Statistical Graphics:

What is Exploratory Data Analysis?

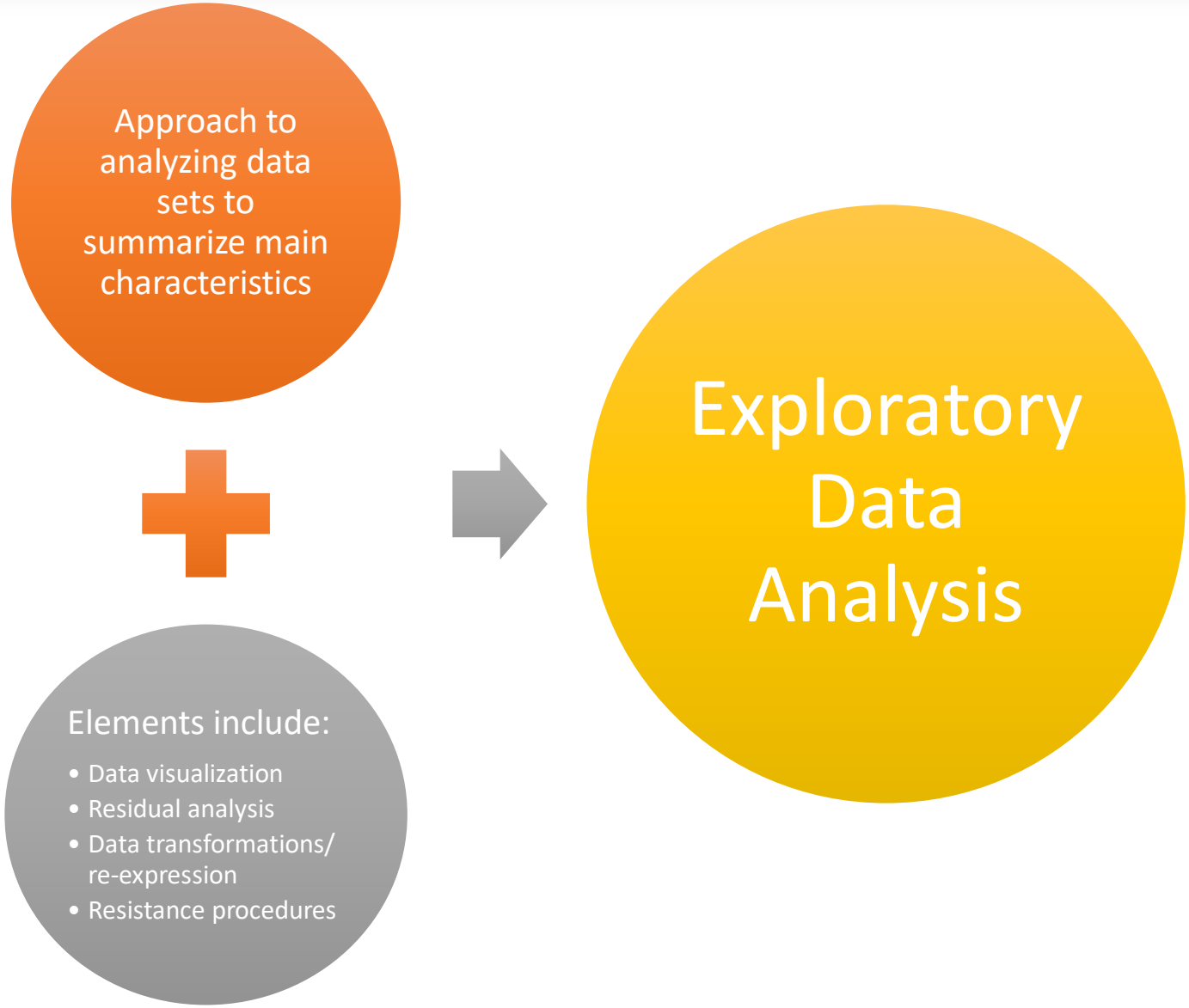
Ira A. Fulton Schools of Engineering
Arizona State University

Objective

Objective

Describe exploratory data analysis

Exploratory Data Analysis



Data Visualization

- Data visualization facilitates advanced data analysis
- Checks distributional and other assumptions
- Observes time-based processing
- Spots outliers
- Examines relationships
- Discriminates clusters
- Compares mean differences

Data Distributions

The type of data distribution affects

- How it should be analyzed
- How it should be visualized

Key step is pre-conditioning data

The Normal Distribution

Normal (Gaussian) Distribution

- Popular
- Fully characterizes with two parameters
- Probability is determined knowing distance from mean
- Many measures and tests are designed for this

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Mean and Standard Deviation

For sample population X = $\{x_1, \ldots, x_n\}$ the mean is defined as:

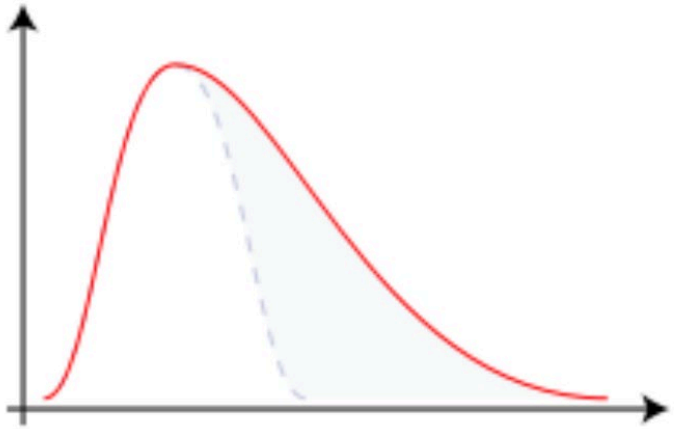
$$\mu = \frac{1}{N}\sum_{i=0}^{N} x_i$$

The standard deviation is defined as:

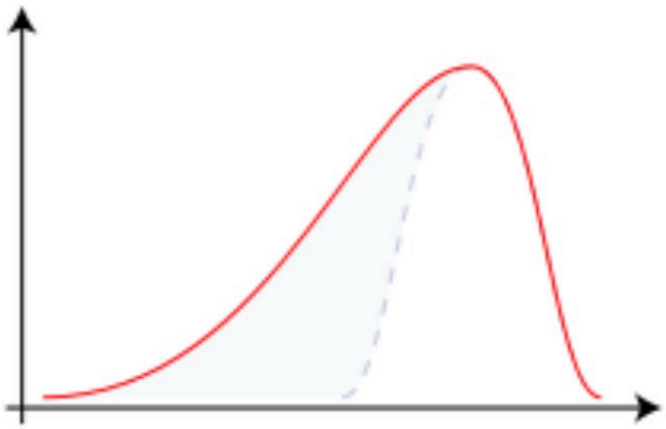
$$\sqrt{\frac{1}{N}\sum_{i=0}^{N}(x_i - \mu)^2}$$

Skewness

Measure of the asymmetry of the probability distribution



Positive Skew



Negative Skew

Skewed Data

For a sample of N values, the sample skewness is:

$$\gamma = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^3}{\left(\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2\right)^{3/2}}$$