PES UNIVERSITY

Getting to know your data

Dr.Mamatha.H.R

Professor

Department of Computer Science and Engineering

PES University

Bangalore

Course material created using various Internet resources and text book

- Knowledge about your data is useful for data preprocessing the first major task of the data mining process.

- You will want to know the following:

- What are the types of *attributes* or fields that make up your data?
- What kind of values does each attribute have?
- Which attributes are discrete, and which are continuous-valued?
- What do the data *look like*?

- How are the values distributed?
- Are there ways we can visualize the data to get a better sense of it all?
- Can we spot any outliers?
- Can we measure the similarity of some data objects with respect to others?
- Gaining such insight into the data will help with the subsequent analysis.

- *“So what can we learn about our data that’s helpful in data preprocessing?”*

- studying the various attribute types.
- These include nominal attributes, binary attributes, ordinal attributes, and numeric attributes.

- Basic *statistical descriptions* can be used to learn more about each attribute’s values
- Given a *temperature* attribute, for example, we can determine its **mean** (average value),**median** (middle value), and **mode** (most common value).
- These are **measures of central tendency**, which give us an idea of the “middle” or center of distribution.

- Knowing such basic statistics regarding each attribute makes it easier to fill in missing values, smooth noisy values, and spot outliers during data preprocessing.
- Knowledge of the attributes and attribute values can also help in fixing inconsistencies incurred during data integration.

- Plotting the measures of central tendency shows us if the data are symmetric or skewed.
- Quantile plots, histograms, and scatter plots are other graphic displays of basic statistical descriptions.
- These can all be useful during data preprocessing and can provide insight into areas for mining.

- The field of data visualization provides many additional techniques for viewing data through graphical means.
- These can help identify relations, trends, and biases “hidden” in unstructured data sets.

- we may want to examine how similar (or dissimilar) data objects are.
- For example, suppose we have a database where the data objects are patients, described by their symptoms. We may want to find the similarity or dissimilarity between individual patients. Such information can allow us to find clusters of like patients within the data set.
- The similarity/dissimilarity between objects may also be used to detect outliers in the data, or to perform nearest-neighbor classification.

Getting to Know Your Data

- Data Objects and Attribute Types
- Basic Statistical Descriptions of Data

Types of Data Sets

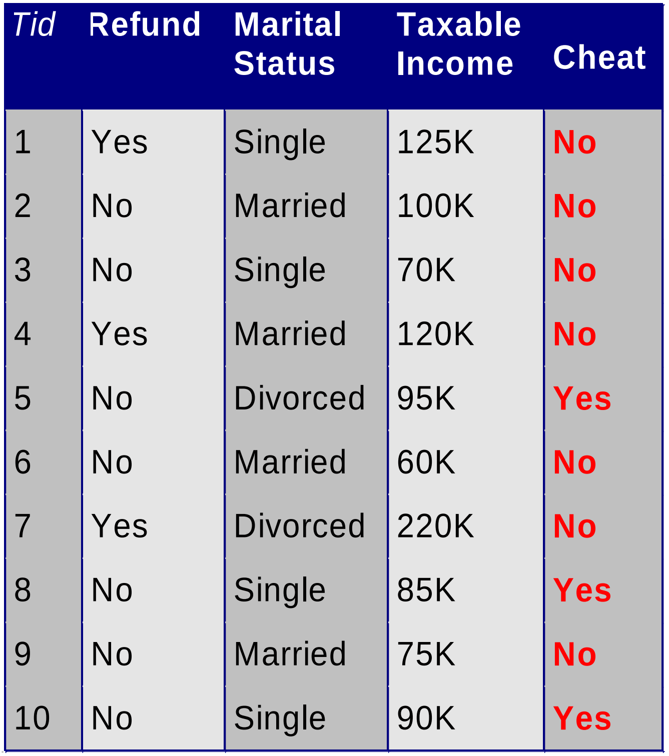
- Record
  - Relational records
  - Data matrix, e.g., numerical matrix, crosstabs
  - Document data: text documents: term-frequency vector
  - Transaction data
- Graph and network
  - World Wide Web
  - Social or information networks
  - Molecular Structures
- Ordered
  - Video data: sequence of images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

	team	coach	play	ball	score	game	win	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

Record Data

- Data that consists of a collection of records, each of which consists of a fixed set of attributes



Tid	**Refund**	**Marital Status**	**Taxable Income**	**Cheat**
1	Yes	Single	125K	**No**
2	No	Married	100K	**No**
3	No	Single	70K	**No**
4	Yes	Married	120K	**No**
5	No	Divorced	95K	**Yes**
6	No	Married	60K	**No**
7	Yes	Divorced	220K	**No**
8	No	Single	85K	**Yes**
9	No	Married	75K	**No**
10	No	Single	90K	**Yes**

Data Matrix

- If data objects have the same fixed set of numeric attributes, then the data objects can be thought of as points in a multi-dimensional space, where each dimension represents a distinct attribute

- Such data set can be represented by an m by n matrix, where there are m rows, one for each object, and n columns, one for each attribute

Projection of x Load	Projection of y load	Distance	Load	Thickness
10.23	5.27	15.22	2.7	1.2
12.65	6.25	16.22	2.2	1.1

Document Data

- Each document becomes a `term' vector,
  - each term is a component (attribute) of the vector,
  - the value of each component is the number of times the corresponding term occurs in the document.

	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

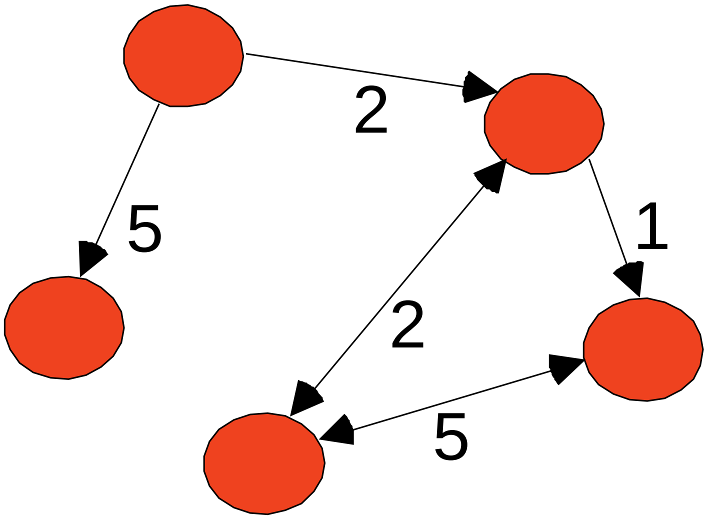
Transaction Data

- A special type of record data, where
  - each record (transaction) involves a set of items.
  - For example, consider a grocery store. The set of products purchased by a customer during one shopping trip constitute a transaction, while the individual products that were purchased are the items.

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

Graph Data

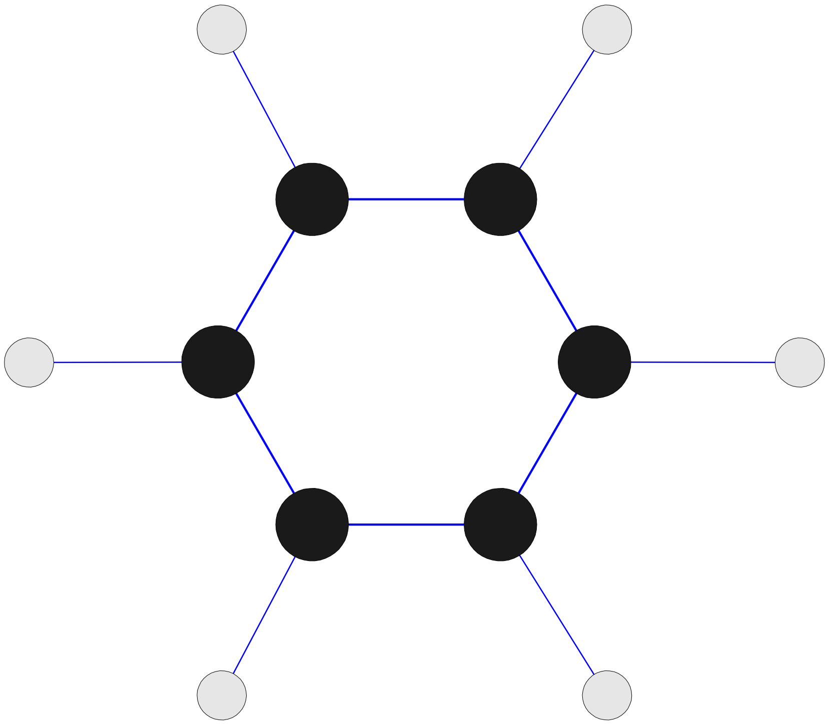
- Examples: Generic graph and HTML Links



<a href="papers/papers.html#bbbb">
Data Mining </a>
<li>
<a href="papers/papers.html#aaaa">
Graph Partitioning </a>
<li>
<a href="papers/papers.html#aaaa">
Parallel Solution of Sparse Linear System of Equations </a>
<li>
<a href="papers/papers.html#ffff">
N-Body Computation and Dense Linear System Solvers

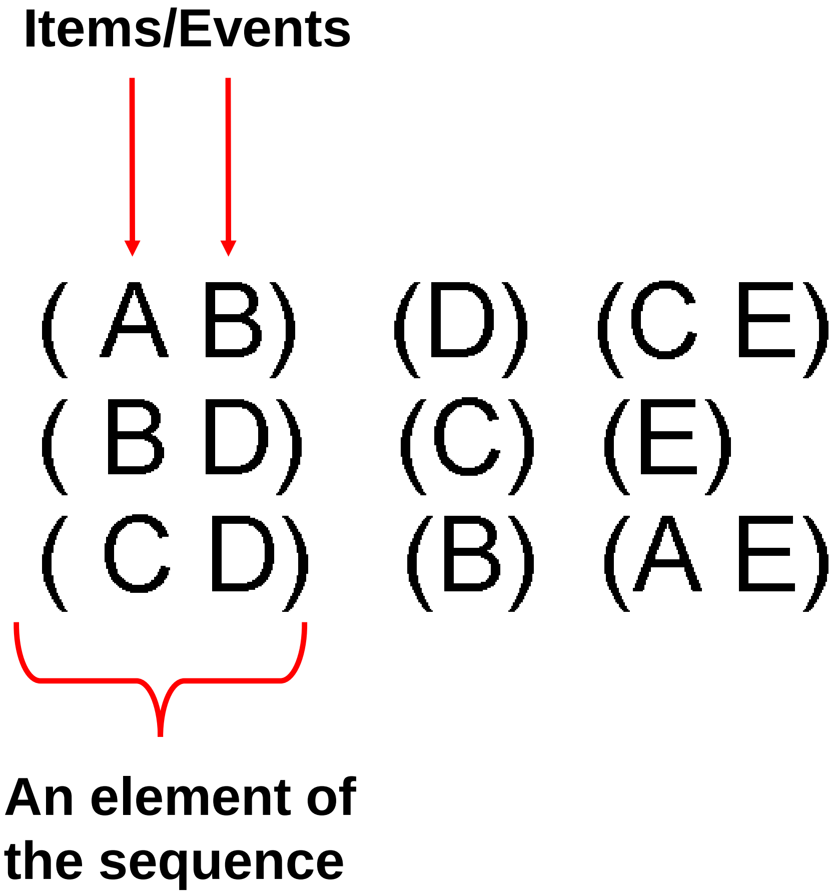
Chemical Data

- Benzene Molecule: C_6H_6



Ordered Data

- Sequences of transactions



Items/Events

(A B) (D) (C E)
(B D) (C) (E)
(C D) (B) (A E)

