



Chapter 2. Getting to Know Your Data: Data Description

Meng Jiang

CSE 40647/60647 Data Science Fall 2017

Introduction to Data Mining

From Data to Knowledge

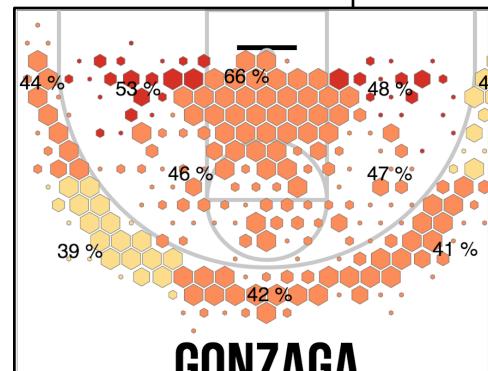
HOME TEAM: Notre Dame 26-9

##	Player Name	TOT-FG			3-PT			REBOUNDS							
		FG-FGA	FG-FGA	FT-FTA	OF	DE	TOT	PF	TP	A	TO	BLK	S	MIN	
03	VJ Beachem.....	f	1-9	0-3	0-0	0	6	6	1	2	3	0	0	1	37
35	<u>Bonzie Colson</u>	f	6-13	0-1	6-10	2	5	7	2	18	2	0	2	1	31
00	Rex Pflueger.....	g	2-3	0-0	0-0	0	2	2	2	4	0	1	0	0	28
05	<u>Matt Farrell</u>	g	6-9	3-5	1-3	0	4	4	2	16	4	3	0	2	36
32	<u>Steve Vasturia</u>	g	3-12	1-2	3-4	3	5	8							
01	<u>Austin Torres</u>		0-1	0-0	0-0	1	0	1							
02	TJ Gibbs.....		0-1	0-0	2-2	0	2	2							
04	<u>Matt Ryan</u>		2-3	0-0	2-2	0	2	2							
23	<u>Martinas Geben</u>		1-1	0-0	0-0	1	0	1							
	TEAM.....						2	1	3						
	Totals.....		21-52	4-11	14-21	9	27	36							

TOTAL FG% 1st Half: 14-30 46.7% 2nd Half: 7-22 31.8%

3-Pt. FG% 1st Half: 2-5 40.0% 2nd Half: 2-6 33.3%

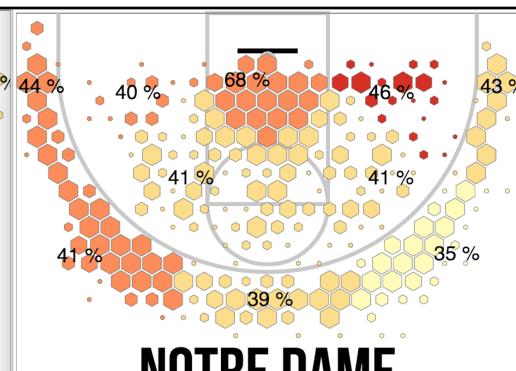
F Throw % 1st Half: 6-8 75.0% 2nd Half: 8-13 61.5%



