Stat review and Data Formats

CSCI 347

- We first look at basic properties of data modeled as a data matrix.

Common Data Formats

- Data can be represented by a *data matrix D*

- $$D = \begin{array}{c|cccc} & X_1 & X_2 & X_3 & X_4 \\ \hline x_1 & 0.2 & 23 & A & 5.7 \\ x_2 & 0.4 & 1 & B & 5.4 \\ x_3 & 1.8 & 0.5 & D & 5.2 \\ x_4 & 5.6 & 50 & C & 5.1 \\ x_5 & -0.5 & 34 & V & 5.3 \\ x_6 & 0.4 & 19 & A & 5.4 \\ x_7 & 1.1 & 11 & B & 5.5 \end{array}$$

Common Data Formats

- Data can be represented by a *data matrix D*
  - Columns represents properties of interest/attributes of data

$$D = \begin{array}{c|cccc} & X_1 & X_2 & X_3 & X_4 \\ x_1 & 0.2 & 23 & A & 5.7 \\ x_2 & 0.4 & 1 & B & 5.4 \\ x_3 & 1.8 & 0.5 & D & 5.2 \\ x_4 & 5.6 & 50 & C & 5.1 \\ x_5 & -0.5 & 34 & V & 5.3 \\ x_6 & 0.4 & 19 & A & 5.4 \\ x_7 & 1.1 & 11 & B & 5.5 \end{array}$$

Common Data Formats

- Data can be represented by a *data matrix D*
  - Columns represents properties/attributes of data

The rows represent entities and their observed values for each attribute

$$D = \begin{array}{c|cccc} & X_1 & X_2 & X_3 & X_4 \\ \hline x_1 & 0.2 & 23 & A & 5.7 \\ x_2 & 0.4 & 1 & B & 5.4 \\ x_3 & 1.8 & 0.5 & D & 5.2 \\ x_4 & 5.6 & 50 & C & 5.1 \\ x_5 & -0.5 & 34 & V & 5.3 \\ x_6 & 0.4 & 19 & A & 5.4 \\ x_7 & 1.1 & 11 & B & 5.5 \end{array}$$

Common Data Formats

› Data can be represented by a ***data matrix D***

  – Columns represents properties/attributes of data

› If there are n rows and d columns, we call that matrix a $n \times d$ matrix.

$$
D = \begin{array}{ccccc}
 & \boldsymbol{X_1} & \boldsymbol{X_2} & \boldsymbol{X_3} & \boldsymbol{X_4} \\
\boldsymbol{x_1} & \mathbf{0.2} & \mathbf{23} & \boldsymbol{A} & \mathbf{5.7} \\
\boldsymbol{x_2} & \mathbf{0.4} & \mathbf{1} & \boldsymbol{B} & \mathbf{5.4} \\
\boldsymbol{x_3} & \mathbf{1.8} & \mathbf{0.5} & \boldsymbol{D} & \mathbf{5.2} \\
\boldsymbol{x_4} & \mathbf{5.6} & \mathbf{50} & \boldsymbol{C} & \mathbf{5.1} \\
\boldsymbol{x_5} & \mathbf{-0.5} & \mathbf{34} & \boldsymbol{V} & \mathbf{5.3} \\
\boldsymbol{x_6} & \mathbf{0.4} & \mathbf{19} & \boldsymbol{A} & \mathbf{5.4} \\
\boldsymbol{x_7} & \mathbf{1.1} & \mathbf{11} & \boldsymbol{B} & \mathbf{5.5}
\end{array}
$$

Common Data Formats

- If there are n rows and d columns, we call that matrix a $n \times d$ matrix.
- x_i denotes the i-th row: $x_i = (x_{i1}, x_{i2}, \dots, x_{id})$
- X_i denotes the j-th column: $X_j = (x_{1j}, x_{2j}, \dots, x_{nj})$
- Depending on the application domain, rows may be referred to as entities, samples, instances, examples, records, transactions, objects, points, feature-vectors, tuples.

$$D = \begin{array}{c|cccc} & X_1 & X_2 & X_3 & X_4 \\ \hline x_1 & 0.2 & 23 & A & 5.7 \\ x_2 & 0.4 & 1 & B & 5.4 \\ x_3 & 1.8 & 0.5 & D & 5.2 \\ x_4 & 5.6 & 50 & C & 5.1 \\ x_5 & -0.5 & 34 & V & 5.3 \\ x_6 & 0.4 & 19 & A & 5.4 \\ x_7 & 1.1 & 11 & B & 5.5 \end{array}$$

