$ and variance $\sigma^2$ (at $M \to \infty$).

Density function of the average $\bar{x} = (x_1 + x_2 + \ldots + x_M)/M$ converges to a Gaussian with $ME=\mu$ and variance $\sigma^2/M$.

# Classic Statistics: Center, 5

- Density function of the average is approximately Gaussian $N(\mu, \sigma/\sqrt{M})$
- Thus, central interval to account for 95% of the area:

$$[\mu - 1.96\sigma/\sqrt{M}, \mu + 1.96\sigma/\sqrt{M}]$$



# Interpretation of clusters by comparing centers with grand means+Bootstrap+iK-means

Contents:

- Interpretation of clusters via centers
- Center: DA perspective
- **Comparing Centers: CS perspective**
  - Validation of center using bootstrap
  - Comparing centers using bootstrap
- K-Means: Complementary criterion and anomalous cluster
- **Home work 2 & 3**

# Classic Statistics: Comparing Centers

Given: average $\mu_1$ of sample of $M_1$ entities and average $\mu_2$ of an independent sample of $M_2$ entities.

Hypothesis: $\mu_1 = \mu_2$ or $\mu = 0$ where $\mu = \mu_1 - \mu_2$.

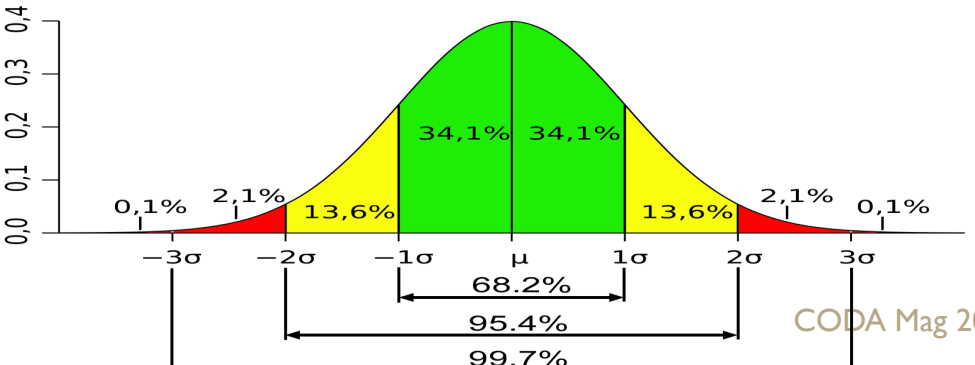
Test is based on: $\mu$ has density $N(0,\sigma)$ where

$\sigma = \sigma_1 / \sqrt{M_1} + \sigma_2 / \sqrt{M_2}$.

- Central interval for 95% area:

A= $[\mu - 1.96\sigma, \mu + 1.96\sigma]$.

- If 0 does not belong to A, the hypothesis is rejected at 95% confidence level



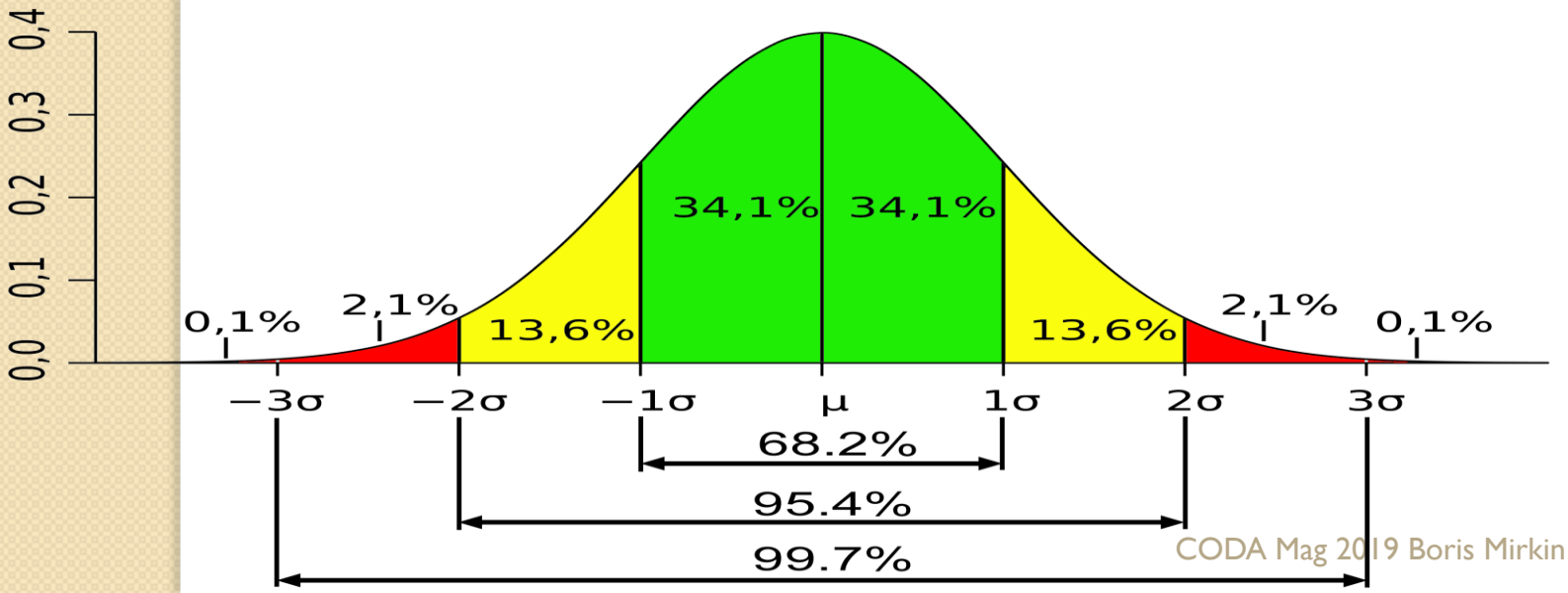
# Interpretation of clusters by comparing centers with grand means+Bootstrap+iK-means