GONZAGA

WWW.SHOTANALYTICS.COM

LOW VOLUME ● ● ● HIGH VOLUME
BELOW AVG ● ● ● ABOVE AVG



NOTRE DAME

WWW.SHOTANALYTICS.COM

LOW VOLUME ● ● ● HIGH VOLUME
BELOW AVG ● ● ● ABOVE AVG

##	Player	gp-gs	min	avg	fg-fga	fg%	3fg-fga	3fg%	ft-fta	ft%	off	def	tot	avg	pf	dq	a	to	blk	stl	pts	avg
35	Colson,Bonzie	36-36	1156	32.1	236-449	.526	26-60	.433	141-180	.783	104	258	362	10.1	81	0	56	44	50	40	639	17.8
03	Beachem,VJ	36-36	1230	34.2	187-443	.422	87-241	.361	61-73	.836	23	123	146	4.1	46	0	31	40	38	33	522	14.5
05	Farrell,Matt	36-36	1238	34.4	172-384	.448	81-193	.420	81-102	.794	9	63	72	2.0	71	0	196	91	5	51	506	14.1
32	Vasturia,Steve	36-36	1244	34.6	162-374	.433	58-162	.358	91-100	.910	25	116	141	3.9	75	1	119	57	4	42	473	13.1
00	Pflueger,Rex	35-11	750	21.4	59-133	.444	27-68	.397	19-29	.655	18	78	96	2.7	60	0	53	24	12	31	164	4.7
02	Gibbs,TJ	36-1	539	15.0	51-136	.375	17-53	.321	49-59	.831	12	41	53	1.5	49	0	62	28	2	26	168	4.7
04	Ryan,Matt	36-0	286	7.9	43-99	.434	36-83	.434	9-10	.900	6	26	32	0.9	29	0	14	13	1	6	131	3.6
23	Geben,Martinas	34-23	421	12.4	42-65	.646	0-0	.000	23-30	.767	42	73	115	3.4	66	3	25	22	11	13	107	3.1
01	Torres,Austin	36-1	261	7.3	21-38	.553	0-0	.000	6-18	.333	23	31	54	1.5	46	0	7	9	8	9	48	1.3
33	Mooney,John	12-0	46	3.8	5-8	.625	2-4	.500	2-2	1.000	6	13	19	1.6	5	0	2	1	1	1	14	1.2
12	Burns,Elijah	11-0	44	4.0	1-4	.250	0-1	.000	7-8	.875	5	5	10	0.9	5	0	1	2	1	2	9	0.8
34	Mazza,Patrick	4-0	4	1.0	1-2	.500	0-0	.000	0-0	.000	0	1	1	0.3	0	0	0	1	1	0	2	0.5
21	Gregory,Matt	5-0	6	1.2	0-4	.000	0-4	.000	0-0	.000	0	0	0	0.0	0	0	0	0	0	0	0	0.0

Chapter 2. Getting to Know Your Data

- **Data Objects and Attribute Types**
- Basic Statistical Descriptions
- Data Visualization
- Measuring Data Similarity and Dissimilarity

Types of Data Sets: (1) Record Data

- Relational records in relational tables: highly structured
- Transaction data
- Document data: Term-frequency matrix of text documents

HOME TEAM: Notre Dame 26-9															
##	Player Name	TOT-FG			3-PT			REBOUNDS							
		FG-FGA	FG-FGA	FT-FTA	OF	DE	TOT	PF	TP	A	TO	BLK	S	MIN	
03	VJ Beachem.....	f	1-9	0-3	0-0	0	6	6	1	2	3	0	0	1	37
35	<u>Bonzie Colson</u>	f	6-13	0-1	6-10	2	5	7	2	18	2	0	2	1	31
00	<u>Rex Pflueger</u>	g	2-3	0-0	0-0	0	2	2	2	4	0	1	0	0	28
05	<u>Matt Farrell</u>	g	6-9	3-5	1-3	0	4	4	2	16	4	3	0	2	36
32	<u>Steve Vasturia</u>	g	3-12	1-2	3-4	3	5	8	0	10	1	0	0	0	37
01	<u>Austin Torres</u>		0-1	0-0	0-0	1	0	1	0	0	0	1	1	0	7
02	TJ Gibbs.....		0-1	0-0	2-2	0	2	2	1	2	0	0	0	0	13
04	<u>Matt Ryan</u>		2-3	0-0	2-2	0	2	2	0	6	0	0	0	0	9
23	<u>Martinas Geben</u>		1-1	0-0	0-0	1	0	1	1	2	0	1	0	0	2
TEAM.....						2	1	3							
Totals.....			21-52	4-11	14-21	9	27	36	9	60	10	6	3	4	200
TOTAL FG% 1st Half: 14-30 46.7%				2nd Half: 7-22 31.8%				Game: 40.4% DEADB							
3-Pt. FG% 1st Half: 2-5 40.0%				2nd Half: 2-6 33.3%				Game: 36.4% REBS							
F Throw % 1st Half: 6-8 75.0%				2nd Half: 8-13 61.5%				Game: 66.7% 3							

Data Objects

- Data sets are made up of data objects.
- A **data object** represents an entity.
- Examples:
 - Sales database: customers, store items, sales.
 - Medical database: patients, treatments.
 - University database: students, professors, courses.
- Also called *samples, examples, instances, data points, objects, tuples*.
- Data objects are described by **attributes**.
- Database rows → data objects; columns → attributes.

