Modern Methods in Data Analysis

Lecture 2, 2019: Feature scales, K-means, and interpretation

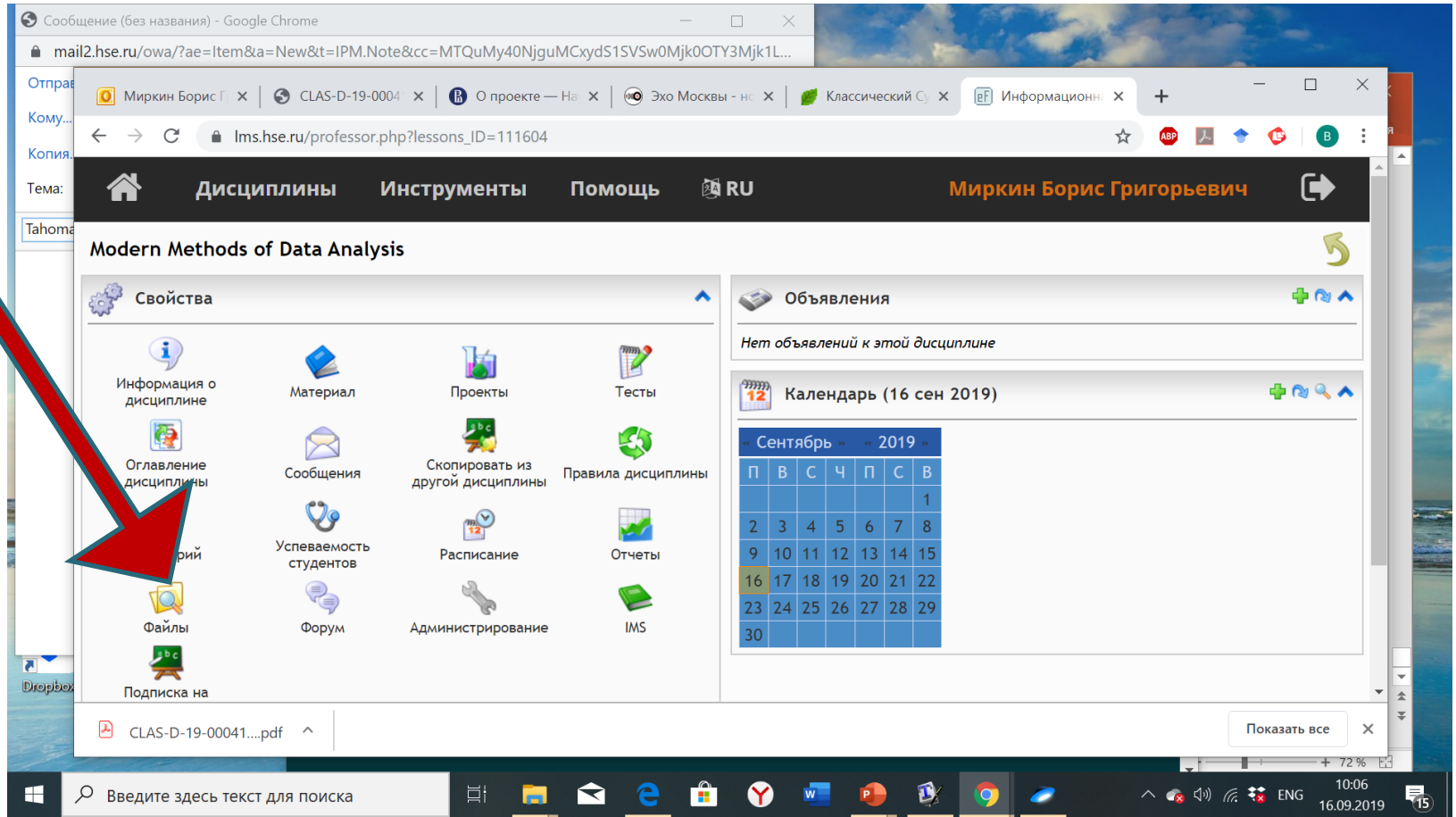
Boris Mirkin

Professor, Data Analysis and AI, NRU HSE, Moscow, bmirkin@hse.ru, **8(963)-7234021**

Professor Emeritus, Computer Science, Birkbeck UL, London UK, mirkin@dcs.bbk.ac.uk

In LMS HSE:

open the class page and go to Files/Mirkin to find lecture slides



Lecture 2 Contents

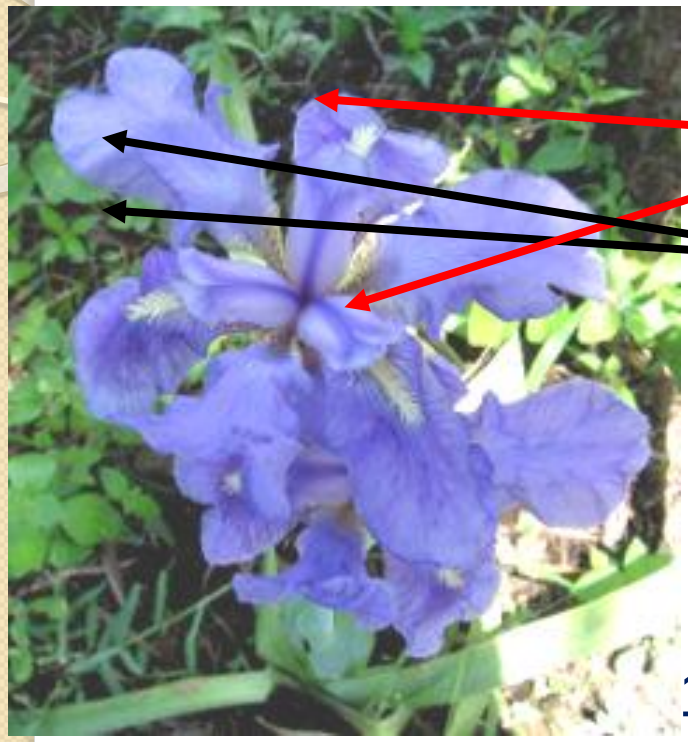
- Two formalizations of the concept of feature: vector (DA) and random variable (Classical Statistics)
- Quantitative and categorical features
- Data standardization
- How to categorize a quantitative feature.
- Partition and its distribution.
- K-means clustering.
- Interpretation of clusters
- The average and its properties
  - Data analysis: Vector view
  - Classical statistics: Density function view
- K-means criterion.

Lecture 2 Contents

- **Two formalizations of the concept of feature: vector and random variable**
- Quantitative and categorical features
- How to categorize a quantitative feature. Partition and its distribution.
- K-means clustering.
- Interpretation of clusters
- The average and its properties
- K-means criterion.

Самая популярная таблица данных: Anderson–Fisher Iris Dataset

Iris flower



Sepal / Чашелистик

Petal / Лепесток

150×4 data of three taxa:

	Taxon
1-50	*Iris setosa* (diploid)
51-100	*Iris versicolor* (tetraploid)
101-150	*Iris virginica* (hexaploid)

Features

W1	**Sepal length**
W2	Sepal width
W3	Petal length
W4	Petal width

Taxa

} Metadata

Iris, features w1, w2, w3, w4

#	Iris			
	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
6	4.9	3.1	1.5	0.2
7	5.0	3.2	1.2	0.2
8	4.6	3.2	1.4	0.2
9	5.0	3.3	1.4	0.2
…	…..			
150	5.1	3.5	1.4	0.2

Consider feature w1. How to model it? Data Science: entries in w1 matter only!!!

What is feature w1? According to Data Science view, just the column w1's contents:

- Index 1 through 9

5.1 4.4 4.4 5.0 5.1 4.9 5.0 4.6 5.0

. . . . . .

- Index 142 through 150

6.7 6.3 6.5 6.5 7.3 6.7 5.6 6.4 6.5

What is this as a mathematical object?

What is the column w1's contents as a mathematical object?

- Index 1 through 9

5.1 4.4 4.4 5.0 5.1 4.9 5.0 4.6 5.0

. . . . . .

- Index 142 through 150

6.7 6.3 6.5 6.5 7.3 6.7 5.6 6.4 6.5

Two different views co-exist in Data Science (like the photon, unit of light, in Quantum Physics: both a particle and a wave)

Two different views, A) and B), at the w1 feature as a mathematical object:

:

- Index 1 through 9

5.1 4.4 4.4 5.0 5.1 4.9 5.0 4.6 5.0

. . . . . .

- Index 142 through 150

6.7 6.3 6.5 6.5 7.3 6.7 5.6 6.4 6.5

A) **Data Analysis: Vector of 150x1 dimension**

B) **Classical Statistics: 150-strong sample from a random variable**

A) Feature as vector, 1:

- Index 1 through 9

5.1 4.4 4.4 5.0 5.1 4.9 5.0 4.6 5.0

. . . . . .

- Index 142 through 150

6.7 6.3 6.5 6.5 7.3 6.7 5.6 6.4 6.5

Math: Given a set I of object indices or names, feature is a mapping $f: I \to R$ where R is the set of all reals, that is, $f=(f_i)$, $i \in I$, an $|I|$-dimensional vector

A) Feature as vector, 2:

- Index 1 through 9

5.1 4.4 4.4 5.0 5.1 4.9 5.0 4.6 5.0

. . . . . .

- Index 142 through 150

6.7 6.3 6.5 6.5 7.3 6.7 5.6 6.4 6.5

Pro: a) Intuitive;

b) Objects are explicit (rows)

c) Linear algebra applies

Con: d) Empirical (depends on I, cannot be extended to the universe)

B) Feature as random variable, 1:

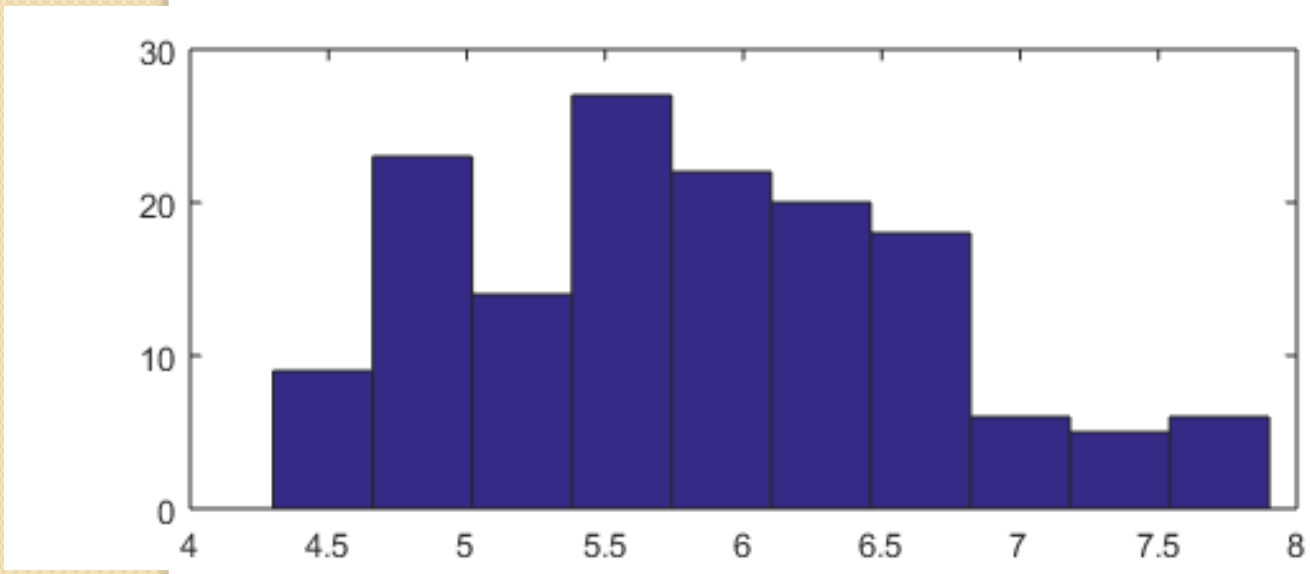
Index 1 through 9

5.1 4.4 4.4 5.0 5.1 4.9 5.0 4.6 5.0

. . . . . .