Example

$$D = \begin{array}{lcccc} & temperature & type & weight & length \\ specimen\ 1 & 1 & A & 62.3 & 21 \\ specimen\ 2 & 23 & C & 45.9 & 0.2 \\ specimen\ 3 & 45 & B & 78.2 & 5 \\ specimen\ 4 & 15 & B & 15.3 & 30 \\ specimen\ 5 & 21 & A & 18.23 & 64 \\ specimen\ 6 & -2 & F & 19.54 & 111 \\ specimen\ 7 & 2.6 & A & 56.23 & 23 \end{array}$$

Real example from UCI Machine learning repository

- Link: https://archive.ics.uci.edu/datasets
- https://archive.ics.uci.edu/dataset/502/online+retail+ii
- Online Retail II data set
  - This Online Retail II data set contains all the transactions occurring for a UK-based and registered, non-store online retail between 01/12/2009 and 09/12/2011. The company mainly sells unique all-occasion gift-ware. Many customers of the company are wholesalers.
  - 1067371 rows (entities), 8 columns (attributes)

InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	*CustomerID*	*Country*
536365	85123*A*	WHITE HANGING HEART T-LIGHT HOLDER	6	12/1/108 : 26	2.55	17850	*UnitedKingdom*
536365	71053	WHITE METAL LANTERN	6	12/1/108 : 26	3.39	17850	*UnitedKingdom*
536365	84406*B*	CREAM CUPID HEARTS COAT HANGER	8	12/1/108 : 26	2.75	17850	*UnitedKingdom*
536365	84029*G*	KNITTED UNION FLAG HOT WATER BOTTLE	6	12/1/108 : 26	3.39	17850	*UnitedKingdom*
536365	84029*E*	RED WOOLLY HOTTIE WHITE HEART.	6	12/1/108 : 26	3.39	17850	*UnitedKingdom*
536365	22752	SET 7 BABUSHKA NESTING BOXES	2	12/1/108 : 26	7.65	17850	*UnitedKingdom*
536365	21730	GLASS STAR FROSTED T-LIGHT HOLDER	6	12/1/108 : 26	4.25	17850	*UnitedKingdom*
536366	22633	HAND WARMER UNION JACK	6	12/1/108 : 28	1.85	17850	*UnitedKingdom*
⋮	⋮	⋮	⋮	⋮	⋮		

More data types

- Strings (DNA, proteins)
- Text
- Time-series (A **time series** is a sequence of data points collected, recorded, or observed at successive points in time, often at regular intervals.)
  - Common in finance, economics, weather data
- Images
- Videos

These types of data may need special technique for analysis.

Common Data Types

- Data is most often numerical or categorical

$$D = \begin{array}{lcccc} & temperature & type & weight & length \\ specimen\ 1 & 1 & A & 62.3 & 21 \\ specimen\ 2 & 23 & C & 45.9 & 0.2 \\ specimen\ 3 & 45 & B & 78.2 & 5 \\ specimen\ 4 & 15 & B & 15.3 & 30 \\ specimen\ 5 & 21 & A & 18.23 & 64 \\ specimen\ 6 & -2 & F & 19.54 & 111 \\ specimen\ 7 & 2.6 & A & 56.23 & 23 \end{array}$$

Numerical attributes

- For numerical attributes arithmetic operations are meaningful.
- Discrete
  - Numeric attribute that can take finite or countably infinite set of values
- Continuous
  - Numeric attributes that can take any real value

Numerical attributes

- Interval-scaled
  - Only differences of the attribute make sense.
  - Ex: Temperature
  - 20°F – 10°F = 10°F
  - 20°F is not twice as hot as 10°F

- Ratio-scaled
  - Can compute both differences as well as ratios between values.
  - Ex: Age
  - 20 – 10 = 10
  - 20 is twice 10
  - 0 means absence of the quantity

Numerical attributes

- Why do we care?
  - Choice of Analytical Methods:
    - You can calculate ratios or percentages with ratio data but not with interval -data.
      - Person A is twice as old as Person B”
      - Multiplication and division may be valid
    - You can apply logarithmic transformations to ratio data, but it makes less sense for interval data without a true zero.
      - Today is 10°C warmer than yesterday

Categorical attributes

- A categorical attribute is one that has a set-valued domain composed of a set of symbols.
- Nominal categorical attribute
  - Categories have no inherent order
    - Domain(Blood_Type) = {A, B, AB, O}
- Ordinal categorical attribute
  - Attribute values are ordered.
  - Domain(Pain_Intensity) = {Mild, Moderate, Severe}
- Why? Choice of statistical methods to use.
  - Mode is always meaningful but does not take the leverage of ordinal data.
  - Median and percentiles require inherent order—good for ordinal data.

What can we learn about numerical data?