An element of the sequence

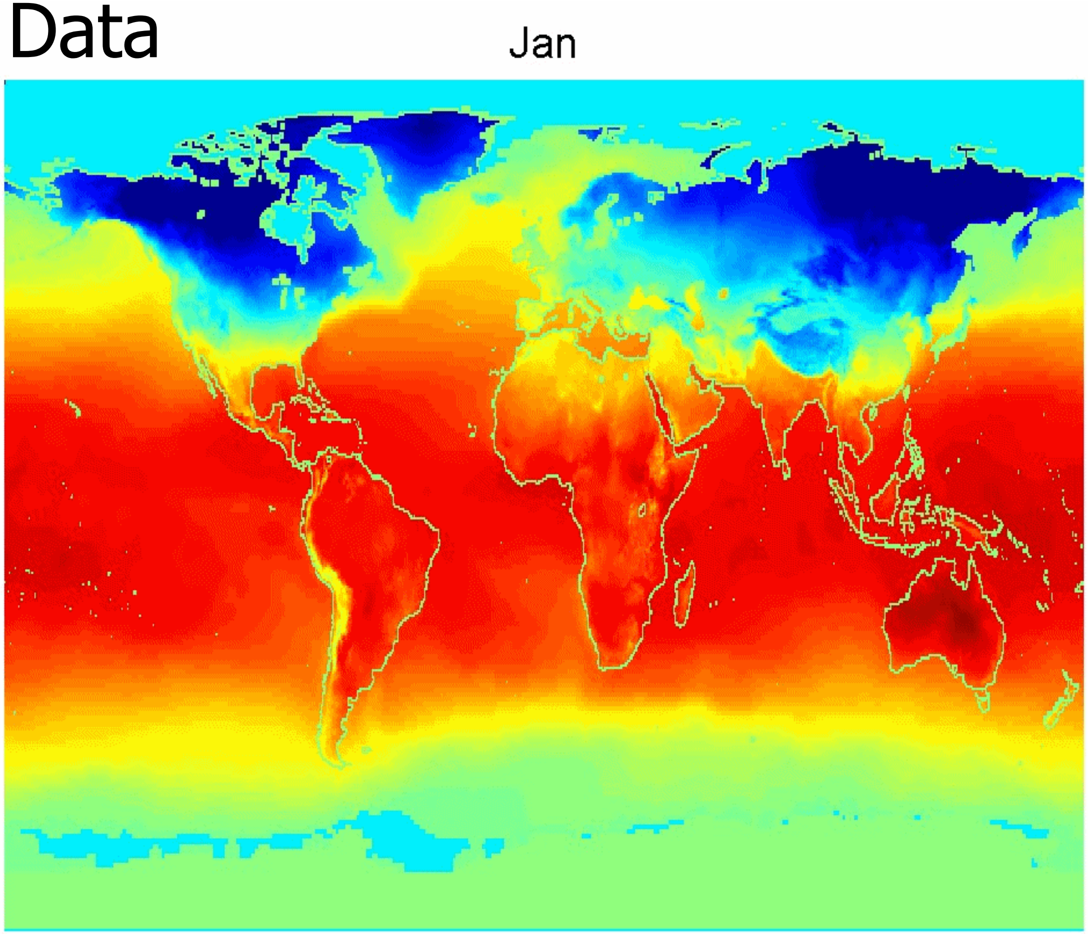
Ordered Data

- Genomic sequence data

```
GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG
```

Ordered Data

- Spatio-Temporal Data



Average Monthly Temperature of land and ocean

Important Characteristics of Data sets

- Dimensionality
  - problem: Curse of Dimensionality
  - Solution: dimensionality reduction

- Sparsity
  - Only presence counts

- Resolution
  - Patterns depend on the scale

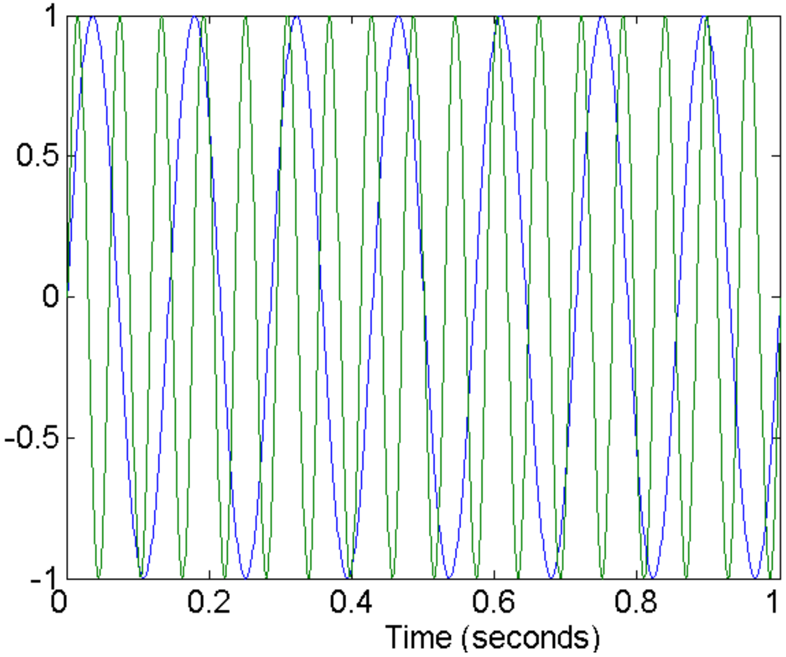
Data Quality

- What kinds of data quality problems?
- How can we detect problems with the data?
- What can we do about these problems?

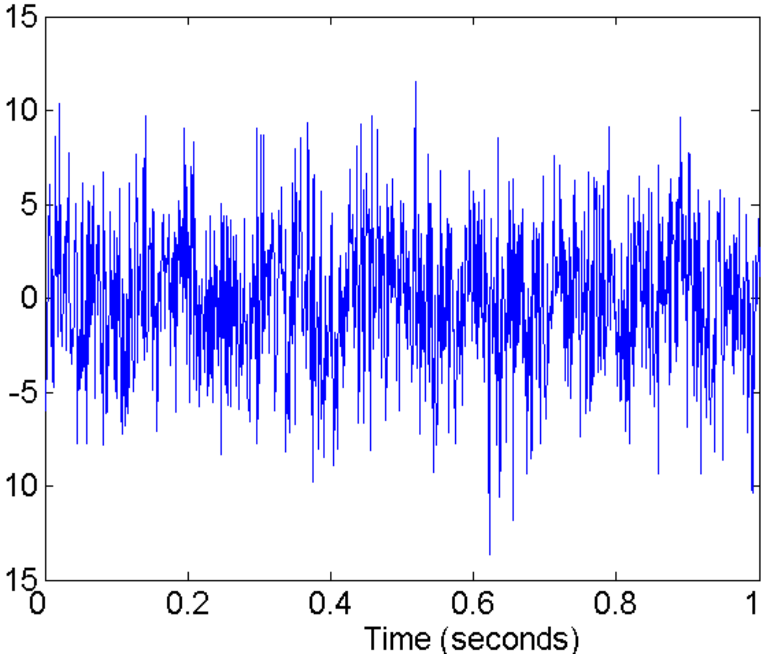
- Examples of data quality problems:
  - Noise and outliers
  - missing values
  - duplicate data

Noise

- Noise refers to modification of original values
  - Examples: distortion of a person's voice when talking on a poor phone and "snow" on television screen



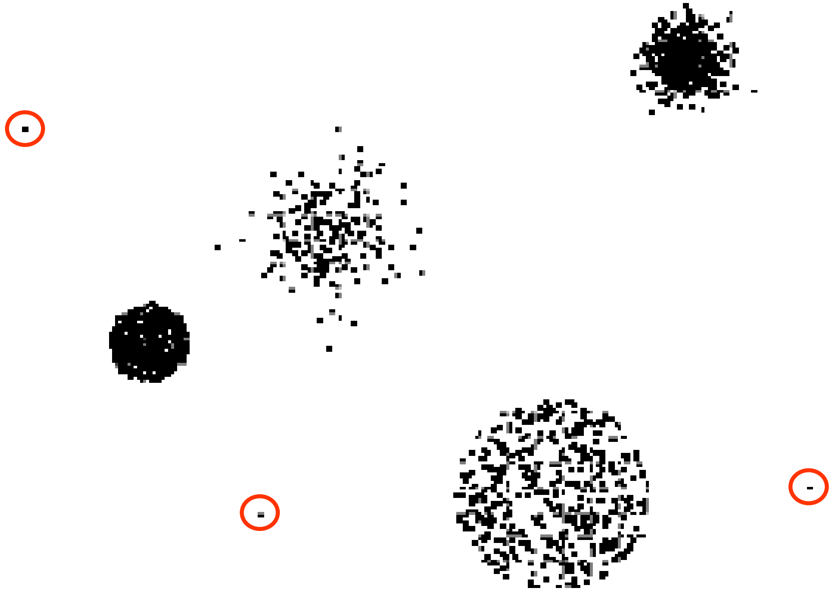
Two Sine Waves



Two Sine Waves + Noise

Outliers

- Outliers are data objects with characteristics that are considerably different than most of the other data objects in the data set



Missing Values

- Reasons for missing values
  - Information is not collected (e.g., people decline to give their age and weight)
  - Attributes may not be applicable to all cases (e.g., annual income is not applicable to children)
- Handling missing values
  - Eliminate Data Objects
  - Estimate Missing Values
  - Ignore the Missing Value During Analysis
  - Replace with all possible values (weighted by their probabilities)

Duplicate Data

- Data set may include data objects that are duplicates, or almost duplicates of one another
  - Major issue when merging data from heterogeous sources

- Examples:
  - Same person with multiple email addresses

- Data cleaning
  - Process of dealing with duplicate data issues

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.
  - *E.g., customer _ID, name, address*
- Types:
  - Nominal
  - Binary
  - Numeric: quantitative
    - Interval-scaled
    - Ratio-scaled

Attributes

Objects

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	**No**
2	No	Married	100K	**No**
3	No	Single	70K	**No**
4	Yes	Married	120K	**No**
5	No	Divorced	95K	**Yes**
6	No	Married	60K	**No**
7	Yes	Divorced	220K	**No**
8	No	Single	85K	**Yes**
9	No	Married	75K	**No**
10	No	Single	90K	**Yes**

Attribute Values

- Attribute values are numbers or symbols assigned to an attribute
- Distinction between attributes and attribute values
  - Same attribute can be mapped to different attribute values
    - Example: height can be measured in feet or meters
  - Different attributes can be mapped to the same set of values
    - Example: Attribute values for ID and age are integers
    - But properties of attribute values can be different
      - ID has no limit but age has a maximum and minimum value

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

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*

Properties of Attribute Values

- The type of an attribute depends on which of the following properties it possesses:

Distinctness:	= ≠
Order:	< >
Addition:	+ -
Multiplication:	* /

  - Nominal attribute: distinctness
  - Ordinal attribute: distinctness & order
  - Interval attribute: distinctness, order & addition
  - Ratio attribute: all 4 properties

Attribute Type	Description	Examples	Operations
Nominal	The values of a nominal attribute are just different names, i.e., nominal attributes provide only enough information to distinguish one object from another. (=, ≠)	zip codes, employee ID numbers, eye color, sex: {*male*, *female*}	mode, entropy, contingency correlation, χ^2 test
Ordinal	The values of an ordinal attribute provide enough information to order objects. (<, >)	hardness of minerals, {*good*, *better*, *best*}, grades, street numbers	median, percentiles, rank correlation, run tests, sign tests
Interval	For interval attributes, the differences between values are meaningful, i.e., a unit of measurement exists. (+, -)	calendar dates, temperature in Celsius or Fahrenheit	mean, standard deviation, Pearson's correlation, *t* and *F* tests
Ratio	For ratio variables, both differences and ratios are meaningful. (*, /)	temperature in Kelvin, monetary quantities, counts, age, mass, length, electrical current	geometric mean, harmonic mean, percent variation

Attribute Level	Transformation	Comments
Nominal	Any permutation of values	If all employee ID numbers were reassigned, would it make any difference?
Ordinal	An order preserving change of values, i.e., *new_value = f(old_value)* where *f* is a monotonic function.	An attribute encompassing the notion of good, better best can be represented equally well by the values {1, 2, 3} or by { 0.5, 1, 10}.
Interval	*new_value =a * old_value + b* where a and b are constants	Thus, the Fahrenheit and Celsius temperature scales differ in terms of where their zero value is and the size of a unit (degree).
Ratio	*new_value = a * old_value*	Length can be measured in meters or feet.