Contents:

- Interpretation of clusters via centers
- Center: DA perspective
- Center:  CS perspective
  - **Validation of center using bootstrap**
  - **Comparing centers using bootstrap**
- K-Means: Complementary criterion and anomalous cluster
- **Home work 2 & 3**

# Classical statistics: Bootstrap

- **Bootstrap: Computational estimate of density function and its central 95% confidence interval A**
- In the case of hypothesis $\mu = \mu_1 - \mu_2 = 0$, the density of $\mu$ is estimated.
- If $0 \notin A$, the hypothesis is rejected at 95% confidence level.



# Bootstrapping for comparing means

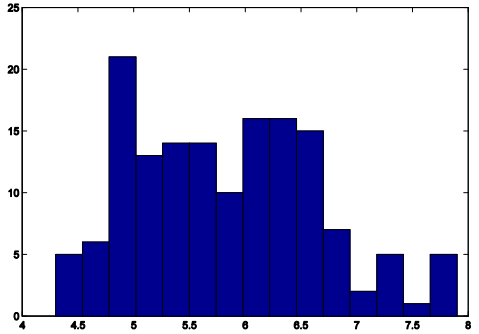
Bootstrap

# Week 2. 1D: Part5 Computational validation of Mean using bootstrap 1

**Consider a feature, say x=iris(:,1) % 1st column of Iris data**

**Its histogram hist(x,15):**

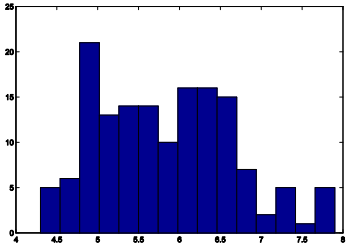
**rather far from Gaussian**




**Its mean m = 5.8433**

**std=0.8253**

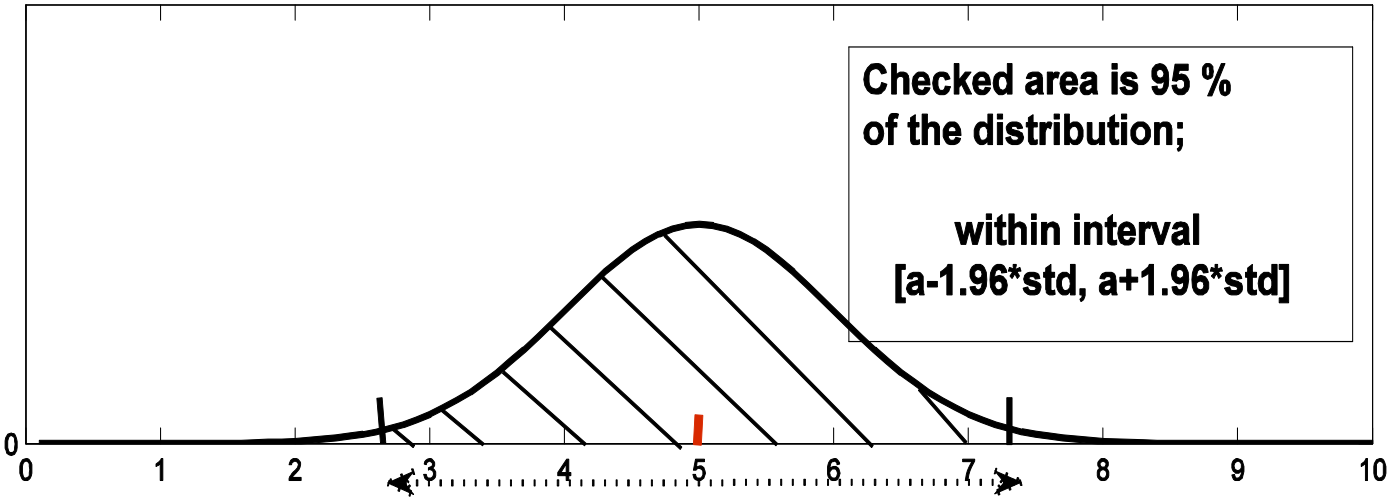
# Computational validation of Mean using bootstrap



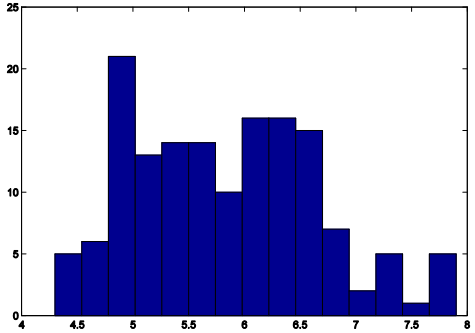
**Plausible boundaries for mean?**

One way to go: using classical math statistics

**According to the Central Limit Theorem applied to random independent samples, the density function of $\bar{x}$ approximates Gaussian distribution with a=5.8433 and σ=0.8253/√N.**



