Building Statistical Models Using StatsModels

EXPLORING STATISTICAL PROPERTIES USING STATSMODELS

Janani Ravi

CO-FOUNDER, LOONYCORN

www.loonycorn.com

Overview

Python package with implementations of statistical models and tests

T-tests to compare population means

One-way ANOVA for multiple categories

Two-way ANOVA for multiple categorical independent variables

Using ANOVA to analyze regression models

Skewness and kurtosis in data

Prerequisites and Course Outline

Prerequisite Courses

Python: Getting Started

Python Fundamentals

Working with Multidimensional Data Using NumPy

Software and Skills

Basic understanding of Python programming using Python3

NumPy, Matplotlib

Working with Jupyter notebooks

Basic understanding of statistics

Course Outline

Statistical data exploration

- Basics of hypothesis testing
- T-test, ANOVA
- Skewness, kurtosis

Linear models

- Weighted Least Squares
- Generalized Linear Models
- Robust Linear Models

Time series models

- ACF and PACF
- Autoregressive and moving average process
- ARMA models

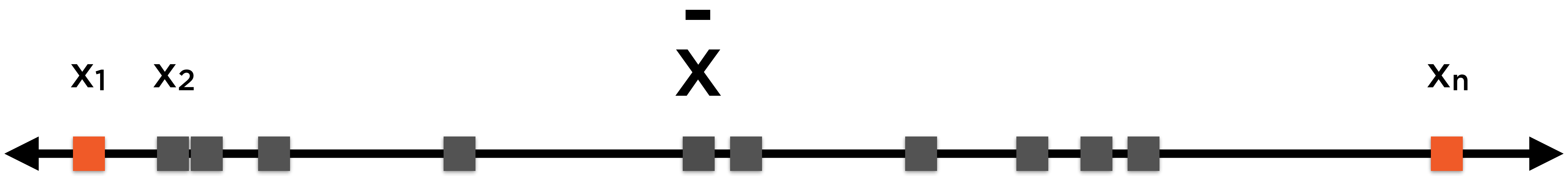
Standardizing Data: Mean and Variance

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

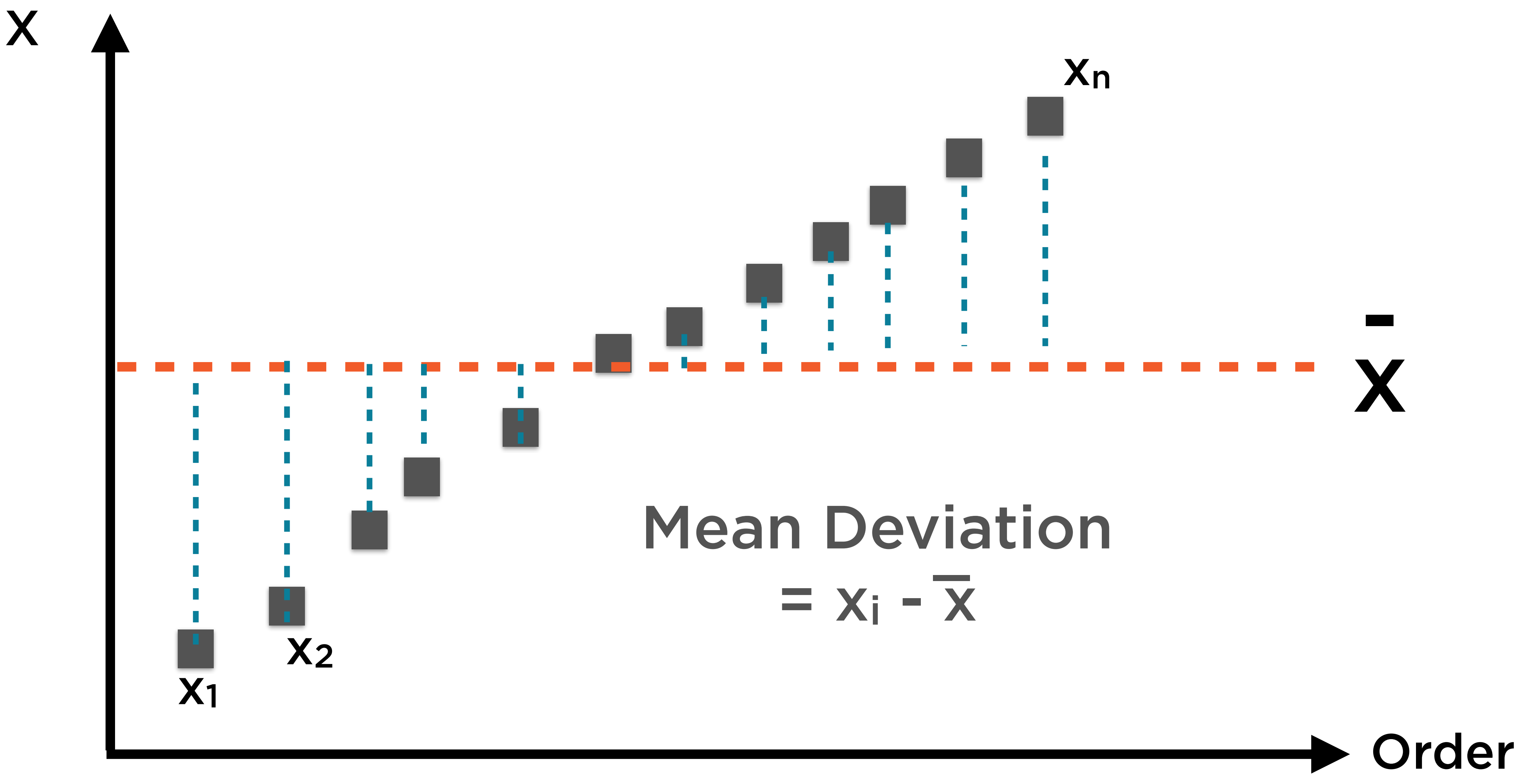


“Do the numbers jump around?”

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

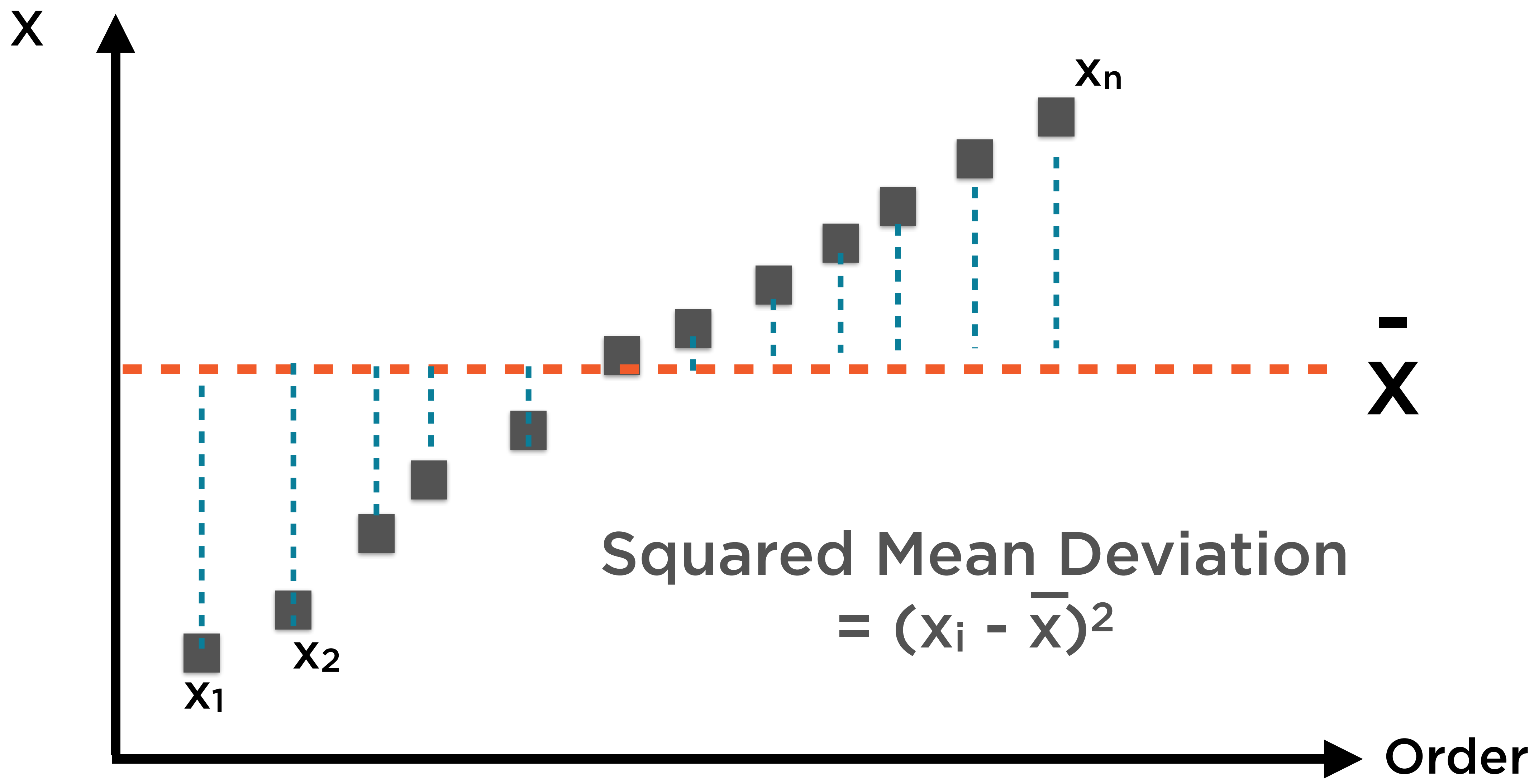
The range ignores the mean, and is swayed by outliers - that’s where variance comes in

Variance as Asterisk



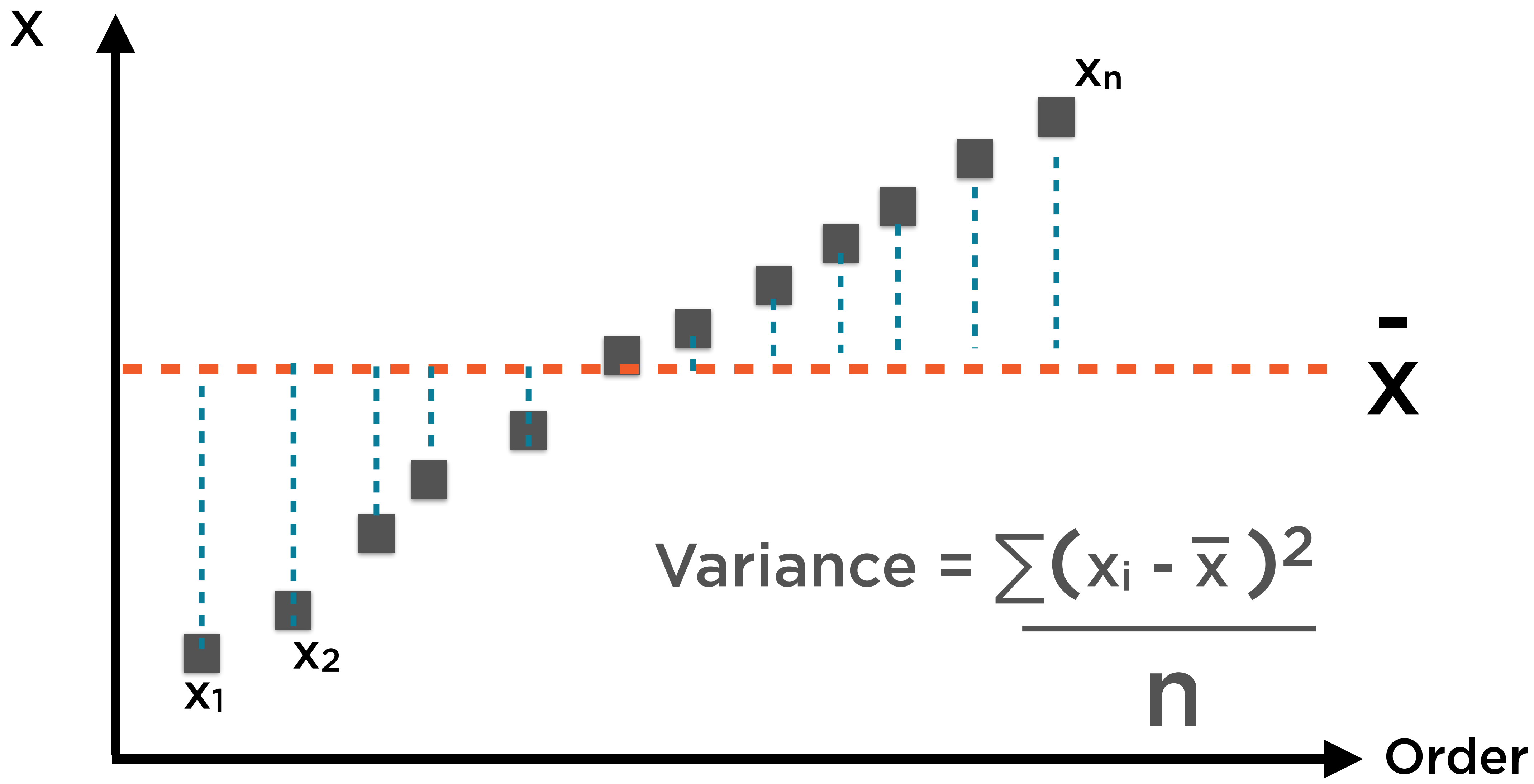
Variance is the second-most important number to summarize this set of data points