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

Getting to Know Your Data

- Data Objects and Attribute Types
- Basic Statistical Descriptions of Data
- Summary

Basic Statistical Descriptions of Data

- Motivation
  - To better understand the data: central tendency, variation and spread
- Data dispersion characteristics
  - median, max, min, quantiles, outliers, variance, etc.

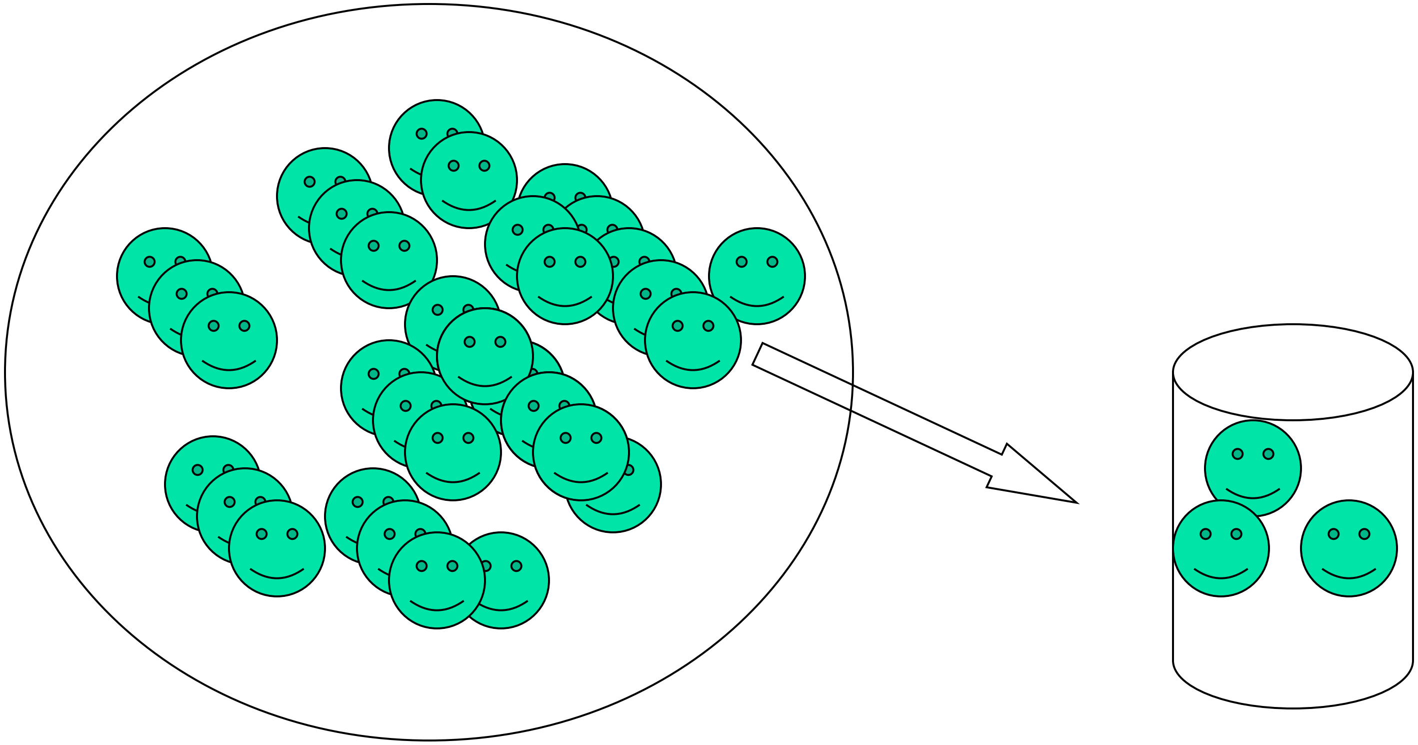
Descriptive Statistics

- Descriptive Statistics are Used by Researchers to Report on Populations and Samples

- Summary descriptions of measurements (variables) taken about a group of people

- By Summarizing Information, Descriptive Statistics Speed Up and Simplify Comprehension of a Group’s Characteristics

Sample vs. Population



Population

Sample

Descriptive Statistics

An Illustration:

Which Group is Smarter?

Class A--IQs of 13 Students		Class B--IQs of 13 Students	
102	115	127	162
128	109	131	103
131	89	96	111
98	106	80	109
140	119	93	87
93	97	120	105
110		109	

Each individual may be different. If you try to understand a group by remembering the qualities of each member, you become overwhelmed and fail to understand the group.

Descriptive Statistics

Which group is smarter now?

Class A--Average IQ	Class B--Average IQ
110.54	110.23

They’re roughly the same!

With a summary descriptive statistic, it is much easier to answer our question.

Descriptive Statistics

Types of descriptive statistics:

- Organize Data
  - Tables
  - Graphs

- Summarize Data
  - Central Tendency
  - Variation

Descriptive Statistics

Types of descriptive statistics:

- Organize Data
    - Tables
        - Frequency Distributions
        - Relative Frequency Distributions
    - Graphs
        - Bar Chart or Histogram
        - Stem and Leaf Plot
        - Frequency Polygon

Descriptive Statistics

Summarizing Data:

- Central Tendency (or Groups' "Middle Values")
  - Mean
  - Median
  - Mode

- Variation (or Summary of Differences Within Groups)
  - Range
  - Interquartile Range
  - Variance
  - Standard Deviation

Distribution - (of a variable) tells us what values the variable takes and how often it takes these values.

- Unimodal - having a single peak
- Bimodal - having two distinct peaks
- Symmetric - left and right half are mirror images.

Frequency Distribution

Consider a data set of 26 children of ages 1-6 years. 1,2,3,4,5,6,1,1,1,3,3,3,2,4,4,5,6,5,5,4,4,3,3,2,1,3 .Then the frequency distribution of variable ‘age’ can be tabulated as follows:

Frequency Distribution of Age

Age	**1**	**2**	**3**	**4**	**5**	**6**
Frequency	**5**	**3**	**7**	**5**	**4**	**2**

Grouped Frequency Distribution of Age:

Age Group	**1-2**	**3-4**	**5-6**
Frequency	**8**	**12**	**6**

Cumulative Frequency

Cumulative frequency of data in previous page

Age	1	2	3	4	5	6
Frequency	5	3	7	5	4	2
Cumulative Frequency	5	8	15	20	24	26

Age Group	1-2	3-4	5-6
Frequency	8	12	6
Cumulative Frequency	8	20	26

Data Presentation

Two types of statistical presentation of data - graphical and numerical.

Graphical Presentation: We look for the overall pattern and for striking deviations from that pattern. Over all pattern usually described by shape, center, and spread of the data. An individual value that falls outside the overall pattern is called an *outlier*.

Bar diagram and Pie charts are used for categorical variables.

Histogram, stem and leaf and Box-plot are used for numerical variable.

Numerical Presentation

- A fundamental concept in summary statistics is that of a *central value* for a set of observations and the extent to which the central value characterizes the whole set of data.

- Measures of central value such as the mean or median must be coupled with measures of data dispersion (e.g., average distance from the mean) to indicate how well the central value characterizes the data as a whole.

To understand how well a central value characterizes a set of observations, let us consider the following two sets of data:

A: 30, 50, 70

B: 40, 50, 60

The mean of both two data sets is 50.

But, the distance of the observations from the mean in data set A is larger than in the data set B.

Thus, the mean of data set B is a better representation of the data set than is the case for set A.

Shape – Center - Spread

- When we gather data, we want to uncover the "information" in it. One easy way to do that is to think of: "Shape –Center- Spread"

- ***Shape*** – What is the shape of the histogram?
- ***Center*** – What is the mean or median?
- ***Spread*** – What is the range or standard deviation?

Key Terms

- Measures of Central Tendency,

The Center

- Mean
  - μ, population; $\bar{x}$, sample
- Weighted Mean
- Median
- Mode

Key Terms

- Measures of Dispersion,

The Spread

- Range
- Variance
- Standard deviation
- Interquartile range

Key Terms

- Measures of Relative Position

- Quantiles
  - Quartiles
  - Percentiles

Key Terms

- Measures of Association

- **Coefficient of correlation, *r***
  - **Direction of the relationship:** direct ($r > 0$) or inverse ($r < 0$)
  - **Strength of the relationship:** When r is close to 1 or −1, the linear relationship between x and y is strong. When r is close to 0, the linear relationship between x and y is weak. When $r = 0$, there is no linear relationship between x and y.
- **Coefficient of determination, r^2**
  - The percent of total variation in y that is explained by variation in x.

Measuring the Central Tendency

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

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

  - Weighted arithmetic mean:
  - Trimmed mean: chopping extreme values

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

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

$$\mu = \frac{\sum x}{N}$$

Problem: Calculate the average number of truck shipments from the United States to five Canadian cities for the following data given in thousands of bags:

Montreal, 64.0; Ottawa, 15.0; Toronto, 285.0;
Vancouver, 228.0; Winnipeg, 45.0

The Center: Weighted Mean

- When what you have is **grouped data**, compute the mean using $\mu = (\Sigma w_i x_i)/\Sigma w_i$

  Problem: Calculate the average profit from truck shipments, United States to Canada, for the following data given in thousands of bags and profits per thousand bags:

Montreal	64.0	Ottawa	15.0	Toronto	285.0
	$15.00		$13.50		$15.50
Vancouver	228.0	Winnipeg	45.0		
	$12.00		$14.00		

- 8946/637

(Ans: $14.04 per thous. bags)

Mean

Class A--IQs of 13 Students

102
128
131
98
140
93
110
115
109
89
106
119
97

$\Sigma Y_i = 1437$

$$\text{Y-bar}_A = \frac{\Sigma Y_i}{n} = \frac{1437}{13} = 110.54$$

Class B--IQs of 13 Students

127
131
96
80
93
120
109
162
103
111
109
87
105

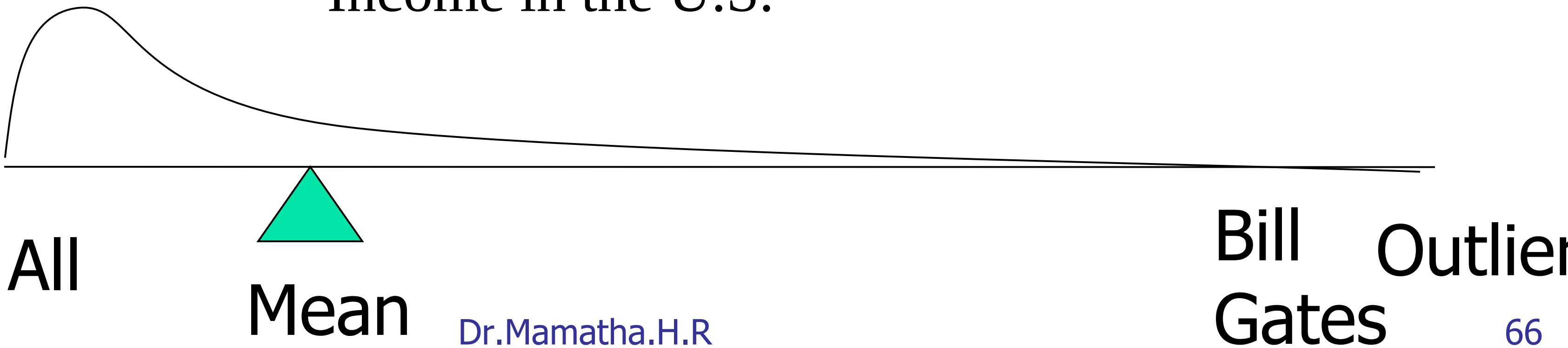
$\Sigma Y_i = 1433$

$$\text{Y-bar}_B = \frac{\Sigma Y_i}{n} = \frac{1433}{13} = 110.23$$

Mean

1. Means can be badly affected by outliers (data points with extreme values unlike the rest)
2. Outliers can make the mean a bad measure of central tendency or common experience

Income in the U.S.