# Computational validation of Mean using bootstrap 2




**Consider a feature, say x=iris(:,1)**

**Its mean m = 5.8433, std=0.8253**

**Plausible boundaries for m? 95%**

One way to go: using classical math statistics

**Assume x is a random independent sample from a Gaussian distribution with a=5.8433 and σ=0.8253:**

**m is Gaussian too, with a=m and σ=std/$N^{1/2}$ *(N=150)***

**Therefore, with 95% confidence**

**Lb= a - 1.96*std /$N^{1/2}$= 5.7108**

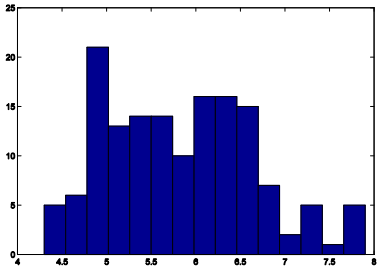
**Rb= a + 1.96*std /$N^{1/2}$= 5.9759**

**Conclusion:**

**m within [5.7108, 5.9759] with confidence 95% (as approximately Gaussian)**

# Computational validation of Mean using bootstrap ,3

**Plausible boundaries for m?**



Another way to go: using computing power

## Bootstrap

**Multiple entity samples of same size *N* (with replacement)**

**Meaning: indices are sampled to form a try**

**MatLab:**

```
>> M=4;K=3;  r=ceil(M*rand(M,K))
```

```
     1    4    4
r=   3    1    4
     3    1    3
     2    3    1
```

**M=4, the number of entities**

**K=3, the number of tries**

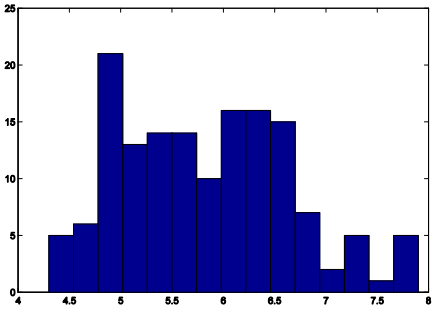
**First try: entities 1, 3 (twice), 2**

**(entity 4 is missed: why?)**

# Computational validation of Mean using bootstrap 4

**Consider a feature, say x=iris(:,1)**

**Its mean = 5.8433, std=0.8253**




**Plausible boundaries for m?**

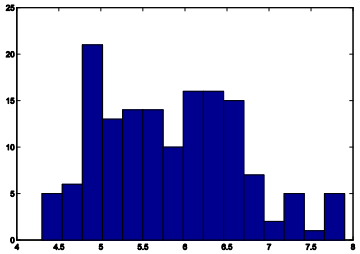
**Bootstrap**

**MatLab:**

```
>> M=150;K=5000;  r=ceil(M*rand(M,K));
>> xr=x(r);
>> mx=mean(xr);
```

**This gives K=5000 means of random samples of x**

# Computational validation of Mean using bootstrap 5

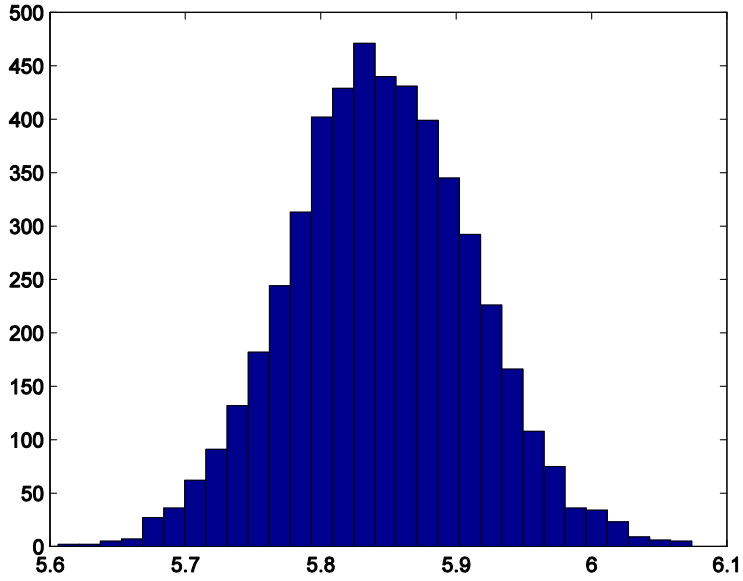


**Plausible boundaries for m?**

## Bootstrap

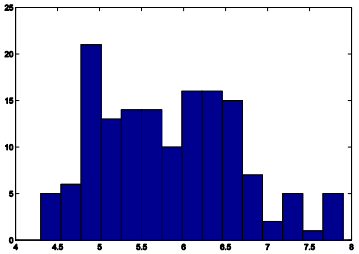
```
>> M=150;K=5000;  r=ceil(M*rand(M,K));
>> xr=x(r); mr=mean(xr);
```

**Histogram of K=5000 means**




**Quite Gaussian, n'est ce pas?**

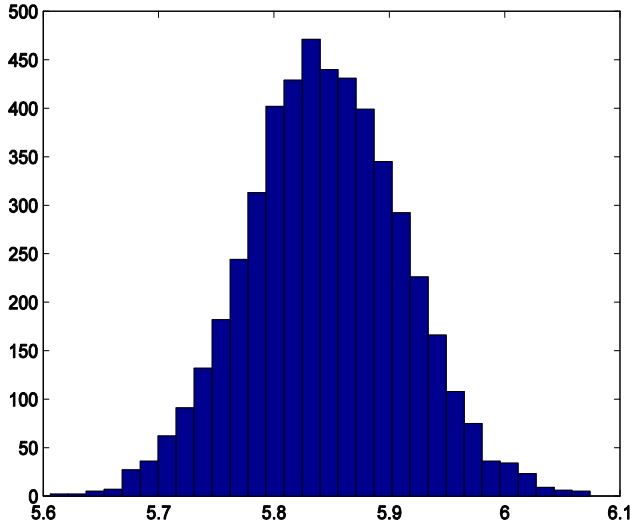
# Computational validation of Mean using bootstrap 6