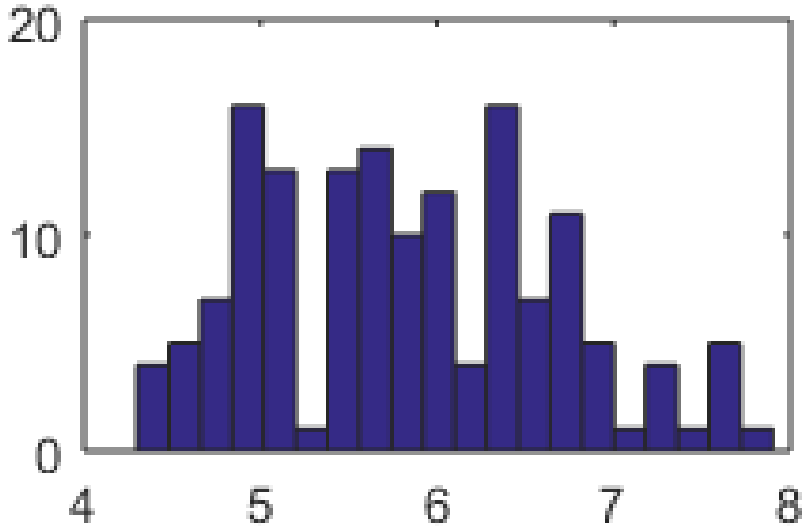
- Index 142 through 150

6.7 6.3 6.5 6.5 7.3 6.7 5.6 6.4 6.5

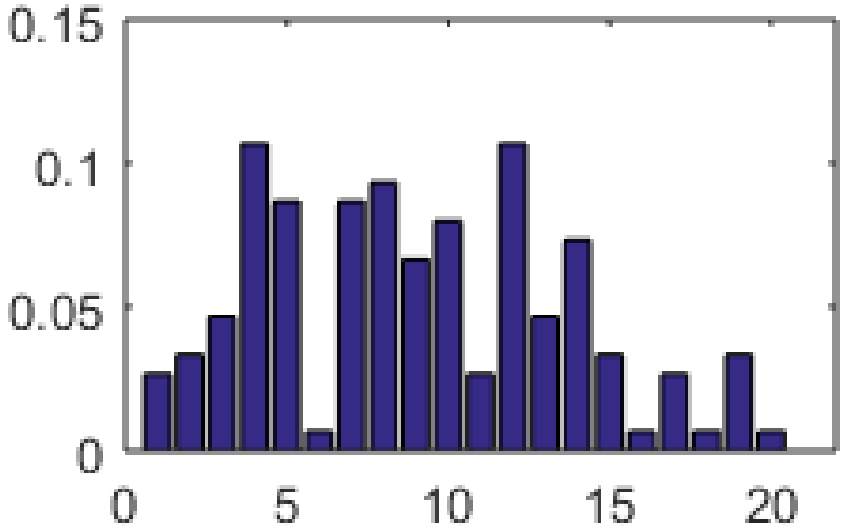


Histogram: range is divided in n(=10) bins; numbers of objects falling in bins are presented by bars.

B) Feature as random variable, 2:



(a)

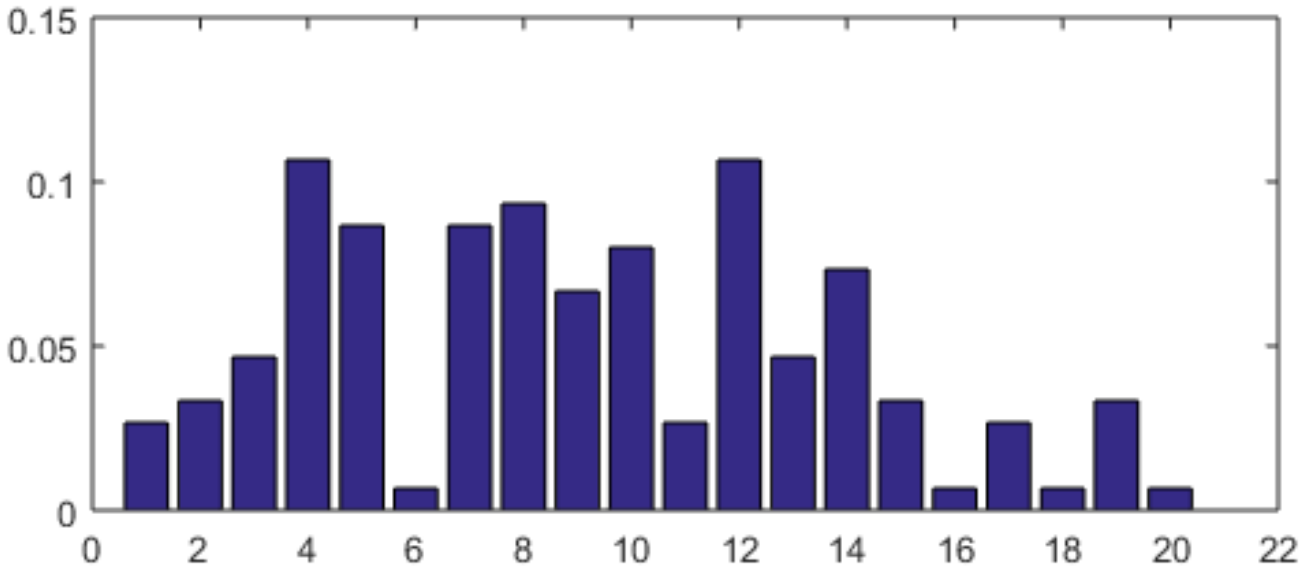


(b)

Histogram: (a) range is divided in n(=20) bins; numbers of objects falling in bins are presented by bars.

Relative histogram: **(b)** bars express proportions of objects in the bins (sum to 1).

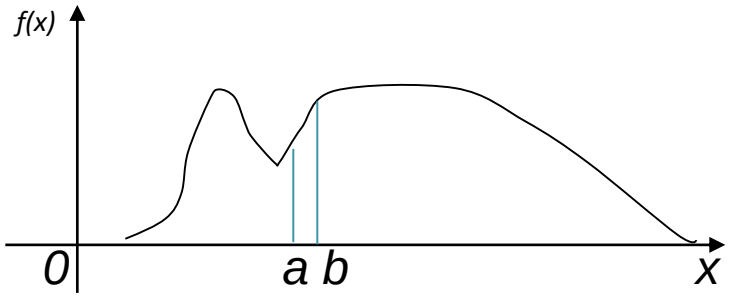
B) Feature as random variable, 3:



Relative histogram: bars express proportions of objects in the *n* bins.

Density function, an abstraction of histogram at *N* and *n* tending to infinity: a measurable nonnegative function (curve) *f(x)* such that $\int_{-\infty}^{+\infty} f(x)dx = 1$.

B) Feature as random variable, 4:



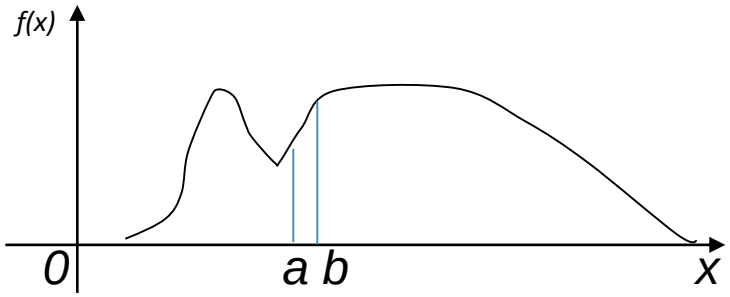
Density function, an abstraction of the relative histogram at *N*, *n* tending to infinity: a nonnegative measurable function *f(x)* such that

$$\int_{-\infty}^{+\infty} f(x)dx = 1$$

$$\int_{a}^{b} f(x)dx =$$

probability of the variable to fall in $[a, b]$

B) Feature as random variable, 5:



Math: Random variable = Density function

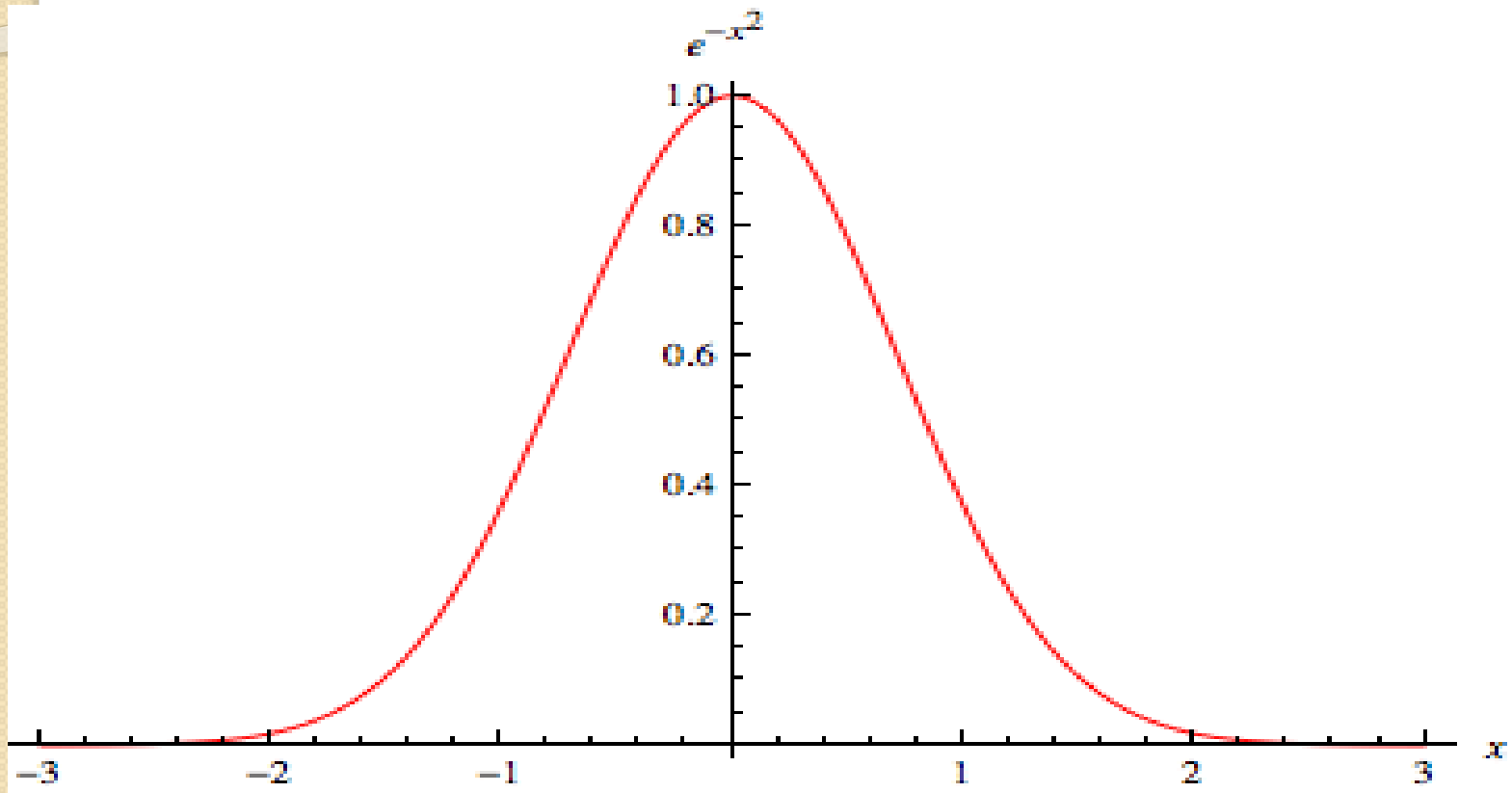
Pro: (a) Universal, does not depend on set I

