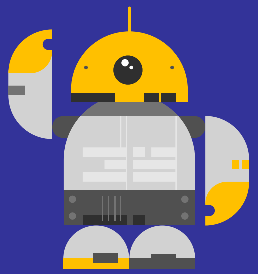


Database Systems

Lecture 4 Cont. - Getting to Know Your Data

Dr. T. Akilan
takilan@lakeheadu.ca

This Session

- Basic Statistical Analysis of Data
    - Central tendency
    - Dispersion
    - Graphical representation

Basic Statistical Analysis of Data

- Purpose:
  - Better understand the data
    - ✓ Identify typical values and properties
    - ✓ Identify outliers and noise
    - ✓ Find the correlation between objects and attributes

- Methods:
  - Central tendency
  - Dispersion
  - Graphical representation

Central Tendency of Data - Mean

- "What is the data typically like?"
  - **Mean**:
    - ✓ N samples $\{x_1, x_2, \ldots, x_N\}$
    - ✓ Assumes all samples are equally important

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

  - **Weighted mean**:
    - ✓ Weights w_i represent some form of relative importance of the samples
    - ✓ E.g. more frequent, more valuable, more significant.

$$\bar{x} = \frac{\sum_{i=1}^{N} w_i x_i}{\sum_{i=1}^{N} w_i} = \frac{w_1 x_1 + w_2 x_2 + \cdots + w_N x_N}{w_1 + w_2 + \cdots + w_N}.$$

Central Tendency of Data - Mean Cont.

- **Major problem:** Sensitivity to extreme values
  - Outliers can corrupt the mean
  - **Recall:** "how much money is in your pockets right now?"
    - ✓ Samples from 10 people: X ={\$10, \$1, \$5, \$20, \$7, \$4, \$12, \$20, \$5, \$**1,000,000**}
    - ✓ On average a person has \$100,008.4 in their pocket

- **Solution:** Trimmed mean
  - Remove a **%** of **maximum** and **minimum** values before computing mean
  - Balance:
    - ✓ Remove high enough % to eliminate outliers
    - ✓ Remove low enough % to not lose information
  - E.g.: remove 20% (low 10% and high 10%) of X.
  - Thus, remove \$1 & \$1M → mean(X_{new}) = \$10.38

Central Tendency of Data - Median

- **Median**
  - Sample value that splits the data in two sets “greater than” and “lesser than” of equal size
    - ✓ Odd number of samples : Middle value
    - ✓ Even number of samples: average of two middle values
  - E.g.:
    - ✓ {$10, $1, $5, $20, $7, $4, $12, $20, $5, $1,000,000}
    - ✓ Median is $8.50
  - **Benefit**: much more tolerant of outliers than mean!

Central Tendency of Data - Mode

- **Mode**
  - The value that occurs the most frequently in the set of objects
  - Applicable for qualitative and quantitative attributes

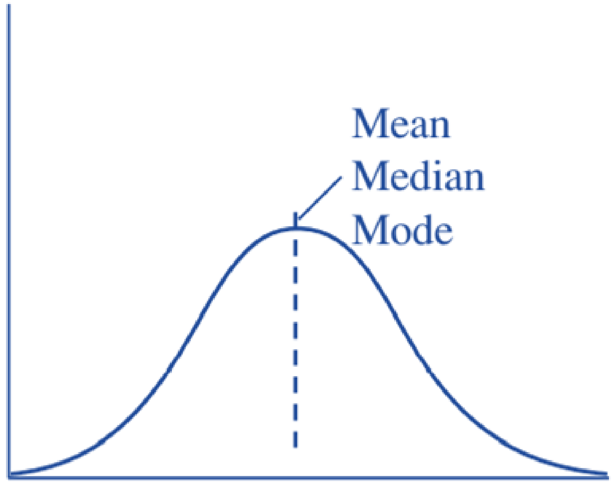
- Several values might have the same maximum frequency (i.e., be modes)
  - One mode: unimodal
  - Two modes: bimodal
  - Three modes: trimodal
  - More than one mode: multimodal
  - No mode - All values occur equally, or each data value occurs only once

Central Tendency of Data Cont.

- **Midrange**
  - The average of the maximum and minimum values in the set

$$midrange(X) = \frac{\min(X) + \max(X)}{2}$$

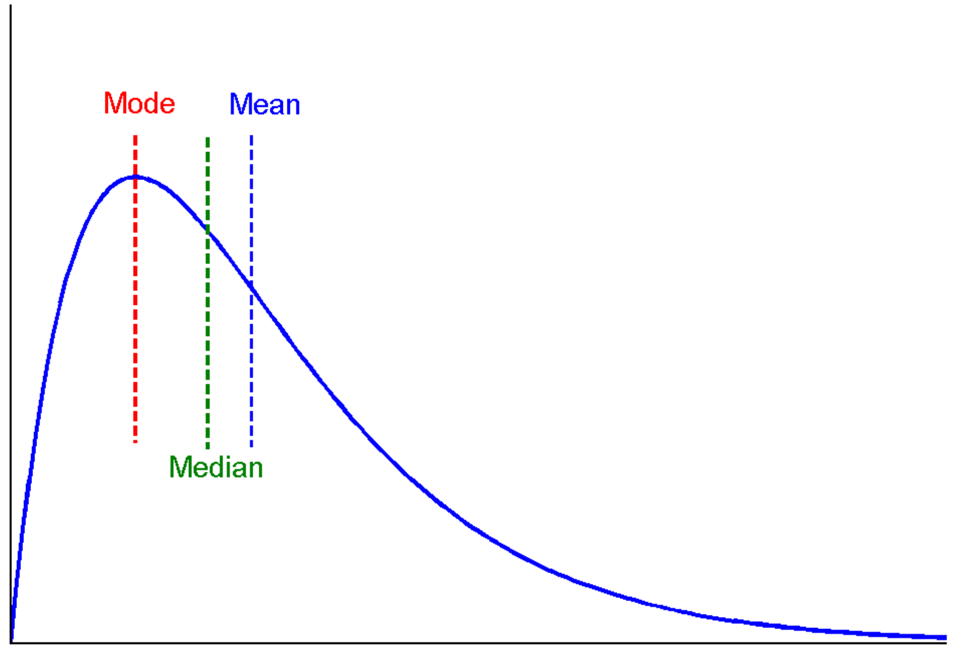
  - Example:
    - ✓ {$10, $1, $5, $20, $7, $4, $12, $20, $5, $1,000,000}
    - ✓ Midrange is $500,000.5