**Feature x=iris(:,1); = 5.8433, std=0.8253**

**Plausible boundaries for m?**

**Bootstrap Histogram of K=5000 means mr**




## A. Pivotal method

**(95% confidence)**

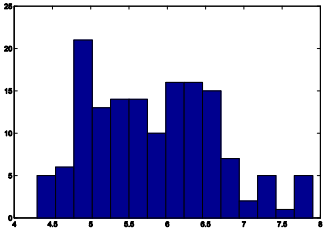
**Assume mr be Gaussian**

```
>> mmr=mean(mr); % 5.8444
>> smr=std(mr); %   0.0675
>> lbp=mmr-1.96*smr; %  5.7121
>> rbp=mmr+1.96*smr; % 5.9767
```

# Computational validation of Mean using bootstrap 7

**Feature x=iris(:,1); = 5.8433, std=0.8253**

**Plausible boundaries for m?**

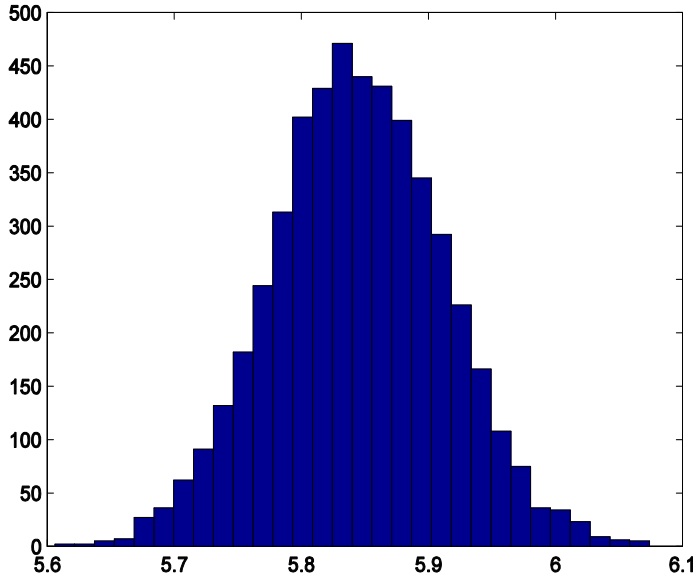


**Bootstrap Histogram of K=5000 means mr**

## B. Non-pivotal method

**(95% confidence)**

**Take 2.5% and 97.5% percentiles as the boundaries**




**1% of 5000 is 50;**

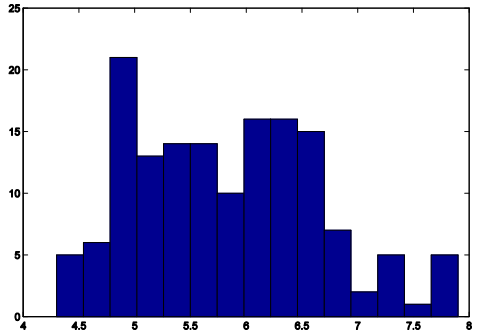
**2.5% is 125; 97.5% is 4875**

```
>> smr=sort(mr); % sorting
>> lbn=somr(126); %    5.7120
>> rbn=somr(4875); % 5.9773
```

# Computational validation of Mean using bootstrap 8

**Consider a feature, say x=iris(:,1)**

**Its mean = 5.8433, std=0.8253**




**Plausible boundaries for m with confidence 95%?**

**Three different methods – m must be within :**

- **[5.7108, 5.9759] (under Gaussian assumption)**
- **[5.7121, 5.9767] (Bootstrap pivotal)**
- **[5.7120, 5.9773] (Bootstrap non-pivotal)**

**with 95% confidence**

**Very much similar…**

# Comparing means using Bootstrap, I

- Compare **mean** Sepal lengths in Taxa 2 and 3:

  - Bootstrap distributions of K trial means in T1 and in T2
  - Differences D=Mean(T1)–Mean(T2) over all K trials
  - 95% confidence interval A for D
  - Checking whether zero is in A or not. If not, one Mean is greater than the other.

# Comparing means using Bootstrap, 11

- Compare mean Sepal length in Taxa 2, 3, All set:

```
>> n=150;m=5000;r=ceil(n*rand(n,m));
>> x=iris(:,1); xr=x(r); mr=mean(xr);%All set
```

In Taxa t{1}, t{2}, t{3}:

```
>> for k=1:5000; y=r(:,k); p=ismember(y,t{2});
n2=sum(p);y2=sum(p.*x(y));m2(k)=y2/n2;end
>> for k=1:5000; y=r(:,k); p=ismember(y,t{3});
n3=sum(p); y3=sum(p.*x(y));m3(k)=y3/n3;end
```

