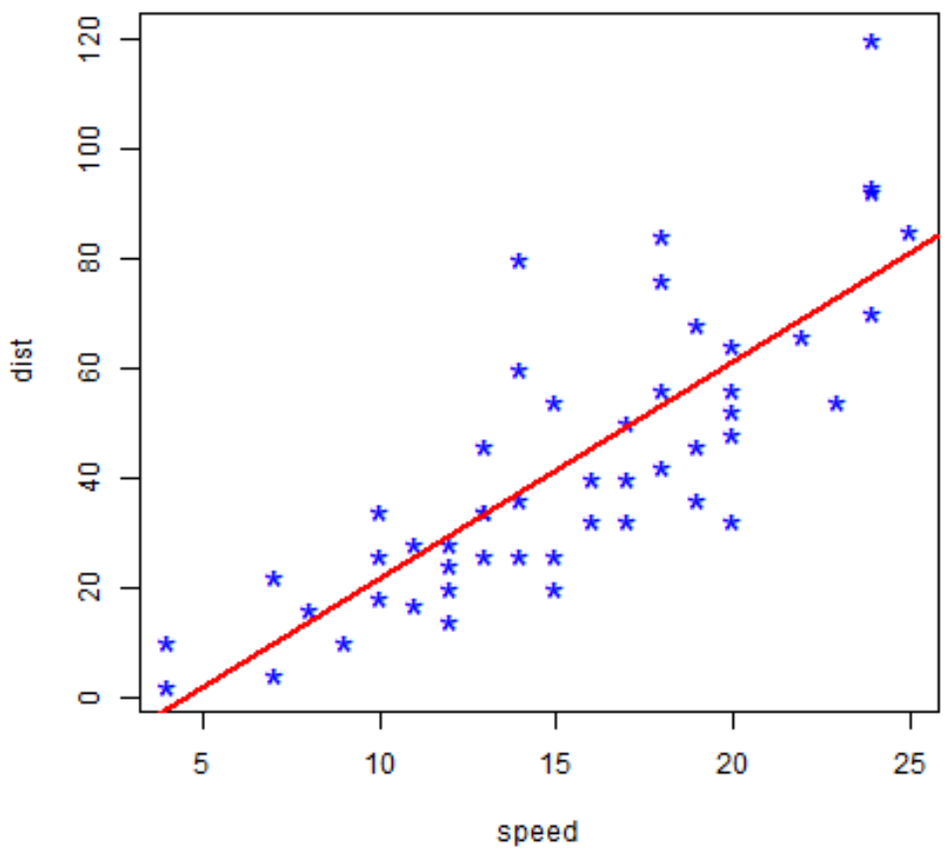
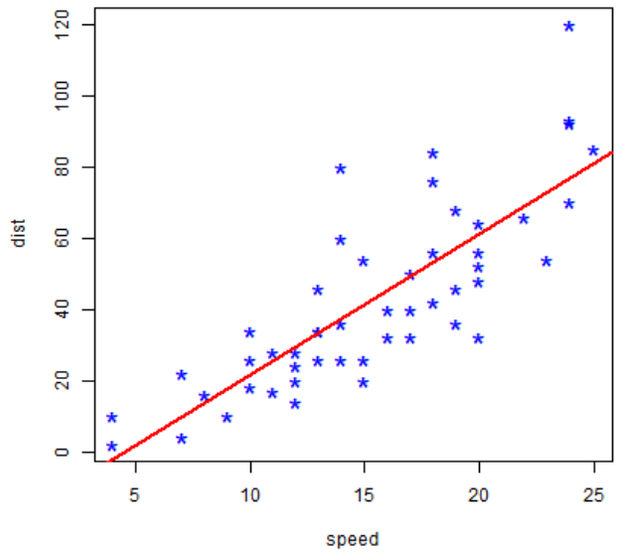
An Introduction to Statistical Modelling

Mustafa Mahfuz

What is a Model?

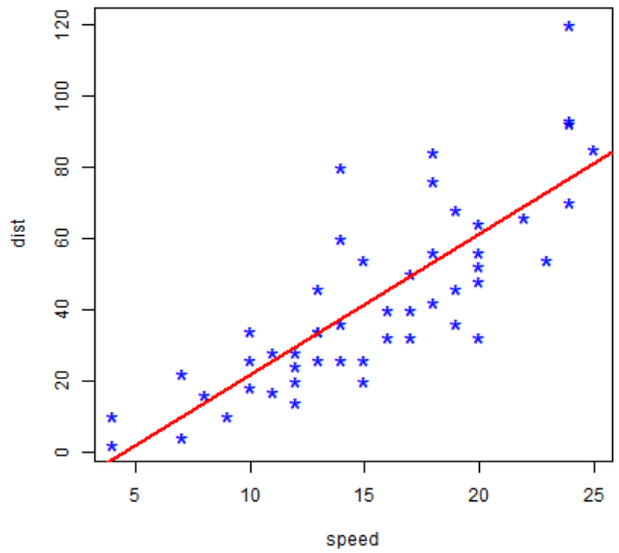


What is a Model?



