CS699
Lecture 2
Data Exploration

Types of Data Sets

- Record
  - Relational records
  - Data matrix, e.g., numerical matrix, crosstabs
  - Document data: text documents: term-frequency vector
  - Transaction data
- Graph and network
  - World Wide Web
  - Social or information networks
  - Molecular Structures
- Ordered
  - Video data: sequence of images
  - Temporal data: time-series
  - Sequential Data: transaction sequences
  - Genetic sequence data
- Spatial, image and multimedia:
  - Spatial data: maps
  - Image data:
  - Video data:

	team	coach	pla y	ball	score	game	wi n	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

TID	*Items*
1	**Bread, Coke, Milk**
2	**Beer, Bread**
3	**Beer, Coke, Diaper, Milk**
4	**Beer, Bread, Diaper, Milk**
5	**Coke, Diaper, Milk**

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 **fields, dimensions, features, variables**): a data field, representing a characteristic or feature of a data object.
  - *E.g., customer _ID, name, address*
- Types:
  - Nominal (or categorical), Binary, Ordinal
  - Numeric: quantitative
    - Interval-scaled
    - Ratio-scaled

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)
  - <u>Symmetric binary</u>: both outcomes equally important
    - e.g., gender
  - <u>Asymmetric binary</u>: 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

Numeric Attribute Types

- Quantity (integer or real-valued)
- **Interval**
    - Measured on a scale of **equal-sized units**
    - Values have order
        - E.g., *temperature in C˚or F˚, calendar dates*
    - No true zero-point
- **Ratio**
    - Inherent **zero-point**
    - We can speak of values as being an order of magnitude larger than the unit of measurement (10 K˚ is twice as high as 5 K˚).
        - e.g., *temperature in Kelvin, length, counts, monetary quantities*

Discrete vs. Continuous Attributes

- **Discrete Attribute**
  - Has only a finite or countably infinite set of values
    - E.g., zip codes, profession, or the set of words in a collection of documents
  - Sometimes, represented as integer variables
  - Note: Binary attributes are a special case of discrete attributes
- **Continuous Attribute**
  - Has real numbers as attribute values
    - E.g., temperature, height, or weight
  - Practically, real values can only be measured and represented using a finite number of digits
  - Continuous attributes are typically represented as floating-point variables

Basic Statistical Descriptions of Data

- Motivation
  - To better understand the data: central tendency, variation and spread
- Central Tendency
  - Location of center of a data distribution
  - mean, median, mode, etc.
- Data dispersion
  - How the data is spread out
  - quartiles, interquartile range, boxplot, standard deviation, variance, etc.

Measuring the Central Tendency

- Mean (algebraic measure) (sample vs. population):

  Note: *n* is sample size and *N* is population size.

  – $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ (sample), $\mu = \frac{\sum x}{N}$ (population)

  – 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

- Median:
  - Middle value if odd number of values, or average of the middle two values otherwise
  - median of <2, 5, 6, 8, 11, 20, 40> is 8
  - median of <2, 5, 6, 8, 20, 40> is 7 (= (6 + 8) / 2)