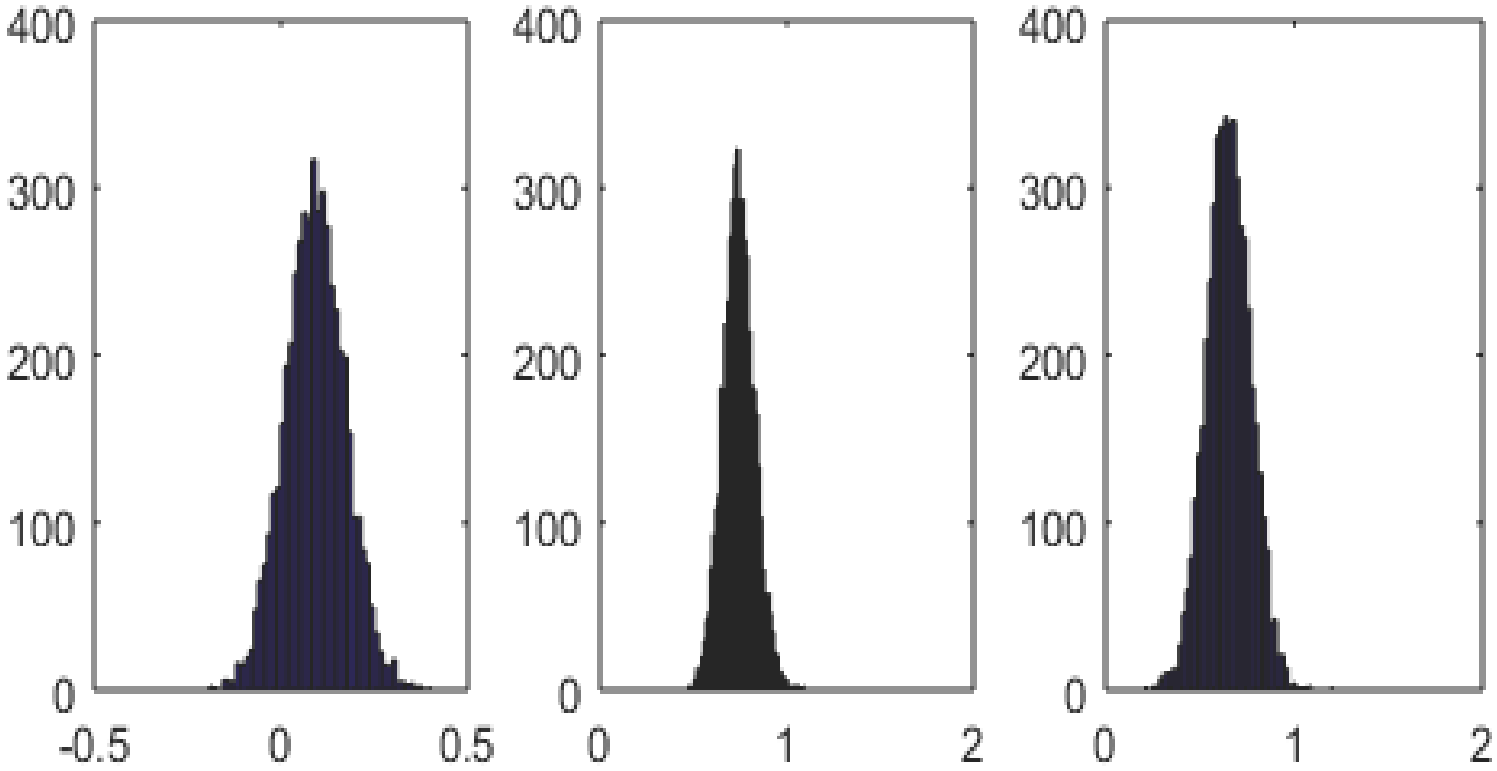
**mr, m2, m3 – 5000-strong bootstrap means**

# Comparing means using Bootstrap,III

- Compare mean Sepal length in Taxa 2, 3, All set:

m2-mr m3-mr m2 -m3

bootstrap means



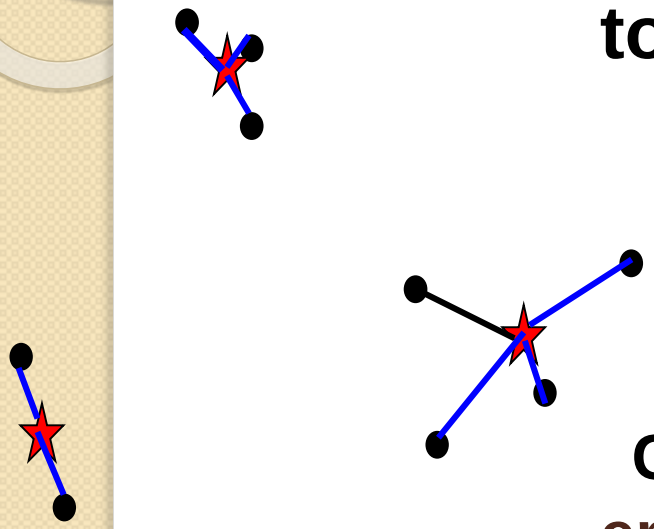
# Interpretation of clusters by comparing centers with grand means+Bootstrap+iK-means

Contents:

- Interpretation of clusters via centers
- Center: DA perspective
- Center: CS perspective
  - Validation of center using bootstrap
  - Comparing centers using bootstrap
- **K-Means: Complementary criterion and anomalous cluster**
- Home work 2 & 3

# Clustering with K-Means

## K-Means criterion:



**Find partition *S* and centers *c* to minimize:**

$$D(S,c) = \sum_{k=1}^{K} \sum_{i \in S_k} d(i, c_k)$$

**Criterion: Sum of distances between entities and centers of their clusters**

**Distance d(.,.) (squared Euclidean):**

X= [ 1, 2, -2]

Y= [ 1, -1, -1]

X-Y=[1-1, 2-(-1), -2-(-1)]=[0, 3, -1]

$d(X,Y)=\langle X - Y, X - Y\rangle = 0^2+3^2+(-1)^2 = 10$

# Pythagorean decomposition

- K-Means criterion:

$$D(S,c) = \sum_{k=1}^{K} \sum_{i \in S_k} \sum_{v=1}^{V} (y_{iv} - c_{kv})^2 =$$

$$= \sum_{k=1}^{K} \sum_{i \in S_k} \sum_{v=1}^{V} (y_{iv}^2 - 2 y_{iv} c_{kv} + c_{kv}^2) =$$

$$= \sum_{i=1}^{N} \sum_{v=1}^{V} y_{iv}^2 - \sum_{k=1}^{K} N_k < c_k, c_k > = T - F(S,c)$$

