2017 DA Lecture 3: Data pre-processing and Interpretation of clusters by comparing centers with grand means

Contents:

Scale types and quantification of categories

Data standardization

What is center: Minkowski's family

Validation of center using bootstrap

Comparing centers using bootstrap

Quantification of categories: Illustrative Data case

Companies characterized by mixed scale features; first three companies making product A, next three making product B, and the last two product C.

Company name	Income, $mln	MShare,%	NSup	AA	Sector
Aversiona	19.0	43.7	2	No	Utility
Antyops	29.4	36.0	3	No	Utility
Astonite	23.9	38.0	3	No	Manufacture
Bayermart	18.4	27.9	2	Yes	Utility
Breaktops	25.7	22.3	3	Yes	Manufacture
Bumchista	12.1	16.9	2	Yes	Manufacture
Civiok	23.9	30.2	4	Yes	Retail
Cyberdam	27.2	58.0	5	Yes	Retail

Company Dataset

Metadata: **Object names, Features and Domain knowledge**

1) Income, $ Mln;
2) MShare - Market share , per cent;
3) NSup - Number of principal suppliers;
4) Affirmative Action (AA) - Yes or No;
5) Sector - (a) Retail, (b) Utility, and (c) Manufacture.

Feature: Maps entities to feature values (unlike variable in math. statistics or m.l.)

Quantitative scale: Arithmetic mean makes sense Examples: 1) Income, 2) MShare, 3) Nsup

Binary scale: 1/0 coding makes it quantitative (mean=proportion)

Company Dataset

Metadata: **Object names, Features and Domain knowledge**

1) Income, $ Mln;
2) MShare - Market share , per cent;
3) NSup - Number of principal suppliers;
4) Affirmative Action (AA) - Yes or No;
5) Sector - (a) Retail, (b) Utility, and (c) Manufacture.

Feature: Maps entities to feature values

Quantitative scale: Arithmetic mean makes sense

Nominal scale: categories are exclusive, no relations (corresponds to partition of the set of objects)

Other scales (orders, non-alternative) not taken into account

Company dataset

Case 1: Companies 5

Company name	Income, $mln	MShare,%	NSup	AA	Sector
Aversiona	19.0	43.7	2	No	Utility
Antyops	29.4	36.0	3	No	Utility
Astonite	23.9	38.0	3	No	Manufacture
Bayermart	18.4	27.9	2	Yes	Utility
Breaktops	25.7	22.3	3	Yes	Manufacture
Bumchista	12.1	16.9	2	Yes	Manufacture
Civiok	23.9	30.2	4	Yes	Retail
Cyberdam	27.2	58.0	5	Yes	Retail

Data analysis issues:

- How to map companies to the screen with their similarity reflected in distances between points? (Summarization/visualization)

- Would clustering of companies reflect the product? What features would be involved then? (Summarization)

- Can rules be derived to predict the product for another company, coming outside of the table? (Correlation)

- Is there any relation between the structural features (Nsup,AA,Sector) and market related features (Income, MShare)? (Correlation)

Company Dataset: Quantification

Case 1: Companies 4

Company name	Income, $mln	MShare,%	NSup	AA	Sector
Aversiona	19.0	43.7	2	No	Utility
Antyops	29.4	36.0	3	No	Utility
Astonite	23.9	38.0	3	No	Manufacture
Bayermart	18.4	27.9	2	Yes	Utility
Breaktops	25.7	22.3	3	Yes	Manufacture
Bumchista	12.1	16.9	2	Yes	Manufacture
Civiok	23.9	30.2	4	Yes	Retail
Cyberdam	27.2	58.0	5	Yes	Retail

Quantitative coding: Each category is made into a 1/0 binary (dummy) feature "Does it hold? 1 if Yes, 0 if No."

Entity	Income	MShar	NSup	AA?	Util?	Manu?	Retail?
1	19.0	43.7	2	0	1	0	0
2	29.4	36.0	3	0	1	0	0
3	23.9	38.0	3	0	0	1	0
4	18.4	27.9	2	1	1	0	0
5	25.7	22.3	3	1	0	1	0
6	12.1	16.9	2	1	0	1	0
7	23.9	30.2	4	1	0	0	1
8	27.2	58.0	5	1	0	0	1

Company data 8×5 converted into quantitative format 8×7

Pre-processing:
- quantification
- filling in missings
- standardization

- **Standardisation:**
  - **shift of the origin** to compare data with a norm
  - **rescaling** to make features comparable
- $$Y_{iv} = (X_{iv} - a_v)/b_v$$
-
- X - original data
- Y – standardized data
- a_v – shift of the origin, typically, the **average**
- b_v – rescaling factor, traditionally the **standard deviation** (from statistics perspective), but **range** may be better

Company Dataset: Standardization

Company data 8×5 converted to the quantitative format 8×7

#	Inco	MSch	NSup	AA	Util	Man	Reta
1	19.0	43.7	2	0	1	0	0
2	29.4	36.0	3	0	1	0	0
3	23.9	38.0	3	0	0	1	0
4	18.4	27.9	2	1	1	0	0
5	25.7	22.3	3	1	0	1	0
6	12.1	16.9	2	1	0	1	0
7	23.9	30.2	4	1	0	0	1
8	27.2	58.0	5	1	0	0	1
Mean	22.45	34.12	3.00	.625	.375	.375	0.25

Three standardizations:

(i). (ii) and (iii)

Why are that many, and what is the need in data standardization?

