ECEN 758 Data Mining and Analysis: Lecture 2, Data and Attributes I

Joshua Peeples, Ph.D.
Assistant Professor
Department of Electrical and Computer Engineering

Announcements

- Assignment #1 will be released next Wednesday (08/28)
  - Due Friday, 09/06
- Updated exam dates
  - 10/14 (Exam 1) and 11/25 (Exam 2)
- Tentative assignment dates

Assignment #	Released	Due
1	08/28	09/06
2	09/18	09/27
3	10/02	10/11
4	11/06	11/15

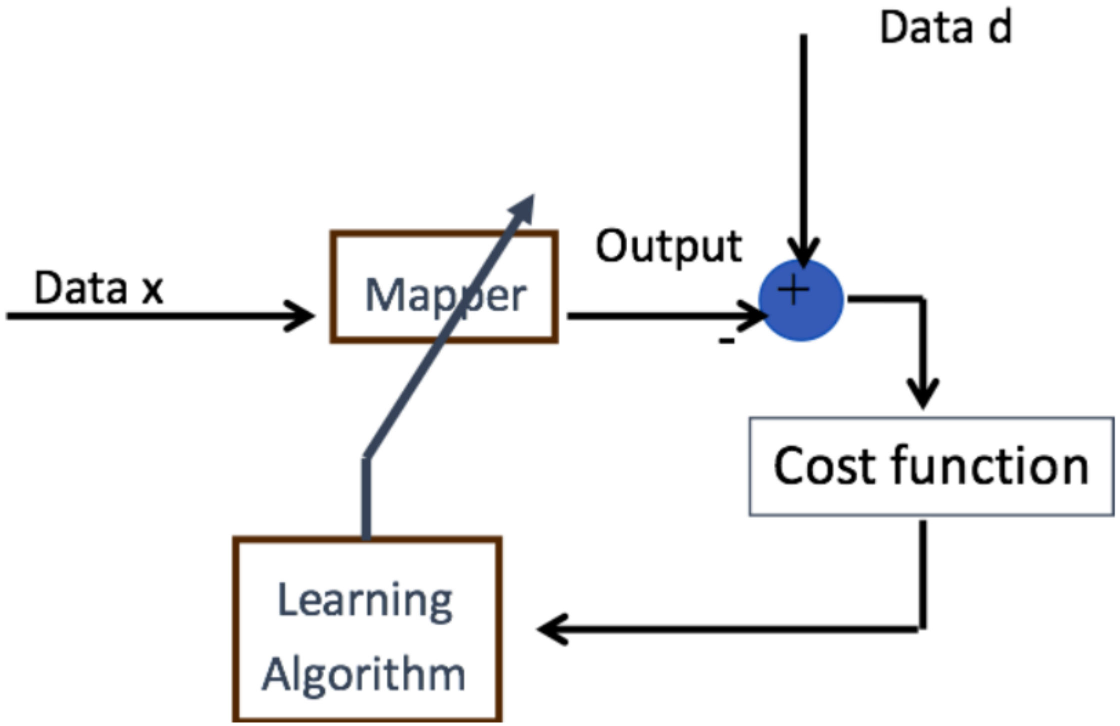
Announcements

- Guest lecture dates
  - October 9th: Dr. John Gottula, Director of Crop Science, Agriculture, SAS
  - November 13th: Dr. Zigfried Hampel-Arias, Research Scientist, Remote Sensing and Data Science, Los Alamos National Laboratory
- Travel planned for November 18th – 21st
  - Lecture slides available for November 18th
  - No class November 20th (Review for exam and work on class project)
- Last day of class: December 2nd (No final exam ☺)

Last Lecture

TEXAS A&M UNIVERSITY Engineering

- Course objectives and Syllabus material
- Introduction to Data Mining and Machine Learning



Today

- Data and attributes
  - **Numerical**
  - Categorical
- Reading: ZM Chapters 2

What can we represent with data?

Data Representations

- Numeric measurements, observations, settings, counts, time intervals, etc. (binary, integer, fixed-point, floating point)
- Text (characters, words, strings, documents)
- Signals (continuous numeric values)
- Time Series (sequence of discrete-time data points often from sensors, communication signals)
- Image and Video (pixel data, series of image data, voxel data, point-clouds)

Data Types

Data Types We Will Use

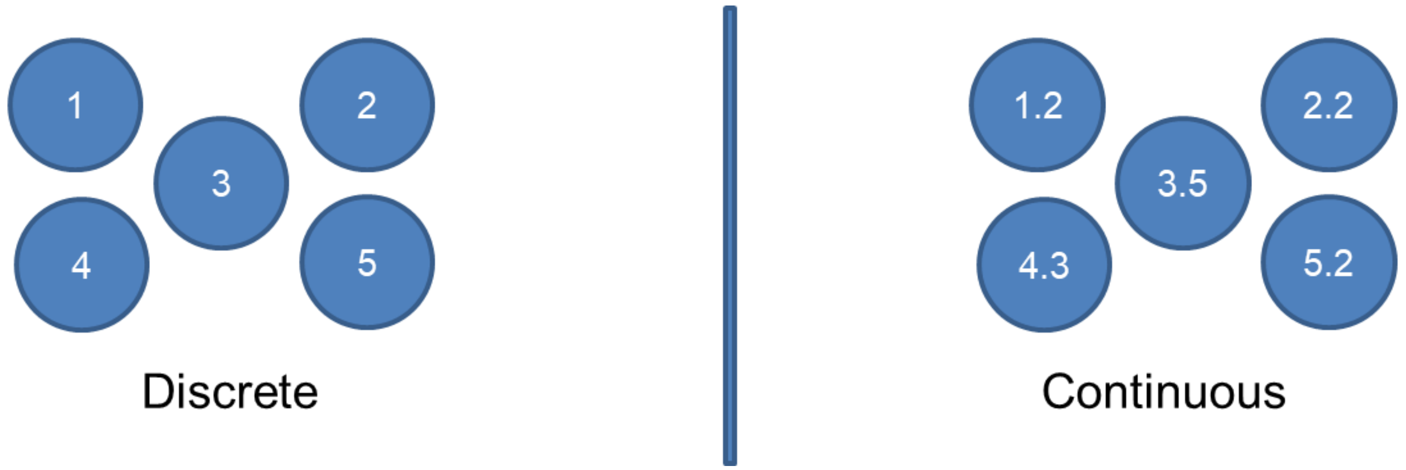
- Data used in Data Mining is generally of two types: Numeric Data and Categorical Data

- Numeric – quantitative, measurable; values are numbers. e.g. 0, 42, 3.1415, 1.602x10^-19

- Categorical – qualitative, recognizable; values are restricted to the possible values in a category and can be represented by a text value or a number. e.g., Tuesday, Medium Rare, Hawaii

Types of Numeric Data

- Discrete – variables can take on only specific values over an interval (e.g., counting numbers, integers)
- Continuous – variable can take on any value over an interval (e.g., real values)



Univariate Analysis

Univariate Data

- Focused on single attribute (e.g., feature)
- Data represented as matrix, **D**
- Each row is a sample and column is an attribute
- X is a random variable
- Each x_i is independent and identically distributed (iid)

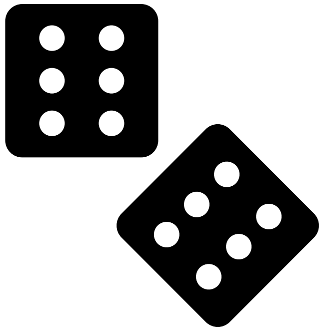
$$\mathbf{D} = \begin{pmatrix} X \\ \hline x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

Probability Review

Sample Space and Probability

- Sample space(Ω, S) : Set of all possible outcomes of an experiment

  Example: Throwing a die, S = {1, 2, 3, 4, 5, 6}



Sample Space and Probability

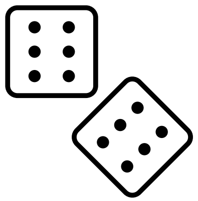
TEXAS A&M UNIVERSITY Engineering

- Sample space(Ω, S) : Set of all possible outcomes of an experiment

  Example: Throwing a die, S = {1, 2, 3, 4, 5, 6}

- Probability law: Assigns a non-negative number to each element of the sample space