Attributes

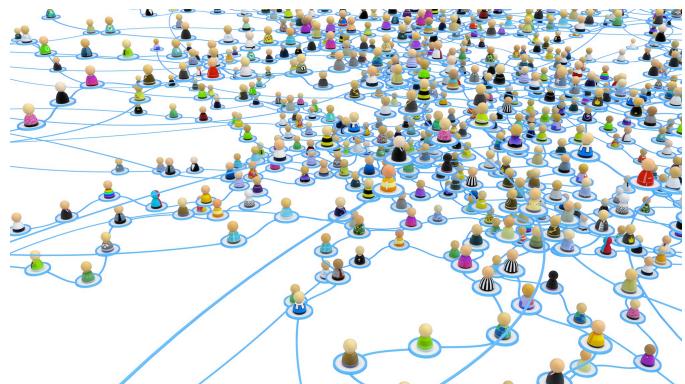
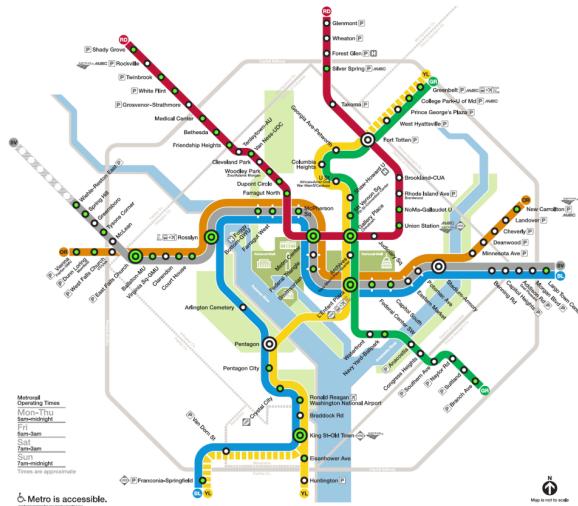
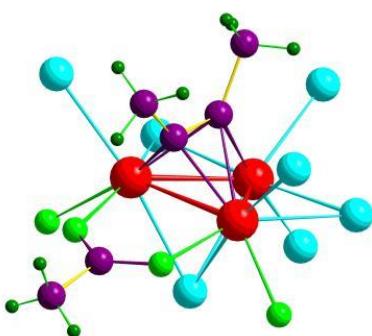
- **Attribute (or dimensions, features, variables)**
 - A data field, representing a characteristic or feature of a data object
- Types:
 - Nominal (e.g., red, blue)
 - Binary (e.g., {true, false})
 - Ordinal (e.g., {freshman, sophomore, junior, senior})
 - Numeric: quantitative
 - Interval-scaled: 100°C is interval scales
 - Ratio-scaled: 100°K is ratio scaled since it is twice as high as 50°K

Attribute Types

- **Nominal:** categories, states, or “names of things”
 - *Hair_color* = {*auburn*, *black*, *blond*, *brown*, *grey*, *red*, *white*}
 - marital status, occupation, ID numbers, zip codes
- **Binary**
 - Nominal attribute with only 2 states (0 and 1)
 - Symmetric binary: both outcomes equally important
 - e.g., gender
 - Asymmetric binary: outcomes not equally important.
 - e.g., medical test (positive vs. negative)
 - Convention: assign 1 to most important outcome (e.g., HIV positive)
- **Ordinal**
 - Values have a meaningful order (ranking) but magnitude between successive values is not known
 - *Size* = {*small*, *medium*, *large*}, grades, army rankings

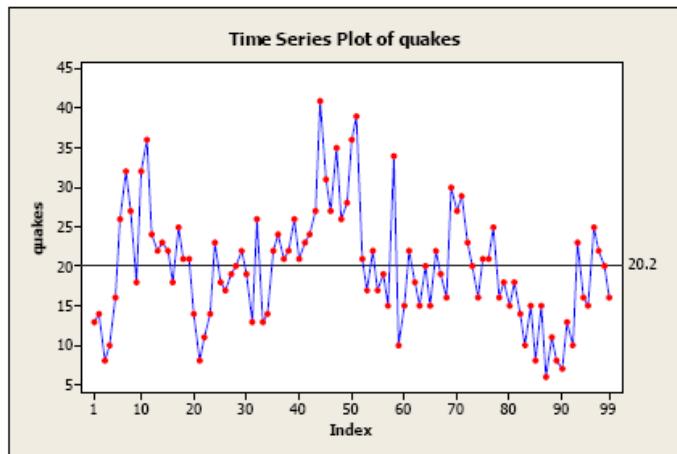
Types of Data Sets: (2) Graphs and Networks

- Transportation networks
- World Wide Web
- Molecular structures
- Social or information networks



Types of Data Sets: (3) Ordered Data

- Video data: sequence of images
- Temporal data: time-series
- Sequential Data: transaction sequences
- Genetic sequence data