Goal: to sharpen the data structure

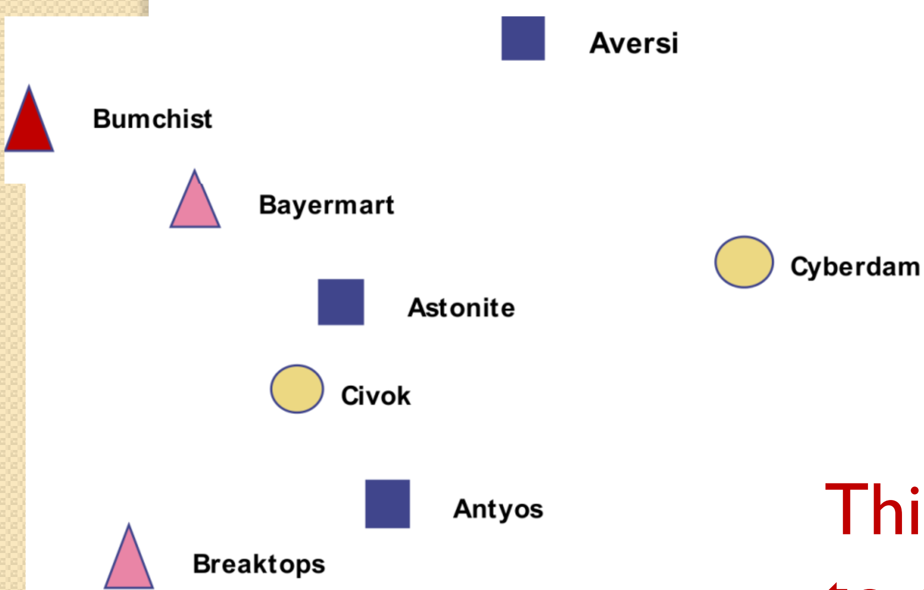
Data standardization in DA (unlike in Math. Stat.):

A. **Feature centering:** to look at feature values against a "normal" backdrop

B. **Feature normalization:** to balance feature weights

Company Dataset: Standardization (i)

-3.45	9.58	-1	-0.62	0.62	-0.38	-0.25
6.95	1.88	0	-0.62	0.62	-0.38	-0.25
1.45	3.88	0	-0.62	-0.38	0.62	-0.25
-4.05	-6.22	-1	0.38	0.62	-0.38	-0.25
3.25	-11.82	0	0.38	-0.38	0.62	-0.25
-10.35	-17.22	-1	0.38	-0.38	0.62	-0.25
1.45	-3.92	1	0.38	-0.38	-0.38	0.75
4.75	23.88	2	0.38	-0.38	-0.38	0.75



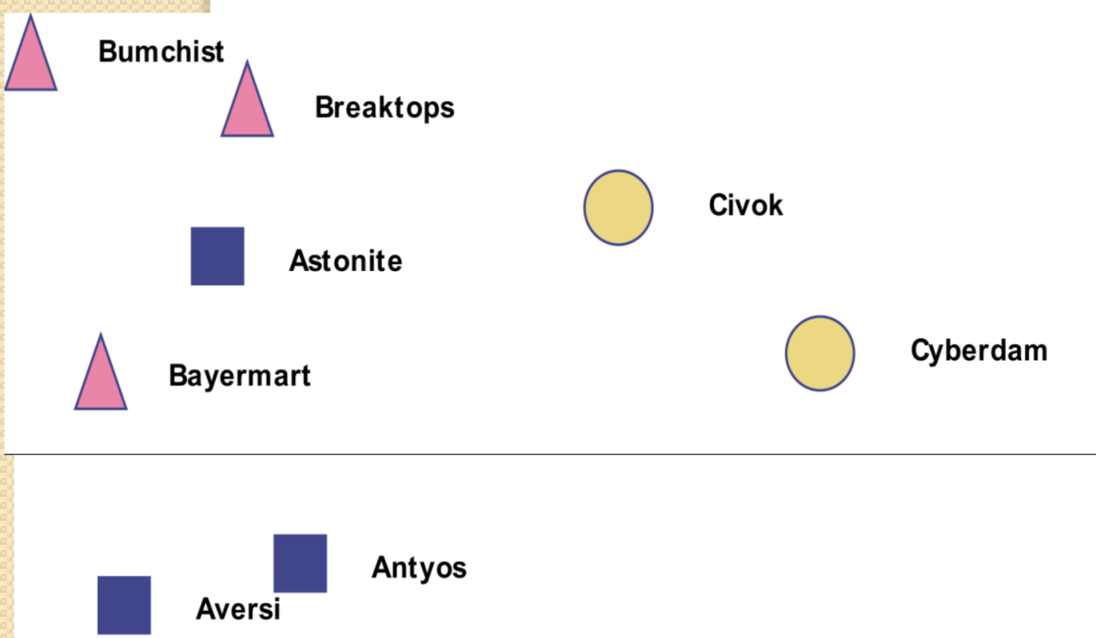
Structure of data at standardization (i): centering