- **T = F(S,c) + D(S,c)**

- *Data_Scatter = "Explained Part"+"Unexplained Part"*

# Clustering with K-Means, 2

## K-Means computation converges

**Minimize *D(S,c)*** **alternatingly:**

$$D(S,c) = \sum_{k=1}^{K} \sum_{i \in S_k} d(i, c_k)$$

**over *S* and *c*.**

**$\text{Min}_S$ *D(S, c)*:**
- Clusters update

**$\text{Min}_c$ *D(S, c)*:**
- Centers update

**Distance (Squared Euclidean):**

X= [ 1, 2, -2]
Y= [ 1, -1, -1]
X-Y= [ 0, 3, -1]

*D(S,c)* decreases at each step:

**Convergence – why? (QUIZ)**

**d(X,Y)=<X - Y, X - Y>=$0^2+3^2+(-1)^2$ = 10**

# Complementary clustering criterion

**K-Means minimizes:**

$$D(S,c) = \sum_{k=1}^{K} \sum_{i \in S_k} d(i, c_k)$$

over *S* and *c*.

**Data scatter** $\sum_{i,v} y_{iv}^2$ **=**

*= D(S,c) + F(S,c)*

**Data scatter is constant while partitioning**

**Complementary criterion:**

**Maximize**

$F(S,c) = \sum_{k=1}^{K} N_k < c_k, c_k >$

**$N_k$ is the number of entities in $S_k$**

**$<c_k, c_k>$ -** ***Euclidean squared distance between 0 and $c_k$***

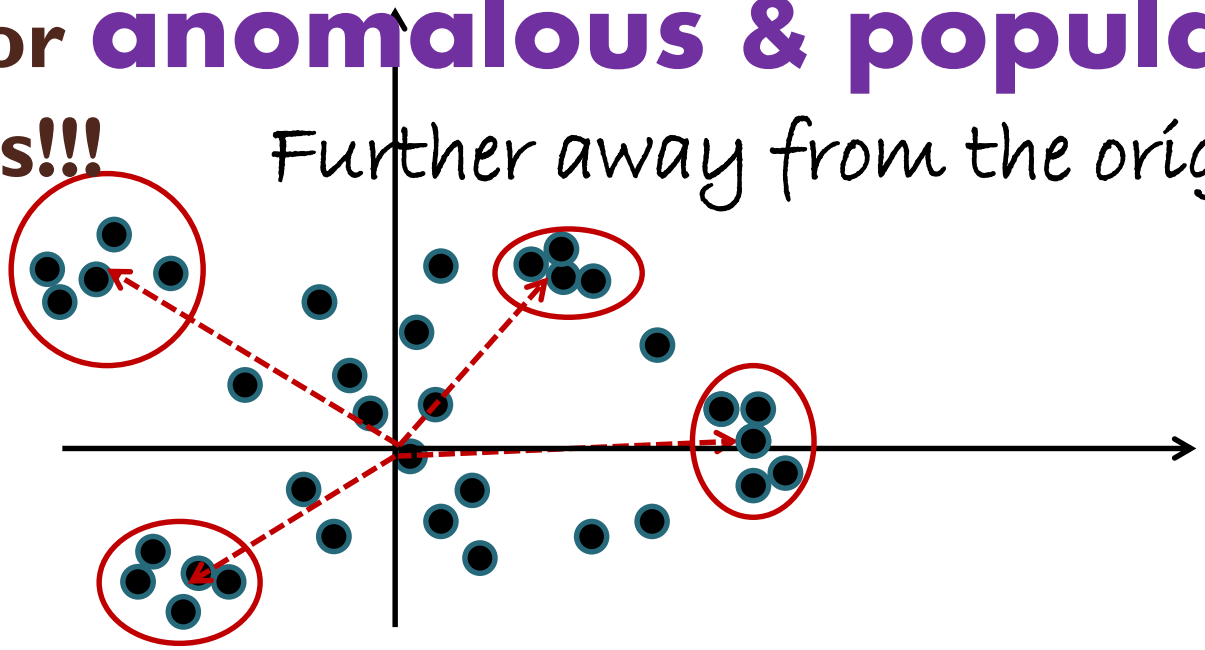
# K-Means complementary criterion

**Maximize** $F(S, c) = \sum_{k=1}^{K} |S_k| < c_k, c_k >$

**Pre-center data:** **0 is grand mean**

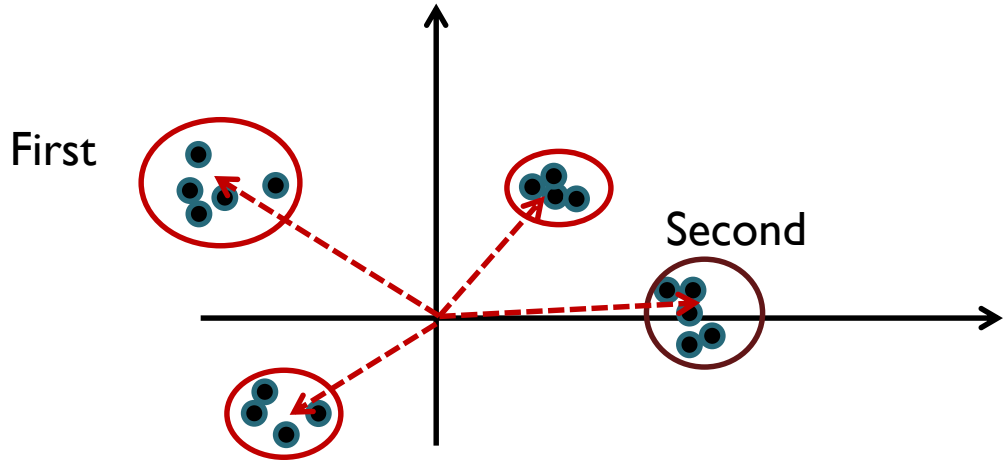
**$<c_k, c_k>$ -** ***Euclidean squared distance*** **0 to $c_k$**

**Look for anomalous & populated clusters!!!**



# Determining the Number of clusters: Two approaches

- A) Extract clusters one-by-one: find an anomalous cluster, then remove it; etc.