Variance as Asterisk



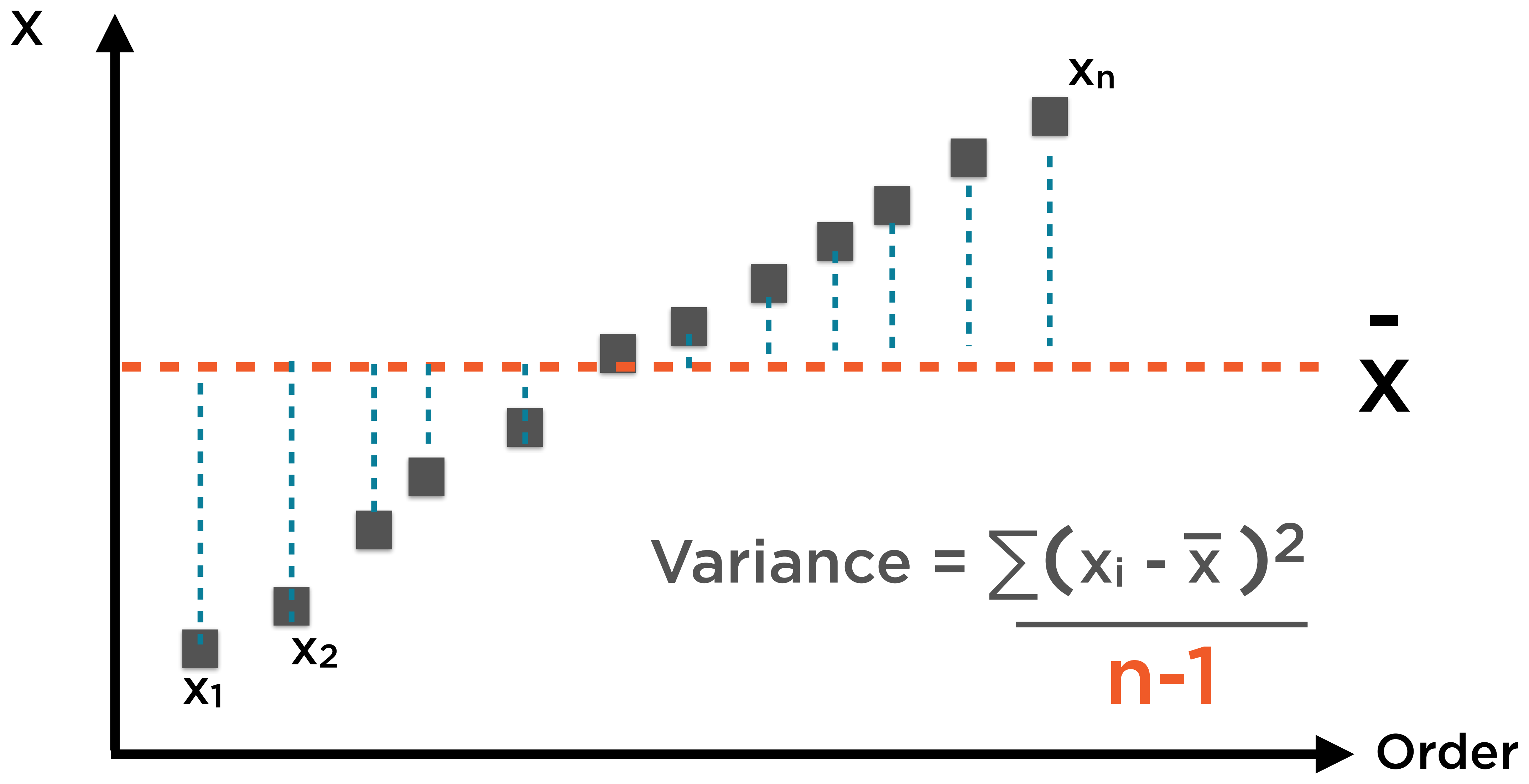
Variance is the second-most important number to summarize this set of data points

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 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

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

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

Standard Deviation as Risk

Standard deviation is the most common way to estimate the uncertainty of a set of outcomes

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

Standardizing Data

$$\begin{bmatrix} X_{11} & & X_{1k} \\ X_{21} & & X_{2k} \\ X_{31} & \cdots & X_{3k} \\ \cdots & & \cdots \\ X_{n1} & & X_{nk} \end{bmatrix}$$

$avg(X_1)$ ... $avg(X_k)$

$stdev(X_1)$ ... $stdev(X_k)$

Standardizing Data

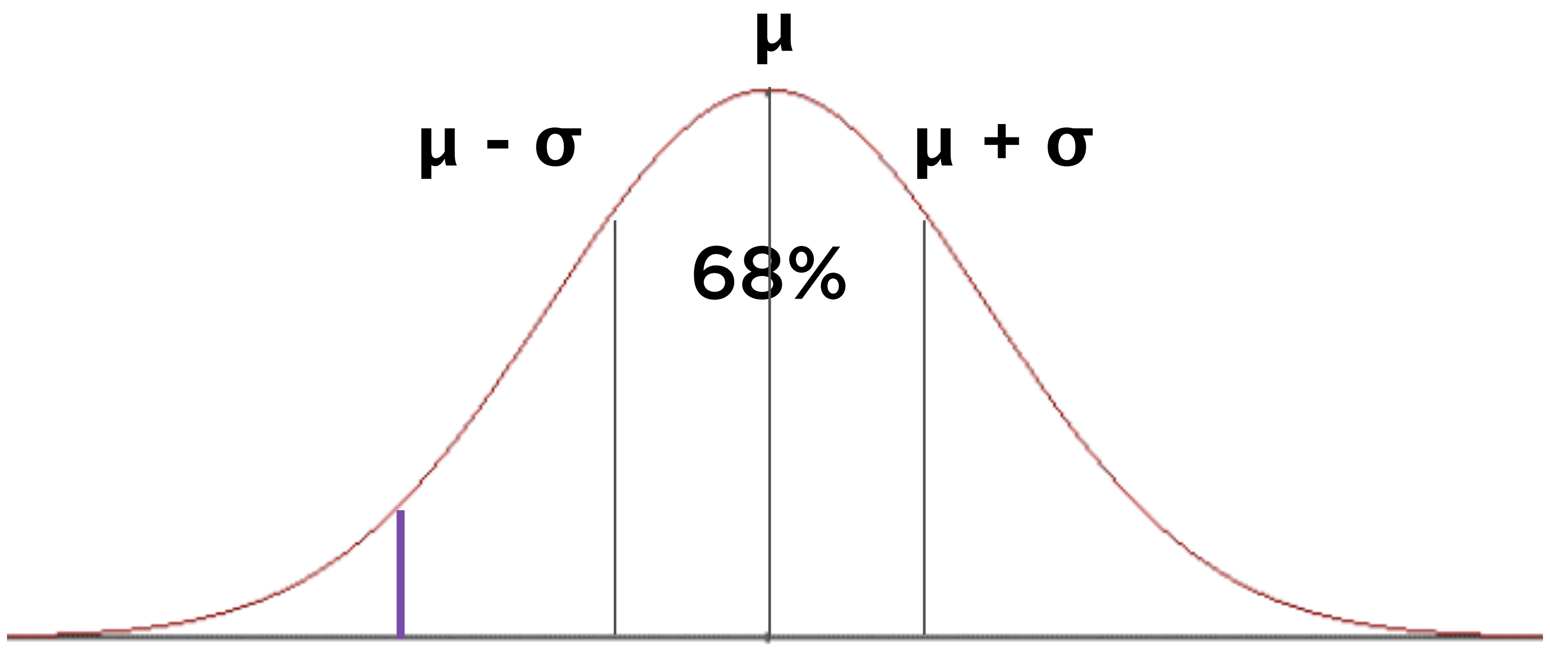
$$\begin{bmatrix} \frac{x_{11} - avg(X_1)}{stdev(X_1)} & \dots & \frac{x_{1k} - avg(X_k)}{stdev(X_k)} \\ \dots & & \dots \\ \frac{x_{n1} - avg(X_1)}{stdev(X_1)} & & \frac{x_{nk} - avg(X_k)}{stdev(X_k)} \end{bmatrix}$$

Each column of the standardized data has mean 0 and variance 1

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

Gaussian distribution

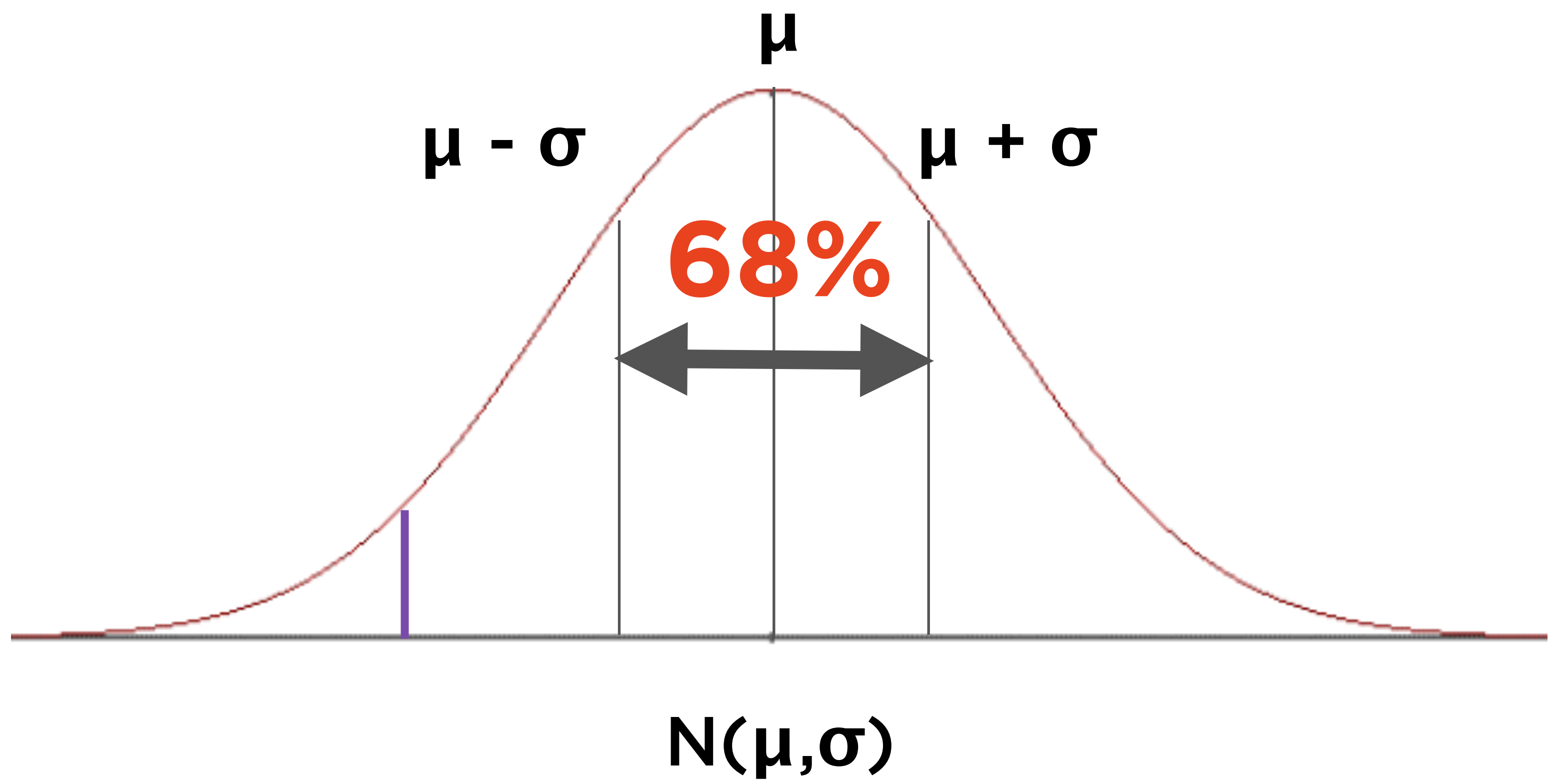
Gaussian Distribution



N(μ,σ)

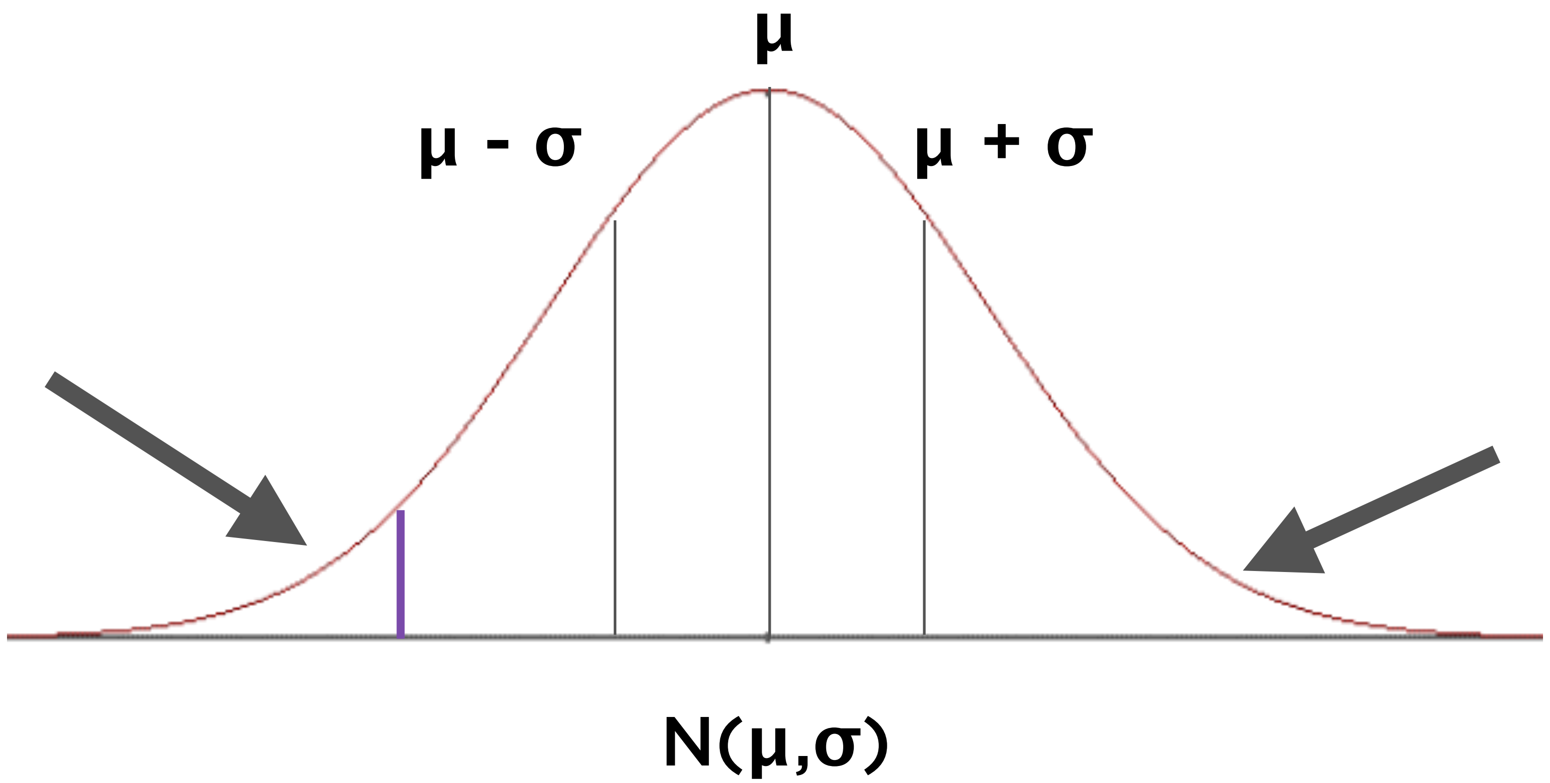
$$N(\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Gaussian Distribution