  Example: P(1) = P(2) = … = P(6) = 1/6



	1	2	3	4	5	6
Counts	100	100	100	100	100	100
Probability	1/6	1/6	1/6	1/6	1/6	1/6

Sample Space and Probability

- Sample space(Ω, S) : Set of all possible outcomes of an experiment

  Example: Throwing a die, S = {1, 2, 3, 4, 5, 6}
- Probability law: Assigns a non-negative number to each element of the sample space

  Example: P(1) = P(2) = … = P(6) = 1/6
- Event (E): A subset of sample space

  Example: Getting an even number, A = {2, 4, 6}

  P(A) = P(2) + P(4) + P(6) = 1/2

Random Variables and Probability Distributions

Random Variables

- <u>Random variables</u>: A mapping from the sample space to some range
  - -Weather: W in {Sunny, Rainy, Foggy}
  - -Temperature: T in {Hot, Cold}; T in [-50, +50]
- Random variables can be <u>discrete</u> or <u>continuous</u>
- Random variables can take finitely many values or infinitely many values

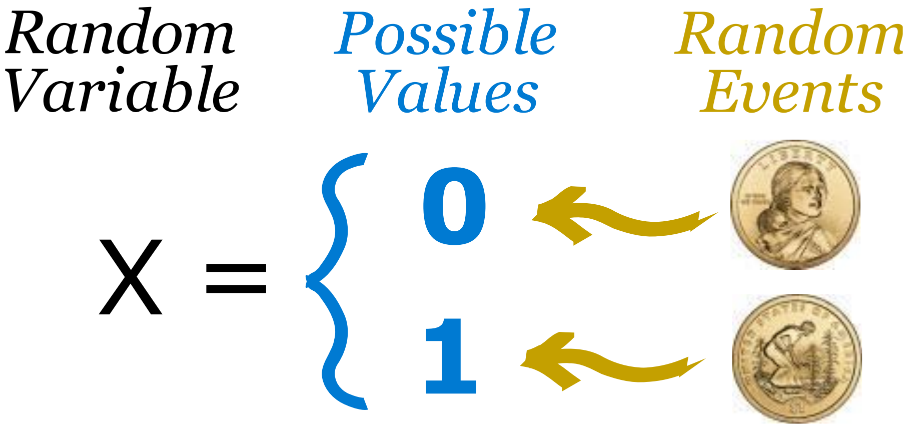
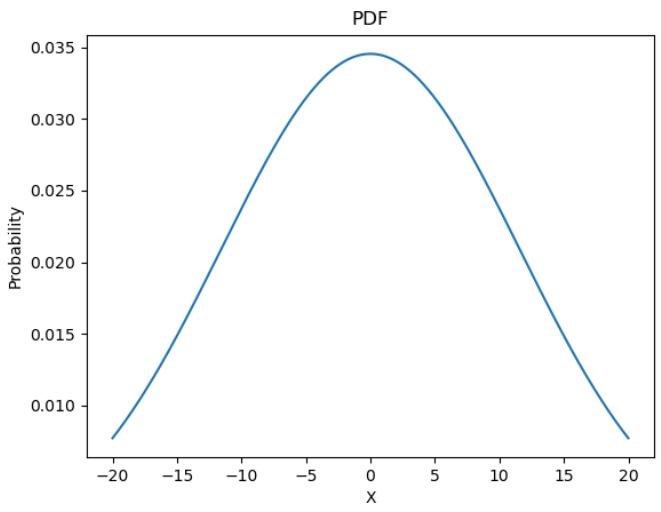
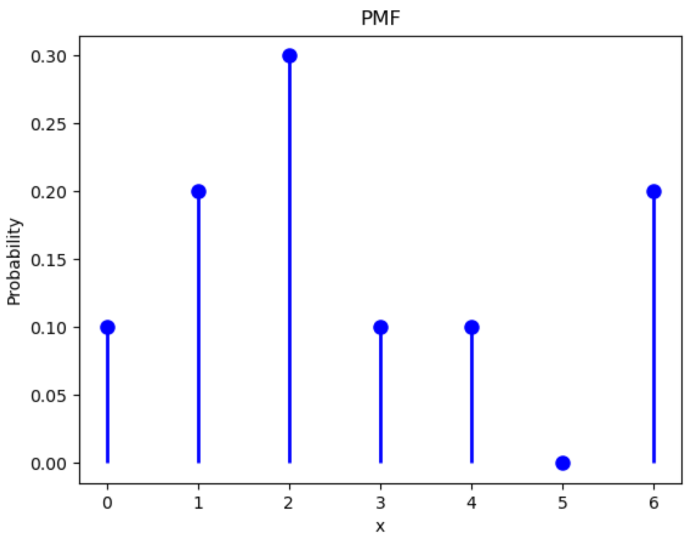


Image from: Math is Fun, Random Variables.

Probability Density Function

- Probability density function (PDF) is an assignment of probability to each possible value of the continuous random variable (RV)
- Probability mass function (PMF) is used for discrete RVs

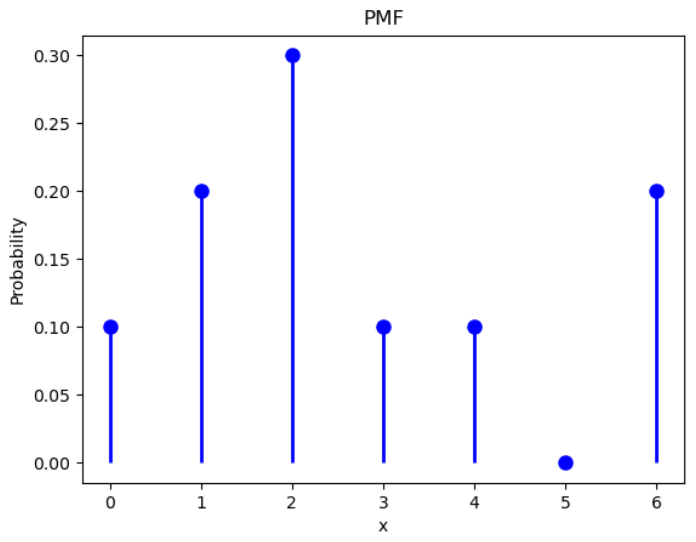




Probability Mass Function

- Probability mass function (PMFs) is used for discrete RVs
- Empirical PMFs assign equal probability to each point

$$\hat{f}(x) = P(X = x) = \frac{1}{n}\sum_{i=1}^{n} I(x_i = x)$$



Cumulative Distribution Function

- Empirical Cumulative Distribution Function (CDF) is the probability that data points (*n*) in the sample are less than or equal to x
- What is the relationship between CDFs and PDFs?

$$\hat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} I(x_i \leq x)$$

Statistical Measures

Populations vs Samples

- Population – set of all data in an area of interest
- Sample – subset of a population

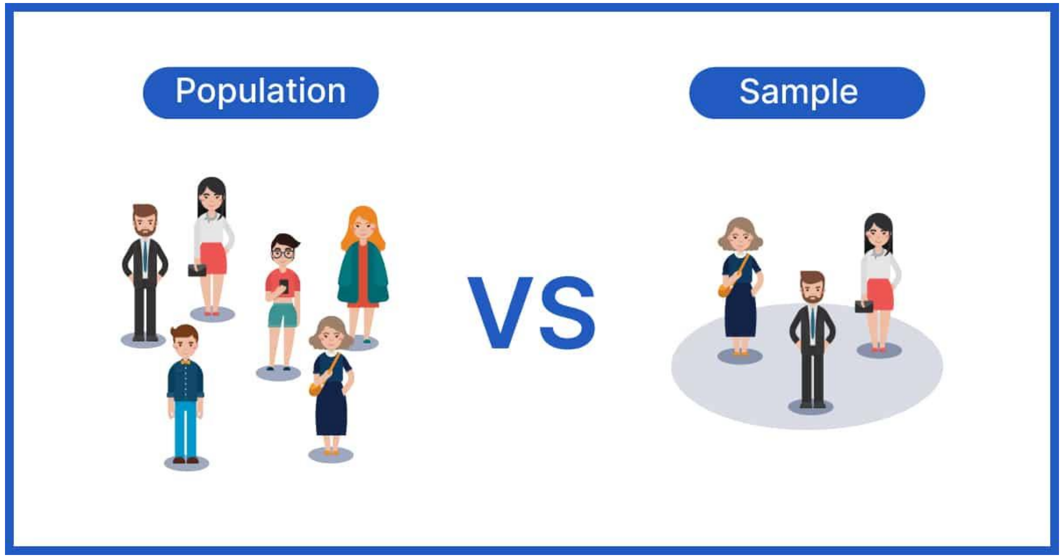


Image from: Voxco. Population vs Sample.