(b) Probability theory can be used

Con: (c) Objects are implicit

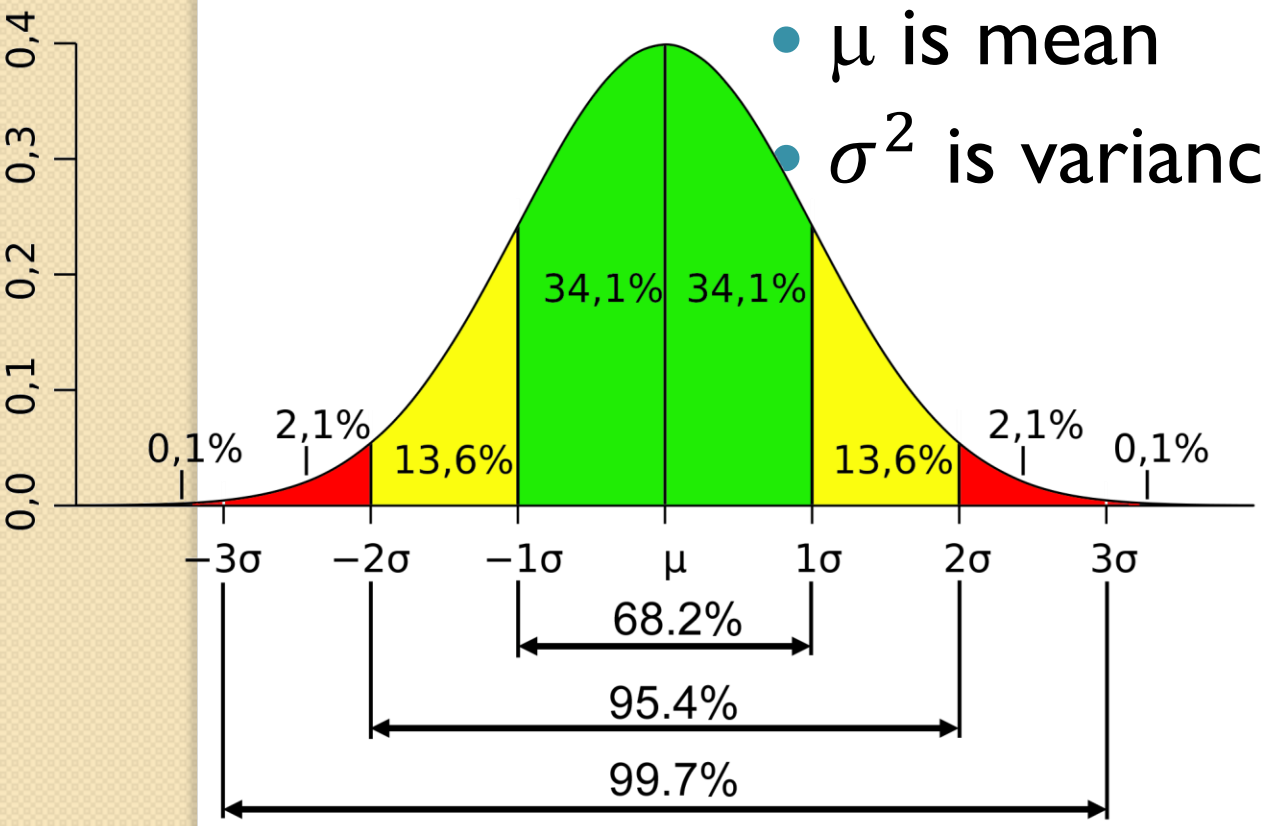
B) Popular density functions: Gaussian N(0,1)

$$f(x) = \exp\{-x^2\}$$



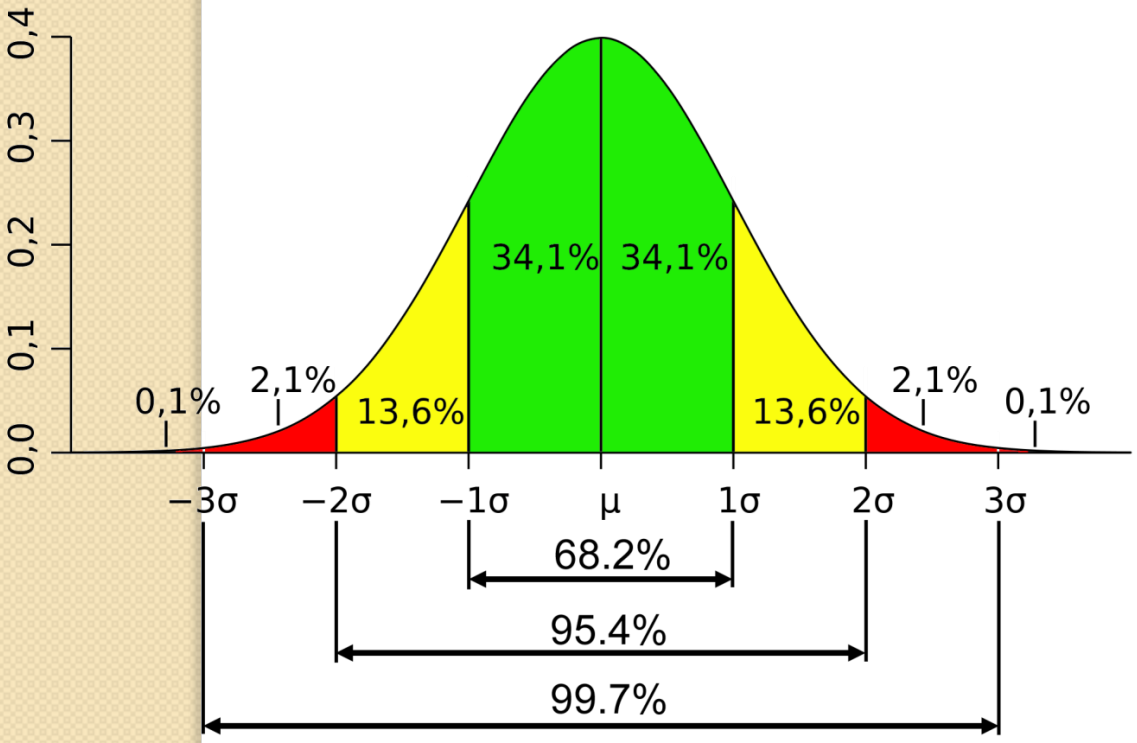
B) Popular density functions: general Gaussian $N(\mu,\sigma)$

- $f(x)=\exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)/\sqrt{2\pi\sigma^2}$
- μ is mean
- σ^2 is variance; σ, the std



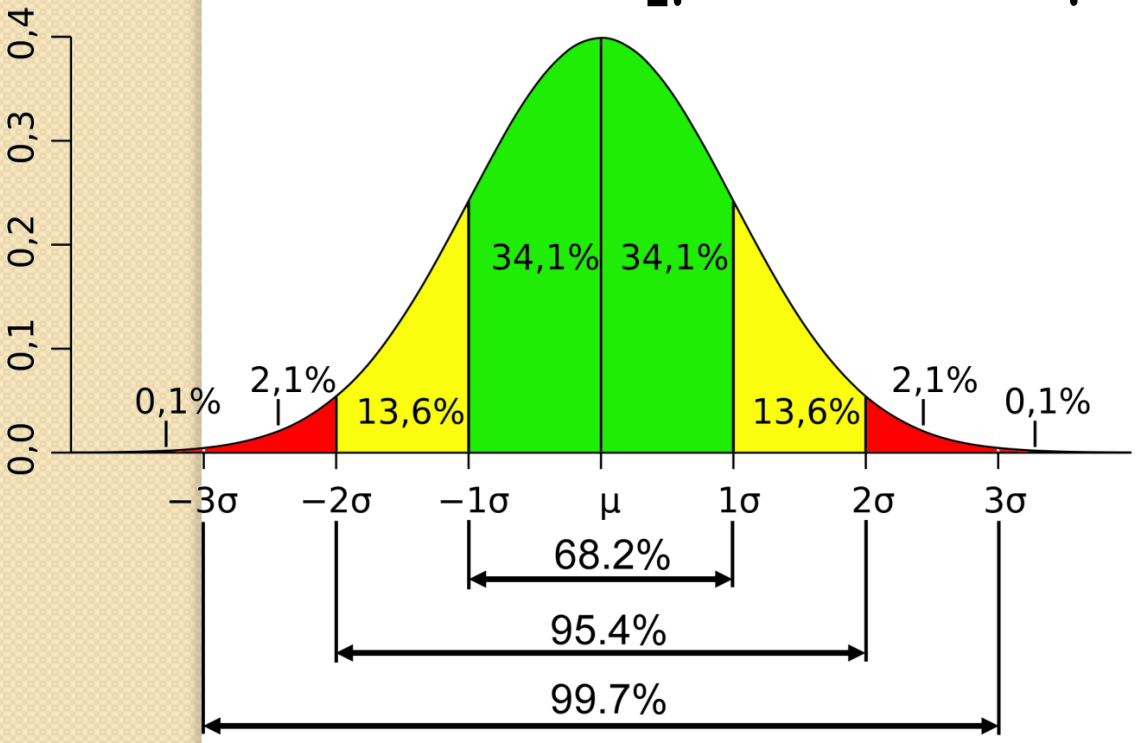
B) General Gaussian N(μ,σ)

- Bell curve (symmetric over μ)
- σ^2 is variance, σ is standard deviation (same scale as feature)
- 2σ rule, 3σ rule



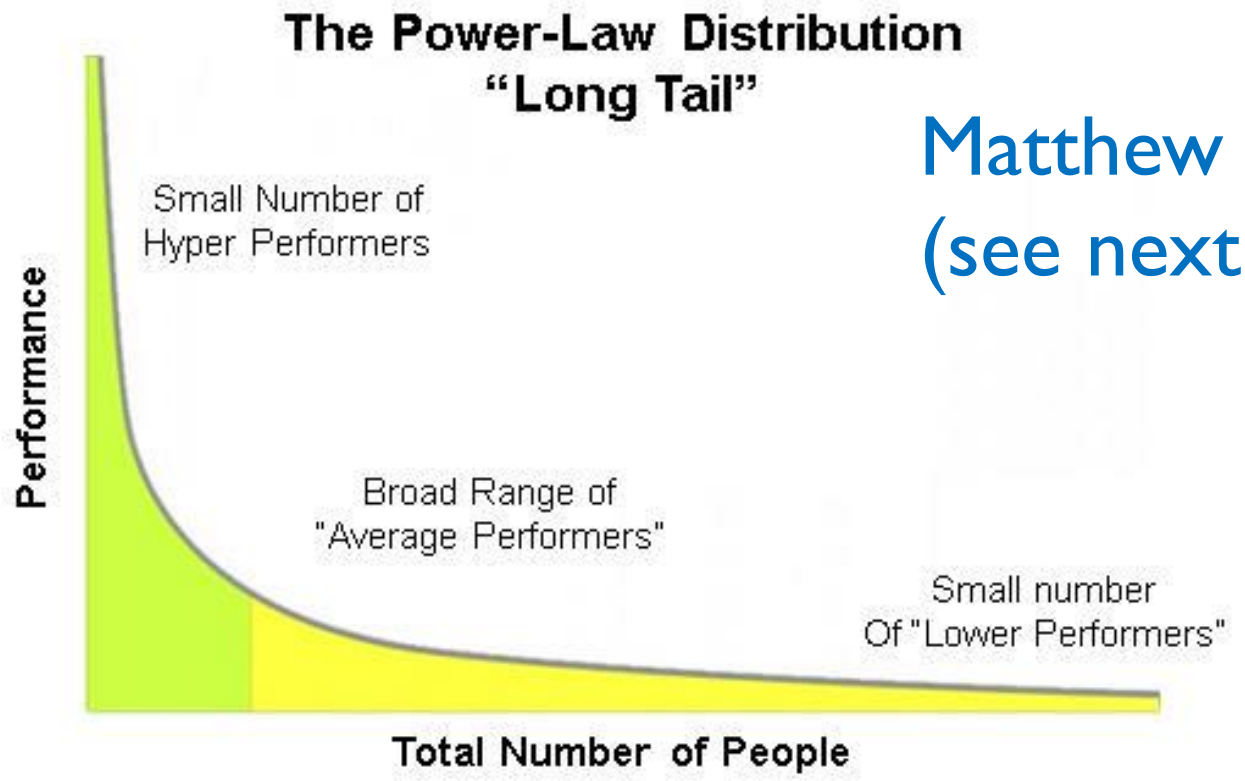
B) General Gaussian N(μ,σ)

- **Bell** curve (symmetric over μ)
- Central interval to account for 0.95=95% of the area:
- $[\mu - 1.96\sigma, \mu + 1.96\sigma]$



