Data Science – Semester 5 – Fall 2022/2023

INTRODUCTION TO DATA MINING & Machine Learning

Lecture 2
Getting to know your data

Outline

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

Types of Datasets

- **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

What is Data

- Collection of data objects and their attributes

- An attribute is a property or characteristic of an object
  - Examples: eye color of a person, temperature, etc.
  - Attribute is also known as variable, field, characteristic, dimension, or feature
- A collection of attributes describe an object
  - Object is also known as record, point, case, sample, entity, or instance

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

Types of Attributes

- There are different types of attributes
  - Nominal
    - Examples: ID numbers, eye color, zip codes
  - Ordinal
    - Examples: rankings (e.g., taste of potato chips on a scale from 1-10), grades, height {tall, medium, short}
  - Interval
    - Examples: calendar dates, temperatures in Celsius or Fahrenheit.
  - Ratio
    - Examples: temperature in Kelvin, length, counts, elapsed time (e.g., time to run a race)

	Attribute Type	Description	Examples	Operations
Categorical Qualitative	Nominal	Nominal attribute values only distinguish. (=, ≠)	zip codes, employee ID numbers, eye color, sex: {*male, female*}	mode, entropy, contingency correlation, $\chi 2$ test
	Ordinal	Ordinal attribute values also order objects. (<, >)	hardness of minerals, {*good, better, best*}, grades, street numbers	median, percentiles, rank correlation, run tests, sign tests
Numeric Quantitative	Interval	For interval attributes, differences between values are meaningful. (+, -)	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, current	geometric mean, harmonic mean, percent variation

Discrete and Continuous Attributes

- **Discrete Attribute**
  - Has only a finite or countably infinite set of values
  - Examples: zip codes, counts, or the set of words in a collection of documents
  - Often represented as integer variables.
  - Note: binary attributes are a special case of discrete attributes
- **Continuous Attribute**
  - Has real numbers as attribute values
  - Examples: 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.

Outline

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

Metrics

- Measures of central tendency: measure the location of the middle or center (average) of a data distribution.
  - given an attribute, where do most of its values fall?
  - mean, median, mode
- Dispersion of the data: how are the data spread out?
  - range, quartiles, and interquartile range
  - five-number summary, boxplots;
  - variance and standard deviation of the data
- graphic displays of basic statistical descriptions to visually inspect our data
  - bar charts, pie charts, and line graphs
  - quantile plots, quantile–quantile plots, histograms, and scatter plots.

Measuring central tendency of data

- Arithmetic mean: most common type of average
  - N values of observations, x_i value of a numeric attribute X
  - Weighted arithmetic mean
  - Trimmed mean: chopping extreme values

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + x_2 + \cdots + x_N}{N}$$

Salaries of 10 staff

Staff	1	2	3	4	5	6	7	8	9	10
Salary	15k	18k	16k	14k	15k	15k	12k	17k	90k	95k

$$mean = \frac{15 + 18 ... + 95}{10} = 30.7$$

- Pros: Simple and intuitive
- Cons: can be skewed by outliers — it doesn't deal well with wildly varying samples. The average of 100, 200 and -300 is 0, which is misleading.

Measuring central tendency of data

- Median: For skewed asymmetric data. The middle value in a set of ordered data values.
  - N values of observations, sorted in increasing order
  - If N is odd → median is the middle value
  - If N is even → average of the two middlemost values

Salaries of 10 staff

Staff	1	2	3	4	5	6	7	8	9	10
Salary	15k	18k	16k	14k	15k	15k	12k	17k	90k	95k

12 14 15 15 15 16 18 17 90 95

$$\frac{15+16}{2}=15.5$$

- Pros: Handles outliers well. Splits data into two groups, each with the same number of items
- Cons: Can be harder to calculate: you need to sort the list first

Measuring central tendency of data

- Mode: value that occurs most frequently in the set.
  - Data sets with one, two, or three modes are respectively called **unimodal**, **bimodal**, and **trimodal**.
  - if each data value occurs only once, then there is no mode.