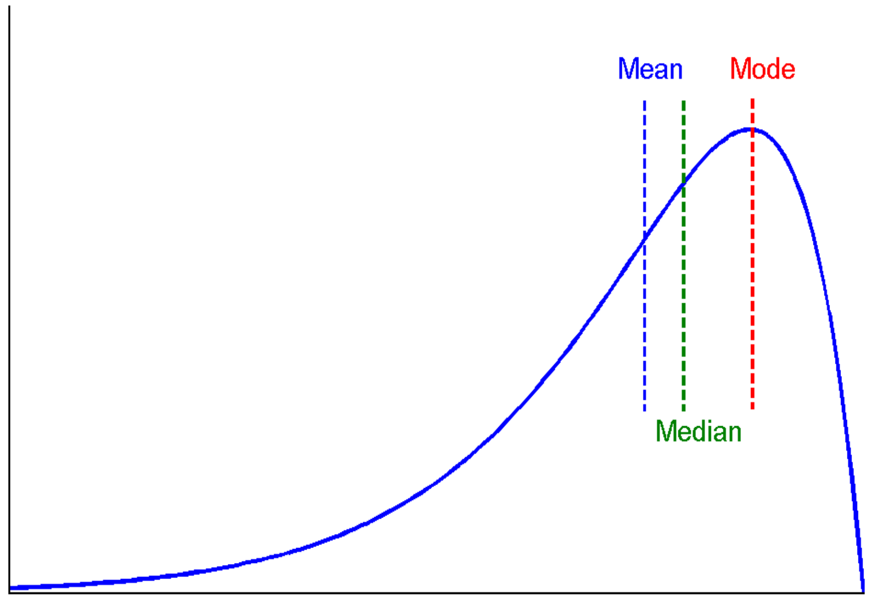
- **Note**: Given unimodal symmetric data, the mean, trimmed means, median, mode and midrange are all the same

Central Tendency of Data

- Data is rarely symmetric, it's usually skewed



Positively skewed (most of the data on the left side):
mode < median < mean



Negatively skewed (most of the data on the right side):
mode > median >mean

Dispersion of Data

- Measuring Dispersion of Data:
  - Range
  - Quartiles
  - Variance
  - Standard Deviation

Dispersion of Data – Range and Quantiles

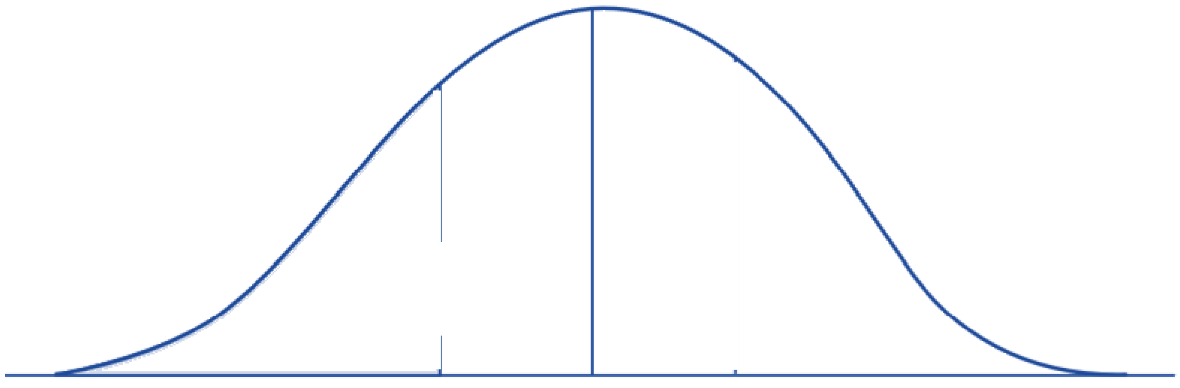
- Assume a set of samples $X = \{x_0, x_1, \ldots, x_{N-1}\}$
  - Sorted in increasing numerical order $x_i \leq x_{i+1}$
- **Range:** difference between the maximum and minimum value: `max(X) - min(X)`

$$range = x_{N-1} - x_0$$

- **Quantiles**: data points taken at regular intervals of a data distribution, dividing it into essentially equal size consecutive sets.

Dispersion of Data – Quantiles and Percentiles

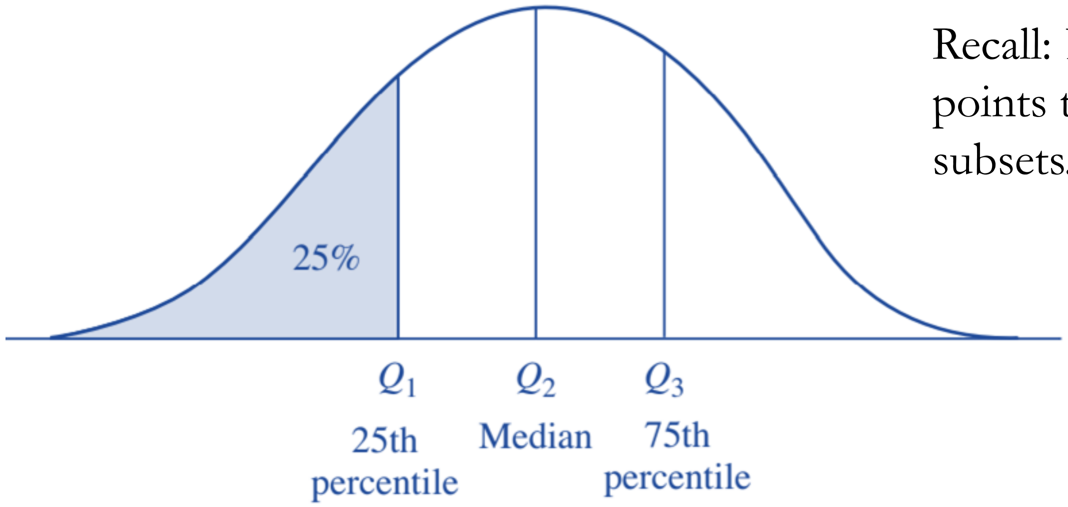
- *Q*-quantile is *Q-1* data points that break the data into *Q* equal subsets
- 2-quantile – the data point that breaks the data into two equal subsets
  - That's the median