B) Popular density functions: **power law**

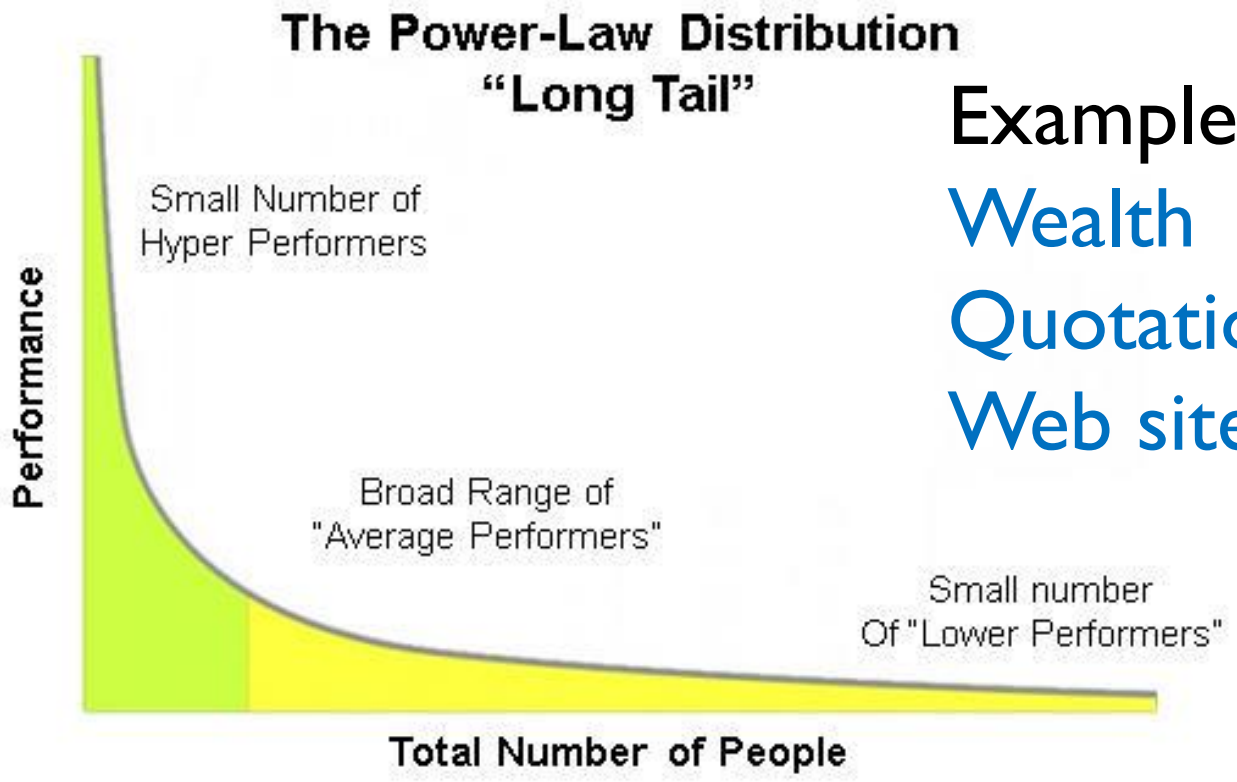
- $f(x)=cx^{-\alpha}$
- α the steepness
- Scale-free (why? Can you tell?)



Matthew effect
(see next slide)

B) Power law: Matthew effect

- **"For unto every one that hath shall be given, and he shall have abundance: but from him that hath not shall be taken even that which he hath."** Matthew Gospel 25:29



Examples:

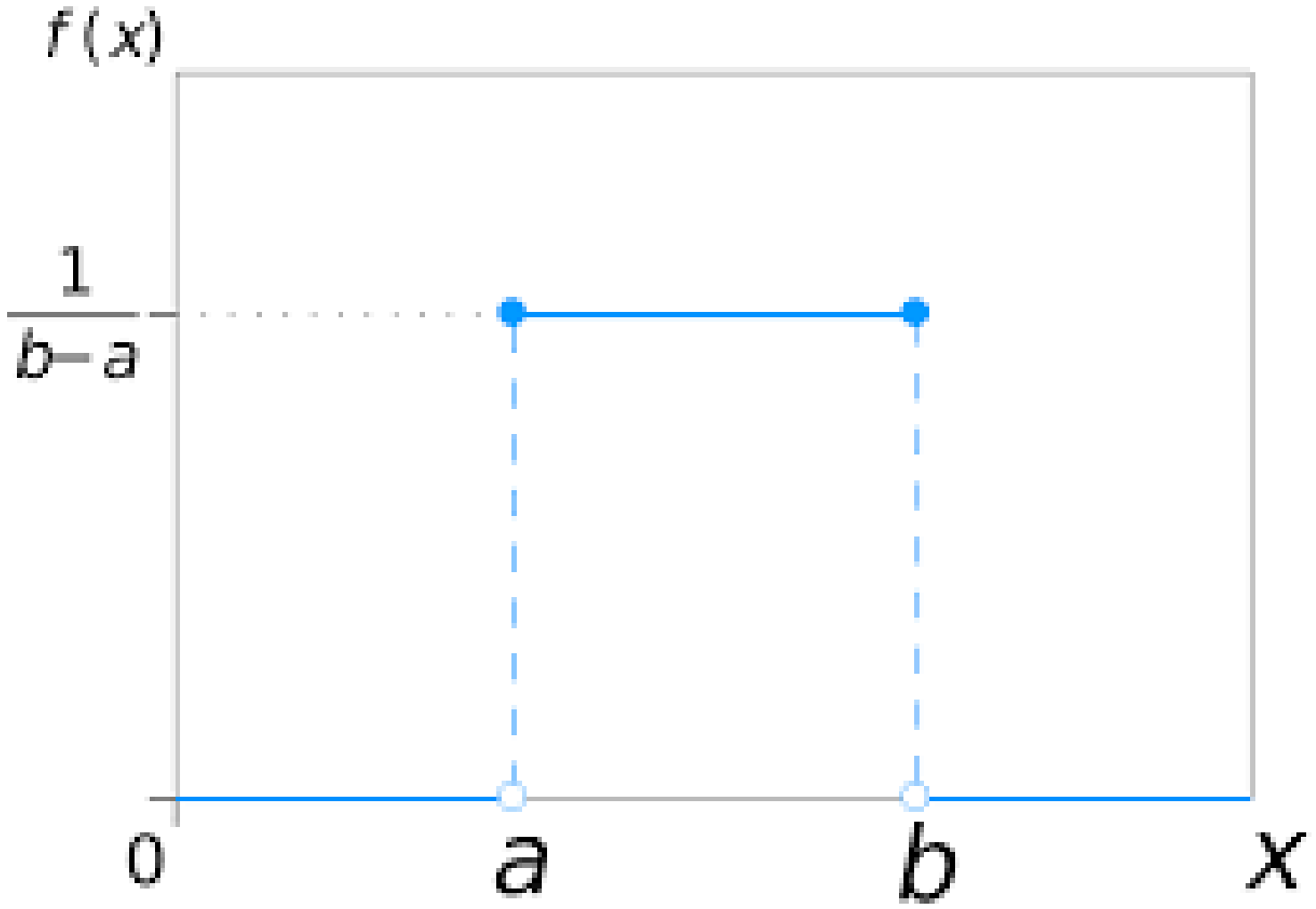
Wealth

Quotations

Web site popularity

B) Popular density functions: **uniform** distribution over [*a*, *b*] interval

Why is that?



Lecture 2 Contents

- Two formalizations of the concept of feature: vector and random variable
- **Quantitative and categorical features**
- Data standardization
- How to categorize a quantitative feature.
- Partition and its distribution.
- K-means clustering.
- Interpretation of clusters
- The average and its properties
- K-means criterion.

Mixed feature scales: Another 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 scales, I

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

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

Feature: Maps entities to feature values

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

Binary scale: 1/0 coding makes it quantitative (mean=proportion) Example: 4) AA

Company Dataset scales,2

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

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

Plus: The names – three products, A, B, and C

Feature: Maps entities to feature values

Nominal scale: categories are exclusive, no relations (corresponds to partition of the set of objects), Example: 5) Sector

Order scale: categories are exclusive, linearly ordered ((corresponds to ranking of the set of objects) Example: ?

Company dataset

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:

Company dataset

Company name	Income, $mln	MShare, %	NSup	AA	Sector

Data analysis issues:

- How to map companies to screen, to reflect similarity in distances between points? (Summarization)
- Would clustering of companies reflect the product? What features would be involved? (Summarization)
- Can rules be derived to predict the product for a 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)

Lecture 2 Contents

- Two formalizations of the concept of feature: vector and random variable
- Quantitative and categorical features
- **Data preprocessing & standardization**
- How to categorize a quantitative feature.
- Partition and its distribution.
- K-means clustering.
- Interpretation of clusters
- The average and its properties
  - Vector view
  - Density function view
- K-means criterion.

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 a quantitative format 8×7

Pre-processing:
- quantification
- filling in missings (not covered)
- standardization

- **Standardization:**
  - **shift of the origin to a** to compare data with a "**norm**"
  - **rescaling by relating to b** 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), (iii) (see next)

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

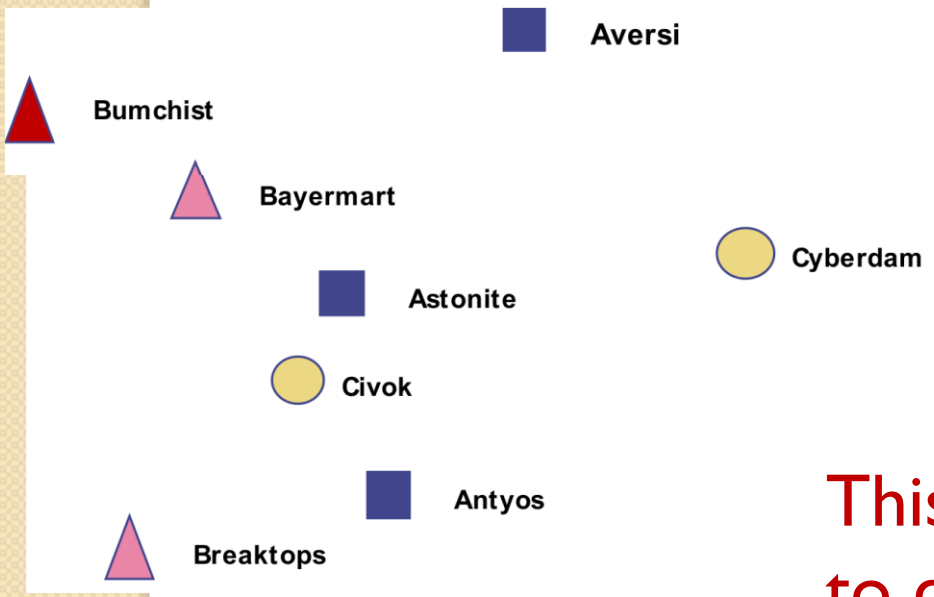
Goal: to sharpen the data structure

Data standardization goals in DA (unlike in CS):