Salaries of 10 staff

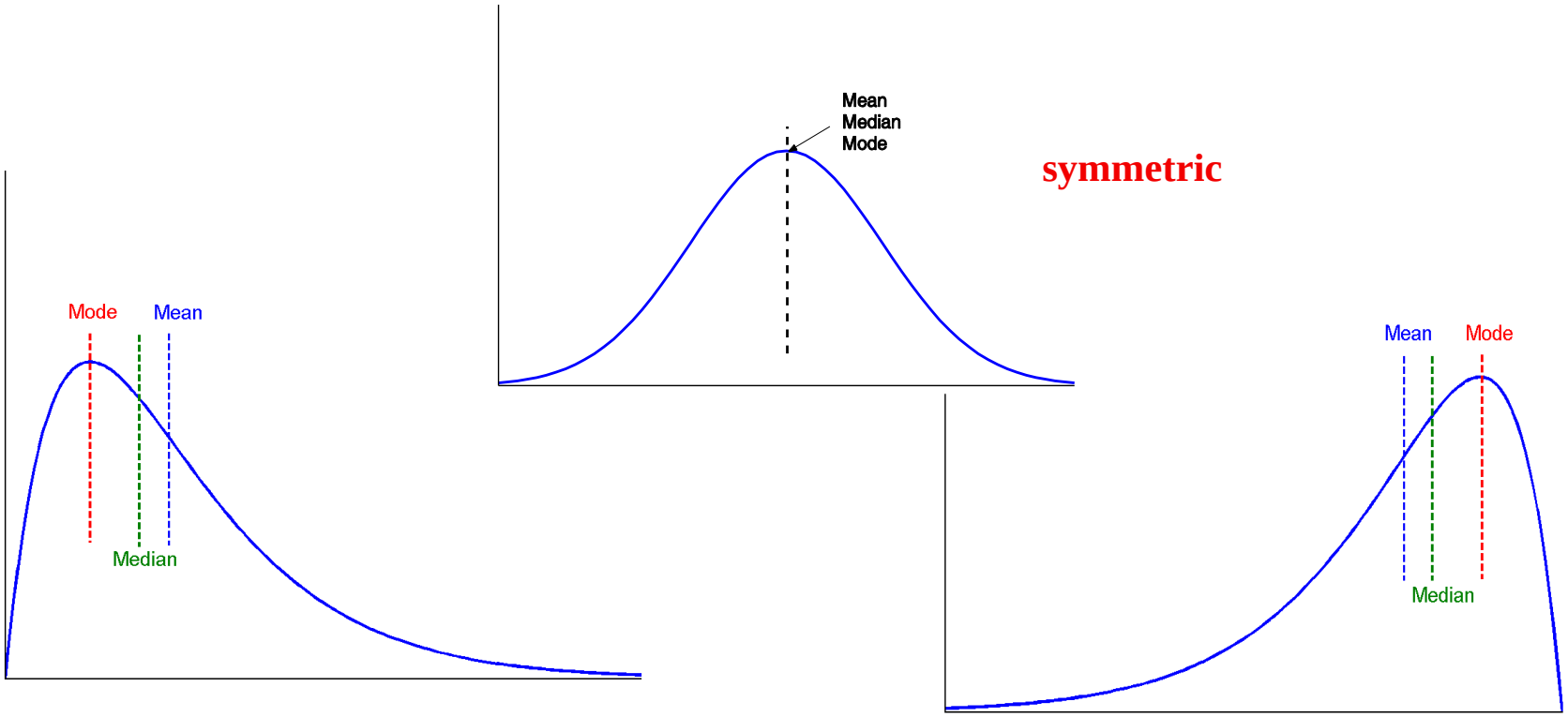
Staff	1	2	3	4	5	6	7	8	9	10
Salary	15k	18k	16k	14k	15k	15k	12k	17k	90k	95k

$$mode = 15$$

- Pros:
  - Works well for exclusive voting situations (this choice or that one; no compromise)
  - Gives a choice that the most people wanted (whereas the average can give a choice that nobody wanted).
- Cons:
  - Requires more effort to compute (have to tally up the votes)

Symmetric vs Skewed data

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

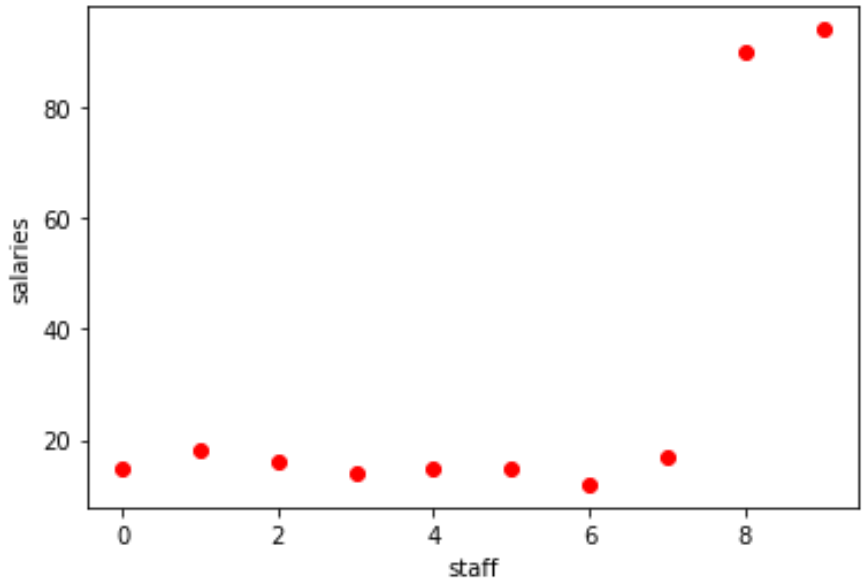


Symmetric vs Skewed data

- Mean = 30.7
- Median = 15.5
- Mode = 15

Salaries of 10 staff

Staff	1	2	3	4	5	6	7	8	9	10
Salary	15k	18k	16k	14k	15k	15k	12k	17k	90k	95k



Measuring dispersion of data

- Measures the dispersion or spread of numeric data
- Answers the question: "How much does my data vary?"