Statistical Measures

- Measures of central tendency
  - Mean, Median, Mode
- Measures of dispersion
  - Range, Interquartile Range, Variance, Standard Deviation

Statistical Measures

- **Measures of central tendency**
  - Mean, Median, Mode
- Measures of dispersion
  - Range, Interquartile Range, Variance, Standard Deviation

Mean

- Arithmetic average of values of X
- Also known as expected value
- f(x) is PMF (discrete) or PDF (continuous)

$$\mu = E[X] = \sum_x x \cdot f(x)$$

Discrete

$$\mu = E[X] = \int_{-\infty}^{\infty} x \cdot f(x)\, dx$$

Continuous

Sample Mean

TEXAS A&M UNIVERSITY Engineering

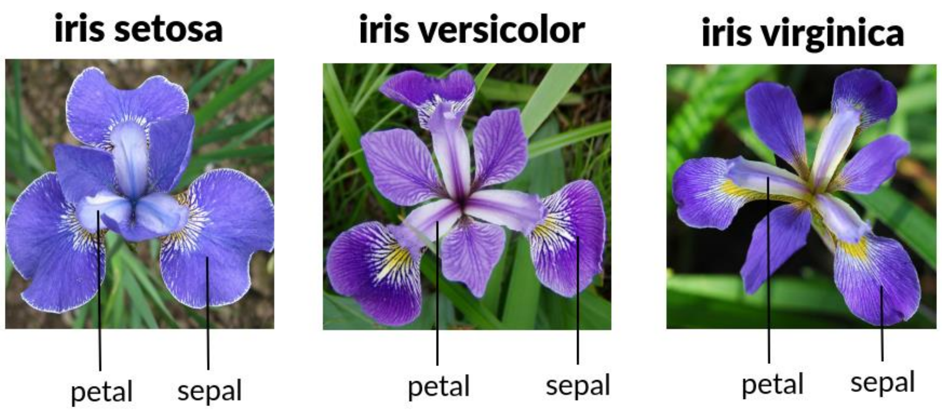
- Statistic defined as average value of x_i
- Estimator for unknown mean value, μ, of X
- **Unbiased** estimator for population mean
- Not a ***robust*** statistic

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

$$E[\hat{\mu}] = E\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right] = \frac{1}{n}\sum_{i=1}^{n} E[x_i] = \frac{1}{n}\sum_{i=1}^{n} \mu = \mu$$

Sample Mean: Iris Sepal Length

- Iris dataset
  - 150 samples
  - 3 classes
  - 4 attributes
    - Petal and Sepal length and width



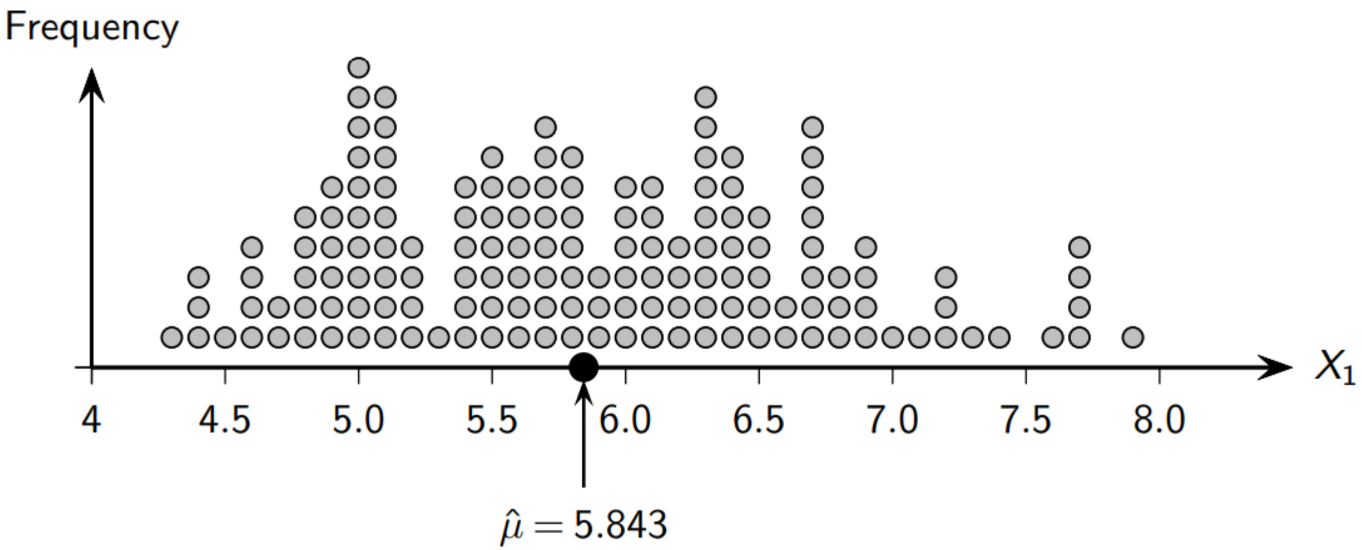


Image from: K. Srivani, IRIS Flowers Classification Using Machine Learning.

Median

- “Middle-most” value
- Half of the values of X are less and half of the values more than median
- Can use CDF or inverse CDF to find median
- Robust statistic

$$P(X \leq m) \geq \frac{1}{2} \text{ and } P(X \geq m) \geq \frac{1}{2}$$

Median, *m*

$$F(m) = 0.5 \text{ or } m = F^{-1}(0.5)$$

Median, *m (CDF)*

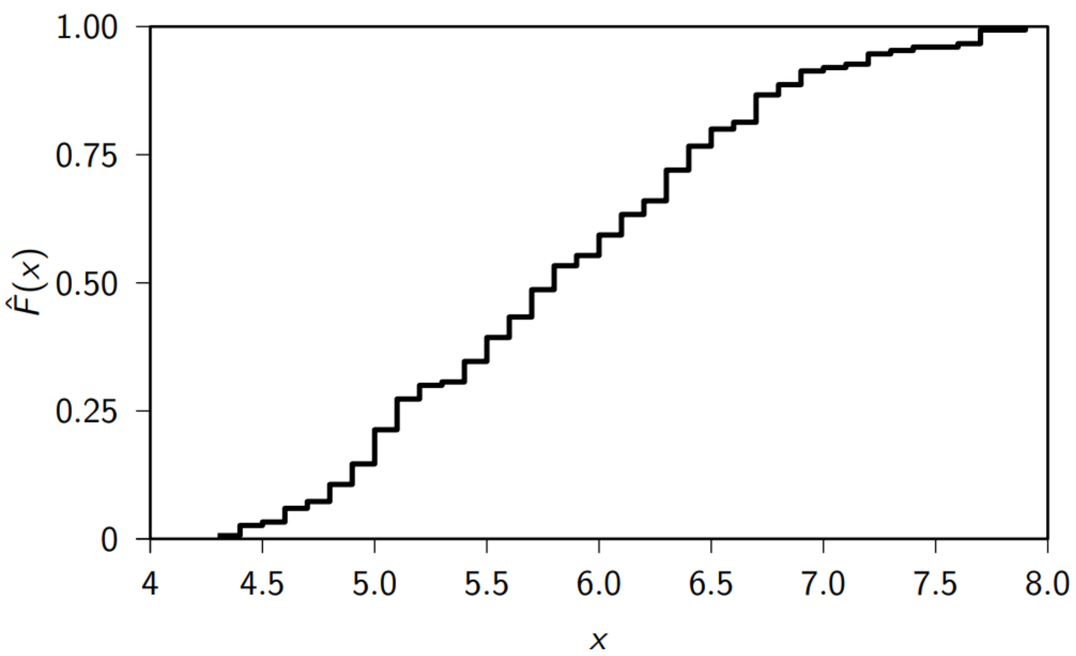
$$\hat{F}(m) = 0.5 \text{ or } m = \hat{F}^{-1}(0.5)$$

