Interpreting Data Using Descriptive Statistics with Python

UNDERSTANDING DESCRIPTIVE STATISTICS

Janani Ravi
CO-FOUNDER, LOONYCORN
www.loonycorn.com

Overview

Descriptive statistics are used to explore and describe data

Measures of central tendency

Measures of dispersion

Confidence intervals of a measure

Skewness and kurtosis

Bivariate measures such as covariance and correlation

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

Understanding descriptive statistics

Working with descriptive statistics using Pandas

Working with descriptive statistics using SciPy and Statsmodels

Statistics in Understanding Data

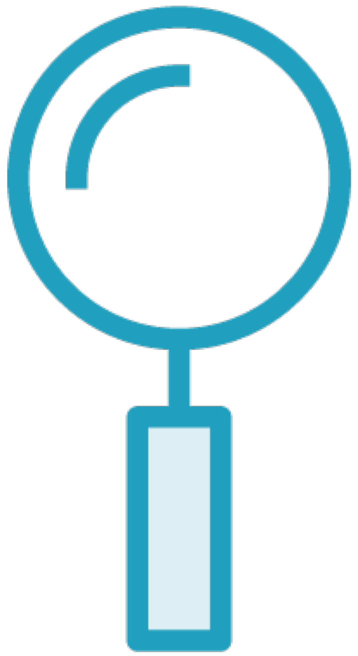
“There are two kinds of statistics, the kind you look up and the kind you make up”

Rex Stout

Statistics

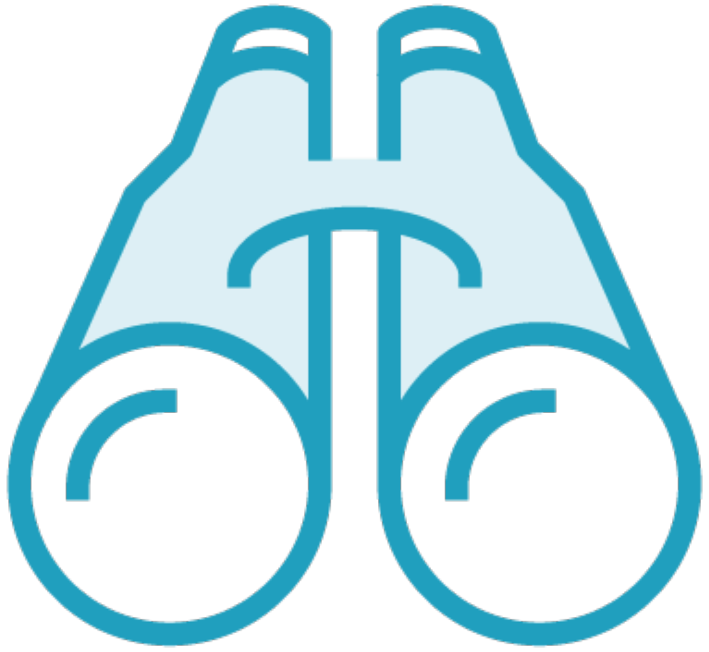
A branch of mathematics that deals with collecting, organizing, analyzing, and interpreting data

Two Sets of Statistical Tools



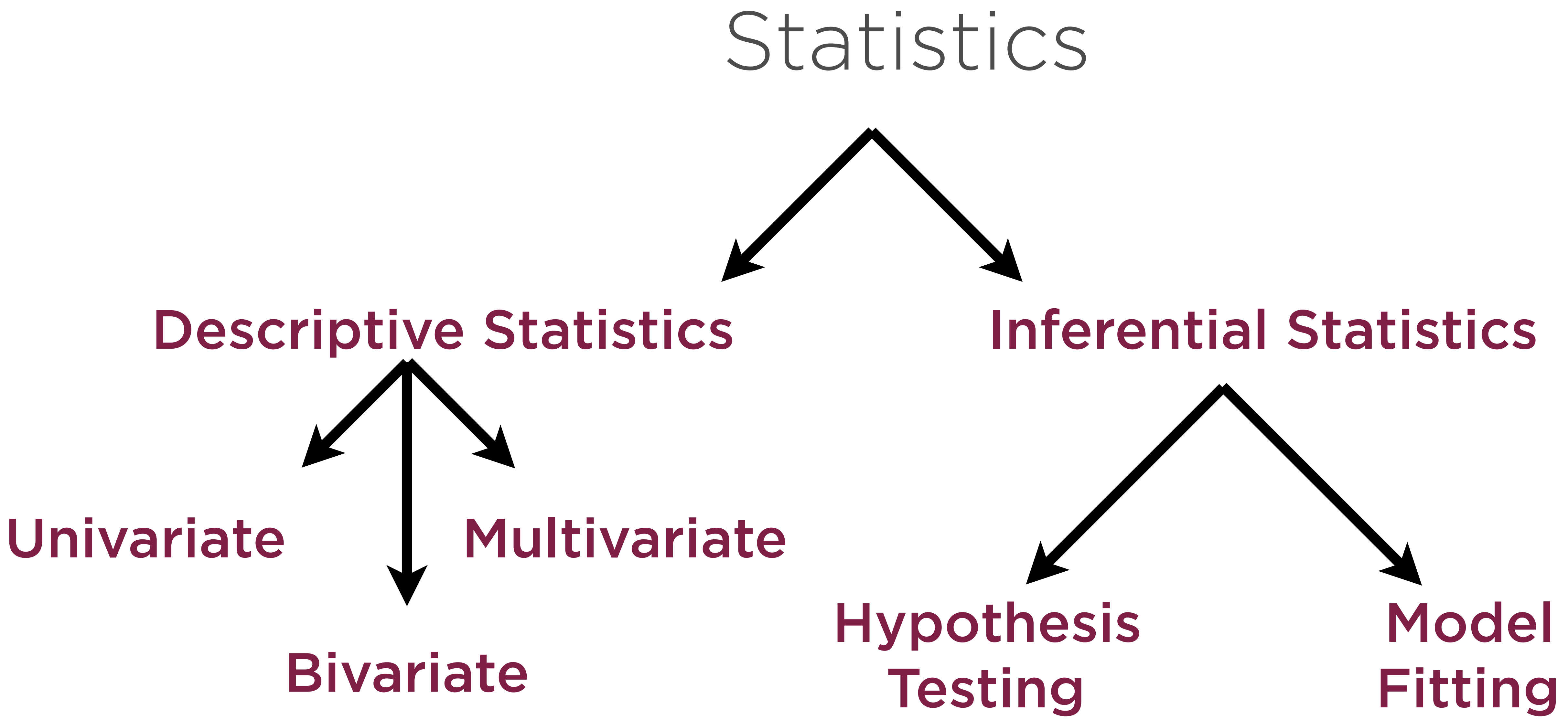
Descriptive Statistics

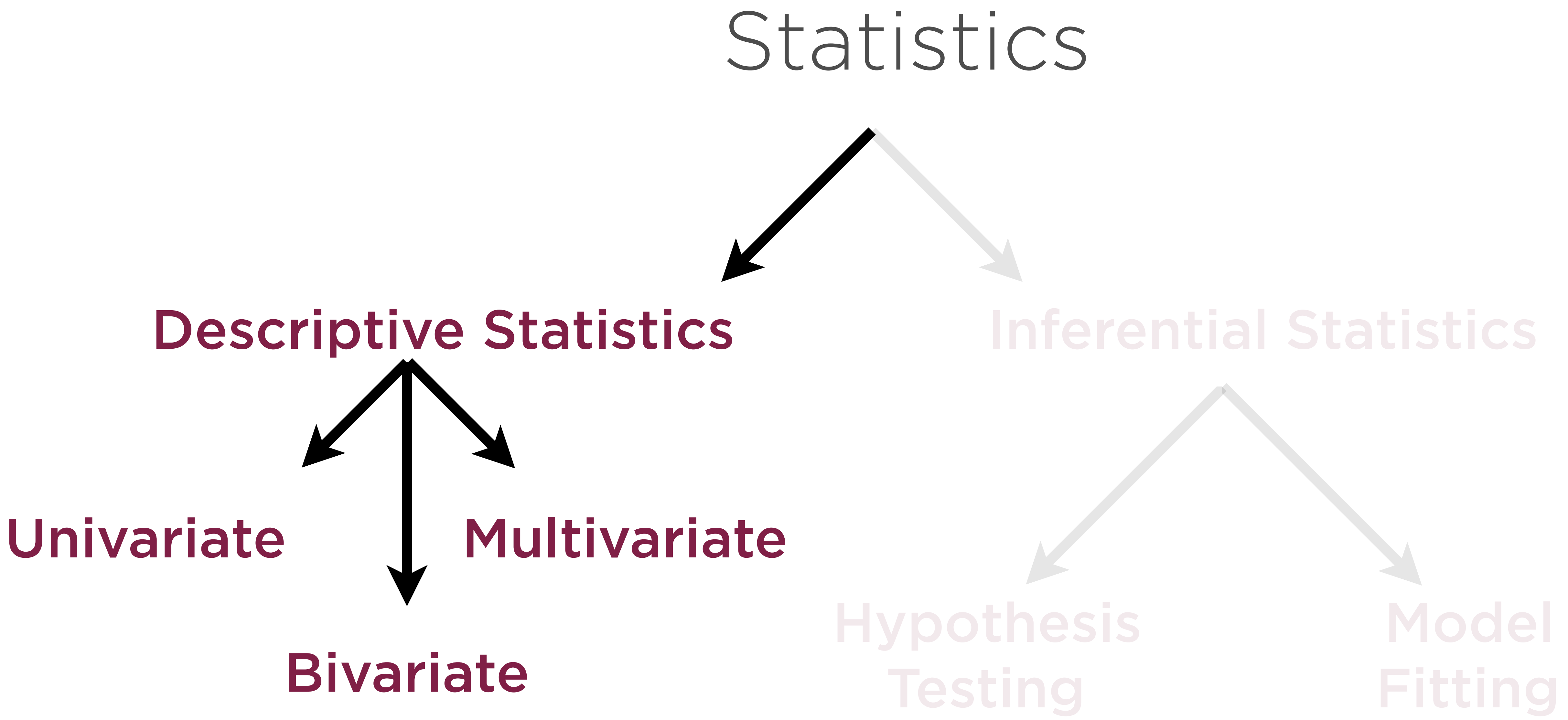
Identify important elements in a dataset



Inferential Statistics

Explain those elements via relationships with other elements


Statistics
Descriptive Statistics
Inferential Statistics
Univariate
Bivariate
Multivariate
Hypothesis Testing
Model Fitting


Statistics
Descriptive Statistics
Inferential Statistics
Univariate
Bivariate
Multivariate
Hypothesis Testing
Model Fitting

Descriptive Statistics

Summarize data as it is

Do not posit any hypothesis about data

Do not try to fit models to data

Descriptive Statistics

Very important initial step

Often neglected

Detect outliers

Plan how to prepare data

Precursor to feature engineering

Descriptive Statistics

Related subjects

- Exploratory data analysis
- Descriptive visualization

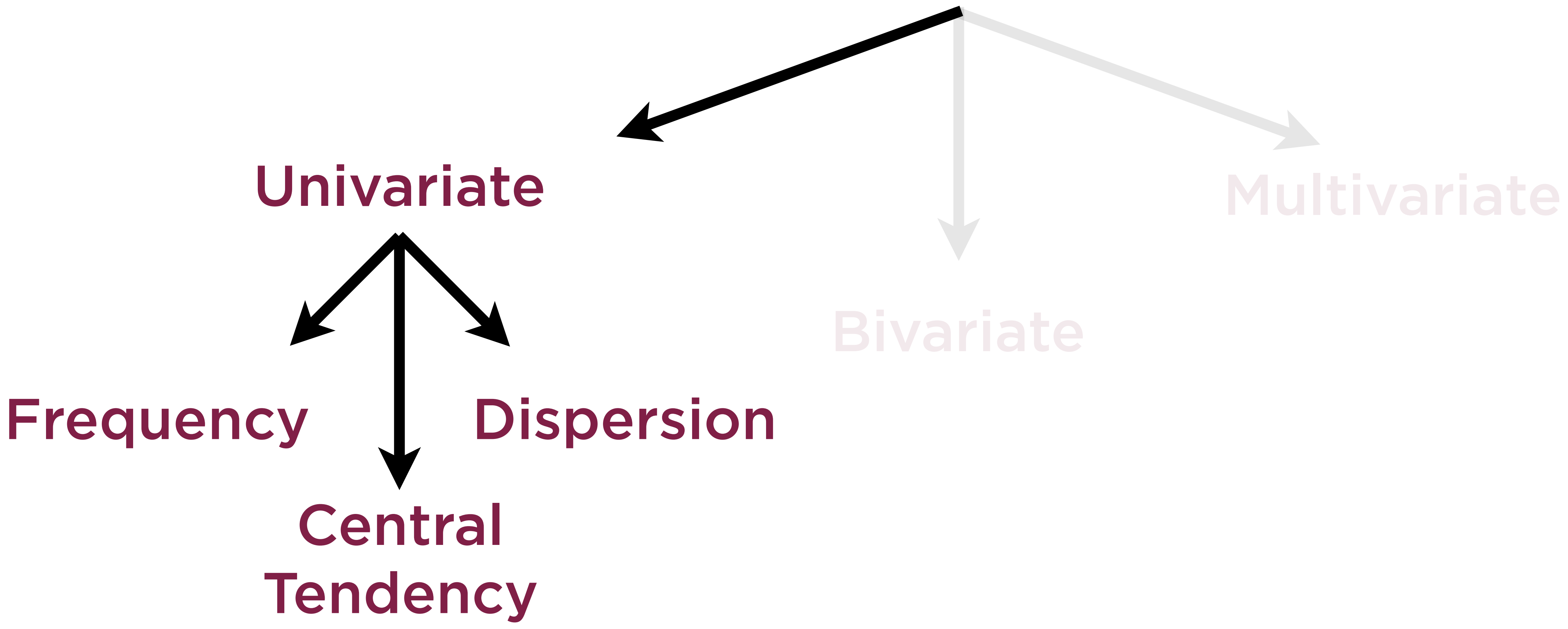
Descriptive Statistics

Univariate

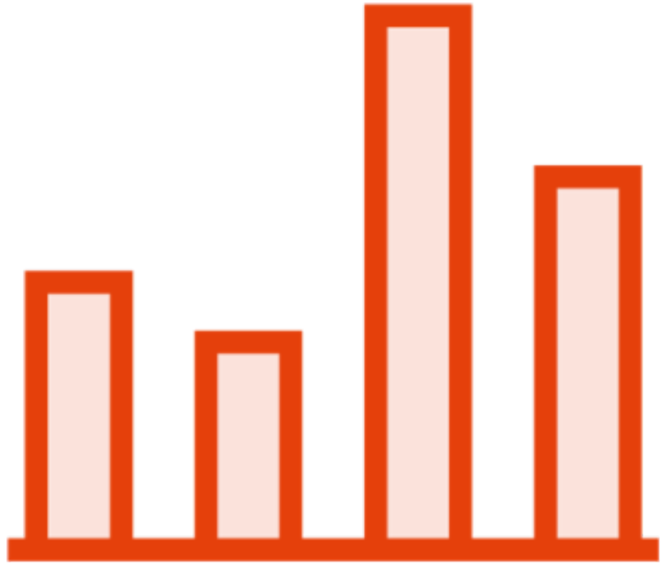
Bivariate

Multivariate

Descriptive Statistics



Measures of Frequency



Frequency tables

Histograms

Measures of Central Tendency

Average (Mean)

Median

Mode

Other infrequently used measures

- Geometric Mean
- Harmonic Mean

Mean

Single best value to represent data

Need not actually be data point itself

Considers every point in data

Discrete as well as continuous data

Vulnerable to outliers

Mean of a Dataset

Data	60	20	10	40	50	30

Mean of a Dataset

Data	60	20	10	40	50	30

$$\bar{x} = \frac{\sum x_i}{n} = \frac{60 + 20 + 10 + 40 + 50 + 30}{6}$$

Mean of a Dataset

Data	60	20	10	40	50	30

$$\bar{x} = \frac{\sum x_i}{n} = \frac{60 + 20 + 10 + 40 + 50 + 30}{6}$$

Mean	35

Impact of Outliers

Data	60	20	10	40	50	30	1000

$$\bar{x} = \frac{\sum x_i}{n} = \frac{60 + 20 + 10 + 40 + 50 + 30 + 1000}{7}$$

Impact of Outliers

Data	60	20	10	40	50	30	1000

$$\bar{x} = \frac{\sum x_i}{n} = \frac{60 + 20 + 10 + 40 + 50 + 30 + 1000}{7}$$

Mean	172.85

Median

Value such that 50% of data on either side

Sort data, then use middle element

For even number of data points, average two middle elements

Median

More robust to outliers than mean

However does not consider every data point

Makes sense for ordinal data (data that can be sorted)

Median of a Dataset

Data	60	20	10	40	50	30

Median of a Dataset

Data	60	20	10	40	50	30

Ordered Data	10	20	30	40	50	60

Even number of data points - average middle two elements

Median of a Dataset

Ordered Data					
10	20	30	40	50	60

Even number of data points - average middle two elements

Middle 2 elements					
10	20	30	40	50	60

Median
35

Impact of Outliers

Data	60	20	10	40	50	30	1000

Impact of Outliers

Data	60	20	10	40	50	30	1000
Ordered Data	10	20	30	40	50	60	1000

Odd number of data points - simply consider middle element

Impact of Outliers

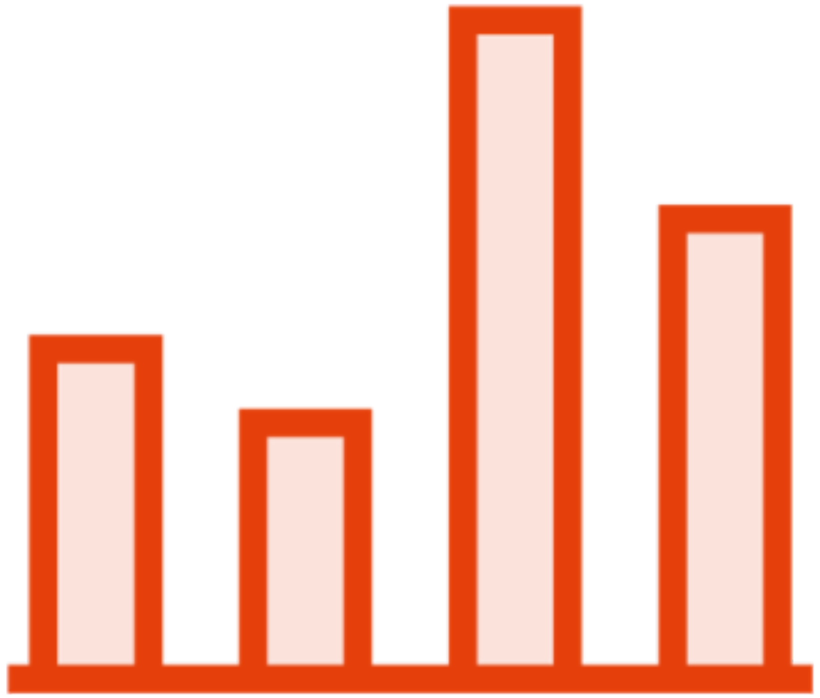
Ordered Data						
10	20	30	40	50	60	1000