› Let’s look at some statistical measures (recap)

What can we learn about numerical data

- Statistics: *Mean*
- *Estimated mean (sample mean) of attribute* $\mathbf{j}: \hat{\mu}_{\mathbf{j}} = \frac{\mathbf{1}}{\mathbf{n}} \cdot \sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{x_{ij}}$
- $$D = \begin{array}{c|cccc} & X_1 & X_2 & X_3 & X_4 \\ x_1 & 0.2 & 23 & A & 5.7 \\ x_2 & 0.4 & 1 & B & 5.4 \\ x_3 & 1.8 & 0.5 & D & 5.2 \\ x_4 & 5.6 & 50 & C & 5.1 \\ x_5 & -0.5 & 34 & V & 5.3 \\ x_6 & 0.4 & 19 & A & 5.4 \\ x_7 & 1.1 & 11 & B & 5.5 \end{array}$$

What can we learn about numerical data

- Statistics: *Mean*
- *Estimated mean (sample mean) of attribute* $\mathbf{j}$: $\hat{\mu}_j = \frac{1}{n} \cdot \sum_{i=1}^{n} x_{ij}$

x_{42}

$$D = \begin{array}{c|cccc} & X_1 & X_2 & X_3 & X_4 \\ x_1 & 0.2 & 23 & A & 5.7 \\ x_2 & 0.4 & 1 & B & 5.4 \\ x_3 & 1.8 & 0.5 & D & 5.2 \\ x_4 & 5.6 & 50 & C & 5.1 \\ x_5 & -0.5 & 34 & V & 5.3 \\ x_6 & 0.4 & 19 & A & 5.4 \\ x_7 & 1.1 & 11 & B & 5.5 \end{array}$$

Activity: Calculate the mean of attribute 2

- Statistics: *Mean*
- *Estimated mean (sample mean) of attribute* $\mathbf{j}$: $\hat{\mu}_{\mathbf{j}} = \frac{\mathbf{1}}{\mathbf{n}} \cdot \sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{x_{ij}}$
- $$D = \begin{array}{c|cccc} & X_1 & X_2 & X_3 & X_4 \\ x_1 & 0.2 & 23 & A & 5.7 \\ x_2 & 0.4 & 1 & B & 5.4 \\ x_3 & 1.8 & 0.5 & D & 5.2 \\ x_4 & 5.6 & 50 & C & 5.1 \\ x_5 & -0.5 & 34 & V & 5.3 \\ x_6 & 0.4 & 19 & A & 5.4 \\ x_7 & 1.1 & 11 & B & 5.5 \end{array}$$

- Recall that $\hat{\mu}_j = \frac{1}{n} \cdot \sum_{i=1}^{n} x_{ij}$
- Therefore, the mean of the attribute 2, $\hat{\mu}_2$ is:

$$\hat{\mu}_2 = \frac{1}{7} \cdot \sum_{i=0}^{7} x_{ij} = x_{12} + x_{22} + x_{32} + x_{42} + x_{52} + x_{62} + x_{72}$$
$$= \frac{1}{7} \cdot (23 + 1 + 0.5 + 50 + 34 + 19 + 11) = 19.78$$

$$D = \begin{array}{c|cccc} & \boldsymbol{X_1} & \boldsymbol{X_2} & \boldsymbol{X_3} & \boldsymbol{X_4} \\ \boldsymbol{x_1} & \mathbf{0.2} & \mathbf{23} & \boldsymbol{A} & \mathbf{5.7} \\ \boldsymbol{x_2} & \mathbf{0.4} & \mathbf{1} & \boldsymbol{B} & \mathbf{5.4} \\ \boldsymbol{x_3} & \mathbf{1.8} & \mathbf{0.5} & \boldsymbol{D} & \mathbf{5.2} \\ \boldsymbol{x_4} & \mathbf{5.6} & \mathbf{50} & \boldsymbol{C} & \mathbf{5.1} \\ \boldsymbol{x_5} & \mathbf{-0.5} & \mathbf{34} & \boldsymbol{V} & \mathbf{5.3} \\ \boldsymbol{x_6} & \mathbf{0.4} & \mathbf{19} & \boldsymbol{A} & \mathbf{5.4} \\ \boldsymbol{x_7} & \mathbf{1.1} & \mathbf{11} & \boldsymbol{B} & \mathbf{5.5} \end{array}$$

What can we learn about numerical data

- Statistics: *Variance*
- **Estimated variance of attribute j:** $\hat{\boldsymbol{\sigma}}_j^2 = \frac{\mathbf{1}}{\mathbf{n-1}} \cdot \sum_{\mathbf{i=1}}^{\mathbf{n}} \left(\boldsymbol{x}_{ij} - \hat{\boldsymbol{\mu}}_j\right)^2$
- $$D = \begin{array}{ccccc} & X_1 & X_2 & X_3 & X_4 \\ x_1 & 0.2 & 23 & A & 5.7 \\ x_2 & 0.4 & 1 & B & 5.4 \\ x_3 & 1.8 & 0.5 & D & 5.2 \\ x_4 & 5.6 & 50 & C & 5.1 \\ x_5 & -0.5 & 34 & V & 5.3 \\ x_6 & 0.4 & 19 & A & 5.4 \\ x_7 & 1.1 & 11 & B & 5.5 \end{array}$$