Approximating reality with a purpose

What is a Model?



Approximating reality with a purpose

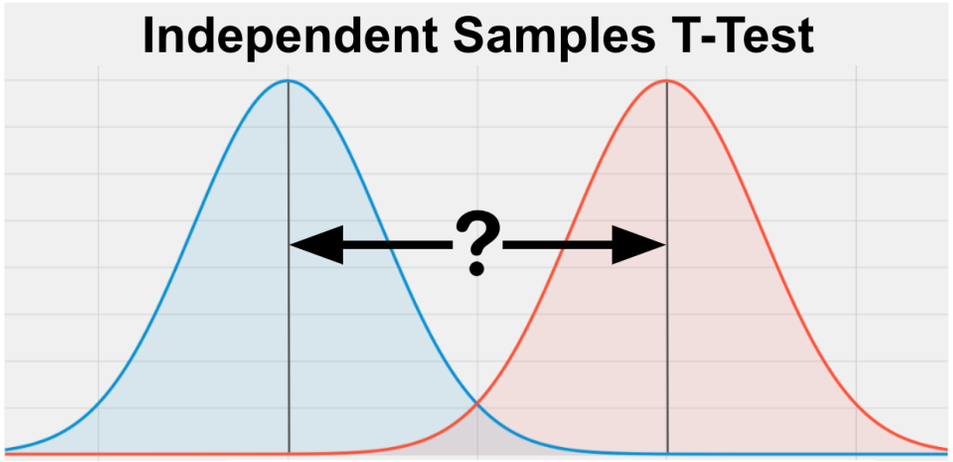
"Modeling is an art, as well as a science and, is directed toward finding a good approximating model … as the basis for statistical inference" – Burnham & Anderson

Statistical model vs. Plain statistics

Statistical modelling

Concepts and techniques with more power to analyze complex system

Traditional methods

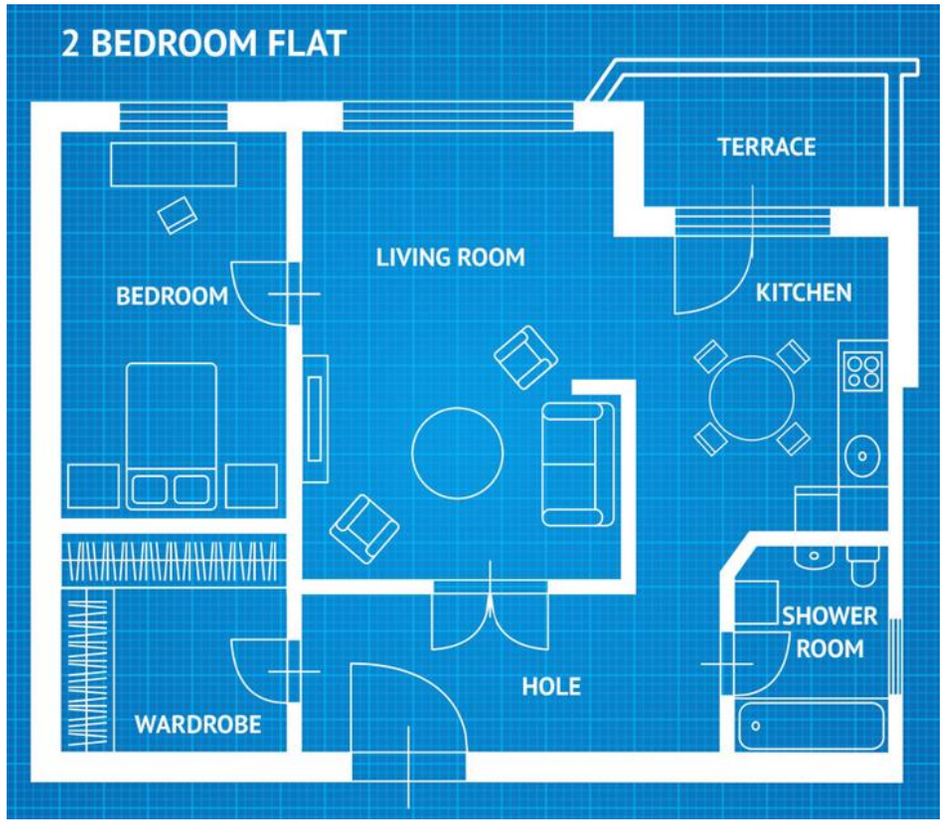


More definition of Model

A model is a representation for a purpose

- **Representation**: it stands for something in the real world
- **Purpose**: Your particular use for the model