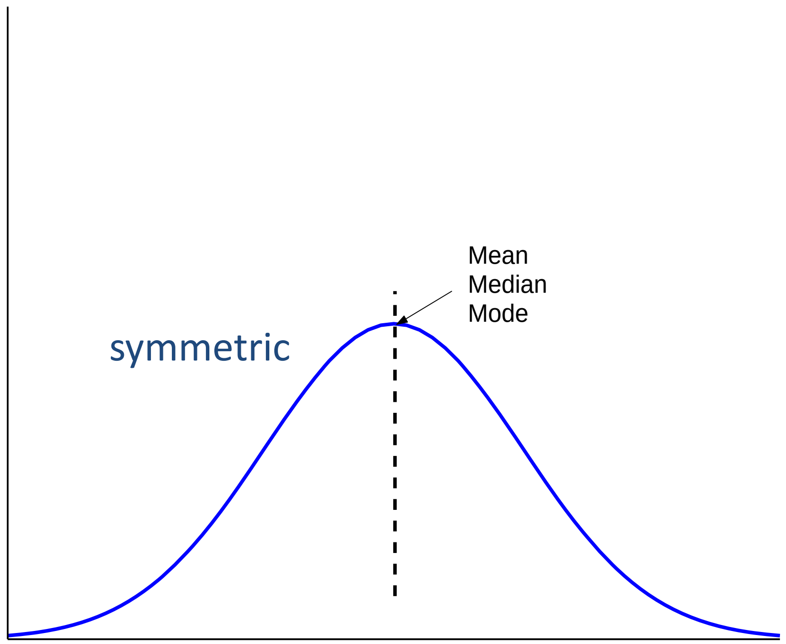
Measuring the Central Tendency

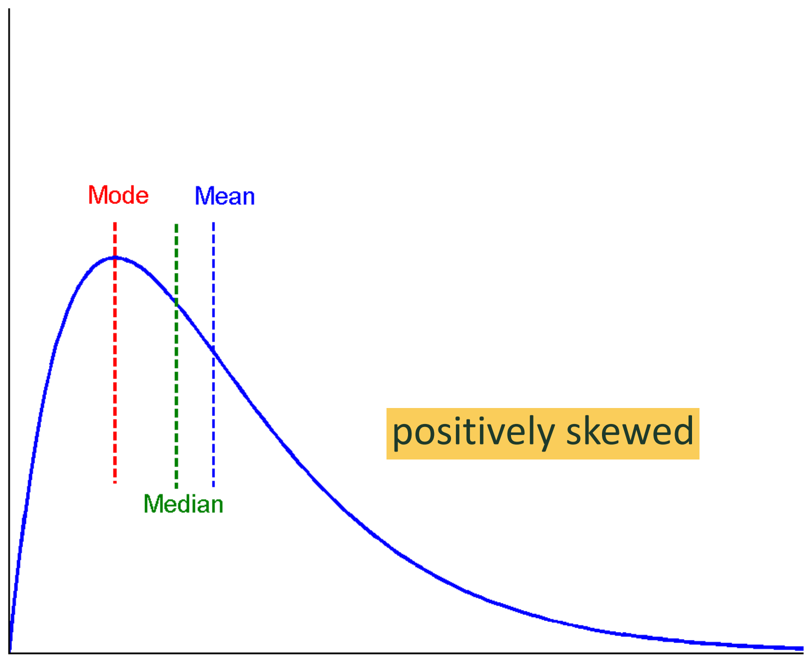
- <u>Mode</u>
  - Value that occurs most frequently in the data
  - Unimodal, bimodal, trimodal
  - mode of <1, 1, 3, 3, 3, 5, 8, 9, 10, 10> is 3 (unimodal)
  - modes of < 1, 1, 3, 3, 3, 5, 8, 9, 10, 10, 10> are 3 and 10 (bimodal)
  - Empirical formula to estimate mode for unimodal, moderately skewed data (given mean and median):

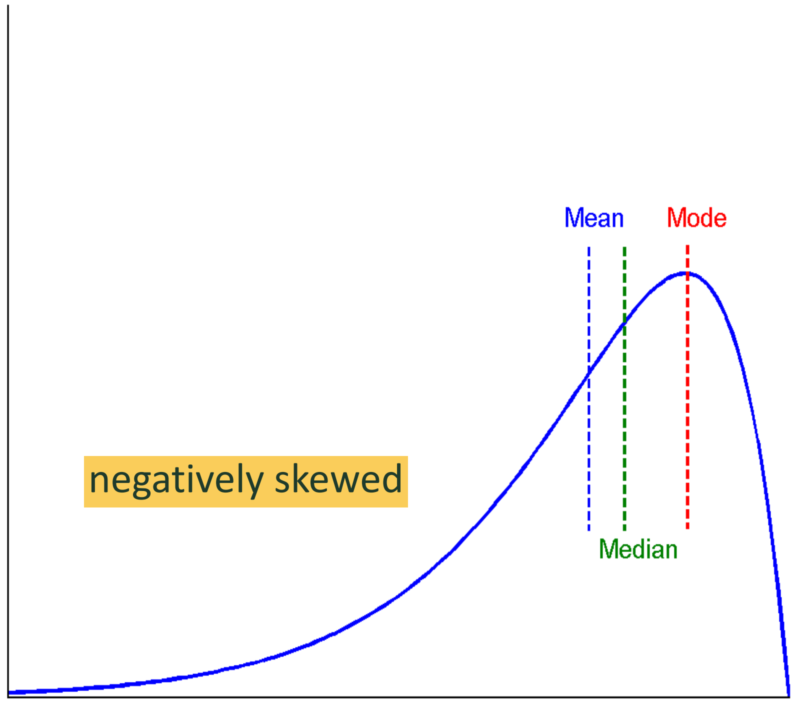
$$mean - mode = 3 \times (mean - median)$$

Symmetric vs. Skewed Data

- Median, mean and mode of symmetric, positively and negatively skewed data







Measuring the Dispersion of Data

- Quartiles, outliers and boxplots
  - **Quartiles**: Q_1 (25th percentile), Q_3 (75th percentile)
  - **Inter-quartile range**: $IQR = Q_3 - Q_1$
  - **Five number summary**: min, Q_1, median, Q_3, max
  - **Boxplot**: ends of the box are the quartiles; median is marked; add whiskers, and plot outliers individually
  - **Outlier**:
    - Less than $Q_1 - 1.5 * IQR$
    - Greater than $Q_3 + 1.5 * IQR$
  - **Note**: There are different ways of determining quartiles.

Measuring the Dispersion of Data

- Example

  D = <2, 10, 12, 15, 17, 20, 53>

  Median = 15

  Q1 = median of lower half <2, 10, 12> = 10 (some include the median, 15, in the lower half)

  Q3 = median of upper half <17, 20, 53> = 20 (some include the median, 15, in the upper half)

  IQR = 20 – 10 = 10

  Q1 – 1.5*IQR = 10 – 15 = -5

  Q3 + 1.5*IQR = 20 + 15 = 35

  So, 53 is an outlier

Measuring the Dispersion of Data

- Variance and standard deviation (*sample: s, population: σ)*
  - **Variance**:

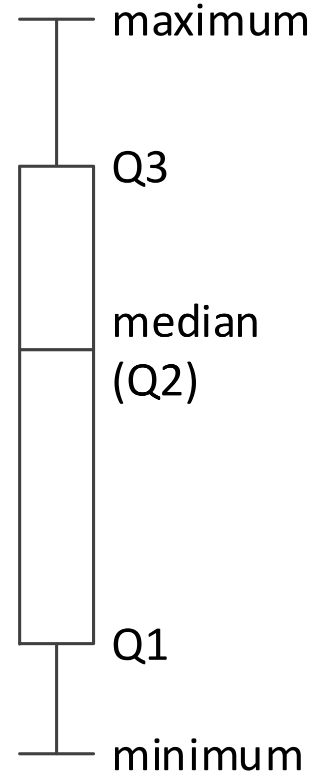
$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}[\sum_{i=1}^{n}{x_i}^2 - \frac{1}{n}(\sum_{i=1}^{n}x_i)^2]$$