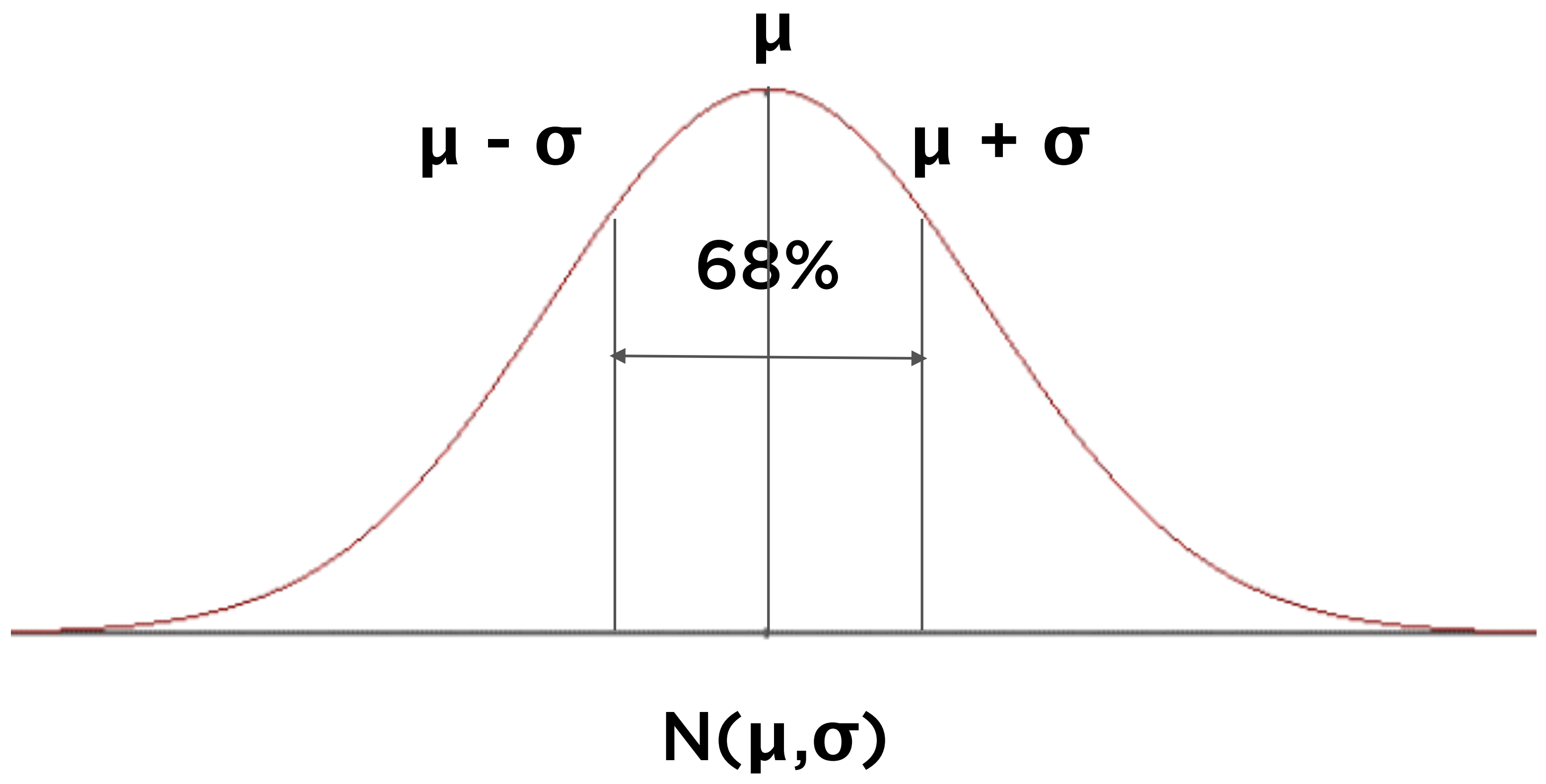
There will be a large number of points close to the average

Gaussian Distribution



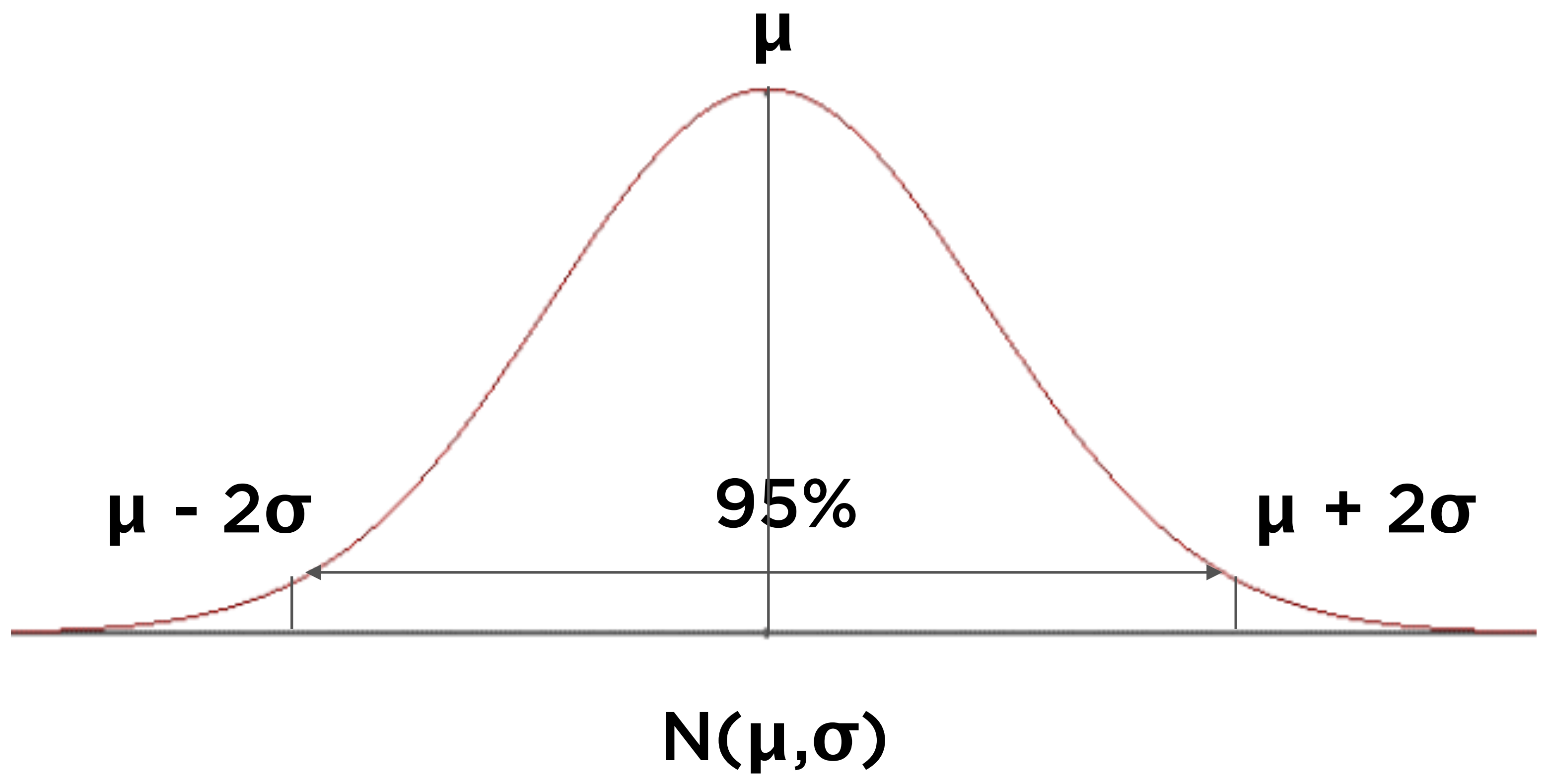
There will be few extreme values - the number of extreme values at either side of the mean will be the same

Gaussian Distribution



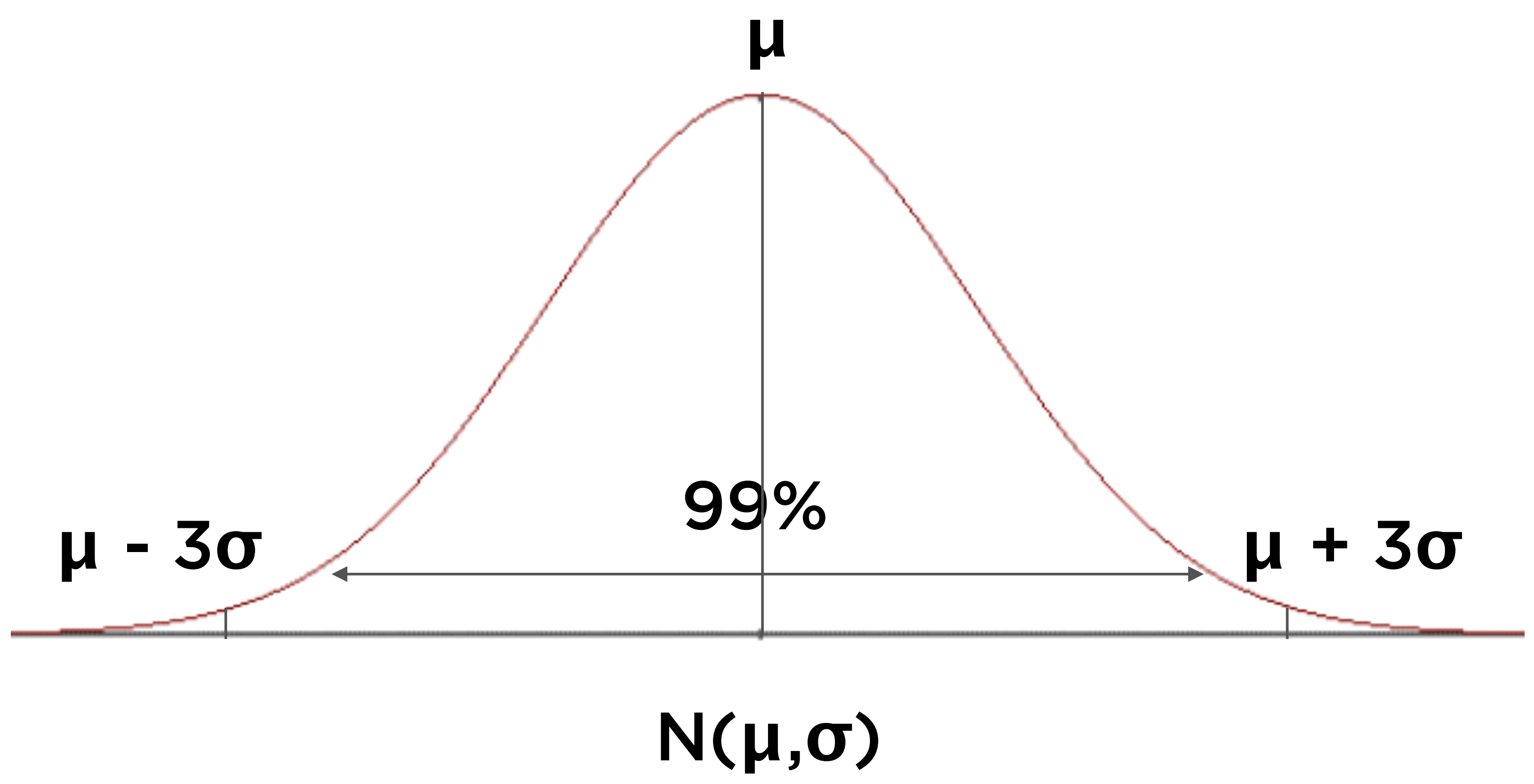
68% within 1 standard deviation of mean

Gaussian Distribution



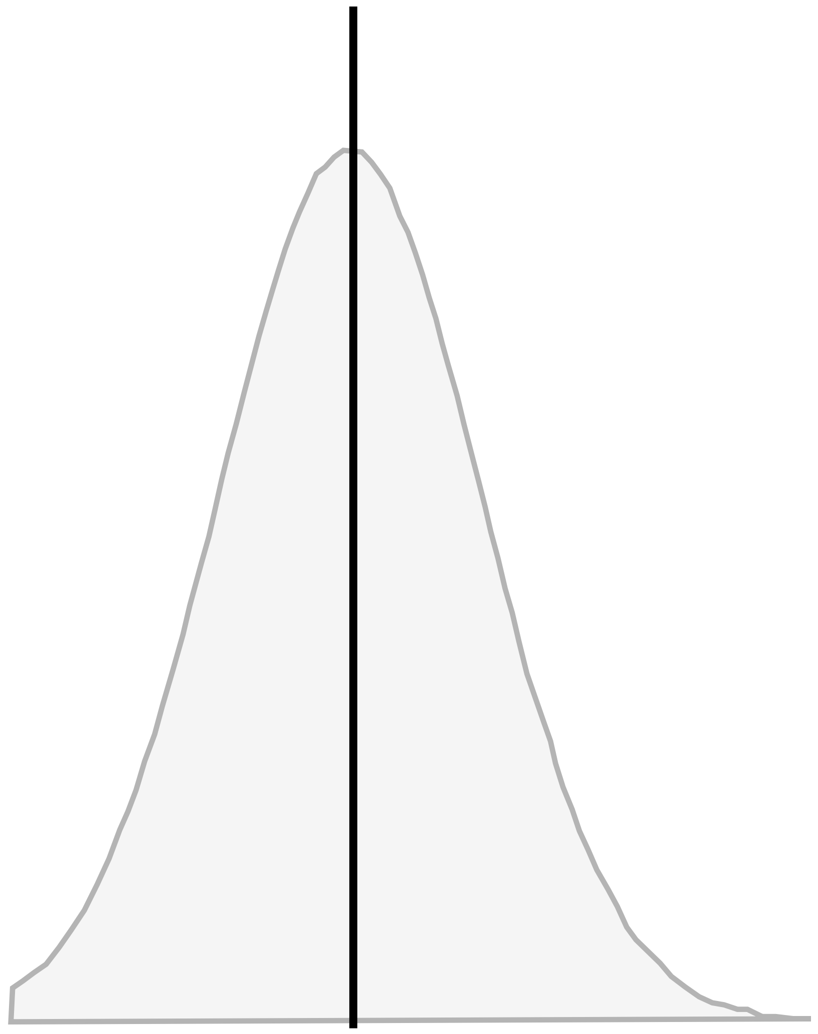
95% within 2 standard deviations of mean