Odd number of data points - simply consider middle element

Middle element						
10	20	30	**40**	50	60	1000

Median
40

Mode



Most frequent value in dataset

Highest bar in histogram

Winner in elections

Typically used with categorical data

Mode of a Dataset

Candidate	Alice	Bob	Charles	Denise	Edgar	Fred
Votes	60	20	10	40	50	30

Mode of a Dataset

Candidate	Alice	Bob	Charles	Denise	Edgar	Fred
Votes	60	20	10	40	50	30

Mode represents the most frequent value in the data

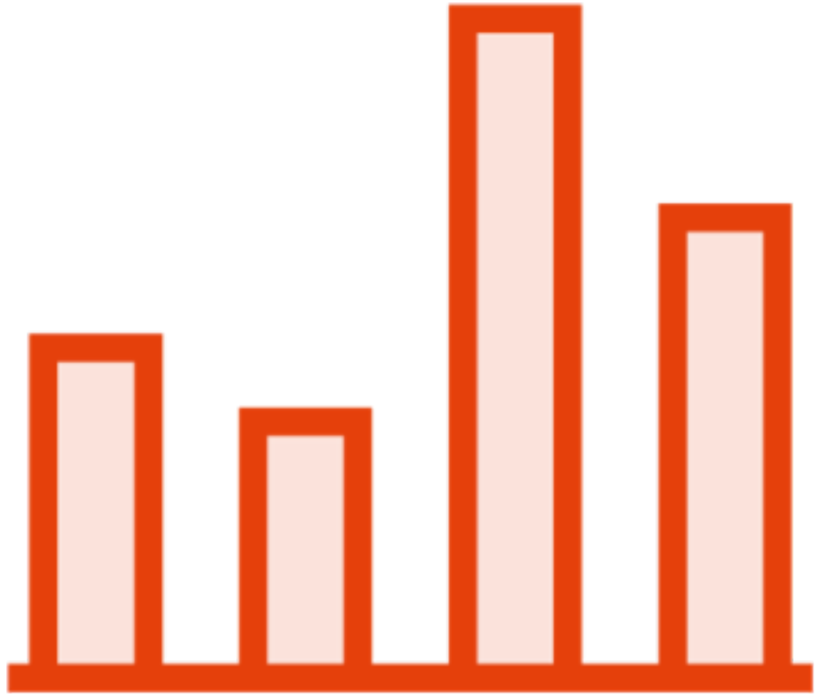
Mode of a Dataset

Candidate	Alice	Bob	Charles	Denise	Edgar	Fred
Votes	60	20	10	40	50	30

Mode represents the most frequent value in the data

Mode	60

Mode

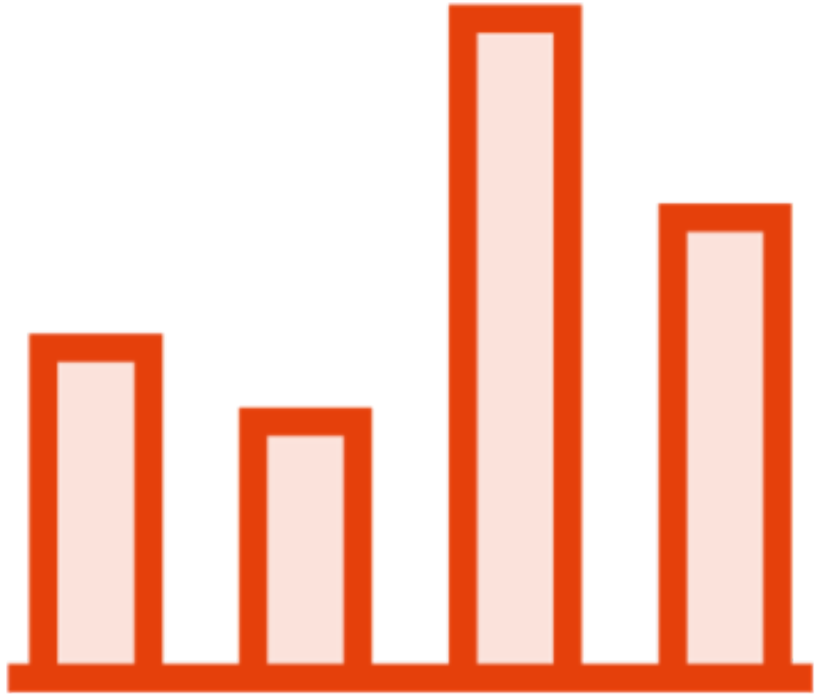


Unlike mean or median, mode need not be unique

Not great for continuous data

Continuous data needs to be discretized and binned first

Other Measures of Central Tendency



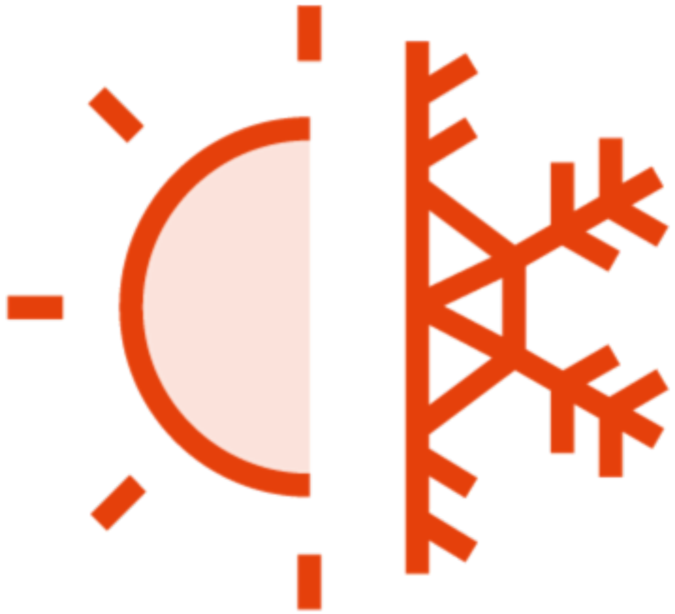
Geometric mean

- Great for summarizing ratios
- Compound Annual Growth Rate (CAGR)

Harmonic mean

- Great for summarizing rates
- Resistors in parallel
- P/E ratios in finance

Measures of Dispersion



Range (max - min)

Inter-quartile range (IQR)

Standard deviation and variance

Univariate Descriptive Statistics

Measures of Frequency

Measures of Central Tendency

Measures of Dispersion

Mean, Variance, and Standard Deviation

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

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

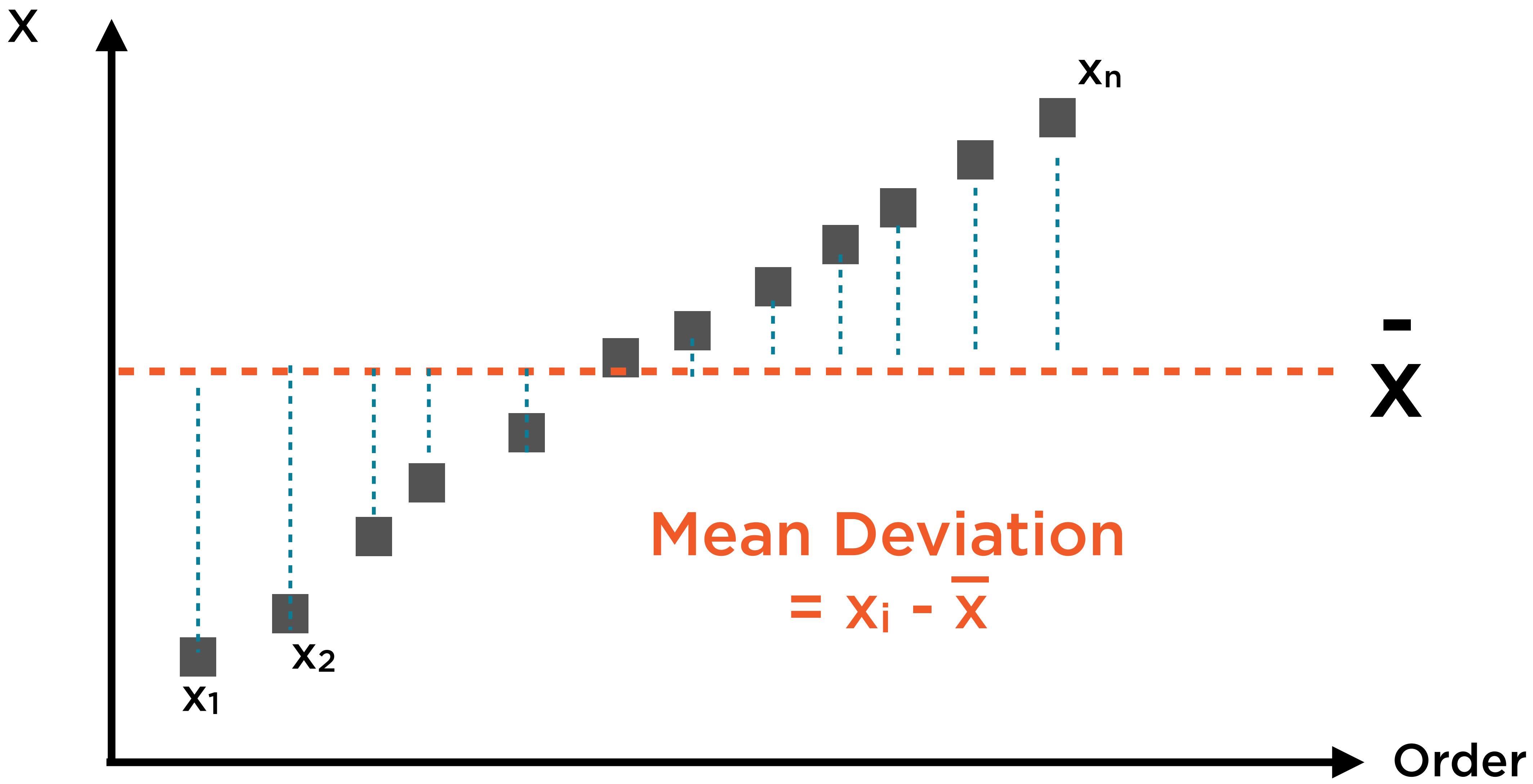
Variation Is Important Too

"Do the numbers jump around?"

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

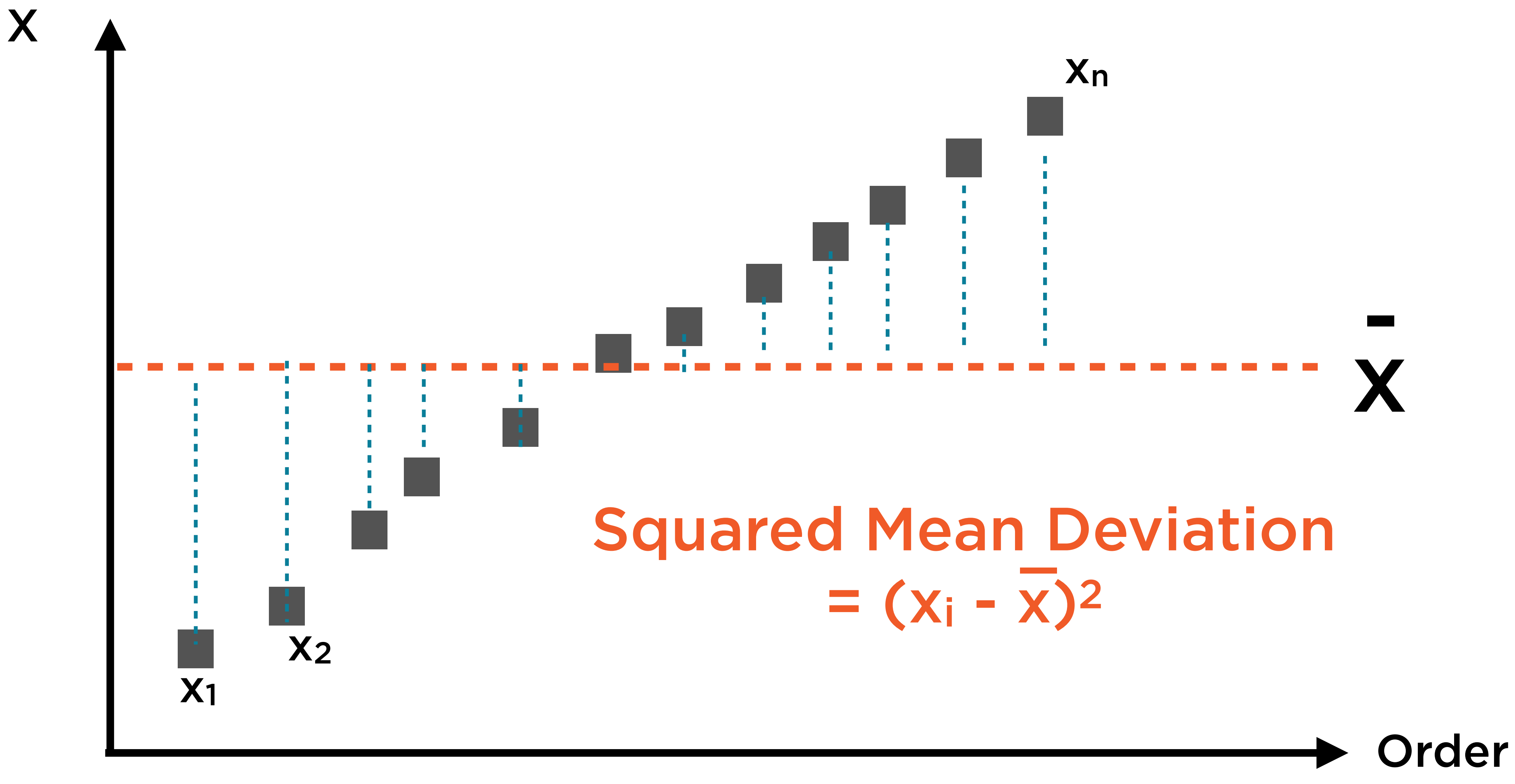
The range ignores the mean, and is swayed by outliers - that's where variance comes in

Variance as Asterisk



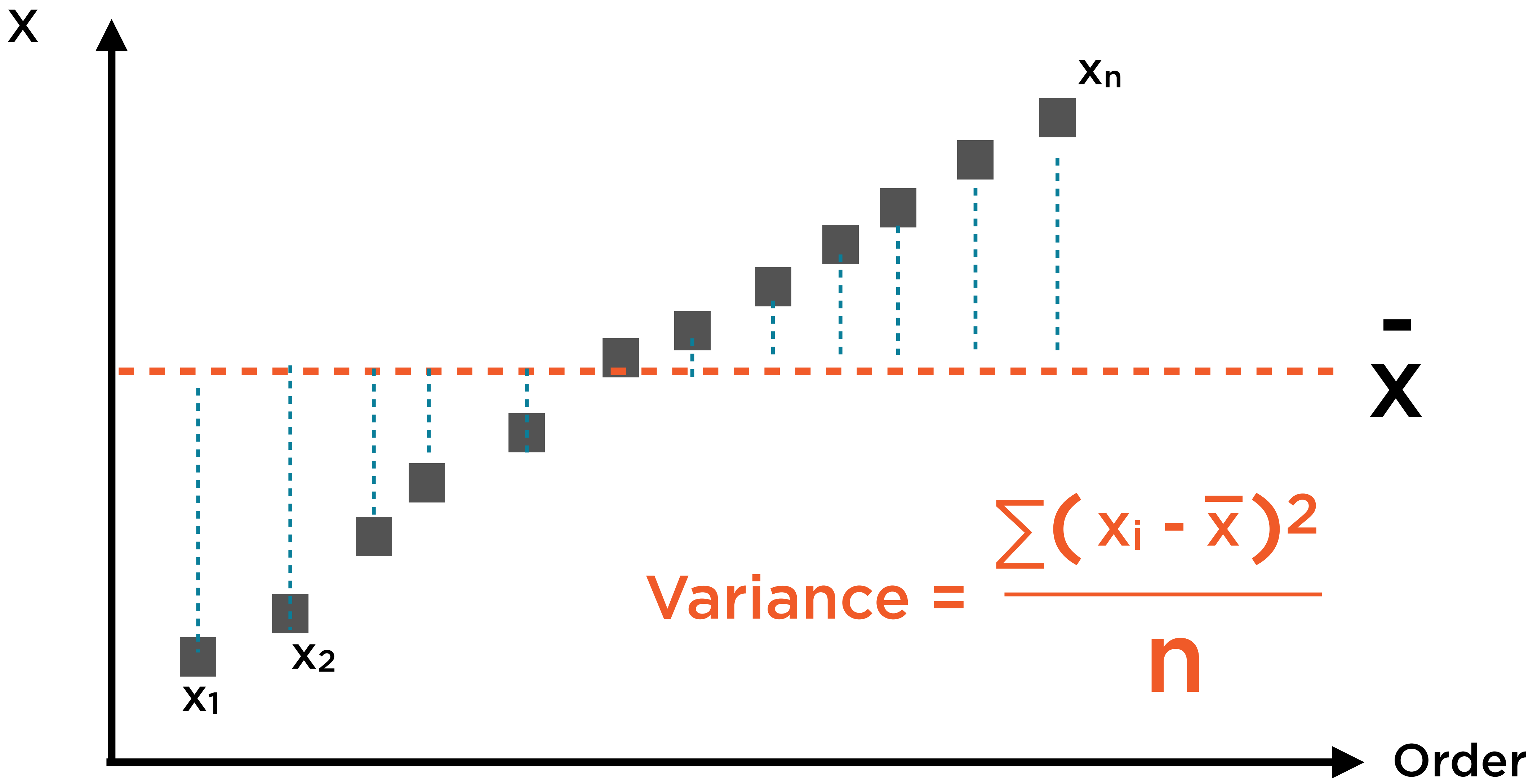
Variance is the second-most important number to summarize this set of data points

Variance as Asterisk



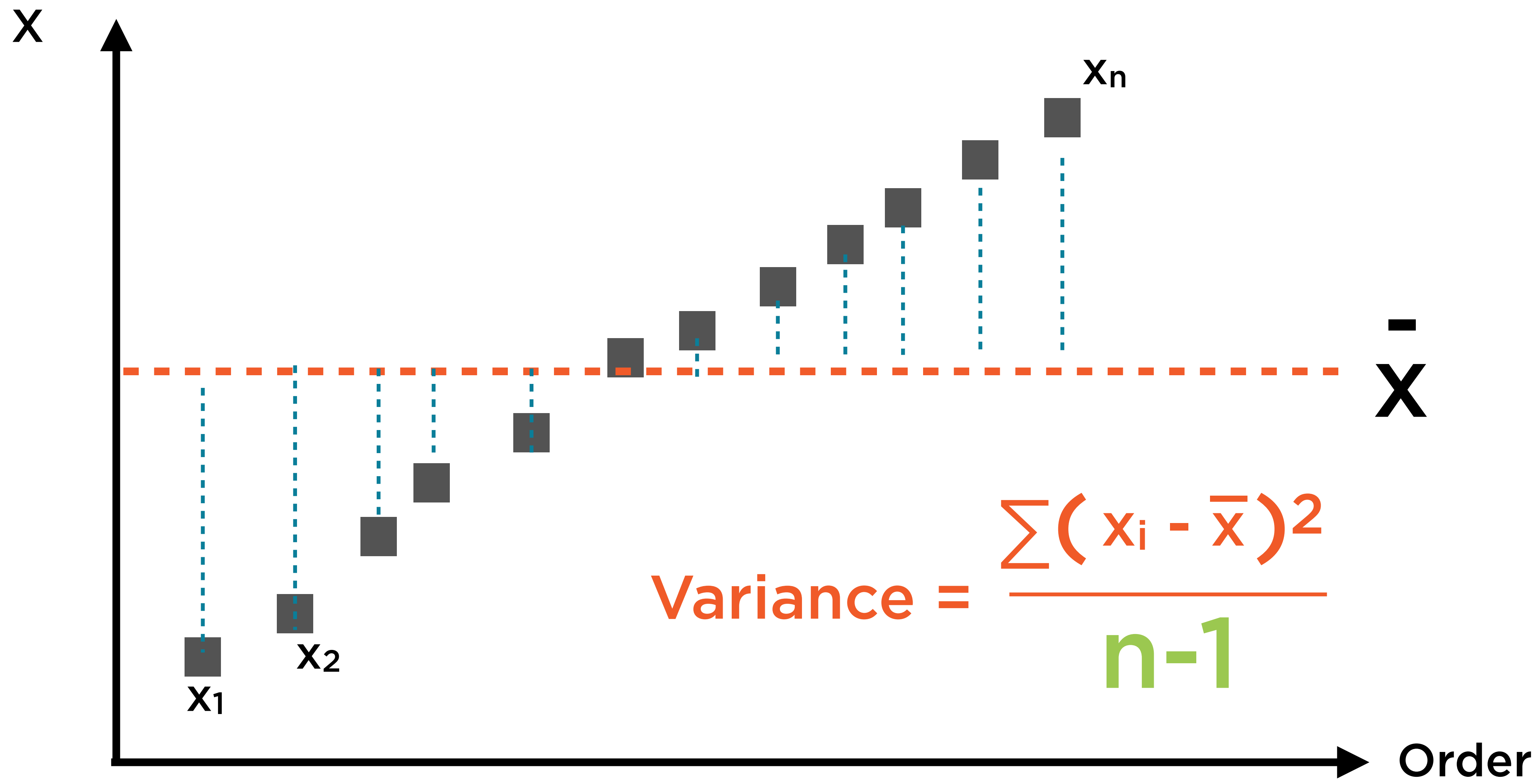
Variance is the second-most important number to summarize this set of data points

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 summarize a set of numbers

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

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

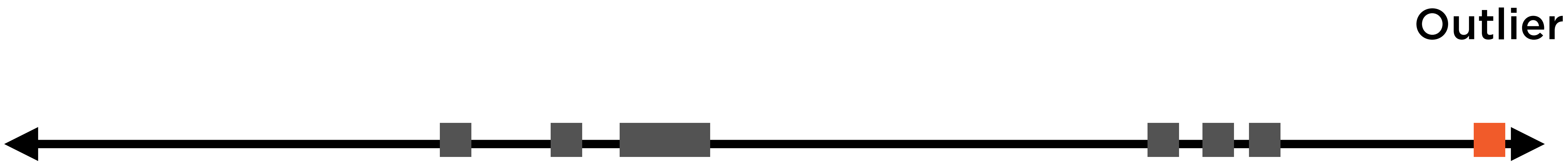
Variance and Standard Deviation

Standard deviation is the square root of variance

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

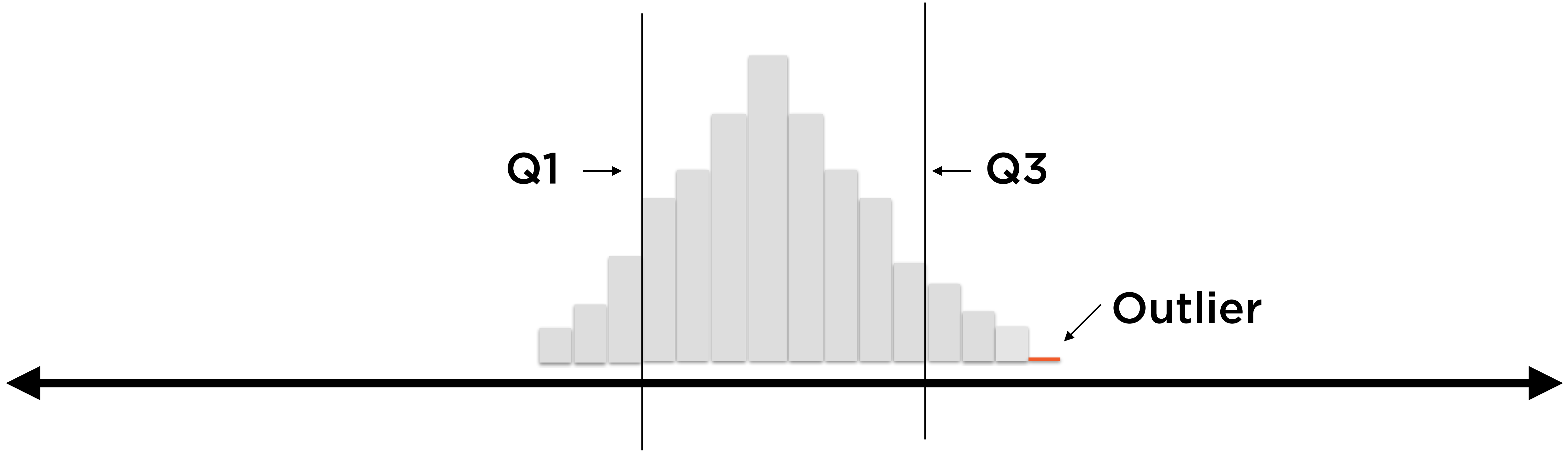
$$\text{Std Dev} = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}}$$