Color/shape corresponds to the product (A,B,C)

This structure has nothing to do with product

Company Dataset: Standardization (ii)

-0.20	0.23	-0.33	-0.62	0.62	-0.38	-0.25
0.40	0.05	0	-0.62	0.62	-0.38	-0.25
0.08	0.09	0	-0.62	-0.38	0.62	-0.25
-0.23	-0.15	-0.33	0.38	0.62	-0.38	-0.25
0.19	-0.29	0	0.38	-0.38	0.62	-0.25
-0.60	-0.42	-0.33	0.38	-0.38	0.62	-0.25
0.08	-0.10	0.33	0.38	-0.38	-0.38	0.75
0.27	0.58	0.67	0.38	-0.38	-0.38	0.75



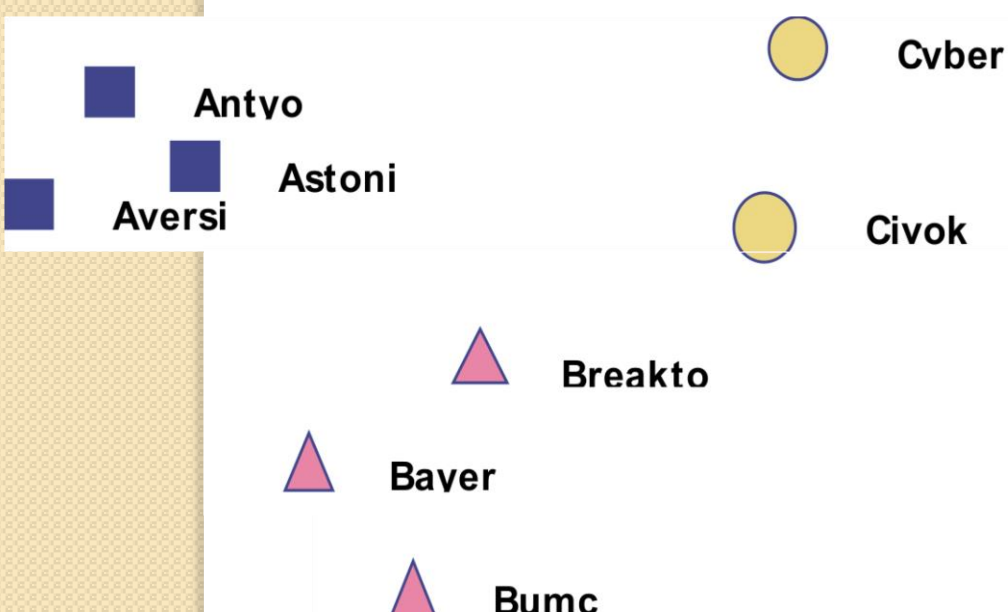
Structure of data at standardization (ii), centering and normalizing by range

Color/shape correspond to the product (A,B,C)

This structure somewhat corresponds to product

Company Dataset: Standardization (iii)

-0.20	0.23	-0.33	-0.62	0.36	-0.22	-0.14
0.40	0.05	0	-0.62	0.36	-0.22	-0.14
0.08	0.09	0	-0.62	-0.22	0.36	-0.14
-0.23	-0.15	-0.33	0.38	0.36	-0.22	-0.14
0.19	-0.29	0	0.38	-0.22	0.36	-0.14
-0.60	-0.42	-0.33	0.38	-0.22	0.36	-0.14
0.08	-0.10	0.33	0.38	-0.22	-0.22	0.43
0.27	0.58	0.67	0.38	-0.22	-0.22	0.43



Structure of data at standardization (iii): (ii)+ further normalizing

Sector features by sqrt(3)

Color/shape corresponds to the product (A,B,C)

This structure corresponds to product **quite well !**

Week2. What is **center**, 1

Consider a feature over *N* entities (transposed)

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

Data analysis view

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

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

at as small residuals e_i as possible

Def. $L_p^{\,p} = [|e_1|^p + |e_2|^p + \ldots + |e_N|^p]/N$

Minkowski criterion: ***min L_p or min $L_p^{\,p}$***

Week2. What is **center**, 2

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

$$\textit{Minimize } L_p^{\,p} = [|c-x_1|^p + |c-x_2|^p + \ldots + |c-x_N|^p]/N$$

with respect to all possible c

At different p, different solutions!

Week2. What is **center**, 2

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

$$\mathit{Minimize}\ L_p^{\,p} = [|c-x_1|^p+|c-x_2|^p+\ldots+|c-x_N|^p]/N$$

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.)

