Finding Relationships in Data with Python

IDENTIFYING AND VISUALIZING COMMON RELATIONSHIPS IN DATA

Janani Ravi
CO-FOUNDER, LOONYCORN

www.loonycorn.com

Overview

Common statistical relationships

Univariate, bivariate and multivariate relationships

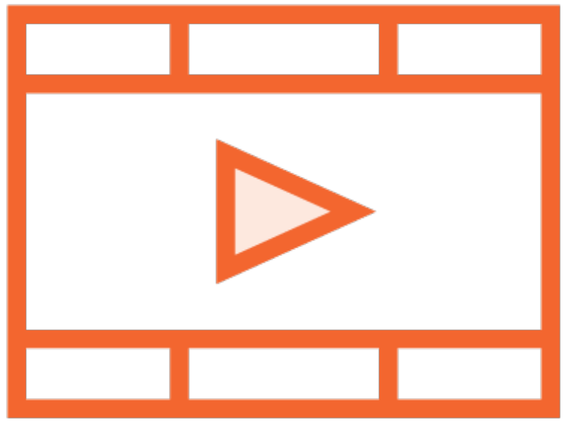
Mean, standard deviation and variance

Covariance and correlation

Autocorrelation

Prerequisites and Course Outline

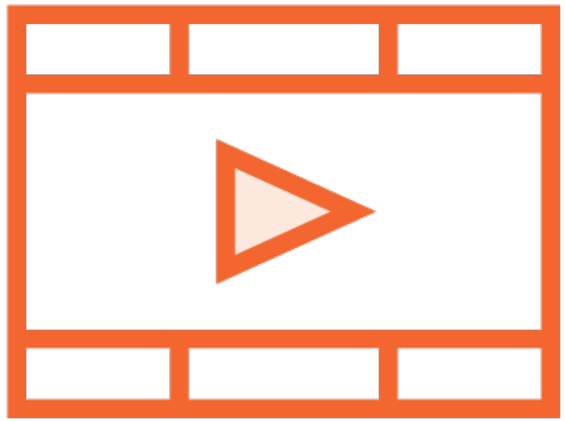
Prerequisites



Basic Python programming

Basic knowledge of math at the level of what an arithmetic mean is

Prerequisites



Python Fundamentals

Course Outline

Identifying and visualizing common relationships in data

Identifying and visualizing probabilistic and statistical relationships

Using interactive visualizations to explore relationships in data

Statistics in Understanding Data

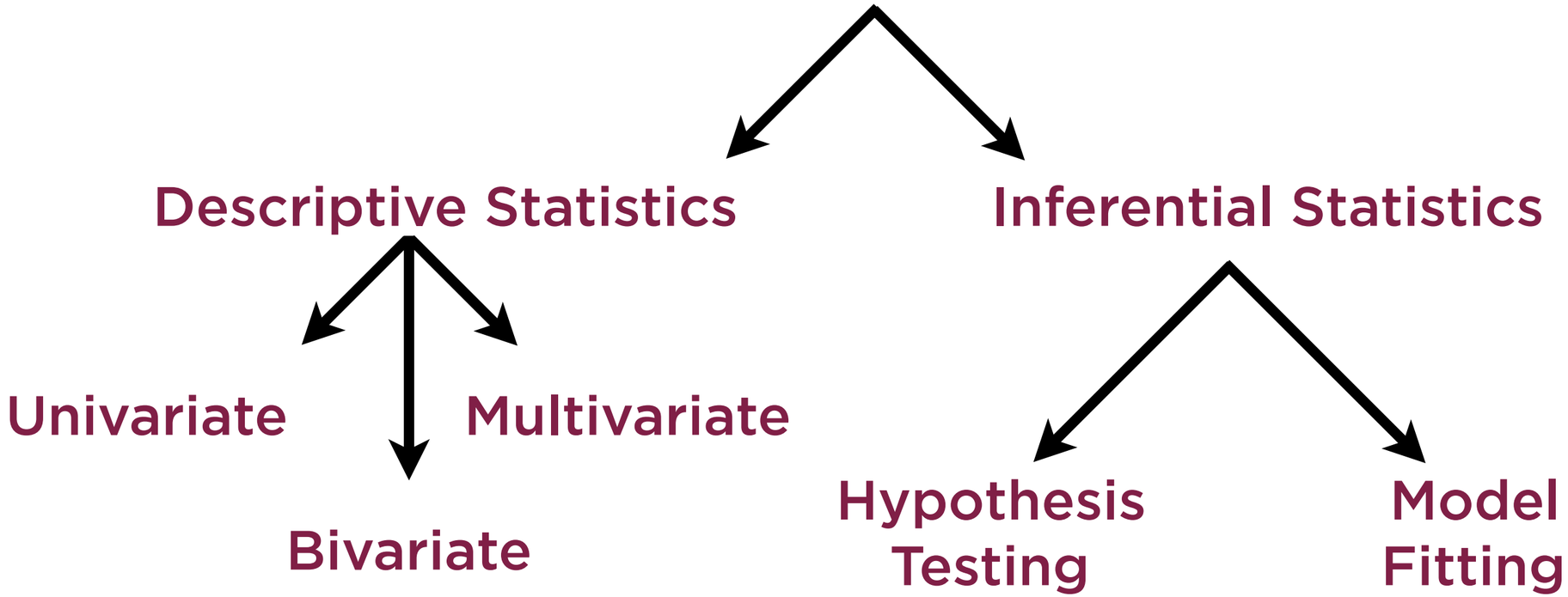
“There are two kinds of statistics, the kind you look up and the kind you make up”

Rex Stout

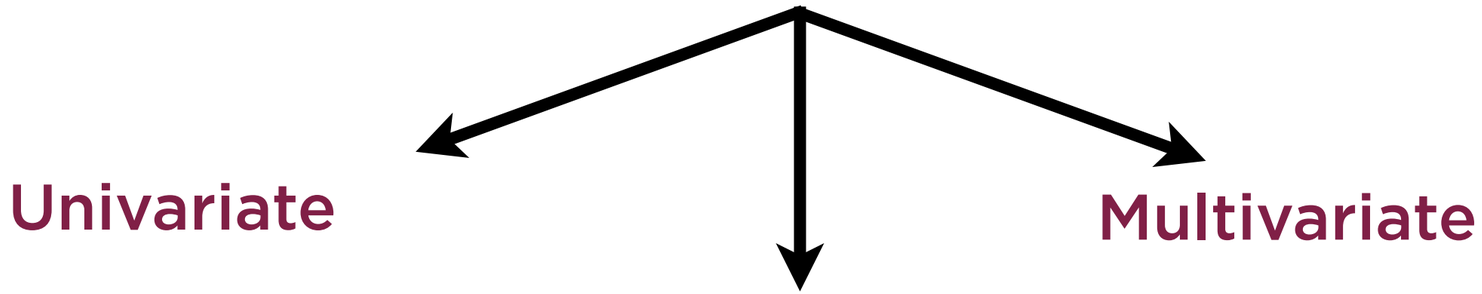
Statistics

A branch of mathematics that deals with collecting, organizing, analyzing, and interpreting data

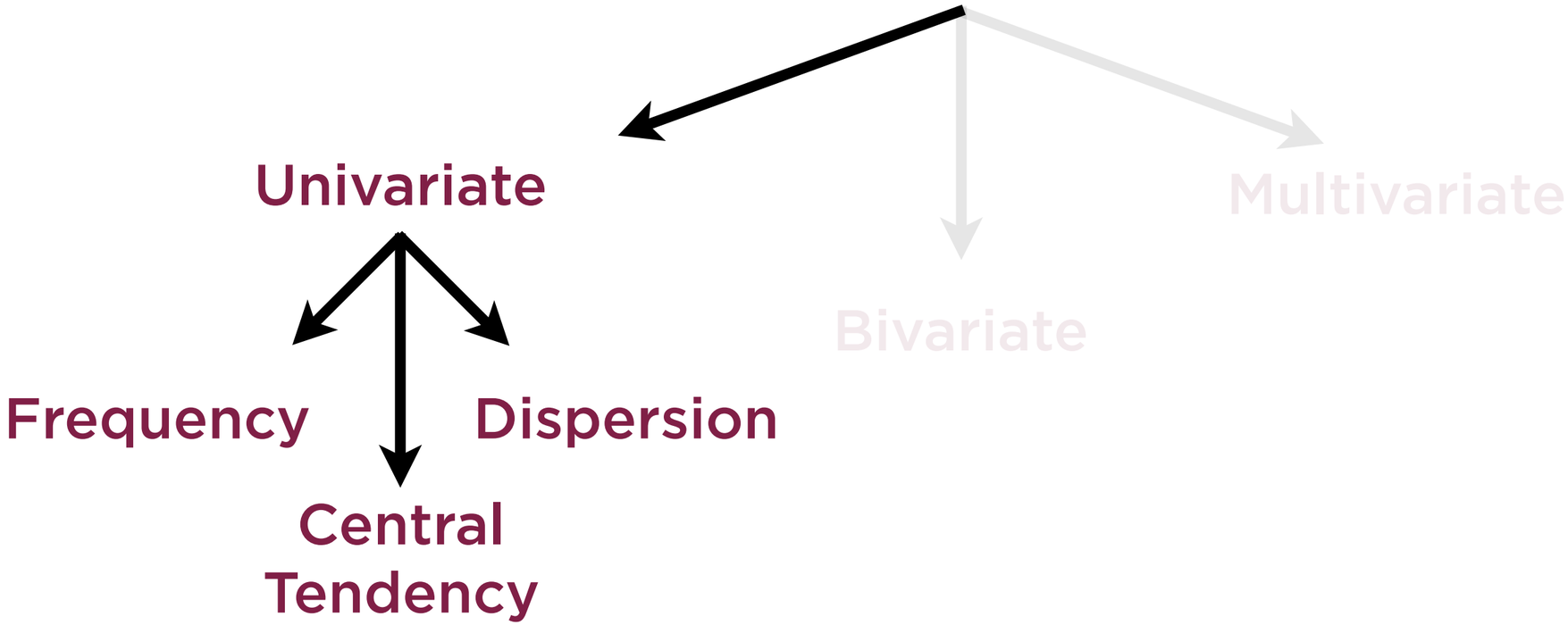
Statistics


Descriptive Statistics
Inferential Statistics
Univariate
Bivariate
Multivariate
Hypothesis Testing
Model Fitting