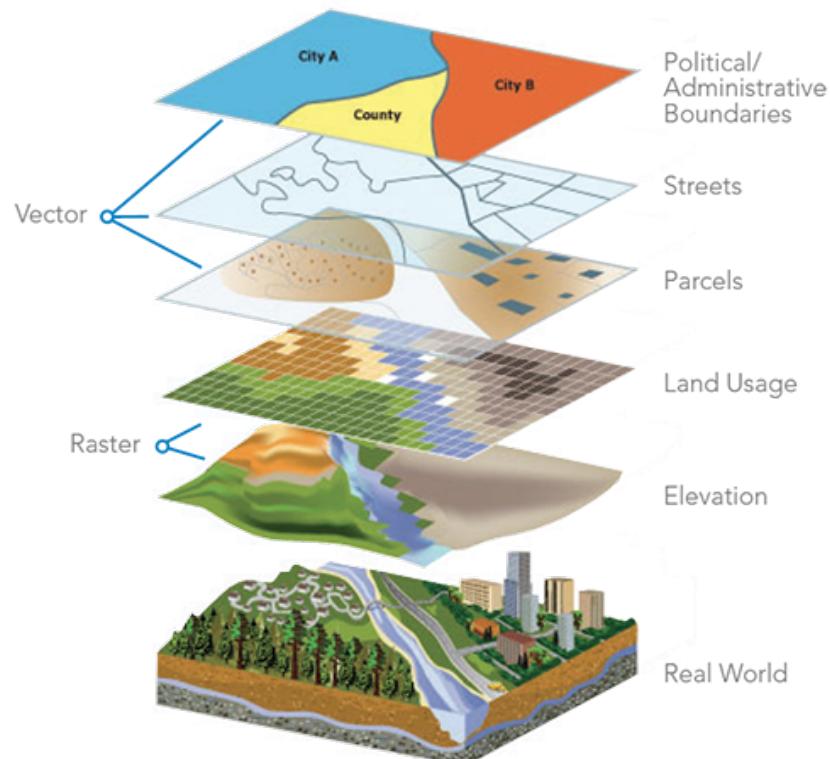


Start

Human	GTTTGAGG	- - ATGTTCAACAAATGCTCCTTCATTCCCTTCTATTACAGACCTGCCGCA
Chimpanzee	GTTTGAGG	- - ATGTTCAATAAATGCTGCTTCACTCCTTCTATTACAGACCTGCCGCA
Macaque	GTTTGAGG	- - ATGCTCAATAAATGCTCCTTCATTCCCTCATTACAAACTGCCGCA
Human	GACAATTCTGCTAGCAGCCTTGTGCTATTATCTGTTTCTAAACCTTAGTAATTGAGTGT	
Chimpanzee	GACAATTCTGCTAGCAGCCTTGTGCTATTATCTGTTTCTAAACCTTAGTAATTGAGTGT	
Macaque	GACAATTCTGCTAGCAGCCTTGTGCTATTATCTGTTTCTAAACCTTAGTAATTGAGTGT	
Human	GATCTGGAGACTAA - CTC TGAAATAAAATAAGCTGATTATTTATTTCTCAAAACAA	
Chimpanzee	GATCTGGAGACTAAACTCTGAAATAAAATAAGCTGATTATTTATTTCTCAAAACAA	
Macaque	TATCTGGAGACTAAACTCTGAAATAAAATAAGCTGATTATTTATTTATTTCTCAAAACAA	
Human	CAGAAATACGATTTAGCAAATTACTCTTAAGATAATTATTTACATTTCTATATTCTCCTA	
Chimpanzee	CAGAAATACGATTTAGCAAATTACTCTTAAGATACTATTACATTTCTATATTCTCCTA	
Macaque	CAGAAATATGATTTAGCAAATTACCTCTTAAGATAATTATTTGCACATTCTATATTCTCCTA	
Human	CCCTGAGTTGATGTGTGAGCAATATGTCACTTTCATAAAAGCCAGGTATAACA - - - TTATG	
Chimpanzee	CCCTGAGTTGATGTGTGAGCGTATGTCACTTTCATAAAAGCCAGGTATAACA - - - TTATG	
Macaque	CCCTGAGTTGATGTGTGAGCAATATGTCACTTCCACAAAGCCAGGTATAATAACATTACG	
Human	GACAGGTAAAGTAAAAAACATATTATTATTCAGTTTGTCCAAAATTTAAATTTC	
Chimpanzee	GACAGGTAAAGTAAAAAACATATTATTATTCAGTTTGTCCAAAAGATTAAATTTC	
Macaque	GACAGGTAAAGTAAAAA-CATATTATTATTCAGTTTGTCCAAAAGAGTTAAATTTC	
Human	AAC TGT TGC CGCGTGT GTGGTAA - - TGT AAA AAC AA AC TC AGT ACA	
Chimpanzee	AAC TGT TGC CGCGTGT GTGGTAA - - TGT AAA AAC AA AC TC AGT ACA	
Macaque	AAC TGT TGT GCA TGT GTGGTAA - - CGT AAA AAC AA AT TC AGT ACA	