The Center: Median

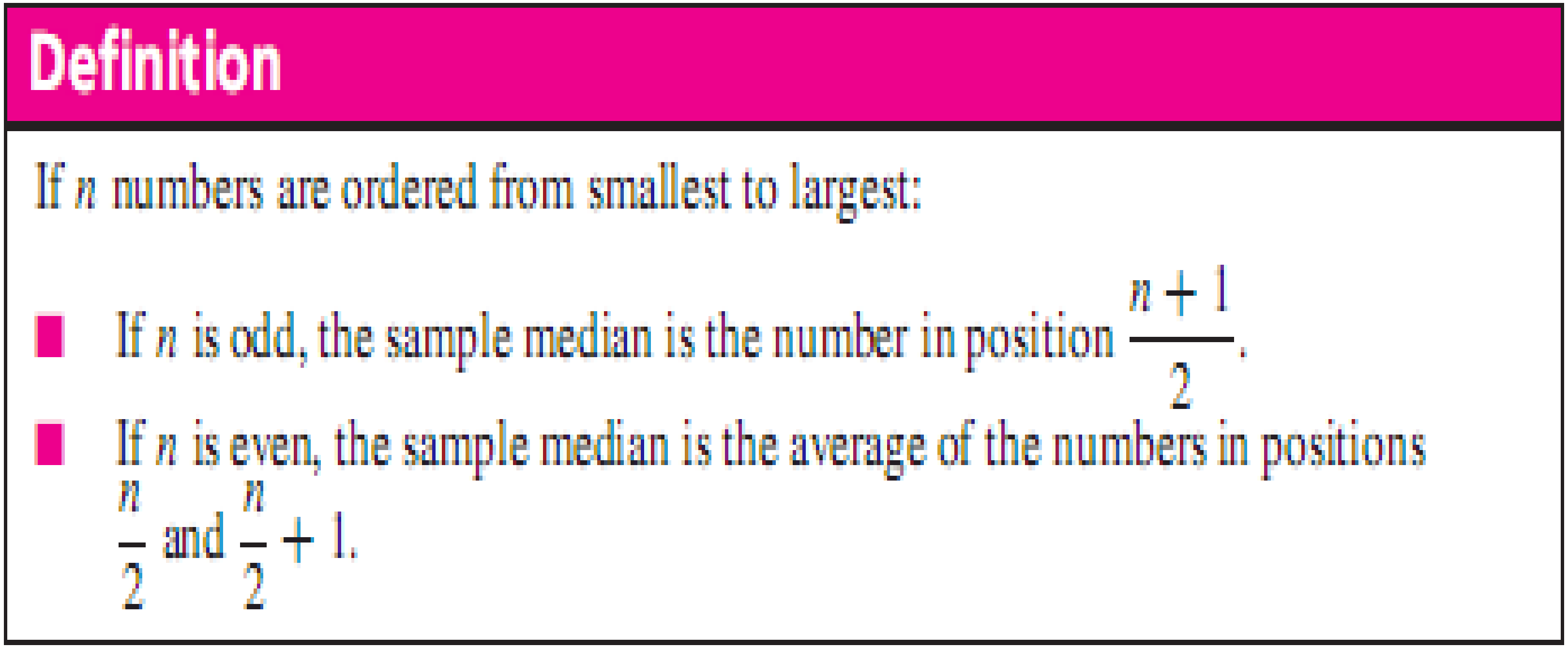
- To find the median:

If the data set has an **ODD** number of numbers, the median is the middle value.

If the data set has an **EVEN** number of numbers, the median is the AVERAGE of the middle two values.

(Note that the median of an even set of data values is not necessarily a member of the set of values.)

- The median is particularly useful if there are outliers in the data set, which otherwise tend to sway the value of an arithmetic mean.



Definition

If n numbers are ordered from smallest to largest:

- If n is odd, the sample median is the number in position $\dfrac{n+1}{2}$.
- If n is even, the sample median is the average of the numbers in positions $\dfrac{n}{2}$ and $\dfrac{n}{2}+1$.

Mean or Median

The median is less sensitive to outliers (extreme scores) than the mean and thus a better measure than the mean for highly skewed distributions,

e.g. family income.

For example mean of 20, 30, 40, and 990

(20+30+40+990)/4 =270.

The median of these four observations is

(30+40)/2 =35.

Here 3 observations out of 4 lie between 20-40. So, the mean 270 really fails to give a realistic picture of the major part of the data. It is influenced by extreme value 990.

Median

Class A--IQs of 13 Students

89

93

97

98

102

106

109

110

115

119

128

131

140

Median = 109

(six cases above, six below)

Median

If the first student were to drop out of Class A, there would be a new median:

~~89~~
93
97
98
102
106
109
110
115
119
128
131
140

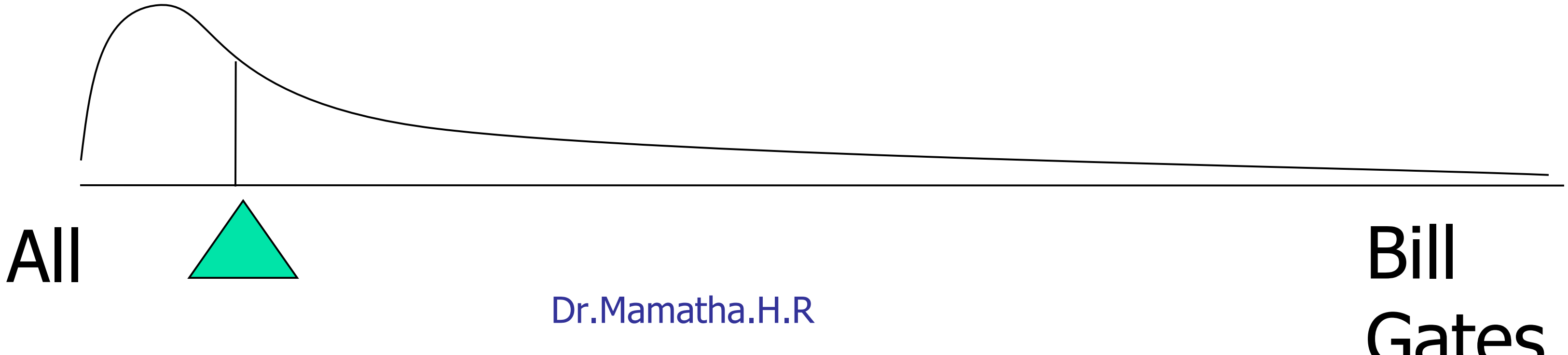
Median = 109.5

109 + 110 = 219/2 = 109.5

(six cases above, six below)

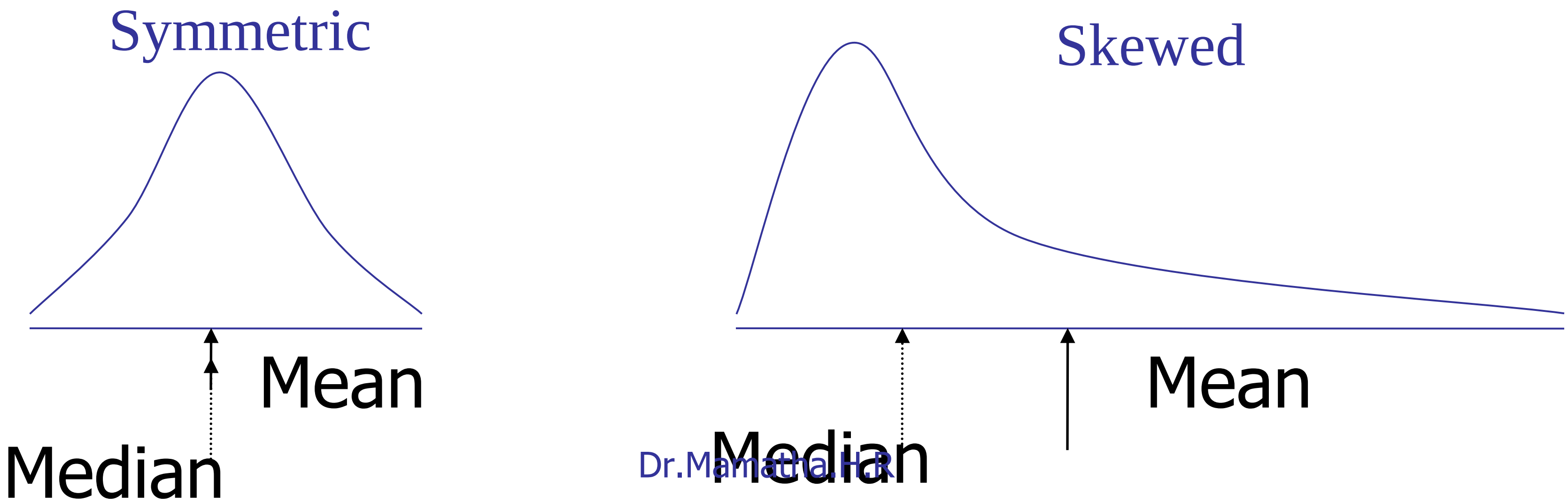
Median

1. The median is unaffected by outliers, making it a better measure of central tendency, better describing the "typical person" than the mean when data are skewed.



Median

2. If the recorded values for a variable form a symmetric distribution, the median and mean are identical.
3. In skewed data, the mean lies further toward the skew than the median.



Median

The middle score or measurement in a set of ranked scores or measurements; the point that divides a distribution into two equal halves.

Data are listed in order—the median is the point at which 50% of the cases are above and 50% below.

The 50th percentile.

Trimmed Mean

- The trimmed mean is computed by arranging the sample values in order, "trimming" an equal number of them from each end, and computing the mean of those remaining.
- If p% of the data are trimmed from each end, the resulting trimmed mean is called the "p% trimmed mean."
- There are no hard-and-fast rules on how many values to trim.
- The most commonly used trimmed means are the 5%, 10%, and 20% trimmed means.

- Note that the median can be thought of as an extreme form of trimmed mean, obtained by trimming away all but the middle one or two sample values.
- If the sample size is denoted by n, and a p% trimmed mean is desired, the number of data points to be trimmed is np/100
- Modified mean or Olympic average-leave maximum and minimum.

- It is used to reduce the effects of outliers on the calculated average.
- This method is best suited for data with large, erratic deviations or extremely skewed distributions.

- For the following data
- 30 75 79 80 80 105 126 138 149 179 179 191
- 223 232 232 236 240 242 245 247 254 274 384 470
- Compute the mean, median, and the 5%, 10%, and 20% trimmed means.

- Solution
- The mean is found by averaging together all 24 numbers, which produces a value of 195.42.
- The median is the average of the 12th and 13th numbers, which is
- (191 + 223)/2 = 207.00.
- To compute the 5% trimmed mean, we must drop 5%
- of the data from each end. This comes to (0.05)(24) = 1.2 observations.
- We round 1.2 to 1, and trim one observation off each end.

- The 5% trimmed mean is the average of the remaining 22 numbers:
- 75 + 79 +···+ 274 + 384/22= 190.45
- To compute the 10% trimmed mean, round off (0.1)(24) = 2.4 to 2.
- Drop 2 observations from each end, and then average the remaining 20:
- 79 + 80 +···+ 254 + 274/20= 186.55
- To compute the 20% trimmed mean, round off (0.2)(24) = 4.8 to 5. Drop 5 observations from each end, and then average the remaining 14:
- 105 + 126 +···+ 242 + 245/14= 194.07

- A figure skating competition produces the following scores
- 6.0,8.1,8.3,9.1,9.9
- Find the average and the 20% trimmed mean

Measuring the Central Tendency

- Mode
  - Value that occurs most frequently in the data
  - Unimodal, bimodal, trimodal
  - Empirical formula:

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

The Center: Mode

- The mode is the most frequent value.
- While there is just one value for the mean and one value for the median, there may be more than one value for the mode of a data set.
- The mode tends to be less frequently used than the mean or the median.

Mode

The most common data point is called the mode.