Gaussian Distribution



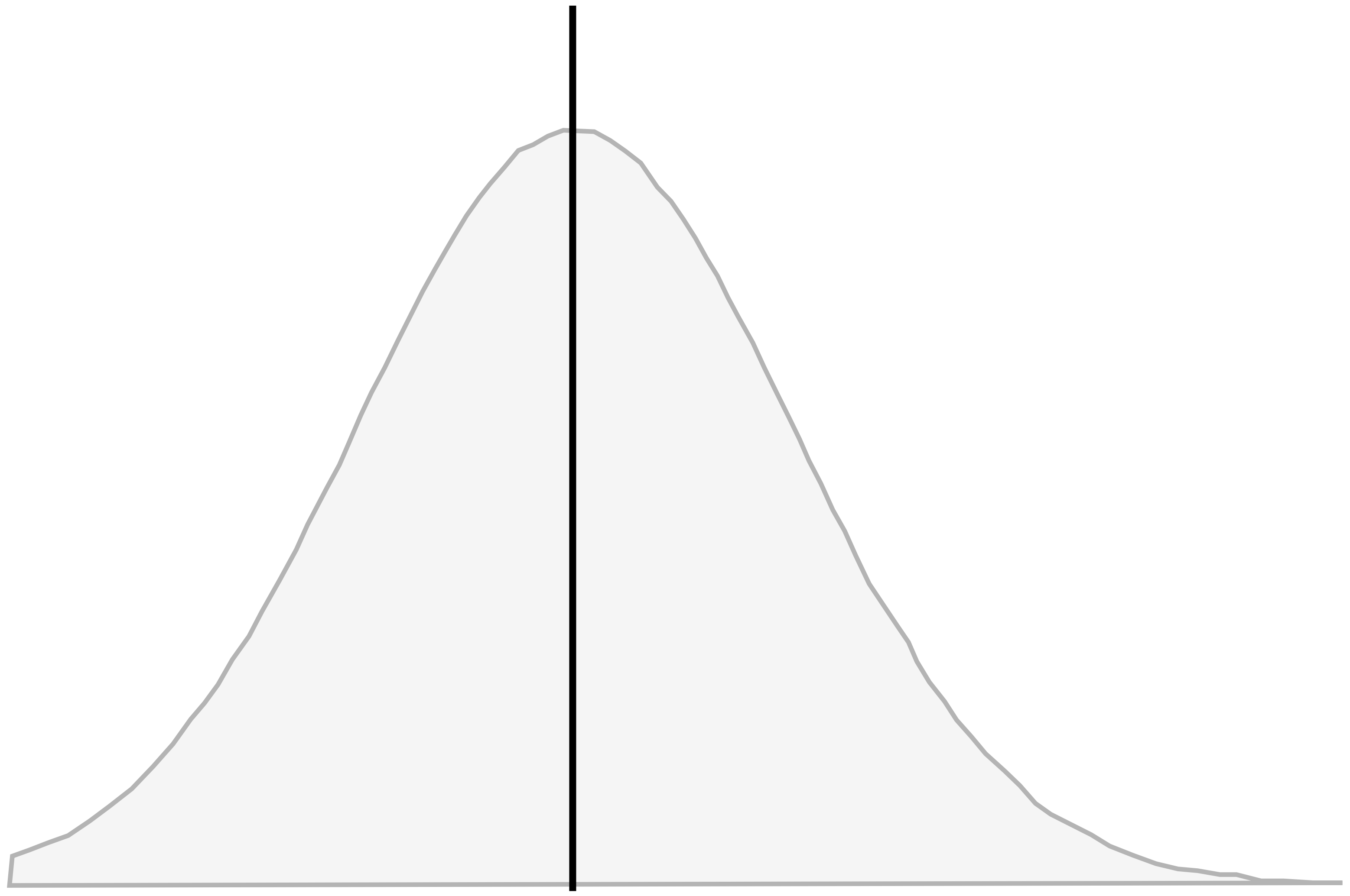
99% within 3 standard deviations of mean

Role of Sigma



Small Standard Deviation

Few points far from the mean



Large Standard Deviation

Many points far from the mean

Hypothesis Testing

Hypothesis

Proposed explanation for a phenomenon

Hypothesis

Proposed explanation

Objectively testable

Singular - hypothesis

Plural - hypotheses

Hypothesis Testing

Null Hypothesis H_0

True until proven false

Usually posits no relationship

Alternative Hypothesis

Negation of null hypothesis

Usually asserts specific relationship

Select Test

Pick from vast library

Know which one to choose

Test Statistic

Convert to p-value

How likely it was just luck?

Significance Level

Usually 1% or 5%

What threshold for luck?

Accept or Reject

Small p-value? Reject

Small: Below significance level

Lady Tasting Tea

Lady tasting tea: famous experiment

Was tea added before or after milk?

Muriel Bristol claimed she could tell

Lady Tasting Tea

Null Hypothesis

(H_0)

The lady cannot tell if milk was poured first

Alternate Hypothesis

(H_1)

The lady can tell if milk was poured first

Lady Tasting Tea

Null Hypothesis

The lady **cannot** tell if the milk was poured first

Alternate Hypothesis

The lady **can** tell if the milk was poured first

It is good practice to assume that the null

Lady Tasting Tea

Null Hypothesis

The lady **cannot** tell if the milk was poured first

Alternate Hypothesis

The lady **can** tell if the milk was poured first

It is good practice to assume that the null hypothesis is correct unless proven otherwise

Lady Tasting Tea

Null Hypothesis H_0
"Lady cannot tell difference"

Can't tell if milk poured first

Alternative Hypothesis
"Lady can tell difference"

Can indeed discern if milk poured first

Select Test
8 cups, 4 of each type

Lady got all 8 correct

Test Statistic
p-value = 1/70 = 1.4%

$^{8}C_{4}$ = 70 combinations

Significance Level
Choose 5% significance level

Part of design of experiment

Accept or Reject
1.4% < 5% => Reject H_0

Lady can indeed tell difference

Lady Tasting Tea

Experiment proved that she could

Conducted by Sir Ronald Fisher

(considered founder of modern statistics)

Errors in Hypothesis Testing

		Decision about Null Hypothesis	
		REJECT	DON'T REJECT
Null Hypothesis is actually	TRUE	Type I error	Correct Inference
	FALSE	Correct Inference	Type II error

Errors in Hypothesis Testing

		Decision about Null Hypothesis	
		REJECT	DON'T REJECT
Null Hypothesis is actually	TRUE	Type I error	Correct Inference
	FALSE	Correct Inference	Type II error

Claim the lady can tell the difference based on spurious test results which are not statistically significant

Errors in Hypothesis Testing

		Decision about Null Hypothesis	
		REJECT	DON'T REJECT
Null Hypothesis is actually	TRUE	Type I error	Correct Inference
	FALSE	Correct Inference	**Type II error**

Fail to realize that the test for the alternative hypothesis was statistically significant

The T-test

Hypothesis Testing

Null Hypothesis H_0

True until proven false

Usually posits no relationship

Alternative Hypothesis

Negation of null hypothesis

Usually asserts specific relationship

Select Test

Pick from vast library

Know which one to choose

Test Statistic

Convert to p-value

How likely it was just luck?

Significance Level

Usually 1% or 5%

What threshold for luck?

Accept or Reject

Small p-value? Reject

Small: Below significance level

Hypothesis Testing

Null Hypothesis H_0

True until proven false

Usually posits no relationship

Select Test

Pick from vast library

Know which one to choose

Significance Level

Usually 1% or 5%

What threshold for luck?

Alternative Hypothesis

Negation of null hypothesis

Usually asserts specific relationship

Test Statistic

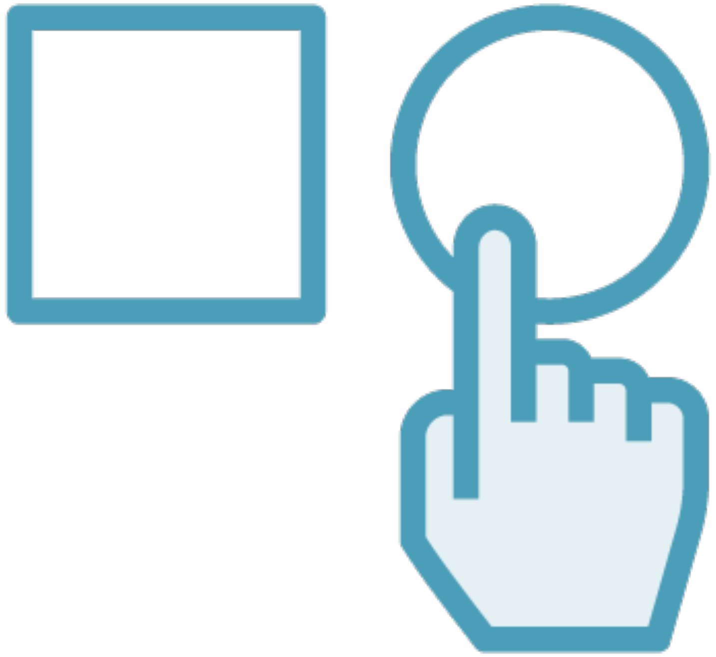
Convert to p-value

How likely it was just luck?

Accept or Reject

Small p-value? Reject

Small: Below significance level



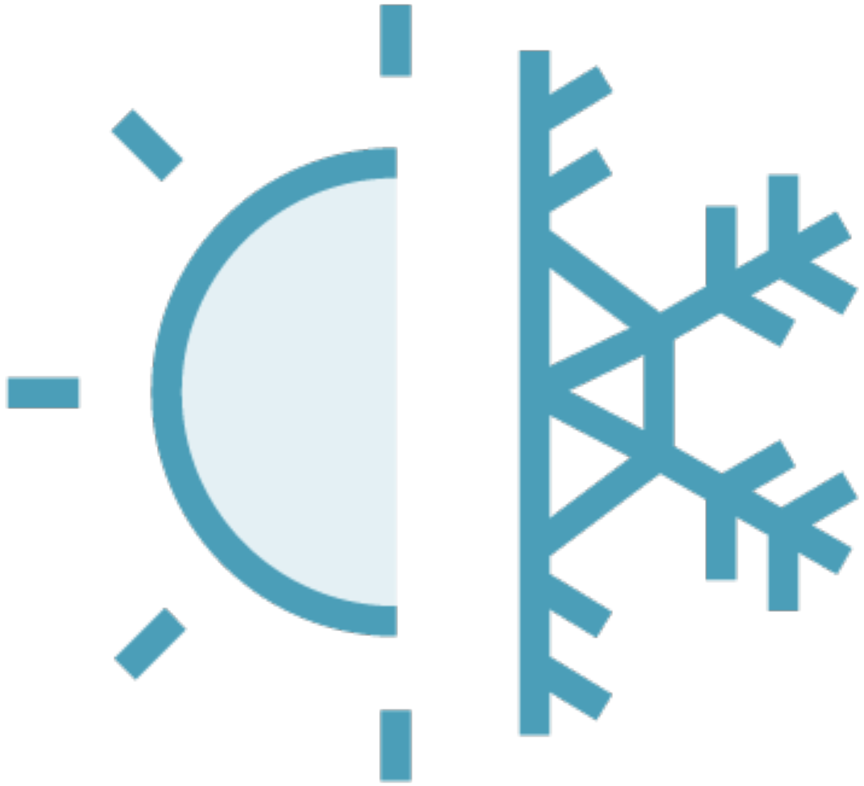
Statistical Test Selection

There are tests for pretty much everything

Developed by statisticians to be sound

Knowing which one to use is hard

Actually using them is relatively easy



T-tests

Most common, simple statistical tests out there

Used to learn about averages across two categories

Also tells whether the differences are significant

T-tests

Average male baby birth weight =
Average female baby birth weight?

Is the difference statistically significant?

T-tests

T-statistic

- Score which indicates the difference in means

P-value

- Whether the T-statistic is significant
- Low p-values of <5% mean the result cannot be due to chance

Types of T-tests

One sample location test

Two sample location test

Paired difference test

Regression coefficient test

One sample location test

One-sample location test

- What is the average weight of babies born in a certain town?
- Is it different from the average of the general population?

Two sample location test

Two-sample location test (independent samples t-test)

- Is the average weight of babies in Town A different from Town B?

Paired difference test

Paired difference test

- Is the average weight of babies born in winter different from babies born in summer?

Regression coefficient test

Regression coefficient test

- Is the coefficient of any of the independent variables > 0?

