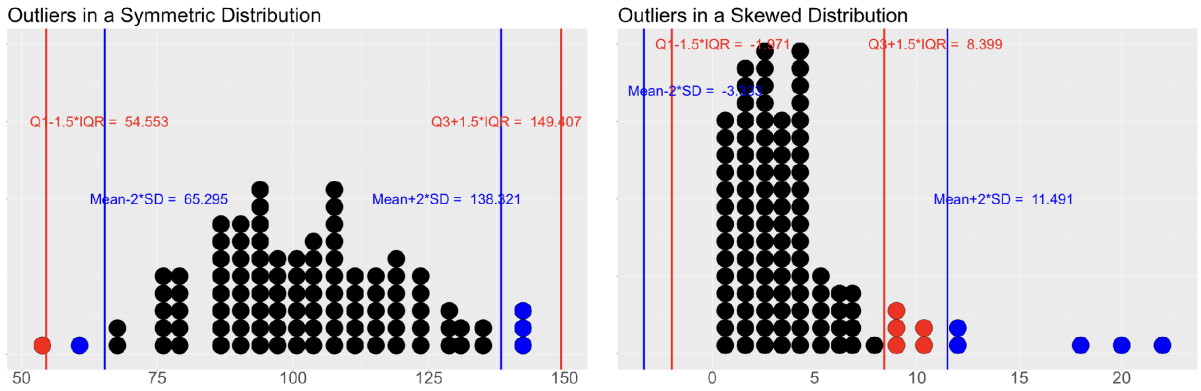
Outliers

Identifying Outliers

- Mild Outliers – Values that slightly deviate from the typical span of values
- Methods to identify outliers
    1. Observation – By looking at values in some visualization, we can sometimes manually pick out outliers
    2. Quartile Method – Uses quartiles and interquartile range (IQR)
        - Mild Outlier
            - Value falls above $Q3 + 1.5 \times IQR$ or below $Q1 - 1.5 \times IQR$
        - Regular Outlier
            - Value falls above $Q3 + 3 \times IQR$ or below $Q1 - 3 \times IQR$
        - Quartile 1 (Q1) and Quartile (Q3) represent the value where 25% and 75% of the values are equal to it or less
        - IQR = Q3 - Q1
    3. Mean / Standard Deviation Method
        - Mild Outlier
            - Value falls above $Mean + 2 \times SD$ or below $Mean - 2 \times SD$
        - Regular Outlier
            - Value falls above $Mean + 3 \times SD$ or below $Mean - 3 \times SD$

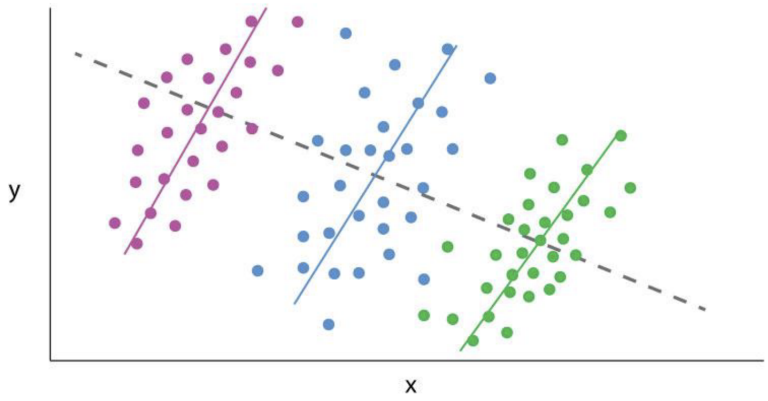


Consideration of Outliers

- In some situations, we may want to remove outliers and they can drastically affect our analysis
- However, sometimes the presence of outliers give us important information as they could be representing rare but important events
    - Rare diagnosis
    - Underrepresented group

Simpson's Paradox

- Simpson's Paradox – Phenomenon where an association between two attributes in a population emerges, but disappears or reevrses when the population is divided into groups
  - Gender bias among graduate student permissions (1978)
    - Overall, men were more likely to be admitted than women
    - However, there were a wide variation of departments being applied to
    - Women tended to apply to more competitive departments with lower rates of admissions
    - In most departments, women were admitted at a higher rate than men



- Consider the example involving the treatment of kidney stones
- The below table shows treatment success for two different treatment types
  - Treatment A – Open surgery
  - Treatment B – Percutaneous Nephrolithotomy

	Treatment A	Treatment B
Stone (diameter $< 2cm$)	81/87 **(93.1%)**	234/270 (86.7%)
Stone (diameter $\geq 2cm$)	192/263 **(73%)**	55/80 (68.8%)
Overall	273/350 (78%)	289/350 **(82.6%)**

Classification

- The task of classification involves assigning a class / label to an instance
- The class / label is some discrete value that represents the instance and that we are interested in (want to predict)
  - Sentiment of a document
  - Name of a species
  - Spam / Not spam
- The class is from a predefined set of classes

Binary vs. Multi-Class Classification

- Binary classification considers only 2 classes
  - Good / Bad
  - Spam / Not Spam
  - Cat / Dog
- Multi-class classification classifies an instance as 1 of 3 or more potential classes
  - Negative / Neutral / Positive
  - Urgent / Normal / Spam
  - Cat / Dog / Bird / Snake / Other

Classifier

- The job of a classifier is to perform the classification task
- Input:
  - A fixed instance d
  - A fixed set of classes $Y = \{y_1, y_2, \ldots, y_j\}$
- Output: A predicted class $\hat{y} \in Y$

Instance as a whole

- Up to this point, we primarily discussed and analyzed individual attributes of an instance
- We will now consider multiple attributes of an instance

Number of Legs	Crustacean	Can Fly
10	Yes	No

- When we use the attribute's values as input to a classifier, we typically refer to them as features
- A feature vector a vector of the different values representing an instance
  - $[1, 4, -2]$
- The feature vector is what will be used to classify the instance