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

  - **Denominator in the formula**
    - $n - 1$ is used for sample
    - N is used for population
  - **Standard deviation** s (or σ) is the square root of variance s^2 (or σ^2)

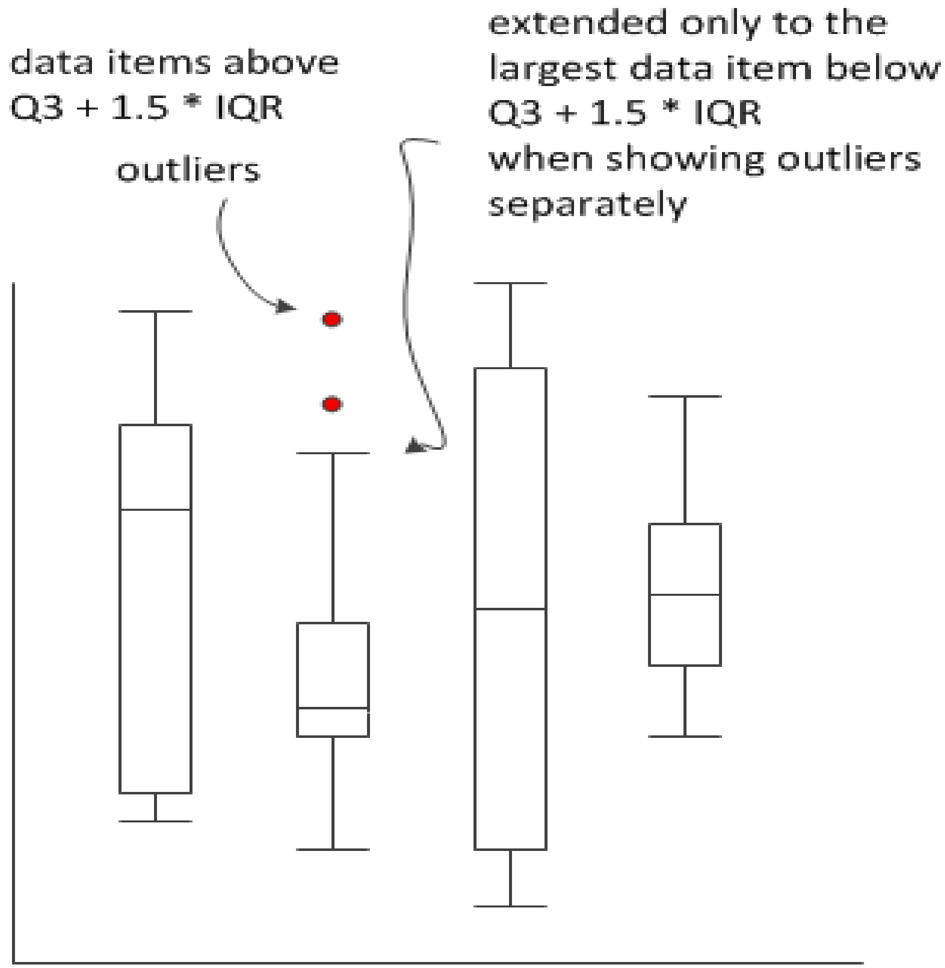
Boxplot Analysis

- **Five-number summary** of a distribution
  - Minimum, Q1, Median, Q3, Maximum
- **Boxplot**
  - Data is represented with a box and lines
  - Can be drawn vertically or horizontally
  - The ends of the box are at the first and third quartiles. So, the height of the box is IQR
  - The median is marked by a line within the box
  - Whiskers: two lines outside the box extended to Minimum and Maximum
  - Outliers: points beyond a specified outlier threshold, plotted individually