Outliers



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

Inter-quartile Range

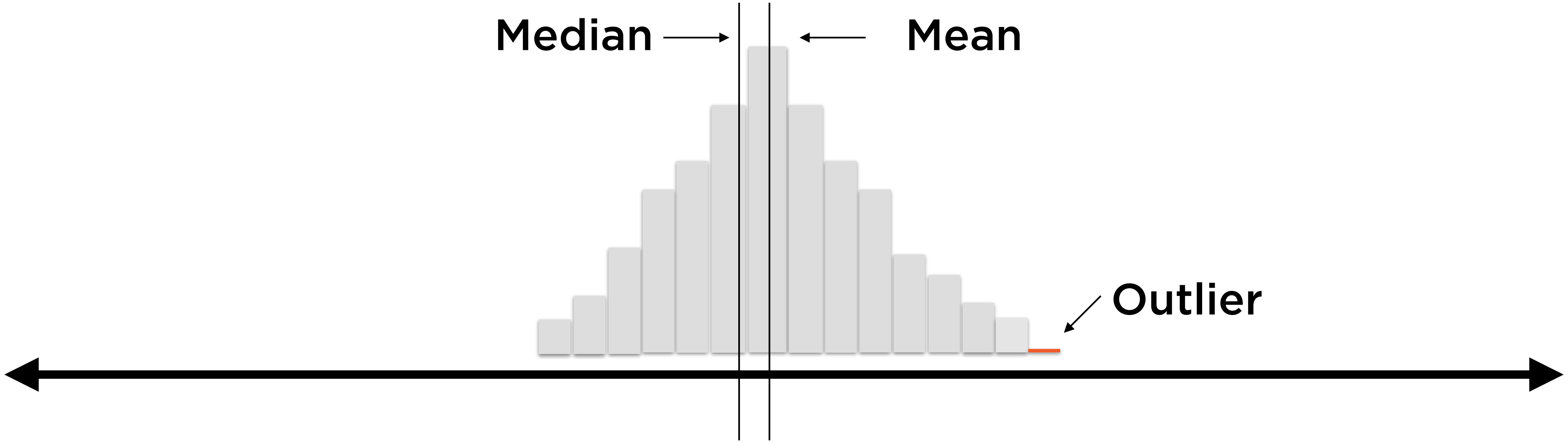


Q3 = 75th percentile: 75% of points smaller than this

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

Understanding Variance

Tossing Two Coins

Coin X

Heads = $1

Tails = -$1

Coin Y

Heads = $1,000

Tails = -$1,000

Small Stakes

Loser pays $1, winner takes $1

High Stakes

Loser pays $1000, winner takes $1000

Tossing Two Coins

Coin X Result	Coin Y Result	Coin X Payoff	Coin Y Payoff
Heads	Heads	$1	$1,000
Heads	Tails	$1	-$1,000
Tails	Heads	-$1	$1,000
Tails	Tails	-$1	-$1,000

Tabulate the possible outcomes (assume each coin is a fair one)

Tossing Two Coins

Coin X Result	Coin Y Result	Coin X Payoff	Coin Y Payoff
Heads	Heads	$1	$1,000
Heads	Tails	$1	-$1,000
Tails	Heads	-$1	$1,000
Tails	Tails	-$1	-$1,000

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = 0$$

Tossing Two Coins

Coin X Result	Coin Y Result	Coin X Payoff	Coin Y Payoff
Heads	Heads	$1	$1,000
Heads	Tails	$1	-$1,000
Tails	Heads	-$1	$1,000
Tails	Tails	-$1	-$1,000

$$\bar{x} = 0$$

Tossing Two Coins

Coin X Result	Coin Y Result	Coin X Payoff	Coin Y Payoff
Heads	Heads	$1	$1,000
Heads	Tails	$1	-$1,000
Tails	Heads	-$1	$1,000
Tails	Tails	-$1	-$1,000

$\bar{x} = 0 \quad \bar{y} = 0$

Tossing Two Coins

Coin X Result	Coin Y Result	Coin X Payoff	Coin Y Payoff
Heads	Heads	$1	$1,000
Heads	Tails	$1	-$1,000
Tails	Heads	-$1	$1,000
Tails	Tails	-$1	-$1,000

$\bar{x} = 0 \quad \bar{y} = 0$

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

Tossing Two Coins

Coin X Result	Coin Y Result	Coin X Payoff	Coin Y Payoff
Heads	Heads	\$1	\$1,000
Heads	Tails	\$1	-\$1,000
Tails	Heads	-\$1	\$1,000
Tails	Tails	-\$1	-\$1,000

$x_i - \bar{x}$	$(x_i - \bar{x})^2$
\$1	1
\$1	1
-\$1	1
-\$1	1

$\bar{x} = 0 \qquad \bar{y} = 0$

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

Tossing Two Coins

Coin X Result	Coin Y Result	Coin X Payoff	Coin Y Payoff
Heads	Heads	$1	$1,000
Heads	Tails	$1	-$1,000
Tails	Heads	-$1	$1,000
Tails	Tails	-$1	-$1,000

$y_i - \bar{y}$	$(y_i - \bar{y})^2$
$1,000	10,00,000
-$1,000	10,00,000
$1,000	10,00,000
-$1,000	10,00,000

$$x = 0 \quad \bar{y} = 0$$

$$\text{Variance} = \frac{\sum (y_i - \bar{y})^2}{n} = 1{,}000{,}000$$

Tossing Two Coins

Coin X Result	Coin Y Result	Coin X Payoff	Coin Y Payoff
Heads	Heads	$1	$1,000
Heads	Tails	$1	-$1,000
Tails	Heads	-$1	$1,000
Tails	Tails	-$1	-$1,000

Var(x) = 1

Var(y) = 1,000,000

As stakes grow, variance gets big faster than the mean

Tossing Two Coins

Coin X → Heads = $1
Coin X → Tails = -$1

Coin Y → Heads = $1,000
Coin Y → Tails = -$1,000

Small Stakes

Loser pays $1, winner takes $1

High Stakes

Loser pays $1000, winner takes $1000

As stakes grow 1000x, variance grows 1,000,000x

Gaussian Normal Distribution

Distribution

A formula which tells how likely a particular value is to occur in your data

Distribution

All values are equally likely

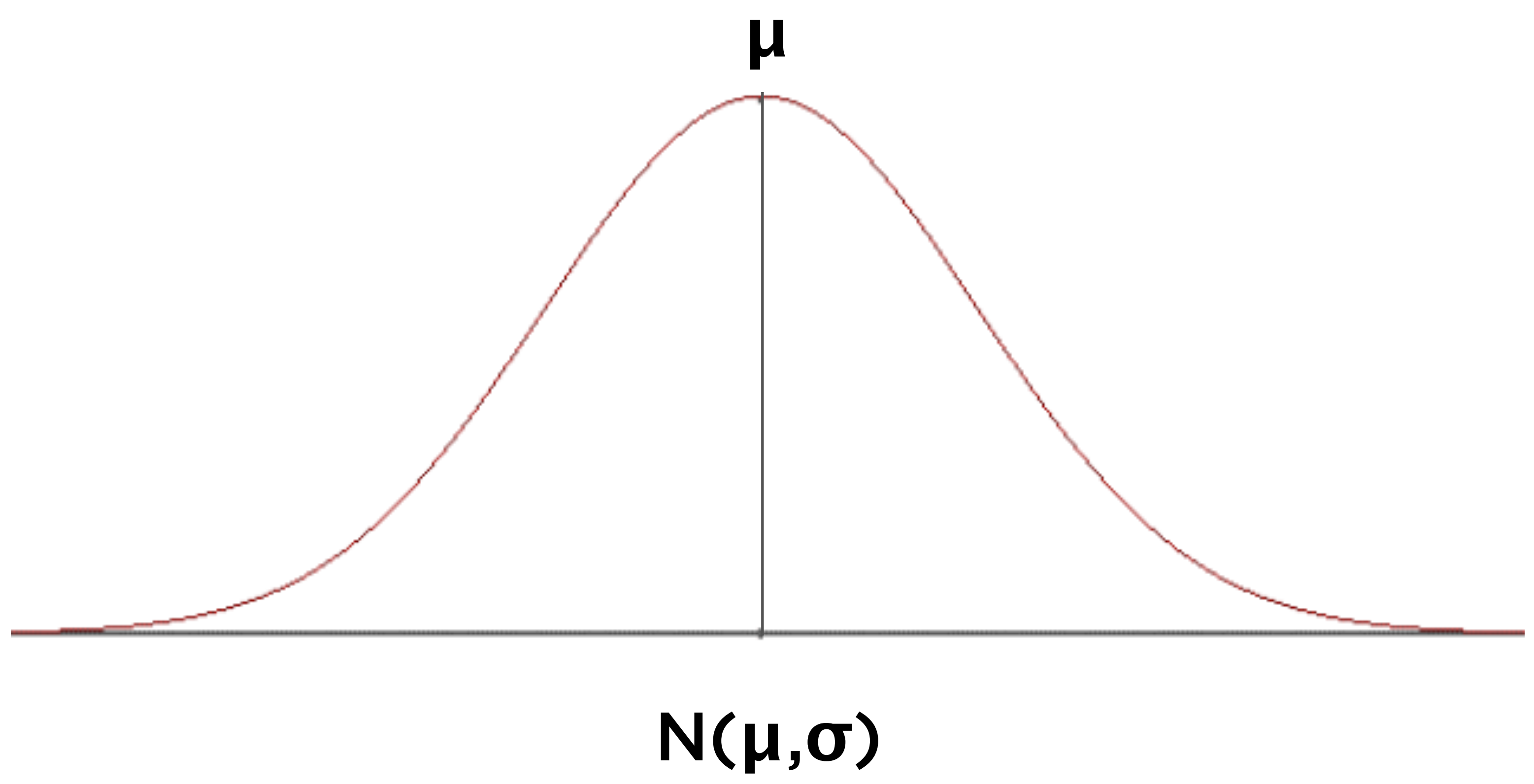
Values close to the mean are more likely

Properties in the real world can be represented by a normal distribution

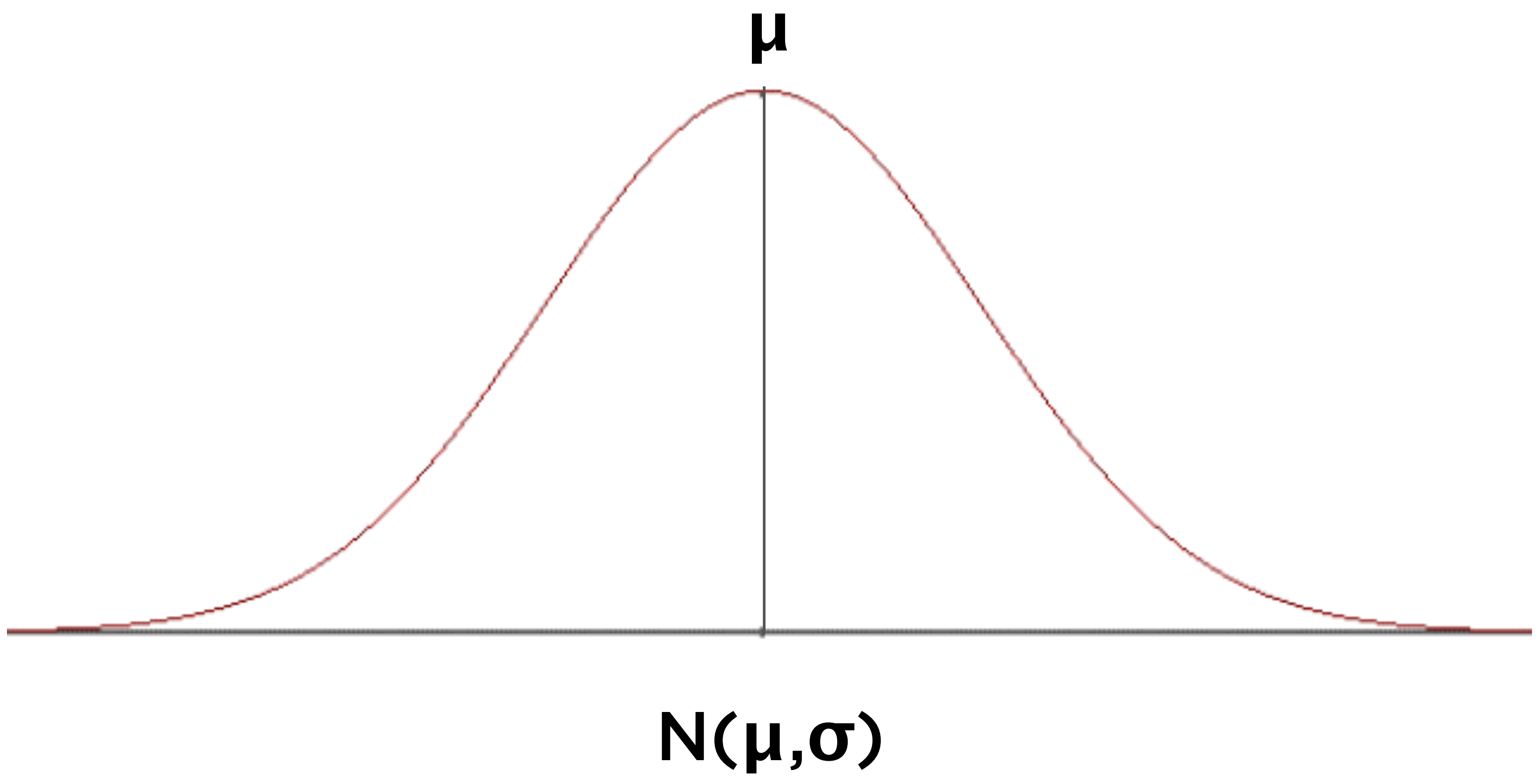
Gaussian distribution

Gaussian Distribution

Gaussian Distribution

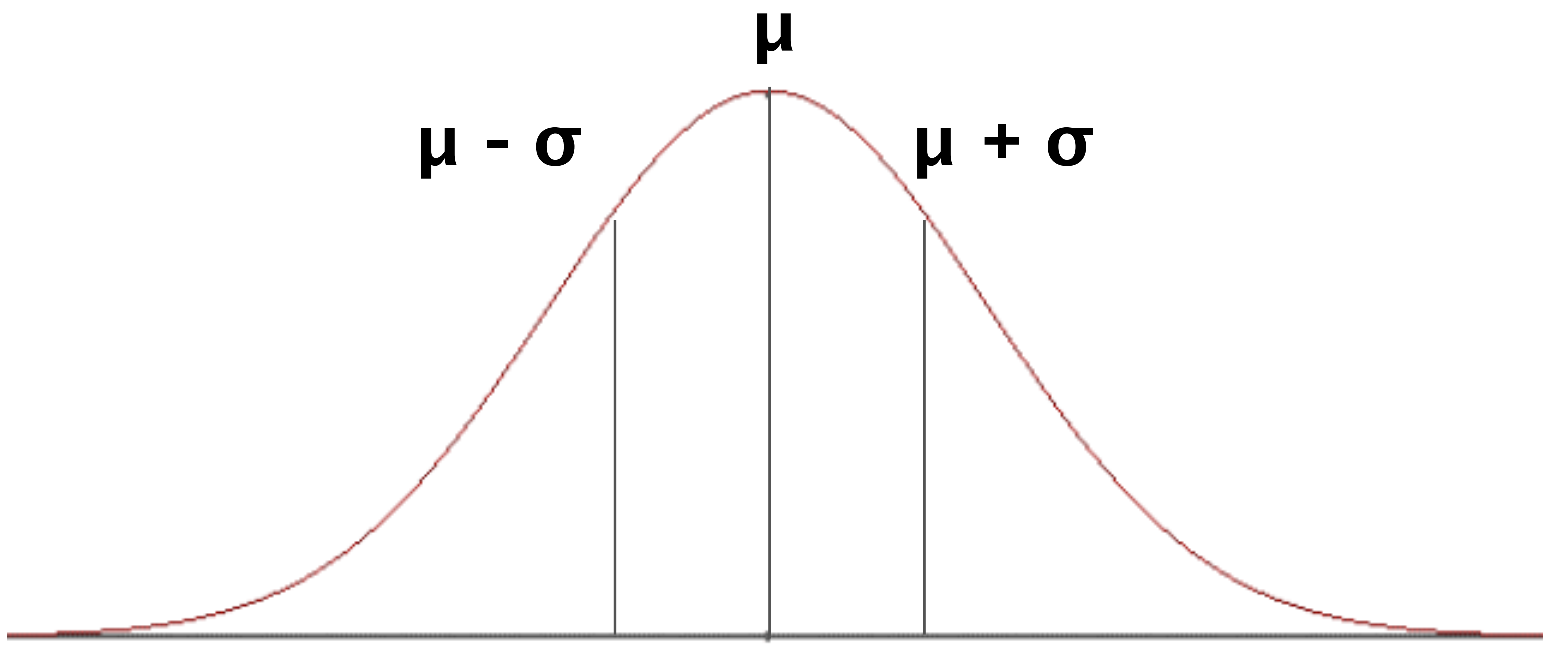


Gaussian Distribution



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

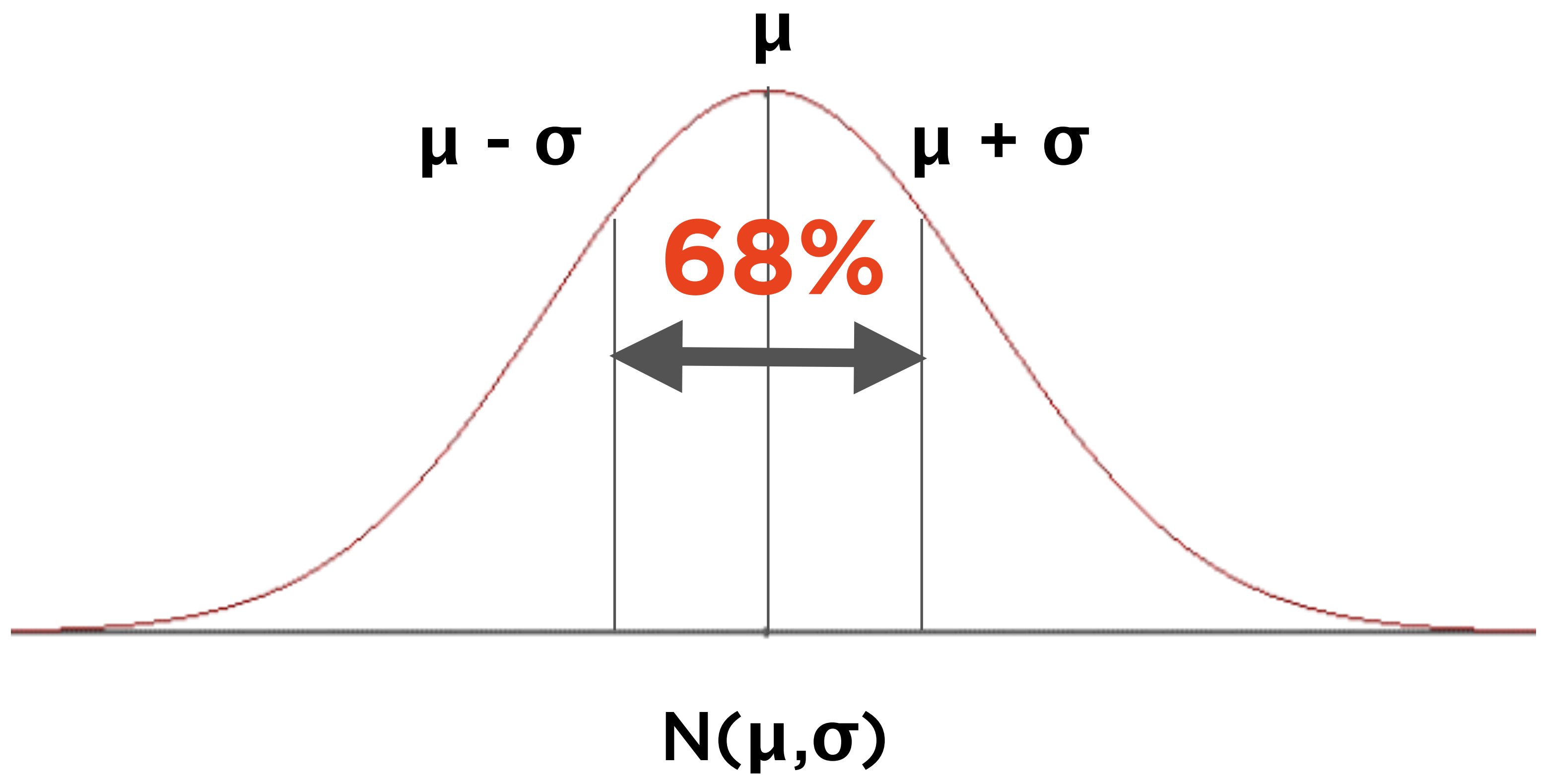
Gaussian Distribution



N(μ,σ)