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

Population

All the data out there in the universe

Sample

A subset - hopefully representative - of the population

From Sample to Population

Population

Representative Sample

Biased Sample

From Sample to Population

Sample Mean

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

Population Mean

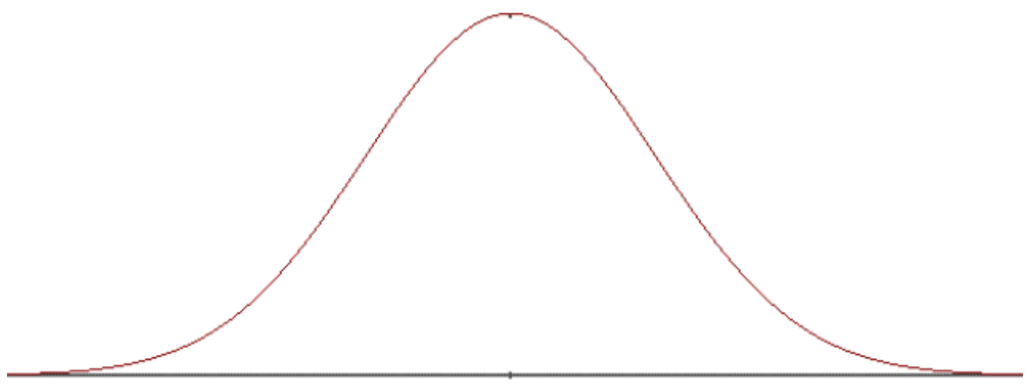
$$\mu = ?$$

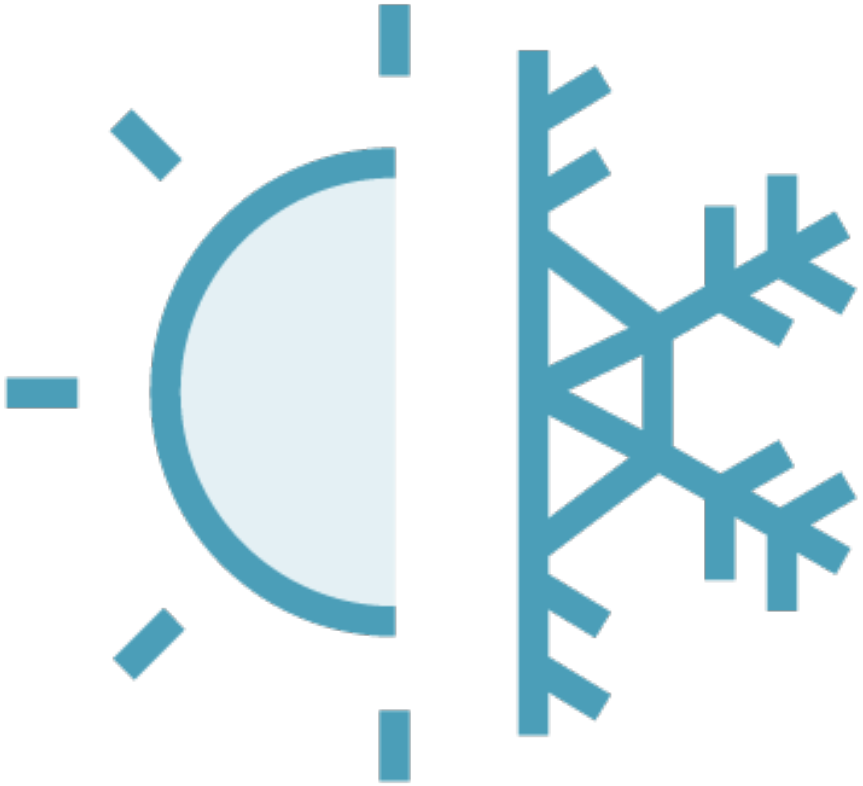
From Sample to Population

Sample Mean

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

Population Mean

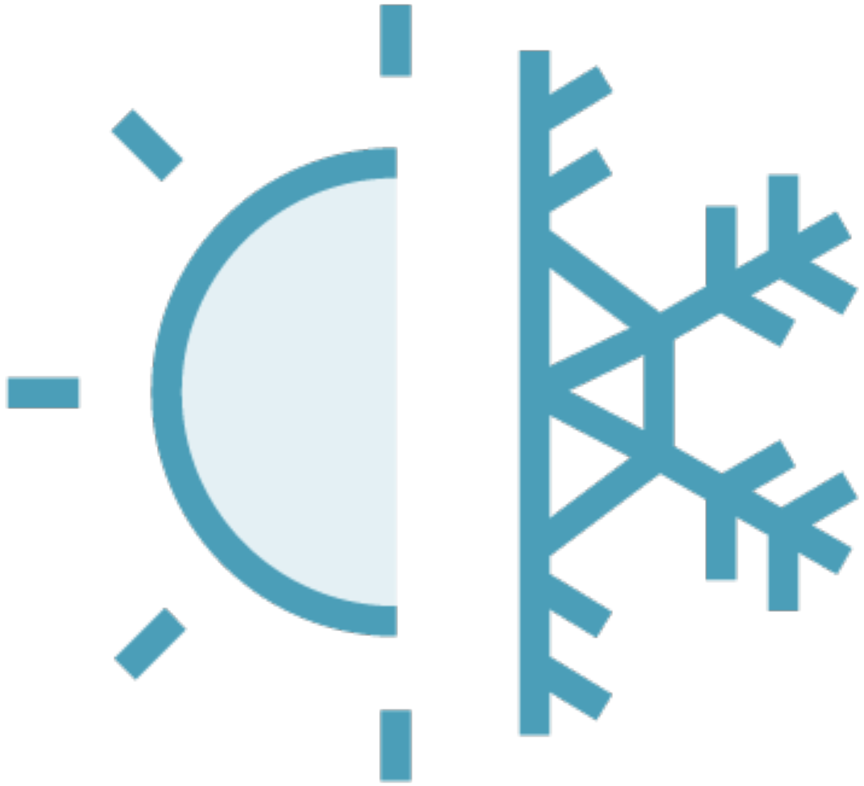




T-tests Assumptions

Notably, that

- populations are normal
- samples are representative
- samples are randomly drawn



T-tests

Work best for two group comparisons

Comparing multiple groups gets tricky

- need many pairwise tests
- increases likelihood of Type 1 error (alpha inflation)

For multiple groups, just use ANOVA

Demo

Performing T-tests

ANOVA

T-tests are useful to compare differences between **two** groups

Running **multiple** significance tests to compare across many groups is **risky**

ANOVA

<u>AN</u>alysis ***<u>O</u>***f ***<u>VA</u>***riance

ANOVA

Looks across multiple groups of populations, compares their means to produce one score and one significance value

ANOVA

Looks across multiple groups of populations, compares their means to produce one score and one significance value

Diabetes Risk

Underweight patients

Normal weight patients

Overweight patients

In order to compare across 3 groups the we’ll need to perform multiple T-tests

Diabetes Risk

Underweight patients	**Normal weight patients**	**Overweight patients**

Perform a single ANOVA test to know whether the risk of diabetes is significantly different between these groups

ANOVA Hypotheses

Null Hypothesis (H_0)	Alternate Hypothesis (H_1)
H_0: All groups of patients are at an equal risk of diabetes	H_0: All groups of patients are NOT at an equal risk of diabetes

ANOVA

Looks across multiple groups of populations, compares their means to produce one score and one significance value

F-statistic

$$F = \frac{\text{Variance between groups}}{\text{Variance within a group}}$$

F-statistic

If the groups are similar, F ~ 1

If the groups are different, F will be large

P-value

Significance of the F-statistic

Smaller p-values indicate that the results are not due to chance

Large F-statistic and small p-value - means the null hypothesis can be rejected

ANOVA Hypotheses

Large F-statistic and small p-values < 0.05 significance level

Accept the alternative hypothesis and reject the null hypothesis

Alternate Hypothesis

(H_1)

H_0: All groups of patients are NOT at an equal risk of diabetes

ANOVA Hypotheses

Null Hypothesis

(H_0)

H_0: All groups of patients are at an equal risk of diabetes

Small F-statistic and large p-values > 0.05 significance level

Accept the null hypothesis and reject the alternative hypothesis

One-way ANOVA helps compare means across two or more groups

A **single** categorical variable is used to split the population into these groups

One-way ANOVA Assumptions

Notably, that

- populations are normal
- samples are representative
- samples are randomly drawn
- variances of the population are constant

Assumptions in ANOVA

Residuals with normal distribution

Independence of errors

Absence of outliers

Homoscedasticity

Assumptions in ANOVA

Residuals with normal distribution	Independence of errors
Absence of outliers	Homoscedasticity

Distance of data points from the fitted values should be normally distributed

Assumptions in ANOVA

Residuals with normal distribution	**Independence of errors**
Absence of outliers	Homoscedasticity

Correlation between errors should be zero

Assumptions in ANOVA

Residuals with normal distribution	Independence of errors
Absence of outliers	Homoscedasticity

The normal distribution of the population implies no major outliers in data

Assumptions in ANOVA

Residuals with normal distribution	Independence of errors
Absence of outliers	**Homoscedasticity**

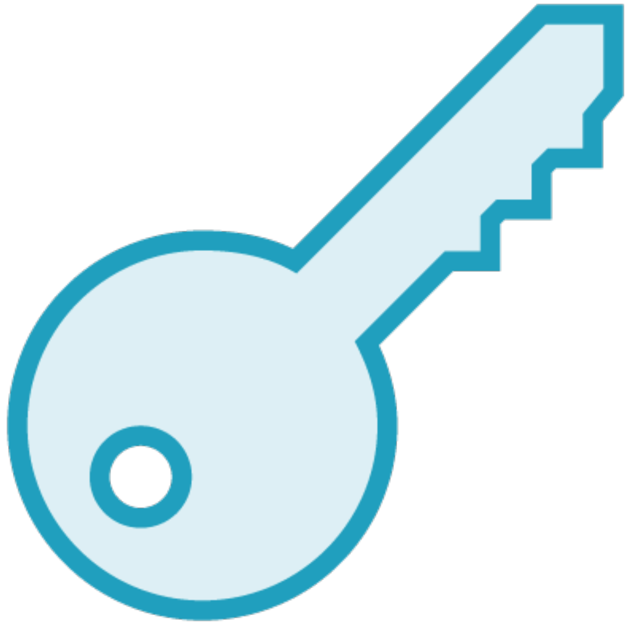
The variance in each group should be constant i.e. the same

Linear Regression

Ordinary Least Squares

Common technique used to find the best-fitting straight line through a set of points