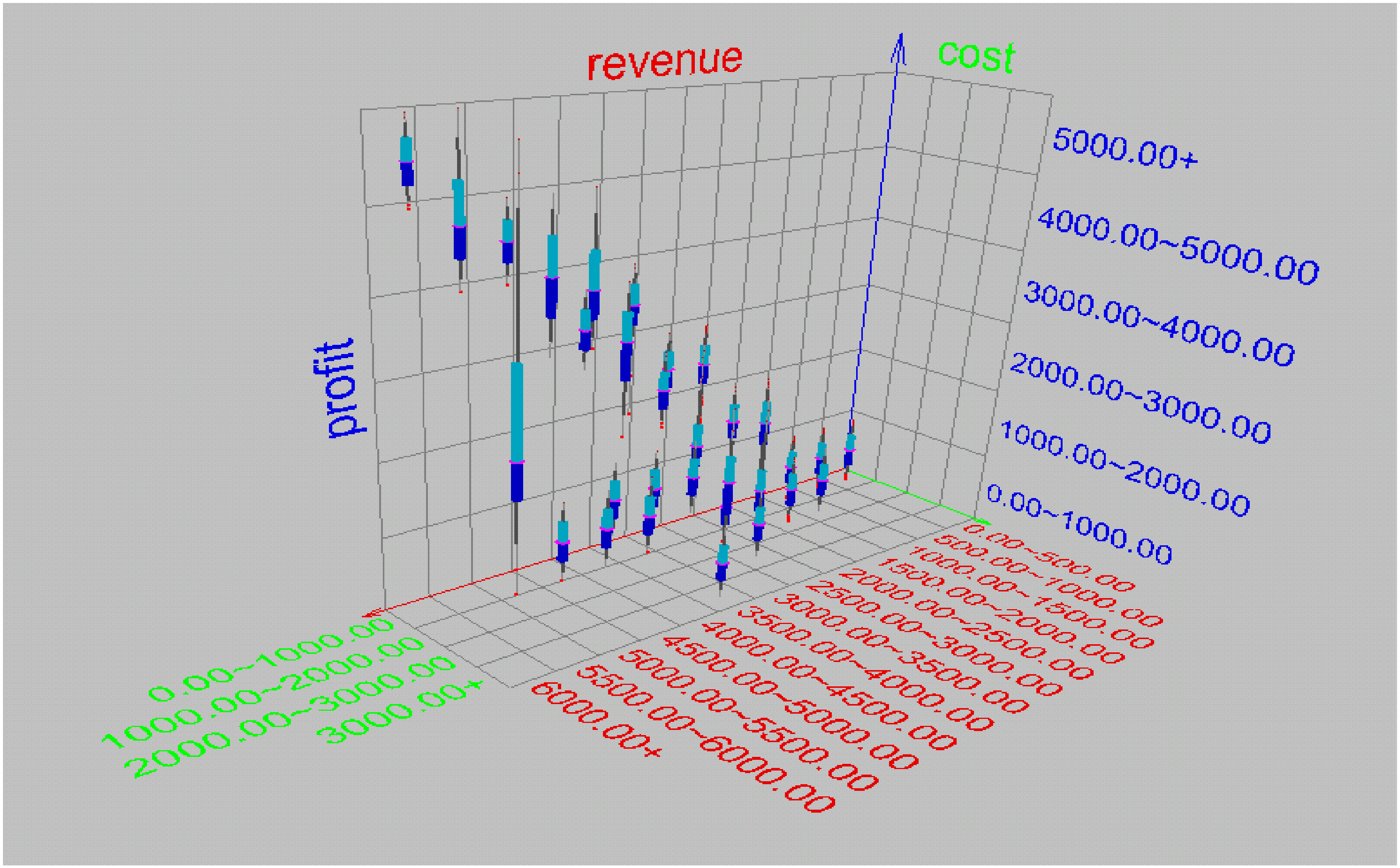
Boxplot Analysis

- Example



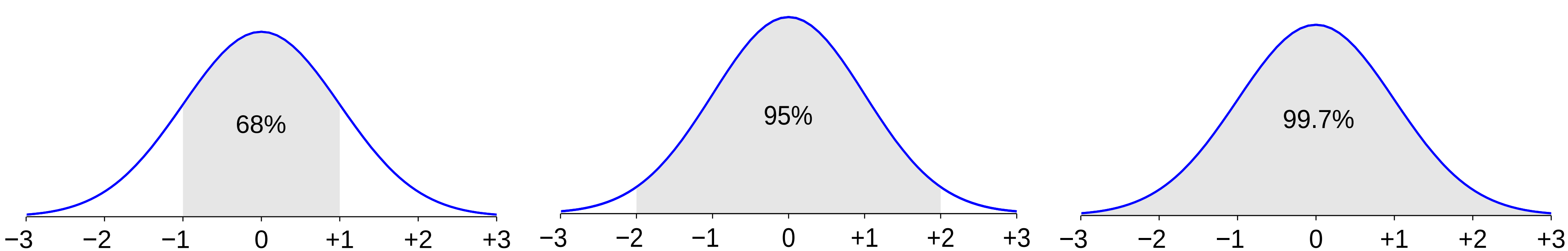
Boxplot: min, Q1, median, Q3, max

Visualization of Data Dispersion: 3-D Boxplots



Properties of Normal Distribution Curve

- The normal (distribution) curve
  - From μ–σ to μ+σ: contains about 68% of the measurements (μ: mean, σ: standard deviation)
  - From μ–2σ to μ+2σ: contains about 95% of it
  - From μ–3σ to μ+3σ: contains about 99.7% of it