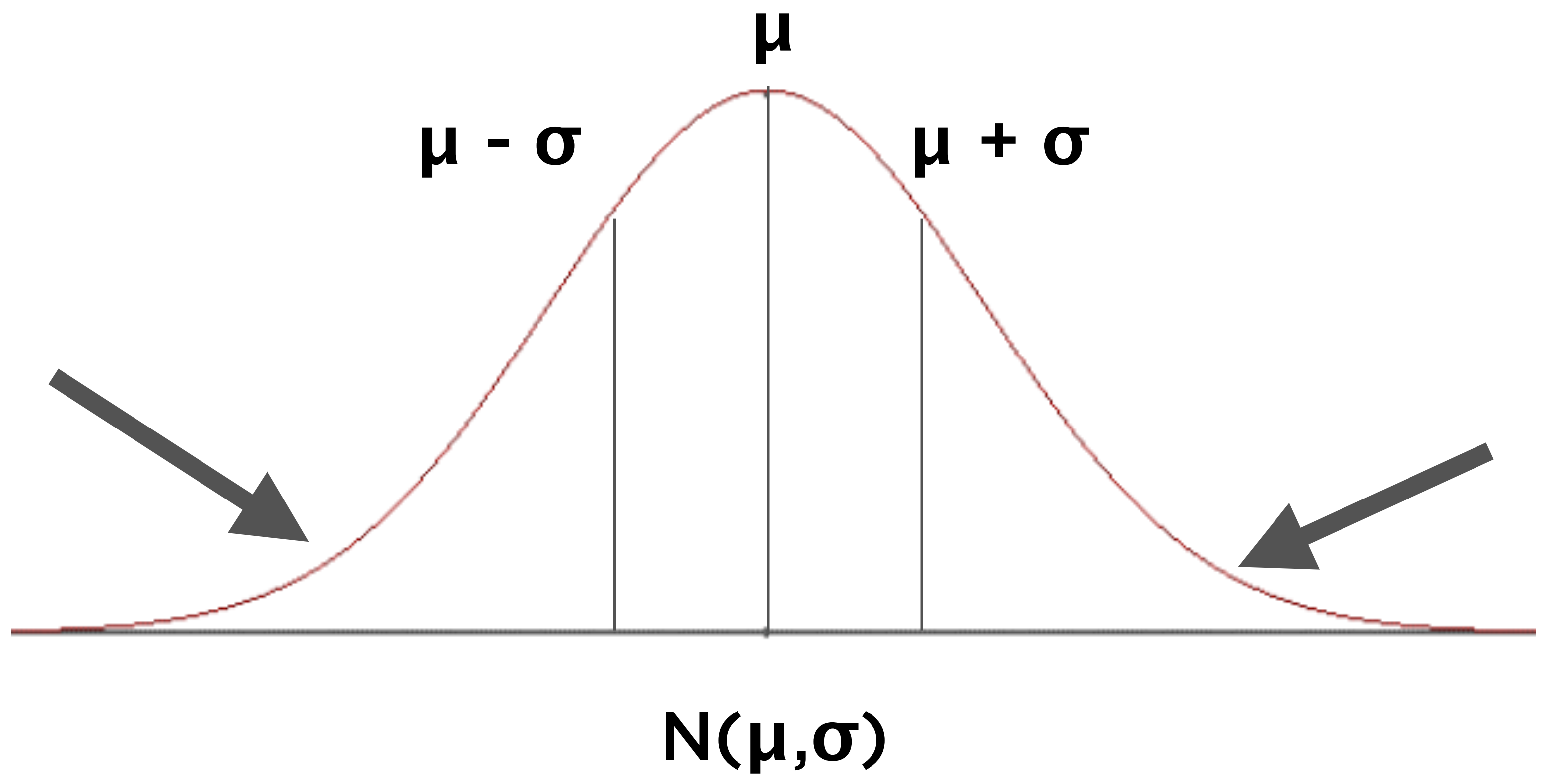
$$N(\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Gaussian Distribution



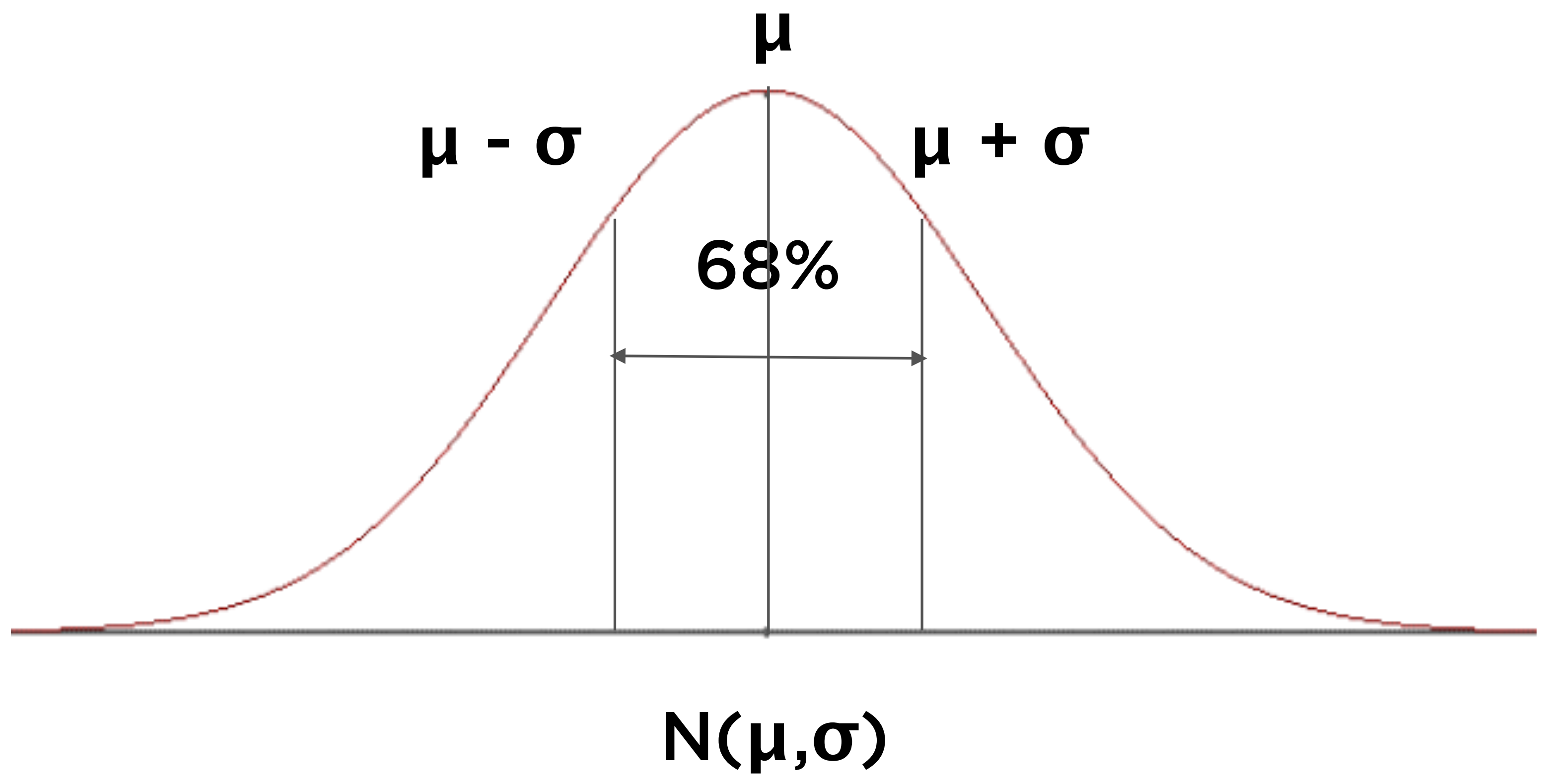
There will be a large number of points close to the average

Gaussian Distribution



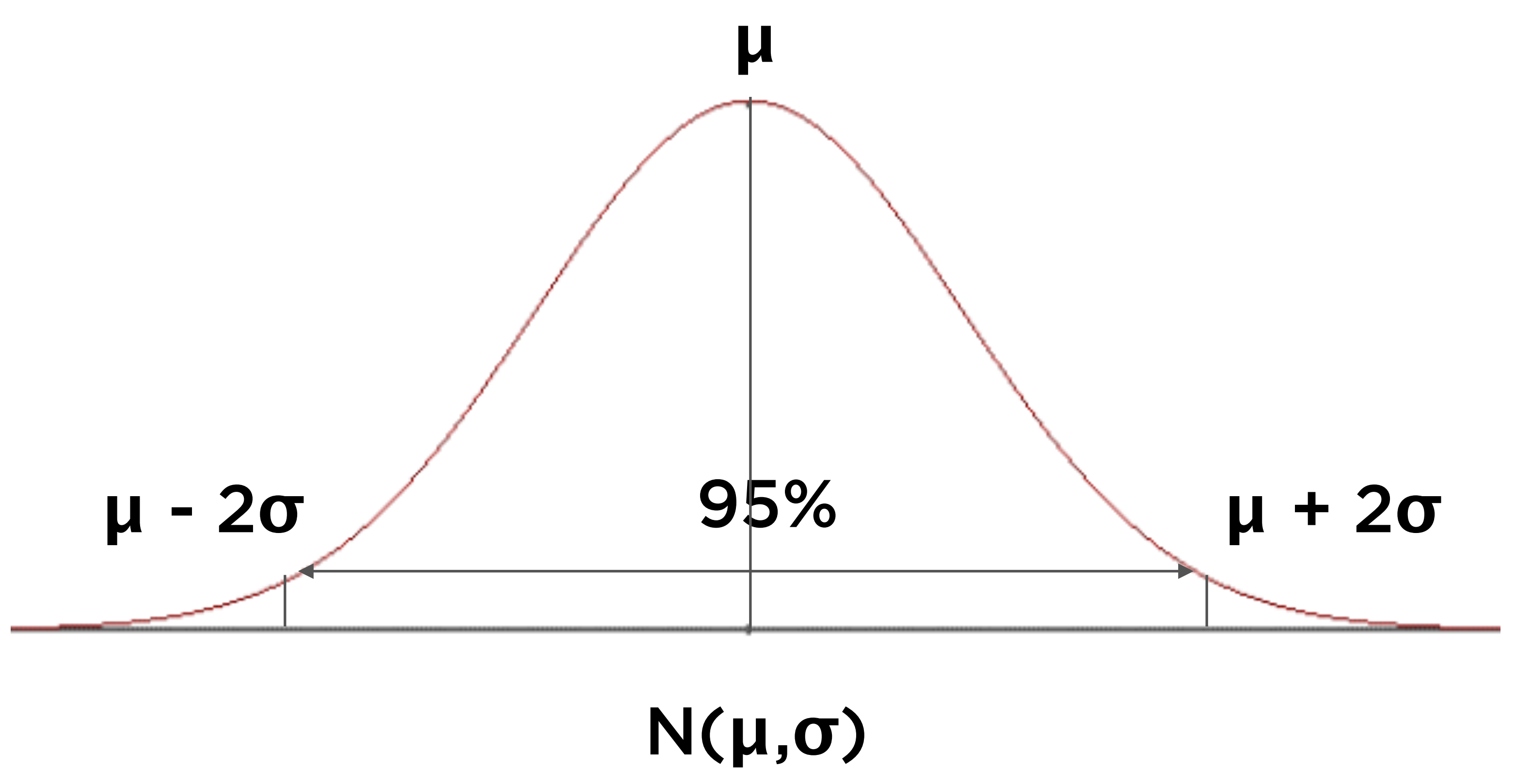
There will be few extreme values - the number of extreme values at either side of the mean will be the same

Gaussian Distribution



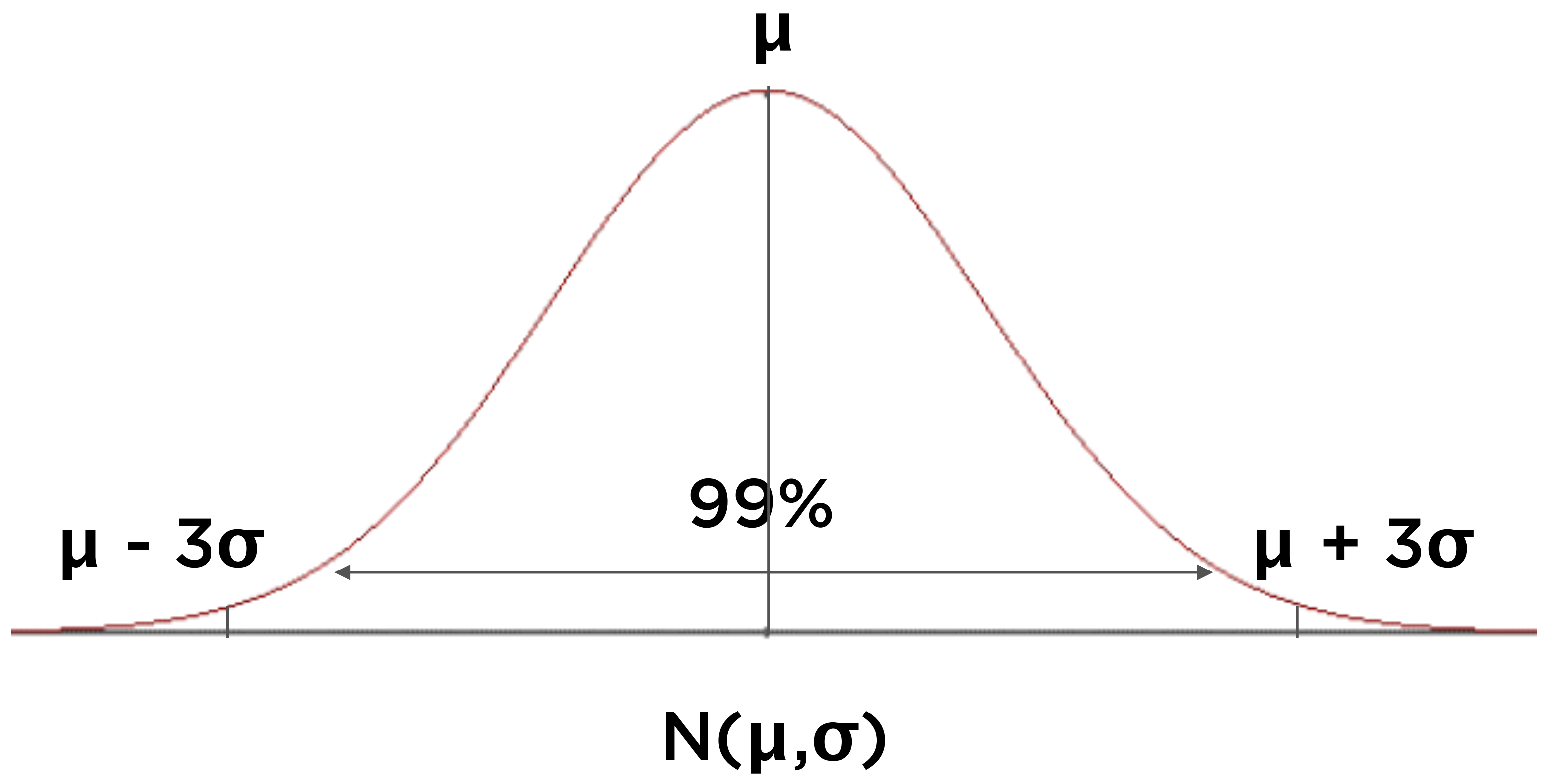
68% within 1 standard deviation of mean

Gaussian Distribution



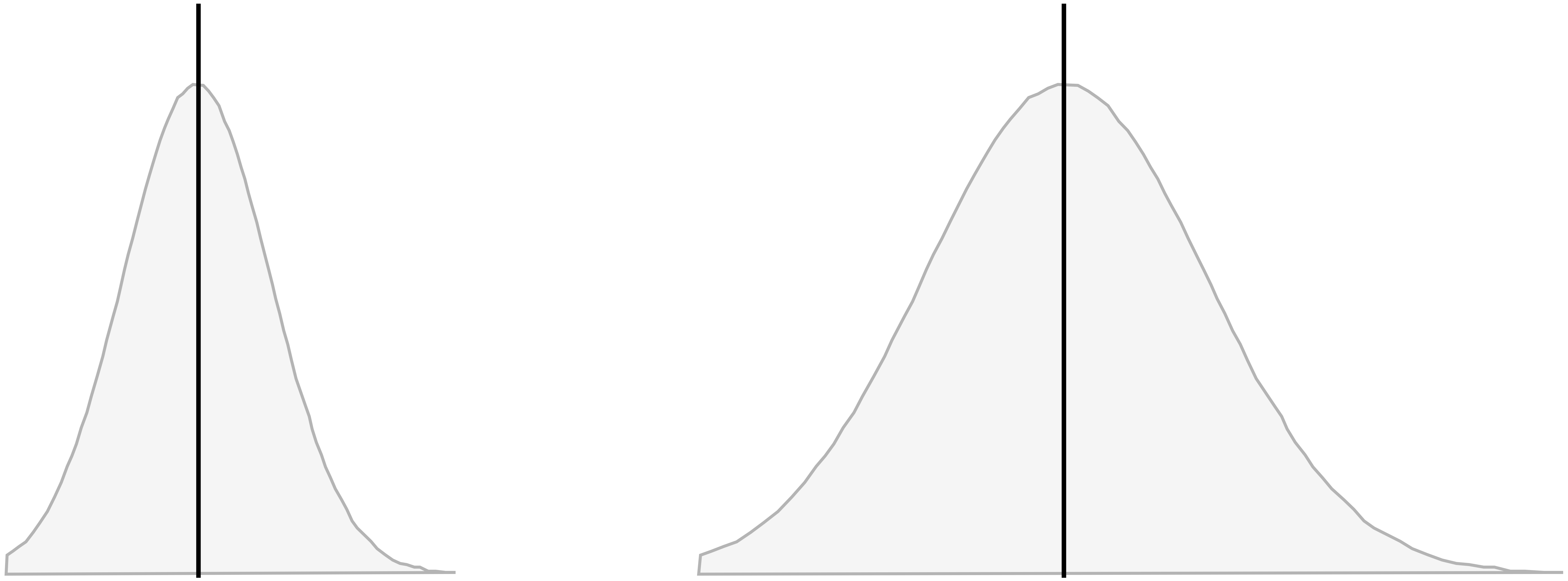
95% within 2 standard deviations of mean

Gaussian Distribution



99% within 3 standard deviations of mean

Role of Sigma



Small Standard Deviation

Few points far from the mean

Large Standard Deviation

Many points far from the mean

Confidence Intervals

From Sample to Population

Population

All the data out there in the universe

Sample

A subset - hopefully representative - of the population

Mean and Variance

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

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

These statistics only apply to the sample of data, and so are known as sample statistics

The corresponding figures for all possible data points out there are called population statistics

From Sample to Population

Sample Mean

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

Population Mean

$$\mu = ?$$

Estimating Population Mean

Aim: Estimate a statistical property (mean) of the population

Will need to do so from a sample

Use properties of sample to estimate property of population

Sampling Distribution

Tricky part is going from properties of sample to property of population

Can't be completely sure of population property

Can however be sure of probability distribution of the population property

This distribution depends on sample alone - Sampling Distribution

Sampling Distribution

Probability distribution of a population statistic (e.g. population mean), given a particular sample.

From Sample to Population

Sample Mean

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

Population Mean

$$\mu = ?$$

From Sample to Population

Sample Mean

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

Population Mean

Sampling Distribution

Sample Mean

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

Population Mean

Estimating Population Mean

Turns out, $\bar{x}$ is the best estimate of μ

Sample mean is best, unbiased estimator of the population mean

Even so, how sure are we of our estimate?