- 100-quantiles is 99 data points that break the data into 100 equal subsets
  - Percentiles

Dispersion of Data – Quantiles and Percentiles

- 4-quantile is three points that break the data into four equal subsets
  - Quartiles (i.e., quarters): Q1, Q2 and Q3



Recall: Percentiles -100-quantiles is 99 points that break the data into 100 equal subsets.

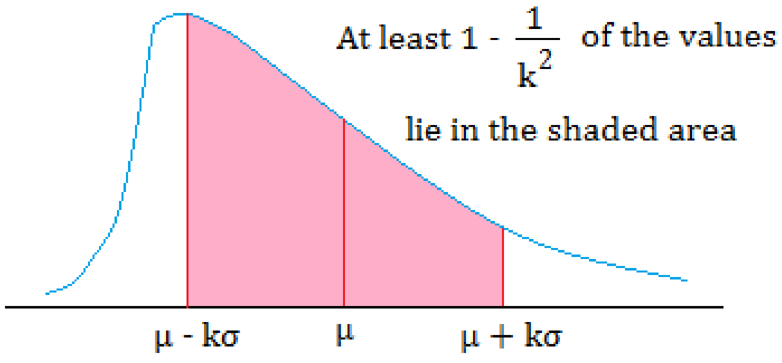
  - Note: **Interquartile range** (IQR) is the **distance between** the **first** and **third quartiles** (i.e., middle half of the data)
    - ✓ `IQR = Q3 - Q1`

Dispersion of Data – Variance and Standard deviation

- **Variance**
  - The variance of N observations, $x_1, x_2, \ldots, x_N$, for a numeric attribute X:

$$\sigma^2 = \frac{\sum_{i=0}^{N-1}\left(x_i - \bar{x}\right)^2}{N} = \frac{\sum_{i=0}^{N-1} x_i^2}{N} - \bar{x}^2$$

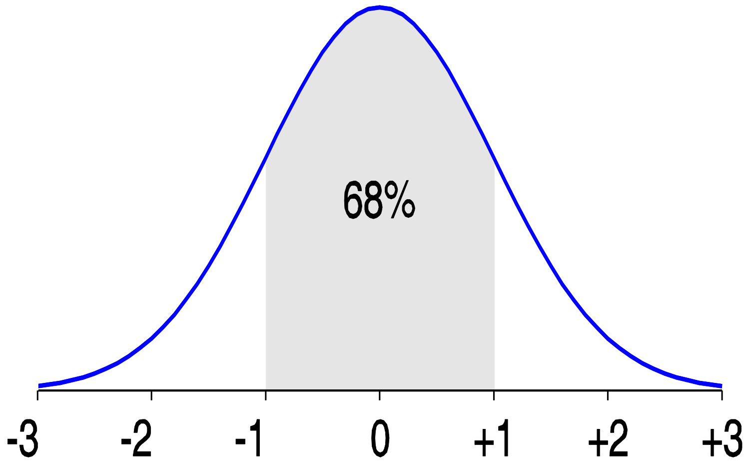
- **Standard deviation** (std) $\sigma = \sqrt{\sigma^2}$
  - Square root of variance
  - Chebyshev's theorem: at least $\left(1 - \frac{1}{k^2}\right) \times 100\%$ of the data points are within k standard deviations from μ.



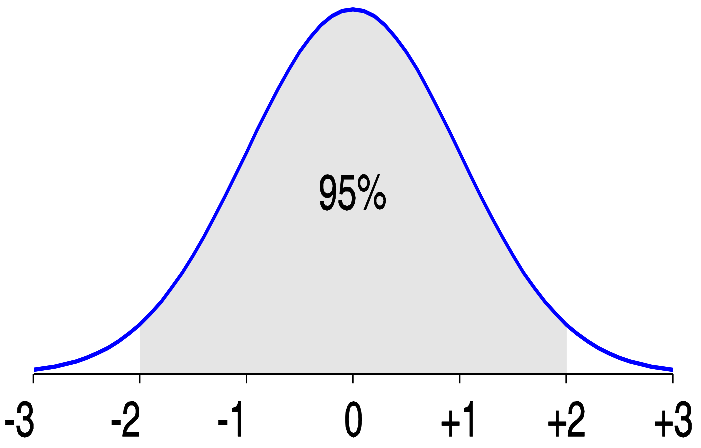
Dispersion of Data - Standard deviation Cont.