Sample Median, $\widehat{m}$ *(CDF)*

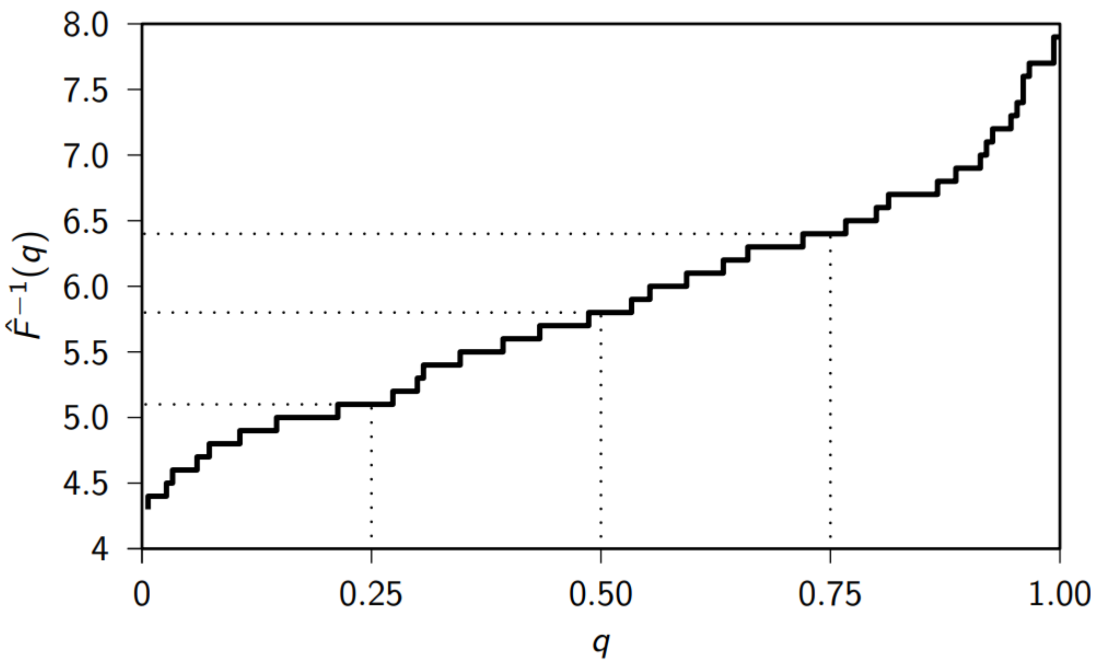
Empirical CDF and Inverse CDF

- Sepal length
- Median is 5.8

$\hat{F}(5.8) = 0.5$



$5.8 = \hat{F}^{-1}(0.5)$



Mode

- Value at which PMF or PDF attains maximum value
- Sample mode computed from empirical PMF

$$\mathrm{mode}(X) = \arg\max_{x} \hat{f}(x)$$

Statistical Measures

- Measures of central tendency
  - Mean, Median, Mode
- **Measures of dispersion**
  - Range, Interquartile Range, Variance, Standard Deviation

Range

- Difference between the maximum and minimum values of X
- Robust statistic?

$$r = \max\{X\} - \min\{X\}$$

Range

$$\hat{r} = \max_{i=1}^{n}\{x_i\} - \min_{i=1}^{n}\{x_i\}$$

Sample range

Interquartile Range (IQR)

- Difference between the 75th and 25th percentiles of X
- More robust than range

$$IQR = F^{-1}(0.75) - F^{-1}(0.25)$$

IQR

$$\widehat{IQR} = \hat{F}^{-1}(0.75) - \hat{F}^{-1}(0.25)$$

Sample IQR

Variance

- Measure of how much values deviate from the expected value of X

$$\boxed{\sigma^2 = \mathrm{var}(X) = E\left[(X-\mu)^2\right]} = \begin{cases} \sum_x (x-\mu)^2 f(x) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\,dx & \text{if } X \text{ is continuous} \end{cases}$$

Sample Variance and Standard Deviation

- Standard deviation is positive square root of variance

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

Sample variance

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

Sample standard deviation

Variance of Sample Mean

- Sample mean is statistic
- Larger values of n will decrease variation of sample mean from mean

$$E[\hat{\mu}] = \mu$$

$$var(\hat{\mu}) = E[(\hat{\mu} - \mu)^2] = \frac{\sigma^2}{n}$$

Variance of Sample Variance

- Sample variance is statistic
- Biased estimator of true population variance
- Larger values of n will reduce bias of estimator

$$E[\hat{\sigma}^2] = \left(\frac{n-1}{n}\right)\sigma^2$$

$$E[\hat{\sigma}^2] \rightarrow \sigma^2 \qquad \text{as } n \rightarrow \infty$$

Bivariate Analysis

Bivariate Data

TEXAS A&M UNIVERSITY Engineering

- Focused on two attributes (e.g., feature)
- Data represented as matrix, **D**
- Each row is a sample and column is an attribute
- **X** is a random variable
- Each $\mathbf{x}_i$ is independent and identically distributed (iid)

$$\mathbf{D} = \begin{pmatrix} X_1 & X_2 \\ \hline x_{11} & x_{12} \\ x_{21} & x_{22} \\ \vdots & \vdots \\ x_{n1} & x_{n2} \end{pmatrix}$$

Bivariate Mean

- Expected value of the vector random variable (**X**)

$$\boldsymbol{\mu} = E[\mathbf{X}] = E\left[\begin{pmatrix} X_1 \\ X_2 \end{pmatrix}\right] = \begin{pmatrix} E[X_1] \\ E[X_2] \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$$

Bivariate mean

$$\hat{\boldsymbol{\mu}} = \sum_{\mathbf{x}} \mathbf{x} \hat{f}(\mathbf{x}) = \sum_{\mathbf{x}} \mathbf{x} \left(\frac{1}{n} \sum_{i=1}^{n} I(\mathbf{x}_i = \mathbf{x}) \right) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i$$

Sample bivariate mean

Covariance

- Measure of association or linear dependence between two attributes (X_1 and X_2)
- Can use to check for independence

$$\begin{aligned}\sigma_{12} &= E[(X_1 - \mu_1)(X_2 - \mu_2)] \\ &= E[X_1 X_2] - E[X_1]E[X_2]\end{aligned}$$

Covariance

$$E[X_1 X_2] = E[X_1] \cdot E[X_2]$$

Independence

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

Sample covariance

Interpreting Covariance

cov(X,Y) > 0 $\longrightarrow$ X and Y are positively correlated

cov(X,Y) < 0 $\longrightarrow$ X and Y are inversely correlated

cov(X,Y) = 0 $\longrightarrow$ X and Y are independent

- *Covariance values are not constrained, and can be from -infinity to +infinity*
- *Covariance is a measure of the directional relationship between variables*

Correction

TEXAS A&M UNIVERSITY Engineering

- Standardized covariance between two attributes (X_1 and X_2)
- Bounded between -1 (negatively correlated) and 1 (positively correlated)