Week2. What is **center: Minkowski p-center (p ≥ 1)**

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

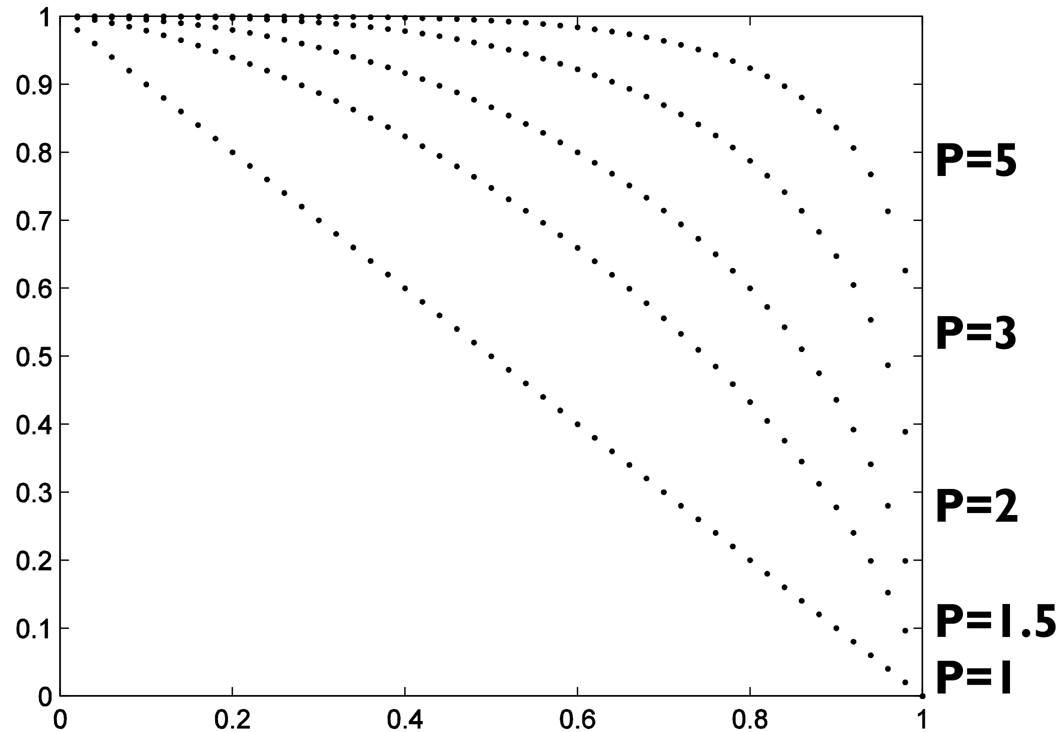
Take p=1. Then

$$L_1 = [|c-x_1| + |c-x_2| + \ldots + |c-x_N|]/N$$

It can be proven that the minimum of L_1 is reached at **c being the median! Then the minimum of L_1 should be used as the corresponding spread.**

Take p tending to infinity and the p-th root of Lp, then c tends to midrange.

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



Week2. What is center, 5

Feature (transposed)

19.0 29.4 23.9 18.4 25.7 12.1 23.9 27.2

Minkowski p	Spread	Center
Infinity	Half-range 8.65	Midrange 20.75
2	Standard deviation 5.26	Mean 22.45
1	Average deviation 4.1	Median 23.9

Week 2. 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

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

Reformulate

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

Week 2. What is **center**, 6

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

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

$N\bar{x}^2$ ***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

Week 2. What is **center**, 7

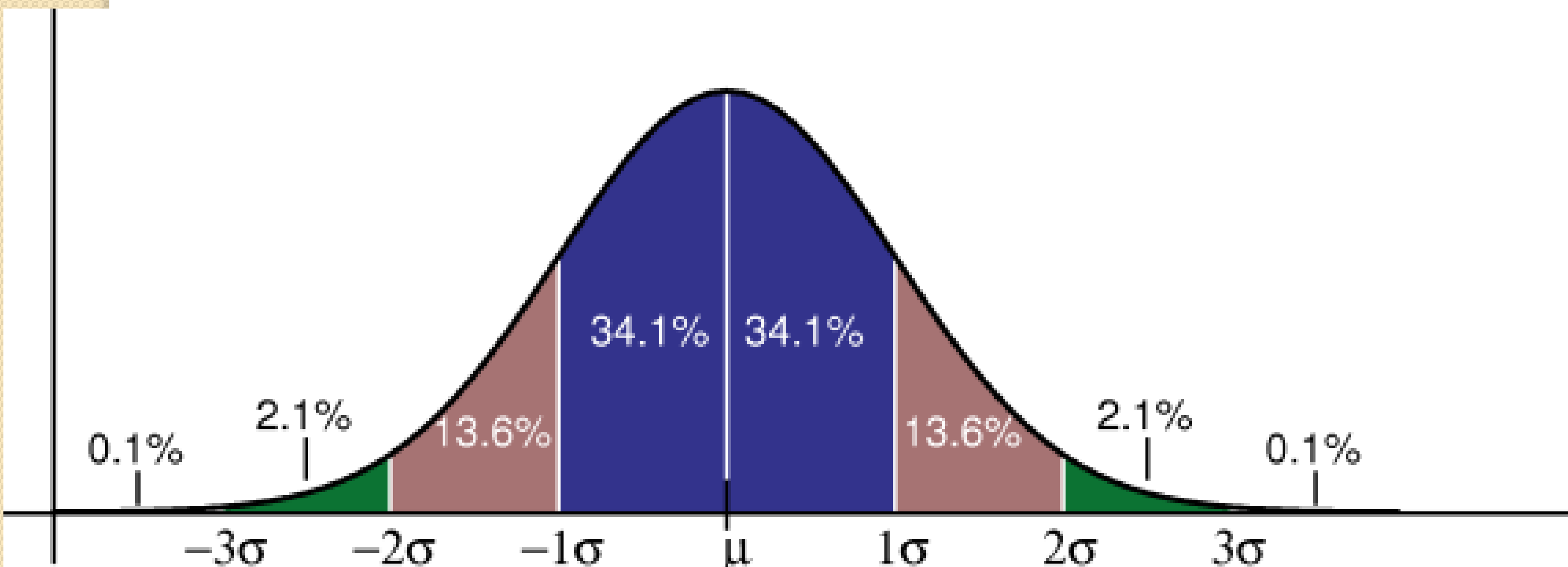
p	**Center**	**Comment**
2	Mean	Intuitive; Gaussian Sensitive to removal/addition of outliers
1	Median	Stable over removal/addition of outliers
∞	Midrange	Does not depend on the distribution shape Sensitive to change of range boundary points