- In a normal distribution curve

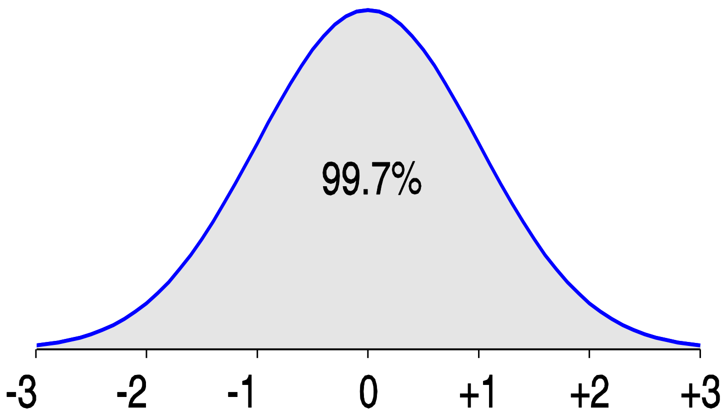
68% of data is within ±1 std of mean



95% of data is within ±2 std of mean

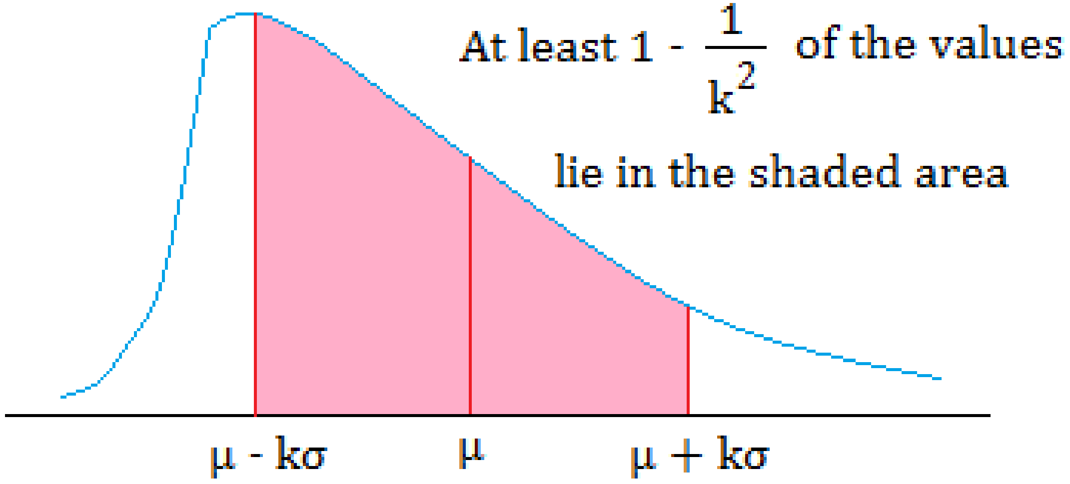


99.7% of data is within ±3 std of mean



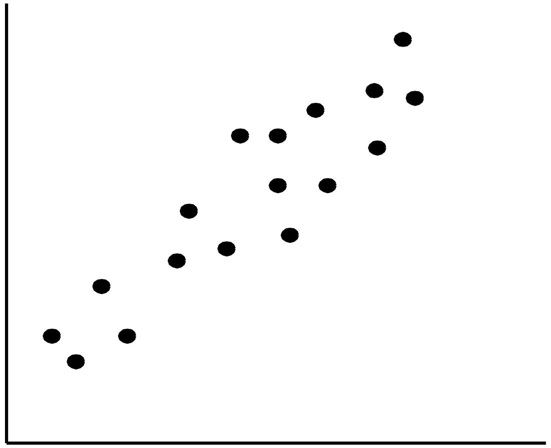
Dispersion of Data – Variance and Standard deviation

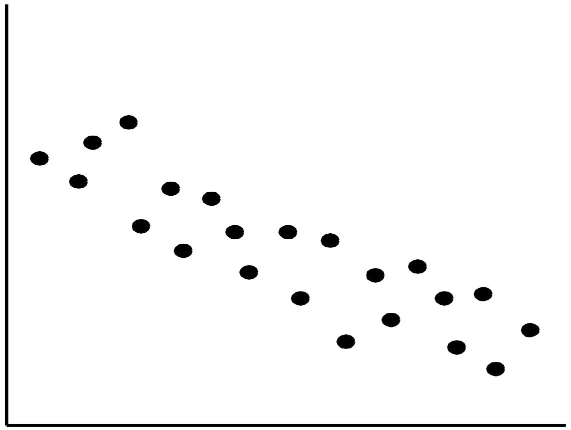
- For any distribution
  - Chebyshev's theorem: at least $\left(1-\frac{1}{k^2}\right)\times 100\%$ of the data points are within k standard deviations from μ.



 - Correlation Analysis of Data

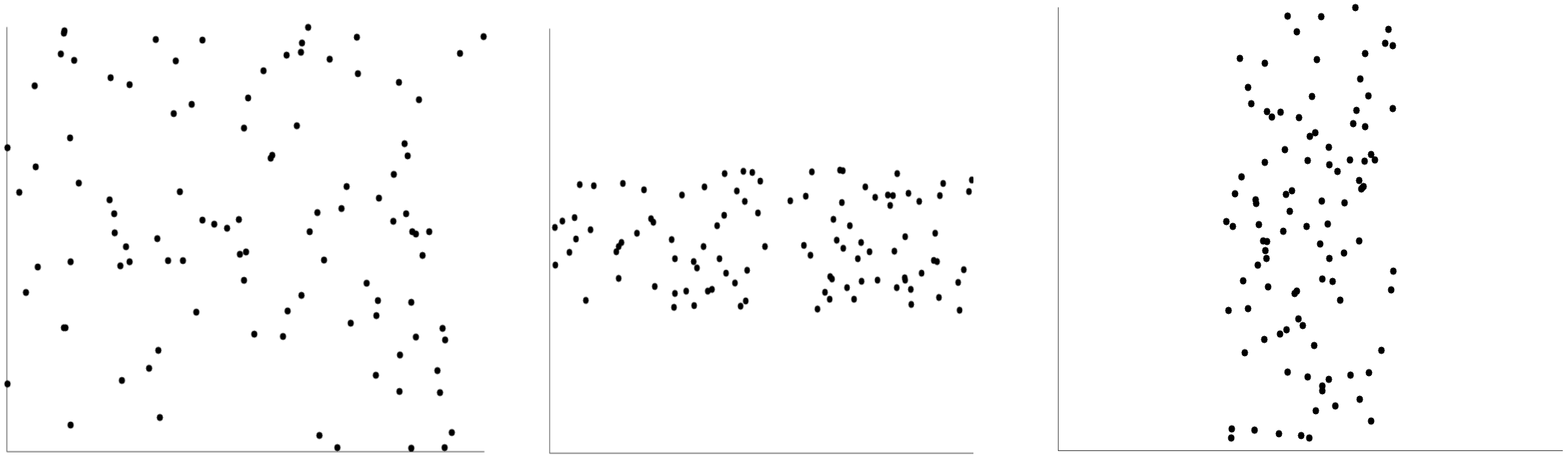
- Discover attributes whose values are related
- High correlation: both attributes values are in sync
  - Nominal, categorical, binary: attributes have values that co-occur together
  - Numerical: attribute values increase or decrease at the same time (same or opposite of each other)
    - ✓ Positive correlation: both values increase together
    - ✓ Negative correlation: one value increases when the other decreases





Correlation Analysis of Data Cont.