X Causes Y



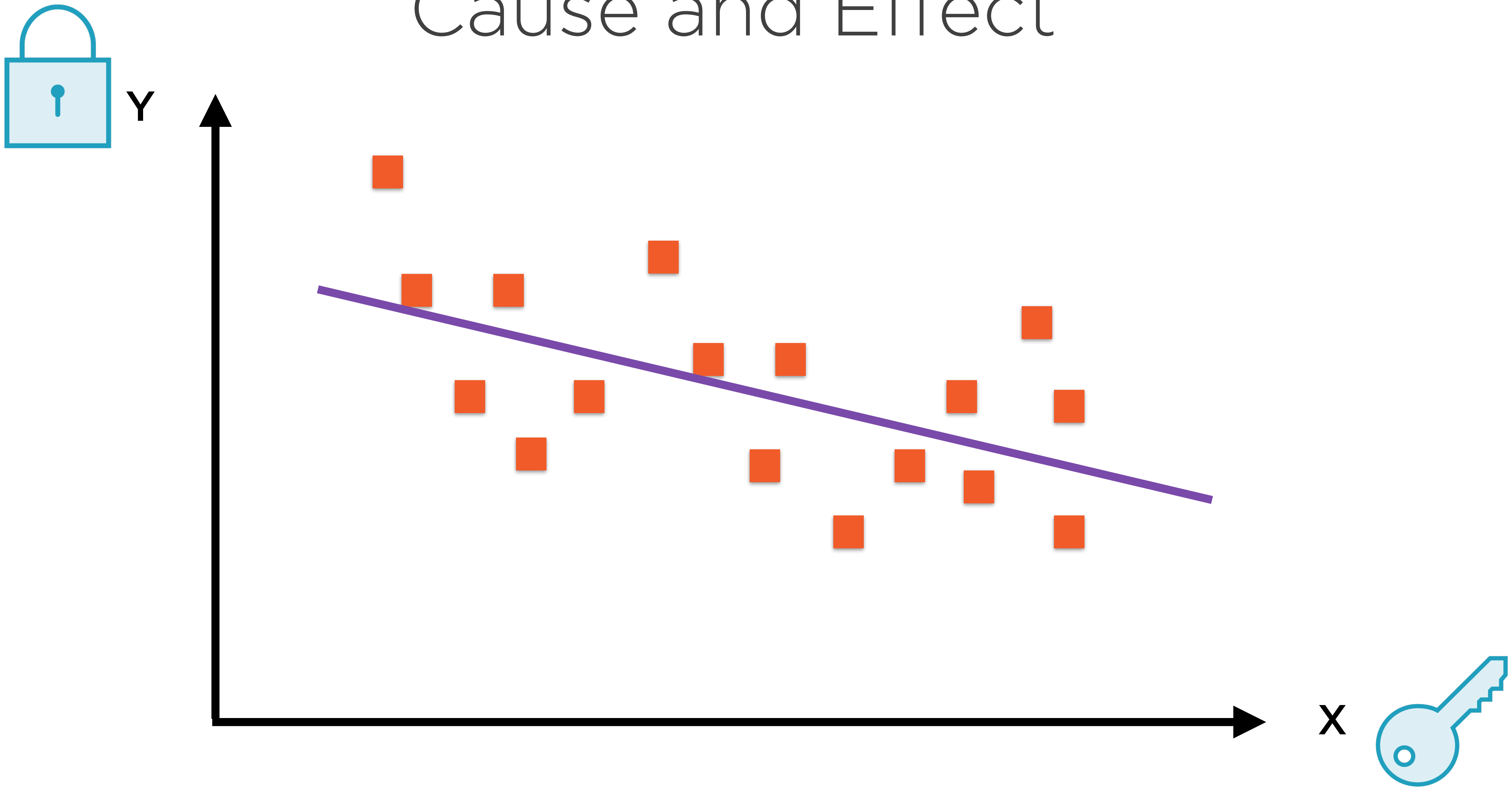
Cause

Explanatory variable

Effect

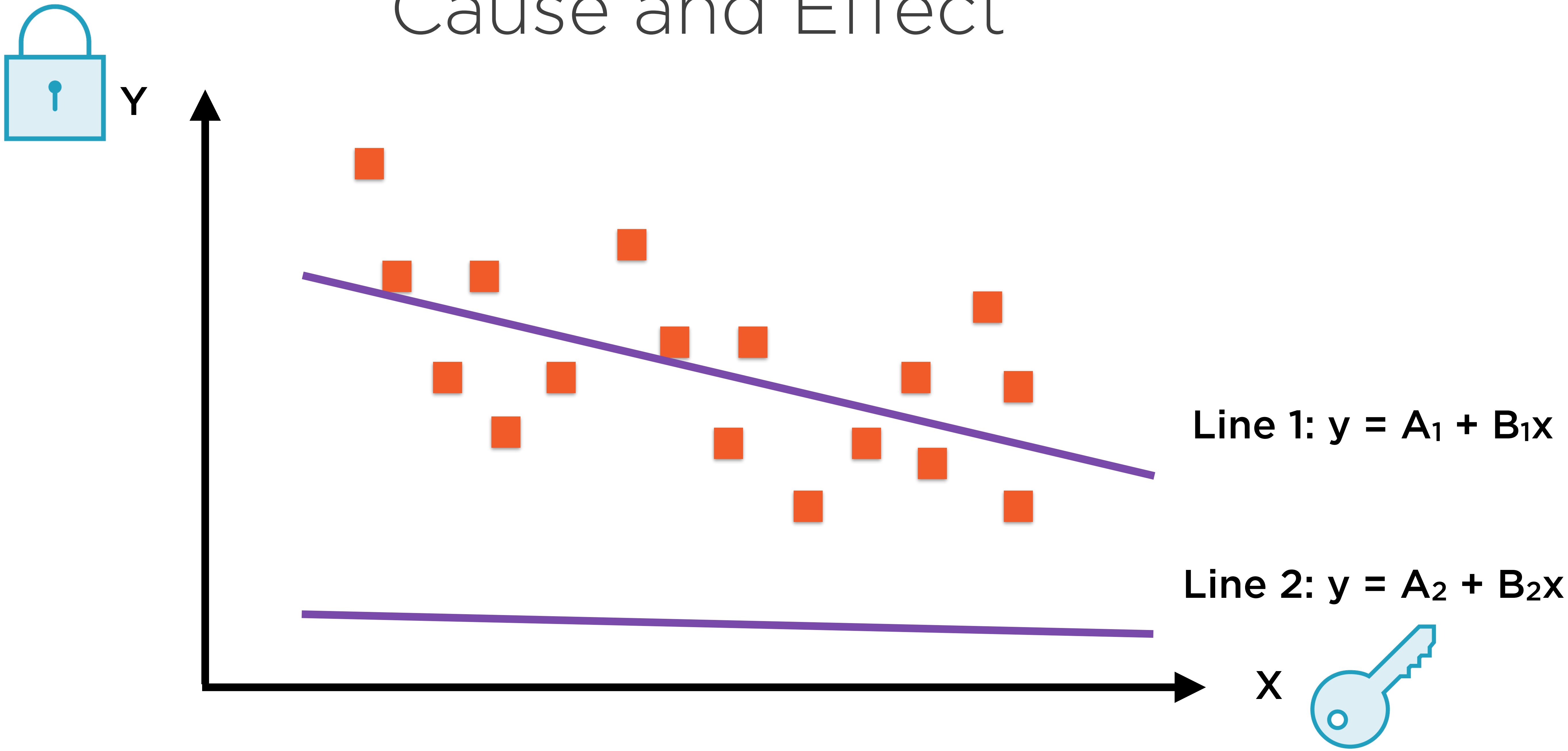
Dependent variable

Cause and Effect



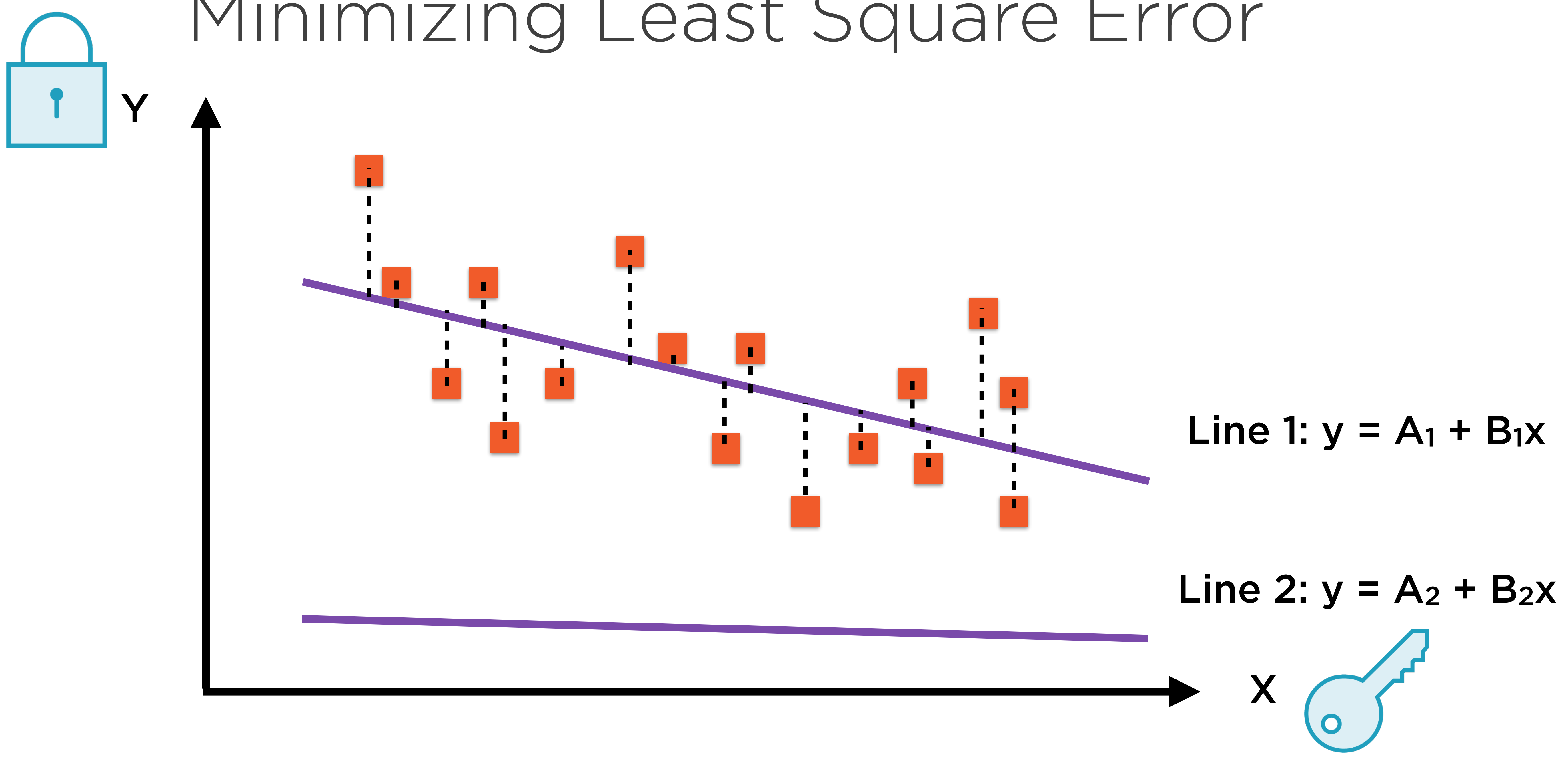
Linear Regression involves finding the “best fit” line

Cause and Effect



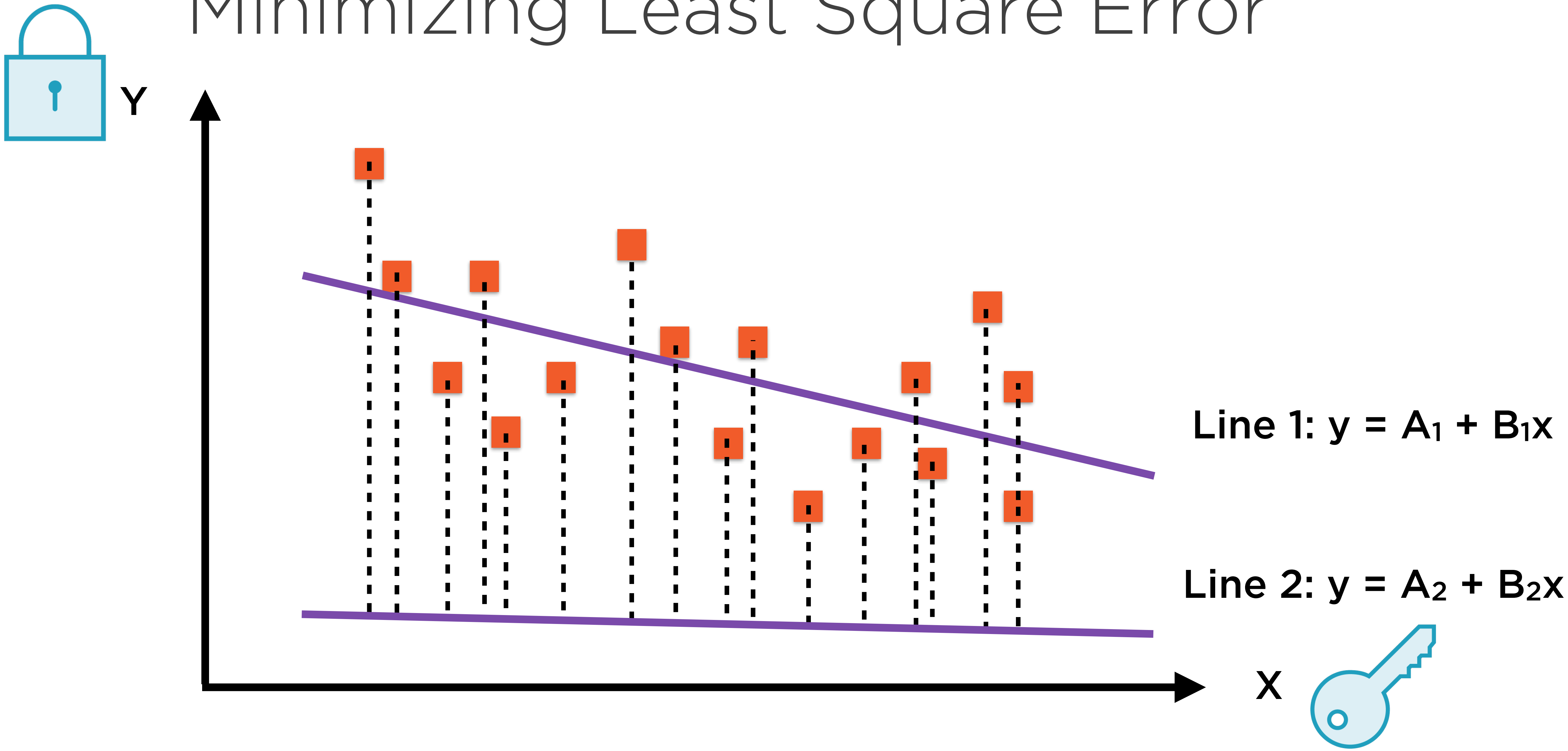
Let’s compare two lines, Line 1 and Line 2

Minimizing Least Square Error



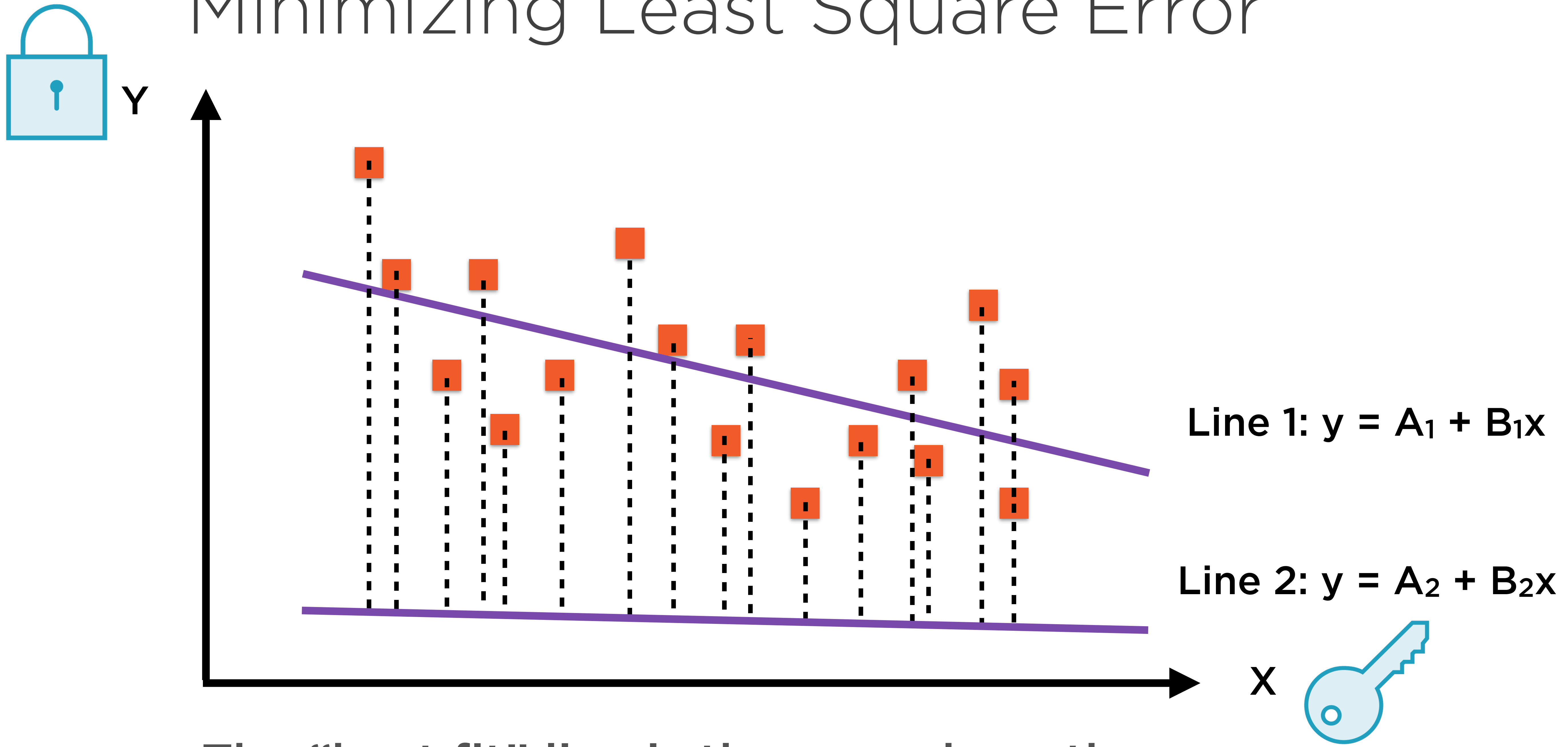
Drop vertical lines from each point to the lines A and B