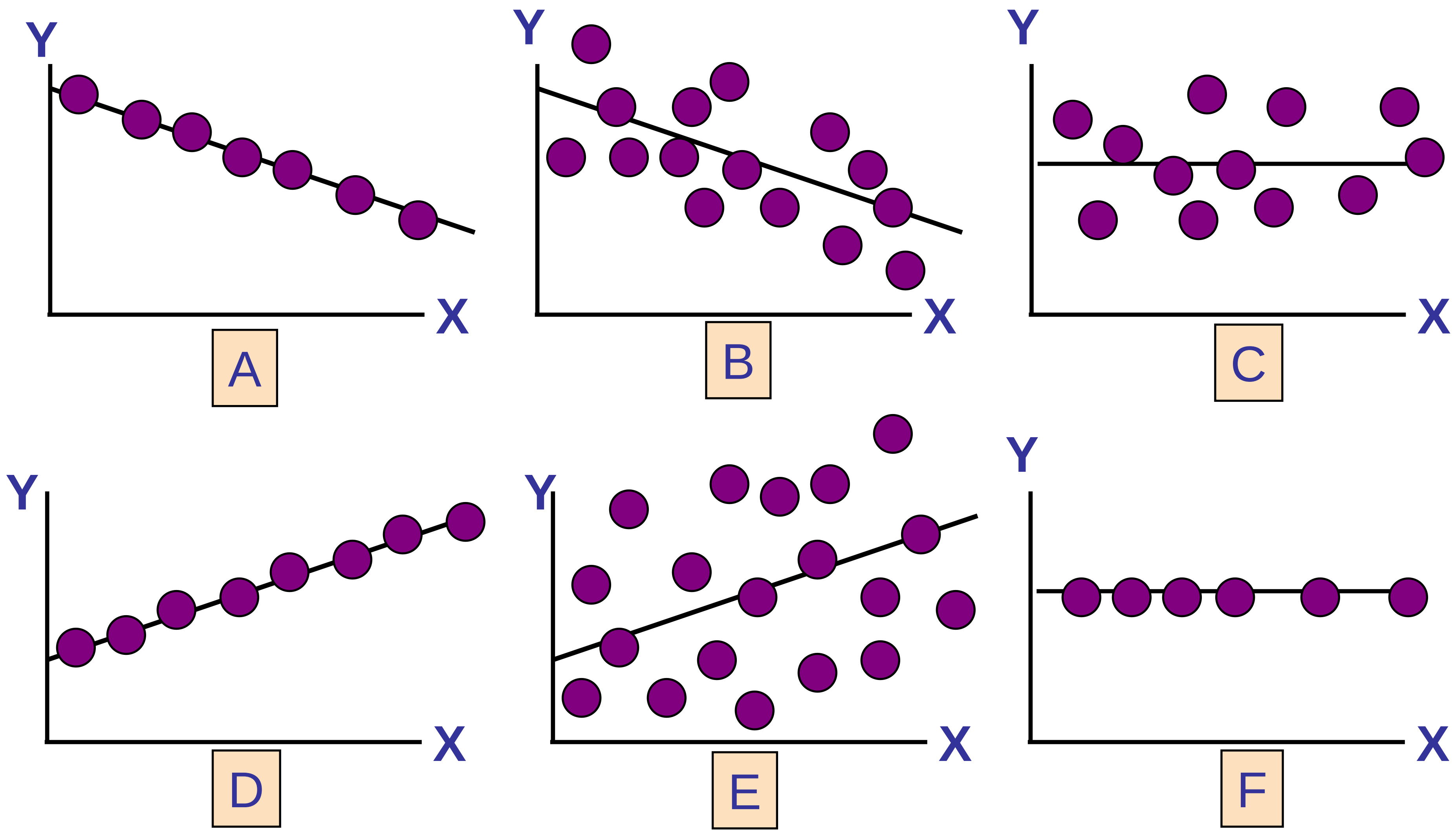
$$\rho_{12} = \frac{\sigma_{12}}{\sigma_1 \sigma_2} = \frac{\sigma_{12}}{\sqrt{\sigma_1^2 \sigma_2^2}}$$

Correlation

$$\hat{\rho}_{12} = \frac{\hat{\sigma}_{12}}{\hat{\sigma}_1 \hat{\sigma}_2} = \frac{\sum_{i=1}^{n} (x_{i1} - \hat{\mu}_1)(x_{i2} - \hat{\mu}_2)}{\sqrt{\sum_{i=1}^{n} (x_{i1} - \hat{\mu}_1)^2 \sum_{i=1}^{n} (x_{i2} - \hat{\mu}_2)^2}}$$

Sample correlation

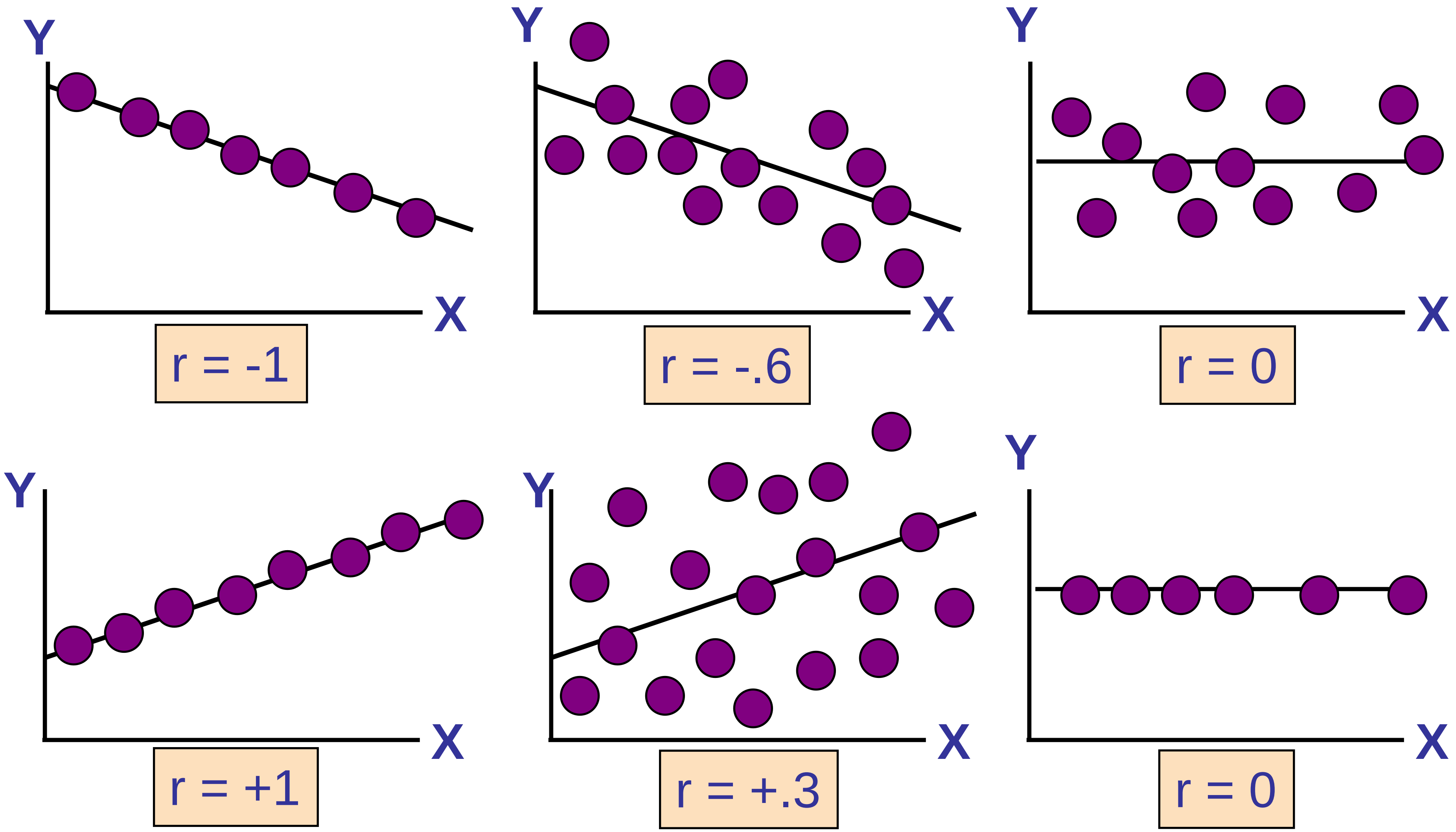
Scatter Plots of Data with Various Correlation Coefficients



Slide from: Statistics for Managers Using Microsoft® Excel 4th Edition, 2004 Prentice-Hall

Scatter Plots of Data with Various Correlation Coefficients

TEXAS A&M UNIVERSITY Engineering



Correlation (Geometric Interpretation)

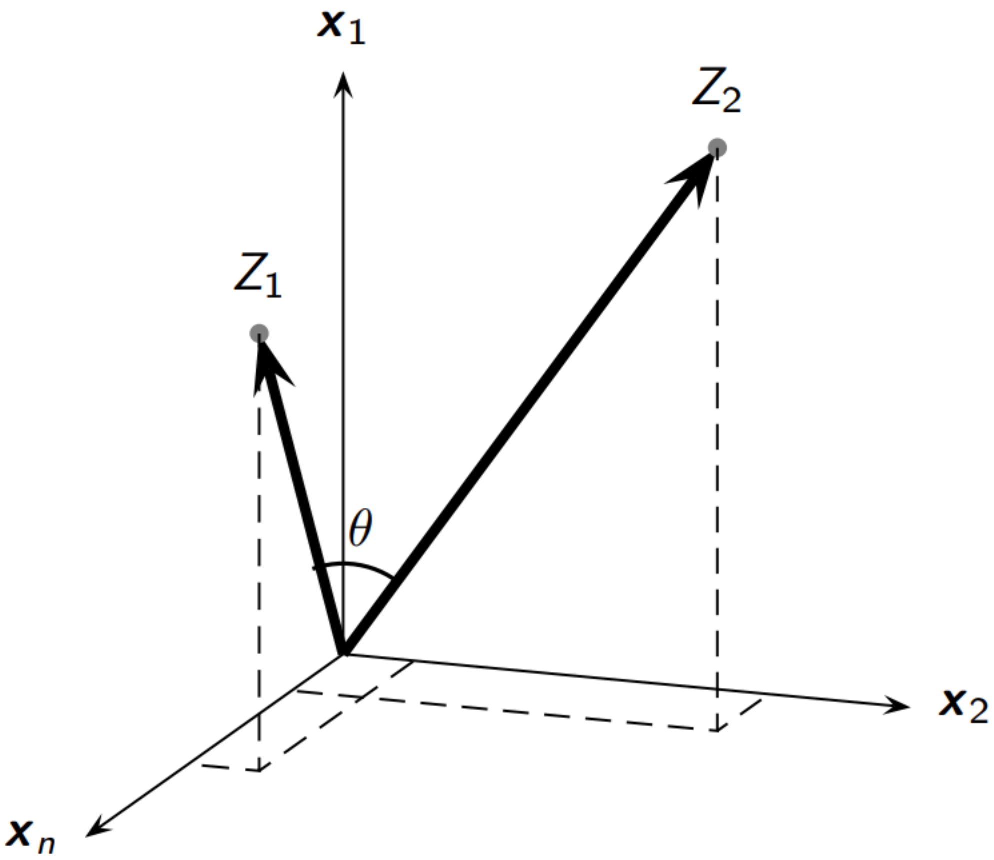
- Cosine of the angle between two centered attribute vectors