Descriptive Statistics



Descriptive Statistics



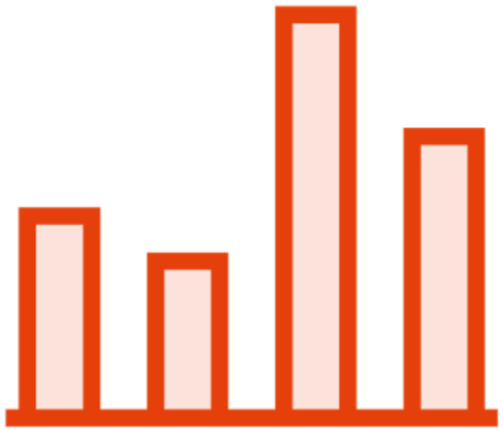
Univariate Descriptive Statistics

Measures of Frequency

Measures of Central Tendency

Measures of Dispersion

Measures of Frequency



Frequency tables

Histograms

Measures of Central Tendency

Average (Mean)

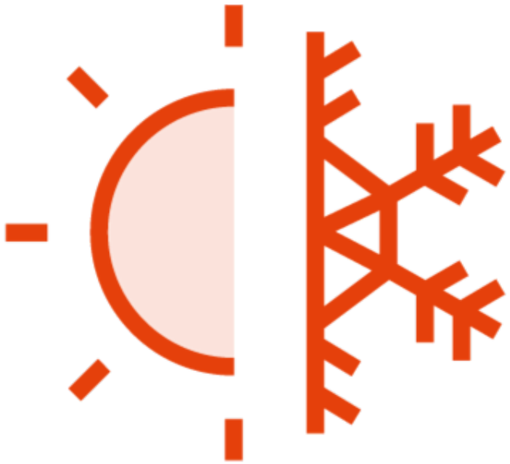
Median

Mode

Other infrequently used measures

- Geometric Mean
- Harmonic Mean

Measures of Dispersion

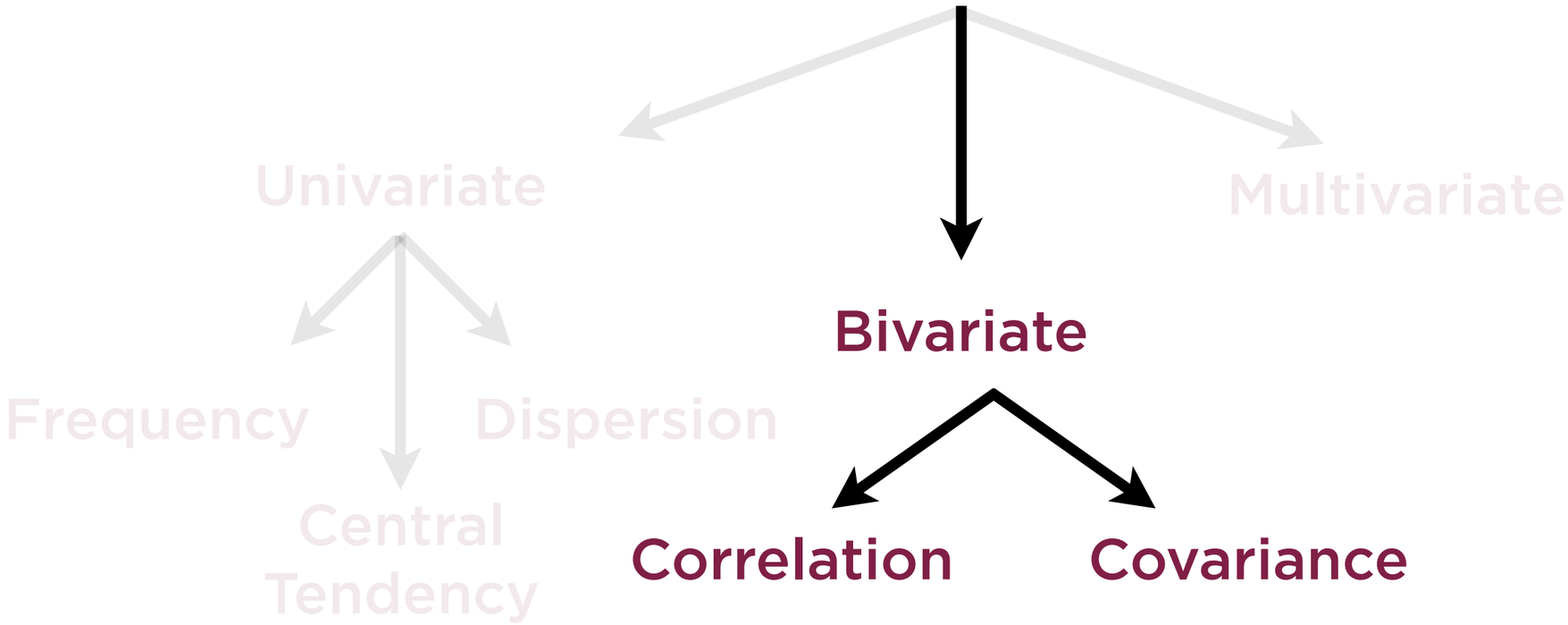


Range (max - min)

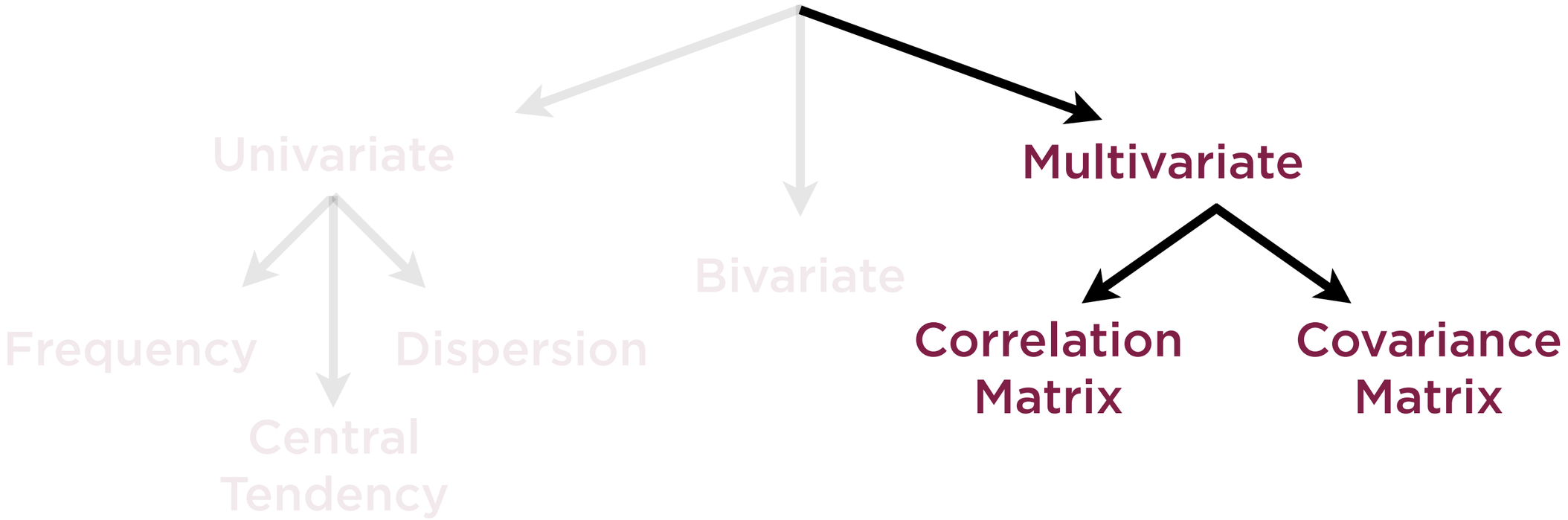
Inter-quartile range (IQR)

Standard deviation and variance

Descriptive Statistics



Descriptive Statistics



Data in One Dimension

Pop quiz: Your thoughtful, fact-based point-of-view on these numbers, please

Mean as Headline

The mean, or average, is the one number that best represents all of these data points