Confidence levels help answer this question

"We can be 99% confident that the average is between ____ and _____"

Confidence Intervals

Variability within Sample

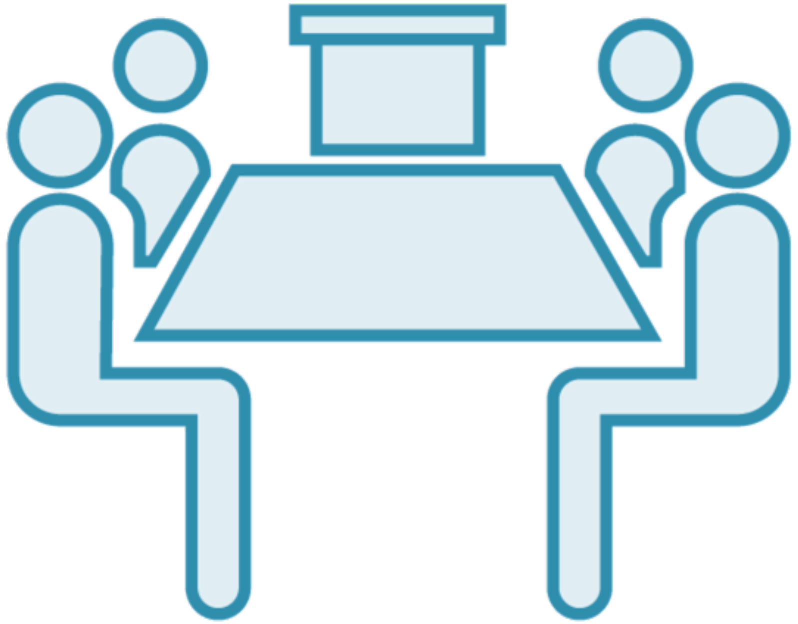
Say we sample 100 points and all of them have the exact same value

- Our confidence in our estimate would be high (intuitively)

Say we sample 100 points and their values vary tremendously

- Our confidence in our estimate would be low (intuitively)

Sample Size Relative to Population



Say we sample 100 million points out of 1 billion and got a sample estimate

- Our confidence in our estimate would be relatively high (intuitively)

Say we sample 100 points out of 1 billion and got a sample estimate

- Our confidence in our estimate would be low (intuitively)

Intuition behind Confidence

Intuitively, confidence in our estimate depends upon

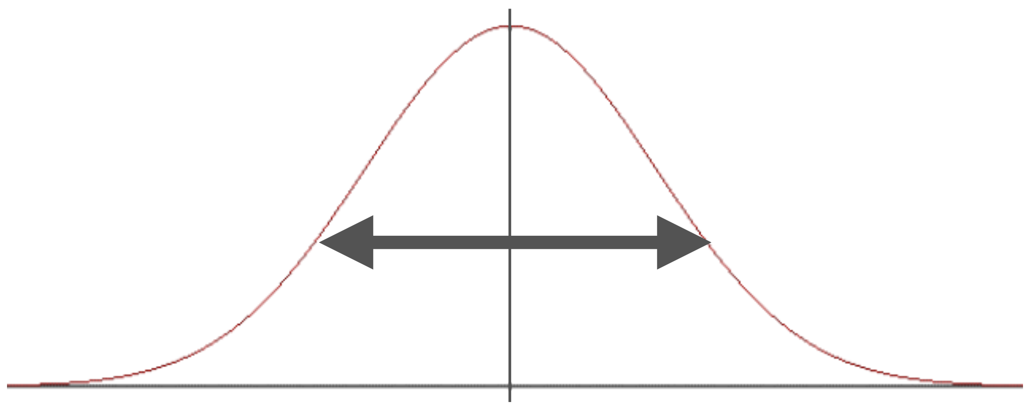
- How much data within the sample varies
- How big the sample size was

Math behind Confidence

Mathematically, confidence in our estimate depends upon

- Sample variance
- Sample size

Sampling Distribution

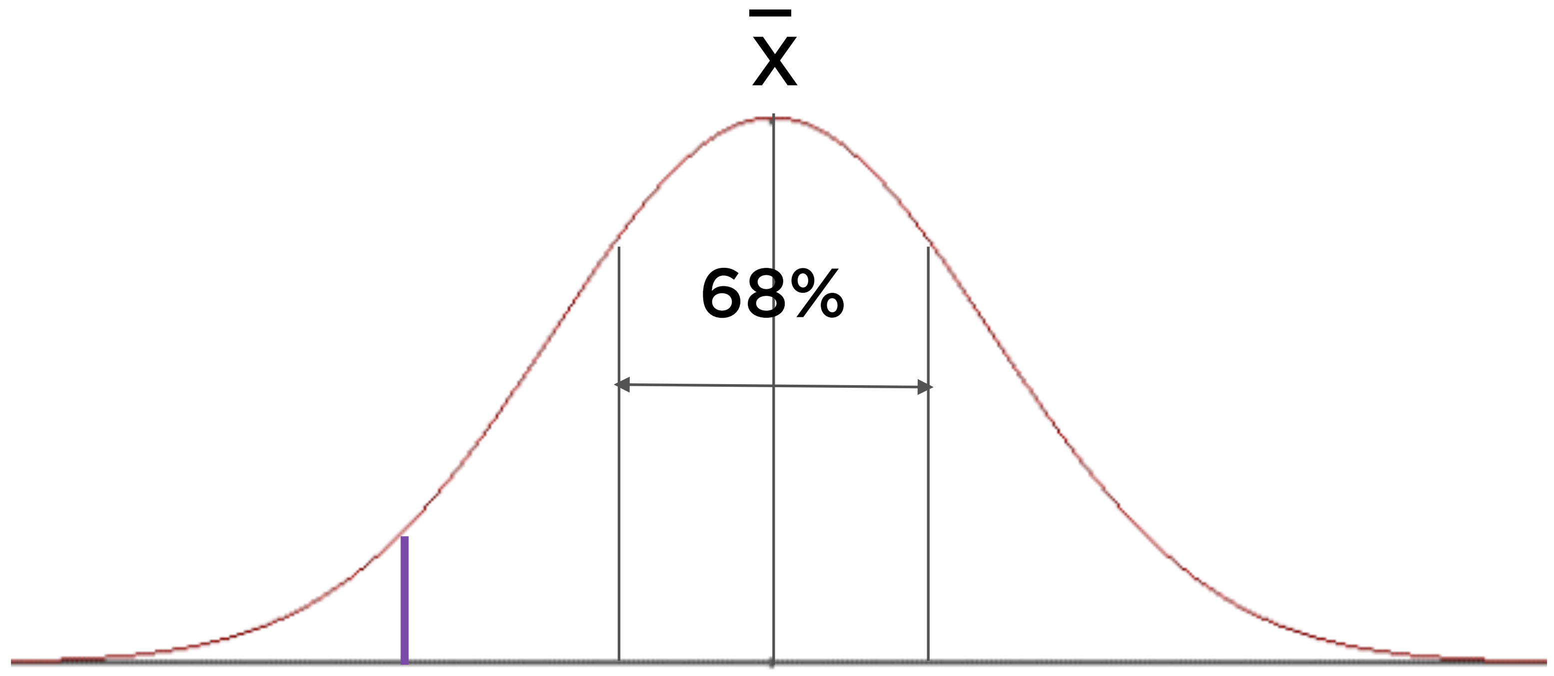


Population mean μ has a distribution called the sampling distribution

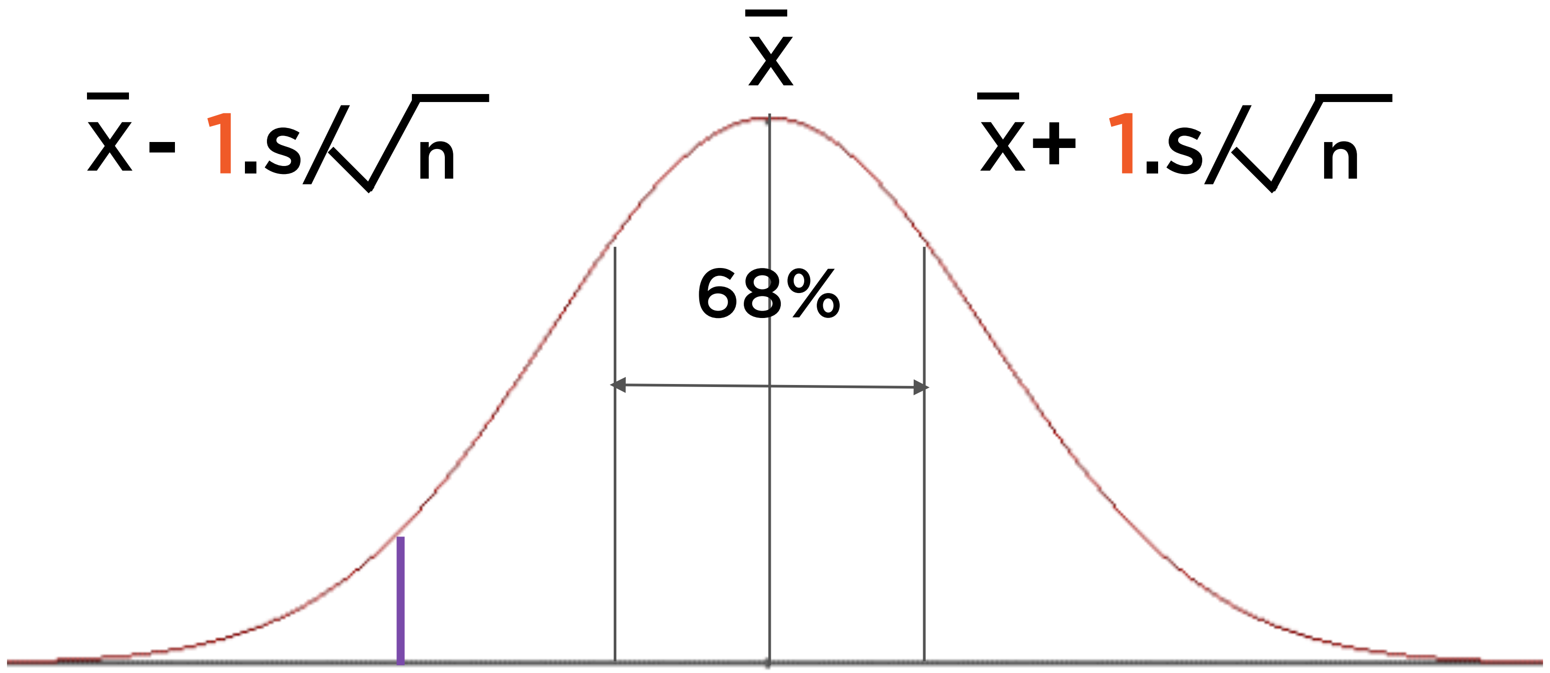
This is a normal distribution

- Mean = Sample mean
- Variance ≈ Sample variance / n
- Std dev. = Sample std dev. / sqrt(n)

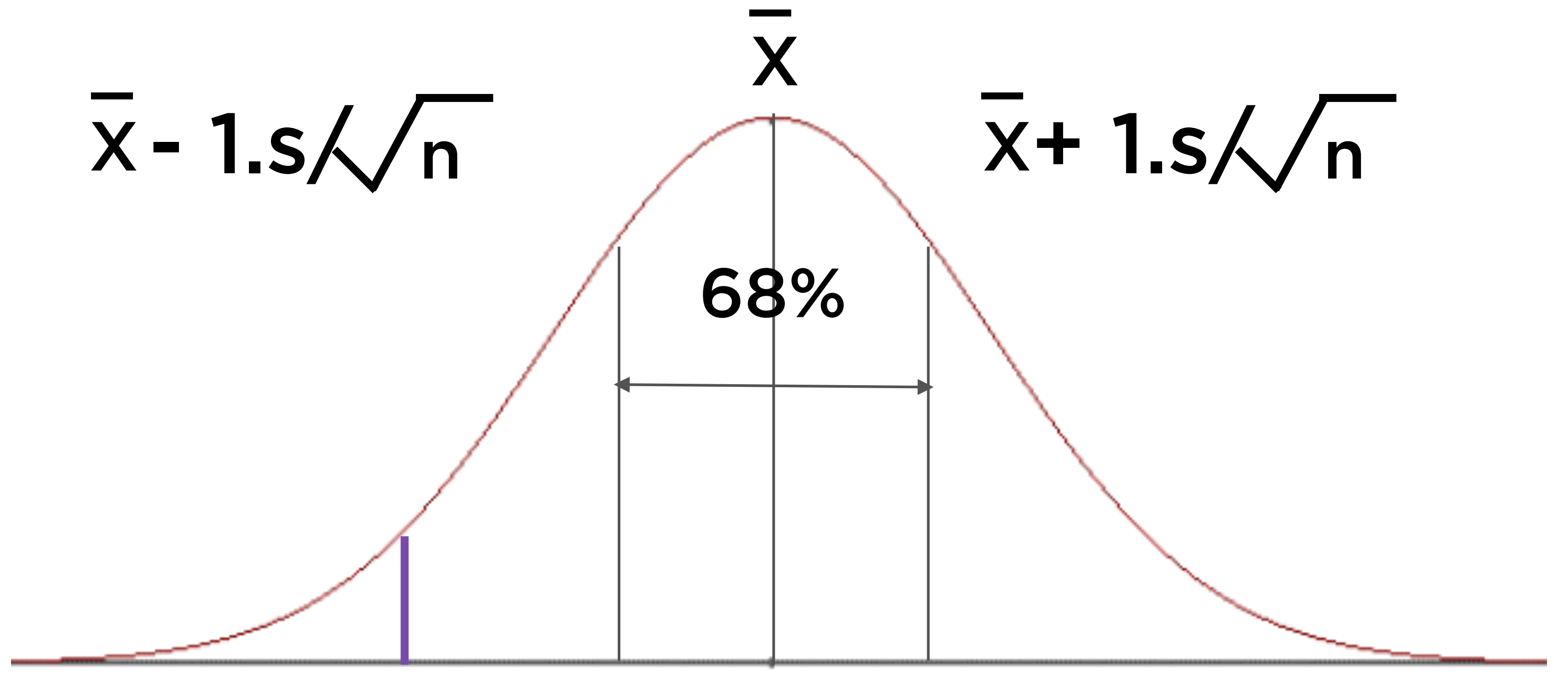
68% Confidence That μ is within 1σ of $\bar{x}$



68% Confidence That **μ** is within 1**σ** of $\bar{x}$

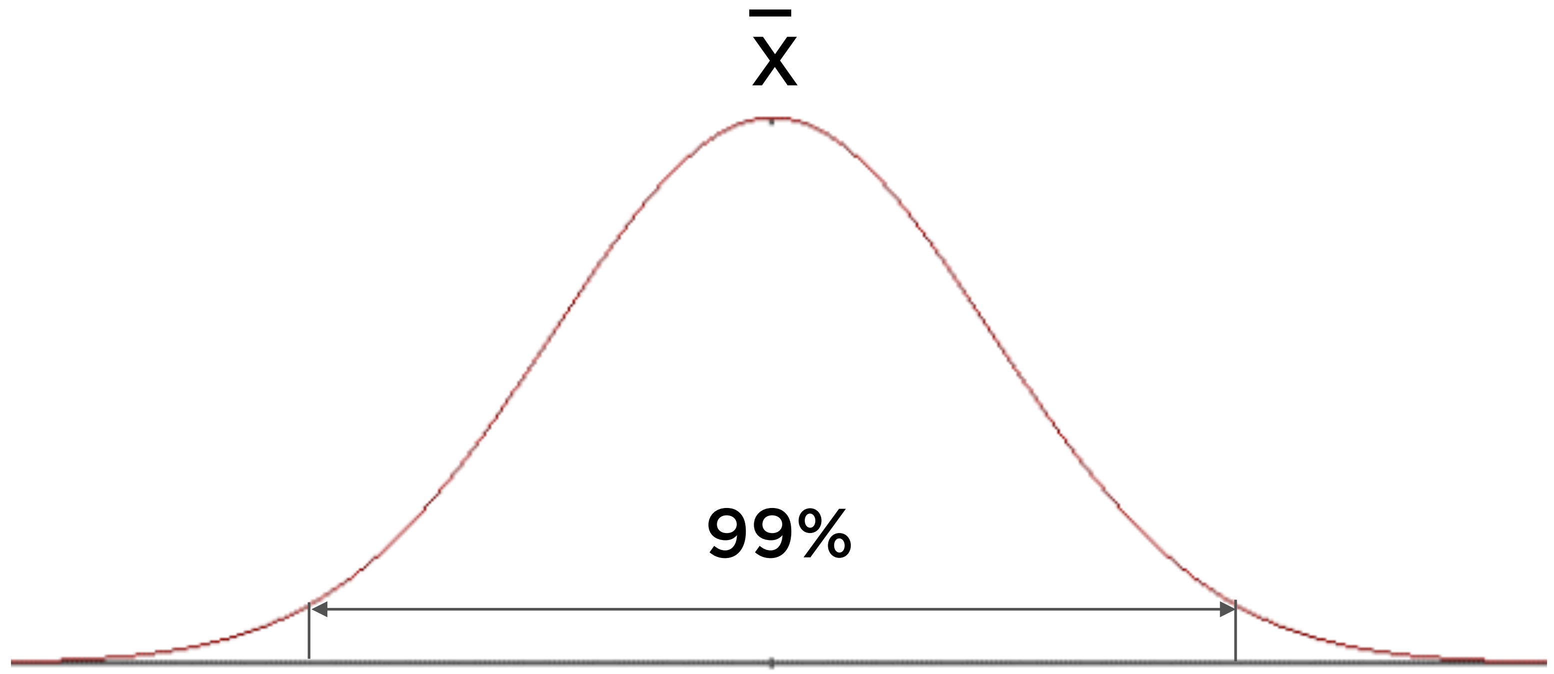


68% Confidence That μ is within 1σ of $\bar{x}$

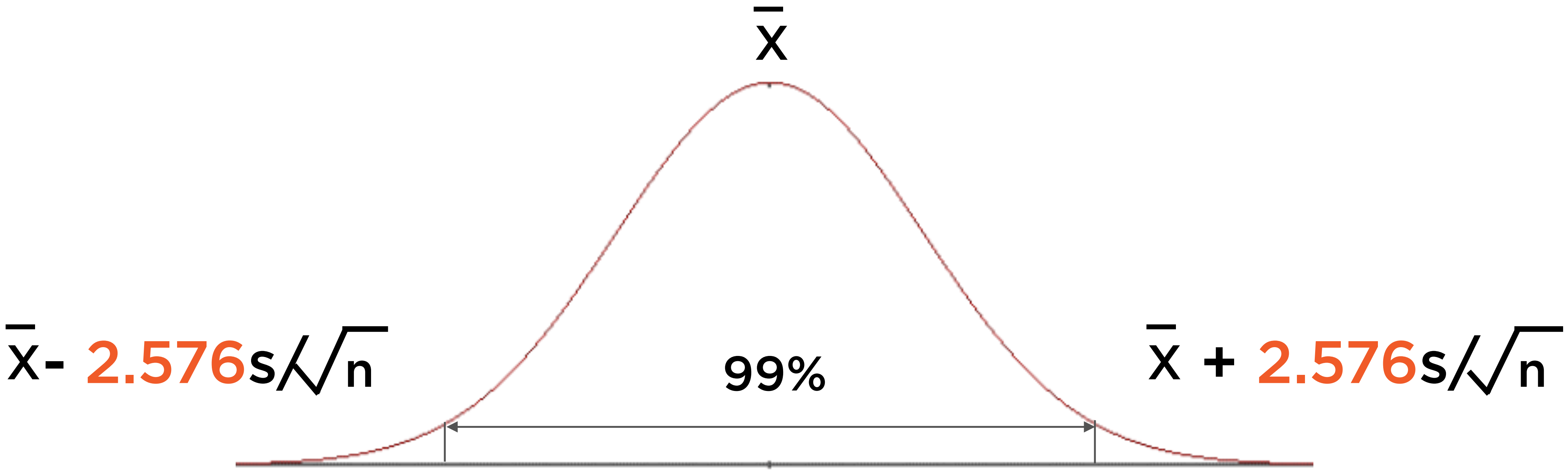


We can state with 68% confidence that the population mean μ lies in the range $\bar{x} - 1.s/\sqrt{n}$ to $\bar{x} + 1.s/\sqrt{n}$

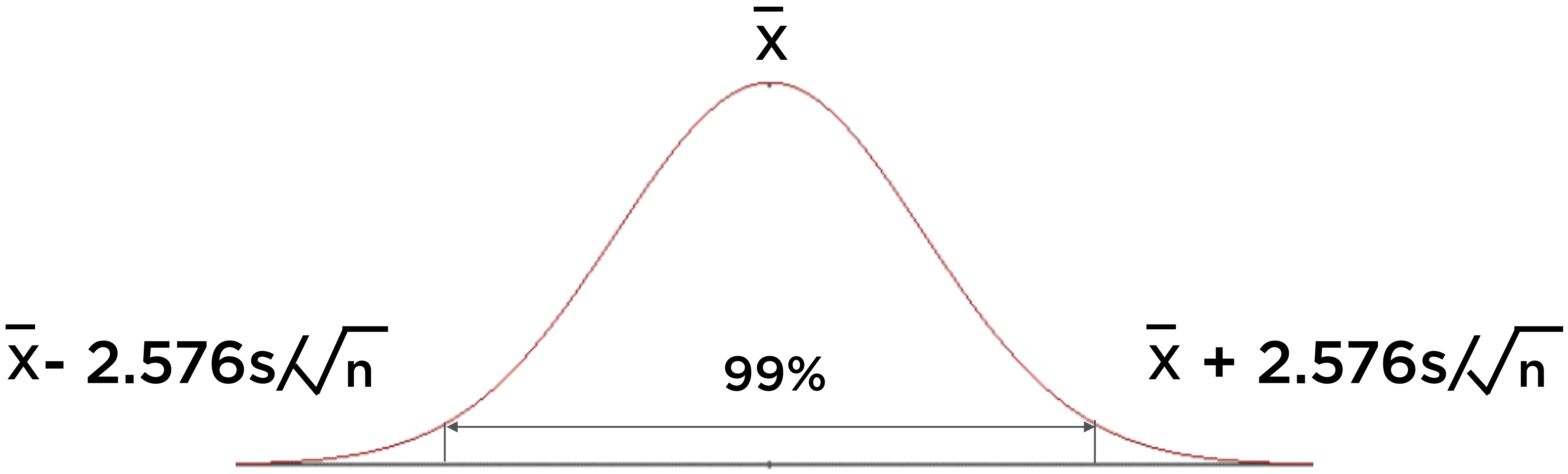
99% Confidence That μ is within 2.57σ of $\bar{x}$