- Uncorrelated data
  - Left: the values are unrelated
  - Centre, right: one attribute increases independently of the other



Redundancy and Correlation Analysis

- **Challenge**: Redundancy is another important issue in data integration.
  - An attribute may be redundant if it can be "derived" from another attribute or set of attributes.

- **Solution**: Some redundancies can be detected by correlation analysis.
  - Given two attributes, such analysis can measure **how strongly one attribute implies the other**, based on the available data.

  - Nominal attributes - χ 2 (chi-square) test.
  - Numeric attributes - correlation coefficient and covariance
    - ✓ express how one attribute's values vary from those of another

Nominal Data Correlation Analysis

- χ^2 correlation test (Chi-Square)
- Two nominal data tuples attributes with **discrete sets of values**

- Let A has c distinct values, namely $a_1, a_2, \ldots, a_c$.
- B has r distinct values, namely $b_1, b_2, \ldots, b_r$.
- The data tuples described by A and B can be shown as a **contingency table**:
  - c number of columns and r number of rows.
  - Let $(\mathbf{A_i}, \mathbf{B_j})$ denote the joint event that attribute A takes on value a_i and attribute B takes on value b_j, that is, where $(A = a_i, B = b_j)$.

Nominal Data Correlation Analysis

- Let A has *c* distinct values, namely $a_1, a_2, ..., a_c$.
- B has *r* distinct values, namely $b_1, b_2, ... , b_r$.
- The data tuples described by A and B can be shown as a **contingency table**:
  - *c* number of columns and *r* number of rows.
  - Let $(\mathbf{A_i}, \mathbf{B_j})$ denote the joint event that attribute A takes on value $\boldsymbol{a_i}$ and attribute B takes on value $\boldsymbol{b_j}$, that is, where $(A = \boldsymbol{a_i}, B = \boldsymbol{b_j})$.
  - Every possible (A_i, B_j) joint event has its own cell (or slot) in the table

Data File

Student ID	Educational Level	Instructional Preference
1	Undergraduate	Online
2	Undergraduate	Face to Face
3	Undergraduate	Face to Face
4	Graduate	Online

A **B**

Contingency Table

	A_1 =UG	A_2 =Grad
B_1= F2F		
B_2= Online		

(A_i, B_j)

Database Systems

Lecture 5 Cont. - Getting to Know Your Data

Dr. T. Akilan
takilan@lakeheadu.ca

This Session

- Basic Statistical Analysis of Data
  - Central tendency
  - Dispersion
  - Graphical representation
  - Correlation Analysis – Nominal Data
    - ✓ Chi-squared test
  - Correlation Analysis – Numerical Data
    - ✓ Correlation coefficient
    - ✓ Covariance
- More on Graphical Representation of Data

Recap: Correlation Analysis – Nominal Data

- **Chi-squared test –** It tests:
  - If two nominal attributes are correlated
  - If there is a statistically substantial relationship between categorical variables

- Four important elements:
  - A null hypothesis (H_o) - the attributes are not correlated (i.e., independent)
  - Chi-square values (χ^2)
  - Degree of freedom (DF)
  - Critical value (the significance level) wrt DF to reject the null hypothesis

Recap: Chi-square Test

- **Chi-squared value**

observed frequency

expected frequency

$$\chi^2 = \sum_{i=1}^{c}\sum_{j=1}^{r}\frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

$$e_{ij} = \frac{count(A = a_i) \times count(B = b_j)}{n}$$

Total number of samples in the dataset

Contingency Table

	A_1	A_2	…	A_c	$Count(B = b_i)$
B_1			…		
B_2			…		
⋮	⋮	⋮		⋮	⋮
B_r			…		
$Count(A = a_j)$			…		n

(A_i, B_j)

observed frequency = (A_i, B_j)

Total number of samples in the dataset

- **Degree of Freedom**

DF = (r − 1) × (c − 1)

r: total number of descriptive labels in attribute B, and

c: total number of descriptive labels in attribute A

Recap: Chi-squared Test – Example

- Data: Survey data of 1500 people on his or her preferred type of reading material, genre was fiction or nonfiction.
- Task: Are **gender** and **preferred reading type** correlated?
- **Step 1:** H_o (null hypothesis)
  - Gender and preferred reading type are independent.

Person ID	Gender	Genre
1		
2		
⋮	⋮	⋮
1499		
1500		

Recap: Chi-squared Test – Example Cont.

- **Step 2:** Summarize the data in the **contingency table** containing the observed frequency (count) of each possible joint event.

	male	female	Total
fiction	250	200	450
non_fiction	50	1000	1050
Total	300	1200	1500

- **Step 3:** Compute the expected frequencies based on the data distribution for both attributes.
- E.g., $E(male, fiction)$:

$$e_{11} = \frac{count(male) \times count(fiction)}{n} = \frac{300 \times 450}{1500} = 90.$$

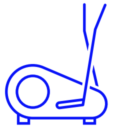
	male	female	Total
fiction	250 (90)	200 (360)	450
non_fiction	50 (210)	1000 (840)	1050
Total	300	1200	1500

Recap: Chi-squared Test – Example Cont.