The combined IQ scores for Classes A & B:

80 87 89 93 93 96 97 98 102 103 105 106 109 109 109 110 111 115 119 120 127 128 131 131 140 162

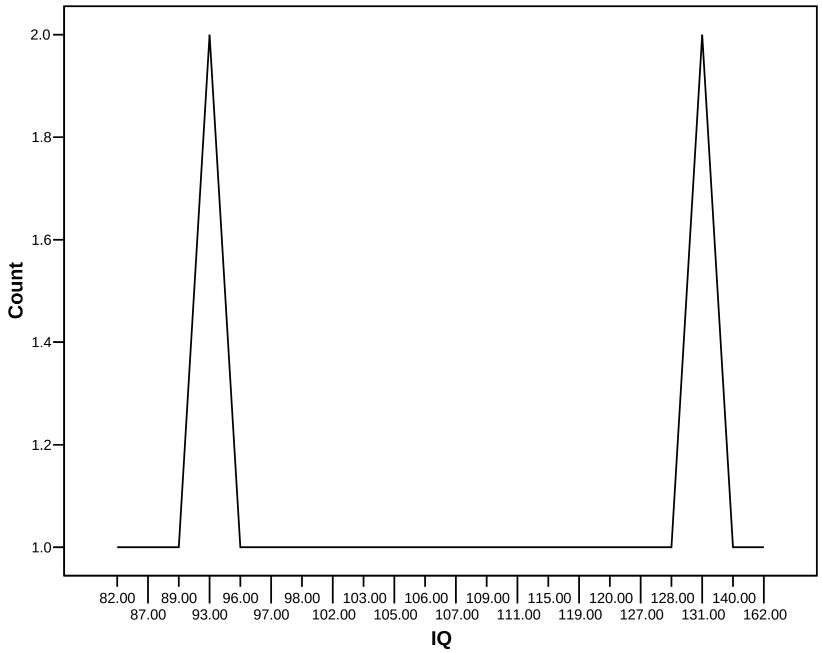
mode!!

It is possible to have more than one mode!

Mode

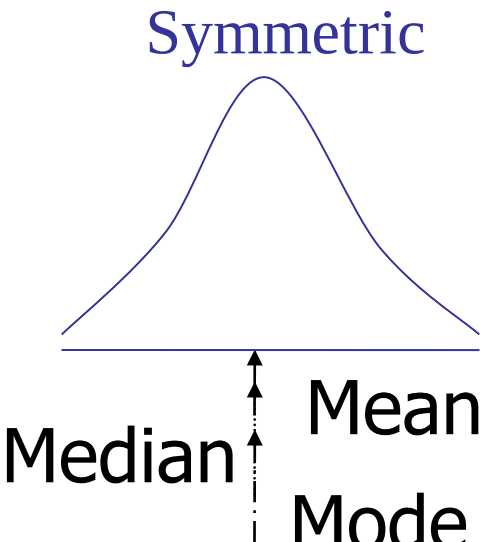
It may not be at the center of a distribution.

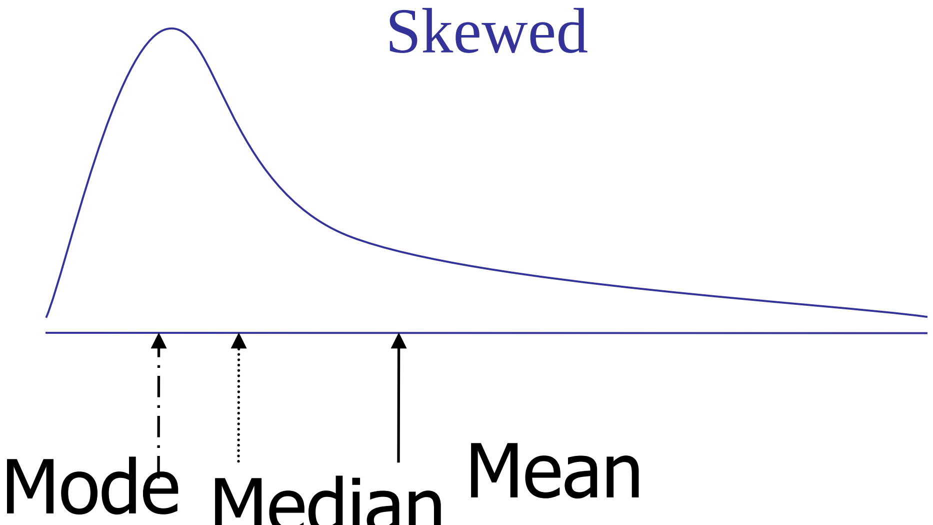
Data distribution on the right is "bimodal"



Mode

1. It may give you the most likely experience rather than the “typical” or “central” experience.
2. In symmetric distributions, the mean, median, and mode are the same.
3. In skewed data, the mean and median lie further toward the skew than the mode.





Shape: The “shape” of the data is called its “distribution”?

- If mean = median = mode, the shape of the distribution is **symmetric**.

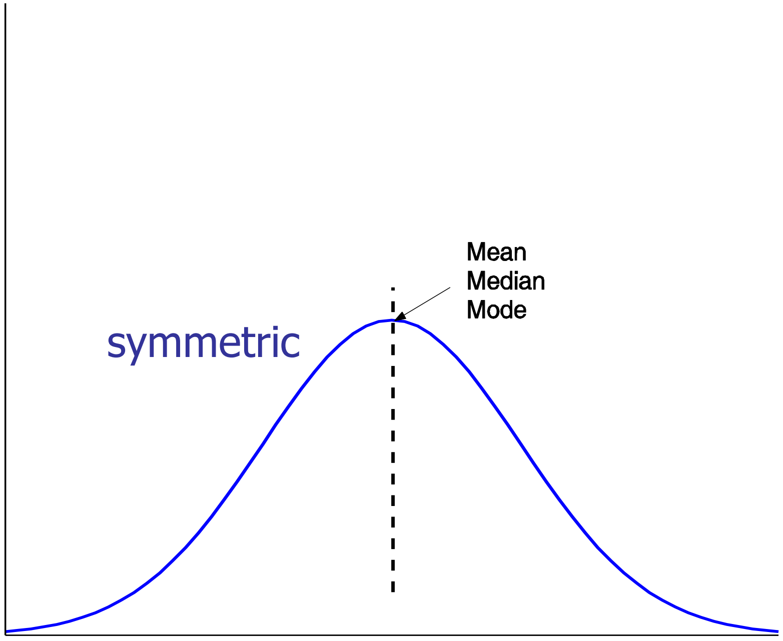
- If mode < median < mean, the shape of the distribution trails to the right, is **positively skewed**.

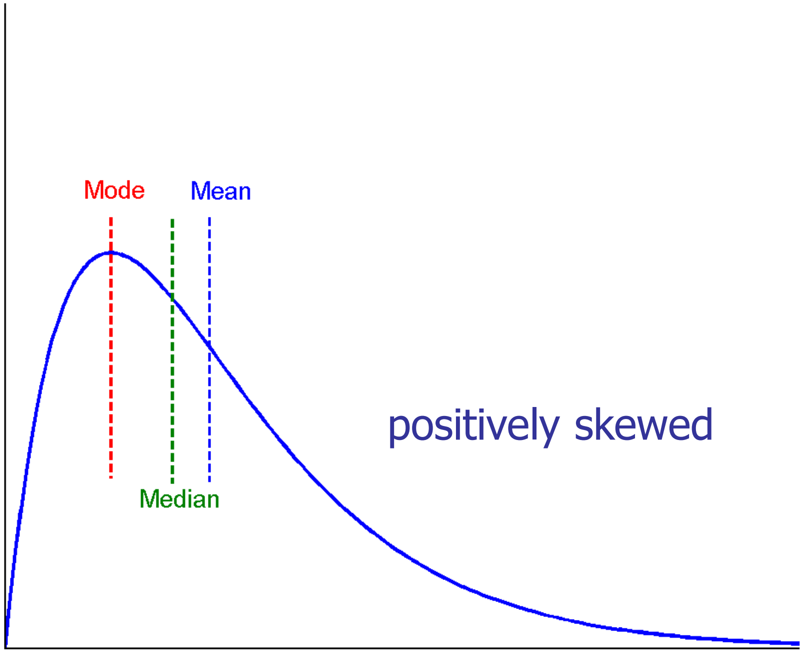
- If mean < median < mode, the shape of the distribution trails to the left, is **negatively skewed.**

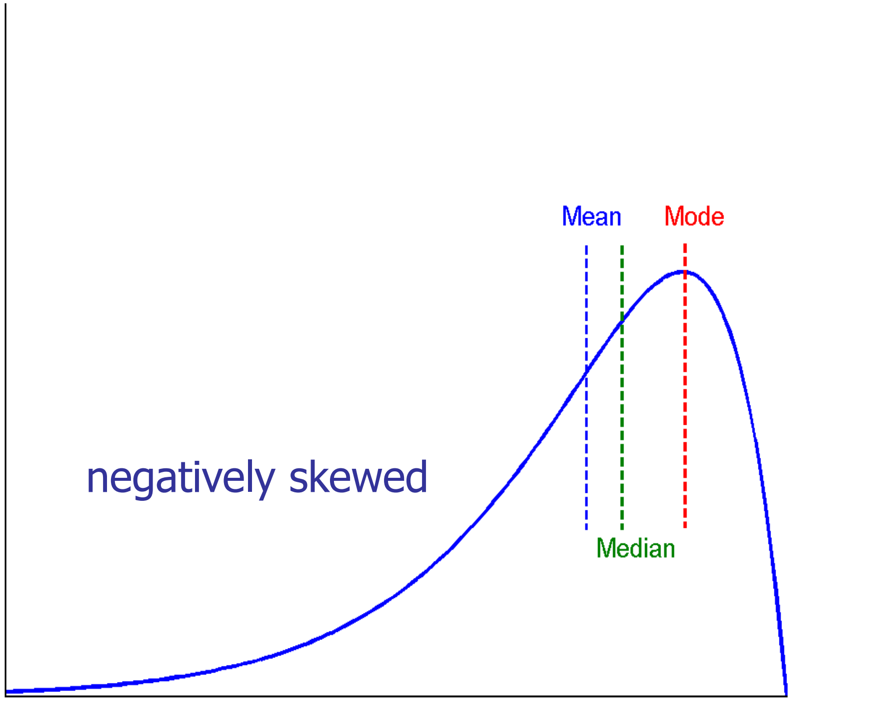
- Distributions of various “shapes” have different properties and names such as the “normal” distribution, which is also known as the “bell curve” (among mathematicians it is called the Gaussian Distribution).

Symmetric vs. Skewed Data

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







- **Example: Alex did a survey of how many games each of his 20 friends owned, and got this:**
- 9, 15, 11, 12, 3, 5, 10, 20, 14, 6, 8, 8, 12, 12, 18, 15, 6, 9, 18, 11
- Find the mean,median and mode

- To find the Mean, add up all the numbers, then divide by how many numbers there are:
- **Mean** = (9+15+11+12+3+5+10+20+14+6+8+ 8+12+12+18+15+6+9+18+11)/20 = **11.1**

- To find the Median, place the numbers in value order and find the middle number (or the mean of the middle two numbers). In this case the mean of the 10th and 11th values:
- 3, 5, 6, 6, 8, 8, 9, 9, 10, 11, 11, 12, 12, 12, 14, 15, 15, 18, 18, 20:
- **Median** = (11 + 11)/2 = **11**

- To find the Mode, or modal value, place the numbers in value order then count how many times each number exists. The Mode is the number which appears most often (there can be more than one mode):
- 3, 5, 6, 6, 8, 8, 9, 9, 10, 11, 11, 12, 12, 12, 14, 15, 15, 18, 18, 20:
- 12 appears three times, more often than the other values, so **Mode = 12**

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**: usually, a value higher/lower than 1.5 x IQR
- Variance and standard deviation (*sample: s, population: σ)*
  - **Variance**: (algebraic, scalable computation)

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n}x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n}x_i\right)^2\right] \qquad \sigma^2 = \frac{1}{N}\sum_{i=1}^{n}(x_i - \mu)^2 = \frac{1}{N}\sum_{i=1}^{n}x_i^2 - \mu^2$$

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

Range

The spread, or the distance, between the lowest and highest values of a variable.