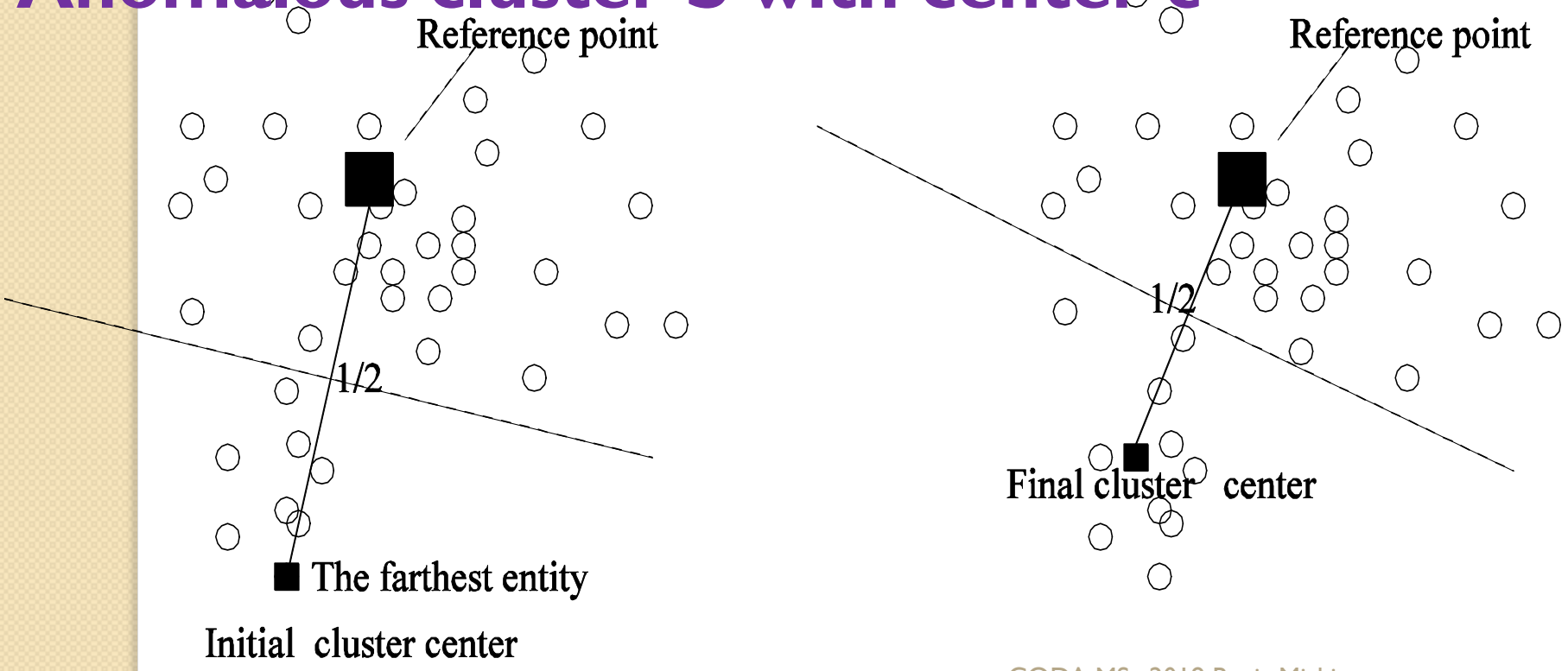

- B) Determine centers/objects, both most distant and representative, in parallel; then K-means

# Finding an ANOMALOUS Cluster (Mirkin 1998, Chiang&Mirkin 2010)

**Anomalous cluster S with center c:**

**max $|S|< c, c >$**

**0 is Reference point (grand mean). Build Anomalous cluster S with center c**



# Finding an Anomalous cluster

1. **Initial** center **c** is object, farthest away from **0**.

2. **Cluster update:** If $d(y_i,c) < d(y_i,0)$, *a*ssign $y_i$ to **S**.

3. **Centroid update:** Within-**S** mean **c'** if **c'** $\neq$ **c**. Go to 2 with **c** $\Leftarrow$ **c'**. Otherwise, halt.

# Anomalous cluster and K-Means

## Anomalous Cluster is (almost) K-Means

**up to:**

**(i) the number of clusters K=2: the "anomalous" one and the "main body" of entities around 0;**

**(ii) center of the "main body" cluster is forcibly always at 0;**

**(iii) natural initialization: $c^0$ is at entity which is the farthest away from 0.**

# ik-means

1. Pre-center the data matrix to grand-mean, set threshold for minimal cluster size t ( =1 by default).
2. Find Anomalous cluster and store its center and size.
3. Remove the Anomalous cluster from data set. Halt if the dataset gets empty, else: go to 2.
4. Initialize k-means with centers of those anomalous clusters whose size $\geq$ t