- **Step 4:** Compute χ^2 using

$$\chi^2 = \sum_{i=1}^{c} \sum_{j=1}^{r} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

	male	*female*	*Total*
fiction	250 (90)	200 (360)	450
non_fiction	50 (210)	1000 (840)	1050
Total	300	1200	1500

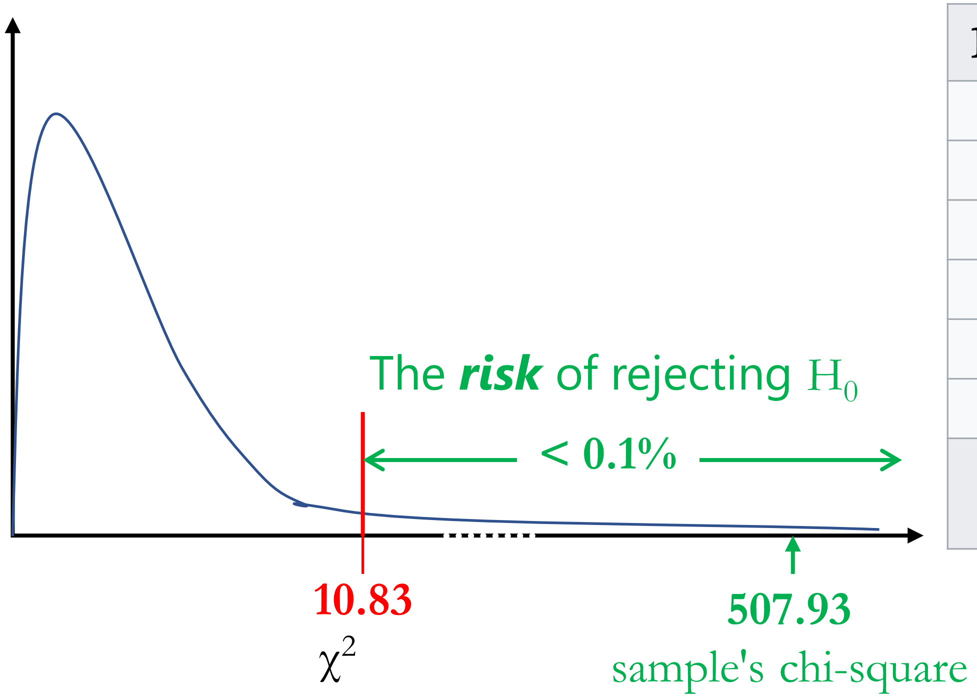


2:00

$$\chi^2 = \frac{(250-90)^2}{90} + \frac{(50-210)^2}{210} + \frac{(200-360)^2}{360} + \frac{(1000-840)^2}{840}$$
$$= 284.44 + 121.90 + 71.11 + 30.48 = 507.93.$$

Recap: Chi-squared Test – Example Cont.

- **Step 5:** Compute DF = (r − 1) × (c − 1) = (2-1) ×(2-1) = 1
- **Step 6:** Let set a significance level for 1 DF to reject the H_0. For e.g., if 0.001 significance level, a χ^2 value > 10.828 will reject the H_0.



DF	X^2 value [1]										
1	0.004	0.02	0.06	0.15	0.46	1.07	1.64	2.71	3.84	6.63	10.83
2	0.10	0.21	0.45	0.71	1.39	2.41	3.22	4.61	5.99	9.21	13.82
3	0.35	0.58	1.01	1.42	2.37	3.66	4.64	6.25	7.81	11.34	16.27
4	0.71	1.06	1.65	2.20	3.36	4.88	5.99	7.78	9.49	13.28	18.47
5	1.14	1.61	2.34	3.00	4.35	6.06	7.29	9.24	11.07	15.09	20.52
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
P	**0.95**	**0.90**	**0.80**	**0.70**	**0.50**	**0.30**	**0.20**	**0.10**	**0.05**	**0.01**	**0.001**

Conclusion: Reject the hypothesis H_0 and conclude that the two attributes are (strongly) correlated for the given group of people.

Correlation Analysis – Numerical Data

- **Correlation coefficient**
- Aka Pearson's product-moment coefficient

$$r_{A,B} = \frac{\sum_{i=1}^{n}(a_i - \bar{A})(b_i - \bar{B})}{n\sigma_A\sigma_B} = \frac{\overbrace{\sum_{i=1}^{n}(a_i b_i)}^{\text{sum of the AB cross-product}} - n\bar{A}\bar{B}}{n\sigma_A\sigma_B}$$

number of data tuples (samples)

value of A in tuple i

value of B in tuple i

mean of A

mean of B

correlation between the two attributes, A and B

standard deviations of A

standard deviations of B

Karl Pearson

Image: wikipedia.org

Correlation Analysis – Numerical Data Cont.

- **Correlation coefficient:**

$$r_{A,B} = \frac{\sum_{i=1}^{n}(a_i - \bar{A})(b_i - \bar{B})}{n\sigma_A\sigma_B} = \frac{\sum_{i=1}^{n}(a_i b_i) - n\bar{A}\bar{B}}{n\sigma_A\sigma_B}$$

- **Properties:**
  - $r_{A,B} \in [-1,1]$
    - ✓ $r_{A,B} < 0$: A and B are negatively correlated
    - ✓ $r_{A,B} > 0$: A and B are positively correlated
    - ✓ $r_{A,B} = 0$: A and B are not correlated (independent)