Other Types of Data Sets

- Spatial data
- Image and multimedia data

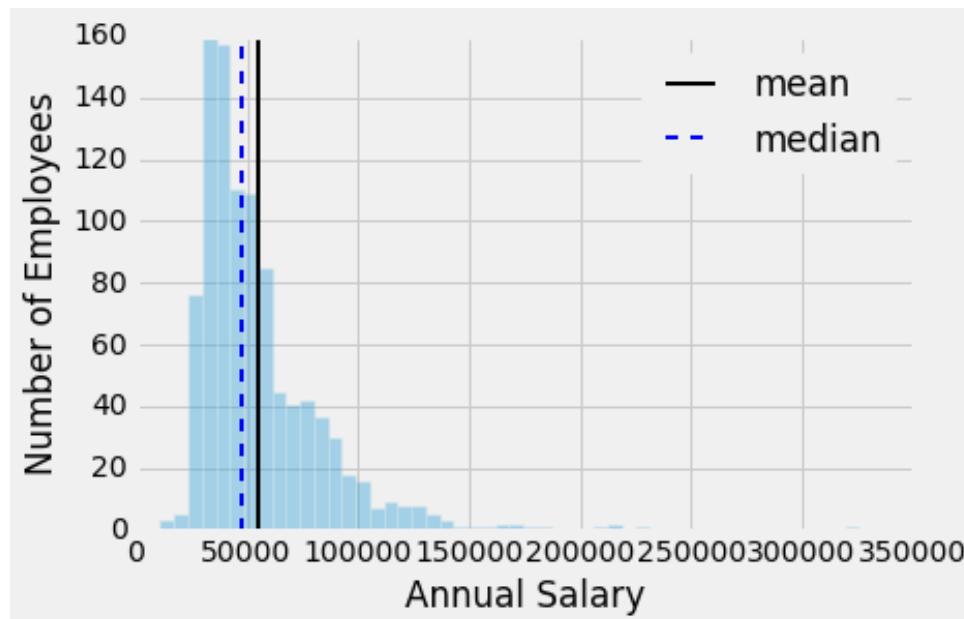


Chapter 2. Getting to Know Your Data

- Data Objects and Attribute Types
- **Basic Statistical Descriptions**
- Data Visualization
- Measuring Data Similarity and Dissimilarity

Basic Statistical Descriptions of Data

- Motivation: to better understand the data
- Data dispersion characteristics
 - Central tendency: Mean, median, mode, max, min ...
 - Outlierness: Variance, standard deviation, Z-score ...



Measuring the Central Tendency:

(1) Mean

- Mean (algebraic measure) (sample vs. population):
 - Note: n is **sample** size and N is **population** size.

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad \mu = \frac{1}{N} \sum_{i=1}^N x_i$$

- Weighted arithmetic mean:

$$\bar{x} = \frac{\sum_{i=1}^n w_i x_i}{\sum_{i=1}^n w_i}$$

- Trimmed mean: Chopping extreme values

Measuring the Central Tendency: (2) Median

- Median:
 - Middle value if odd number of values, or average of the middle two values otherwise
- Estimated by interpolation (for grouped data):

Finding the **median** from a grouped frequency table

Example: Parcels

Calculate an **estimate of the median** weight, shown to the nearest gram, in the following grouped frequency table:

Weight (g)	1-10	11-20	21-30	31-40	41-50
Frequency	10	13	28	15	9

Which interval does the median lie?

Cumulative frequency

Finding the median from a grouped frequency table

Weight (g)	1-10	11-20	21-30	31-40	41-50
Frequency	10	13	28	15	9
Cumulative Frequency	10	23	51	66	75

The median lies in the $\frac{1}{2} (75) = 37.5$ therefore:
 38^{th} position

Therefore the **median** lies in the **21 – 30** class.