99% Confidence That μ is within 2.57σ of $\bar{x}$

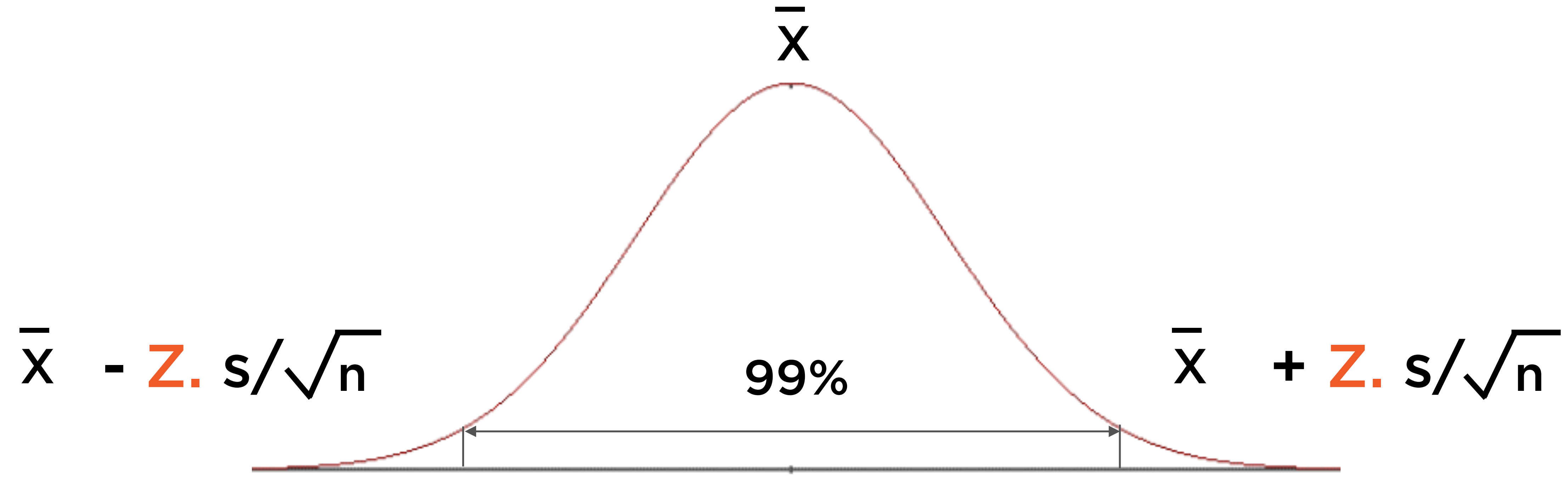


99% Confidence That μ is within 2.57σ of $\bar{x}$

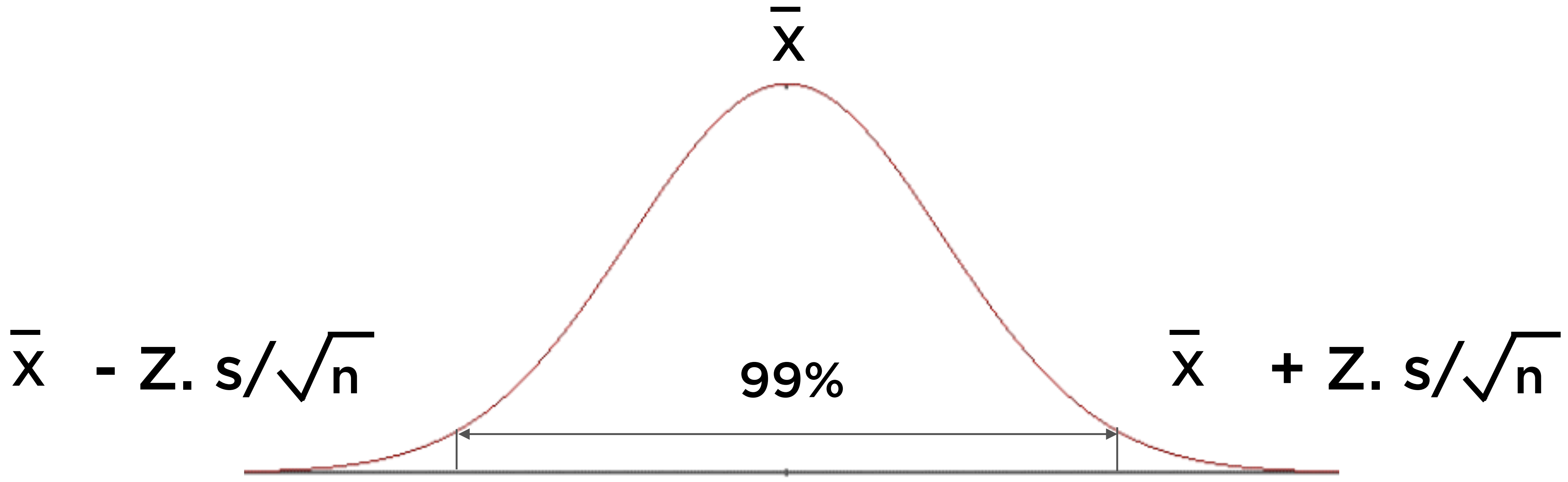


We can state with 99% confidence that the population mean μ lies in the range $\bar{x} - 2.576s/\sqrt{n}$ to $\bar{x} + 2.576s/\sqrt{n}$

(100-p)% Confidence That **μ** is within Z**σ** of $\bar{x}$

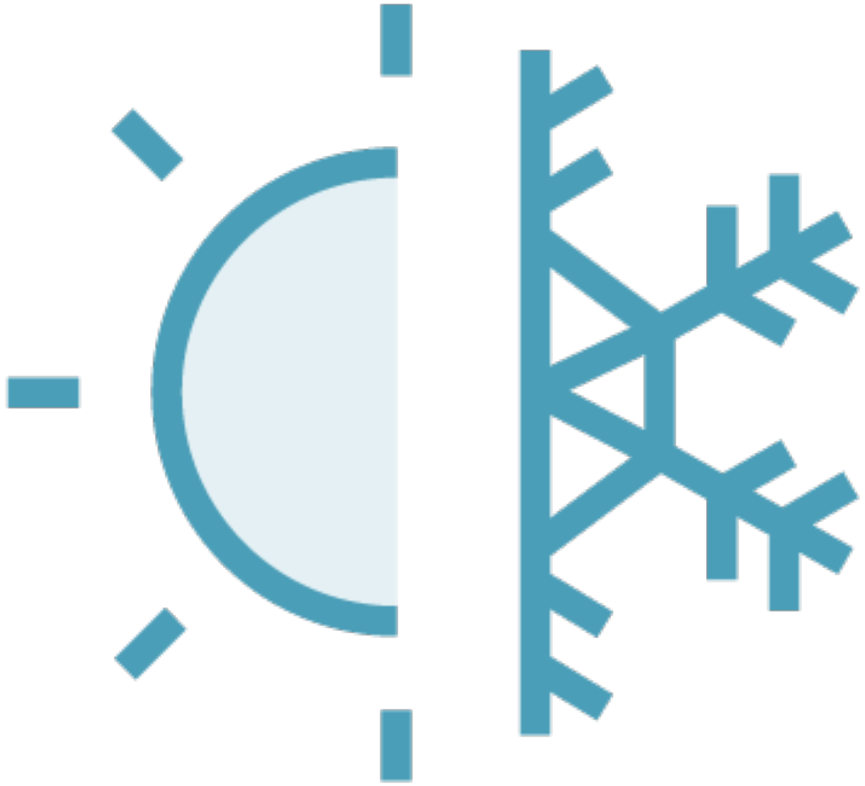


(100-p)% Confidence That μ is within $Z\sigma$ of $\bar{x}$



We can state with (100- p)% confidence that the population mean μ lies in the range $\bar{x} - Z.s/\sqrt{n}$ **to** $\bar{x} + Z.s/\sqrt{n}$

Sampling Distribution

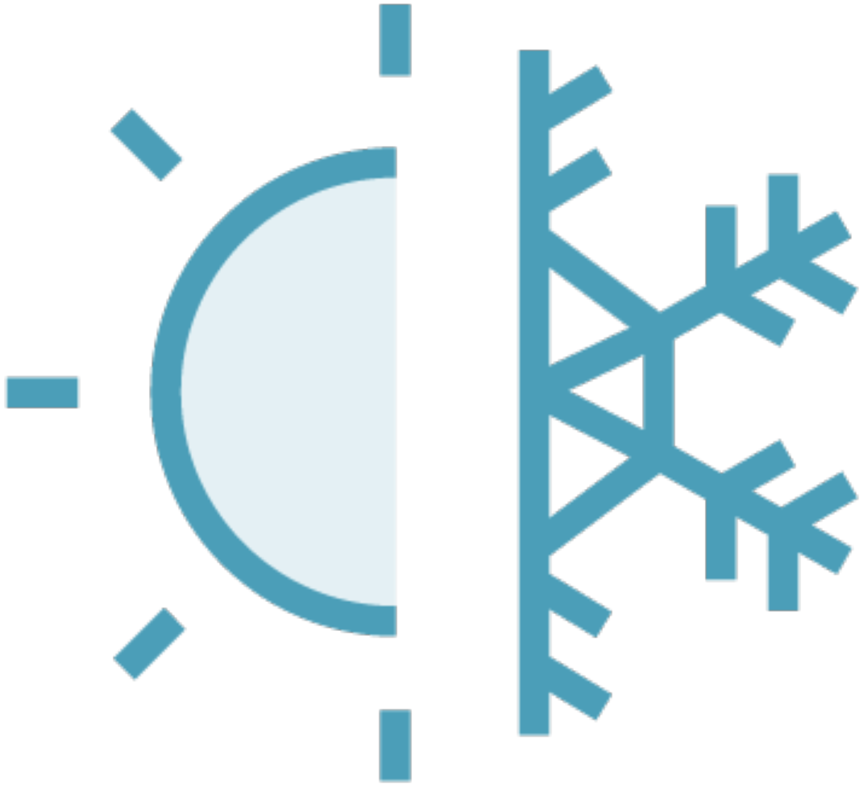


p is the level of significance

Z is the number of standard deviations from the mean corresponding to p

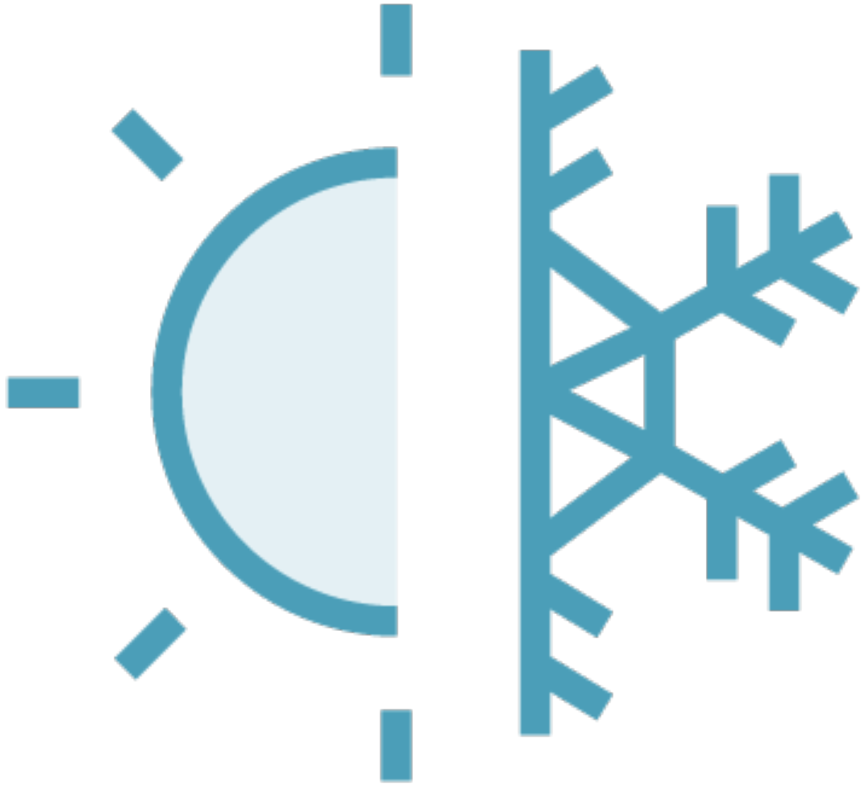
s and $\bar{x}$ are calculated from the sample properties

Sampling Distribution



Confidence Interval	Z
80%	1.282
85%	1.440
90%	1.645
95%	1.960
99%	2.576
99.5%	2.807
99.9%	3.291

Sampling Distribution



Range is centered around sample mean

Extends symmetrically on both sides

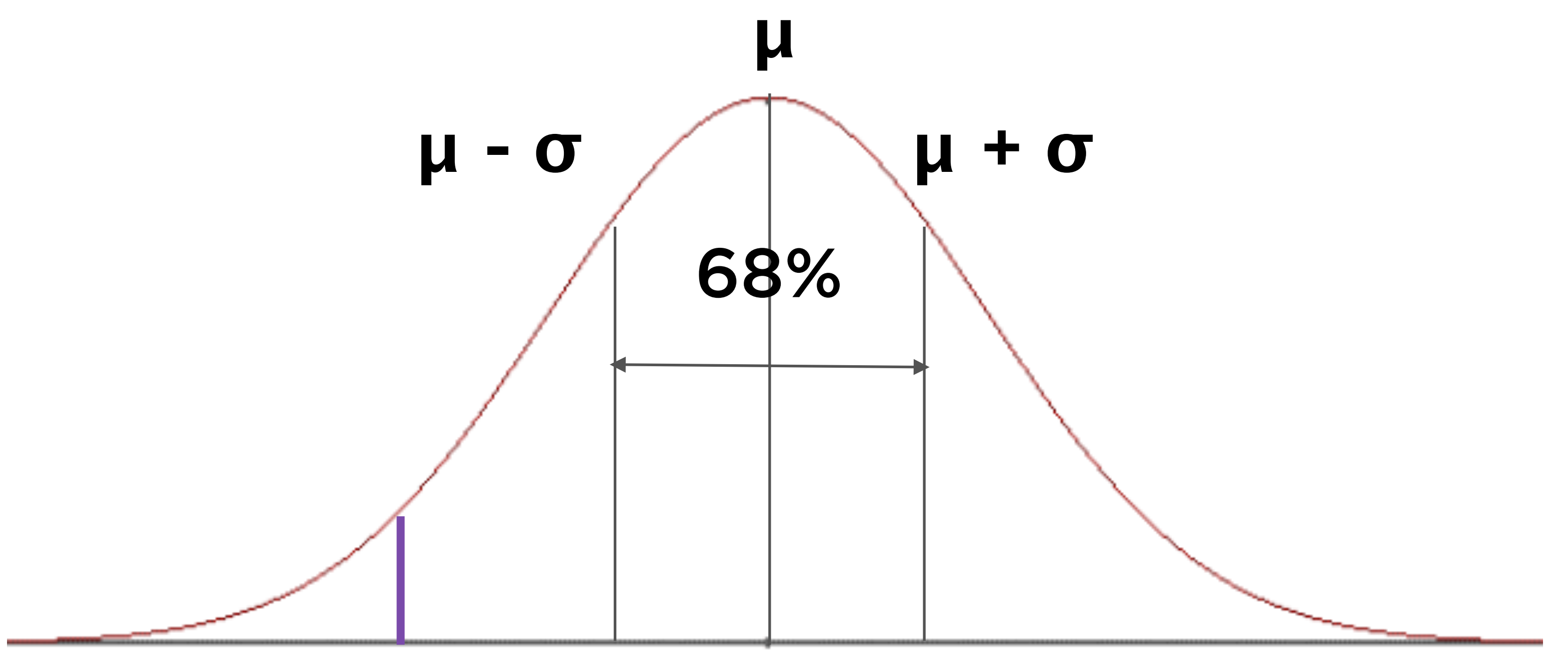
Greater the range, the greater our confidence that estimate lies within it

Skewness and Kurtosis

Skewness

A measure of asymmetry around the mean

Gaussian Distribution



N(μ,σ)