- Example: A measure of spread gives us an idea of how well the mean, for example, represents the data. If the spread of values in the data set is large, the mean is not as representative of the data as if the spread of data is small.

- **Range, Quartiles, and Interquartile Range**
- **Five-Number Summary, Boxplots, and Outliers**
- **Variance and standard deviation**

Measuring dispersion of data

- Range = max – min

Salaries of 10 staff

Staff	1	2	3	4	5	6	7	8	9	10
Salary	15k	18k	16k	14k	15k	15k	12k	17k	90k	95k

range = 95 – 12 = 83

- Range alone doesn’t provide much information
- But when combined with other metrics (e.g. mean, median, mode), it may tell us how spread our data is

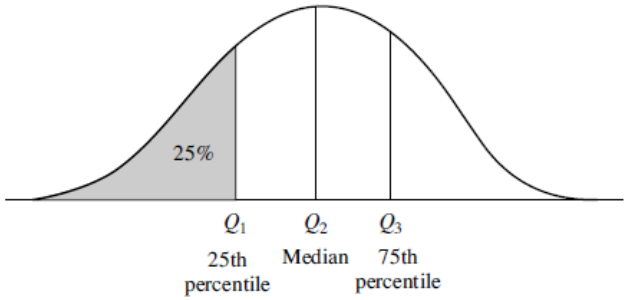
Measuring dispersion of data

- Percentiles, Quartiles and Interquartile range:
- we can consider the maximum value of a distribution in another way.
- think of it as the value in a set of data that has 100% of the observations at or below it. We call it 100^{th} percentile
- From this perspective, the median, which has 50% of the observations at or below it, is the 50^{th} percentile
- p^{th} percentile of a distribution is the value such that p percent of the observations fall at or below it
- most commonly used percentiles other than the median are 25^{th} percentile and the 75^{th} percentile

Measuring dispersion of data

- Percentiles, Quartiles and Interquartile range:

- 25th percentile → first quartile
- 50th percentile (median) → second quartile
- 75th percentile → third quartile

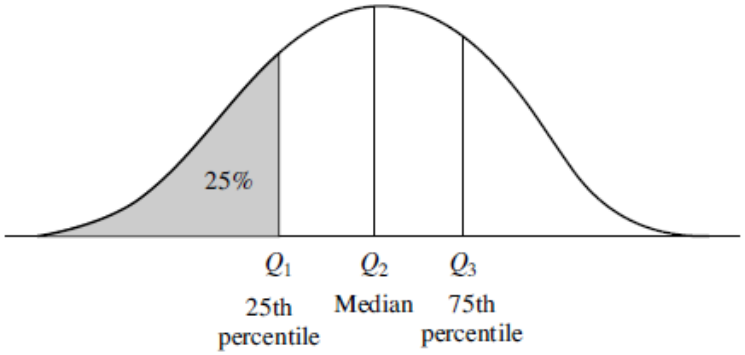


- Quartiles are a useful measure of spread because they are much less affected by outliers or a skewed data set than the equivalent measures of mean and standard deviation.
- Quartiles are often reported along with the median as the best choice of measure of spread and central tendency, respectively, when dealing with skewed and/or data with outliers

Measuring dispersion of data

- Percentiles, Quartiles and Interquartile range:

- Common way of expressing quartiles is an Interquartile range
- IQR = Q3 – Q1



Measuring dispersion of data

- Percentiles, Quartiles and Interquartile range

Salaries of 10 staff

Staff	1	2	3	4	5	6	7	8	9	10
Salary	15k	18k	16k	14k	15k	15k	12k	17k	90k	95k

Sorted values

12 14 15 15 15 16 17 18 90 95

- Q2 = 50th percentile = median = 15.5
- Q1 = 25th percentile: 15
- Q3 = 75th percentile = 18

- IQR = 18 - 15 = 3 → which suggests that our data is closer together

Measuring dispersion of data

- Five-Number Summary, Boxplots, and Outliers

- No single numeric measure of spread is very useful for describing skewed distributions.
- The five-number summary of a distribution consists of:
  - Minimum, Q1, Median, Q3, Maximum

- A common rule of thumb for identifying suspected outliers is to single out values falling:
  - Below Q1 – 1.5*IQR
  - Above Q3 + 1.5*IQR

  - In our example:
    - Q1 – 1.5*IQR = 15 - 1.5*3 = 10.5 → no outliers
    - Q3 + 1.5*IQR = 18 + 1.5*3 = 22.5 → outliers are 90k and 95k