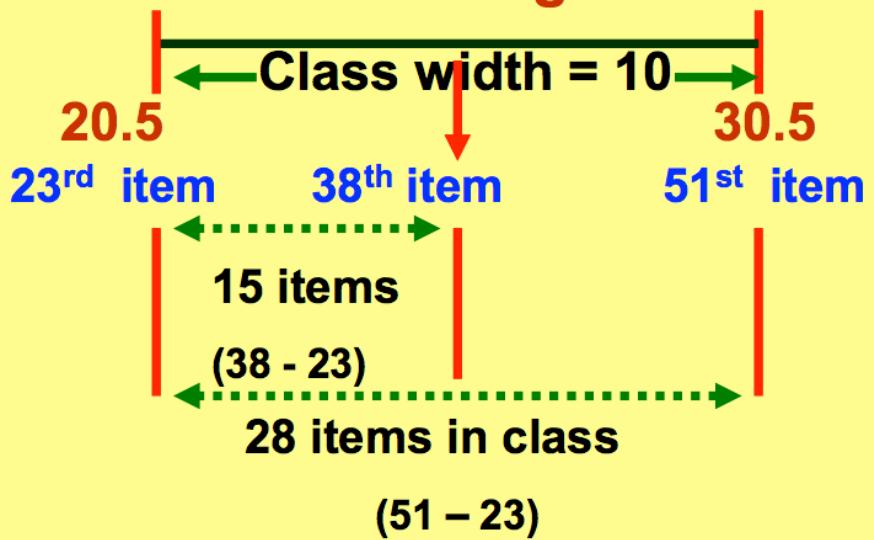
BUT

Remember it was rounded:

ACTUAL CLASS BOUNDARY: 20.5 – 30.5

Lower class boundary + number of items up to median x class width
 Number of items in the class

Useful diagram:



Assumption:

data are evenly spread over each class interval

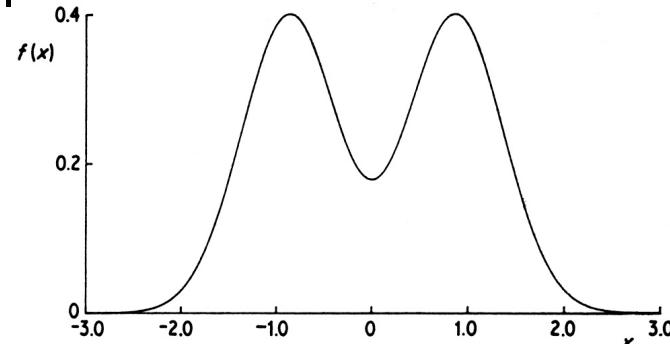
$$\begin{aligned}
 & 20.5 + \frac{15}{28} \times 10 \\
 & = 25.85714.... \\
 & = 25.9(1d.p.)
 \end{aligned}$$

$$\text{median} = L_1 + \left(\frac{n/2 - (\sum \text{freq})_l}{\text{freq}_{\text{median}}} \right) \text{width}$$

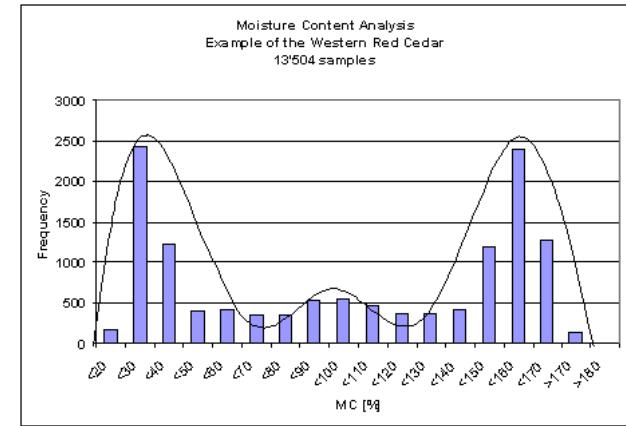
Measuring the Central Tendency:

(3) Mode

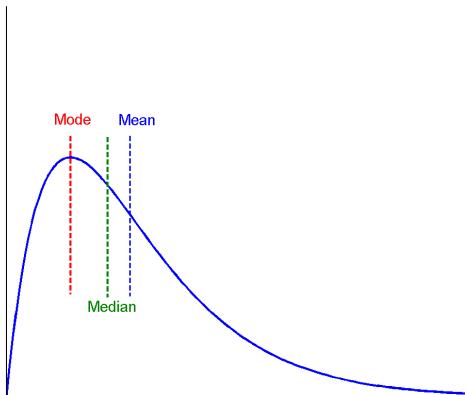
- Mode: Value that occurs most frequently in the data
- Multi-modal
 - Bimodal
 - Trimodal



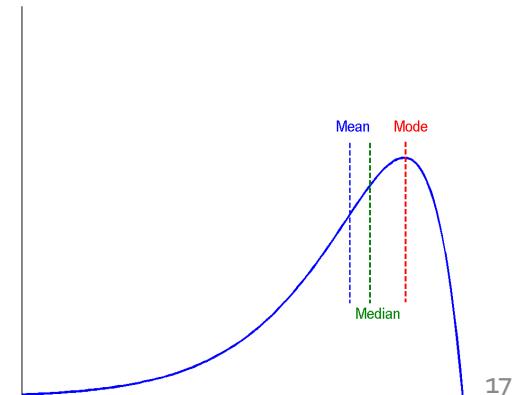
Symmetric data



Positively skewed data



Negatively skewed data



Measuring the Outlierness: Variance and Standard Deviation

- Variance and standard deviation (sample: s , population: σ)
 - Variance: (algebraic, scalable computation)

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \quad \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

Why?