$$\bar{X}_1 = X_1 - \mathbf{1}\cdot\hat{\mu}_1 = \begin{pmatrix} x_{11}-\hat{\mu}_1 \\ x_{21}-\hat{\mu}_1 \\ \vdots \\ x_{n1}-\hat{\mu}_1 \end{pmatrix} \qquad \bar{X}_2 = X_2 - \mathbf{1}\cdot\hat{\mu}_2 = \begin{pmatrix} x_{12}-\hat{\mu}_2 \\ x_{22}-\hat{\mu}_2 \\ \vdots \\ x_{n2}-\hat{\mu}_2 \end{pmatrix}$$

$$\hat{\rho}_{12} = \frac{\bar{X}_1^T\bar{X}_2}{\sqrt{\bar{X}_1^T\bar{X}_1}\sqrt{\bar{X}_2^T\bar{X}_2}} = \frac{\bar{X}_1^T\bar{X}_2}{\|\bar{X}_1\|\,\|\bar{X}_2\|} = \left(\frac{\bar{X}_1}{\|\bar{X}_1\|}\right)^T\left(\frac{\bar{X}_2}{\|\bar{X}_2\|}\right) = \cos\theta$$

Correlation (Geometric Interpretation)

- Cosine of the angle between two centered attribute vectors



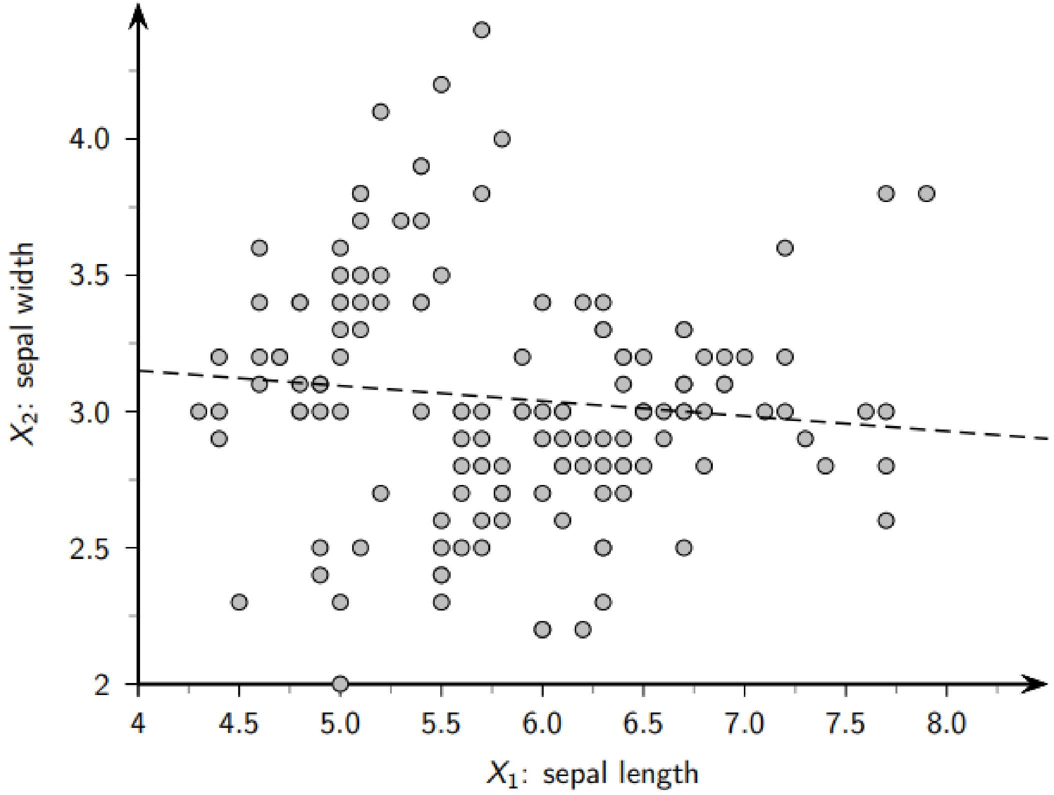
Covariance Matrix

- Summary of variance-covariance information
- Symmetric matrix
- Total variance is trace of covariance matrix

$$\Sigma = E[(\boldsymbol{X} - \mu)(\boldsymbol{X} - \mu)^T]$$
$$= \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix}$$

$$\boxed{var(\boldsymbol{D}) = tr(\Sigma) = \sigma_1^2 + \sigma_2^2}$$

Correlation and Covariance: Iris Dataset



The sample mean is

$$\hat{\boldsymbol{\mu}} = \begin{pmatrix} 5.843 \\ 3.054 \end{pmatrix}$$

The sample covariance matrix is

$$\widehat{\Sigma} = \begin{pmatrix} 0.681 & -0.039 \\ -0.039 & 0.187 \end{pmatrix}$$

The sample correlation is

$$\hat{\rho}_{12} = \frac{-0.039}{\sqrt{0.681 \cdot 0.187}} = -0.109$$

Multivariate Analysis

Multivariate Data

TEXAS A&M UNIVERSITY Engineering

- Focused on d attributes (e.g., feature)
- Data represented as matrix, **D**
- Each row is a sample and column is an attribute
- X is a random variable
- Each x_i is independent and identically distributed (iid)

$$\mathbf{D} = \begin{pmatrix} X_1 & X_2 & \cdots & X_d \\ \hline x_{11} & x_{12} & \cdots & x_{1d} \\ x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nd} \end{pmatrix}$$

Multivariate Mean

TEXAS A&M UNIVERSITY Engineering

- Expected value of the vector random variable (**X**)

$$\boldsymbol{\mu} = E[\mathbf{X}] = \begin{pmatrix} \mu_1 & \mu_2 & \cdots & \mu_d \end{pmatrix}^T$$

Mean Vector

$$\boxed{\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i}$$

Sample mean vector

Covariance Matrix

- Symmetric
- Positive, semi-definite (PSD)
- Total variance is sum of diagonal (trace)

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1d} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2d} \\ \cdots & \cdots & \cdots & \cdots \\ \sigma_{d1} & \sigma_{d2} & \cdots & \sigma_d^2 \end{pmatrix} \qquad \widehat{\Sigma} = \begin{pmatrix} \hat{\sigma}_1^2 & \hat{\sigma}_{12} & \cdots & \hat{\sigma}_{1d} \\ \hat{\sigma}_{21} & \hat{\sigma}_2^2 & \cdots & \hat{\sigma}_{2d} \\ \cdots & \cdots & \cdots & \cdots \\ \hat{\sigma}_{d1} & \hat{\sigma}_{d2} & \cdots & \hat{\sigma}_d^2 \end{pmatrix}$$

$$\boxed{tr(\boldsymbol{\Sigma}) = \sigma_1^2 + \sigma_2^2 + \cdots + \sigma_d^2}$$

Sample Covariance Matrix: Inner and Outer Product

- Pairwise inner or dot product of centered attribute vectors normalized by sample size

$$\hat{\Sigma} = \frac{1}{n}\left(\overline{\mathbf{D}}^T\ \overline{\mathbf{D}}\right) = \frac{1}{n}\begin{pmatrix} \overline{X}_1^T\overline{X}_1 & \overline{X}_1^T\overline{X}_2 & \cdots & \overline{X}_1^T\overline{X}_d \\ \overline{X}_2^T\overline{X}_1 & \overline{X}_2^T\overline{X}_2 & \cdots & \overline{X}_2^T\overline{X}_d \\ \vdots & \vdots & \ddots & \vdots \\ \overline{X}_d^T\overline{X}_1 & \overline{X}_d^T\overline{X}_2 & \cdots & \overline{X}_d^T\overline{X}_d \end{pmatrix}$$

- Sum of rank-one matrices calculated as outer product of each centered point

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} \overline{\mathbf{x}}_i \cdot \overline{\mathbf{x}}_i^T$$

Data Normalization

Min-Max or Range Normalization

- Each sample scaled by the sample range
- Features normalized between 0 and 1

$$x_i' = \frac{x_i - \min_i\{x_i\}}{\hat{r}} = \frac{x_i - \min_i\{x_i\}}{\max_i\{x_i\} - \min_i\{x_i\}}$$

Standard Score Normalization

- Z-normalization
- Scales data to be centered (zero mean) and unit variance

$$x_i' = \frac{x_i - \hat{\mu}}{\hat{\sigma}}$$

Range vs Standard Score Normalization

•Range normalization can be useful where distribution of the data is unknown and in algorithms that do not make assumptions of distribution of the data.