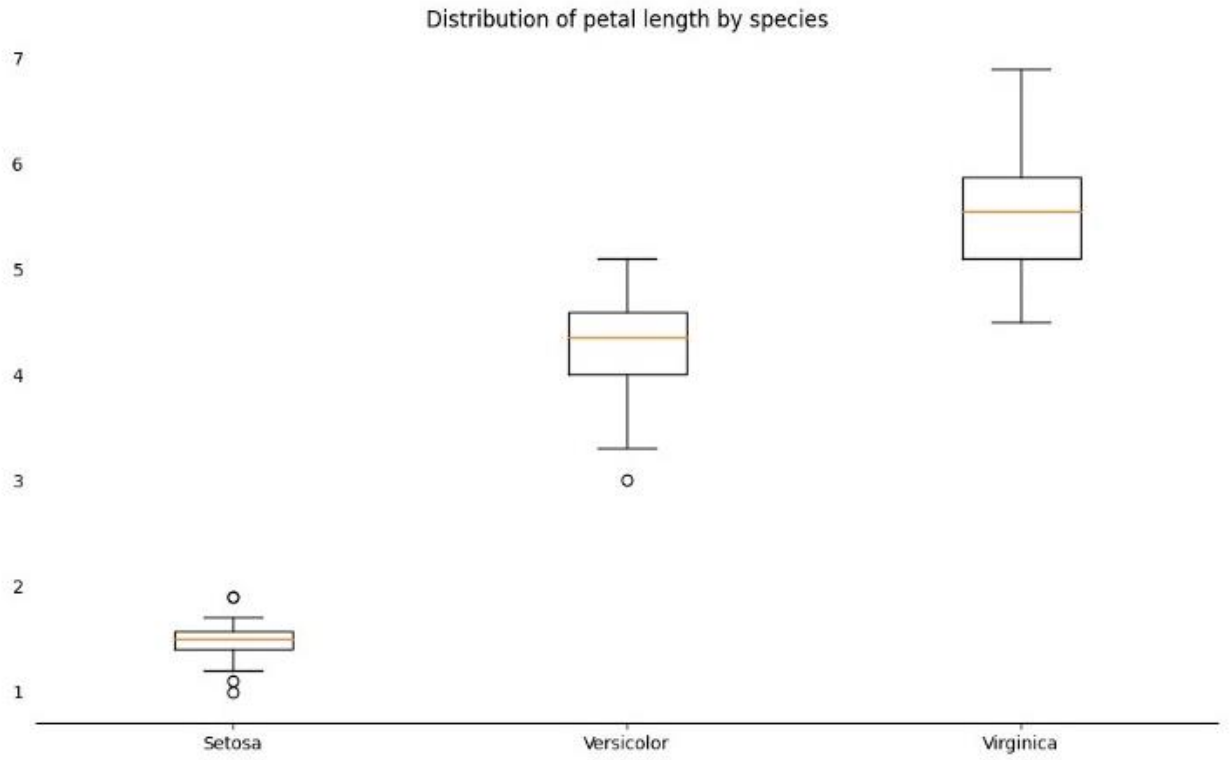
Measuring dispersion of data

- Five-Number Summary, Boxplots, and Outliers

- Boxplots are a popular way of visualizing a distribution. A boxplot incorporates the five-number summary as follows:
  - ends of the box are at the quartiles so that the box length is the interquartile range.
  - median is marked by a line within the box.
  - Two lines (called whiskers) outside the box extend to minimum and maximum
  - Outliers: points beyond a specified outlier threshold, plotted individually