What can we learn about numerical data

- Statistics: *Variance*
- Estimated variance of attribute 2:

$$\hat{\sigma}_2^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_{i2} - \hat{\mu})^2$$

Thus, the estimated variance of X_2 in the following example is:

$$\hat{\sigma}_2^2 = \frac{1}{6} \cdot ((\mathbf{23} - 19.78)^2 + (\mathbf{1} - 19.78)^2 + (\mathbf{0.5} - 19.78)^2 + (\mathbf{50} - 19.78)^2 + (\mathbf{34} - 19.78) + (\mathbf{19} - 19.78)^2 + (\mathbf{11} - 19.78)^2)$$

$$= \mathbf{321.3214}$$

- $$D = \begin{array}{c|cccc} & X_1 & X_2 & X_3 & X_4 \\ x_1 & 0.2 & 23 & A & 5.7 \\ x_2 & 0.4 & 1 & B & 5.4 \\ x_3 & 1.8 & 0.5 & D & 5.2 \\ x_4 & 5.6 & 50 & C & 5.1 \\ x_5 & -0.5 & 34 & V & 5.3 \\ x_6 & 0.4 & 19 & A & 5.4 \\ x_7 & 1.1 & 11 & B & 5.5 \end{array}$$

- Note that if the above statistic formula is for estimated sample variance.
- If you are given the whole population, then the variance can be calculated using:
- $\boldsymbol{\sigma_2^2 = \frac{1}{n} \cdot \sum_{i=1}^{n}(x_{i2} - \mu)^2}$

What can we learn about numerical data

- Statistics: *Standard deviation*
- Estimated standard deviation of attribute $\mathbf{j}$: $\hat{\boldsymbol{\sigma}}_{\boldsymbol{j}} = \sqrt{\hat{\sigma}_j^2}$ (square root of the estimated variance)
- Variance is mathematically convenient; standard deviation is variance made interpretable.

What can we learn about numerical data

- Statistics: *Standard deviation*
- Estimated standard deviation of attribute j: $\hat{\sigma}_j = \sqrt{\hat{\sigma}_j^2}$ (square root of the estimated variance)
- The estimated standard deviation of X_2 in the following example is:
- $\hat{\sigma}_2 = \sqrt{\hat{\sigma}_2^2} = \sqrt{321.3214} = 17.925$
- $$D = \begin{array}{ccccc} & X_1 & X_2 & X_3 & X_4 \\ x_1 & 0.2 & 23 & A & 5.7 \\ x_2 & 0.4 & 1 & B & 5.4 \\ x_3 & 1.8 & 0.5 & D & 5.2 \\ x_4 & 5.6 & 50 & C & 5.1 \\ x_5 & -0.5 & 34 & V & 5.3 \\ x_6 & 0.4 & 19 & A & 5.4 \\ x_7 & 1.1 & 11 & B & 5.5 \end{array}$$

What can we learn about numerical data

- Statistics: *Multi-dimensional mean*
- Estimated mean of entire (numerical) sample data set.

$$\hat{\mu} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$$

- $$D = \begin{array}{c|cccc} & X_1 & X_2 & X_3 & X_4 \\ x_1 & 0.2 & 23 & A & 5.7 \\ x_2 & 0.4 & 1 & B & 5.4 \\ x_3 & 1.8 & 0.5 & D & 5.2 \\ x_4 & 5.6 & 50 & C & 5.1 \\ x_5 & -0.5 & 34 & V & 5.3 \\ x_6 & 0.4 & 19 & A & 5.4 \\ x_7 & 1.1 & 11 & B & 5.5 \end{array}$$

What can we learn about numerical data

- Statistics: *Multi-dimensional mean*
- Estimated mean of entire (numerical) data set.

$$\hat{\mu} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$$

- $$D = \begin{array}{cccc} & X_1 & X_2 & X_3 \\ x_1 & 0.2 & 23 & 5.7 \\ x_2 & 0.4 & 1 & 5.4 \\ x_3 & 1.8 & 0.5 & 5.2 \\ x_4 & 5.6 & 50 & 5.1 \\ x_5 & -0.5 & 34 & 5.3 \\ x_6 & 0.4 & 19 & 5.4 \\ x_7 & 1.1 & 11 & 5.5 \end{array}$$

$$\hat{\mu} = \frac{1}{7}\big((0.2\ 23\ 5.7) + (0.4\ 1\ 5.4\ ) + (1.8\ 0.5\ 5.2) + (5.6\ 50\ 5.1) + (-0.5\ 34\ 5.3) + (0.4\ 19\ 5.4) + (1.1\ 11\ 5.5)\big)$$
$$= (\mathbf{1.3} \quad \mathbf{19.8} \quad \mathbf{5.4})$$