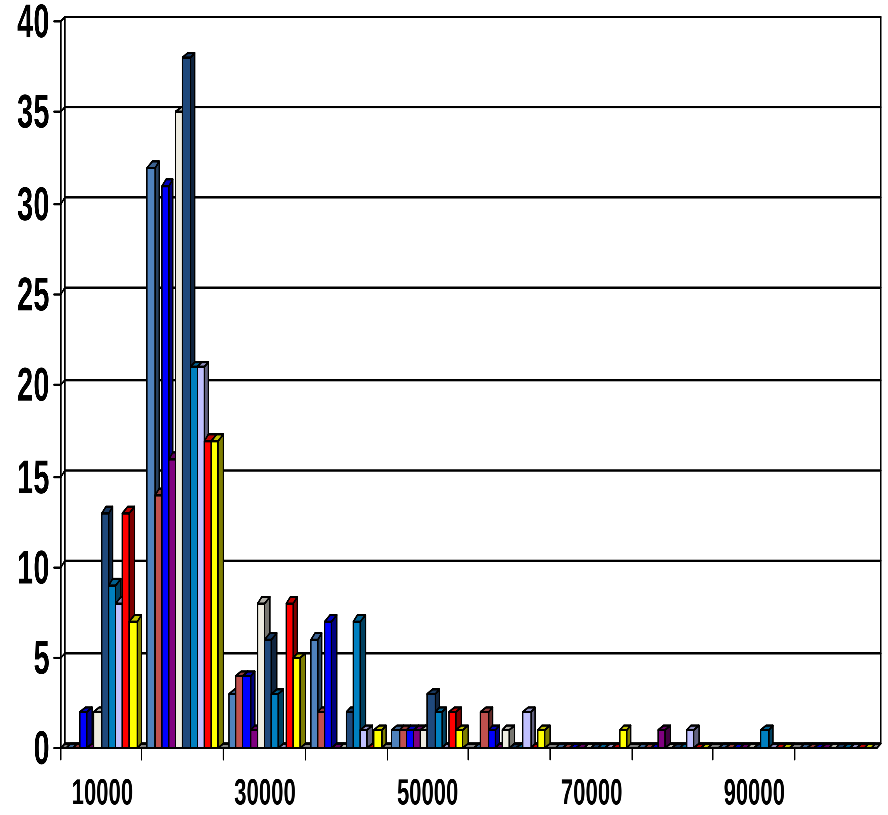


Graphic Displays of Basic Statistical Descriptions

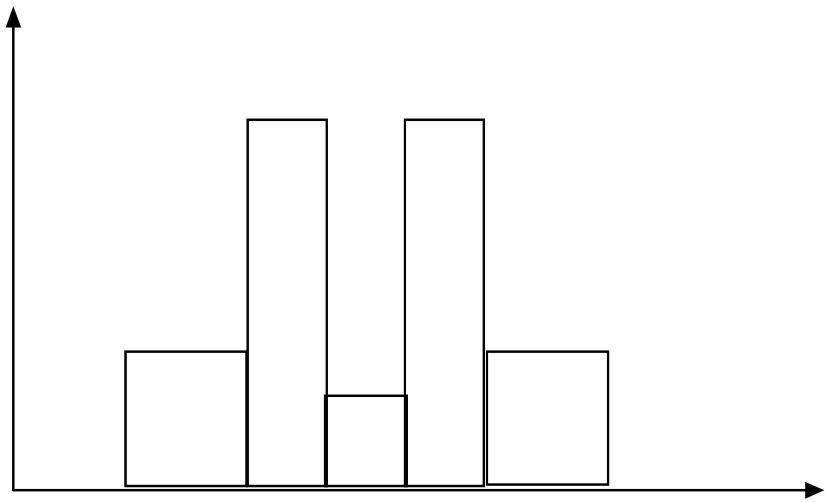
- **Boxplot**: graphic display of five-number summary
- **Histogram**: x-axis are values, y-axis represents frequencies
- **Scatter plot**: each pair of values is a pair of coordinates and plotted as points in the plane

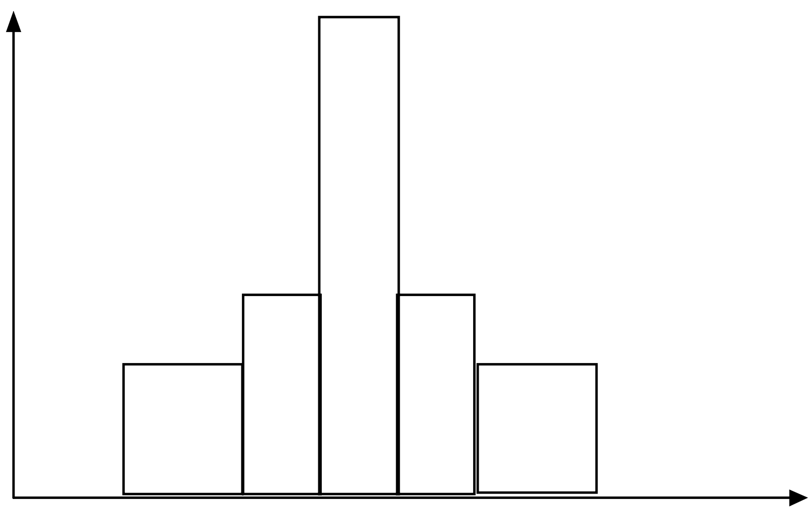
Histogram Analysis

- Histogram: Graph display of tabulated frequencies, shown as bars
- It shows what proportion of cases fall into each of several categories
- The categories are usually specified as non-overlapping intervals of some variable. The categories (bars) must be adjacent