To get the range for a variable, you subtract its lowest value from its highest value.

Class A--IQs of 13 Students		Class B--IQs of 13 Students	
102	115	127	162
128	109	131	103
131	89	96	111
98	106	80	109
140	119	93	87
93	97	120	105
110		109	

Class A Range = 140 - 89 = 51 **Class B Range = 162 - 80 = 82**

The Range

- The Range is the difference between the lowest and highest values.
- Example: In **{4, 6, 9, 3, 7}** the lowest value is 3, and the highest is 9.
- So the range is 9-3 = **6**.

The Range Can Be Misleading

- The range can sometimes be misleading when there are extremely high or low values.

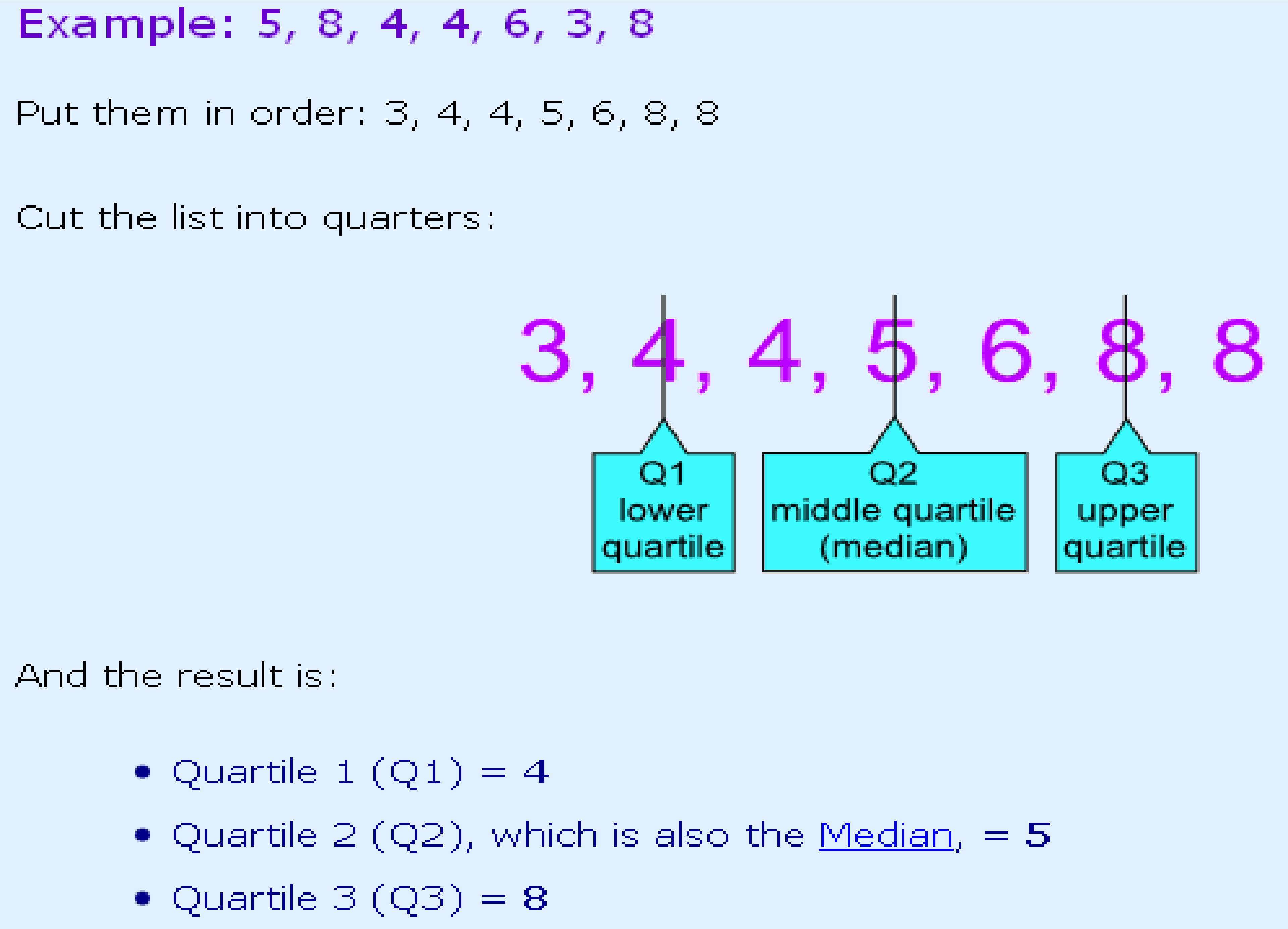
- Example: In **{8, 11, 5, 9, 7, 6, 3616}**:
- the lowest value is 5,
- and the highest is 3616,
- So the range is 3616-5 = **3611**.
- The single value of 3616 makes the range large, but most values are around 10.

- It is a measure of spread, but it is rarely used, because it depends only on the two extreme values and provides no information about the rest of the sample.

Quartiles

- Quartiles are the values that divide a list of numbers into quarters.
- **First** put the list of numbers in order
- **Then** cut the list into four equal parts
- The Quartiles are at the "cuts"

- The simplest method of computing quartiles by hand is as follows:
- Let n represent the sample size.
- Order the sample values from smallest to largest.
- To find the first quartile, compute the value 0.25(n +1).
- If this is an integer, then the sample value in that position is the first quartile.
- If not, then take the average of the sample values on either side of this value.
- The third quartile iscomputed in the same way, except that the value 0.75(n+1) is used.
- The second quartile uses the value 0.5(n + 1). The second quartile is identical to the median.



Example: 5, 8, 4, 4, 6, 3, 8

Put them in order: 3, 4, 4, 5, 6, 8, 8

Cut the list into quarters:

3, 4, 4, 5, 6, 8, 8

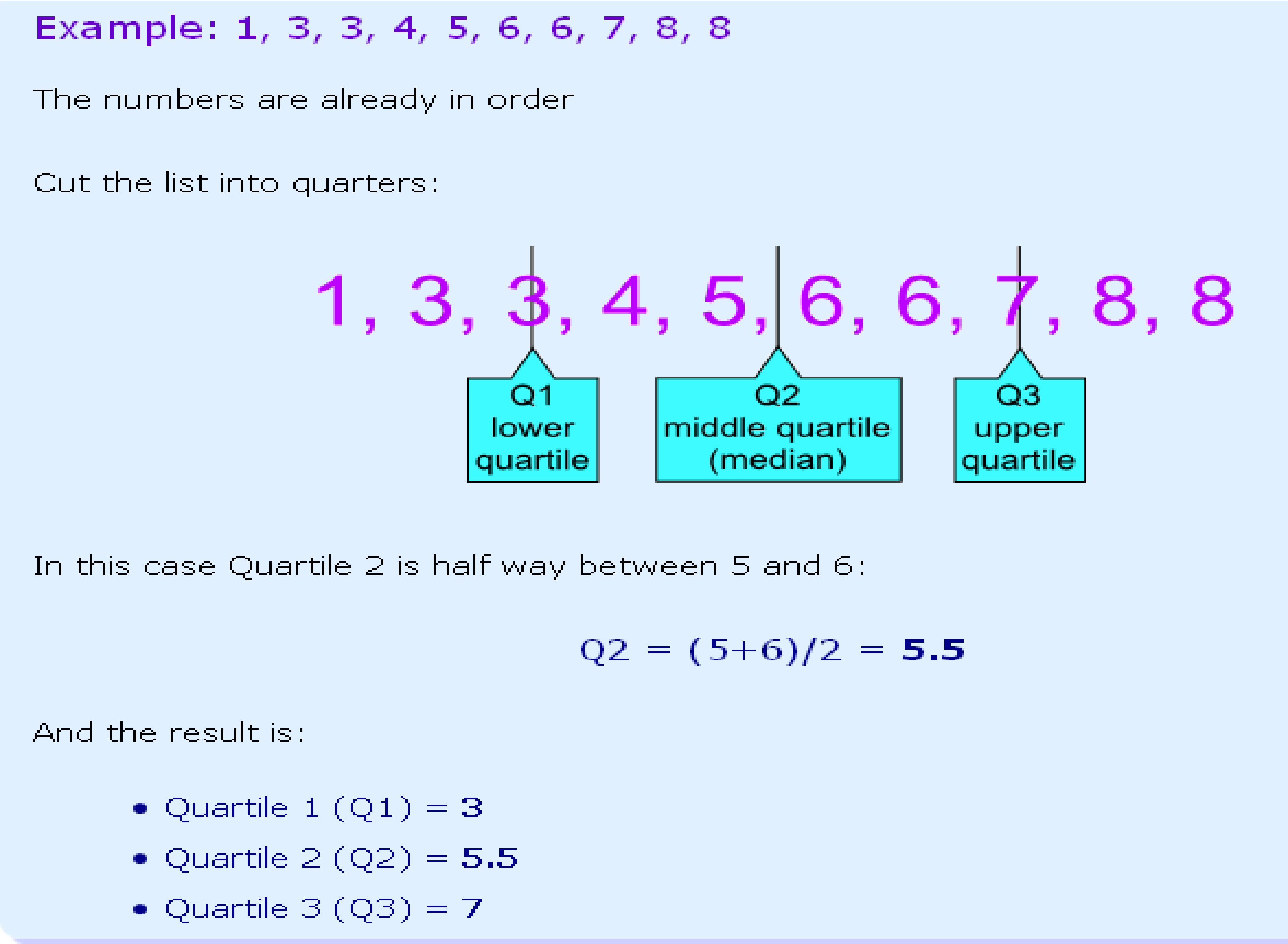
Q1
lower
quartile

Q2
middle quartile
(median)

Q3
upper
quartile

And the result is:

- Quartile 1 (Q1) = **4**
- Quartile 2 (Q2), which is also the Median, = **5**
- Quartile 3 (Q3) = **8**



Example: 1, 3, 3, 4, 5, 6, 6, 7, 8, 8

The numbers are already in order

Cut the list into quarters:

1, 3, 3, 4, 5, 6, 6, 7, 8, 8

Q1
lower
quartile

Q2
middle quartile
(median)

Q3
upper
quartile

In this case Quartile 2 is half way between 5 and 6:

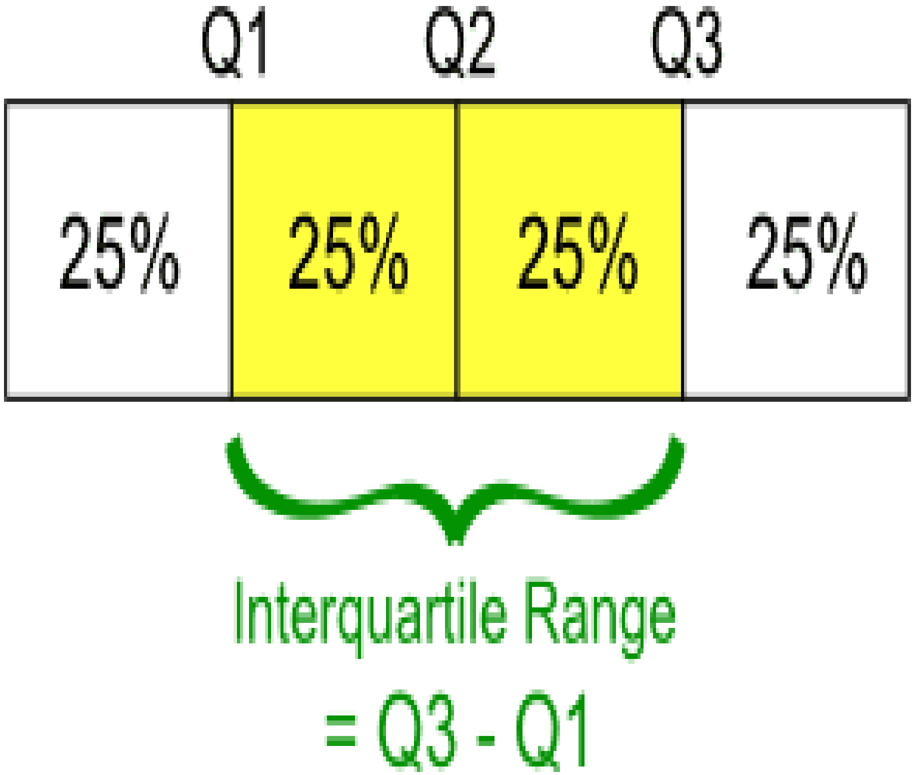
Q2 = (5+6)/2 = **5.5**

And the result is:

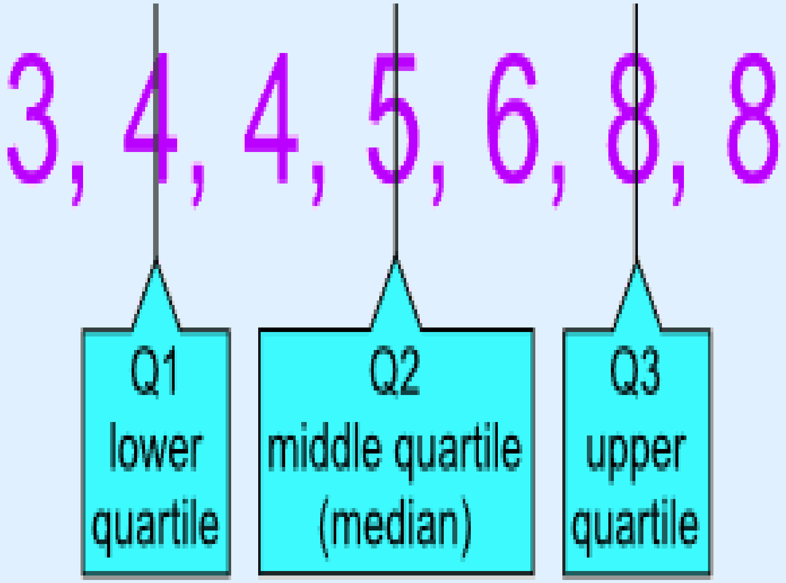
- Quartile 1 (Q1) = **3**
- Quartile 2 (Q2) = **5.5**
- Quartile 3 (Q3) = **7**

Interquartile Range

The "Interquartile Range" is from Q1 to Q3:



Example:


3, 4, 4, 5, 6, 8, 8

Q1
lower
quartile

Q2
middle quartile
(median)

Q3
upper
quartile

The **Interquartile Range** is:

Q3 - Q1 = 8 - 4 = **4**

Interquartile Range

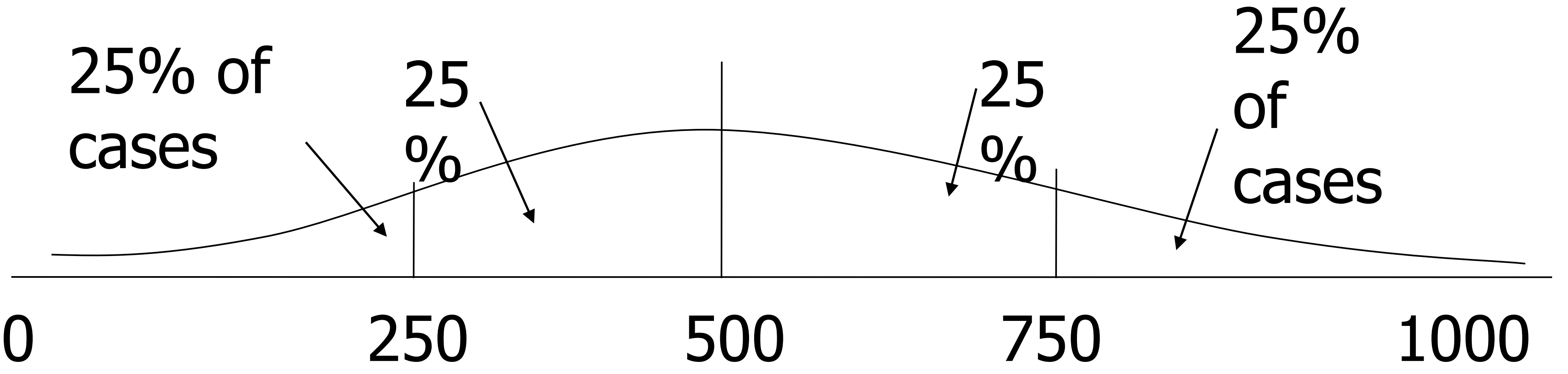
A quartile is the value that marks one of the divisions that breaks a series of values into four equal parts.

The median is a quartile and divides the cases in half.

25th percentile is a quartile that divides the first ¼ of cases from the latter ¾.

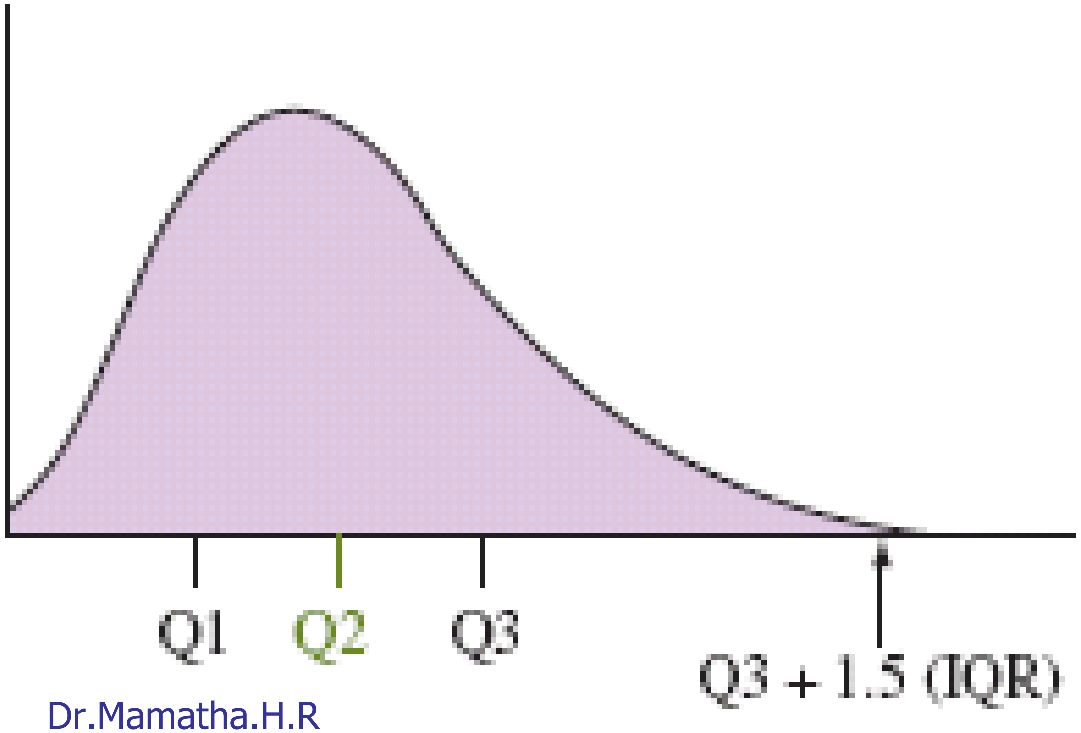
75th percentile is a quartile that divides the first ¾ of cases from the latter ¼.

The interquartile range is the distance or range between the 25th percentile and the 75th percentile. Below, what is the interquartile range?



Criteria for Identifying an Outlier

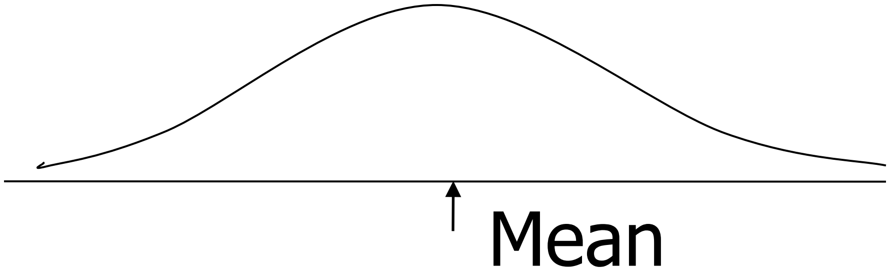
An observation is a potential outlier if it falls more than *1.5 x IQR* below the first or more than *1.5 x IQR* above the third quartile.



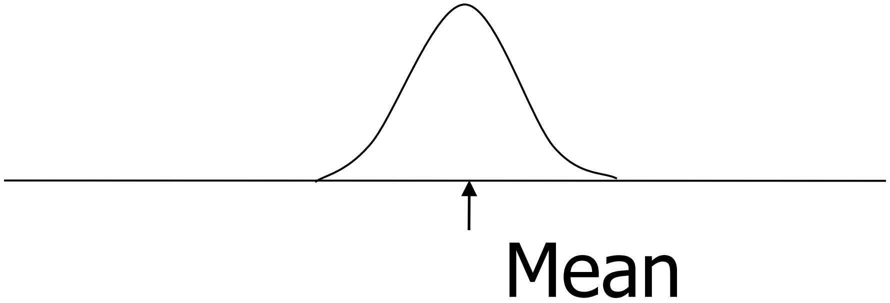
Variance

A measure of the spread of the recorded values on a variable. A measure of dispersion.

The larger the variance, the further the individual cases are from the mean.



The smaller the variance, the closer the individual scores are to the mean.



Variance

Variance is a number that at first seems complex to calculate.

Calculating variance starts with a "deviation."

A deviation is the distance away from the mean of a case's score.

Y*i* – Y-bar

If the average person's car costs $20,000, my deviation from the mean is - $14,000!

6K - 20K = -14K

Variance

The deviation of 102 from 110.54 is?   Deviation of 115?

Class A--IQs of 13 Students

102

128

131

98

140

93

110

115

109

89

106

119

97

Y-bar$_A$ $= 110.54$

Variance

The deviation of 102 from 110.54 is? Deviation of 115?

102 - 110.54 = -8.54 115 - 110.54 = 4.46

Class A--IQs of 13 Students

102
128
131
98
140
93
110
115
109
89
106
119
97

Y-bar$_A$ ***= 110.54***

Variance

- We want to add these to get total deviations, but if we were to do that, we would get zero every time. Why?
- We need a way to eliminate negative signs.

Squaring the deviations will eliminate negative signs...

A Deviation Squared: $(Y_i – \text{Y-bar})^2$

Back to the IQ example,

A deviation squared for 102 is: of 115:

$(102 - 110.54)^2 = (-8.54)^2 = 72.93$ $(115 - 110.54)^2 = (4.46)^2 = 19.89$

Variance

If you were to add all the squared deviations together, you'd get what we call the

"Sum of Squares."

Sum of Squares (SS) = $\Sigma\,(Y\mathit{i} - \text{Y-bar})^2$

SS = $(Y1 - \text{Y-bar})^2 + (Y2 - \text{Y-bar})^2 + \ldots + (Yn - \text{Y-bar})^2$

Variance

Class A, sum of squares:

$(102 – 110.54)^2 + (115 – 110.54)^2 +$

$(126 – 110.54)^2 + (109 – 110.54)^2 +$

$(131 – 110.54)^2 + (89 – 110.54)^2 +$

$(98 – 110.54)^2 + (106 – 110.54)^2 +$

$(140 – 110.54)^2 + (119 – 110.54)^2 +$

$(93 – 110.54)^2 + (97 – 110.54)^2 +$

$(110 – 110.54) = SS = 2825.39$

Class A--IQs of 13 Students

102	115
128	109
131	89
98	106
140	119
93	97
110	

Y-bar = 110.54

Variance

The last step…

The approximate average sum of squares is the variance.

SS/N = Variance for a population.

SS/n-1 = Variance for a sample.

Variance = $\Sigma(Y_i - \text{Y-bar})^2 / n - 1$

Why divide by n – 1 instead of n (Bessel's Correction)

- Standard deviation is used to estimate the amount of spread in the population from which sample was drawn.
- Ideally we should compute deviations from population mean instead of deviation from sample mean.
- Since population mean is unknown, sample mean is used in its place.
- Mathematical fact : deviation around sample mean is smaller than the deviations around population mean.
- Bessel's correction is an approach to reduce the bias error due to finite sample count.