What can we learn about numerical data

- Statistics: Estimated *Covariance*
- What is the estimated covariance between two attributes in a numerical data set?
- Ex: Estimated Covariance between attribute 1 and 2?

$$\hat{\sigma}_{12} = \frac{\mathbf{1}}{\boldsymbol{n}-\mathbf{1}} \cdot \sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{n}} (\boldsymbol{x}_{\boldsymbol{i}\mathbf{1}} - \widehat{\boldsymbol{\mu}}_{\mathbf{1}}) \cdot (\boldsymbol{x}_{\boldsymbol{i}\mathbf{2}} - \widehat{\boldsymbol{\mu}}_{\mathbf{2}})$$

- $$D = \begin{array}{cccc} & X_1 & X_2 & X_3 \\ x_1 & 0.2 & 23 & 5.7 \\ x_2 & 0.4 & 1 & 5.4 \\ x_3 & 1.8 & 0.5 & 5.2 \\ x_4 & 5.6 & 50 & 5.1 \\ x_5 & -0.5 & 34 & 5.3 \\ x_6 & 0.4 & 19 & 5.4 \\ x_7 & 1.1 & 11 & 5.5 \end{array}$$

What can we learn about numerical data

- Statistics: *Covariance*
- What is the estimated covariance between two attributes in a numerical data set?
- Ex: Estimated Covariance between attribute 1 and 2?

$$\hat{\sigma}_{12} = \frac{\mathbf{1}}{\boldsymbol{n-1}} \cdot \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} (\boldsymbol{x_{i1}} - \hat{\boldsymbol{\mu}}_{\mathbf{1}}) \cdot (\boldsymbol{x_{i2}} - \hat{\boldsymbol{\mu}}_{\mathbf{2}})$$

- $$D = \begin{array}{c|ccc} & X_1 & X_2 & X_3 \\ x_1 & 0.2 & 23 & 5.7 \\ x_2 & 0.4 & 1 & 5.4 \\ x_3 & 1.8 & 0.5 & 5.2 \\ x_4 & 5.6 & 50 & 5.1 \\ x_5 & -0.5 & 34 & 5.3 \\ x_6 & 0.4 & 19 & 5.4 \\ x_7 & 1.1 & 11 & 5.5 \end{array}$$

What are the possible values for the covariance?

1. Only positive values
2. Between -1 to +1
3. Between $-\infty$ to $+\infty$

What can we learn about numerical data

- Statistics: *Covariance*
- What is the estimated covariance between two attributes in a numerical data set?
- Ex: Estimated Covariance between attribute 1 and 2?

$$\hat{\sigma}_{12} = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_{i1} - \hat{\mu}_1) \cdot (x_{i2} - \hat{\mu}_2)$$

- $$D = \begin{array}{c|ccc} & X_1 & X_2 & X_3 \\ x_1 & 0.2 & 23 & 5.7 \\ x_2 & 0.4 & 1 & 5.4 \\ x_3 & 1.8 & 0.5 & 5.2 \\ x_4 & 5.6 & 50 & 5.1 \\ x_5 & -0.5 & 34 & 5.3 \\ x_6 & 0.4 & 19 & 5.4 \\ x_7 & 1.1 & 11 & 5.5 \end{array}$$

What are the possible values for the covariance?

1. Only positive values
2. Between -1 to +1
3. **Between $-\infty$ to $+\infty$**

- Positive covariance (> 0): Indicates that two variables tend to increase or decrease together.
- Negative covariance (< 0): Indicates that when one variable increases, the other tends to decrease.
  - For example, outdoor temperature and heating cost often have negative covariance.
- Zero covariance ($= 0$): Indicates no linear relationship between the variables

In class activity

- Statistics: *Covariance*
- **Ex: Covariance between attribute 1 and 2?**

$$\hat{\sigma}_{12} = \frac{\mathbf{1}}{\boldsymbol{n-1}} \cdot \sum_{\boldsymbol{i=0}}^{\boldsymbol{n}} (\boldsymbol{x_{i1}} - \boldsymbol{\hat{\mu}_1}) \cdot (\boldsymbol{x_{i2}} - \boldsymbol{\hat{\mu}_2})$$