Histograms Often Tell More than Boxplots

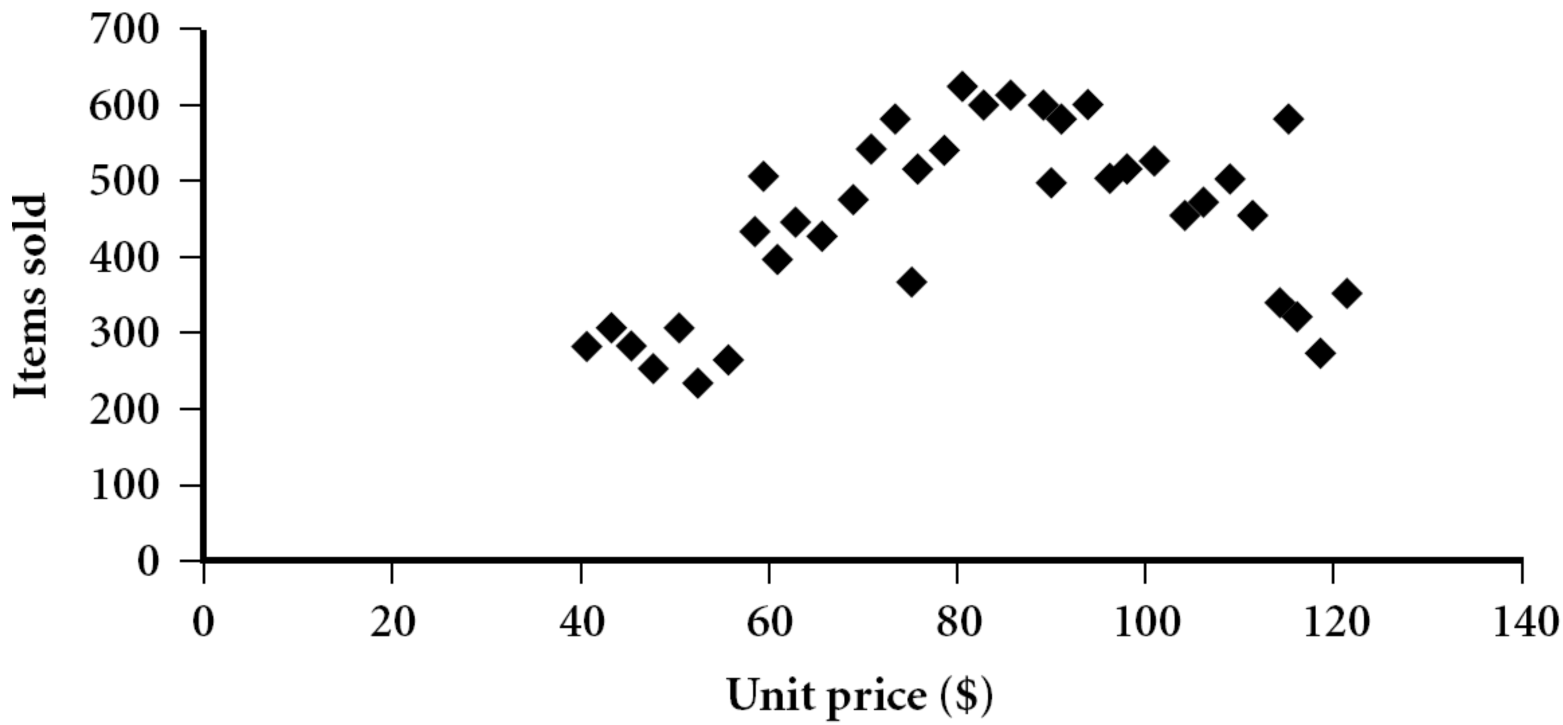




- The two histograms shown in the left may have the same boxplot representation
  - The same values for: min, Q1, median, Q3, max
- But they have rather different data distributions

Scatter plot

- Provides a first look at bivariate data to see clusters of points, outliers, etc
- Each pair of values is treated as a pair of coordinates and plotted as points in the plane



Similarity and Dissimilarity

- **Similarity**
  - Numerical measure of how alike two data objects are
  - Value is higher when objects are more alike
  - Often falls in the range [0,1]
- **Dissimilarity** (e.g., distance)
  - Numerical measure of how different two data objects are
  - Lower when objects are more alike
  - Minimum dissimilarity is often 0
  - Upper limit varies
- **Proximity** refers to a similarity or dissimilarity

Data Matrix and Dissimilarity Matrix

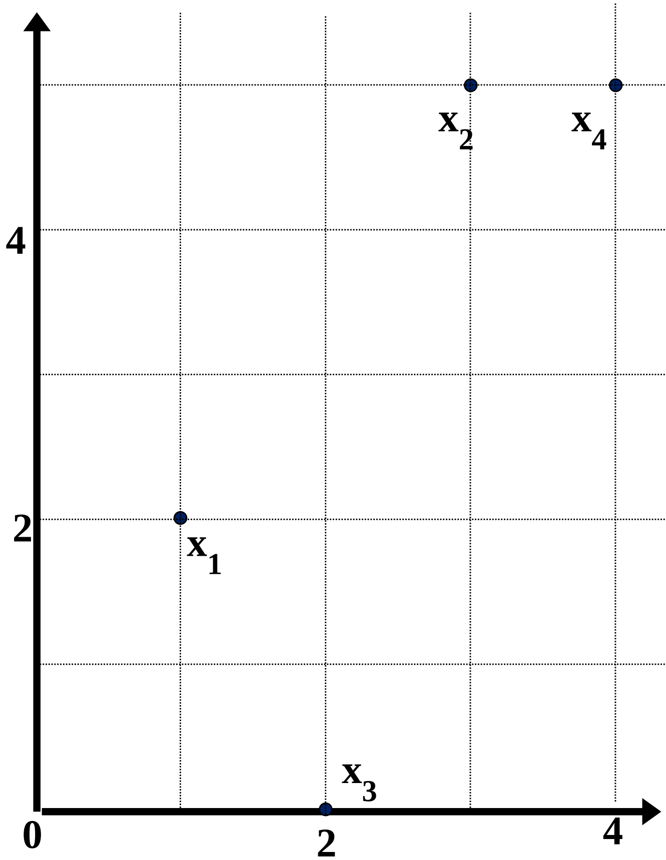
- Data matrix
  - n data points with p dimensions
  - Two modes

$$\begin{bmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{bmatrix}$$

- Dissimilarity matrix
  - n data points, but registers only the distance
  - A triangular matrix
  - Single mode