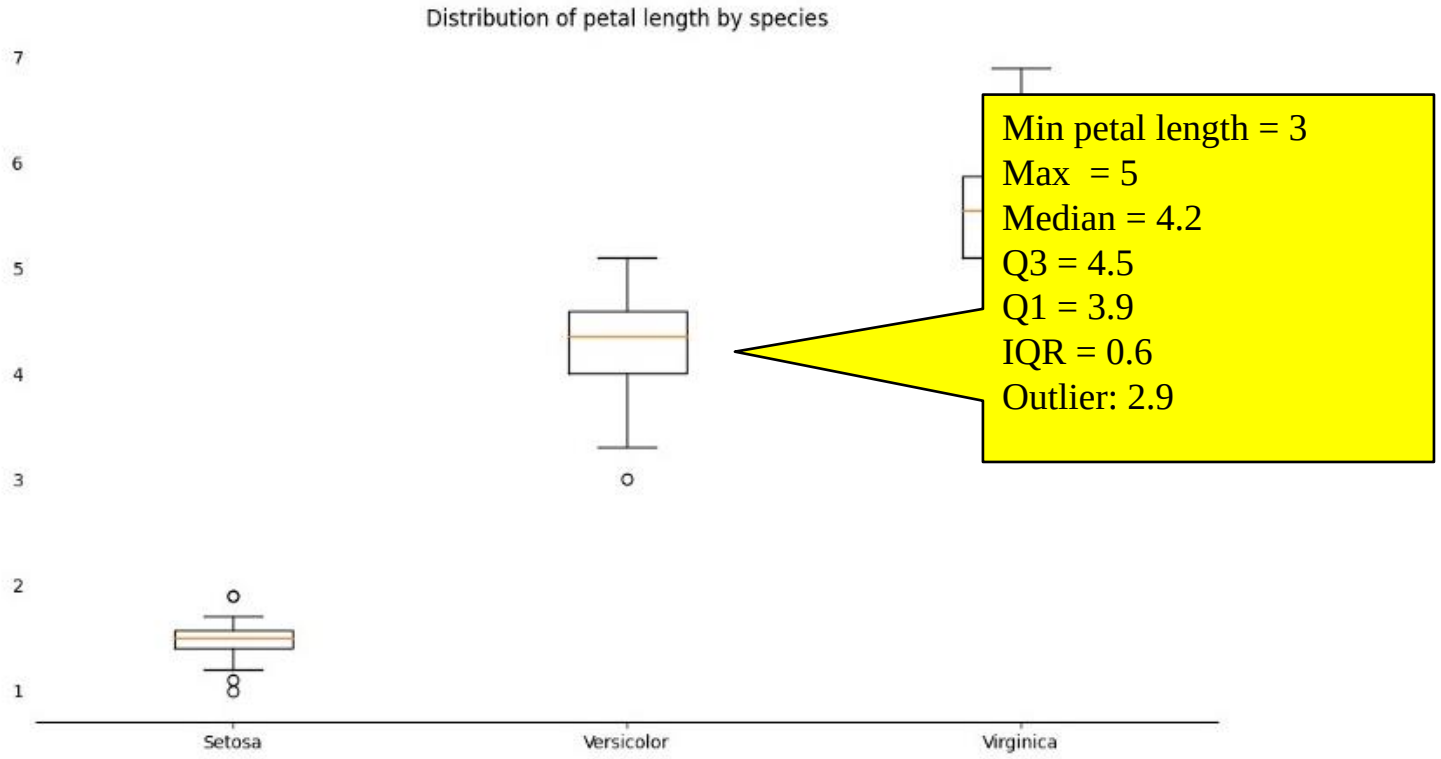
Measuring dispersion of data

- Five-Number Summary, Boxplots, and Outliers
  - Example: distribution of petal length by species (from Iris dataset)



Measuring dispersion of data

- Five-Number Summary, Boxplots, and Outliers
  - Example: distribution of petal length by species (from Iris dataset)



Measuring dispersion of data

- Variance and standard deviation

- A low standard deviation means that the data observations tend to be very close to the mean, while a high standard deviation indicates that the data are spread out over a large range of values.

- Variance: $$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 = \left(\frac{1}{N}\sum_{i=1}^{N}x_i^2\right) - \bar{x}^2,$$

- Standard deviation: σ (square root of variance)

Measuring dispersion of data

- Variance and standard deviation

Salaries of 10 staff

Staff	1	2	3	4	5	6	7	8	9	10
Salary	15k	18k	16k	14k	15k	15k	12k	17k	90k	95k

- Mean = 37.5
- Variance = $\frac{1}{10} \times \left((15-37.5)^2 + (18-37.5)^2 + \ldots + (95-37.5)^2\right) = 958.41$
- Standard deviation: std = 30.96

Measuring dispersion of data

- Variance and standard deviation

- Some properties about standard deviation:

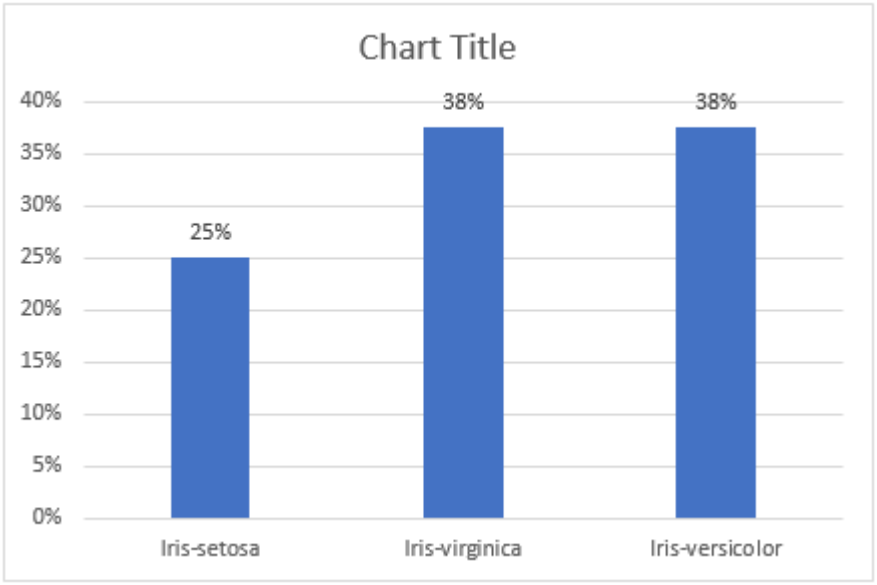
- measures spread about the mean and should be considered only when the mean is chosen as the measure of center.
- Std = 0 only when there is no spread, that is, when all observations have the same value. Otherwise, std > 0.