Some everyday models



Purpose of a Statistical Model

- Separate **'signal'** from '**noise'**
- Understand **trends** and **patterns**
- Identify which variables are **related** to a response variable
- **Quantify** this relationship
- Make **predictions** about future observations
- Understand underlying, **unexplained variability** in the patterns
- Compare results against other statistical models

Mathematical model

Constructed out of mathematical entities

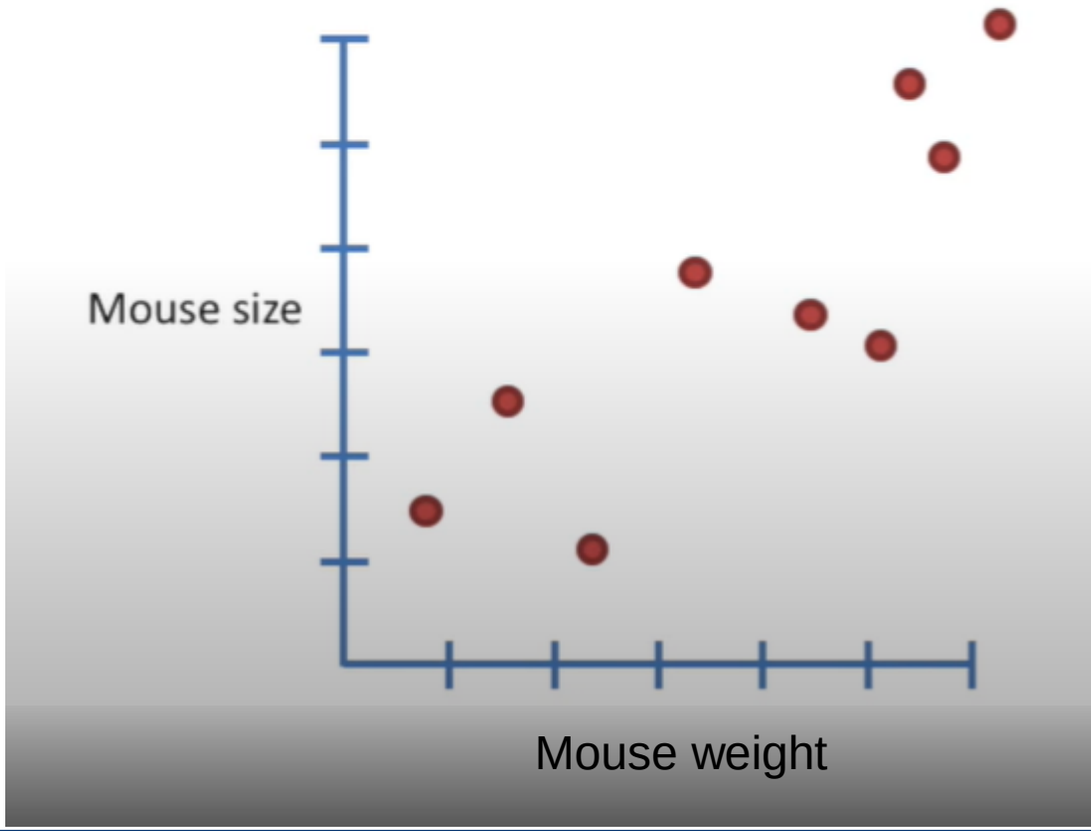
- Numbers
- Model formulas
- Equations
- **Deterministic**