Variance

For Class A, Variance = 2825.39 / n - 1

= 2825.39 / 12 = 235.45

How helpful is that???

Standard Deviation

To convert variance into something of meaning, let's create standard deviation.

The square root of the variance reveals the average deviation of the observations from the mean.

$$\text{s.d.} = \sqrt{\frac{\Sigma(Y\textit{i} - \text{Y-bar})^2}{n - 1}}$$

Standard Deviation

For Class A, the standard deviation is:

$$\sqrt{235.45} = 15.34$$

The average of persons’ deviation from the mean IQ of 110.54 is 15.34 IQ points.

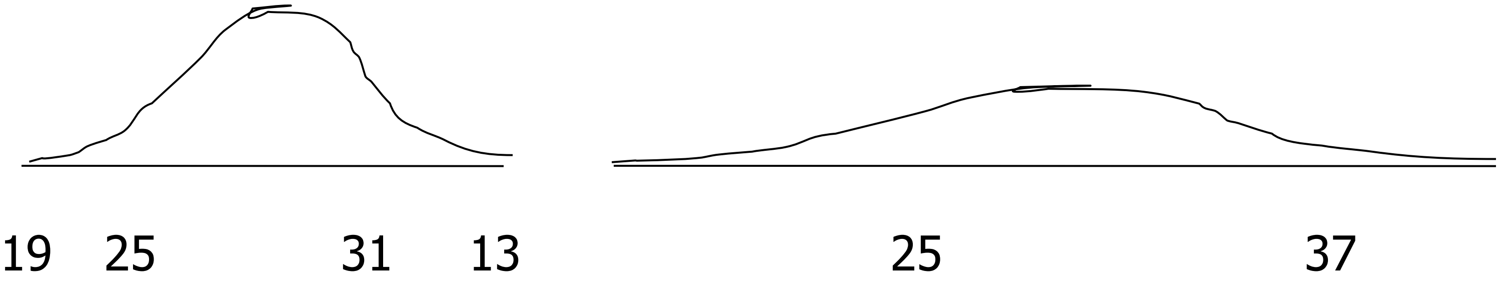
Review:

1. Deviation
2. Deviation squared
3. Sum of squares
4. Variance
5. Standard deviation

Standard Deviation

1. Larger s.d. = greater amounts of variation around the mean.

   For example:



$\bar{Y} = 25$

s.d. = 3

$\bar{Y} = 25$

s.d. = 6

2. s.d. = 0 only when all values are the same (only when you have a constant and not a "variable")
3. If you were to "rescale" a variable, the s.d. would change by the same magnitude—if we changed units above so the mean equaled 250, the s.d. on the left would be 30, and on the right, 60
4. Like the mean, the s.d. will be inflated by an outlier case value.

- Standard deviation is a measure of the variability of a single item.
- The standard deviation does not decline as the sample size increases.
- The estimate of the standard deviation becomes more stable as the sample size increases.

Practical Application for Understanding Variance and Standard Deviation

Even though we live in a world where we pay real dollars for goods and services (not percentages of income), most American employers issue raises based on percent of salary.

Why do supervisors think the most fair raise is a percentage raise?

Answer: 1) Because higher paid persons win the most money.

2) The easiest thing to do is raise everyone's salary by a fixed percent.

If your budget went up by 5%, salaries can go up by 5%.

The problem is that the flat percent raise gives unequal increased rewards. . .

Practical Application for Understanding Variance and Standard Deviation

Acme Toilet Cleaning Services

Salary Pool: $200,000

Incomes:

President: $100K; Manager: 50K; Secretary: 40K; and Toilet Cleaner: 10K

Mean: $50K

Range: $90K

Variance: $1,050,000,000

Standard Deviation: $32.4K

These can be considered "measures of inequality"

Now, let's apply a 5% raise.

Practical Application for Understanding Variance and Standard Deviation

After a 5% raise, the pool of money increases by $10K to $210,000

Incomes:

President: $105K; Manager: 52.5K; Secretary: 42K; and Toilet Cleaner: 10.5K

Mean: $52.5K – went up by 5%

Range: $94.5K – went up by 5%

Variance: $1,157,625,000

Standard Deviation: $34K –went up by 5%

Measures of Inequality

The flat percentage raise increased inequality. The top earner got 50% of the new money. The bottom earner got 5% of the new money. Measures of inequality went up by 5%.

Last year's statistics:

Acme Toilet Cleaning Services annual payroll of $200K

Incomes:

$100K, 50K, 40K, and 10K

Mean: $50K

Range: $90K; Variance: $1,050,000,000; Standard Deviation: $32.4K

Practical Application for Understanding Variance and Standard Deviation

The flat percentage raise increased inequality. The top earner got 50% of the new money. The bottom earner got 5% of the new money. Inequality increased by 5%.

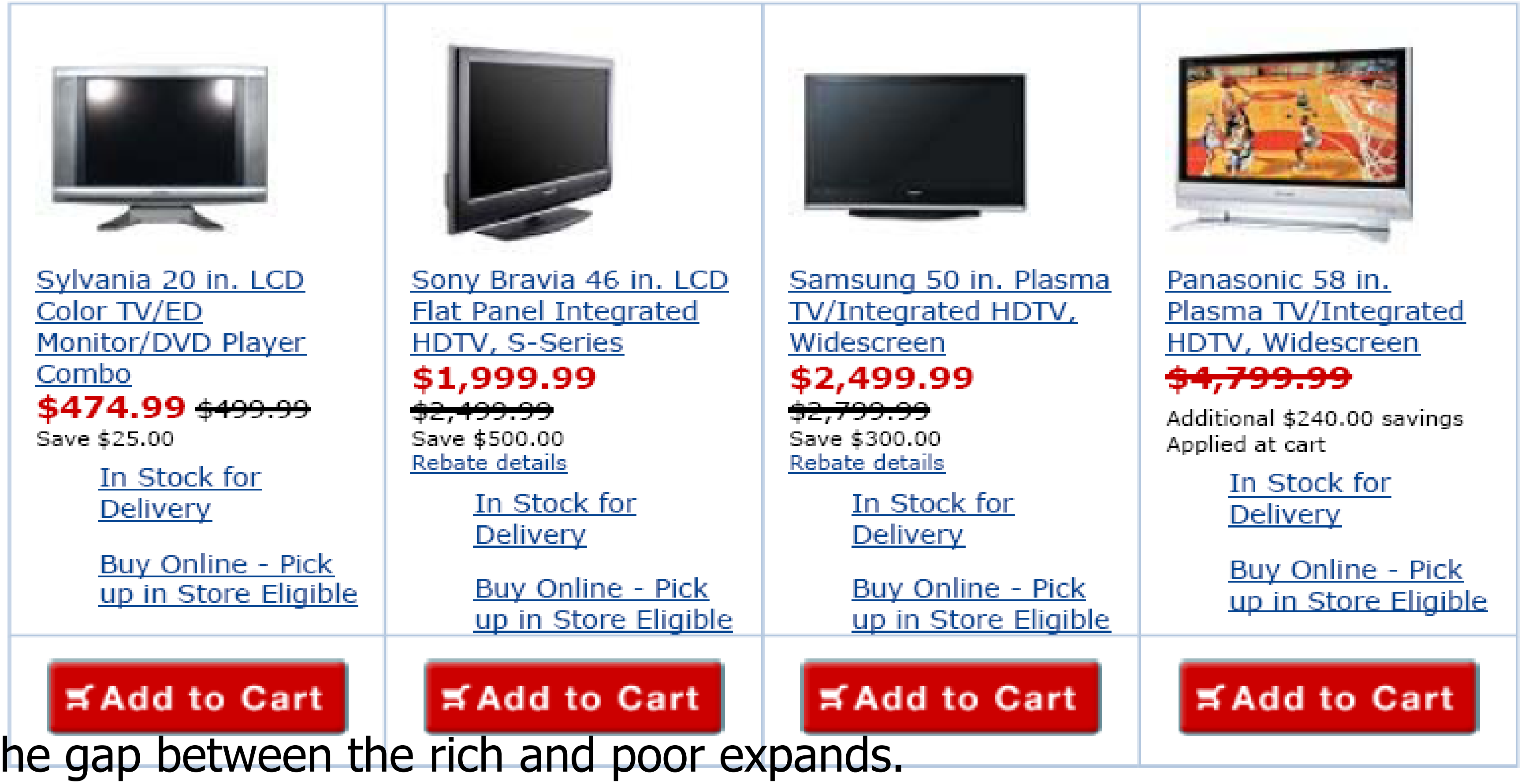
Since we pay for goods and services in real dollars, not in percentages, there are substantially more new things the top earners can purchase compared with the bottom earner for the rest of their employment years.

Acme Toilet Cleaning Services is giving the earners $5,000, $2,500, $2,000, and $500 more respectively ***each and every year forever***.

What does this mean in terms of compounding raises?

Acme is essentially saying: "Each year we'll buy you a new TV, in addition to everything else you buy, here's what you'll get:"

Practical Application for Understanding Variance and Standard Deviation



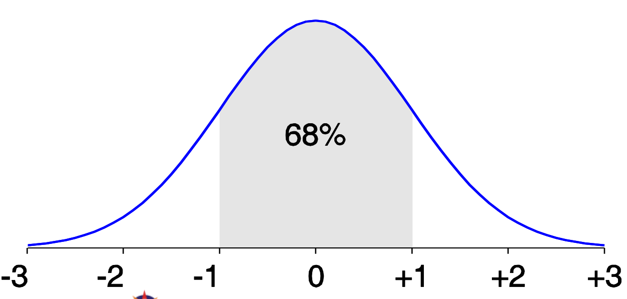
Toilet Cleaner	Secretary	Manager	President
Sylvania 20 in. LCD Color TV/ED Monitor/DVD Player Combo $474.99 ~~$499.99~~ Save $25.00 In Stock for Delivery Buy Online - Pick up in Store Eligible	Sony Bravia 46 in. LCD Flat Panel Integrated HDTV, S-Series $1,999.99 ~~$2,499.99~~ Save $500.00 Rebate details In Stock for Delivery Buy Online - Pick up in Store Eligible	Samsung 50 in. Plasma TV/Integrated HDTV, Widescreen $2,499.99 ~~$2,799.99~~ Save $300.00 Rebate details In Stock for Delivery Buy Online - Pick up in Store Eligible	Panasonic 58 in. Plasma TV/Integrated HDTV, Widescreen ~~$4,799.99~~ Additional $240.00 savings Applied at cart In Stock for Delivery Buy Online - Pick up in Store Eligible
Add to Cart	Add to Cart	Add to Cart	Add to Cart

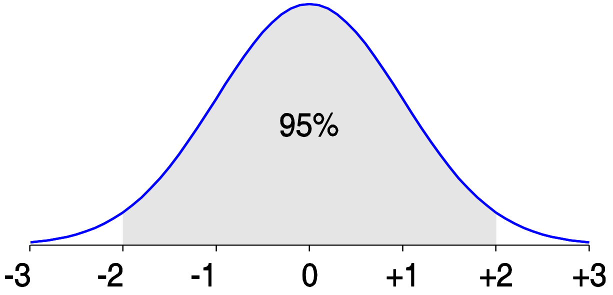
The gap between the rich and poor expands.

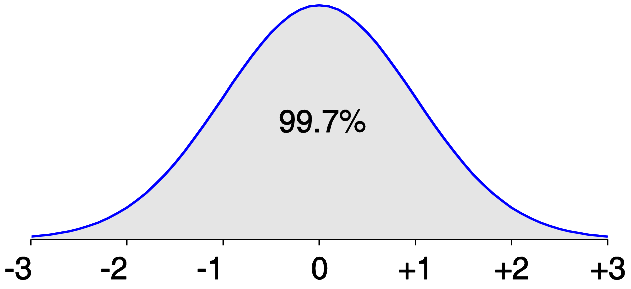
This is why some progressive organizations give a percentage raise with a flat increase for lowest wage earners. For example, 5% or $1,000, whichever is greater.

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







124