- $$D = \begin{array}{cccc} & X_1 & X_2 & X_3 \\ x_1 & 0.2 & 23 & 5.7 \\ x_2 & 0.4 & 1 & 5.4 \\ x_3 & 1.8 & 0.5 & 5.2 \\ x_4 & 5.6 & 50 & 5.1 \\ x_5 & -0.5 & 34 & 5.3 \\ x_6 & 0.4 & 19 & 5.4 \\ x_7 & 1.1 & 11 & 5.5 \end{array}$$

First find $\hat{\mu}_1$ and $\hat{\mu}_2$:
$\hat{\mu}_1 = 1.3$, $\hat{\mu}_2 = 19.78$

› Next, we use $\hat{\mu}_1$ and $\hat{\mu}_2$ to calculate $\hat{\sigma}_{12}$

$$\hat{\sigma}_{12} =$$

$$\frac{1}{6} \cdot \big((0.2 - 1.3) \cdot (23 - 19.8) + (0.4 - 1.3) \cdot (1 - 19.8) + (1.8 - 1.3) \cdot (0.5 - 19.8) + (5.6 - 1.3) \cdot (50 - 19.8)\big)$$
$$+(-0.5 - 1.3) \cdot (34 - 19.8) + (0.4 - 1.3) \cdot (19 - 19.8) + (1.1 - 1.3) \cdot (11 - 19.8)$$

$$\hat{\sigma}_{12} = 18.4$$

What can we learn about numerical data

- Statistics: *Correlation coefficient*
- The correlation coefficient between attribute j and k is:

$$\hat{\rho}_{xj} = \frac{\hat{\sigma}_{xj}}{\hat{\sigma}_x \cdot \hat{\sigma}_j} = \frac{cov(x,y)}{std(x) \cdot std(y)}$$

- $$D = \begin{array}{cccc} & X_1 & X_2 & X_3 \\ x_1 & 0.2 & 23 & 5.7 \\ x_2 & 0.4 & 1 & 5.4 \\ x_3 & 1.8 & 0.5 & 5.2 \\ x_4 & 5.6 & 50 & 5.1 \\ x_5 & -0.5 & 34 & 5.3 \\ x_6 & 0.4 & 19 & 5.4 \\ x_7 & 1.1 & 11 & 5.5 \end{array}$$

- What is the possible range of values for correlation coefficient?
  - Only positive values
  - Between -1 to +1
  - Between $-\infty$ to $+\infty$

- What is the possible range of values for correlation coefficient?
  - Only positive values
  - **Between -1 to +1**
  - Between $-\infty$ to $+\infty$

- What does correlation coefficient of 1 mean?
  - Variables move in the same direction
  - A straight line with positive slope fits all points perfectly.
- What does correlation coefficient of -1 mean?
  - Variables move in exactly opposite directions
  - A straight line with negative slope fits all points perfectly.
    - Temperature in Celsius vs. distance from a heat source
- Zero correlation (0)
  - No linear relationship between variables.
  - Note: Could still have non-linear relationship.

› What is the correlation coefficient between attribute 1 and 2?

› $D = \begin{array}{cccc} & X_1 & X_2 & X_3 \\ x_1 & 0.2 & 23 & 5.7 \\ x_2 & 0.4 & 1 & 5.4 \\ x_3 & 1.8 & 0.5 & 5.2 \\ x_4 & 5.6 & 50 & 5.1 \\ x_5 & -0.5 & 34 & 5.3 \\ x_6 & 0.4 & 19 & 5.4 \\ x_7 & 1.1 & 11 & 5.5 \end{array}$

$$\hat{\rho}_{12} = \frac{\hat{\sigma}_{12}}{\hat{\sigma}_1 \cdot \hat{\sigma}_2} = \frac{(18.4)}{(4.14) \cdot (17.925)} = 0.247$$

Questions

› If $\hat{\rho}_{xy} = \ 0.7$ which one of the following statement(s) are true?
  - A: An increment in x will causes an increase in y
  - B: An increment in y will causes an increase in x
  - C: x and y move together
  - D: All above

› Is $\hat{\rho}_{xy} = \hat{\rho}_{yx}$?
  - True
  - False

Questions

- If $\hat{\rho}_{xy} = 0.7$ which one of the following statement(s) are true?
  - A: An increment in x will cause an increase in y
  - B: An increment in y will cause an increase in x
  - C: x and y move together
  - D: All above

- Is $\hat{\rho}_{xy} = \hat{\rho}_{yx}$?
  - True
  - False

Correlation and Causality

- Correlation DOES NOT imply causality!
- Check spurious-correlation
- Correlation doesn’t have direction, but causality has direction