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & & \\ d(n,1) & d(n,2) & \cdots & \cdots & 0 \end{bmatrix}$$

Example:
Data Matrix and Dissimilarity Matrix



Data Matrix

point	attribute1	attribute2
x1	1	2
x2	3	5
x3	2	0
x4	4	5

Dissimilarity Matrix
(with Euclidean Distance)

	x1	***x2***	***x3***	***x4***
x1	0			
x2	3.61	0		
x3	2.24	5.1	0	
x4	4.24	1	5.39	0

Proximity Measure for Nominal Attributes

- Can take 2 or more states, e.g., red, yellow, blue, green (generalization of a binary attribute)
- Distance : $$d(i, j) = \frac{p - m}{p}$$

  where *m*: # of matches, *p*: total # of variables

  (or distance = #mismatches / all)

- Example

Object	Income	Housing	Zip	Marital_status
Molly	high	own	02215	yes
Greg	medium	own	02215	yes

distance(Molly, Greg) = 1/4 or 0.25

Dissimilarity between Binary Variables

- For symmetric binary variables, use the same method that is used for nominal attributes: distance = #mismatches / all
- Example

Name	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	Y	N	P	N	N	N
Mary	Y	N	P	N	P	N
Jim	Y	P	N	N	N	N

$$d(jack, mary) = \frac{1}{6} = 0.17$$

$$d(jack, jim) = \frac{2}{6} = 0.33$$

$$d(jim, mary) = \frac{3}{6} = 0.5$$

Standardizing Numeric Data

- Z-score: $$z = \frac{x-\mu}{\sigma}$$
  - *X*: raw score to be standardized, μ: mean, σ (or s): standard deviation
  - the distance between the raw score and the population mean in units of the standard deviation
  - negative when the raw score is below the mean, "+" when above
- An alternative way: Use **mean absolute deviation**, S_f, instead of σ

$$s_f = \frac{1}{n}(|x_{1f} - \mu| + |x_{2f} - \mu| + \ldots + |x_{nf} - \mu|)$$

  - standardized measure (*z-score*):

$$z_{if} = \frac{x-\mu}{s_f}$$

- Using mean absolute deviation is more robust than using standard deviation

Standardizing Numeric Data

- Example

  Data = (5, 8, 3, 12, 7)

  $\mu = 7$, $s = 3.391$

  $S_f = (|5 - 7| + |8 - 7| + |3 - 7| + |12 - 7| + |7 - 7|) / 5 = 2.4$

  Standardizing 5 and 8 using standard deviation:

  $(5 - 7) / 3.391 = -0.590$, $(8 - 7) / 3.391 = 0.295$

  Standardizing 5 and 8 using mean absolute deviation:

  $(5 - 7) / 2.4 = -0.833$, $(8 - 7) / 2.4 = 0.417$

Distance on Numeric Data: Minkowski Distance

- *Minkowski distance*: A popular distance measure

$$d(i, j) = \sqrt[h]{|x_{i1} - x_{j1}|^h + |x_{i2} - x_{j2}|^h + \cdots + |x_{ip} - x_{jp}|^h}$$

where $i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ and $j = (x_{j1}, x_{j2}, \ldots, x_{jp})$ are two *p*-dimensional data objects, and *h* is the order (the distance so defined is also called L-*h* norm)

- Properties
  - $d(i, j) > 0$ if $i \neq j$, and $d(i, i) = 0$ (Positive definiteness)
  - $d(i, j) = d(j, i)$ (Symmetry)
  - $d(i, j) \leq d(i, k) + d(k, j)$ (Triangle Inequality)
- A distance that satisfies these properties is called metric

Special Cases of Minkowski Distance

- h = 1: Manhattan distance (city block, L_1 norm)
  - E.g., the Hamming distance: the number of bits that are different between two binary vectors

$$d(i,j) = |x_{i_1} - x_{j_1}| + |x_{i_2} - x_{j_2}| + \dots + |x_{i_p} - x_{j_p}|$$

- h = 2: Euclideean distance (L_2 norm)

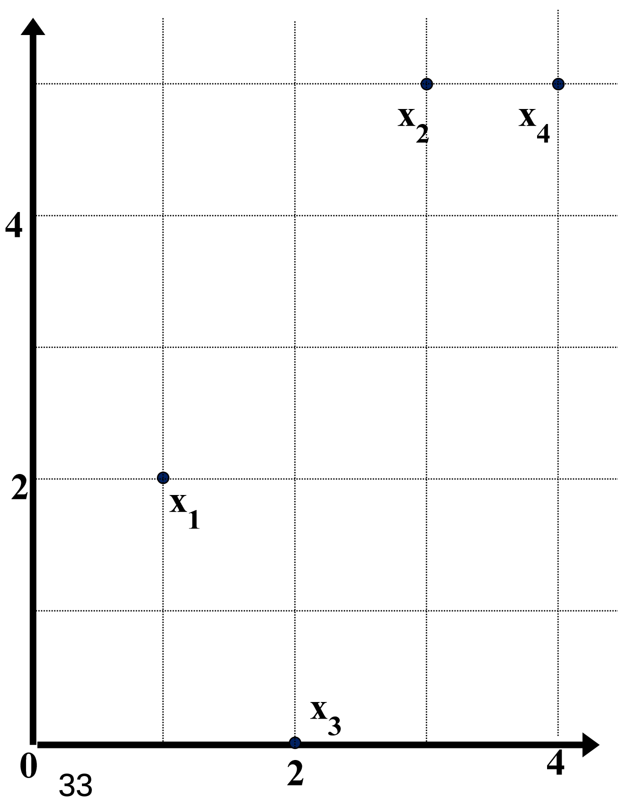
$$d(i,j) = \sqrt{(|x_{i_1} - x_{j_1}|^2 + |x_{i_2} - x_{j_2}|^2 + \dots + |x_{i_p} - x_{j_p}|^2)}$$