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

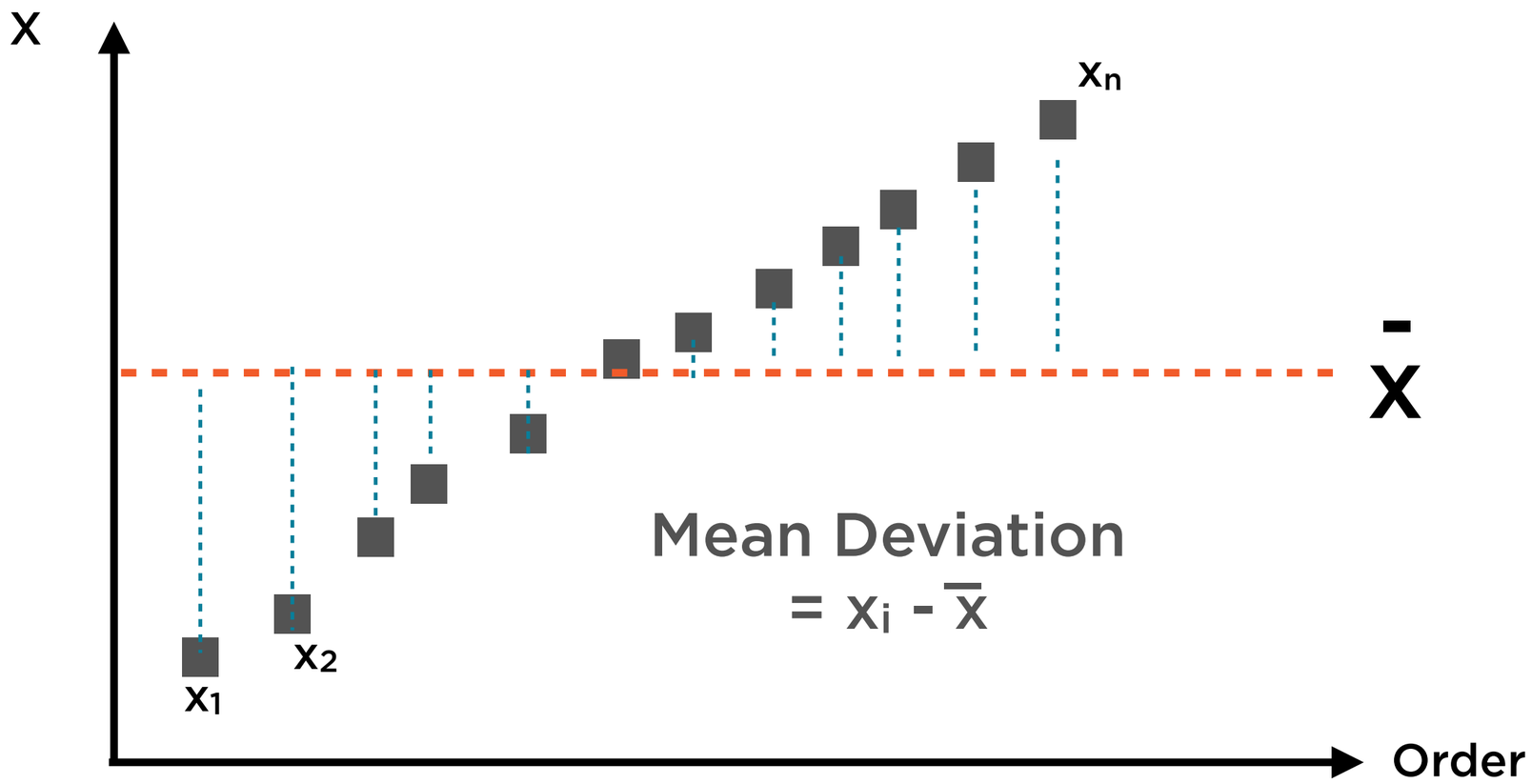
Variation Is Important Too

"Do the numbers jump around?"

$$\text{Range} = X_{max} - X_{min}$$

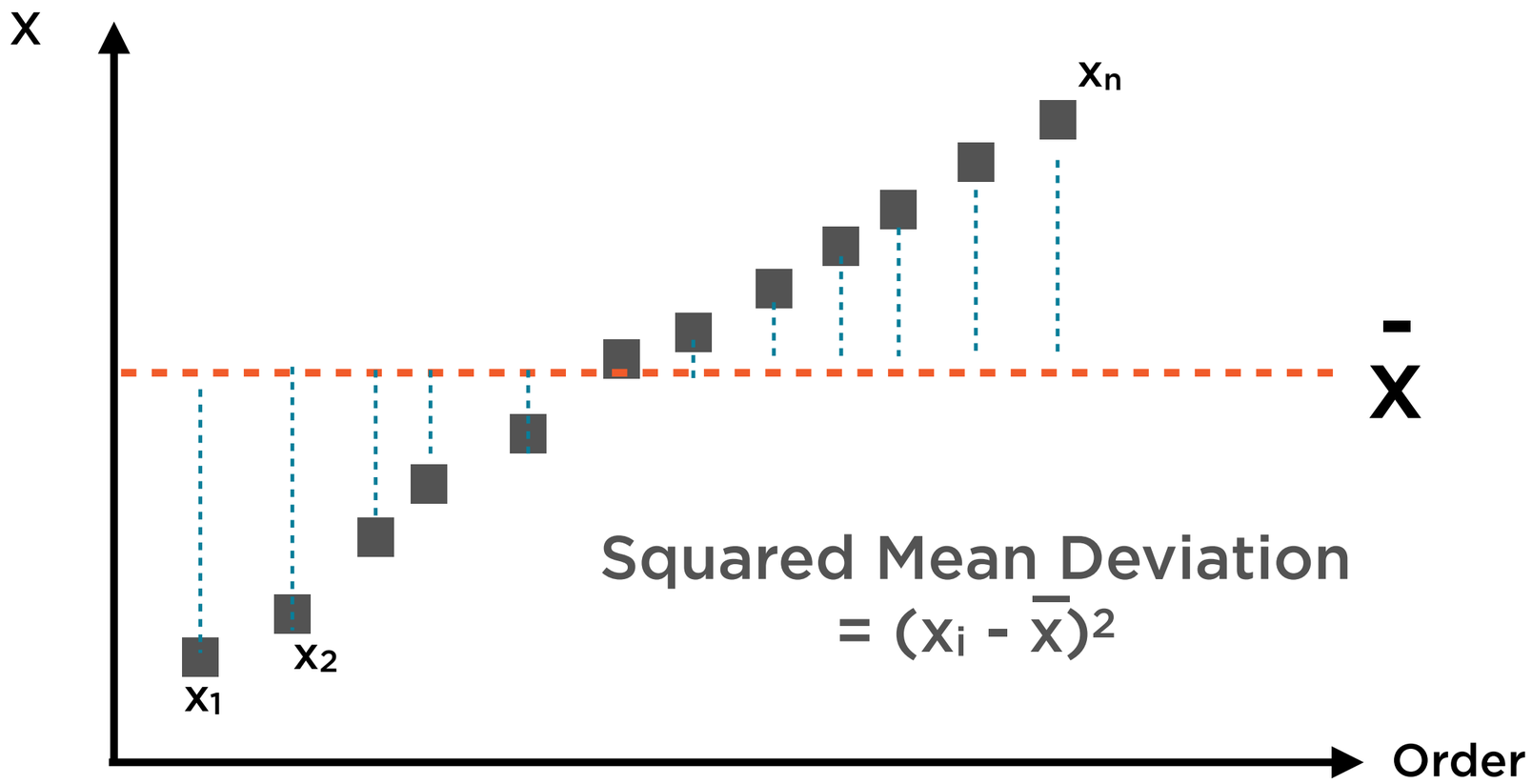
The range ignores the mean, and is swayed by outliers - that's where variance comes in