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

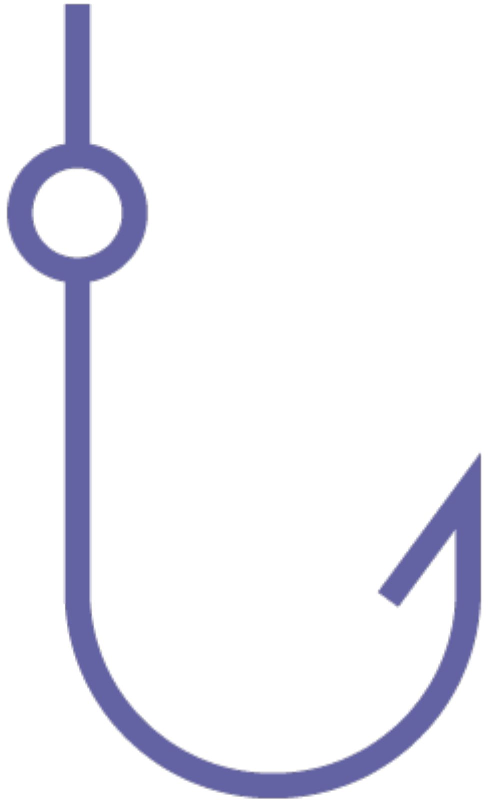
Skewness

Normally distributed data: skewness = 0

Extreme values are equally likely on both sides of the mean

Symmetry about the mean

Positive Skewness



Consider incomes of individuals

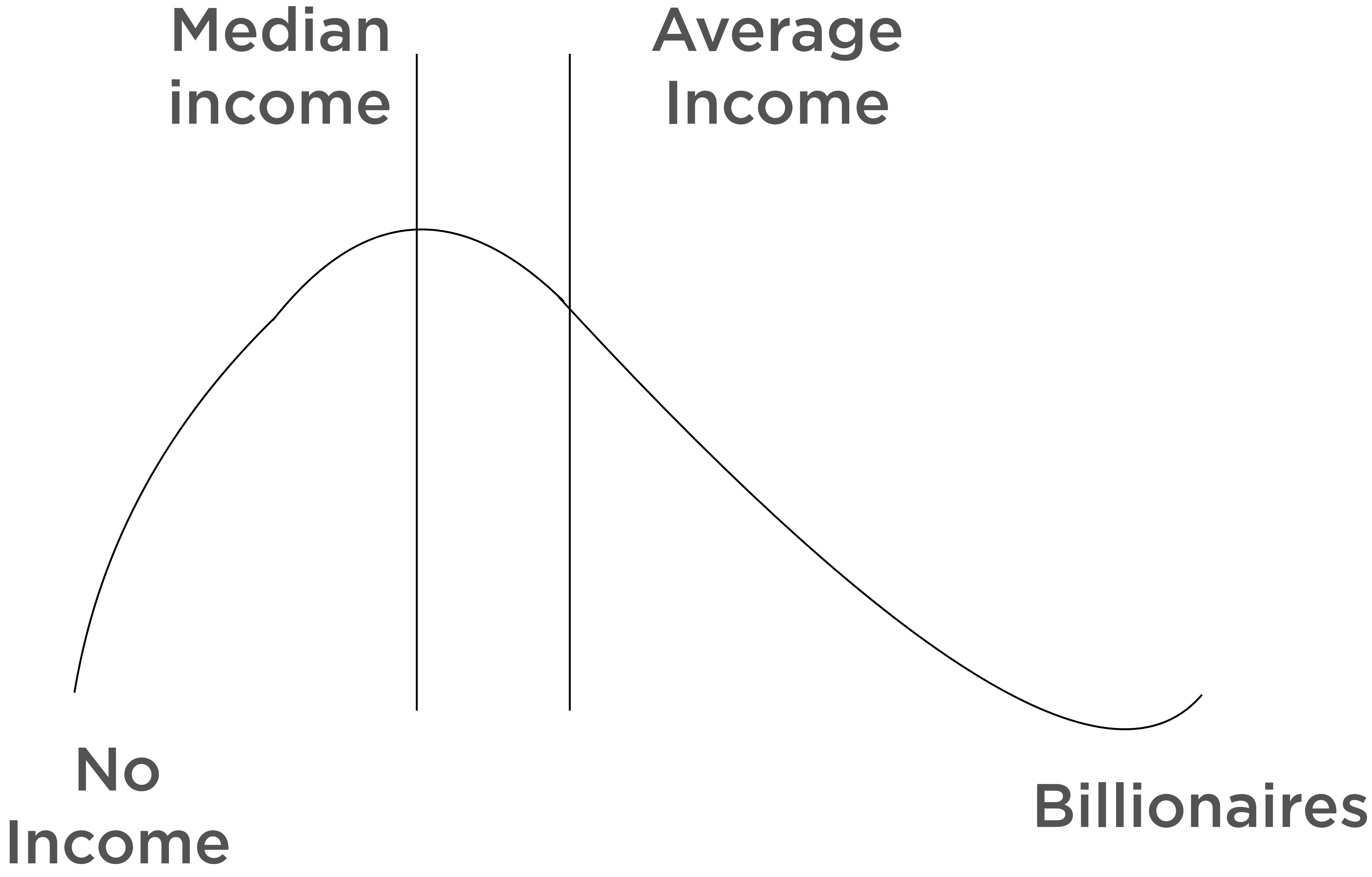
Billionaires: positive skew

Outliers greater than mean more likely than outliers less than mean

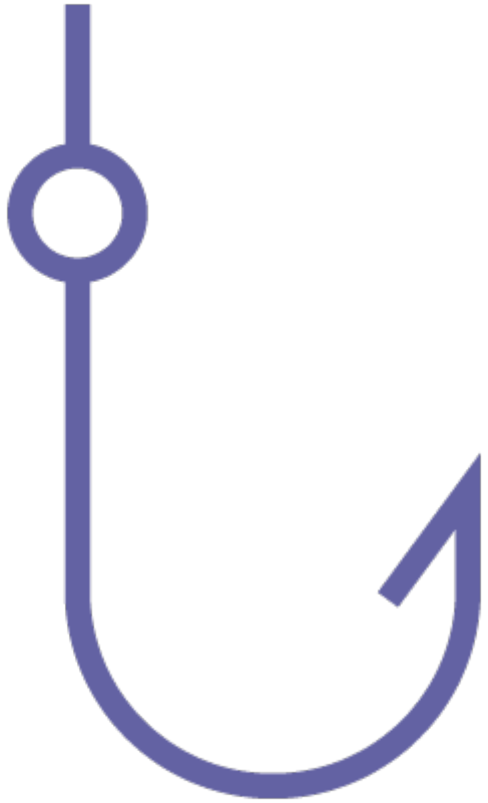
Right-skewed distribution

Often seen when lower bound but no upper bound

Positive Skewness



Negative Skewness



Consider losses from storms

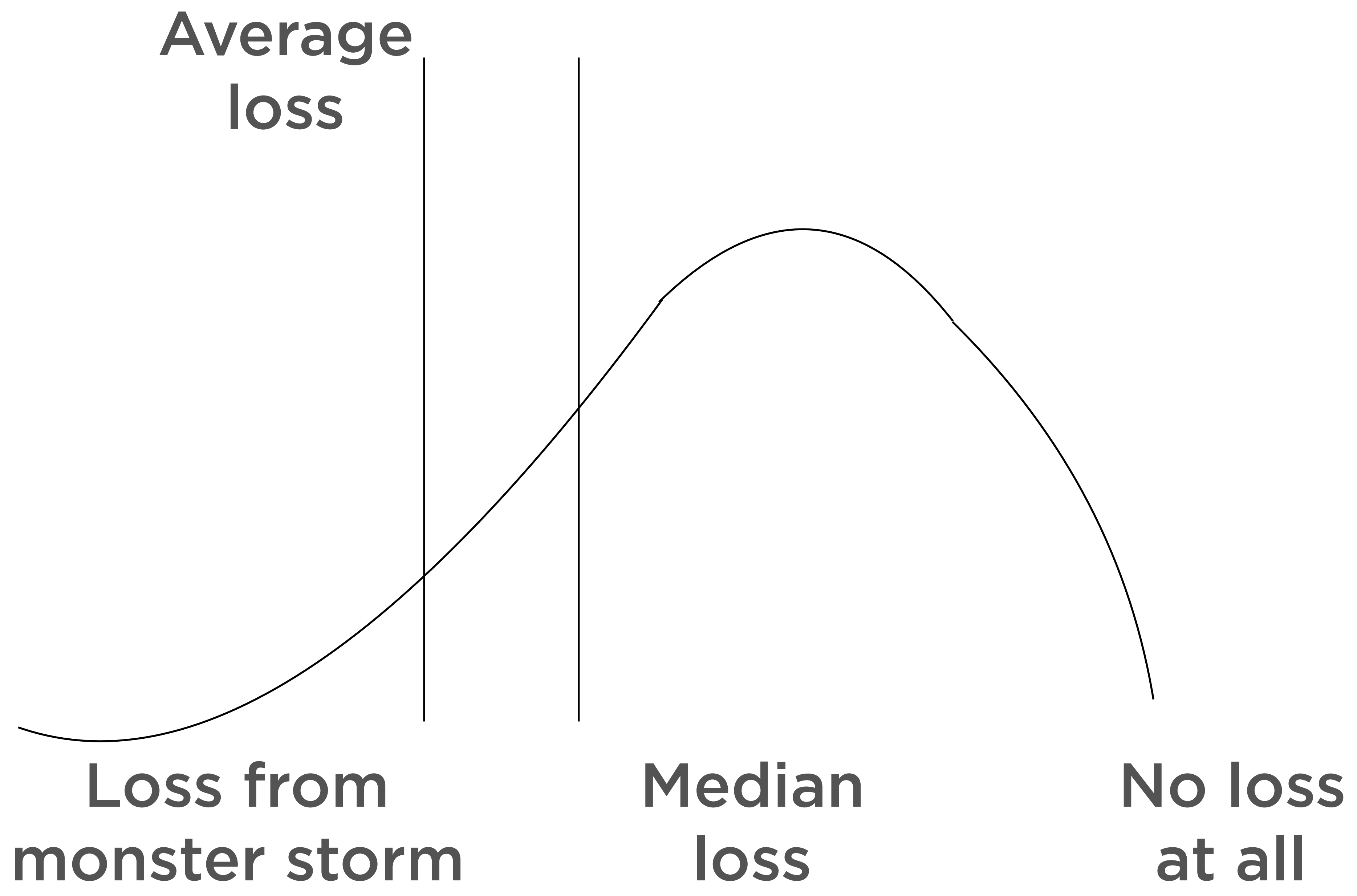
Usually minor, then a monster storm hits

Outliers worse than mean more likely than outliers greater than mean

Left-skewed distribution

Often seen when upper bound but no lower bound

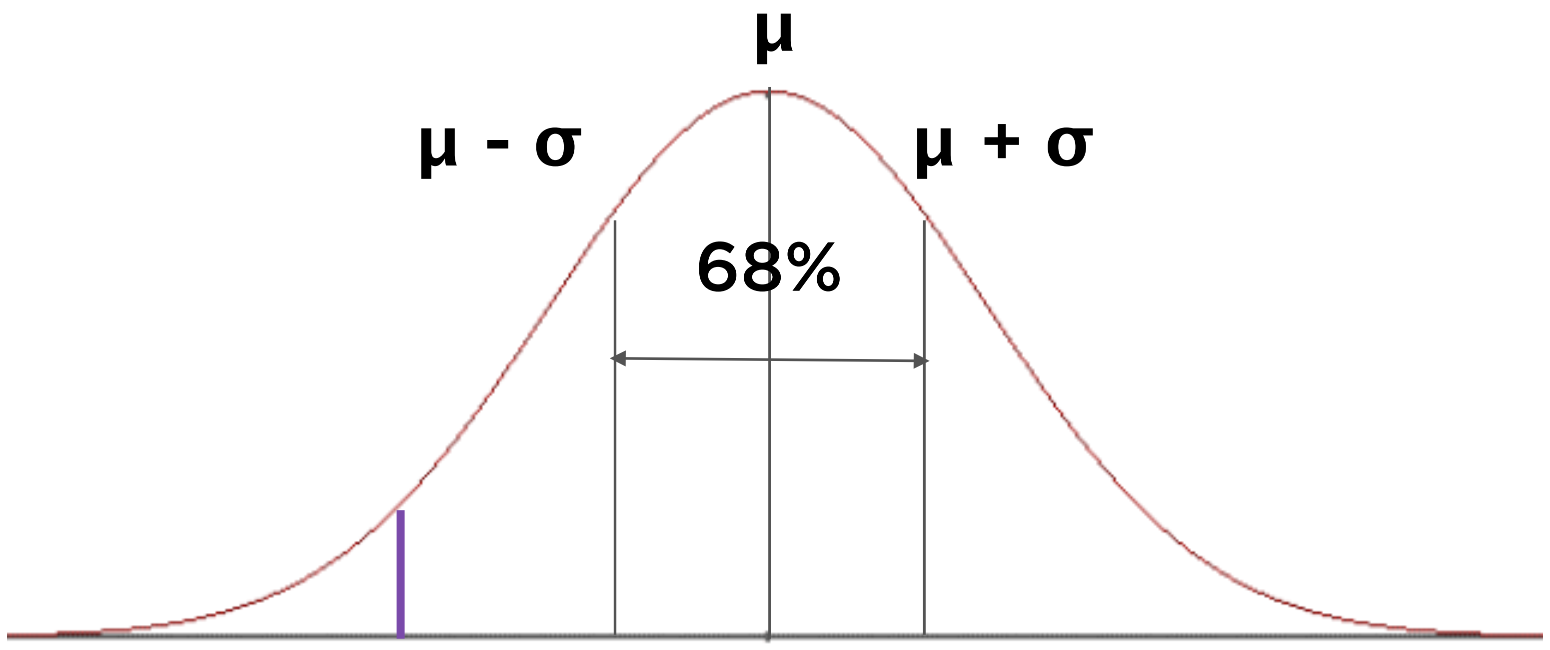
Negative Skewness



Kurtosis

Measure of how often extreme values (on either side of the mean) occur

Gaussian Distribution



N(μ,σ)

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

Kurtosis

Normally distributed data: kurtosis = 3

Excess kurtosis = kurtosis - 3

Kurtosis

Kurtosis ~ Tail risk

High kurtosis => extreme events more likely than in normal distribution

Kurtosis

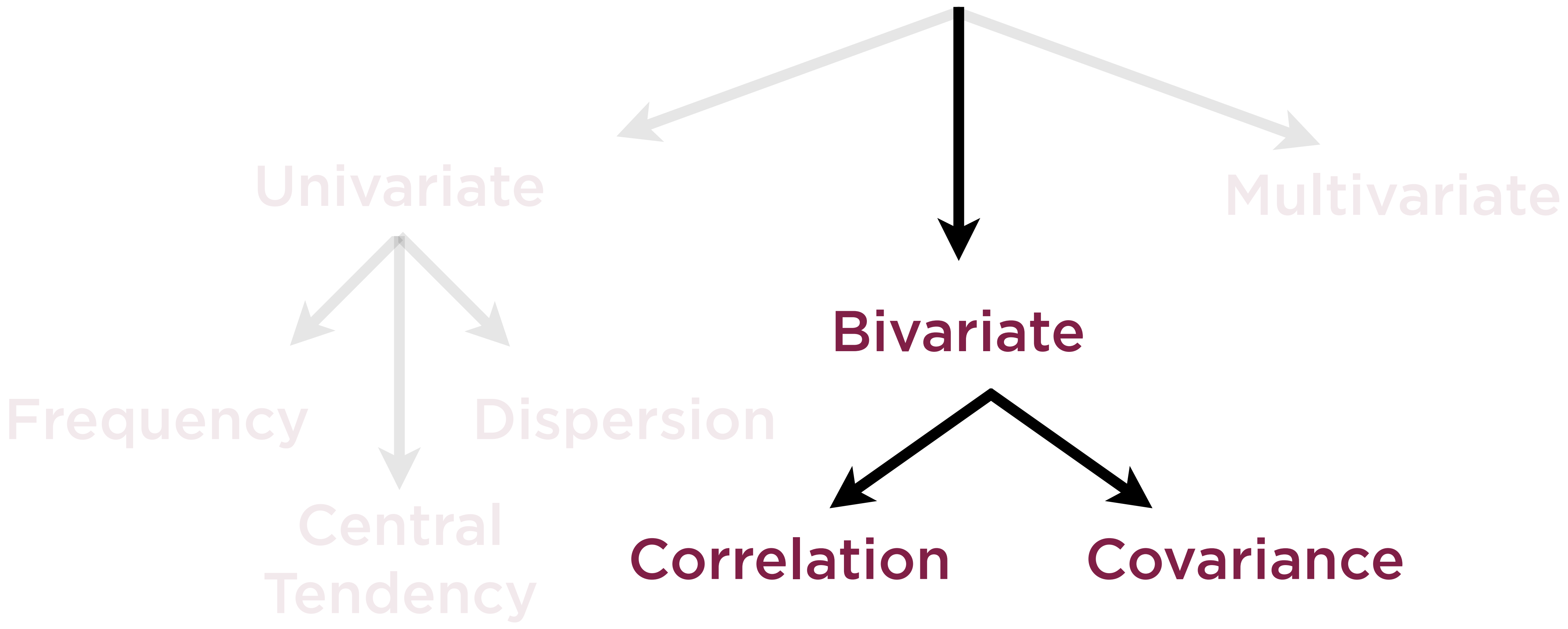
2008 Financial Crisis:

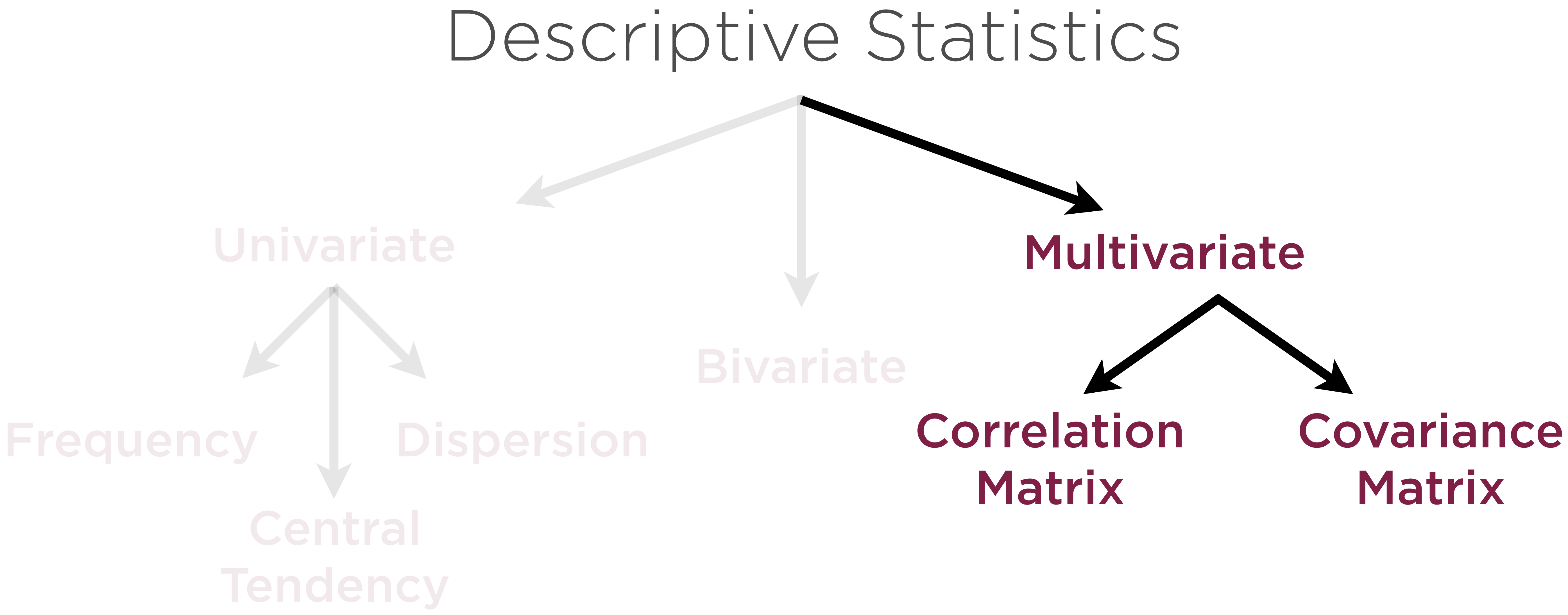
Several once-in-a-century events, all in 1 month

- Risk models were incorrectly assuming markets are normal
- In reality, market returns display significant excess kurtosis

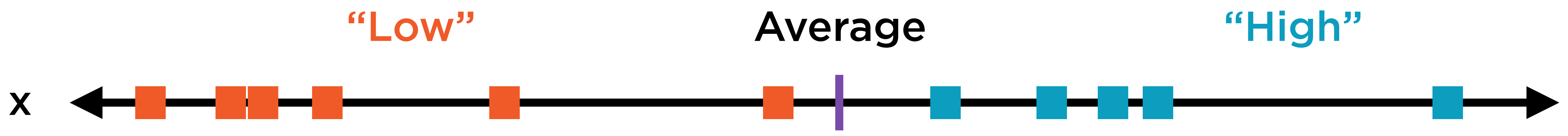
Covariance and Correlation

Descriptive Statistics




Descriptive Statistics
Univariate
Bivariate
Multivariate
Frequency
Dispersion
Central Tendency
Correlation Matrix
Covariance Matrix

Data in One Dimension



Unidimensional data is analyzed using statistics such as mean, median, standard deviation