- $h \to \infty$. "supremum" distance (L_{max} norm, L_∞ norm)
  - This is the maximum difference between any component (attribute) of the vectors

$$d(i,j) = \lim_{h \to \infty} \left(\sum_{f=1}^{p} |x_{if} - x_{jf}|^h \right)^{\frac{1}{h}} = \max_{f}^{p} |x_{if} - x_{jf}|$$

Example: Minkowski Distance

point	attribute 1	attribute 2
x1	1	2
x2	3	5
x3	2	0
x4	4	5



Manhattan (L_1)

L	x1	x2	x3	x4
x1	0			
x2	5	0		
x3	3	6	0	
x4	6	1	7	0

Euclidean (L_2)

L2	x1	x2	x3	x4
x1	0			
x2	3.61	0		
x3	2.24	5.1	0	
x4	4.24	1	5.39	0

Supremum

L_∞	x1	x2	x3	x4
x1	0			
x2	3	0		
x3	2	5	0	
x4	3	1	5	0

Cosine Similarity

- A **document** can be represented by thousands of attributes, each recording the *frequency* of a particular word (such as keywords) or phrase in the document.

Document	*team*	*coach*	*hockey*	*baseball*	*soccer*	*penalty*	*score*	*win*	*loss*	*season*
Document1	5	0	3	0	2	0	0	2	0	0
Document2	3	0	2	0	1	1	0	1	0	1
Document3	0	7	0	2	1	0	0	3	0	0
Document4	0	1	0	0	1	2	2	0	3	0

- Other vector objects: gene features in micro-arrays, …
- Applications: information retrieval, biologic taxonomy, gene feature mapping, ...
- Cosine measure: If d_1 and d_2 are two vectors (e.g., term-frequency vectors), then

  $$\cos(d_1, d_2) = (d_1 \bullet d_2) / ||d_1|| \, ||d_2|| ,$$

  where $\bullet$ indicates vector dot product, $||d||$: the length of vector d
- Between 0 and 1, inclusive; Closer to 0: less similar; Closer to 1: more similar

Example: Cosine Similarity

- $\cos(d_1, d_2) = (d_1 \bullet d_2) / ||d_1||\ ||d_2||$,
  where $\bullet$ indicates vector dot product, $||d||$ is the length of vector d

- Ex: Find the **similarity** between documents 1 and 2.

$d_1 =$ (5, 0, 3, 0, 2, 0, 0, 2, 0, 0)
$d_2 =$ (3, 0, 2, 0, 1, 1, 0, 1, 0, 1)

$d_1 \bullet d_2 = 5*3+0*0+3*2+0*0+2*1+0*1+0*0+2*1+0*0+0*1 = 25$
$||d_1|| = (5*5+0*0+3*3+0*0+2*2+0*0+0*0+2*2+0*0+0*0)^{0.5} = (42)^{0.5} = 6.481$
$||d_2|| = (3*3+0*0+2*2+0*0+1*1+1*1+0*0+1*1+0*0+1*1)^{0.5} = (17)^{0.5} = 4.12$

$\cos(d_1, d_2) = 25 / (6.481 * 4.12) = 0.94$

Attributes of Mixed Types

- Distance between object 1 and object 2.
- A1 and A2: interval-scaled; A3, A4, and A5: asymmetric binary (P is more important than N); A6 and A7: nominal; A8 is ordinal (ranks are gold = 3, silver = 2, bronze = 1); "?" indicates a missing value.
- A1: |8 – 21| / (21 – 6) = 0.867
- A2:|17 – 6| / (21 – 6) = 0.733
- A3: 1, A6: 0, A7: 1
- A8: |1 – 0.5| / (1 – 0) = 0.5
- *d*(O1,O2)

= (0.87 + 0.73 + 1 + 0 + 1 + 0.5) / 6

= 0.68

OID	A1	A2	A3	A4	A5	A6	A7	A8
1	8	17	N	N	N	two	4wd	gold
2	21	6	P	?	N	two	fwd	silver
3	10	10	P	P	N	two	fwd	bronze
4	16	12	P	N	Y	four	4wd	gold
5	12	14	P	N	Y	four	fwd	gold
6	13	11	N	P	N	two	fwd	silver
7	10	8	P	N	N	four	4wd	bronze
8	6	21	N	P	Y	four	fwd	gold

References

- Han, J., Kamber, M., Pei, J., "Data mining: concepts and techniques," 3rd Ed., Morgan Kaufmann, 2012
- http://www.cs.illinois.edu/~hanj/bk3/