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

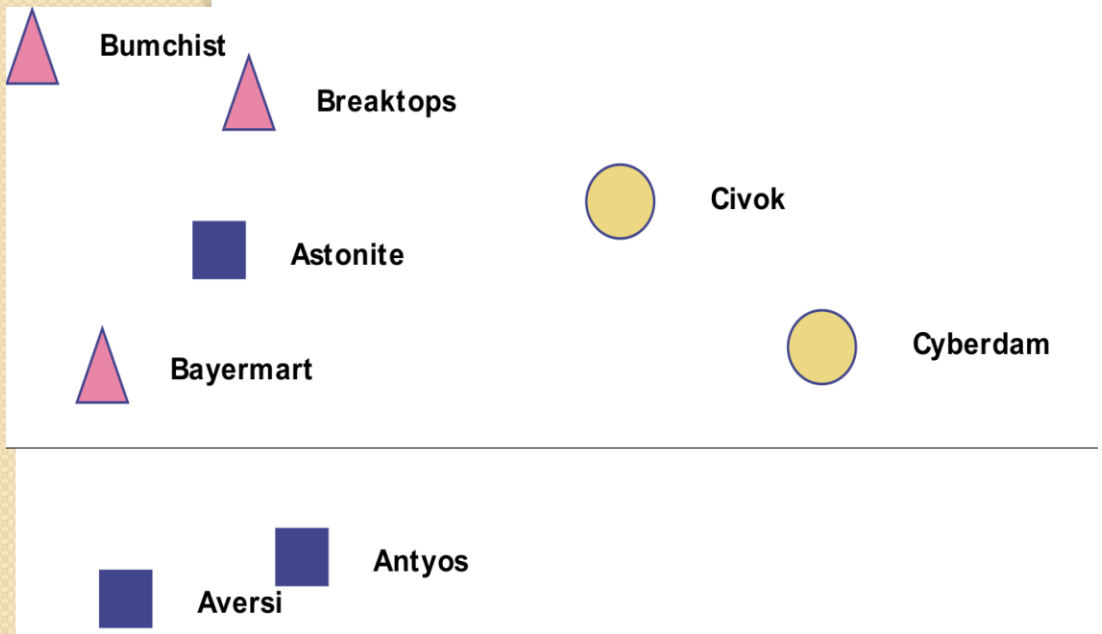
(i): centering

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

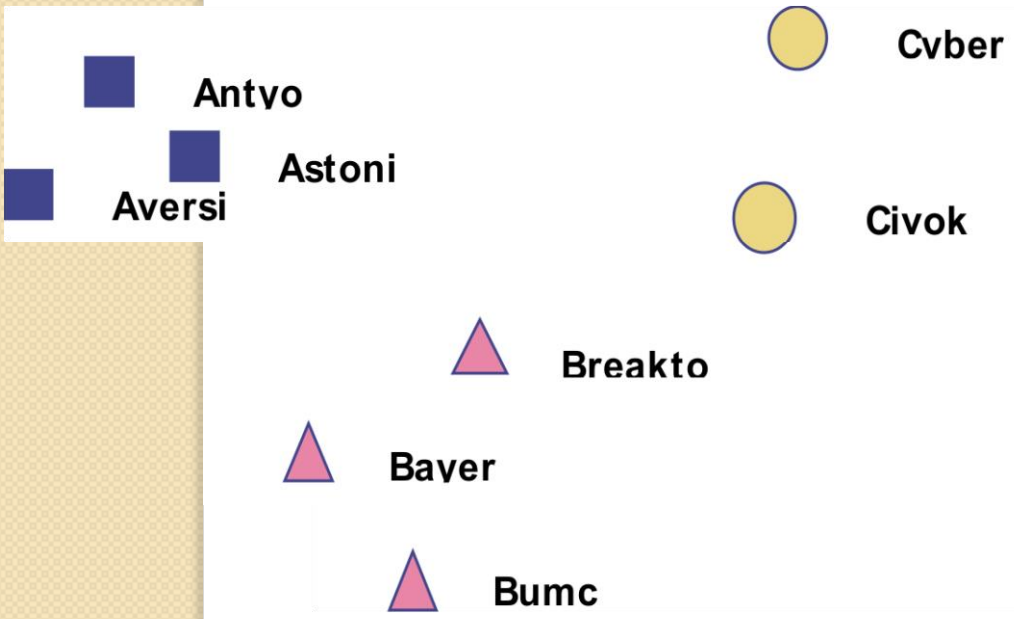
(ii) centering and normalizing by range

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 dummies by $\sqrt{3}$ = sqrt(3)

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

Lecture 2 Contents

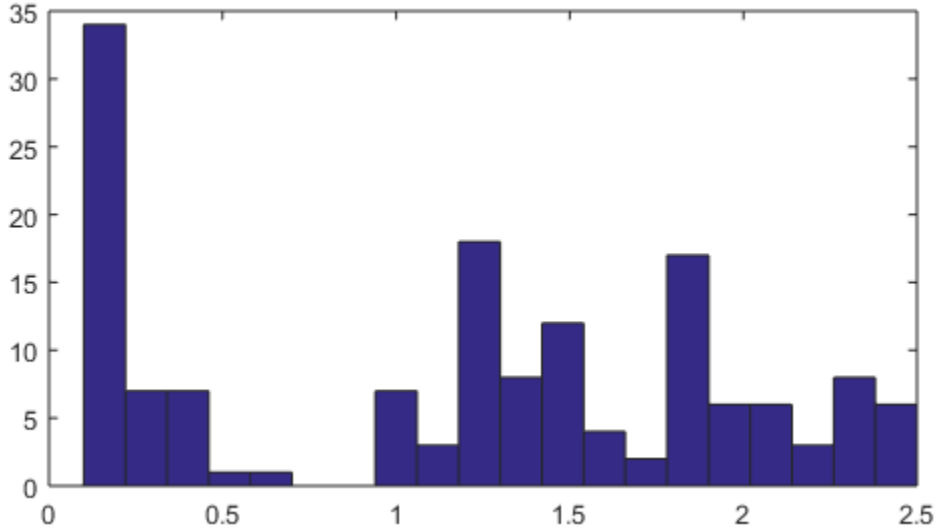
- Two formalizations of the concept of feature: vector and random variable
- Quantitative and categorical features
- **How to categorize a quantitative feature. Partition and its distribution.**
- K-means clustering.
- Interpretation of clusters
- The average and its properties
- K-means criterion.

Quantitative feature

>>iris=load('Data/iris.dat') % 150x4 dataset

>>w=iris(:,4); %4th feature, Petal width

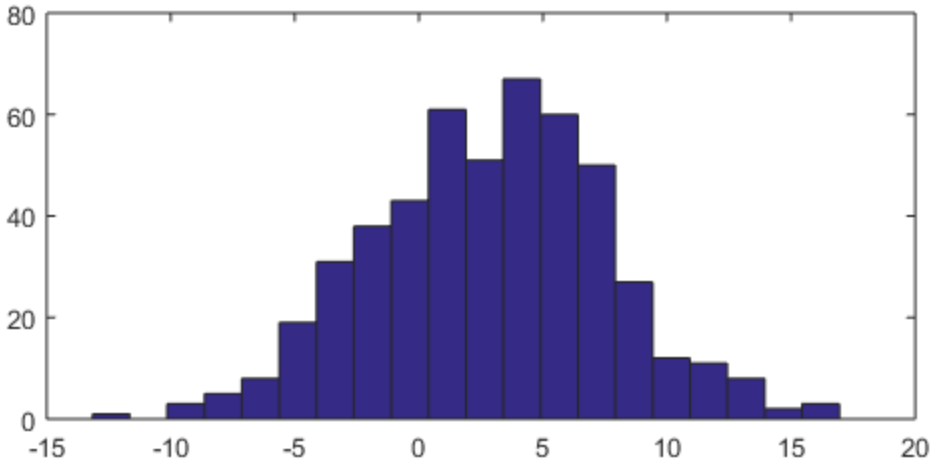
>> hist(w,20); %Histogram with 20 bins



This sample is not homogeneous. In what sense?

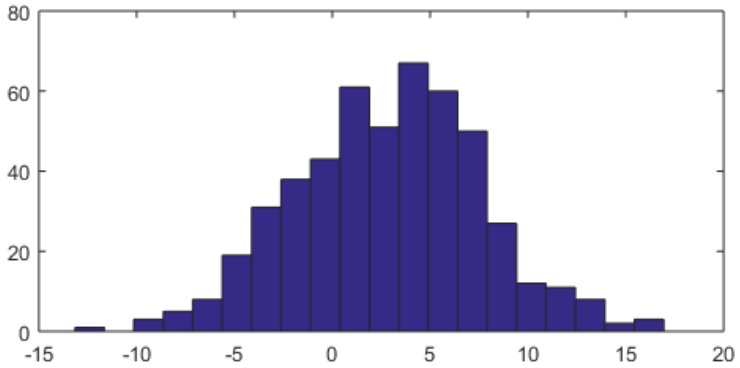
A homogeneous sample

- >> x=5*randn(500,1)+3;hist(x,20);

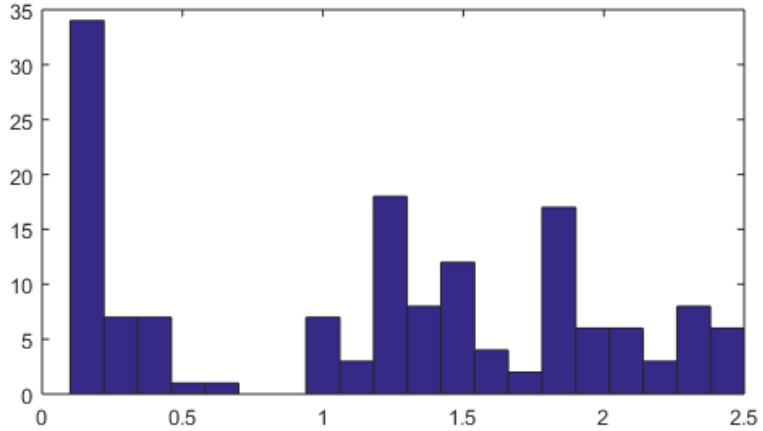


- No deep minima

Homogeneity and non-homogeneity

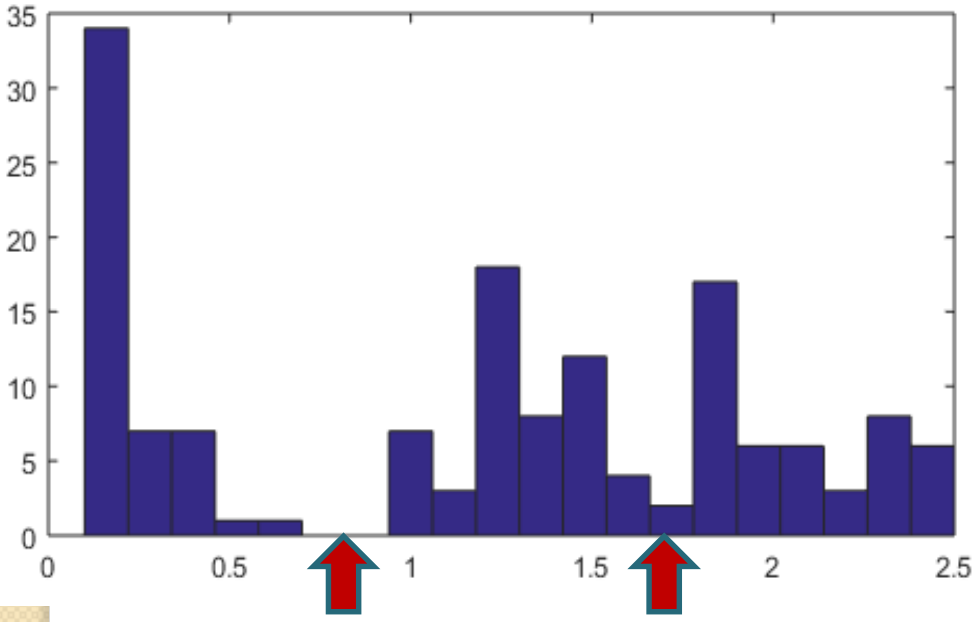


No deep minima (homogeneity)



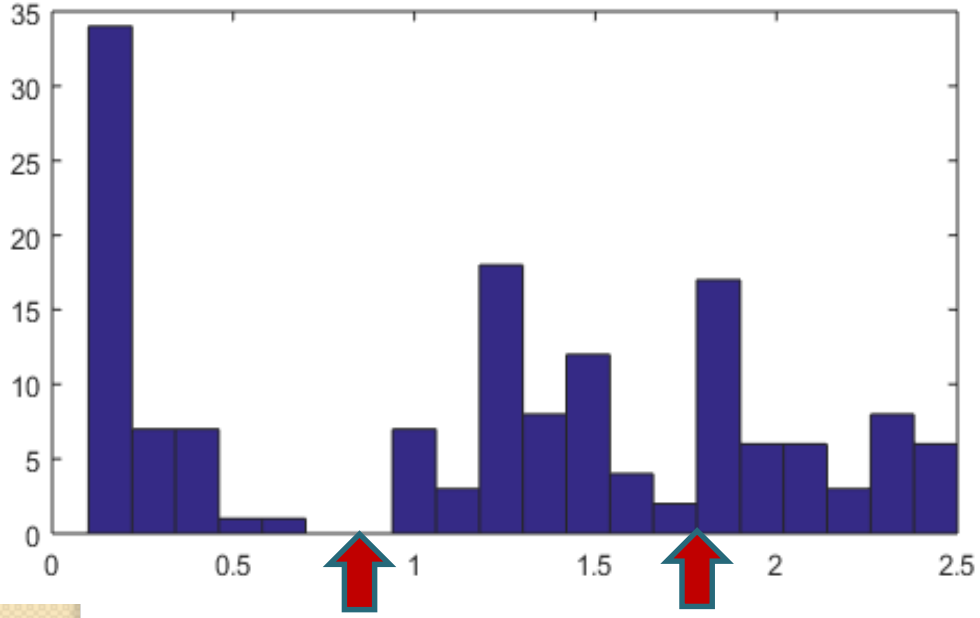
There are deep minima (non-homogeneity)

Homogeneous groups in a non-homogeneous set



- Homogeneous groups in intervals
- A=[0, 0.8], B=[0.8, 1.7], C=[1.7 2.5] -
- - Define a nominal (or order) feature with 3 categories, A, B C.