Graphic displays

- **Boxplot**: graphic display of five-number summary
- **Histogram**
- **Quantile plot, Quantile-quantile (q-q) plot**
- **Scatter plot**

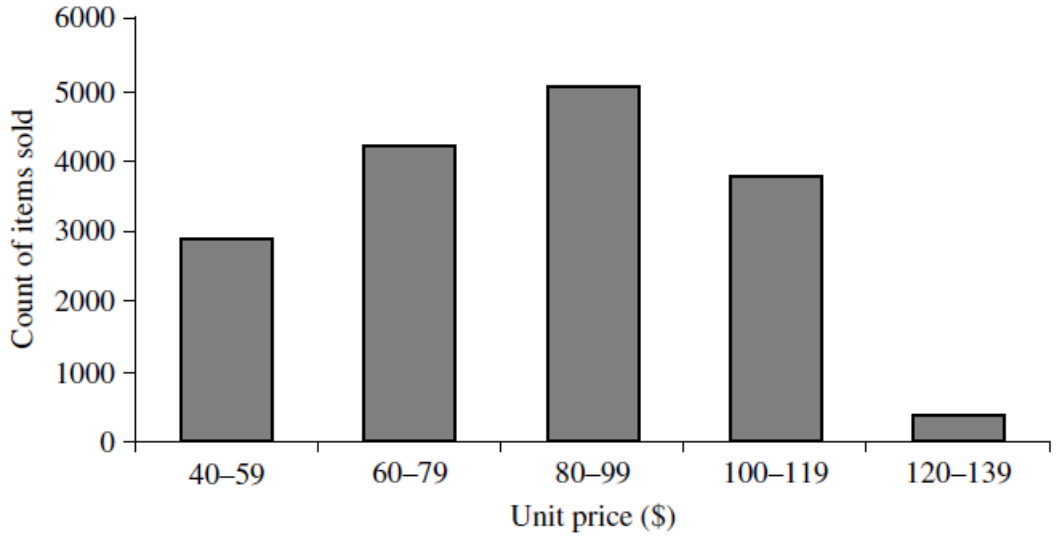
Histogram

- graphical method for summarizing the distribution of a given attribute, *X*.
- If X is nominal such as *automobile model* or *item type* (bar chart):
  - A vertical bar is drawn for each known value of *X*
  - The height of the bar indicates the frequency (i.e., count) of that *X* value.



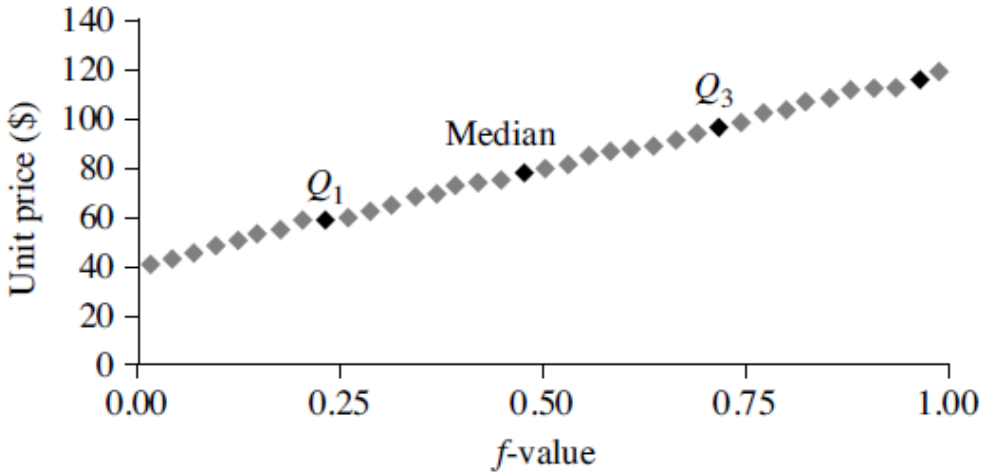
Histogram

- graphical method for summarizing the distribution of a given attribute, *X*.
- If X is numerical:
  - The range of values for X is partitioned into disjoint consecutive subranges (buckets or bins)



Quantile plot

- Displays all of the data (allowing the user to assess both the overall behavior and unusual occurrences)
- Plots quantile information
- For a sorted data
- Each x_i has a percentage f_i which indicates that approximately $100*f_i$% of the data are below or equal to the value x_i