**Extensive experimentation by Chiang and Mirkin (J of CI, 2010) demonstrated superiority of iK-Means over competition**

# Clustering with K-Means

## Anomalous cluster

**Anomalous Cluster applied to Iris dataset just centered (no further normalization):**

**Initial center: the furthest away entity 132**

**c0=(1.8567 -0.4573 3.1420 1.1007)**

**- 27 entities are closer to c0 than to 0; their center**

**c1=(1.1641 0.0390 2.1716 0.9377)**

**- 47 entities are closer to c1 than to 0; their center**

**c2=(0.8865 -0.0361 1.8399 0.8156)**

**- 58 entities are closer to c2 than to 0; their center**

**c3=(0.7618 -0.0729 1.7023 0.7593)**

**- 60 entities are closer to c3 than to 0; their center**

**c4=(0.7600 -0.0773 1.6737 0.7407)**

**convergence**

# Clustering with K-Means

## Anomalous cluster iterated

**Iris dataset just centered (no further normalization)**

**AC ITERATIVELY to those yet unclustered:**

| | | | | | |
|---|---|---|---|---|---|
| AnomClus 1 | | | | | **What are these?** |
| 60 entities | c=(0.7600 | -0.0773 | 1.6737 | 0.7407) | **34.6%** |
| AnomClus 2 | | | | | |
| 50 entities | c=(-0.8373 | 0.3707 | -2.2960 | -0.9533) | **51.5%** |
| AnomClus 3 | | | | | |
| 31 entities | c=(-0.1853 | -0.4122 | 0.3872 | 0.0684) | **1.6%** |
| AnomClus 4 | {67} | singleton | | | **0.2%** |
| AnomClus 5 | 5 entities | | | | **0.6%** |
| AnomClus 6 | {98} | singleton | | | Less 0.1% |
| AnomClus 7 | {99} | singleton | | | Less 0.1% |
| AnomClus 8 | {55} | singleton | | | Less 0.1% |

# Clustering with K-Means

## Iterated Anomalous cluster

**Anomalous Cluster ITERATED :**

| | |
|---|---|
| AnomClus 1 | |
| 60 entities | **34.6%** |
| AnomClus 2 | |
| 50 entities | **51.5%** |
| AnomClus 3 | |
| 31 entities | **1.6%** |
| AnomClus 4 | **0.2%** |
| AnomClus 5 | **0.6%** |
| AnomClus 6 | Less 0.1% |
| AnomClus 7 | Less 0.1% |
| AnomClus 8 | Less 0.1% |

**Because the algorithm is local**

**Maximize total contribution to data scatter**

$$F(S,c) = \sum_{k=1}^{K} N_k < c_k, c_k >$$

$N_k = |S_k|$

$<c_k, c_k>$ - *squared distance between 0 , $c_k$*

**Contribution of a cluster to the data scatter (WHY?)**

$$100 * N_k < c_k, c_k > / \sum_v x_{vk}^2$$

# Anomalous cluster and iK-Means

**Anomalous Clusters ITERATIVELY :**

| | |
|---|---|
| AnomClus 1 | |
| 60 entities | **34.6%** |
| AnomClus 2 | |
| 50 entities | **51.5%** |
| AnomClus 3 | |
| 31 entities | **1.6%** |
| AnomClus 4 | **0.2%** |
| AnomClus 5 | **0.6%** |
| AnomClus 6 | Less 0.1% |
| AnomClus 7 | Less 0.1% |
| AnomClus 8 | Less 0.1% |

**Intelligent K-Means**

0. Standardize data by centering and, if needed, normalization
1. Iteratively find all Anomalous clusters
2. Choose the largest K among them or, if K is difficult to specify, apply threshold on the minimum cardinality of a cluster (say $N_k$<10 for Iris).
3. Apply K-Means initialized at centers of largest Anomalous Clusters

# Clustering with K-Means at IRIS

**1. Anomalous Cluster applied to Iris dataset just centered (no further normalization) ITERATIVELY to those unclustered:**

AnomClus 1
60 entities c1=(0.7600 -0.0773 1.6737 0.7407) **34.6%**
AnomClus 2
50 entities c2=(-0.8373 0.3707 -2.2960 -0.9533) **51.5%**
AnomClus 3
31 entities c3=(-0.1853 -0.4122 0.3872 0.0684) **1.6%** **Etc.**

**2. Leave those $N_k$ >10: then $K$=3 and initial centers above.**

**3. Apply K-Means:**

**14+3=17 errors**

| Taxa: | T1 | T2 | T3 | Total |
|---|---|---|---|---|
| Cl3 | 0 | 47 | 14 | 61 |
| Cl2 | 50 | 0 | 0 | 50 |
| Cl1 | 0 | 3 | 36 | 39 |

# Quiz for the courageous:

- Give an algorithm for finding Minkowski's center at any p>1.
- Prove that the median is a Minkowski's center at p=1.
- Consider a zero-one feature f; given a cluster partition of the object set, put down a formula for cluster centers.
- Can you explain the meaning of a confidence interval to a user at large?
- I recommend comparing within-cluster centers with grand mean

# Interpretation of clusters by comparing centers with grand means+Bootstrap+iK-means

Contents:

- Interpretation of clusters via centers
- Center: DA perspective
- Center:  CS perspective
  - Validation of center using bootstrap
  - Comparing centers using bootstrap
- K-Means: Complementary criterion and anomalous cluster
- **Home work 2 & 3**

# HomeWork 2, I

- 1. Choose 3-6 features, Explain the choice, Apply K-means:
  - At K=5
  - At K=9
  - In both cases: 10 or more random initializations, chose the best over the K-means criterion
  - 2. Interpret each found partition by using features from the data table. Explain why you consider one of them better than the other in this perspective.

# HomeWork 2, II

- 3. Take one of the partitions
  - 3.1. Compare one of the features between two clusters with using bootstrap
  - 3.2. Take a feature, find the 95% confidence interval for its grand mean by using bootstrap
  - 3.3. Take a cluster, and compare the grand mean with the within-cluster mean for the feature by using bootstrap
  - Note: each application of bootstrap should be done in both, pivotal and non-pivotal, versions