Categorized feature x, 1

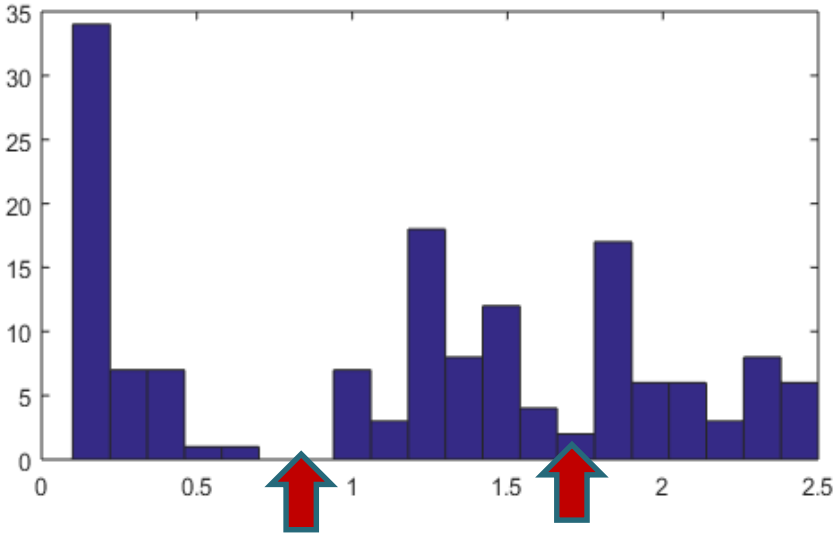


Non-homogeneity

- Nominal feature with 3 categories:
- A=[0, 0.8], B=[0.8, 1.7], C=[1.7 2.5]
- Corresponds to **partition** on the object set: $S=\{S_A, S_B, S_C\}$
- $S_A=\{i: 0<x(i)<=0.8\}$,

$S_B=\{i: 0.8<x(i)<=1.7\}$, $S_C=\{i: 1.7<x(i)<=2.5\}$

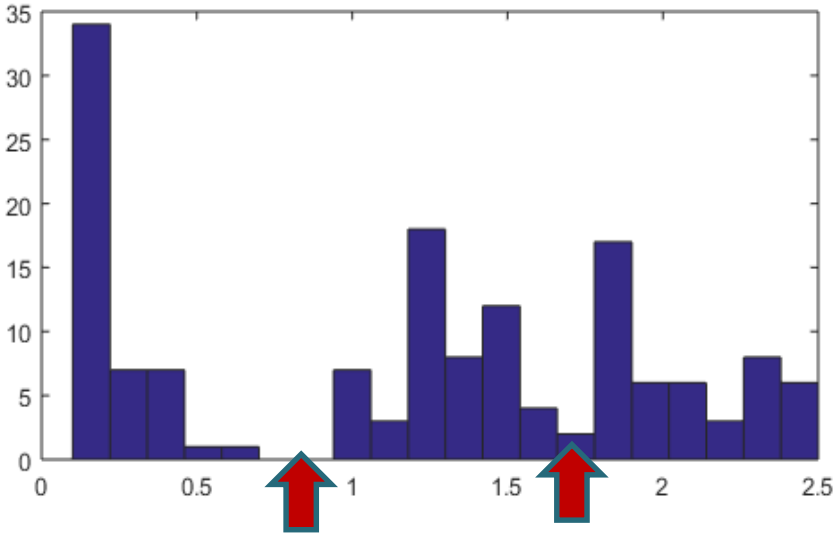
Categorized feature x, 2



- Nominal/order feature with 3 categories:
- A=[0, 0.8], B=[0.8, 1.7], C=[1.7 2.5]
- Corresponds to (ordered) **partition** on the object set:
- S={S{1}, S{2}, S{3}} **What is partition?**
- **Code:**

```
>>f=[0 0.8 1.7 2.5];
>>for k=1:3; s{k}=find(x>f(k) & x<=f(k+1)); end;
```

Categorized feature x, 3: Distribution



- Nominal/order feature with 3 categories:
- A=[0, 0.8], B=[0.8, 1.7], C=[1.7 2.5]
- Corresponds to **partition** of the object set:
  - S={S{1}, S{2}, S{3}}
- Frequencies: [|S{1}|, |S{2}|, |S{3}|}={50, 54, 46}
- Relative frequencies:
- p = {50, 54, 46}/150= {0.33, 0.36, 0.31}

Lecture 2 Contents

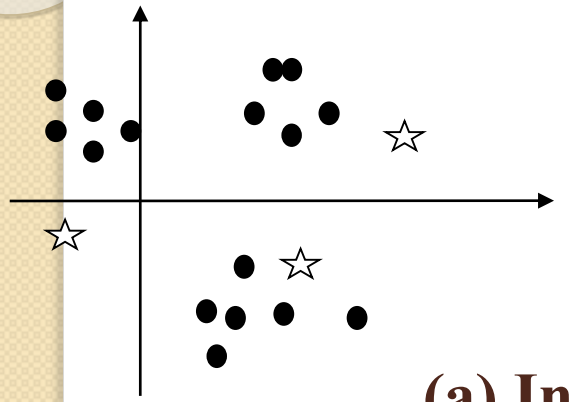
- Two formalizations of the concept of feature: vector and random variable
- Quantitative and categorical features
- Data standardization
- How to categorize a quantitative feature
- Partition and its distribution
- **K-means clustering**
- Interpretation of clusters
- The average and its properties
  - Vector view
  - Density function view
- K-means criterion.

K-Means clustering

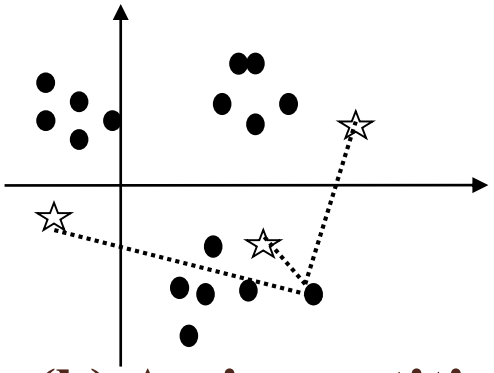
- Method
- Issues
- Interpretation of clusters
- Criterion of the method

Clustering with K-Means, I
K-Means iterations illustrated

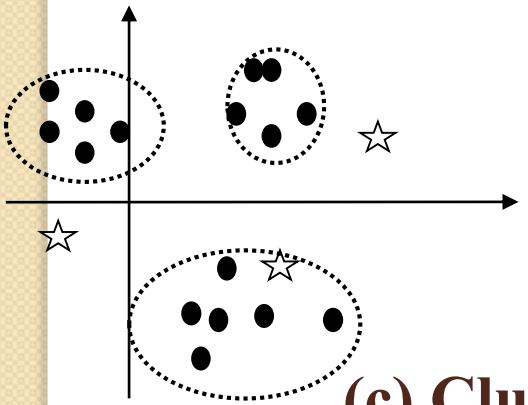
Cluster k: center c_k and set S_k ($k=1,\ldots,K$)



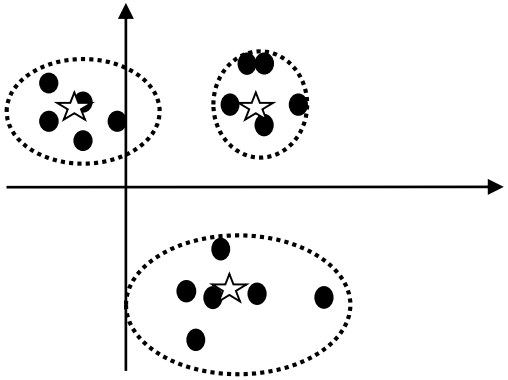
(a) Initialize



(b) Assign entities to nearest center



(c) Cluster update



(d) Center update

Clustering with K-Means, 2
K-Means iterations formulated

Cluster k: center c_k and set S_k ($k=1,\ldots,K$)

K-Means method:

0. Specify K, number of clusters, and initial centers c_k ($k=1,\ldots,K$)

1. **Update sets S_k ($k=1,\ldots,K$) using Minimum distance rule**
2. **Update centers c_k ($k=1,\ldots,K$) as means of S_k**
3. **If new centers coincide with the previous ones, stop. Else go to 1.**

Clustering with K-Means, 3

Explanation of the mean

Let S be in R^4 and consist of 3 objects

i1 =	(2,	1,	2,	0)
i2 =	(1,	2,	0,	1)
i3 =	(3,	0,	1,	5)

Mean:	6/3	3/3	3/3	6/3
Sum/Nk	(2,	1,	1,	2)

Clustering with K-Means, 4

Explanation of the distance

Euclidean squared distance:

dot product of the difference by itself

$$d(i, c_k) = \langle i - c_k, i - c_k \rangle$$

object $i = (2, 1, 2, 0)$

center $c_k = (1, 2, 0, 1)$

$i - c_k = (2-1, 1-2, 2-0, 0-1) = (1, -1, 2, -1)$

Distance: $d(i,c_k) = (1)^2+(-1)^2+(2)^2+(-1)^2 = 7$

Clustering with K-Means, 5

Applying K-Means method to Iris dataset

Step –1. Standardization (options):

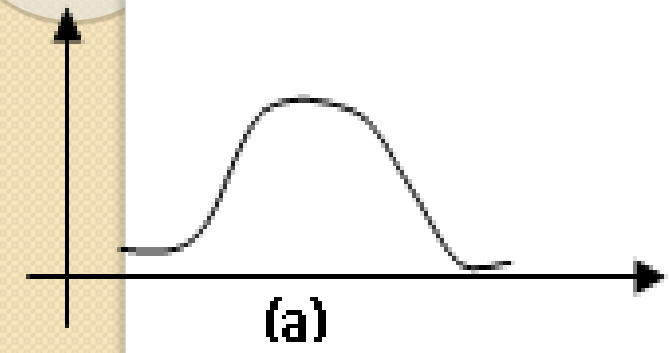
A. No pre-processing
(Why? All measurements relate to elements of the same flower and use the same unit)

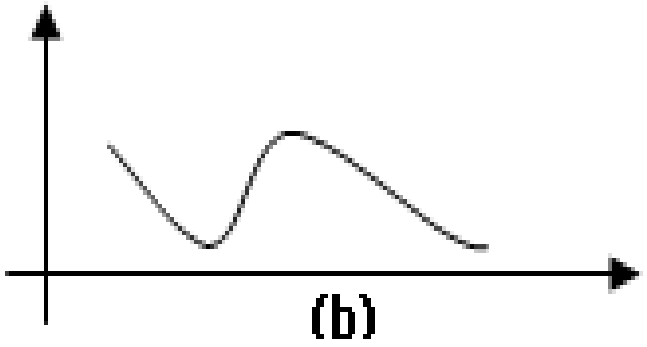
B. Z-scoring: Each feature centered by its mean and normalized by its standard deviation
(Why? Everybody does so)

C. Each feature centered by its mean and normalized by its range (Why? Dividing by range is better than by the standard deviation – see next slide)

Clustering with K-Means , 6

Why dividing by range can be better than dividing by std for clustering





$$\sigma_b >> \sigma_a \quad \Rightarrow \quad \frac{x-m}{\sigma_b} << \frac{x-m}{\sigma_a}$$

Counter-intuitive: (b) divides dataset, thus better, (a) not, thus worse for clustering

Clustering with K-Means,7 Preprocessing options at Iris dataset

0. Specify K=3 and specimens 1, 51, 101 as initial centers (because of preliminary knowledge). Then run iterations of K-Means:

A. (No preprocessing) Converged in 4 iterations

B. (Z scoring) Converged in 7 iterations

C. (Normalizing by range) Converged in 5 iterations

Clustering with K-Means, 8

Confusion regarding Ground Truth

Partition of Iris dataset using K-Means method:

Ground Truth: Taxa	T1	T2	T3	Total	
A. (No preprocessing)	50	0	0	50	Cl1
16 errors	0	48	14	62	Cl2
	0	2	36	38	Cl3
B. (Z scoring)	50	0	0	50	Cl1
28 errors	0	39	17	56	Cl2
	0	11	33	44	Cl3
C. (Norm. by range)	50	0	0	50	Cl1
17 errors	0	47	14	61	Cl2
	0	3	36	39	Cl3

Clustering with K-Means, 10

Advantages

- K-Means computations model typology making (who knows: what is typology?)
- Computation is intuitive
- Computation is fast and requires no additional memory
- Computation is easy to parallelize (big data)

Clustering with K-Means, I I

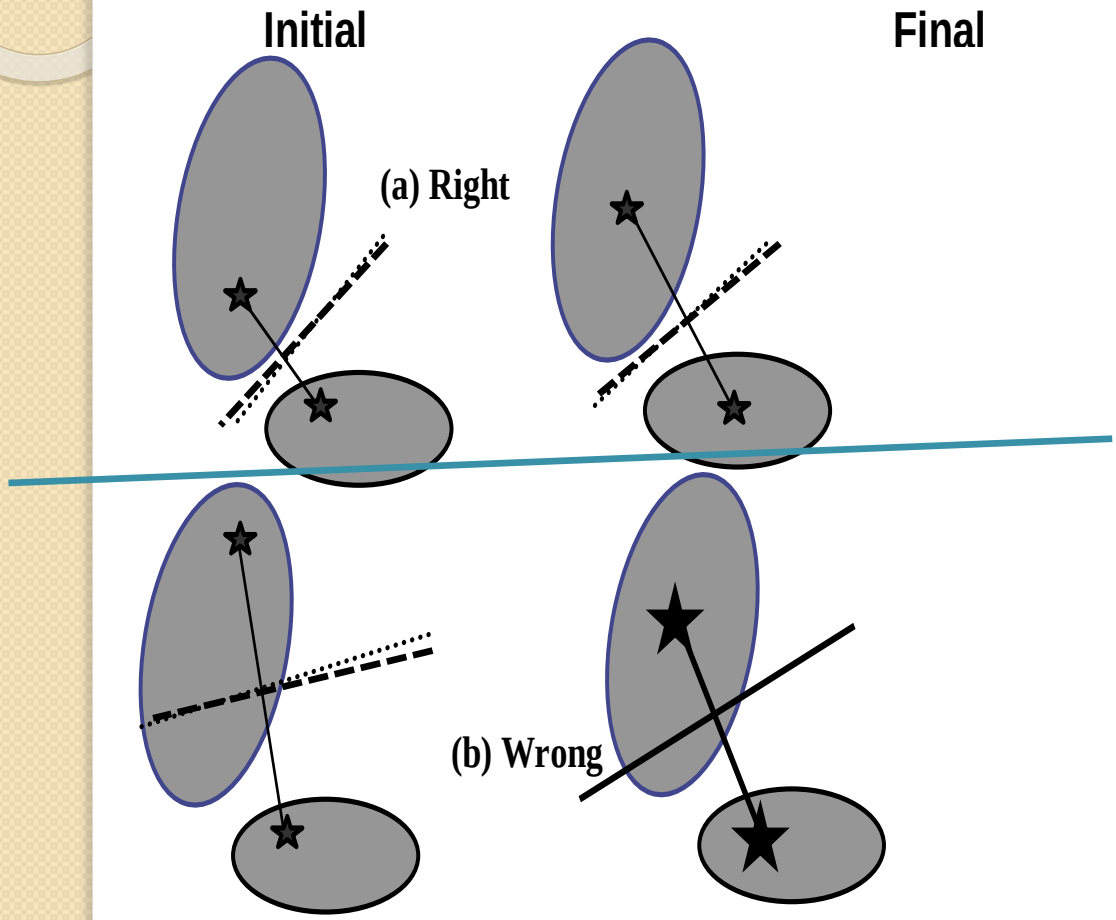
Issues of K-Means

- Would the K-Means computations ever stop?
- Results depend of the initialization (see next slide): How should one initialize?
- Choosing the number of clusters K?

To address, let us do a bit of theory

Clustering with K-Means, 12

Results heavily depend of the initialization



Two clearly visible clusters case.

Top:
Reasonable location of initial centroids

Bottom:
Asymmetric initial centers lead to wrong clustering results

Lecture 2 Contents

- Two formalizations of the concept of feature: vector and random variable
- Quantitative and categorical features
- Data standardization
- How to categorize a quantitative feature.
- Partition and its distribution.
- K-means clustering.
- **Interpretation of clusters through means**
- The average and its properties
  - Vector view
  - Density function view
- K-means criterion.

Clustering with K-Means, 9

Cluster interpretation

Center c_k for Interpretation of cluster S_k:

Iris taxon T1: Specimens number 1, 2, …, 50

Taxon T1 Interpretation: SMALL PETAL

	SLength	SWidth	PLength	PWidth
Within-cluster mean	**5.006**	**3.428**	**1.462**	**0.246**
Grand mean	5.843	3.057	3.758	1.199
Difference	–0.837	0.371	–2.296	–0.953
Difference, %	–14.3	+12.1	**–61.1**	**–79.5**

Lecture 2 Contents

- Two formalizations of the concept of feature: vector and random variable
- Quantitative and categorical features
- Data standardization
- How to categorize a quantitative feature.
- Partition and its distribution.
- K-means clustering.
- Interpretation of clusters through means
- **The average/mean and its properties**
  - **Vector view**
  - **Density function view**
- K-means criterion.

What is **a vector feature center,** I

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**

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

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

What is **a feature vector center,** 2

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

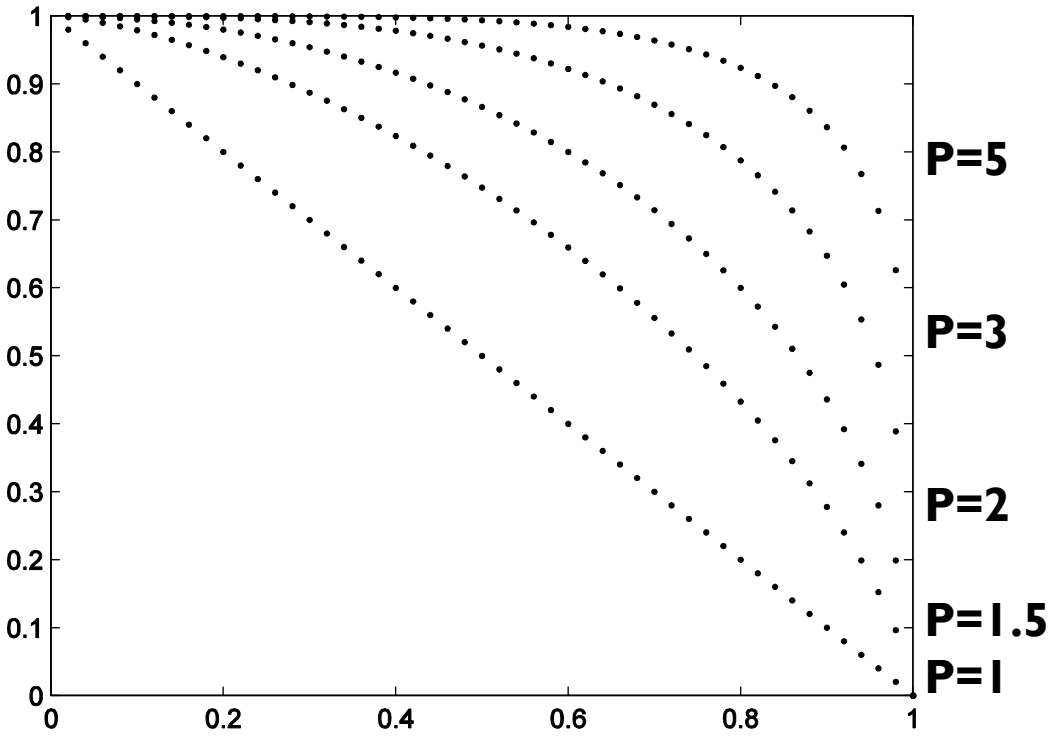
Minimize $L_p=[|c-x_1|^p+|c-x_2|^p+\ldots+|c-x_N|^p\ ]/N$

with respect to all possible c

At different p, different solutions!

L_p is a measure of spread of the feature around center

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



What is **center, 3**

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 $L_p = L_2$ is quadratic. First-order minimum condition can be applied leading to optimal

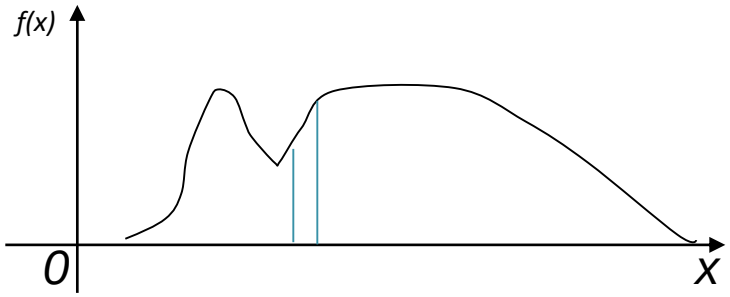
$$c=\text{Mean}(x) = \frac{1}{N}\sum_{i=1}^{N} x_i$$

At this c,

L_2 is the variance, the squared standard deviation!

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

What is center, 4: Classical statistics view

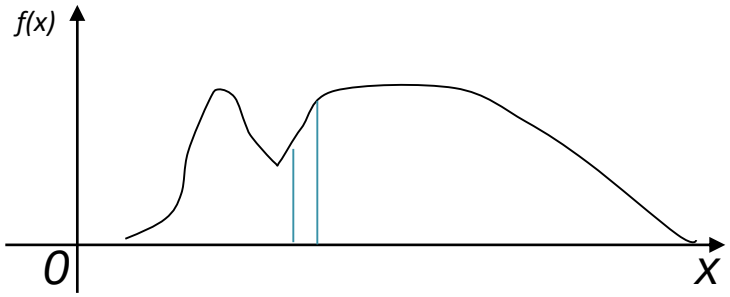


Unidimensional data $x=(x_i)$ is a set of N independent random variables with the same density function $f(x)$.

Denote its mathematical expectation by μ and variance by σ^2. The former is defined as $ME(f) = \int_{-\infty}^{+\infty} \boldsymbol{x f(x) dx}$; the latter as $ME([f(x)-ME(f)]^2)$.

In contrast to simple-minded expectation, the density of $x_1 + x_2$ is not $2f(x)$ but a much more complicated function, the so-called convolution of $f(x)$ with itself.

What is center, 5: Classical statistics view

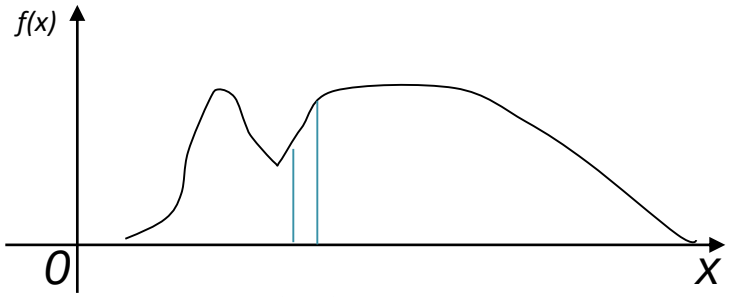


Unidimensional data $x=(x_i)$ is a set of N independent random variables with the same density function $f(x)$.

Denote its mathematical expectation by μ and variance by σ^2. The former is defined as $ME(f) = \int_{-\infty}^{+\infty} \boldsymbol{x f(x) dx}$; the latter as $ME([f(x)-ME(f)]^2)$.

Relation to DA view: **the DA's mean is unbiased estimate of the Mathematical Expectation!**

What is center, 6: Classical statistics view



Unidimensional data $x=(x_i)$ is considered a set of N independent random variables with the same density function $f(x)$.