Data in Two Dimensions

It's often more insightful to view data in relation to some other, related data

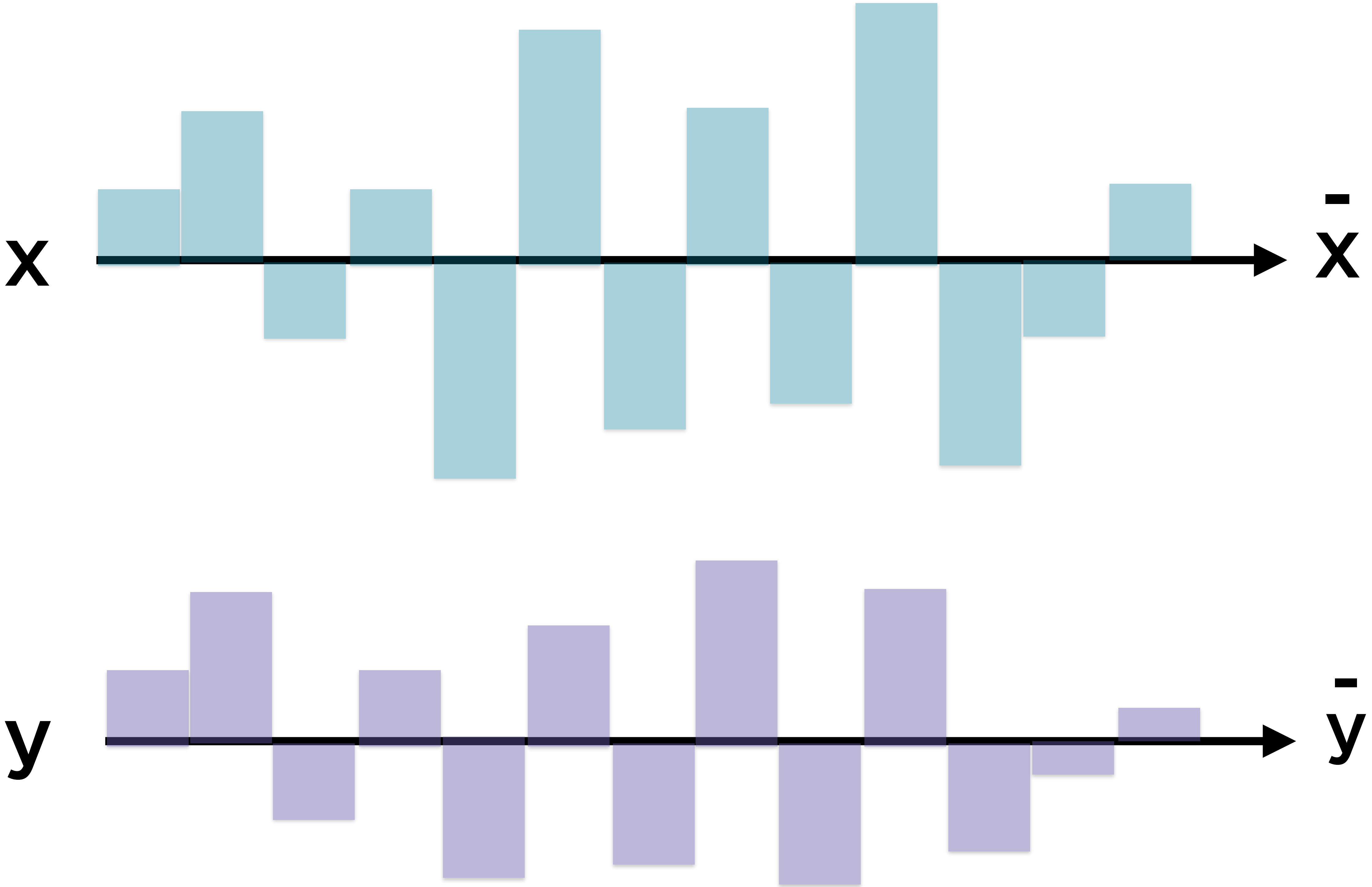
Covariance

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

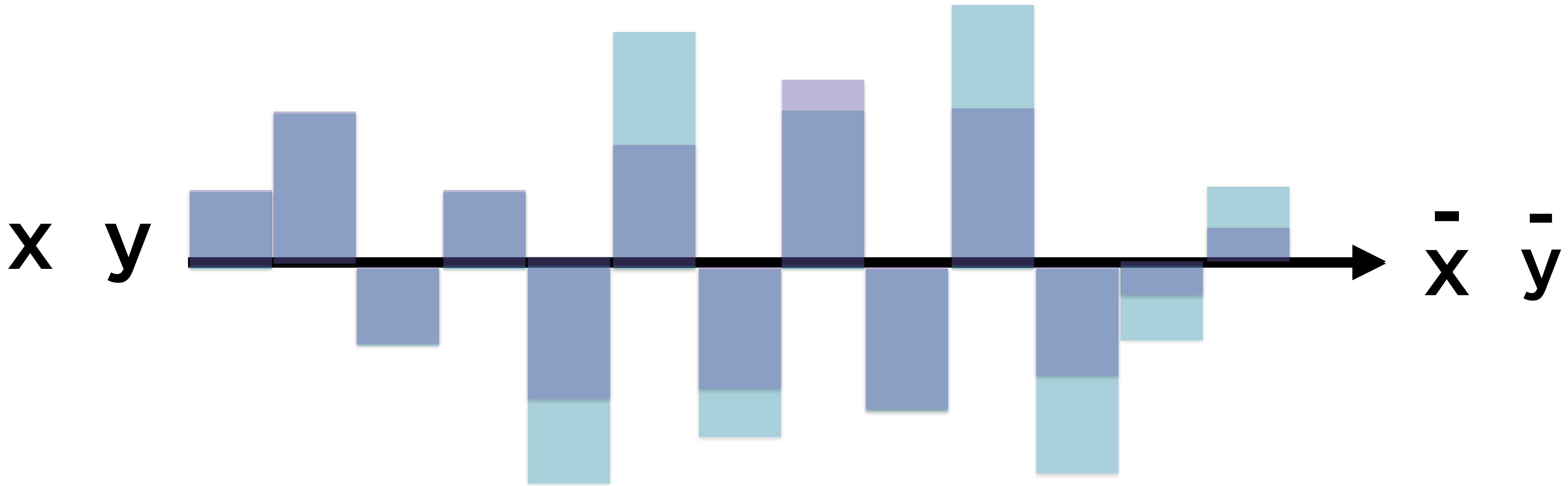
Covariance

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

Intuition: Positive Covariance

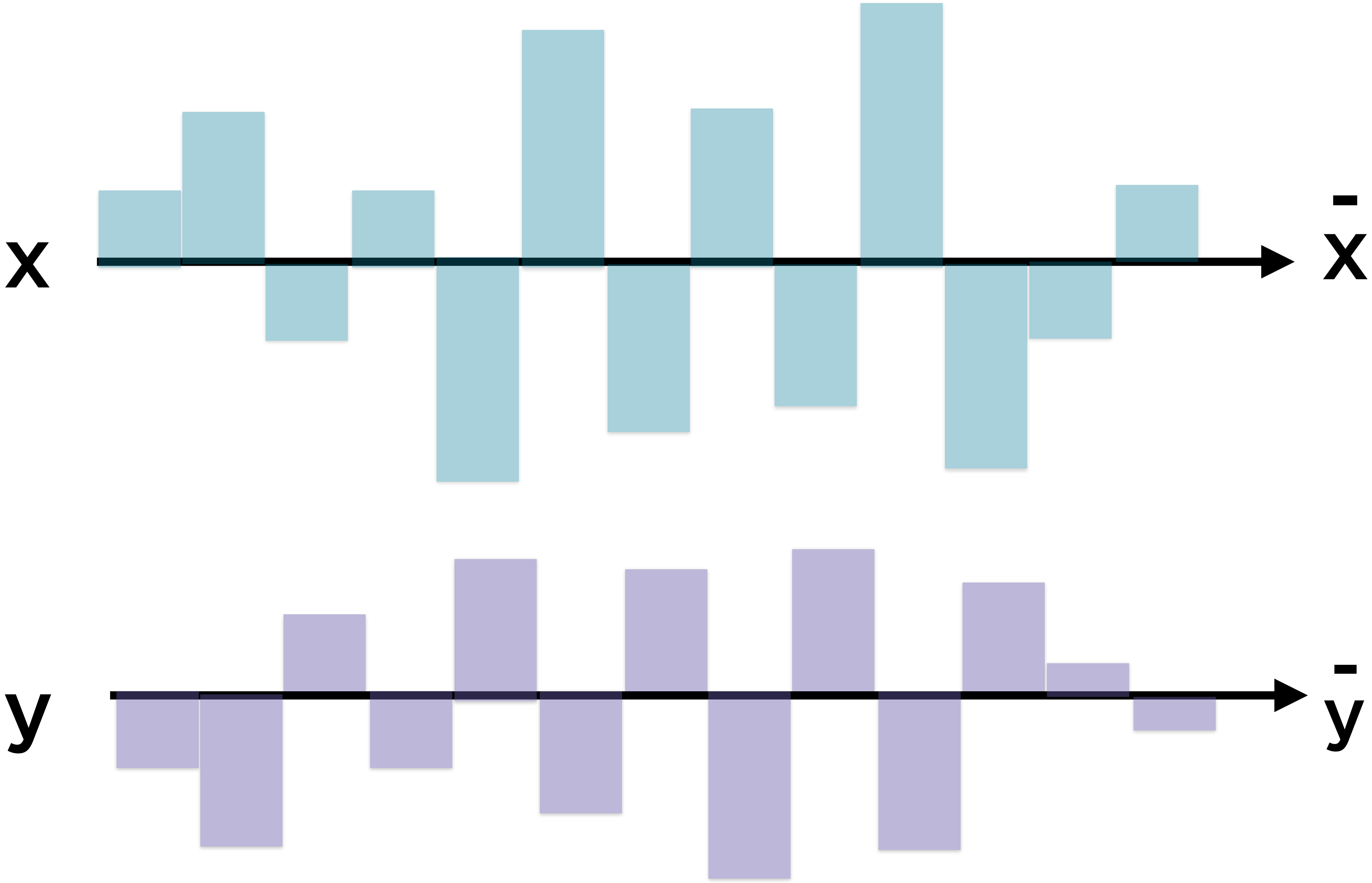


Intuition: Positive Covariance

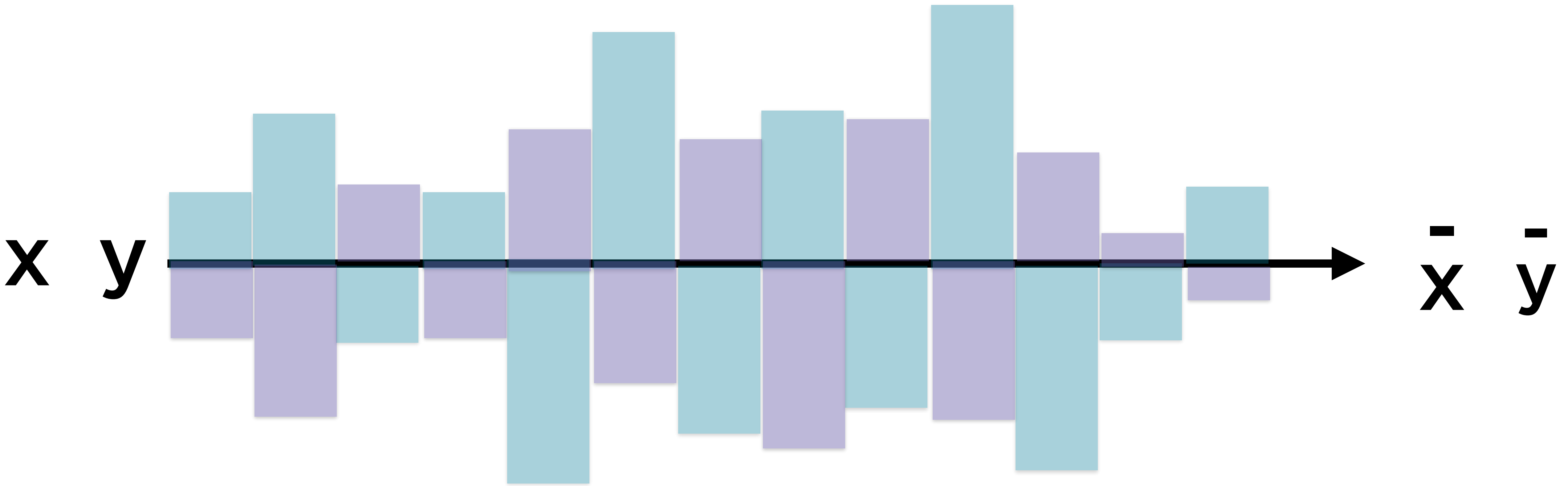


The deviations around the means of the two series are in sync

Intuition: Negative Covariance

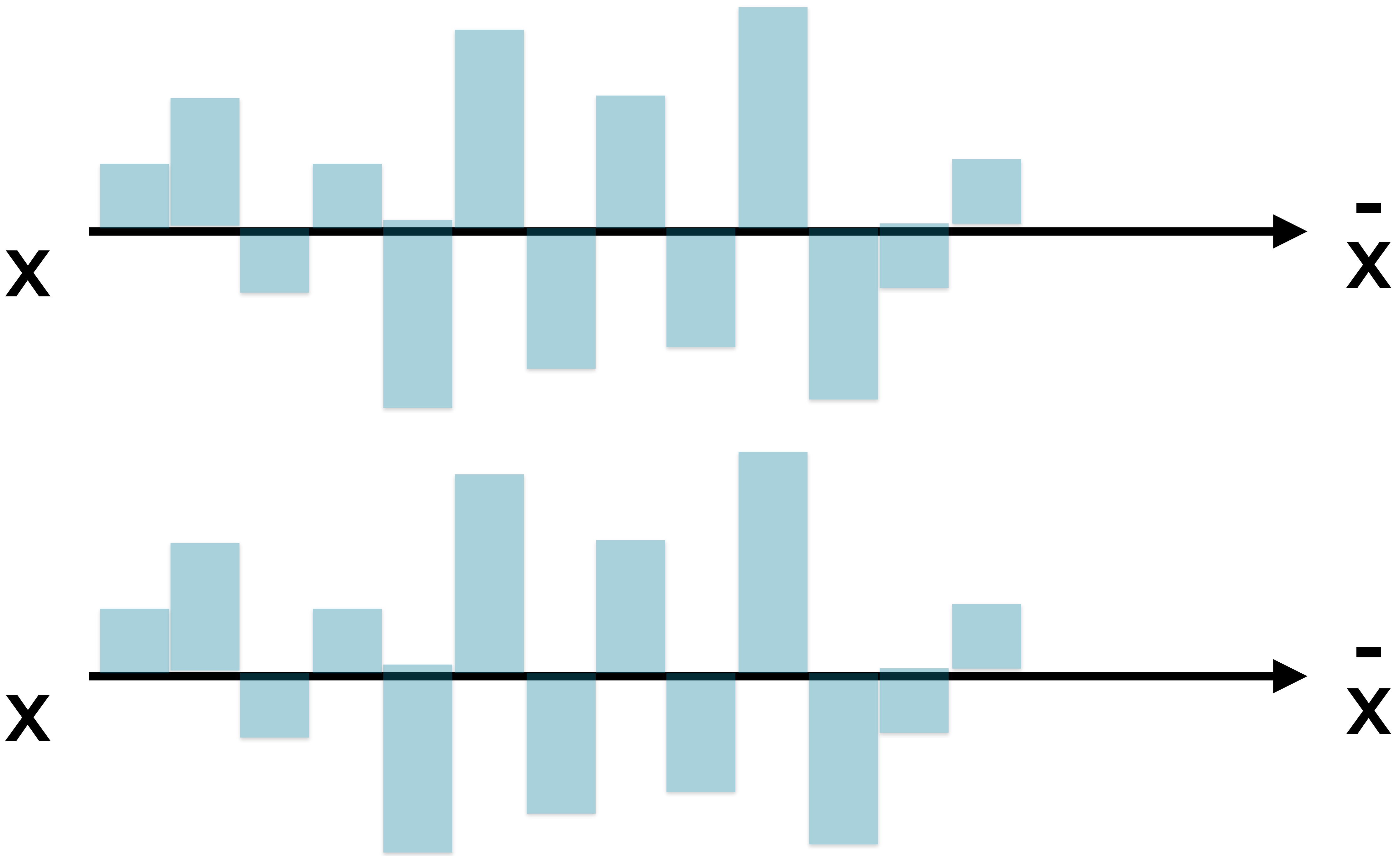


Intuition: Negative Covariance

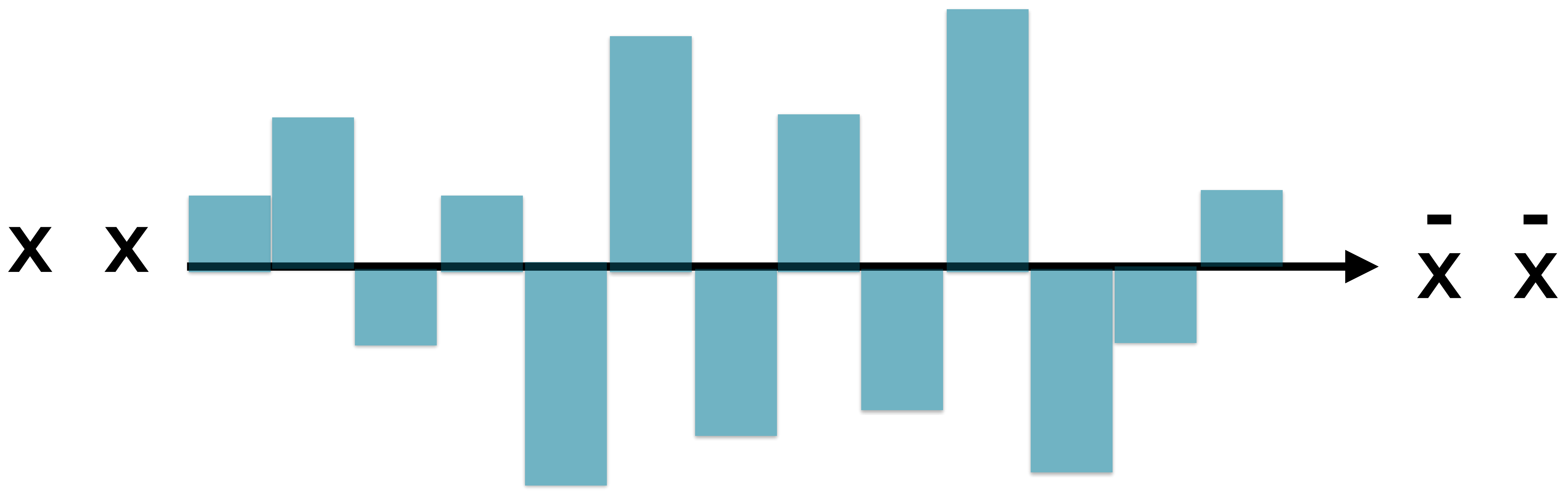


The deviations around the means of the two series are out of sync

Intuition: Covariance and Variance



Intuition: Positive Covariance



Variance is the covariance of a series with itself

A covariance matrix summarizes the covariances of columns in a data matrix

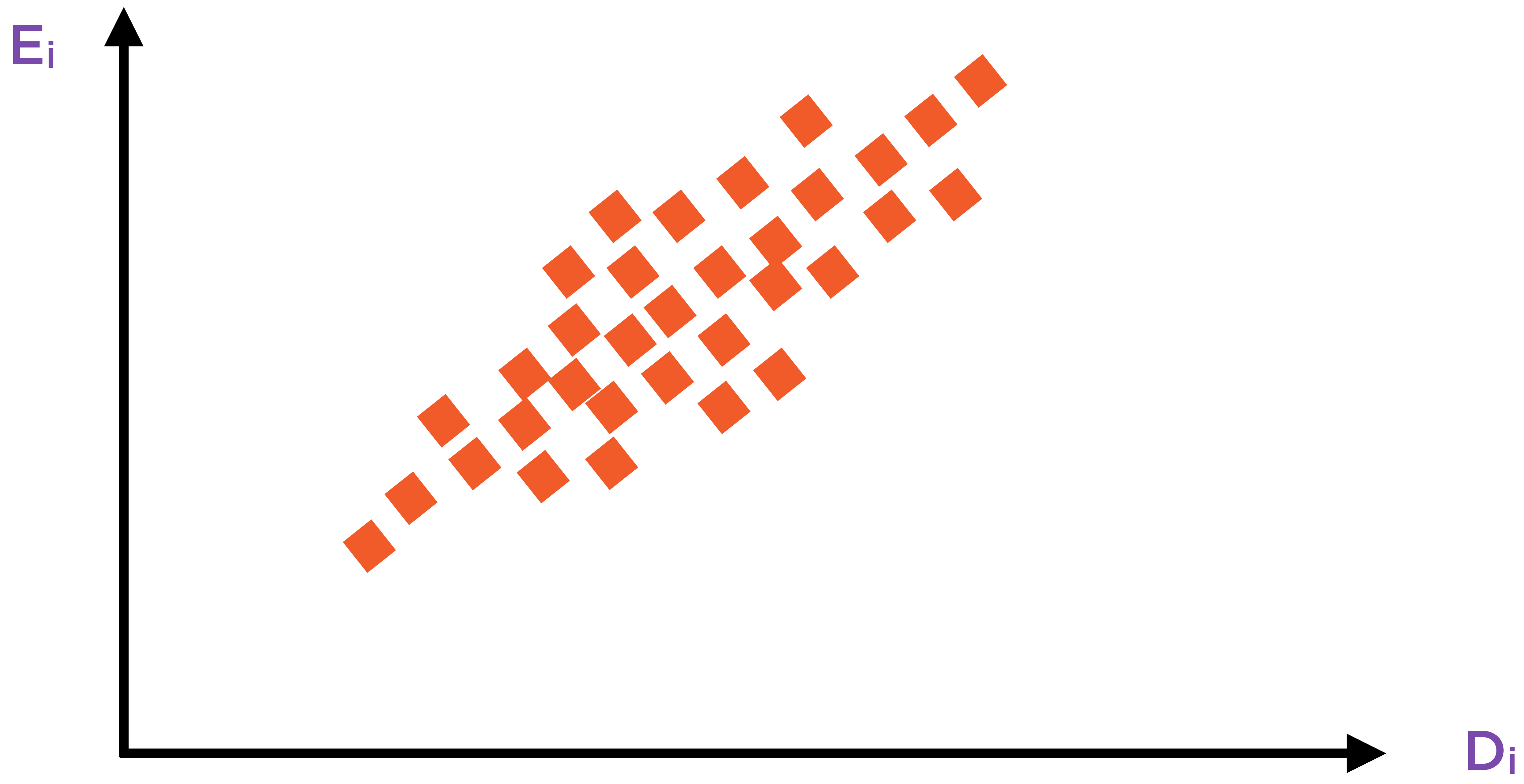
Correlation

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.

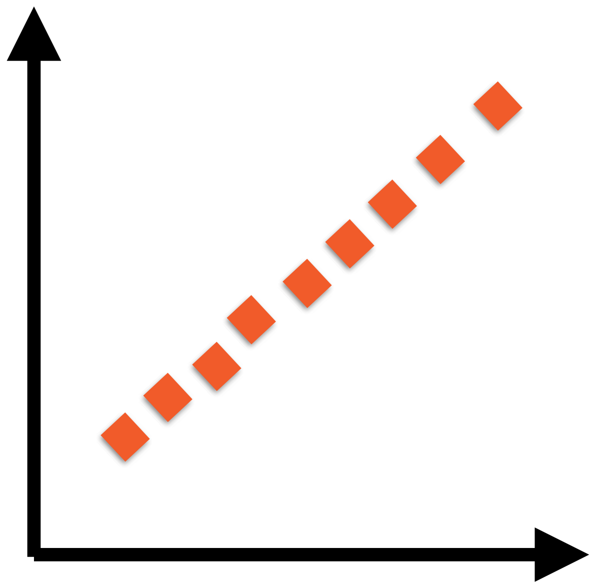
Correlation

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.

Correlated Random Variables

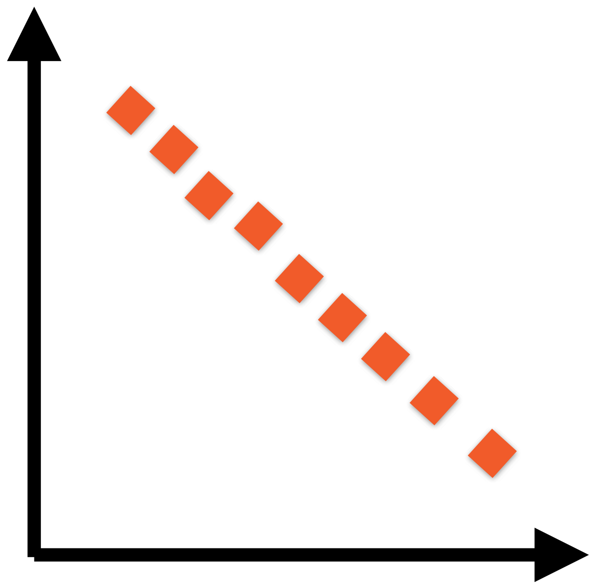


Correlation Captures Linear Relationships



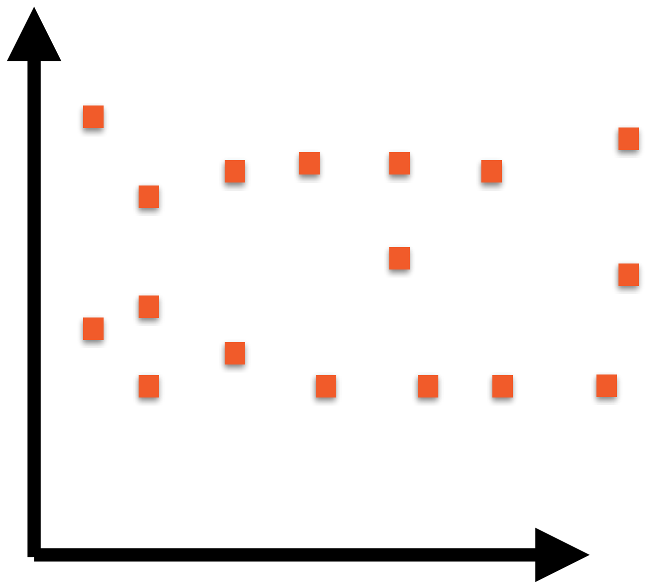
Correlation = +1

As X increases, Y increases linearly



Correlation = -1

As X increases, Y decreases linearly



Correlation = 0

Changes in X independent* of changes in Y

Correlation and Covariance

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

Independent variables have zero covariance and zero correlation

Summary

Descriptive statistics are used to explore and describe data

Measures of central tendency

Measures of dispersion

Confidence intervals of a measure

Skewness and kurtosis

Bivariate measures such as covariance and correlation