Covariance matrix

› The covariance matrix Σ stores the covariance between each pair of attributes, as well as the variance for each attribute:

$$D = \begin{array}{cccc} & X_1 & X_2 & X_3 \\ x_1 & 0.2 & 23 & 5.7 \\ x_2 & 0.4 & 1 & 5.4 \\ x_3 & 1.8 & 0.5 & 5.2 \\ x_4 & 5.6 & 50 & 5.1 \\ x_5 & -0.5 & 34 & 5.3 \\ x_6 & 0.4 & 19 & 5.4 \\ x_7 & 1.1 & 11 & 5.5 \end{array}$$

$$\Sigma = \begin{pmatrix} \hat{\boldsymbol{\sigma}}_1^2 & \hat{\boldsymbol{\sigma}}_{12} & \hat{\boldsymbol{\sigma}}_{13} \\ \hat{\boldsymbol{\sigma}}_{21} & \hat{\boldsymbol{\sigma}}_2^2 & \hat{\boldsymbol{\sigma}}_{23} \\ \hat{\boldsymbol{\sigma}}_{31} & \hat{\boldsymbol{\sigma}}_{32} & \hat{\boldsymbol{\sigma}}_3^2 \end{pmatrix}$$

Covariance matrix

› The covariance matrix Σ stores the covariance between each pair of attributes, as well as the variance for each attribute:

$$D = \begin{array}{cccc} & X_1 & X_2 & X_3 \\ x_1 & 0.2 & 23 & 5.7 \\ x_2 & 0.4 & 1 & 5.4 \\ x_3 & 1.8 & 0.5 & 5.2 \\ x_4 & 5.6 & 50 & 5.1 \\ x_5 & -0.5 & 34 & 5.3 \\ x_6 & 0.4 & 19 & 5.4 \\ x_7 & 1.1 & 11 & 5.5 \end{array}$$

$$\Sigma = \begin{pmatrix} \mathbf{4.1} & \mathbf{18.4} & \mathbf{-0.26} \\ \mathbf{18.42} & \mathbf{321.32} & \mathbf{-1.09} \\ \mathbf{-0.26} & \mathbf{-1.09} & \mathbf{0.04} \end{pmatrix}$$

What can we learn about numerical data

- Statistics: *Total Variance*
- **What is the total variance in a numerical data set?.**

$$\boldsymbol{Var(D) = \hat{\sigma}_1^2 + \hat{\sigma}_2^2 + \hat{\sigma}_3^2 + \cdots + \hat{\sigma}_n^2}$$

$$D = \begin{array}{cccc} & X_1 & X_2 & X_3 \\ x_1 & 0.2 & 23 & 5.7 \\ x_2 & 0.4 & 1 & 5.4 \\ x_3 & 1.8 & 0.5 & 5.2 \\ x_4 & 5.6 & 50 & 5.1 \\ x_5 & -0.5 & 34 & 5.3 \\ x_6 & 0.4 & 19 & 5.4 \\ x_7 & 1.1 & 11 & 5.5 \end{array}$$

$$\boldsymbol{Var(D) = 4.1 + 321.32 + 0.04 = 325.44}$$

Total variance

- Provides insight about overall variability or spread of the dataset .
- Higher variance mean data points are more spread out.
- Lower variance mean the data points are closer to the mean, indicating less variability (or consistency)

Other basic statistical measures

- Mode
  - the value that appears most often in a set of data values
- Median
  - The median is the “middle value” of a dataset when all values are sorted in order.

Mean Centering

- Mean-centering shifts the data matrix mean to 0.
- $Z_i = x_i - \widehat{\mu_i}$
- For each attribute subtract the mean from the instance value.

$$D = \begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ x_7 \end{array} \begin{array}{ccc} X_1 & X_2 & X_3 \\ 0.2 & 23 & 5.7 \\ 0.4 & 1 & 5.4 \\ 1.8 & 0.5 & 5.2 \\ 5.6 & 50 & 5.1 \\ -0.5 & 34 & 5.3 \\ 0.4 & 19 & 5.4 \\ 1.1 & 11 & 5.5 \end{array}$$