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

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

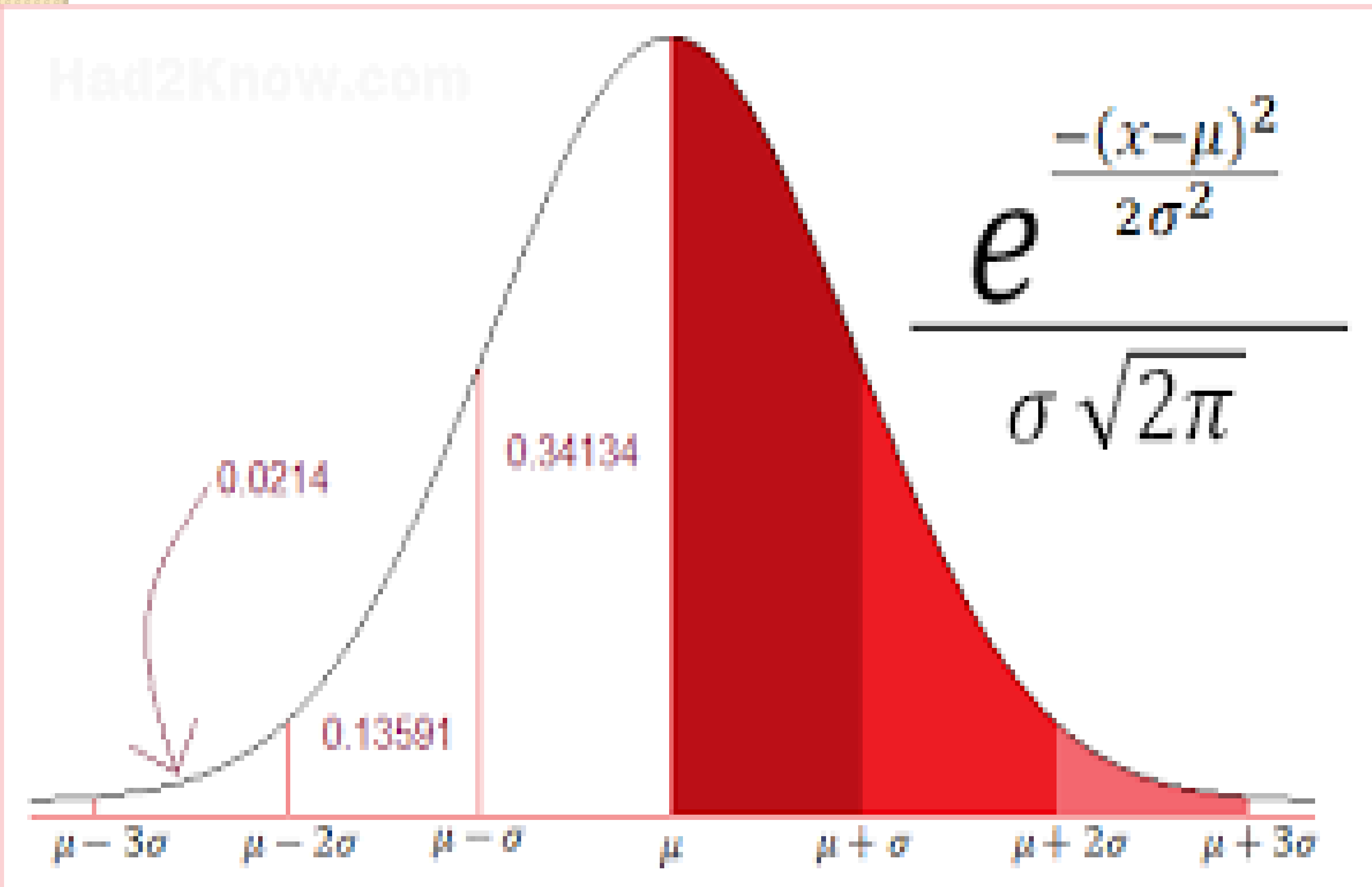
Gaussian density function

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



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

Gaussian density function



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

Estimates of parameters in the Gaussian density

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

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

Variance σ^2 (Standard deviation squared)