Statistical model

A special type of mathematical model

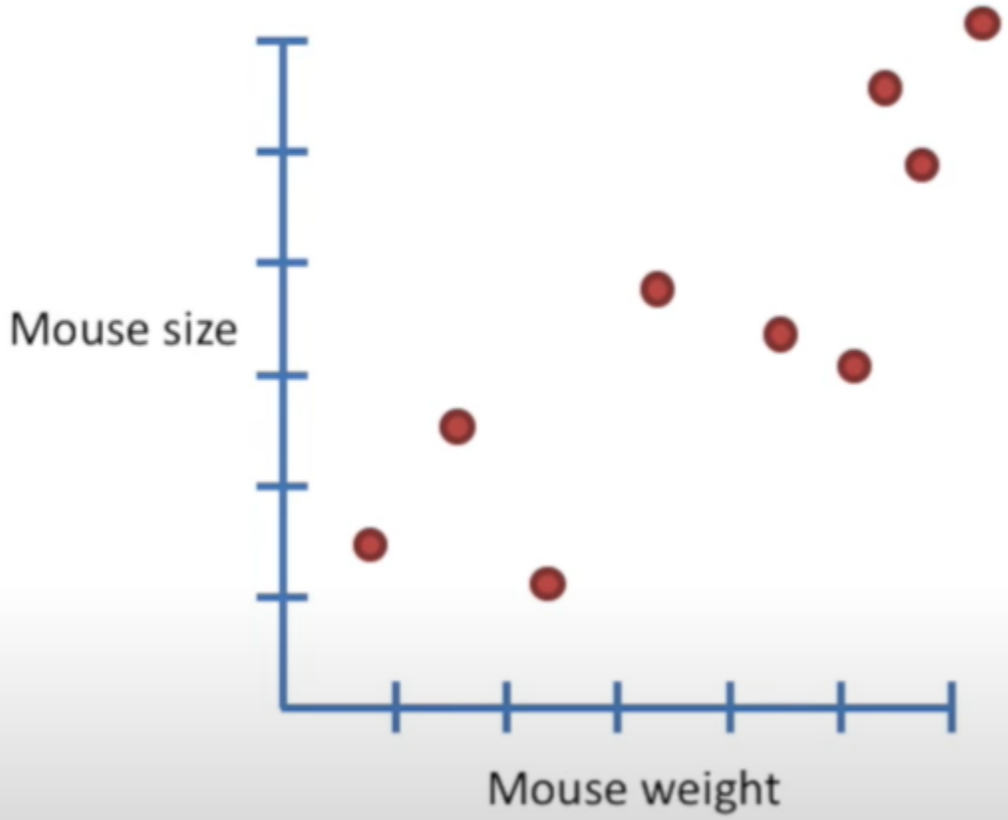
- Informed by data
- Incorporates uncertainties and randomness
- **Stochastic**

“Model” is used in a lot of contexts.

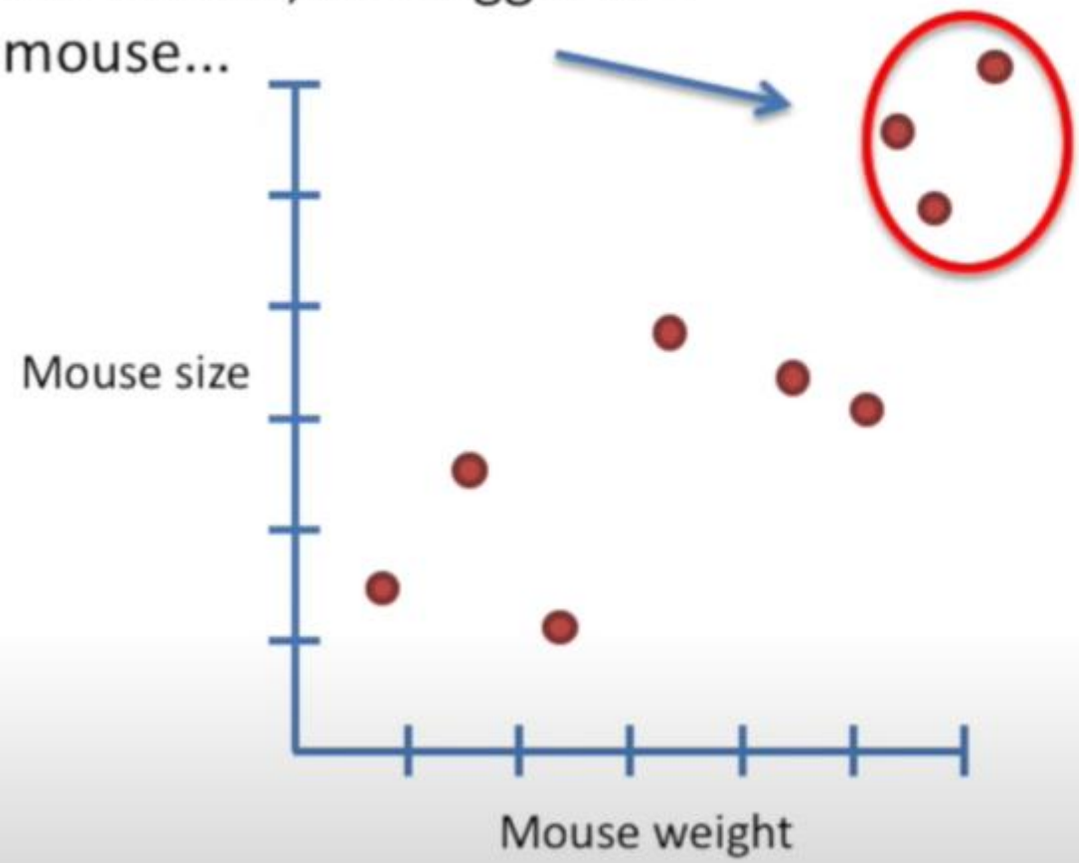


For example, I might “model mouse size with mouse weight”.