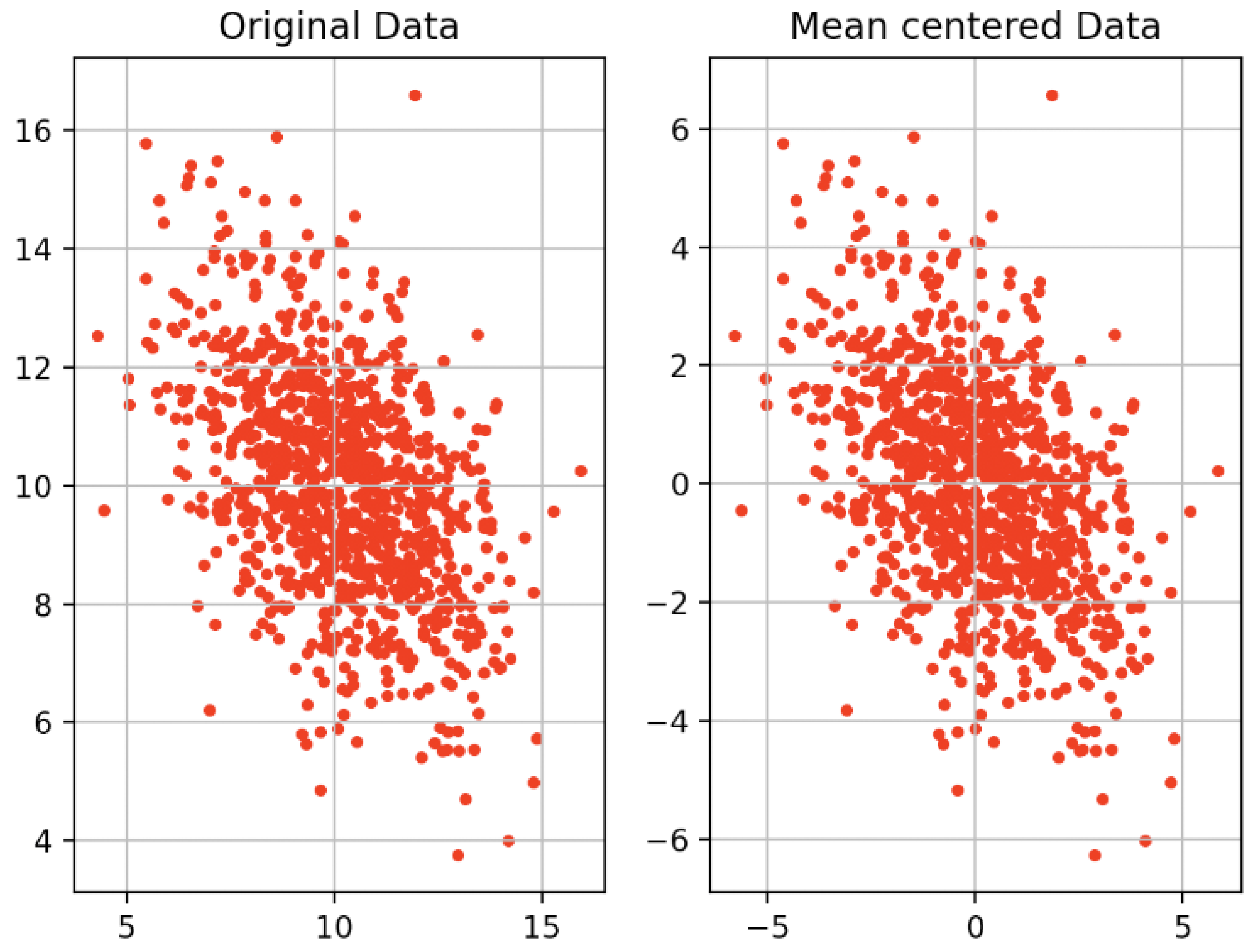
$$\boldsymbol{z_{11} = x_{11} - \widehat{\mu}_1 = 0.2 - 1.3 = -1.1}$$

Mean Centering

- Mean-centering shifts the data matrix mean to 0.
- $Z_i = x_i - \widehat{\mu_i}$
- For each attribute subtract the mean from the instance value.

$$D = \begin{array}{c|ccc} & X_1 & X_2 & X_3 \\ x_1 & 0.2 & 23 & 5.7 \\ x_2 & 0.4 & 1 & 5.4 \\ x_3 & 1.8 & 0.5 & 5.2 \\ x_4 & 5.6 & 50 & 5.1 \\ x_5 & -0.5 & 34 & 5.3 \\ x_6 & 0.4 & 19 & 5.4 \\ x_7 & 1.1 & 11 & 5.5 \end{array} \longrightarrow Z = \begin{array}{c|ccc} & X_1 & X_2 & X_3 \\ z_1 & -1.09 & 3.21 & 0.33 \\ z_2 & -0.89 & -18.79 & 0.03 \\ z_3 & 0.51 & -19.29 & -0.17 \\ z_4 & 4.31 & 30.21 & -0.27 \\ z_5 & -1.79 & 14.21 & -0.07 \\ z_6 & -0.89 & -0.79 & 0.03 \\ z_7 & -0.19 & -8.79 & 0.13 \end{array}$$

Mean Centering



Why mean centering?

- Removing bias of the mean
  - Ensures the analysis focuses on the variability and relationship within the data rather than the absolute values of the variable
  - For many statistical techniques magnitude of the data does not matter, but their relative relationships do.
- Numerical stability
  - By mean centering values of the data typically becomes smaller and balanced, which improves the stability of mathematical operations.
- Improves visualization of data. (data is centered around the origin)
  - Refer previous figure.
- Does not change the variance