Central limit theorem:

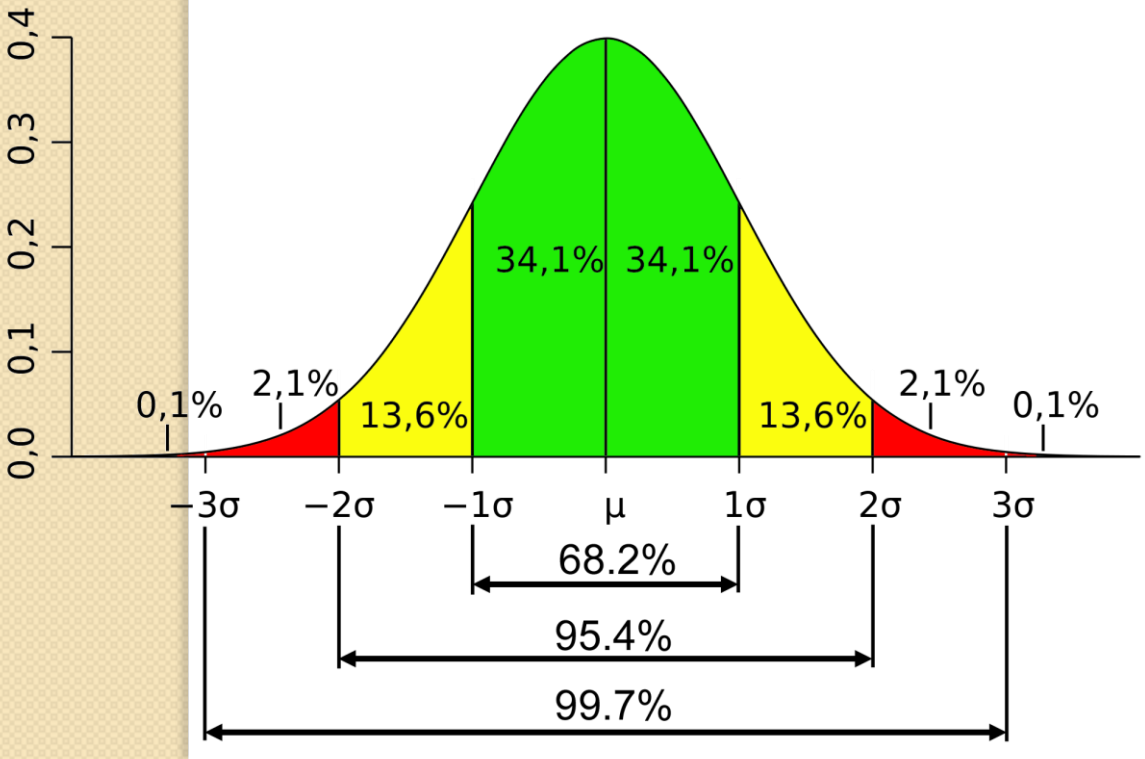
The distribution (density function) of sum $x_1 + x_2 + \ldots + x_N$ converges to Gaussian distribution with the mathematical expectation μ and variance σ^2.

The density function of the average $\bar{x} = (x_1 + x_2 + \ldots + x_N)/N$ converges to Gaussian with $ME=\mu$ and variance σ^2/N .

Center, 7: Classical statistics view

- The average's density function approximates Gaussian $N(\mu,\sigma/\sqrt{N})$
- Thus, central interval to account for 95% of the area:

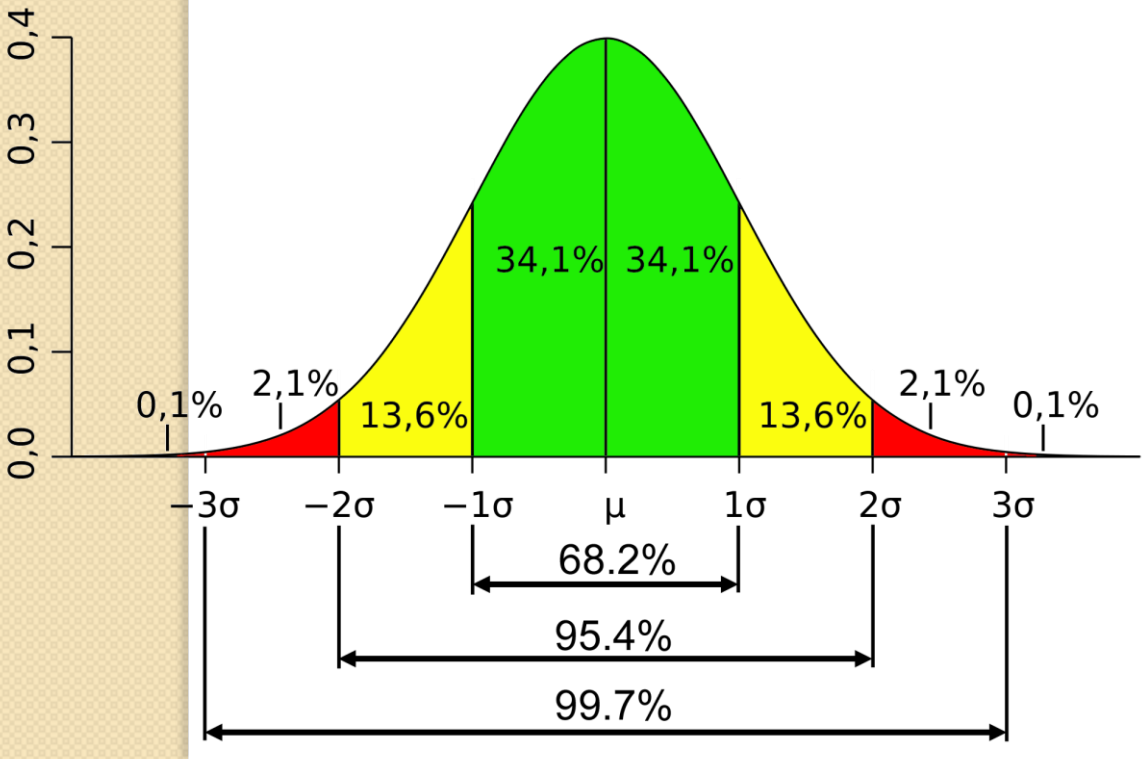
$[\mu - 1.96\sigma/\sqrt{N}, \mu + 1.96\sigma/\sqrt{N}]$



Center, 8: Classical statistics view

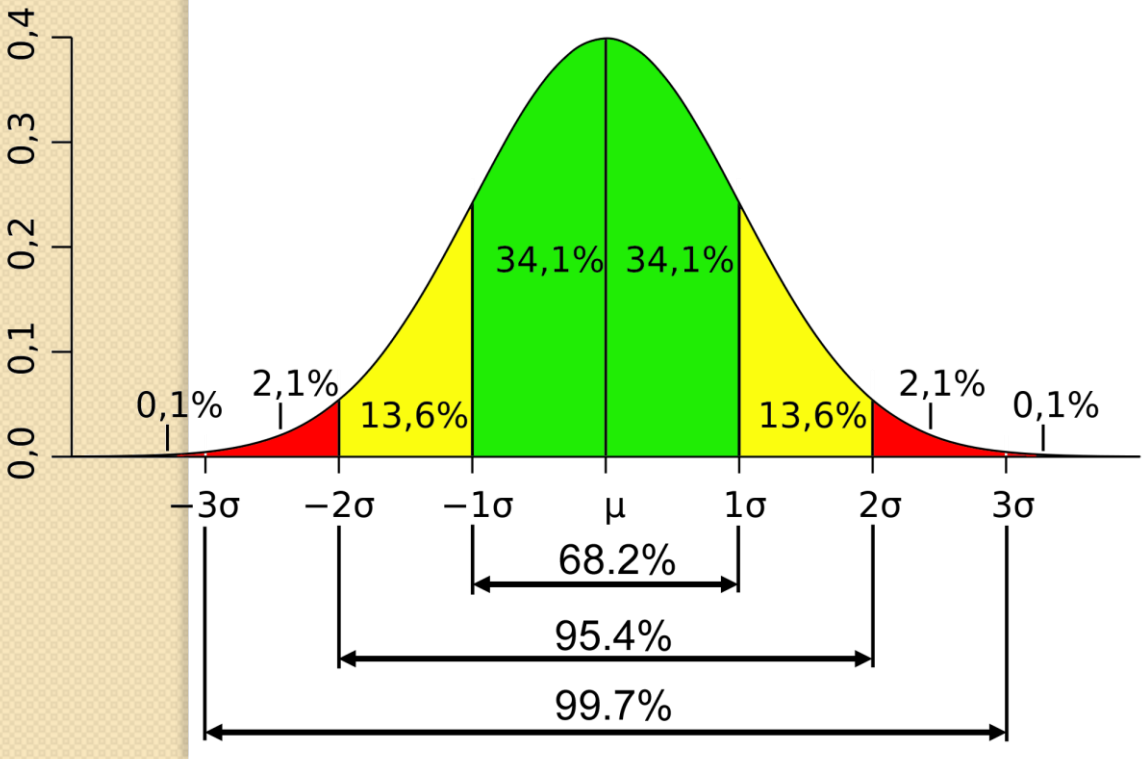
- To compare the averages μ_1 of sample of N_1 objects and μ_2 of independent sample of N_2 objects, consider the difference $\mu = \mu_1 - \mu_2$ and. If the hypothesis is true, $\mu = 0$. To test this, notice that μ is $N(0,\sigma)$ where $\sigma = \sigma_1/\sqrt{N_1} + \sigma_2/\sqrt{N_2}$.
- The central interval to account for 95% of the area:

$[\mu - 1.96\sigma/\sqrt{N}, \mu + 1.96\sigma/\sqrt{N}]$. If 0 does not belong to this, the hypothesis is rejected with 95% confidence.



Center, 9: Classical statistics view

- The distribution density function and its central interval to account for 95% of the area can be estimated computationally via what is called **bootstrap** (will be studied later)
- If 0 does not belong to this, the hypothesis is rejected with 95% confidence.

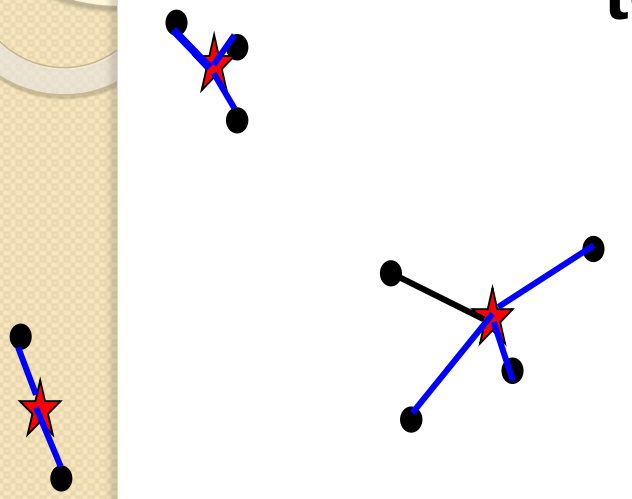


Lecture 2 Contents

- Two formalizations of the concept of feature: vector and random variable
- Quantitative and categorical features
- Data standardization
- How to categorize a quantitative feature.
- Partition and its distribution.
- K-means clustering.
- Interpretation of clusters
- The average and its properties
  - DA: Vector view
  - CS: Density function view
- **K-means criterion and convergence**

Clustering with K-Means, 13

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]

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

K-Means criterion, I

K-Means = alternating minimization of *D(S,c)*

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*.

Min_S *D(S, c)*:
- Clusters update

Min_c *D(S, c)*:
- Centers update

D(S,c) decreases at each step:

Convergence – why? (QUIZ)

Distance (Squared Euclidean):

X= [1, 2, -2]
Y= [1, -1, -1]
X-Y= [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"*

Keeping up: How to prepare yourself to the next lecture:

- After the lecture, put down **main concepts** that have been discussed in the lecture and think a few minutes of what do they mean

- Just before the next lecture: Take a few minutes and look through the slides of the previous lecture

Home work 1:

- **1.** Each to form/join a team of one or two; the team **finds a meaningful dataset** of their liking **on the internet** as advised in Lecture 1
- Number of entities ≥ 100, of features ≥ 7
- **No missings**
- **No Irivine ML repository**
- **The dataset is to be approved by the instructor (myself).**

- **2. Start writing a team's report file**
- Project title page
- Section 1.
  - Explanation of the choice of the dataset
  - Information of the dataset: features, number of entities, source address, examples of problems