Variance as Asterisk



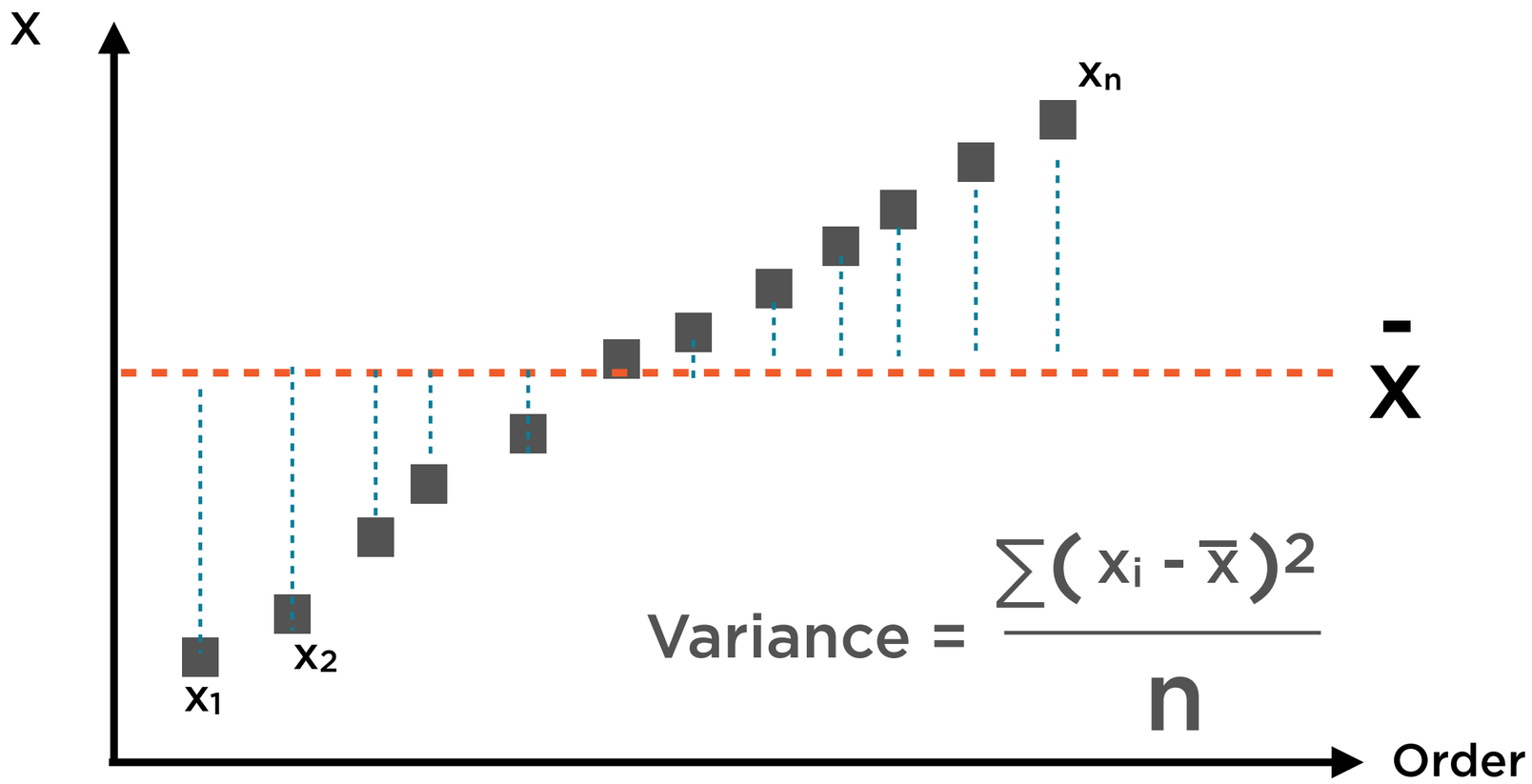
Variance is the second-most important number to summarize this set of data points

Variance as Asterisk



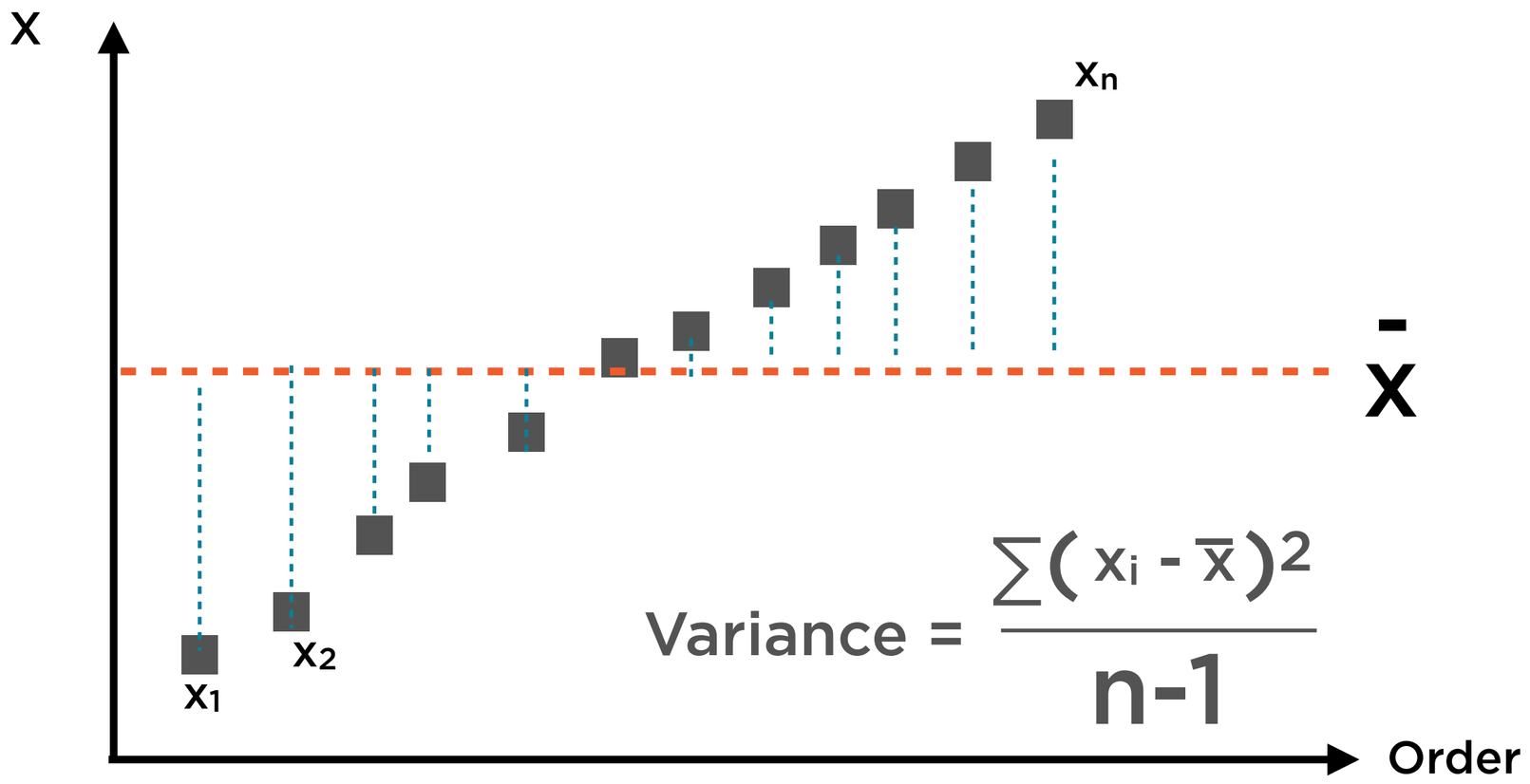
Variance is the second-most important number to summarize this set of data points

Variance as Asterisk



Variance is the second-most important number to summarize this set of data points

Variance as Asterisk



We can improve our estimate of the variance by tweaking the denominator - this is called Bessel's Correction

Mean and Variance

Mean and variance succinctly summarise a set of numbers

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

$$\text{Variance} = \frac{\sum (x_i - \bar{x})^2}{n}$$

Variance and Standard Deviation

Standard deviation is the square root of variance

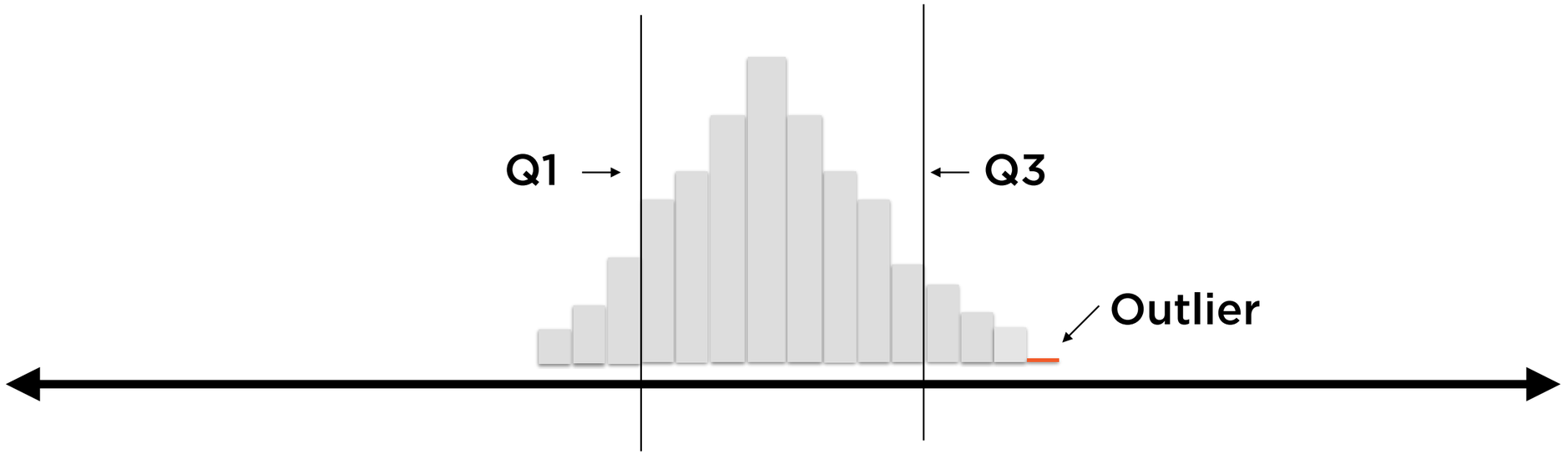
$$\text{Variance} = \frac{\sum(x_i - \bar{x})^2}{n-1}$$

$$\text{Std Dev} = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}}$$

Outliers

Outliers might represent data errors, or genuinely rare points legitimately in dataset