Minimizing Least Square Error



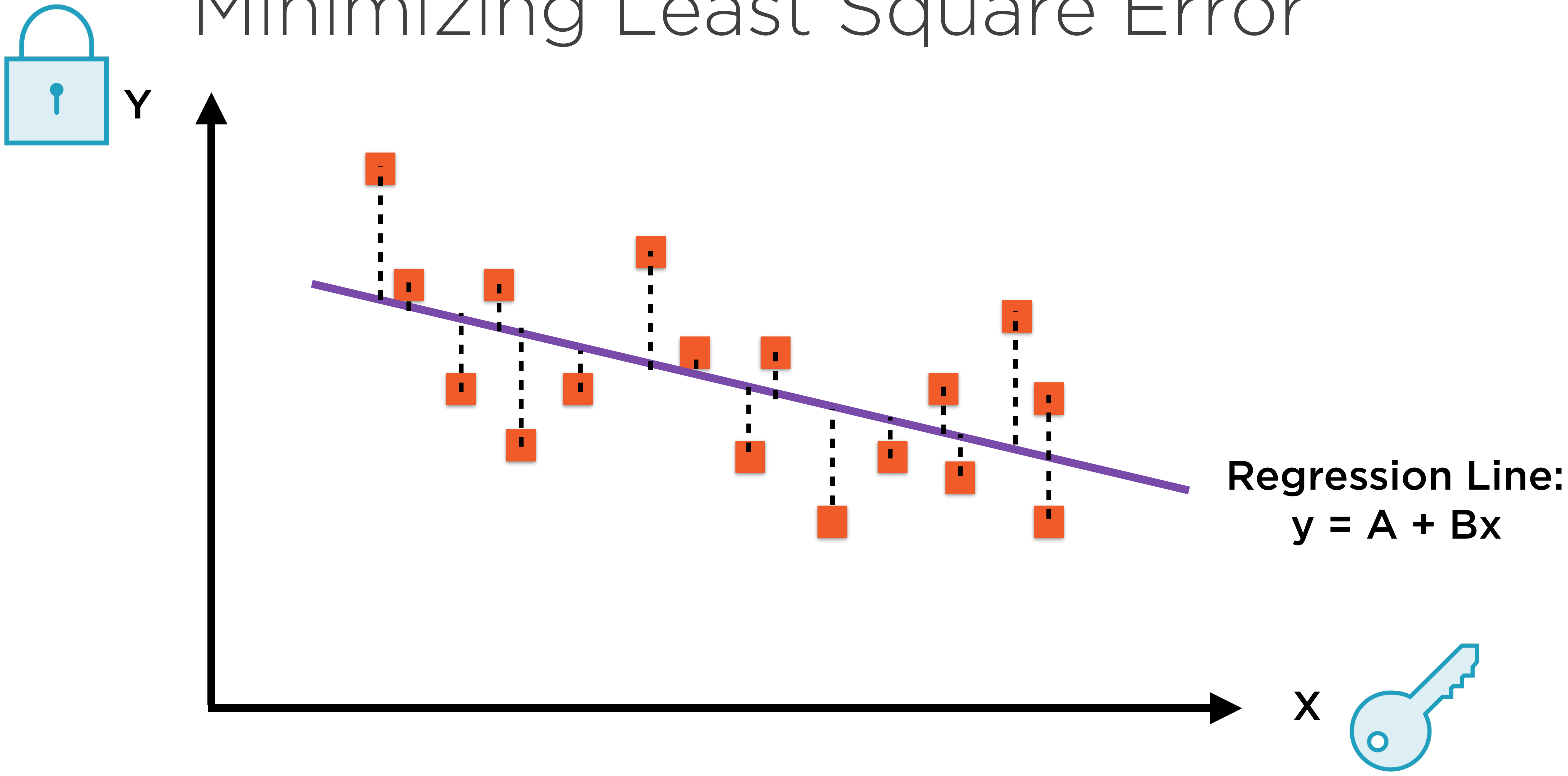
Drop vertical lines from each point to the lines A and B

Minimizing Least Square Error



The “best fit” line is the one where the sum of the squares of the lengths of these dotted lines is minimum

Minimizing Least Square Error



The “best fit” line is called the regression line

R-square

How well does the line represent the data?

How much of the variance in the data is captured by the line?

R-square

$$R^2 = \frac{\text{Explained variance}}{\text{Total variance}}$$

R-square

A higher R-square value indicates that a lot of the underlying variance is captured

Better-fit line

Two-way ANOVA

Two-way ANOVA

Examines the influence of two different independent variables on one continuous dependent variable

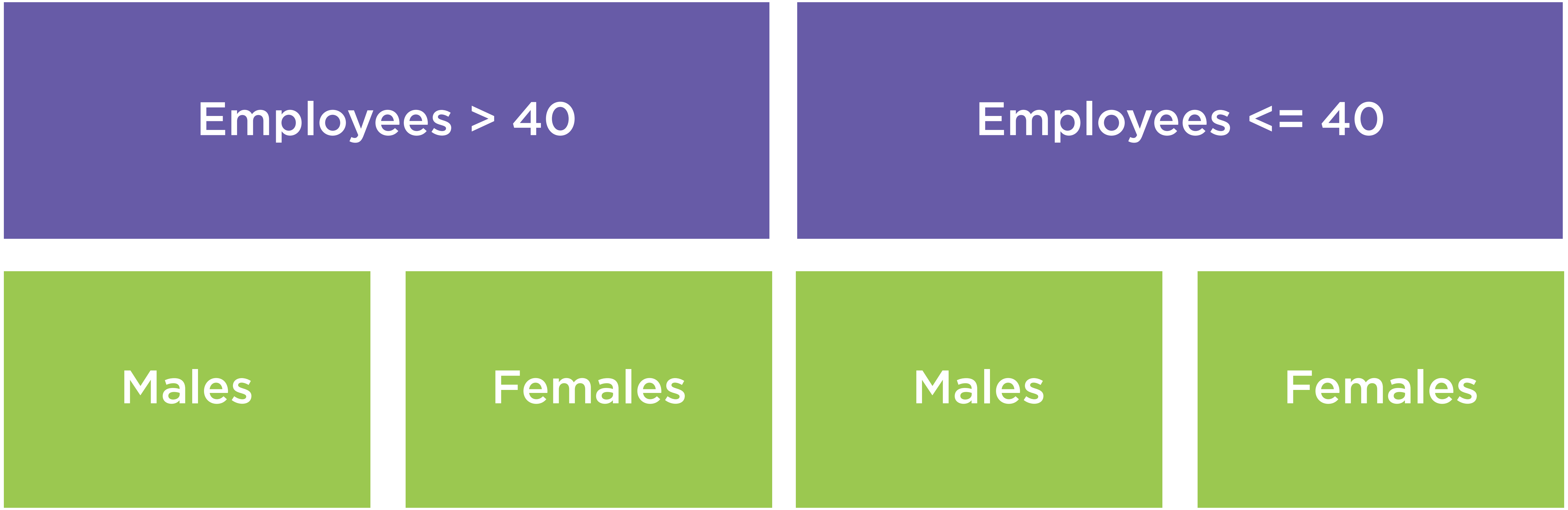
Two-way ANOVA

Examines the influence of two different independent variables on one continuous dependent variable

Two-way ANOVA

Employees > 40	Employees <= 40
Males	Females

Two-way ANOVA



Two-way ANOVA Hypotheses

Null Hypothesis (H_{01})

Null Hypothesis (H_{02})

Null Hypothesis (H_{03})

H_{01}: All groups have equal levels of stress

H_{02}: All ages have equal levels of stress

H_{03}: There is no interaction between age and gender

F-statistic

Calculate an F-statistic and get the p-value for each hypothesis

Accept or reject each hypothesis

Demo

Perform OLS regression

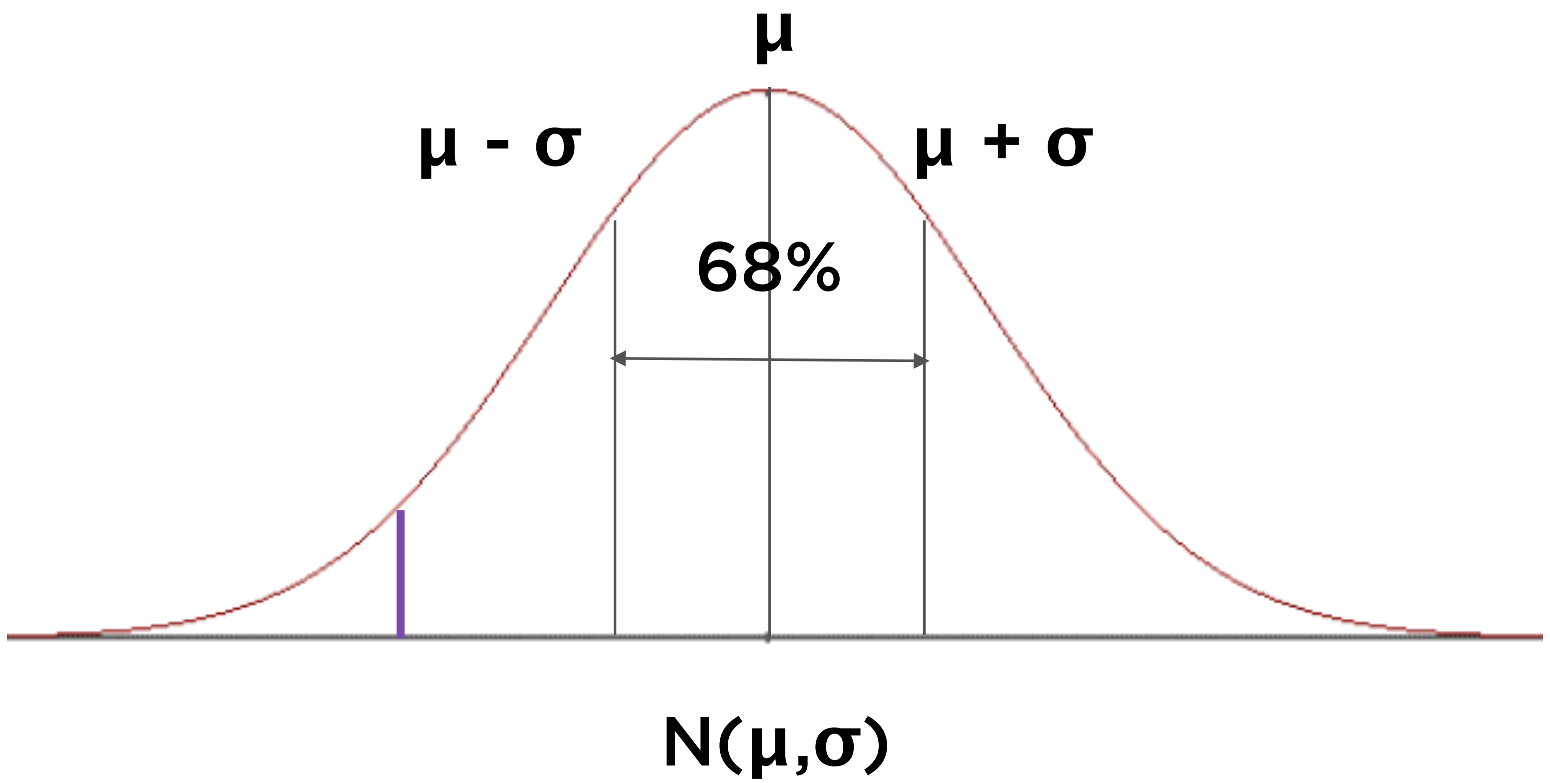
Test significance of regression results using one-way and two-way ANOVA

