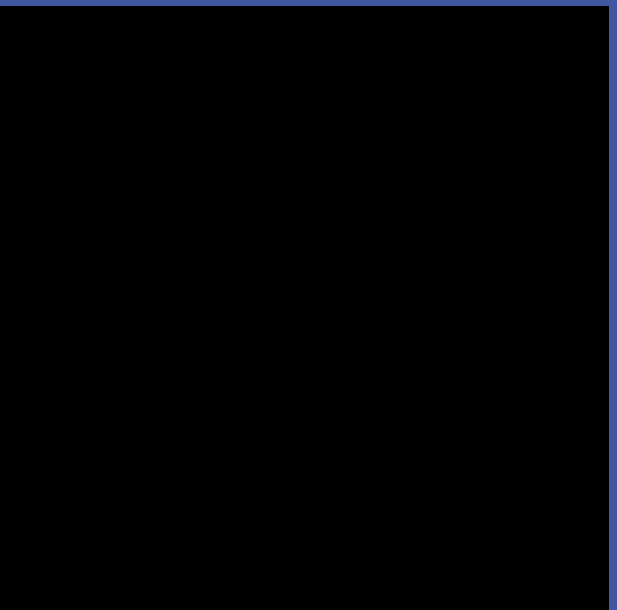
Data Operations

Lesson



AI Academy

Why is it important to know the type of an attribute?

Importance of Attribute Types

Why is it important to know the type of an attribute?

This dictates:

- What **statistics and tests** can be calculated
- What **transformations** can be performed
- How **learning algorithms** can use the attribute

Mode

- The **most common value** for an attribute
- Can be calculated for **all data types**

-3 2 2 4 5 **7** **7** **7** 10

Median

- The Median is the **middle-ranked** value
  - Sort all your values, take the middle one
  - If the count is even, average the two middle values
- Summary of distribution that is **robust to outliers**
- Can be calculated for the following data types:
  - Ordinal
  - Interval
  - Ratio

-3 2 2 4 **5** 7 7 7 10

Mean

Unweighted arithmetic mean

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Weighted arithmetic mean

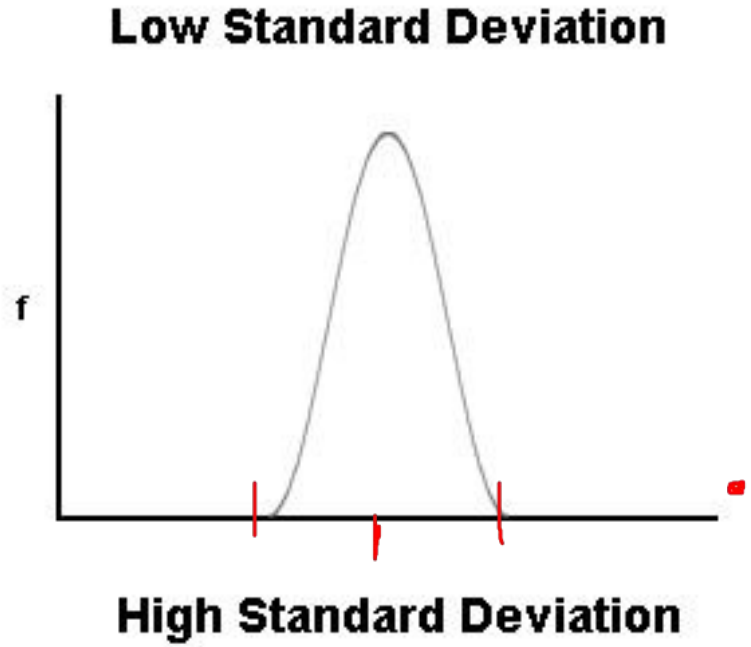
$$\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

- Can be calculated for **Interval and Ratio** data types
- *Can you take the mean of an ordinal attribute?*

Spread

- **Range:** max-min
- **Variance:** s^2

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$



- **Standard Deviation:** s
  - Square root of the variance
  - Measures spread about the mean
  - Zero if and only if all the values are equal
- **Median Absolute Deviation**
  - Median of distances from the mean

Transformation: Z-score Normalization

- Values normalized by mean and standard deviation

$$z(x_i) = \frac{x_i - \overline{x}}{s}$$

- Centered around 0
- Distance in units of standard deviation
- Can be calculated for **Ratio and Interval** data
- *Remember Z-values when we talk about normalization*

More Summary Statistics & Tests

- **Nominal:** mode, entropy, Chi-squared (χ^2) test
- **Ordinal:** median, percentiles, rank correlation, rank sum test, sign test
- **Interval:** mean, standard deviation, Pearson's correlation, *t*-test, *z*-score
- **Ratio:** Geometric mean, harmonic mean, percent variation
- Each type can use the stats above too (e.g. Ratio data can have a median)

Learning Objectives: Operations

You now should be able to:

Identify **operations** and **transformations** that can applied to different types of data

AI Academy
NC STATE

Data Operations

Exercises

AI Academy

Practice Question: Operations

Which of these following operations can be applied to **ordinal** data?

A. Mean
B. Median
C. Standard Deviation
D. Mode