Inter-quartile Range

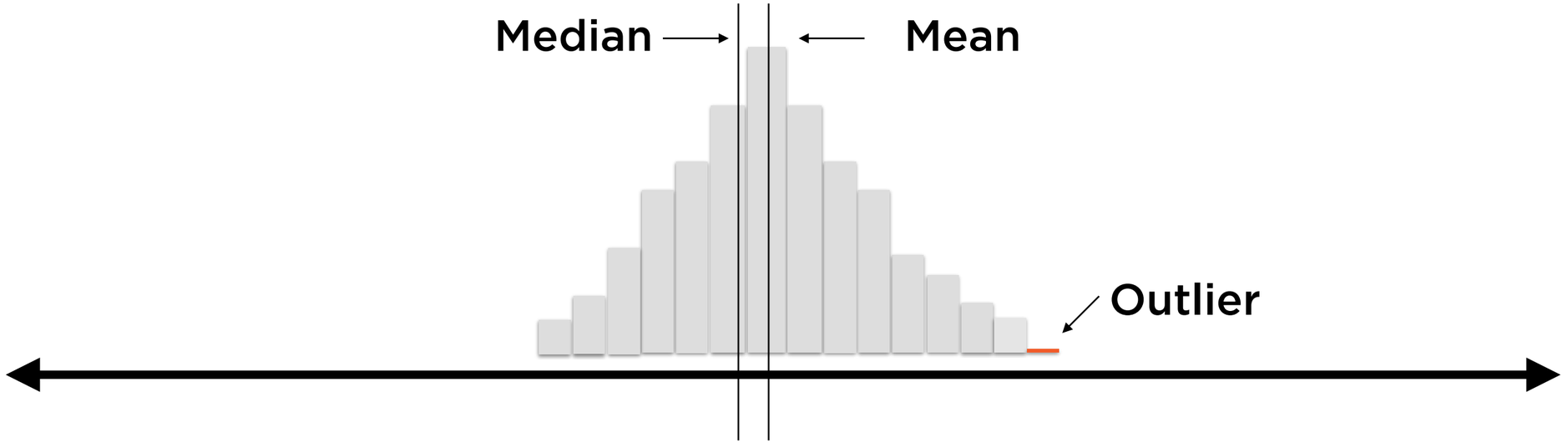


Q3 = 75th percentile: 75% of points smaller than this

Q1 = 25th percentile: 25% of points smaller than this

Inter-quartile Range (IQR) = 75th percentile - 25th percentile

Median



Median = 50th percentile: 50% of points on either side

Unlike mean, median changes little due to outliers

Bivariate Descriptive Statistics

Correlation

Covariance

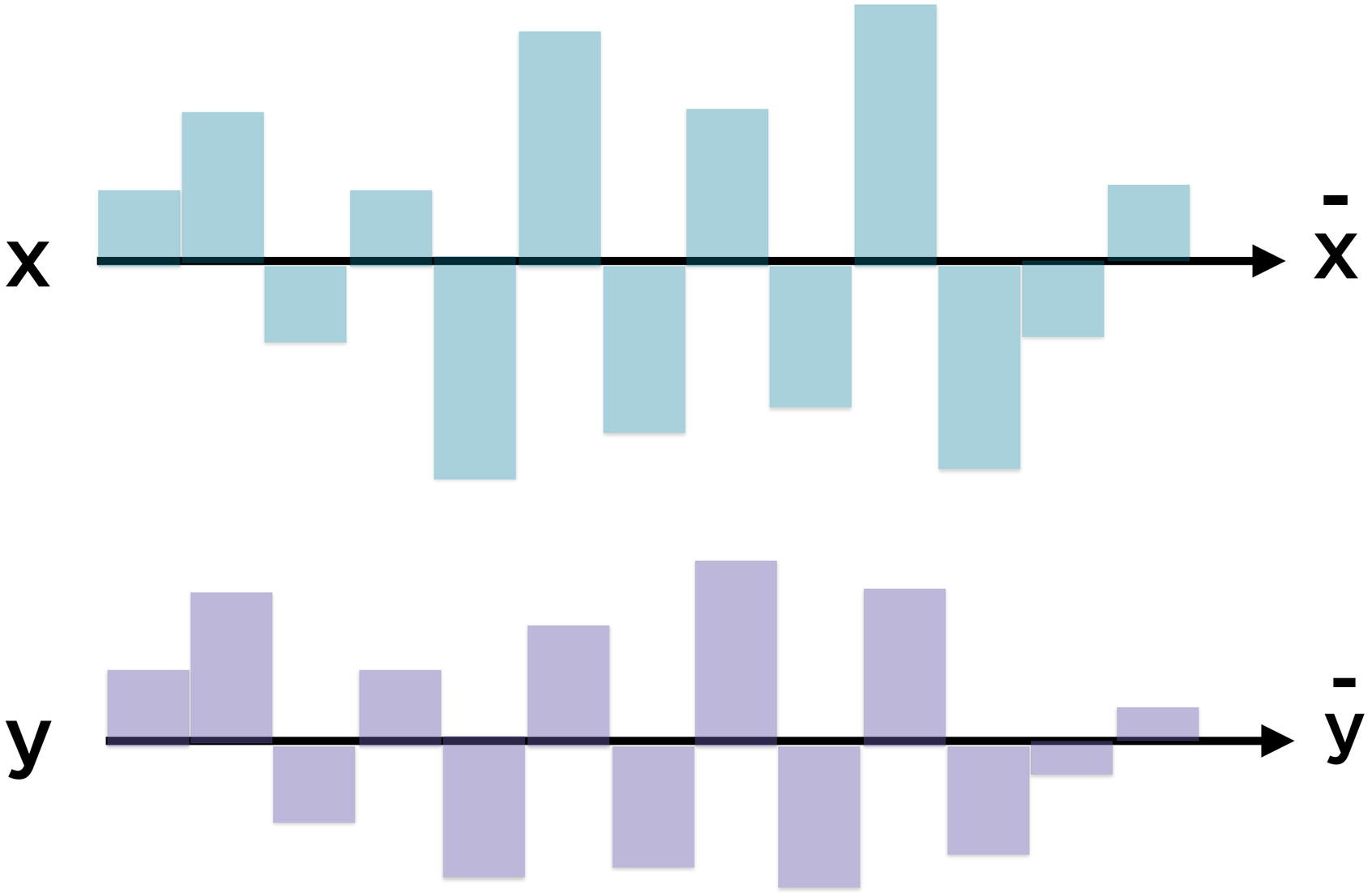
Covariance

Measures relationship between two variables, specifically whether greater values of one variable correspond to greater values in the other.

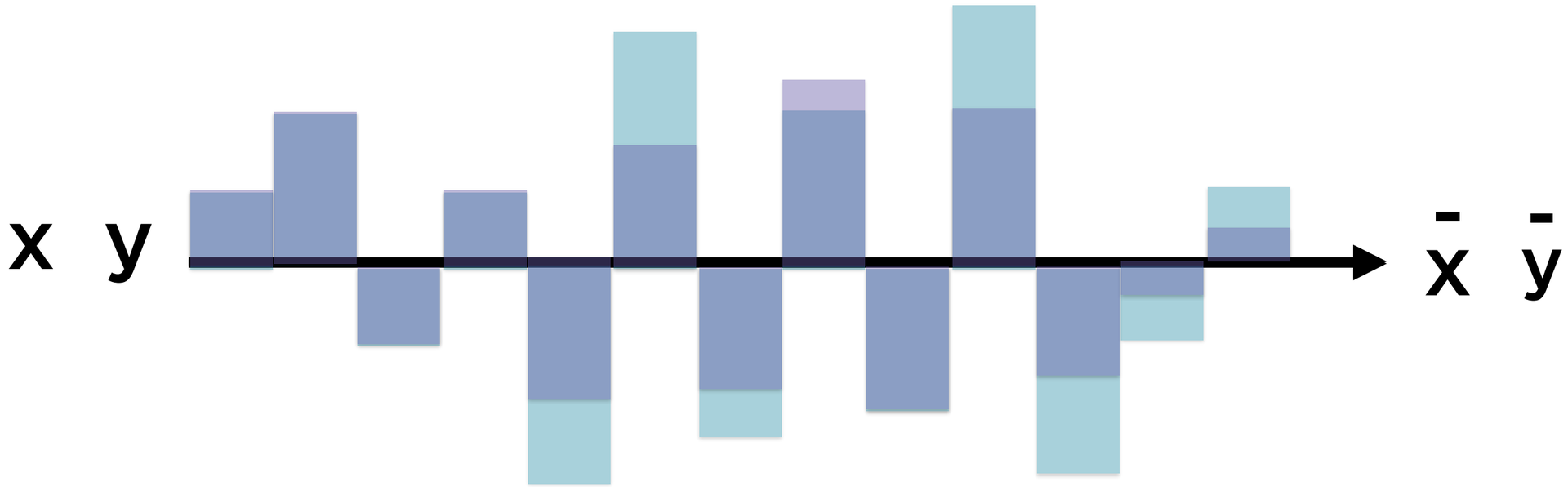
Covariance

Measures relationship between two variables, specifically whether greater values of one variable correspond to greater values in the other.