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2 \quad \mu = \frac{1}{N} \sum_{i=1}^N x_i$$

- Standard deviation s (or σ) is **square root** of variance s^2 (or σ^2)

Biased Sample Variance

$$\begin{aligned}\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 &= \frac{1}{n} \sum_{i=1}^n [(X_i - \mu) + (\mu - \bar{X})]^2 \\ \text{Biased} &= \frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2 + \frac{2}{n} \sum_{i=1}^n (X_i - \mu)(\mu - \bar{X}) + \frac{1}{n} \sum_{i=1}^n (\mu - \bar{X})^2 \\ &= \frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2 + 2(\bar{X} - \mu)(\mu - \bar{X}) + (\mu - \bar{X})^2 \\ &= \frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2 - (\mu - \bar{X})^2\end{aligned}$$

Unbiased

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

Bessel's Correction: 3 alternative proofs of correctness

Measuring the Outlierness: Variance and Standard Deviation

- Variance and standard deviation (sample: s , population: σ)
 - Variance: (algebraic, scalable computation)
 - **Q: Can you compute it incrementally and efficiently?**

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

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

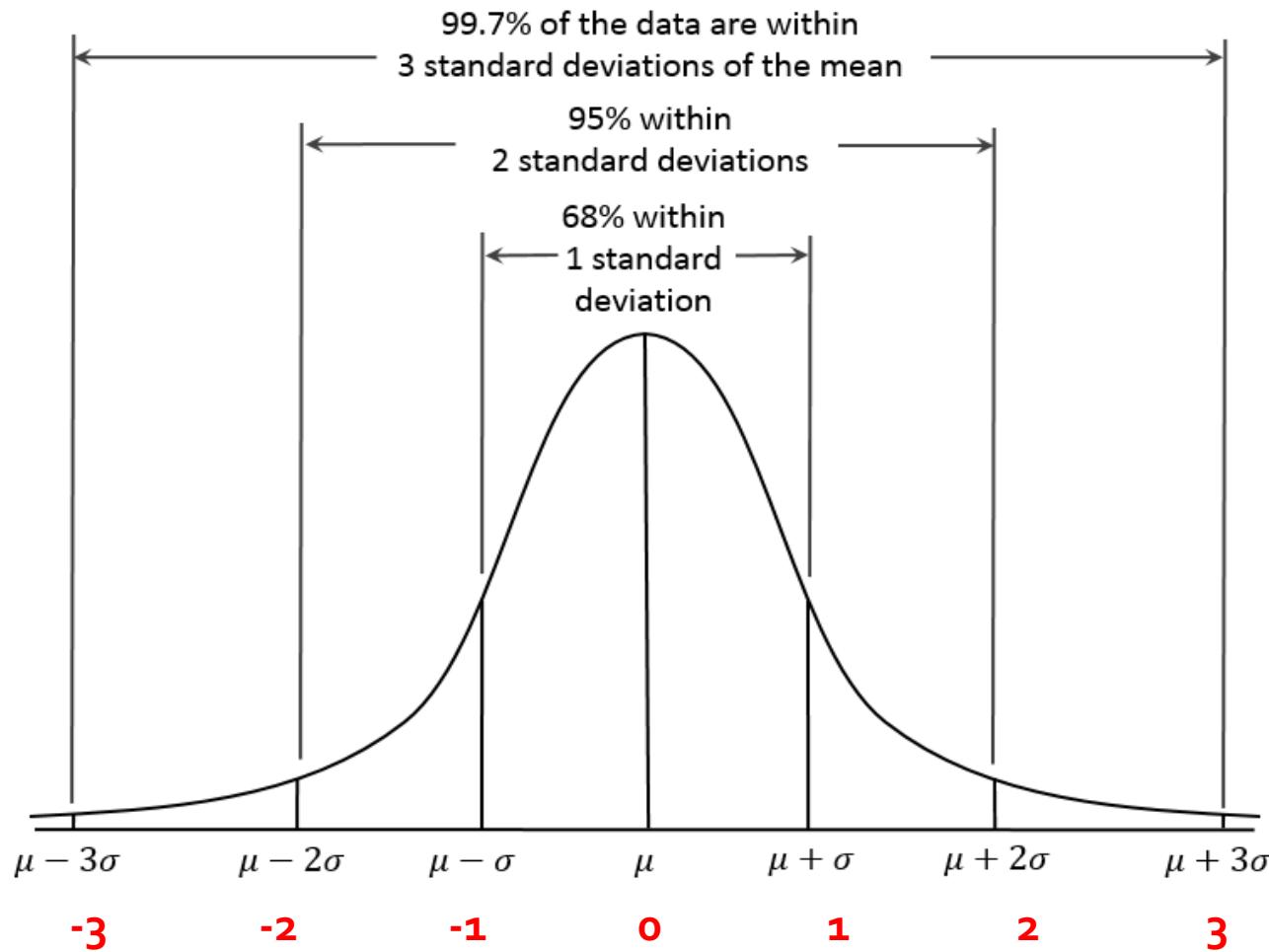
$$\sigma^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2$$