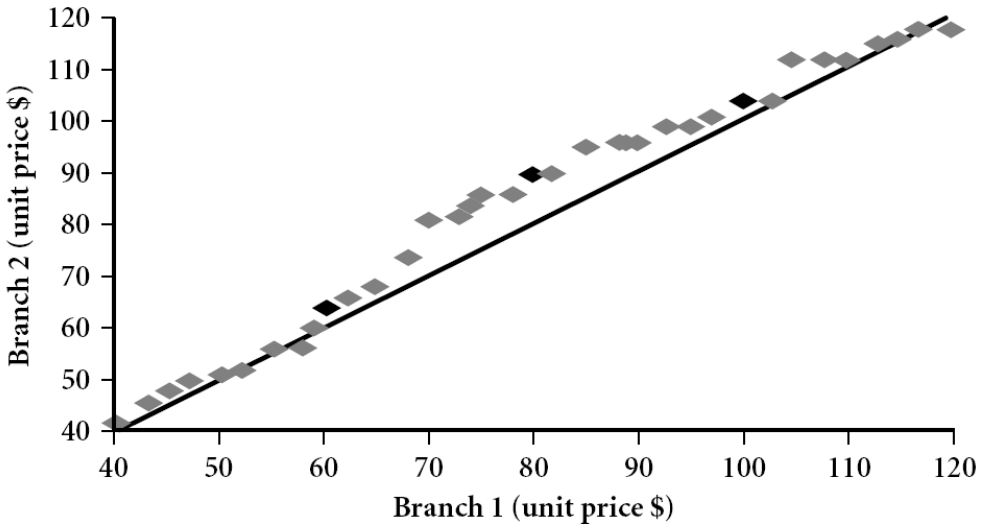


Quantile-Quantile plot (q-q plot)

- plots the quantiles of one univariate distribution against the corresponding quantiles of another
- Test if two data sets come from populations with a common distribution
- a 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line.
- The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions

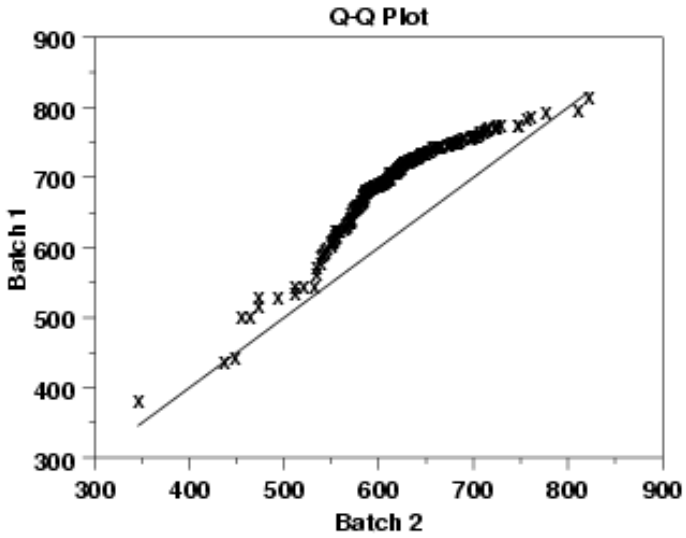
Quantile-Quantile plot (q-q plot)

- Example shows unit price of items sold at Branch 1 vs. Branch 2 for each quantile. Unit prices of items sold at Branch 1 tend to be lower than those at Branch 2.



Quantile-Quantile plot (q-q plot)

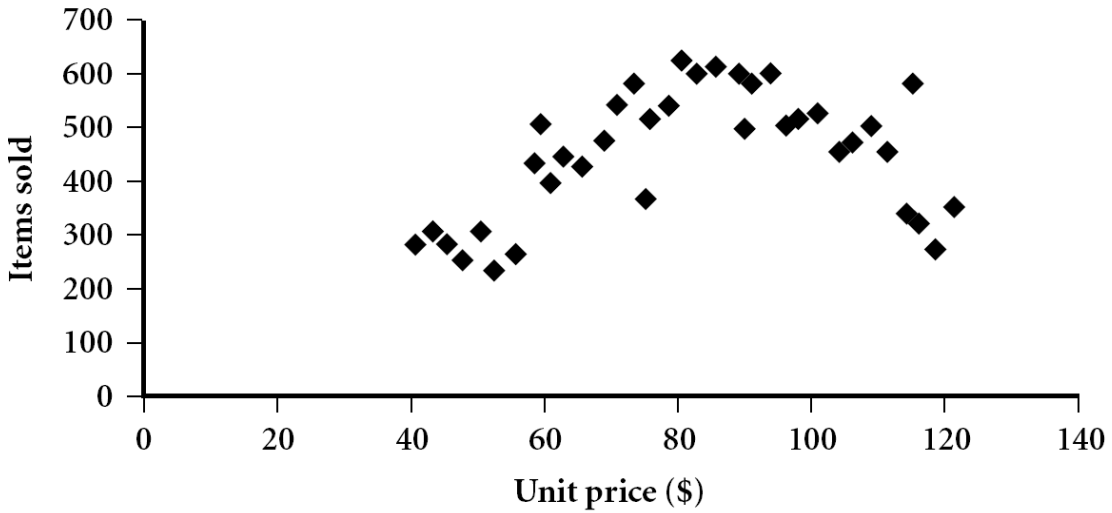
- What does this q-q plot indicates?



- These 2 batches do not appear to have come from populations with a common distribution.
- The batch 1 values are significantly higher than the corresponding batch 2 values.
- The differences are increasing from values 525 to 625. Then the values for the 2 batches get closer again.