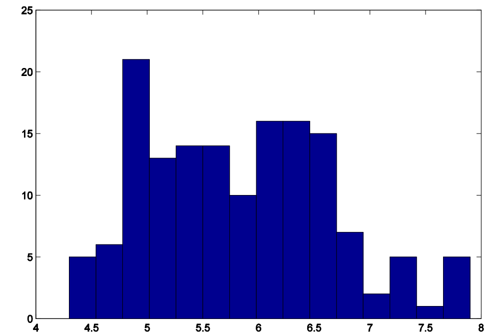
$$s^2=\frac{1}{N}\sum_{i=1}^{N}(x_i-a)^2 \quad \textbf{or}$$

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

Bootstrapping

Week 2. 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

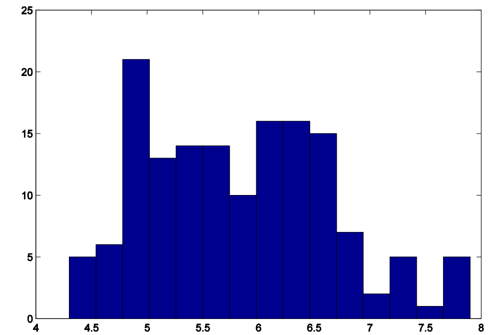
std = 0.8253

Week 2. Computational validation of Mean using bootstrap 1

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

Its mean m = 5.8433

std=0.8253

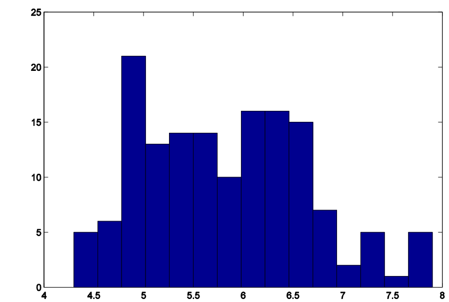


If one wants to reasonably speculate of plausible boundaries within which Mean should be expected at any possible set of iris specimen,

what should they suggest?

m±std? Or m±2*std? Or m±3*std? Or what?

Week 2. Computational validation of Mean using bootstrap 2

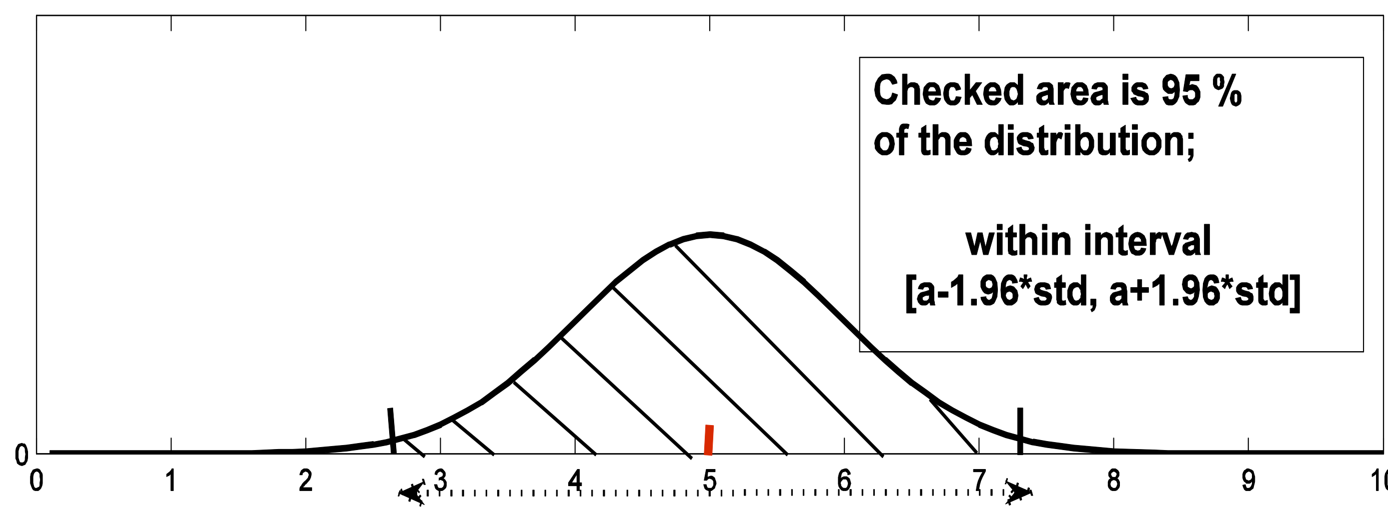


Plausible boundaries for mean?