Skewness and Kurtosis

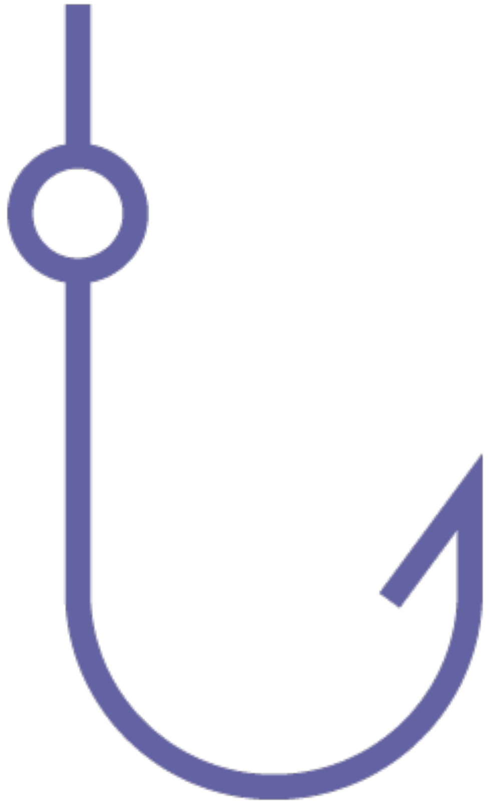
Skewness

A measure of asymmetry around the mean

Gaussian Distribution



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

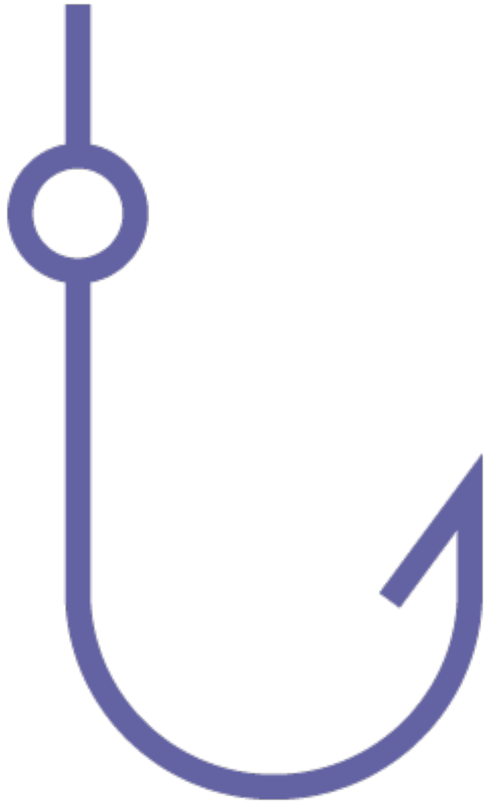


Skewness

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

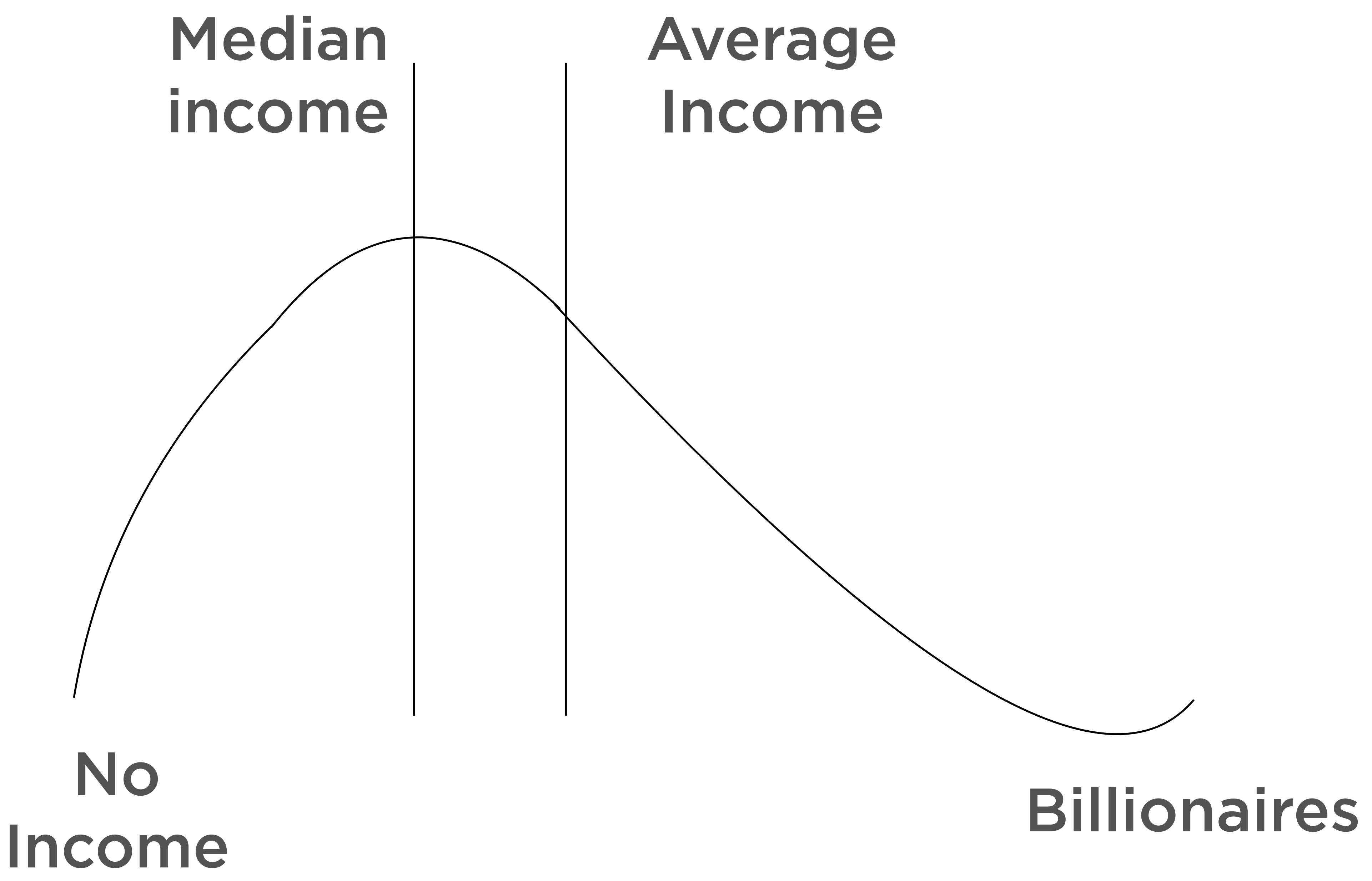
Billionaires: positive skew

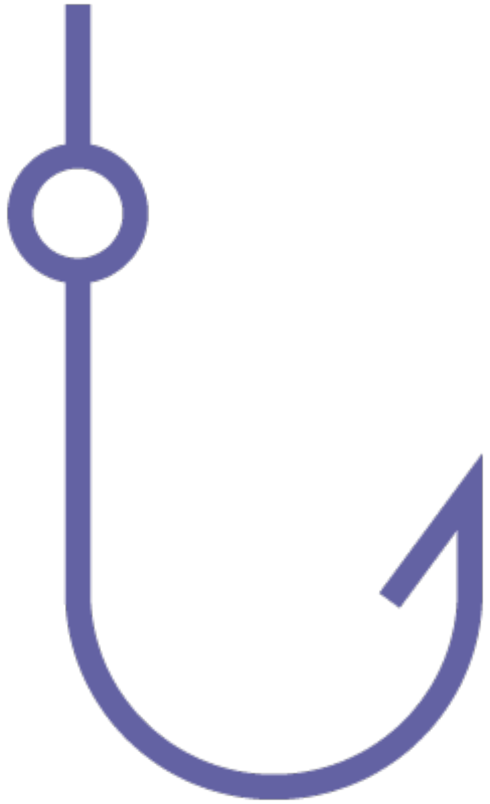
Outliers greater than mean more likely than outliers less than mean

Right-skewed distribution

Often seen when lower bound but no upper bound

Positive Skewness





Negative Skewness

Consider losses from storms

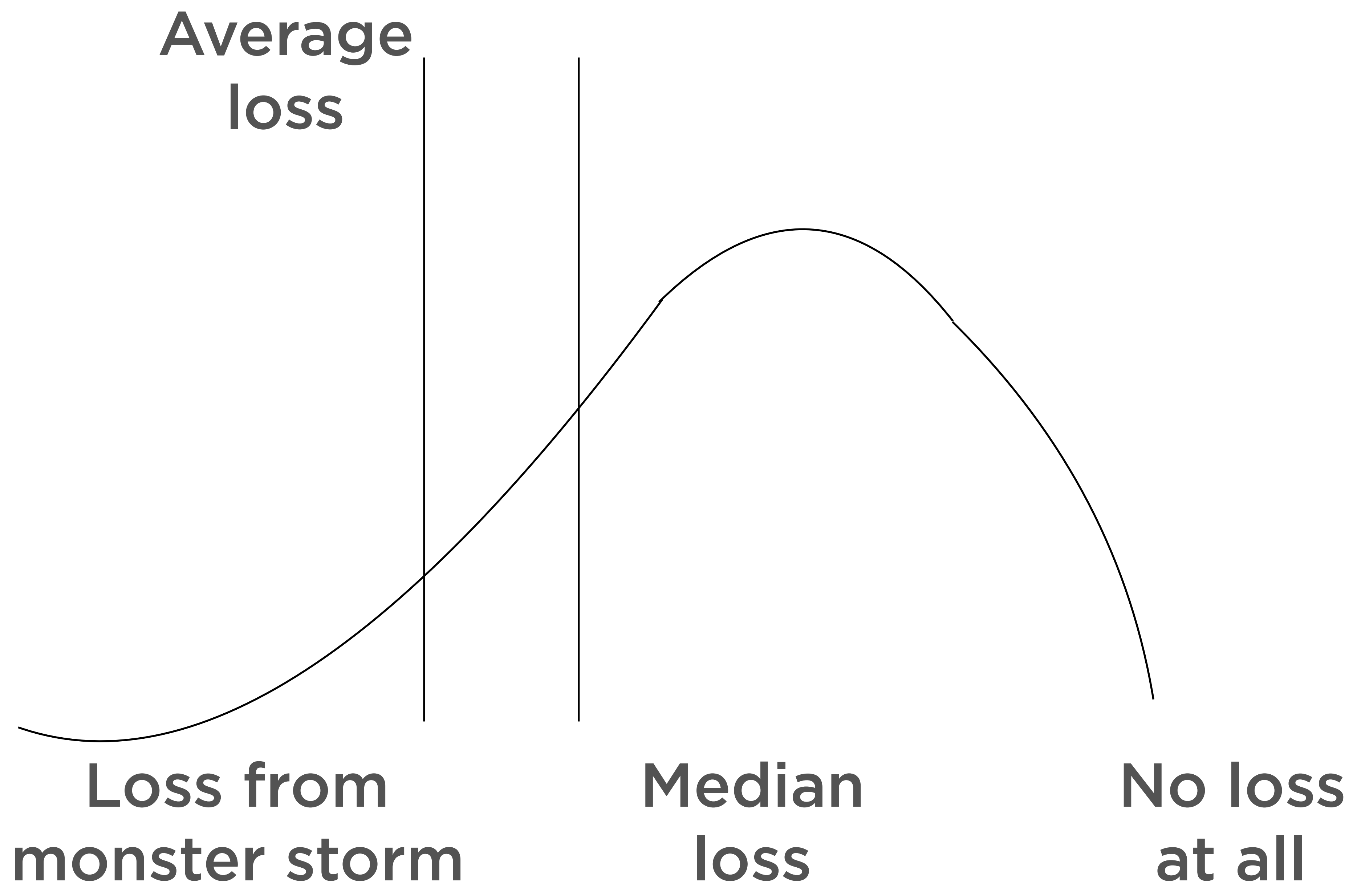
Usually minor, then a monster storm hits

Outliers worse than mean more likely than outliers greater than mean

Left-skewed distribution

Often seen when upper bound but no lower bound

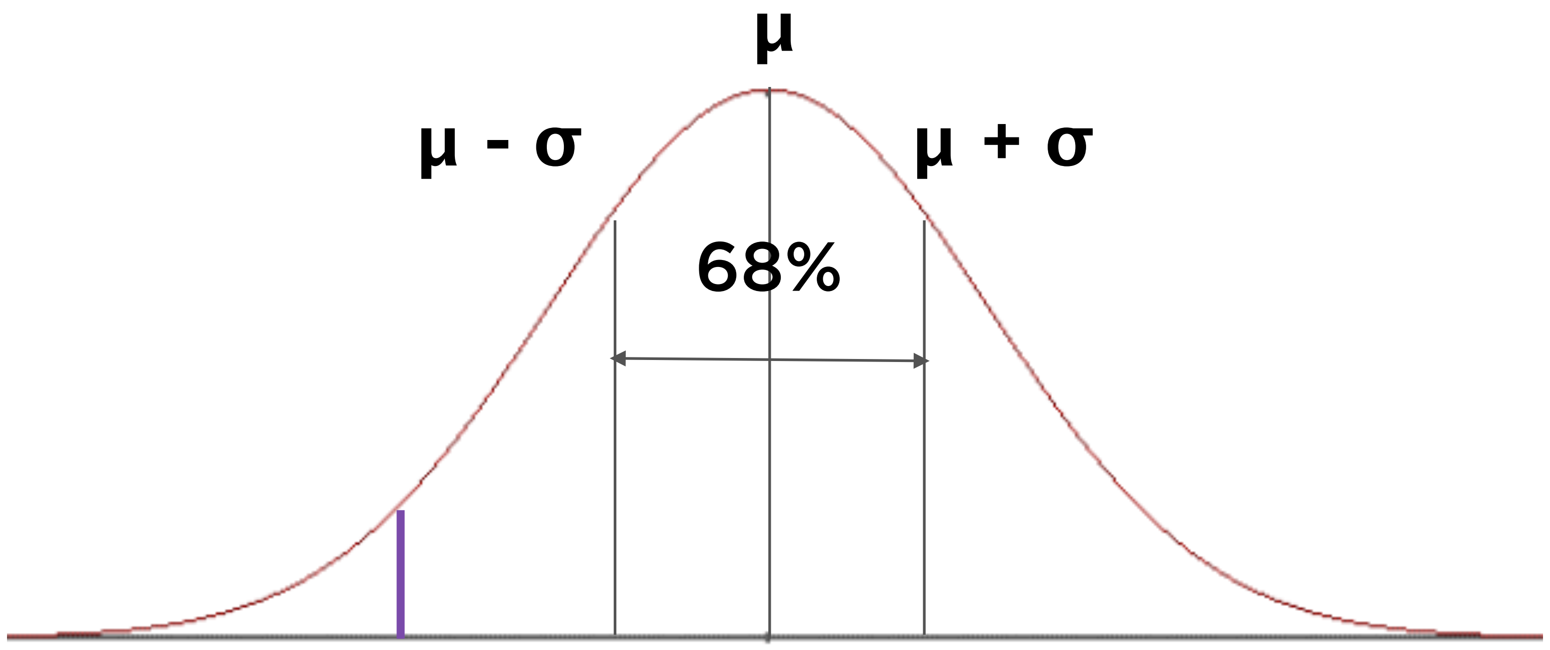
Negative Skewness



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

Demo

Analyzing skewness and kurtosis

Summary

Python package with implementations of statistical models and tests

T-tests to compare population means

One-way ANOVA for multiple categories

Two-way ANOVA for multiple categorical independent variables

Using ANOVA to analyze regression models

Skewness and kurtosis in data