•Standardization is well suited to data that is characterized by a Normal (aka Gaussian) distribution. Its application is not just restricted to such data. Standardization is more robust to outliers.

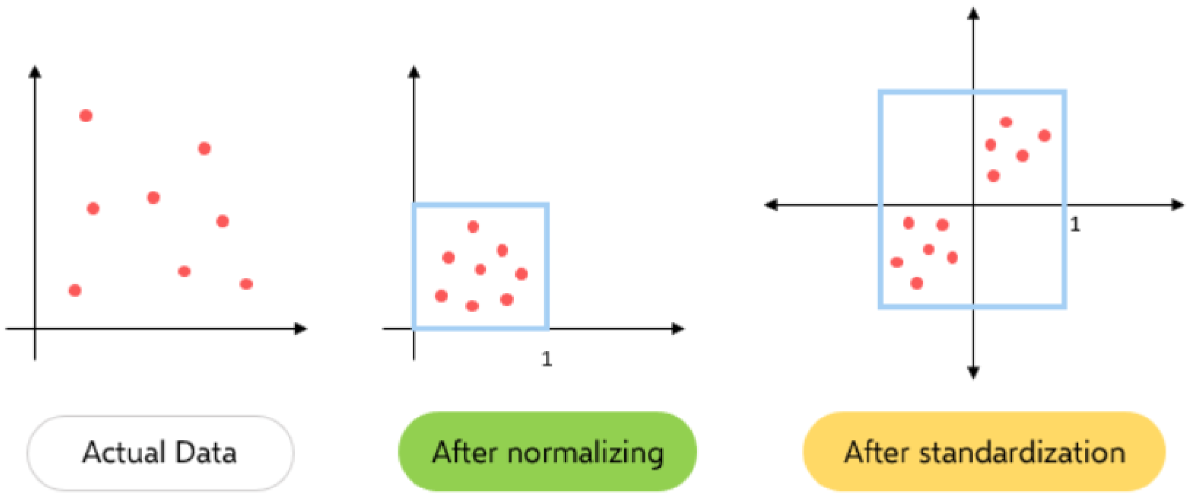
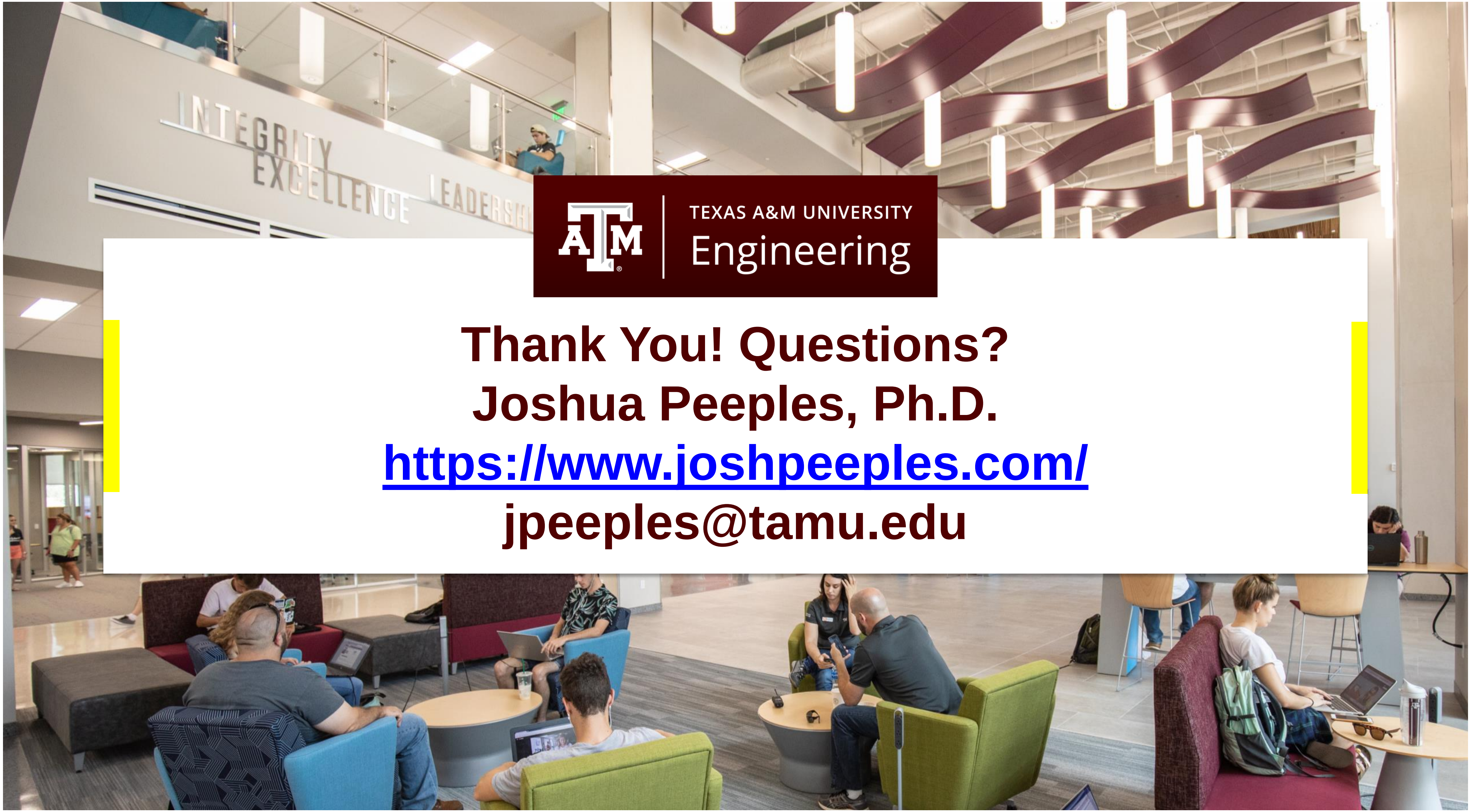


Image from: Someka, How to Normalize Data in Excel.

Next class

- Data and attributes
  - Numerical
    - Normal distribution
  - Categorical



TEXAS A&M UNIVERSITY
Engineering

Thank You! Questions?

Joshua Peeples, Ph.D.

https://www.joshpeeples.com/

jpeeples@tamu.edu

TEXAS A&M UNIVERSITY
Engineering

Supplemental Slides

Tools for Statistics

- **Python** **statistics** is a built-in Python library for descriptive statistics. You can use it if your datasets are not too large or if you can't rely on importing other libraries.