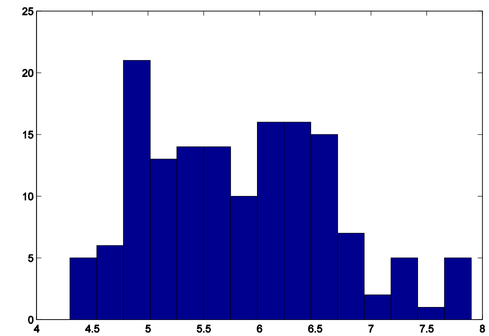
One way to go: using classical math statistics

Say, assume x is a random independent sample from a Gaussian distribution with a=5.8433 and σ=0.8253

Proven: is Gaussian too, with a= and σ=std/ √N



Week 2. Computational validation of Mean using bootstrap 3



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

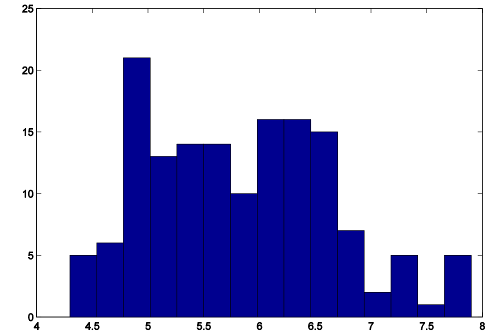
Lb= a - 1.96*std /$N^{1/2}$= 5.7108

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

Conclusion:

m within [5.7108, 5.9759] with confidence 95% (under Gaussian assumption)

Week 2. Computational validation of Mean using bootstrap 4



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

MatLab:

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

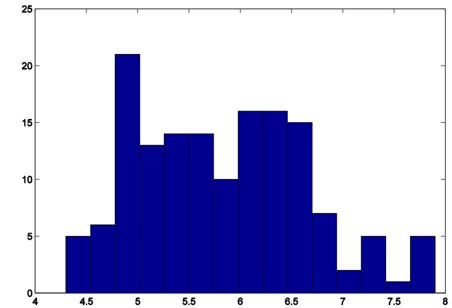
1 4 4 **N, the number of entities**

r= 3 1 4 **M, the number of samples**

3 1 3 **First sample: entity 1, entity 3 (twice),**

2 3 1 **entity 2 (entity 4 is missed: why?)**

Week 2. Computational validation of Mean using bootstrap 5



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

Its mean = 5.8433, std=0.8253

Plausible boundaries for m?

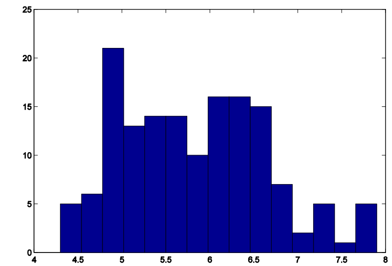
Bootstrap

MatLab:

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

This gives M=5000 means of random samples of x

Week 2. Computational validation of Mean using bootstrap 6

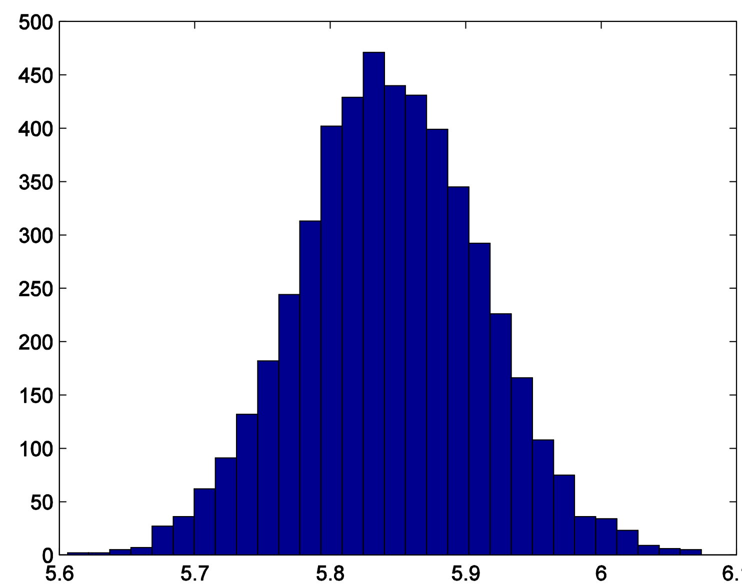


Plausible boundaries for m?

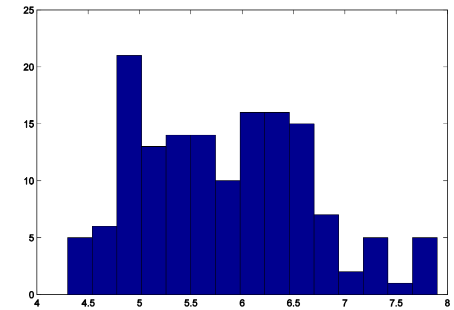
Bootstrap

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

Histogram of M=5000 means



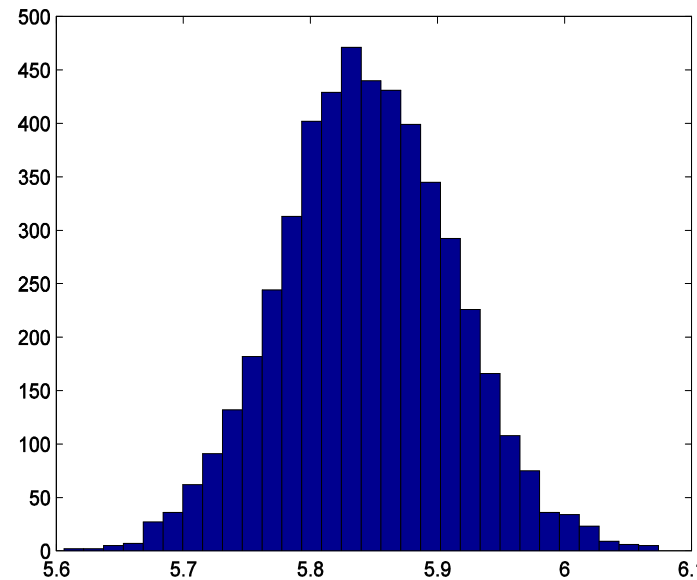
Week 2. Computational validation of Mean using bootstrap 8



Feature x=iris(:,1); = 5.8433, std=0.825

Plausible boundaries for m?