Intuition: Positive Covariance

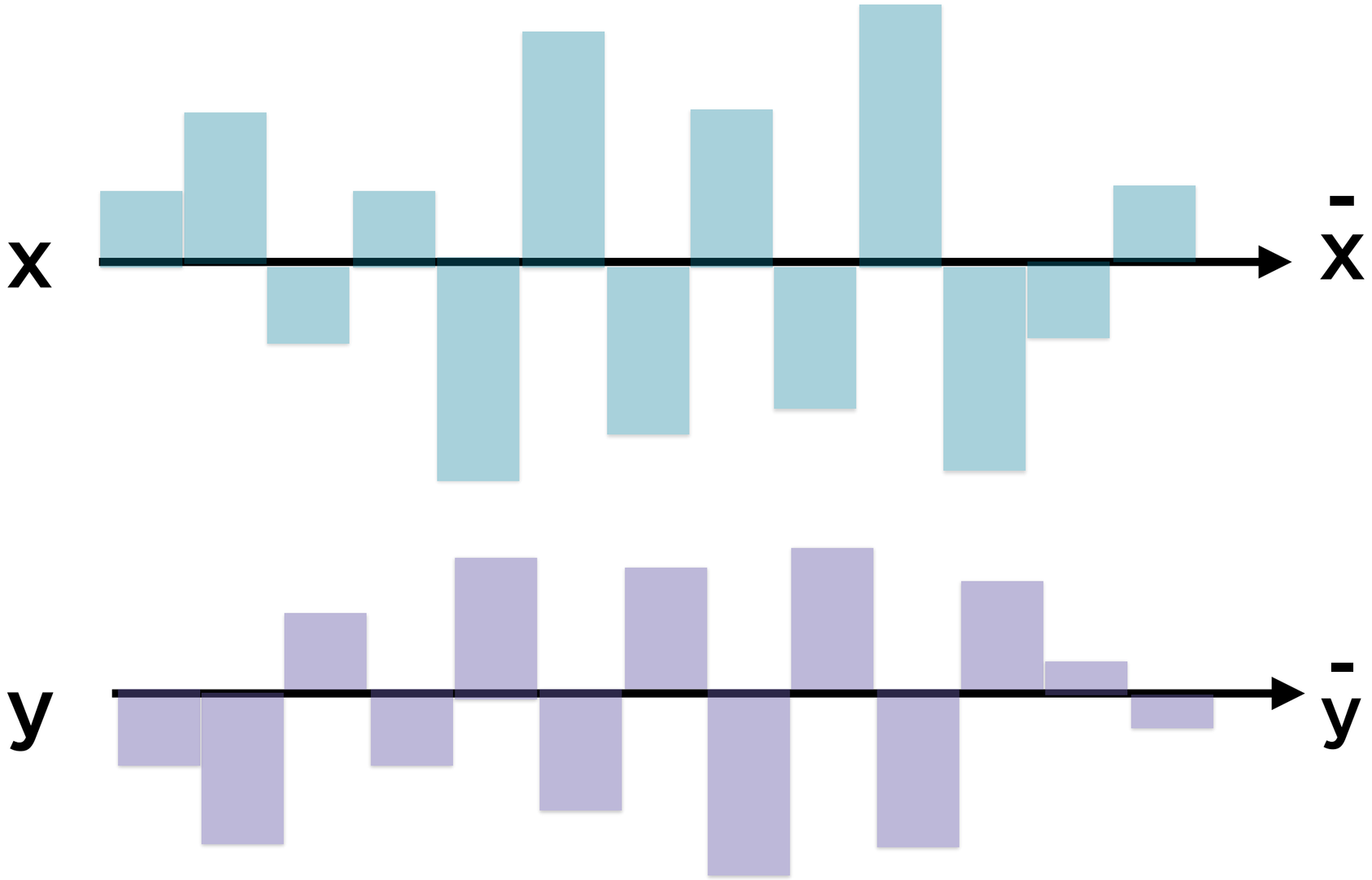


Intuition: Positive Covariance

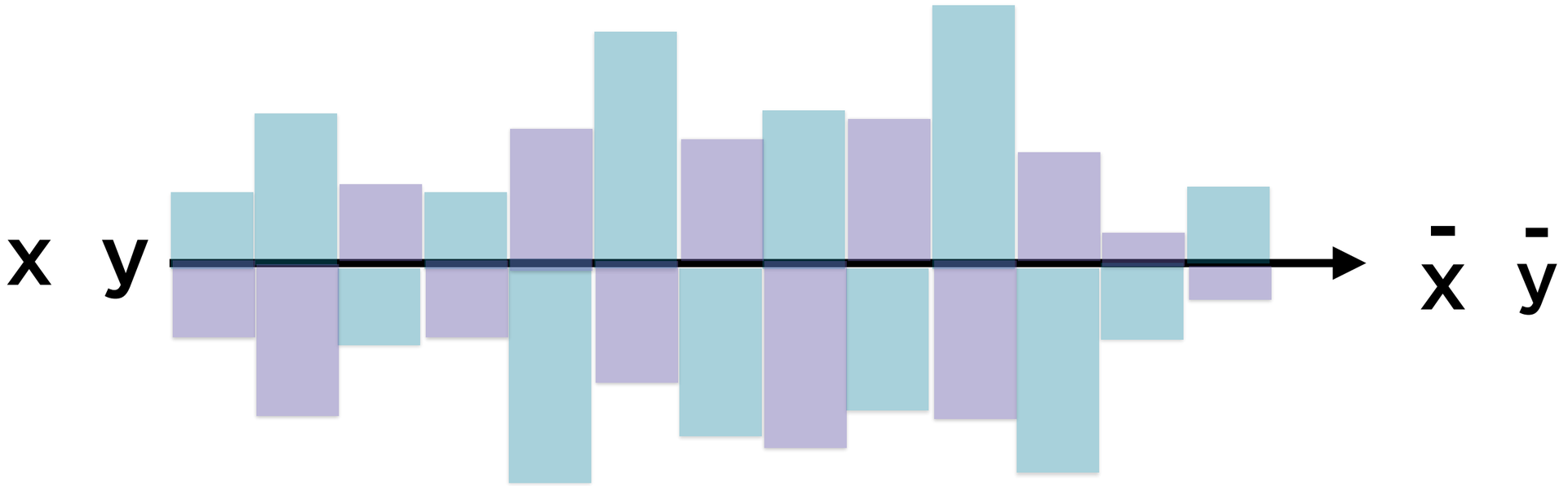


The deviations around the means of the two series are in-sync

Intuition: Negative Covariance



Intuition: Negative Covariance



The deviations around the means of the two series are out-of-sync

Correlation

Similar to covariance; measures whether greater values of one variable correspond to greater values in the other. Scaled to always lie between +1 and -1.

Correlation

Similar to covariance; measures whether greater values of one variable correspond to greater values in the other. Scaled to always lie between +1 and -1.

Correlation

A measure of whether a linear relationship exists between two variables; ranges from +1 (positive linear relationship) to -1 (negative linear relationship). Independent variables exhibit zero correlation.

Correlation

A measure of whether a linear relationship exists between two variables; ranges from +1 (positive linear relationship) to -1 (negative linear relationship). Independent variables exhibit zero correlation.

Correlation

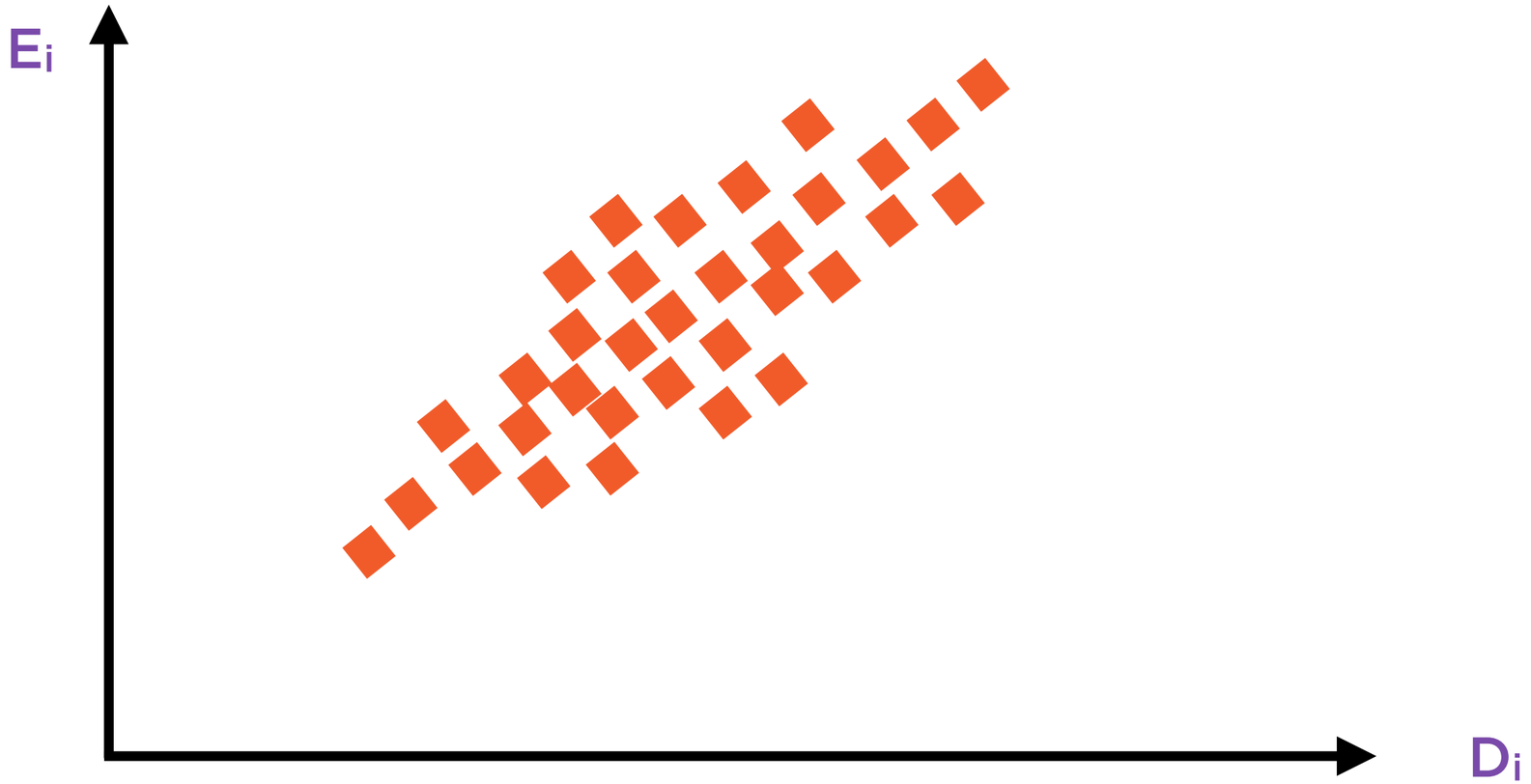
A measure of whether a linear relationship exists between two variables; ranges from +1 (positive linear relationship) to -1 (negative linear relationship).