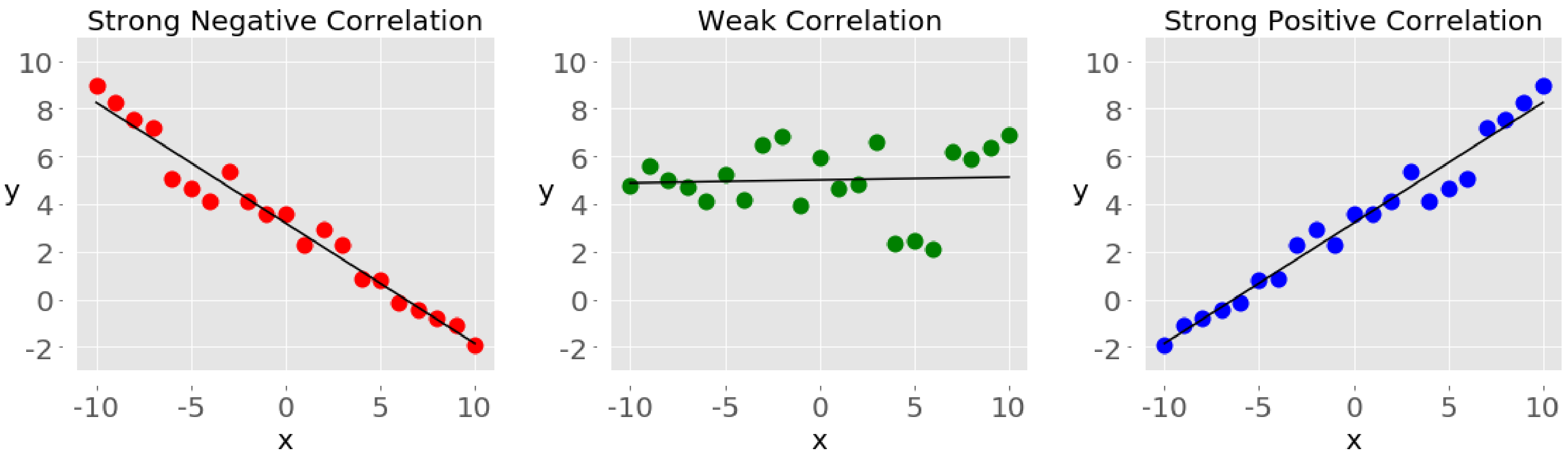


Image from: Python Statistics Fundamentals: How to Describe Your Data

Tools for Statistics

- [NumPy](#) is a library for numerical computing, optimized for working with single- and multi-dimensional arrays. Its primary type is the array type called ndarray. This library contains many routines for statistical analysis.
- **SciPy** is a library for scientific computing based on NumPy. It offers additional functionality compared to NumPy, including scipy.stats for statistical analysis.

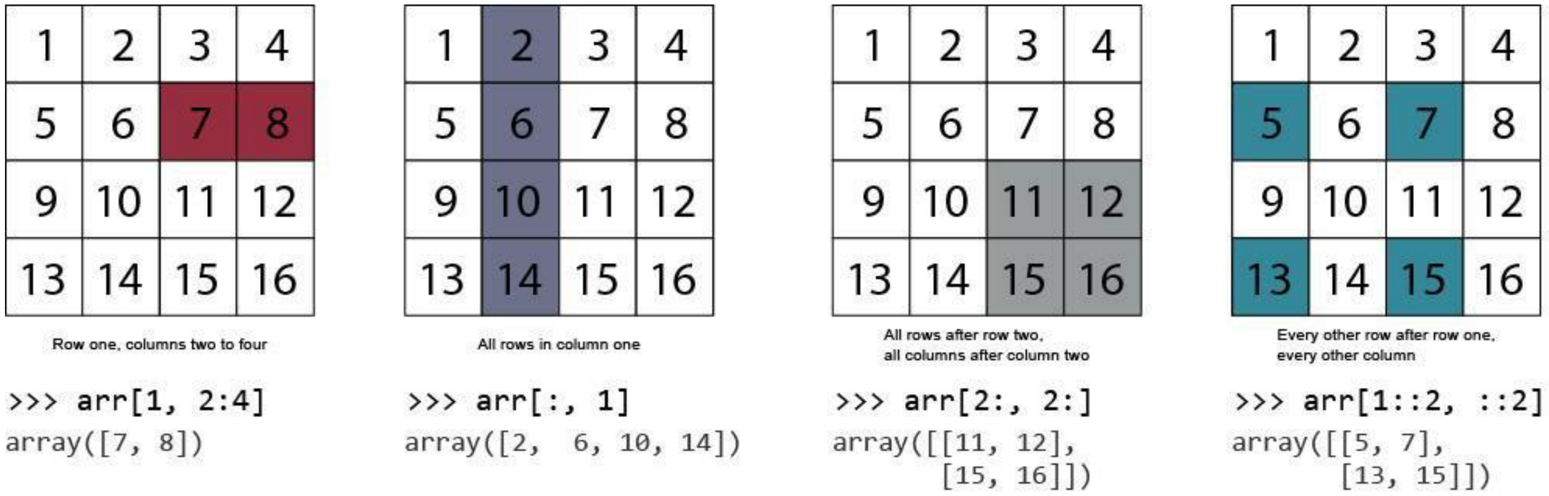


Image from: Python NumPy Tutorial: An Applied Introduction for Beginners

Tools for Statistics

- **[Pandas]** is a library for numerical computing based on NumPy. It excels in handling labeled one-dimensional (1D) data with Series objects and two-dimensional (2D) data with DataFrame objects.

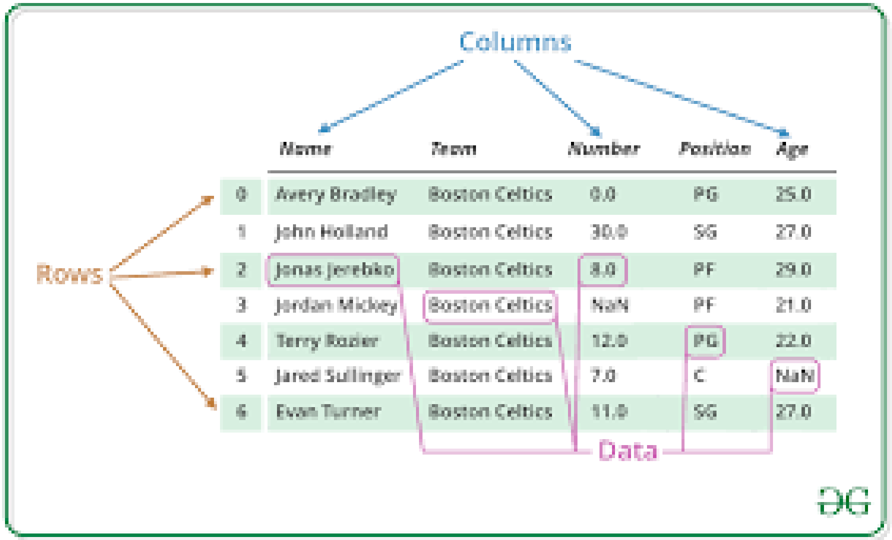


Image from: GeeksforGeeks. Python | Pandas DataFrame

Tools for Statistics

TEXAS A&M UNIVERSITY Engineering

- **Matplotlib** is a library for data visualization. It works well in combination with NumPy, SciPy, and Pandas.

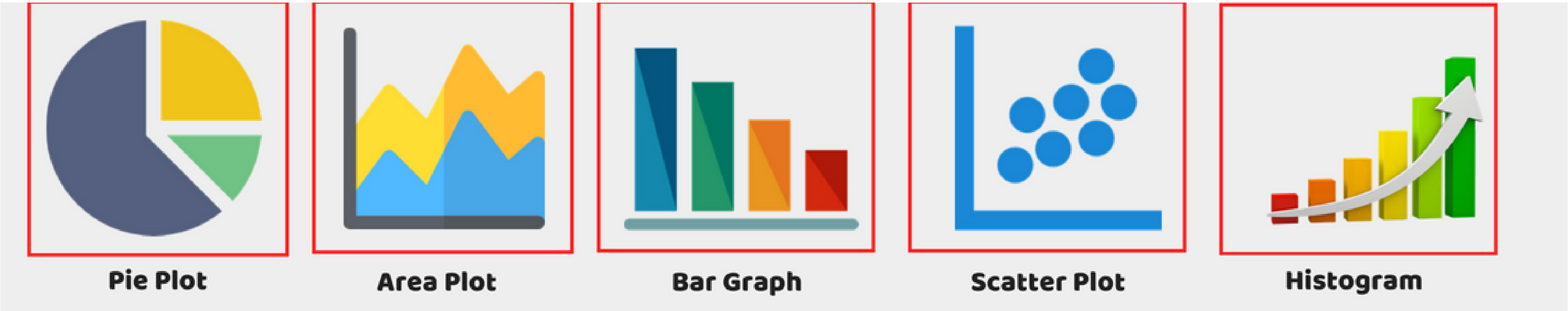


Image from: Matplotlib Library Tutorial with Examples – Python | Datascience+

Tools for Statistics

- Seaborn pair-plots give us a good way to view correlations between pairs of variables (features):