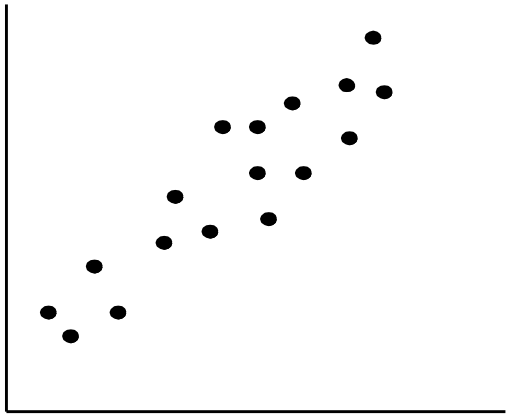
Scatter Plot and Data Correlation

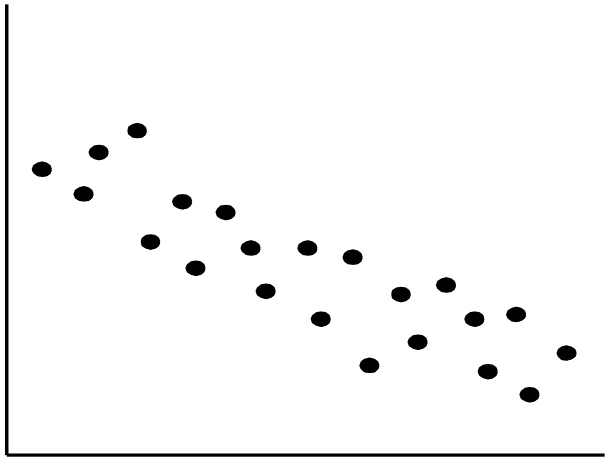
- Scatter plot: one of the most effective graphical methods for determining if there appears to be a relationship, pattern, or trend between two numeric attributes.
- Provides a first look at bivariate data to see clusters of points, outliers, etc
- Each pair of values is treated as a pair of coordinates and plotted as points in the plane

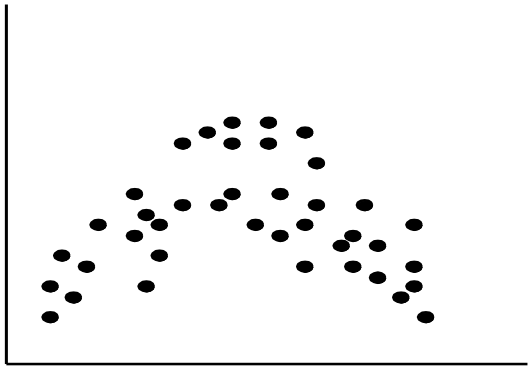


Scatter Plot and Data Correlation

- Positively and negatively correlated data



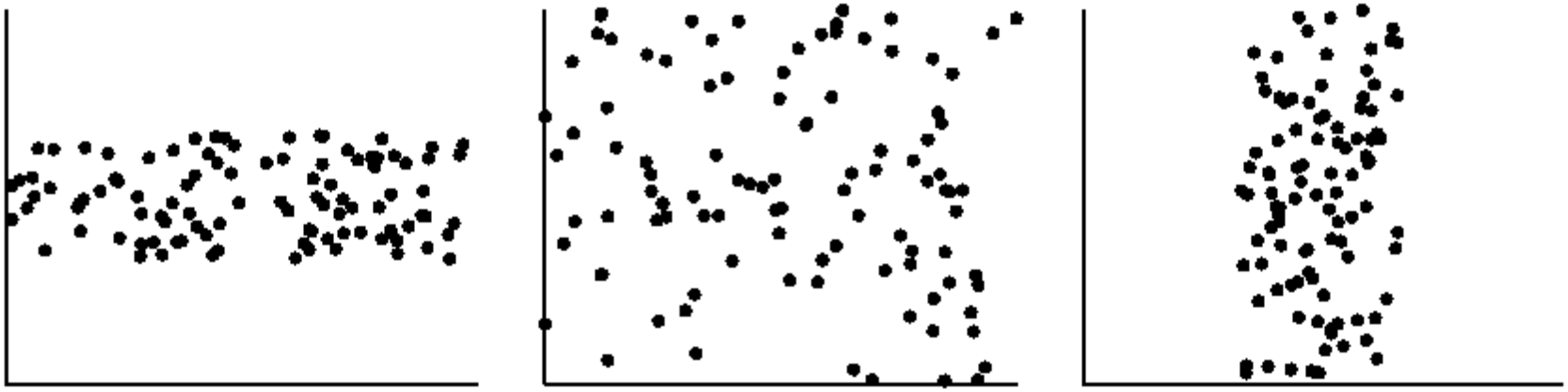




- The left half fragment is positively correlated
- The right half is negative correlated

Scatter Plot and Data Correlation

- Uncorrelated data



Conclusion

- Data attribute types: nominal, binary, ordinal, interval-scaled, ratio-scaled
- Many types of data sets, e.g., numerical, text, graph, Web, image.
- Basic data descriptions include:
  - measures of central tendency
  - measures of dispersion and
  - graphic statistical displays (e.g., quantile plots, histograms, and scatter plots)
- Basic data descriptions provide valuable insight into the overall behavior of your data.
- By helping to identify noise and outliers, they are especially useful for data cleaning.