Independent variables exhibit zero correlation.

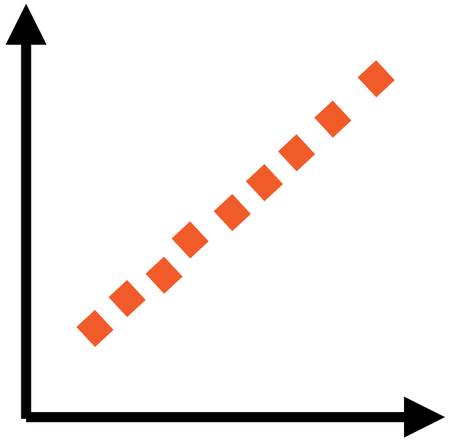
Correlation and Covariance

$$\text{Correlation}(x,y) = \frac{\text{Covariance}(x,y)}{\sqrt{\text{Variance}(x)}\ \sqrt{\text{Variance}(y)}}$$

Correlated Random Variables

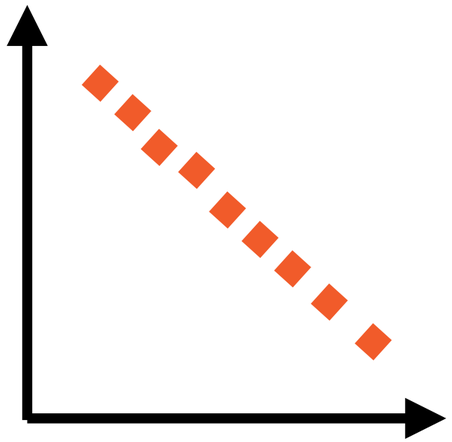


Correlation Captures Linear Relationships



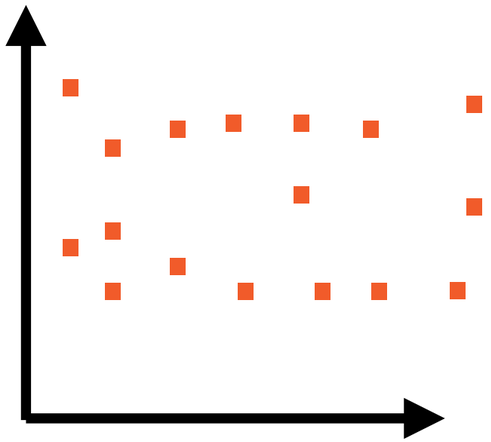
Correlation = +1

As X increases, Y increases linearly



Correlation = -1

As X increases, Y decreases linearly



Correlation = 0

Changes in X independent* of changes in Y

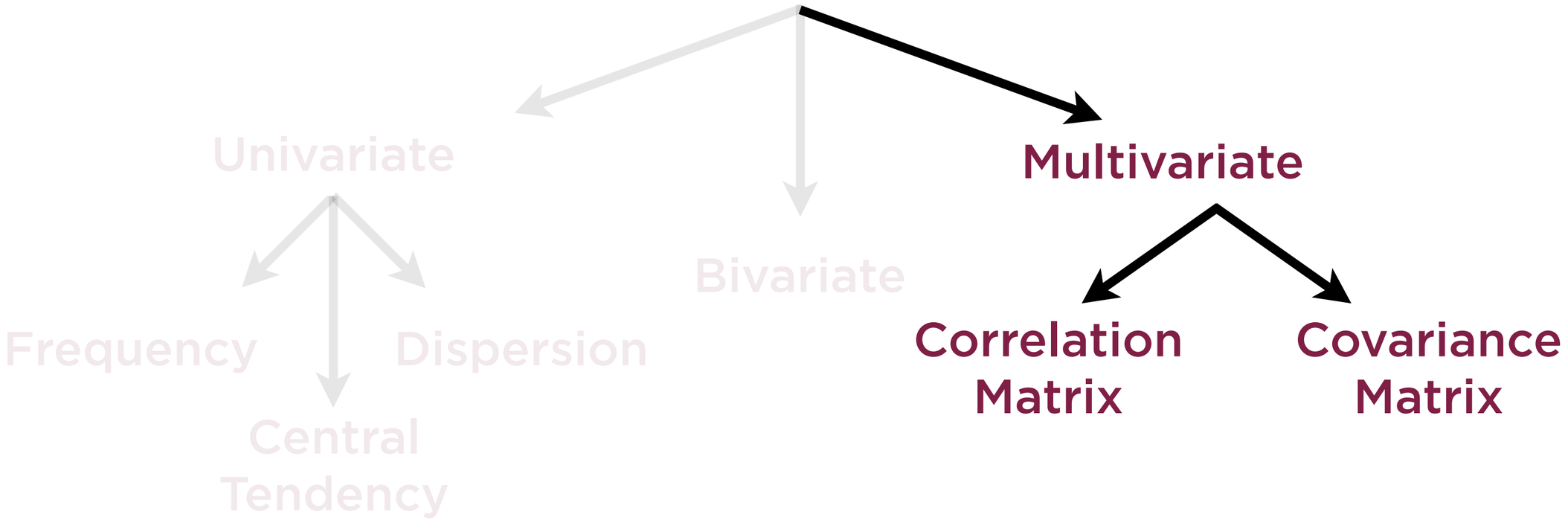
Independent variables have zero covariance and zero correlation

Multivariate Descriptive Statistics

Correlation Matrices

Covariance Matrices

Descriptive Statistics



Demo

Loading, cleaning, and preparing data for exploratory data analysis

Demo

Exploring and visualizing relationships in data

Demo

Calculating and visualizing correlations and linear relationships

Autocorrelation

self

Autocorrelation

Autocorrelation

Measures the relationship between a variable's current value and past value

Autocorrelation

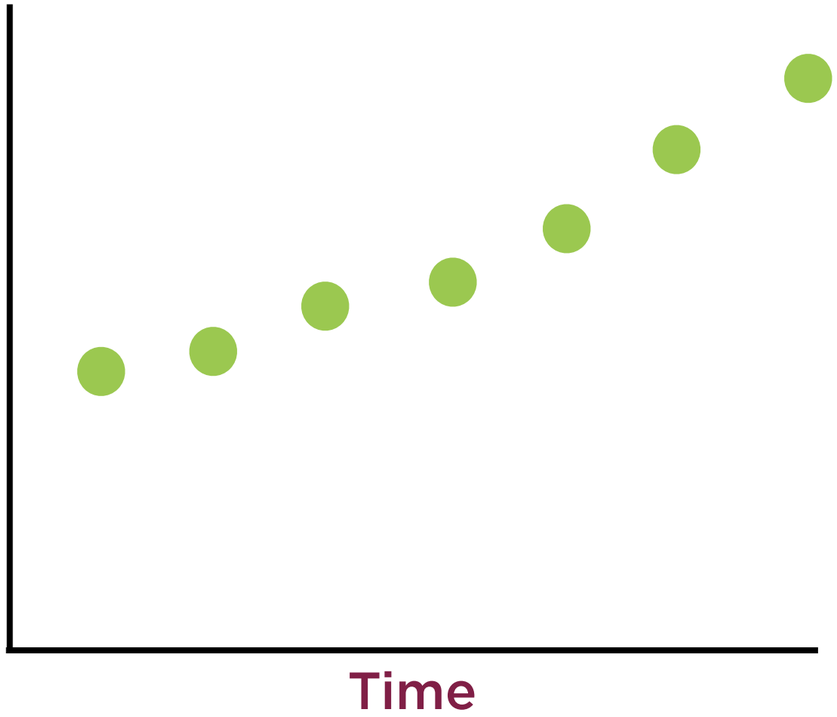
Measures the relationship between a variable's current value and past value