Bootstrap Histogram of M=5000 means mr



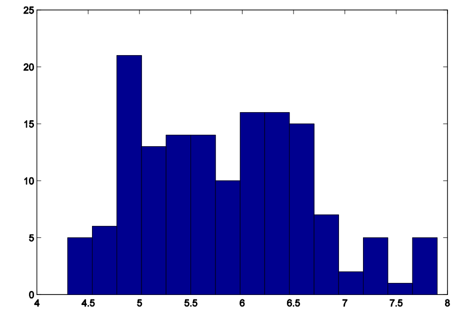
A. Pivotal method

(95% confidence)

Assume mr be Gaussian

```
>> mmr=mean(mr); % 5.8444
>> smr=std(mr); %   0.0675
>> lbp=mmr-1.96*smr; %   5.71
>> rbp=mmr+1.96*smr; %  5.9767
```

Week 2. Computational validation of Mean using bootstrap 8

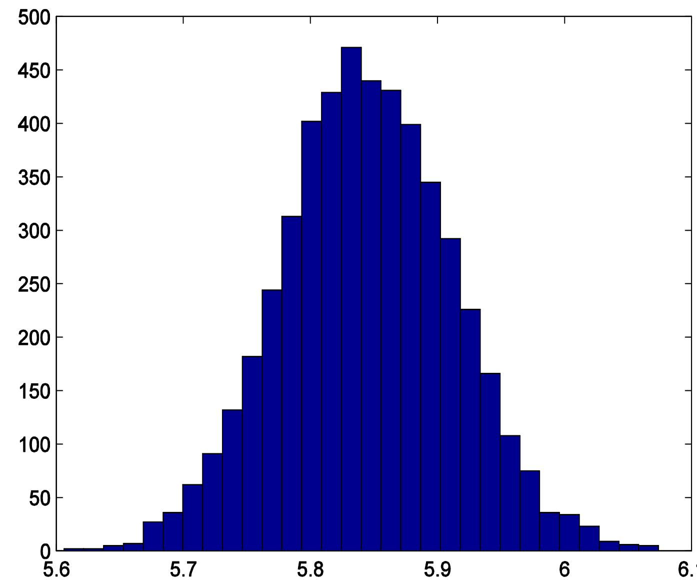


Feature x=iris(:,1); = 5.8433, std=0.825

Plausible boundaries for m?

Bootstrap Histogram of M=5000 means mr

B. Nonpivotal method (95% confidence)



Take 2.5% and 97.5% percentil as the boundaries

1% of 5000 is 50;

2.5% is 125; 97.5% is 4875

>> smr=sort(mr); % sorting

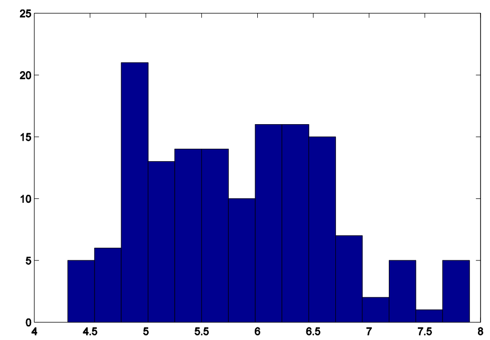
>> lbn=somr(126); % 5.7120

>> rbn=somr(4875); % 5.9773

Week 2. Computational validation of Mean using bootstrap 9

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 nonpivotal)

with 95% confidence

I can see no difference in these; there is an issue with the choice of 95%, too…

Comparison of means using Bootstrap

- Compare mean Sepal width in Taxons 2 and 3:

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

Lecture's summary

- Three scale types
- Quantification of a mixed scale data table
- Data standardization
- Minkowski's center
- Minkowski's centers at p=1, 2, ∞
- Decomposition of the data scatter at p=2
- Gaussian distribution and its parameters
- Confidence interval
- Bootstrap: what is it?
- Bootstrap for validating the mean: pivotal, nonpivotal
- Bootstrap for comparing two means

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 the user?
- I recommend comparing within-cluster centers with grand mean. Is there any problem about deriving a 95% confidence interval for the difference between them?

Home-work 2

- Reasonably select several features in your dataset
- Apply K-Means at two or three different K
- Take a clustering of your liking and interpret all the clusters by comparing their centers with grand mean
- Answer this question in your report: is it an interesting partition?

Home-work 3

- Reasonably select two clusters in the clustering you dealt with in HW2
- Take one of the features at one of the clusters and validate its within-cluster mean using bootstrap
- Take one more cluster and compare the within-cluster means of the feature by using bootstrap