$$\mu = \frac{1}{N} \sum_{i=1}^N x_i$$

- Standard deviation s (or σ) is **square root** of variance s^2 (or σ^2)

Measuring the Outlierness: Properties of Normal Distribution Curve

Z-score: The distance between the raw score and the population mean in the unit of the standard deviation



Z-score Normalization

- The normalized value of X_i is calculated as:

$$Z_i = \frac{X_i - \bar{X}}{S}$$

$$\mathbf{y} = \begin{bmatrix} 35 \\ 36 \\ 46 \\ 68 \\ 70 \end{bmatrix} \quad s = \sqrt{\frac{(35-51)^2 + (36-51)^2 + (46-51)^2 + (68-51)^2 + (70-51)^2}{5-1}}$$
$$= \frac{1}{2} \sqrt{(-16)^2 + (-15)^2 + (-5)^2 + 17^2 + 19^2}$$
$$= 17.$$

$$\mathbf{z} = \begin{bmatrix} \frac{35-51}{17} \\ \frac{36-51}{17} \\ \frac{46-51}{17} \\ \frac{68-51}{17} \\ \frac{70-51}{17} \end{bmatrix} = \begin{bmatrix} -\frac{16}{17} \\ -\frac{15}{17} \\ -\frac{5}{17} \\ \frac{17}{17} \\ \frac{19}{17} \end{bmatrix} = \begin{bmatrix} -0.9412 \\ -0.8824 \\ -0.2941 \\ 1.0000 \\ 1.1176 \end{bmatrix}$$

vs. Min-Max Normalization:

$$[0, 1/35, 11/35, 33/35, 1] = [0, 0.0286, 0.3143, 0.9429, 1.0]$$

Discussion

- Can you use Z-score to automatically find phrases?
 - If we have 1,000 “matrix” and 1,000 “factorization” in 1,000,000 words, and we assume independency, we should have only one “matrix factorization” (expected).
 - But actually we have more! - Outlierness

Jingbo Shang, Jialu Liu, Meng Jiang, Xiang Ren, Clare R Voss, Jiawei Han.
“Automated Phrase Mining from Massive Text Corpora”. Submitted to
Transactions on Knowledge and Data Engineering.

References

- W. Cleveland, Visualizing Data, Hobart Press, 1993
- T. Dasu and T. Johnson. Exploratory Data Mining and Data Cleaning. John Wiley, 2003
- U. Fayyad, G. Grinstein, and A. Wierse. Information Visualization in Data Mining and Knowledge Discovery, Morgan Kaufmann, 2001
- L. Kaufman and P. J. Rousseeuw. Finding Groups in Data: an Introduction to Cluster Analysis. John Wiley & Sons, 1990.
- H. V. Jagadish et al., Special Issue on Data Reduction Techniques. Bulletin of the Tech. Committee on Data Eng., 20(4), Dec. 1997
- D. A. Keim. Information visualization and visual data mining, IEEE trans. on Visualization and Computer Graphics, 8(1), 2002
- D. Pyle. Data Preparation for Data Mining. Morgan Kaufmann, 1999
- S. Santini and R. Jain, "Similarity measures", IEEE Trans. on Pattern Analysis and Machine Intelligence, 21(9), 1999
- E. R. Tufte. The Visual Display of Quantitative Information, 2nd ed., Graphics Press, 2001
- C. Yu, et al., Visual data mining of multimedia data for social and behavioral studies, Information Visualization, 8(1), 2009