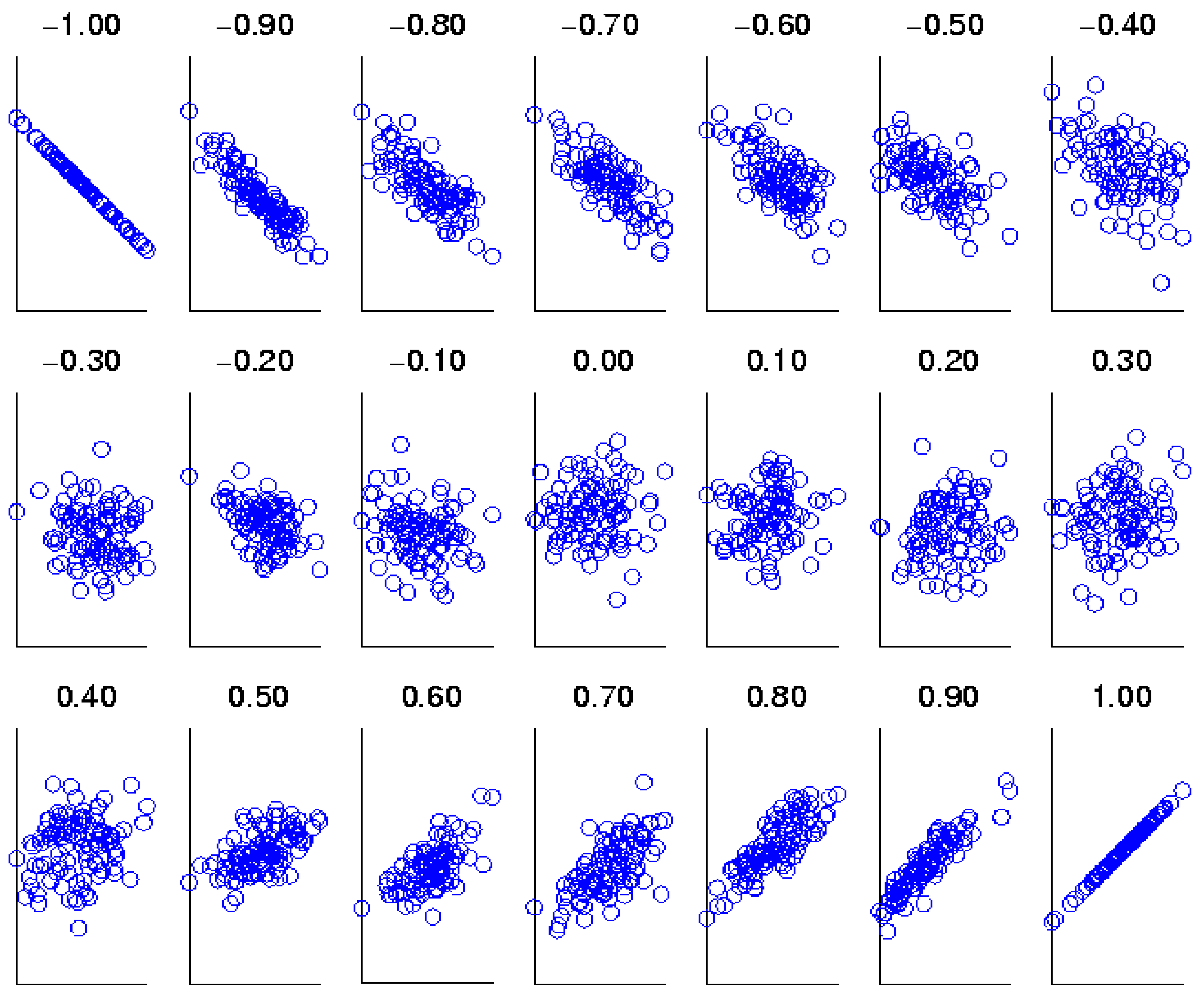
Redundancy alert: A (or B) may be removed as a redundancy

- **Decide wisely:**
  - correlation does not imply causality.
  - if A and B are correlated, this does not necessarily imply that A causes B or that B causes A.
  - E.g., demographic database: attributes representing the number of hospitals and the number of car thefts in a region can be correlated.
  - This does not mean that one causes the other. But both are causally linked to a third attribute, population.

Correlation Analysis – Numerical Data Cont.

- Recall that we can visualize the correlations between attributes using Scatter plots.

- Figure shows from negative correlation to positive correlation between two attributes.



Correlation Analysis – Numerical Data Cont.

- **Covariance - Cov()**

- Consider two numeric attributes A and B, and a set of *n* observation: $\{(a_1,b_1), \ldots, (a_n,b_n)\}$.

value of A in tuple i

mean of A

value of B in tuple i

$$Cov(A,B) = \frac{\sum_{i=1}^{n}(a_i - \bar{A})(b_i - \bar{B})}{n}$$

mean of B

number of data tuples (samples)

- Correlation coefficient ($r_{A,B}$) vs covariance *Cov*(A,B)

$$r_{A,B} = \frac{\sum_{i=1}^{n}(a_i - \bar{A})(b_i - \bar{B})}{n\sigma_A\sigma_B}$$

$$r_{A,B} = \frac{Cov(A,B)}{\sigma_A\sigma_B}$$

Correlation Analysis – Numerical Data Cont.

- Simplification of covariance computation → $Cov(A,B) = E(A \cdot B) - \bar{A}\bar{B}$

- **Properties:**
  - $cov(A,B) > 0$: if A and B change together
  - $cov(A,B) < 0$: if A and B change opposite of each other
  - If A and B are independent then $cov(A,B) = 0$

- **Note:**
  - random variables (attributes) may have cov. of 0, but they are not independent!
  - multivariate normal distribution holds cov. of 0 for independent data pairs.
  - So, we assume the data follow multivariate normal distributions.

Numerical Data Correlation Analysis – Covariance: Working Example

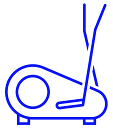
- Decide whether the attributes, number of rooms and house prices are independent using covariance analysis on the dataset.

Time Stamp	number of rooms	house prices (in M)
t1	6	20
t2	5	10
t3	4	14
t4	3	5
t5	2	5

$$E(\#\ of\ \ rooms) = \frac{6+5+4+3+2}{5} = \frac{20}{5} = \$4$$

$$E(house\ price) = \frac{20+10+14+5+5}{5} = \frac{54}{5} = \$10.80.$$

$$Cov(A,B) = E(A \cdot B) - \bar{A}\bar{B}$$



$$Cov(\#\ of\ \ rooms,\ house\ price) = \frac{6\times20+5\times10+4\times14+3\times5+2\times5}{5} - 4\times10.80$$

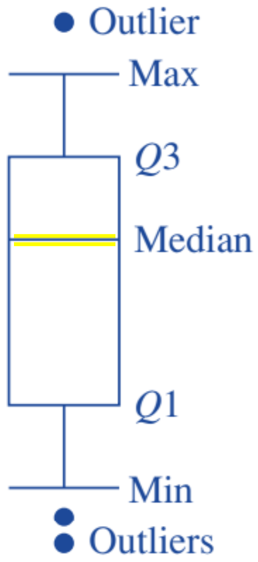
$$= 50.2 - 43.2 = 7.$$

Graphical Representation of Data

- Boxplot
  - Graphic display of five-number summary
- Histogram
  - x-axis are values, y-axis are occurrence frequencies of values
- Quantile plot
  - x-axis are %, y-axis are attribute values
- Scatter plot
  - Cartesian graph; x and y values are attributes and each data sample is plotted as a point.