Autocorrelation

Measures the relationship between a variable's current value and past value

Autocorrelation

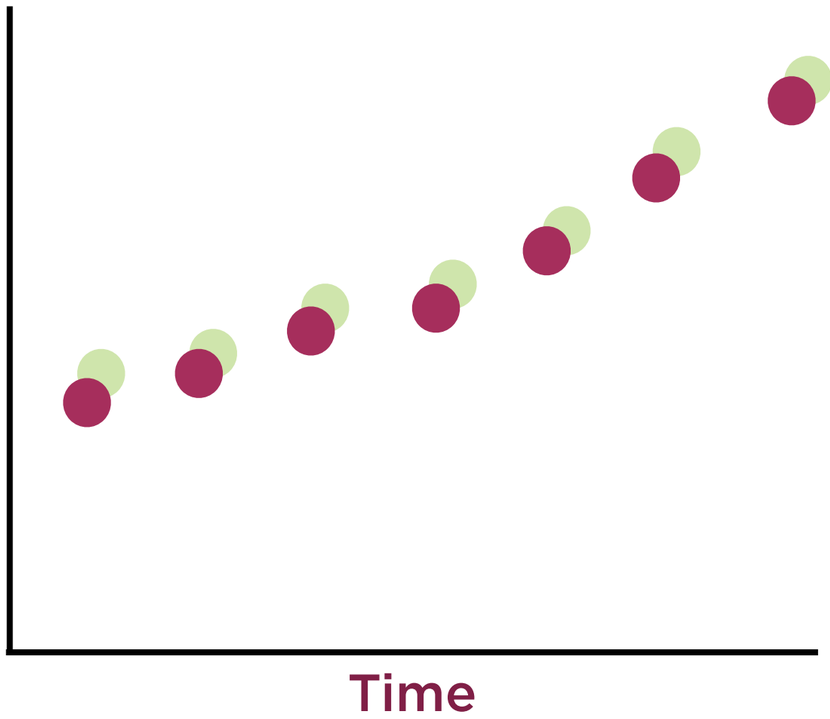
Same time series is used twice



Original form

Autocorrelation

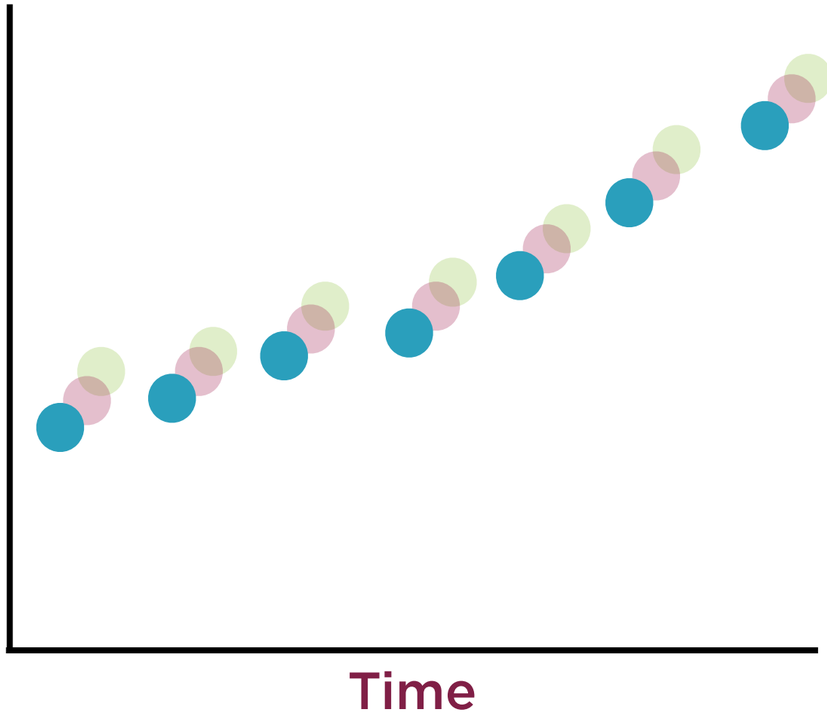
Same time series is used twice



Lagged over one or more time periods

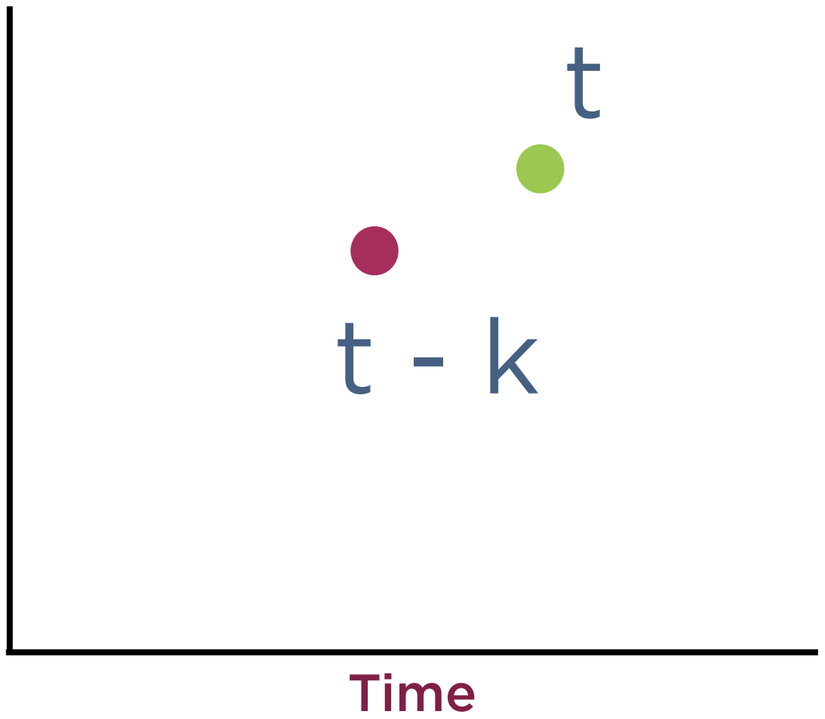
Autocorrelation

Same time series is used twice



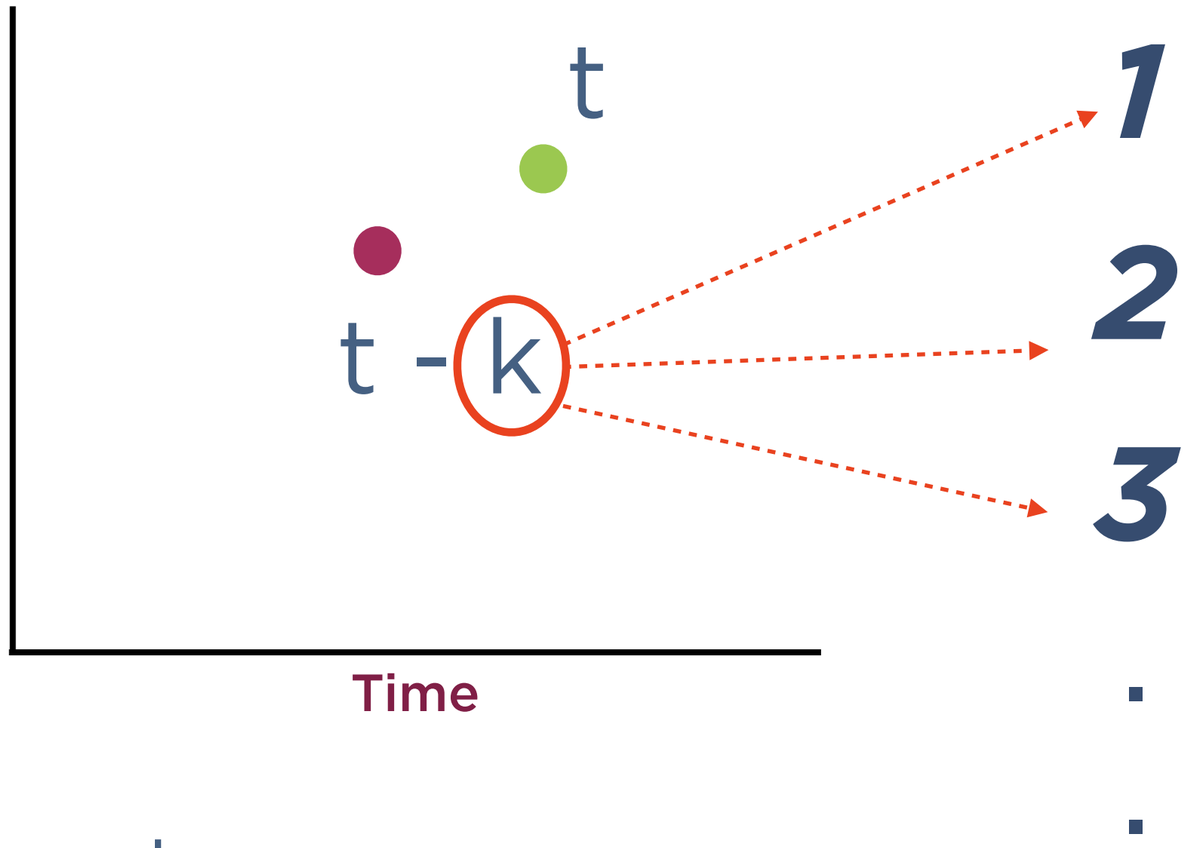
Lagged over one or more time periods

Autocorrelation



Lagged over one or more time periods

Autocorrelation



Lagged over one or more time periods

Autocorrelation

Ranges between

Autocorrelation

Perfect positive correlation

1

Autocorrelation

-1

Perfect negative correlation

Correlation

The measure of the relationship between two items or variables

Autocorrelation

More likely

Today

Tomorrow

Autocorrelation

Less likely

Today

Tomorrow

Demo

Calculating and visualizing autocorrelations with time lags

Demo

Exploring different visualizations to learn relationships in data

Summary

Common statistical relationships

Univariate, bivariate and multivariate relationships

Mean, standard deviation and variance

Covariance and correlation

Autocorrelation