Graphical Representation of Data – Boxplot

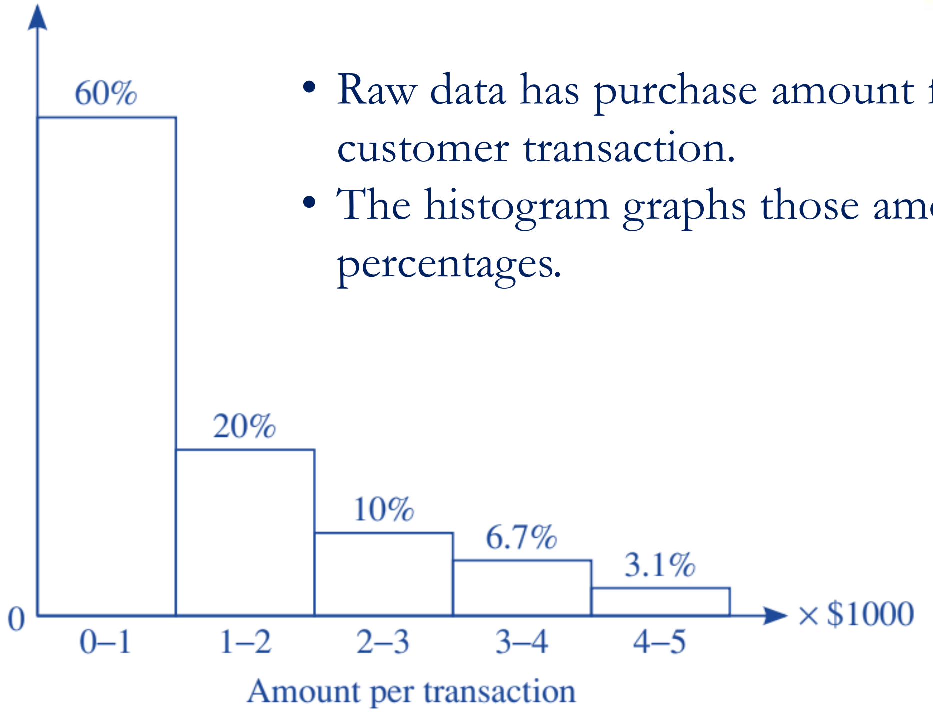
- Data is represented with a box with **five-number summary** of the distribution: Minimum, Q1, Median, Q3, Maximum



- **Properties:**
  - **Quartiles**: The ends of the box are at the first (Q1) and third quartiles (Q1)
  - Height of the box is interquartile range (IQR), i.e., the distance between the lower and upper quartiles (middle half of the data)
  - 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

Graphical Representation of Data – Histogram

- Histogram (bar chart)
  - Divide the data into discrete, disjoint subsets
    - ✓ "Buckets" or "bins"
- Plot the occurrence frequency of each bin. E.g., 60% of the transaction amounts are between $0.00 and $1000.



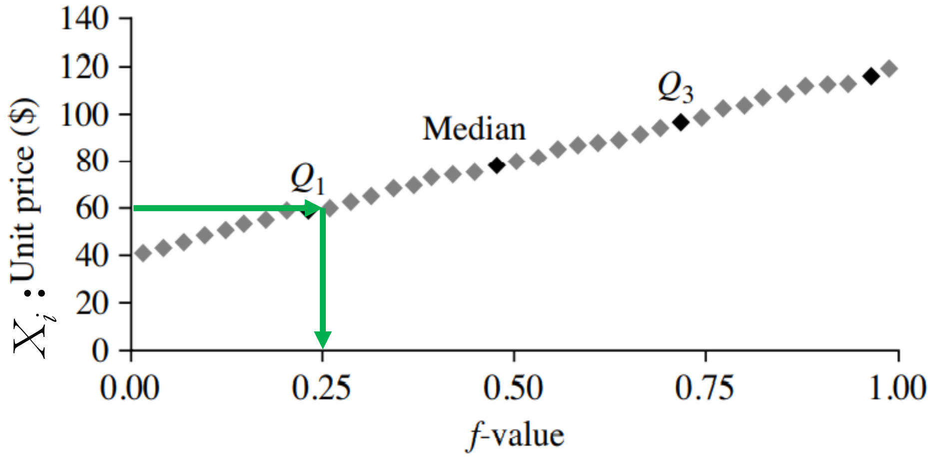
- Raw data has purchase amount for every customer transaction.
- The histogram graphs those amounts as percentages.

- We can use the histogram as a nonparametric statistical model to **capture outliers**.
  - For e.g., a transaction amount of $7500 can be regarded as an outlier.
  - Only $1 - (60\% + 20\% + 10\% + 6.7\% + 3.1\%) = 0.2\%$ of transactions have an amount > $5000.
  - A transaction amount of $385 can be treated as normal because it falls into the bin holding 60% of the transactions.

Graphical Representation of Data - Quantile Plot

Unit price ($)	Count of items sold
40	275
43	300
47	250
—	—
74	360
75	515
78	540
—	—
115	320
117	270
120	350

- Break the data into quantiles, and measure an attribute for each quantile
- Let X be some ordinal or numeric attribute, where x_i be the data sorted in increasing order, for $i = 1$ to N.
  - x_1 - the smallest observation
  - x_N - the largest for.
  - **Quantile plot** - each observation, x_i, is paired with a percentage, f_i, which indicates that approximately $f_i \times 100\%$ of the data are below the value, xi.



- **Question**: What is percentage of items sold under a unit price of $60?

$$f_i = \frac{i - 0.5}{N}$$

increases in equal steps of $1/N$,
ranging from $1/2N$ (which is slightly above 0) to $1 - 1/2N$ (which is slightly below 1)