In this context, “model” refers to a relationship.



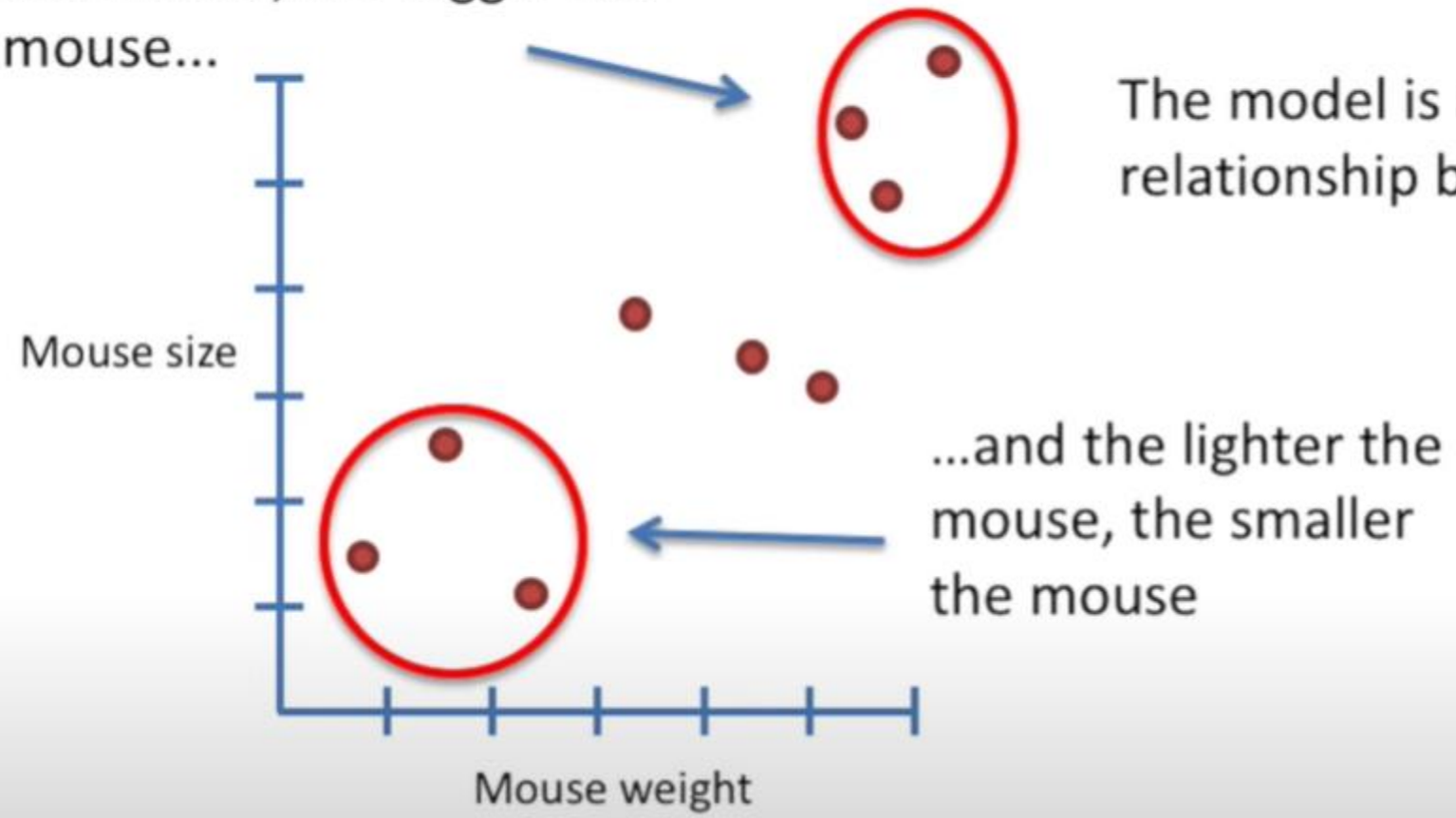
In this case, the relationship is pretty obvious. The heavier the mouse, the bigger the mouse...



In this context, “model” refers to a relationship.

The model is a way to explore the relationship between weight and size.

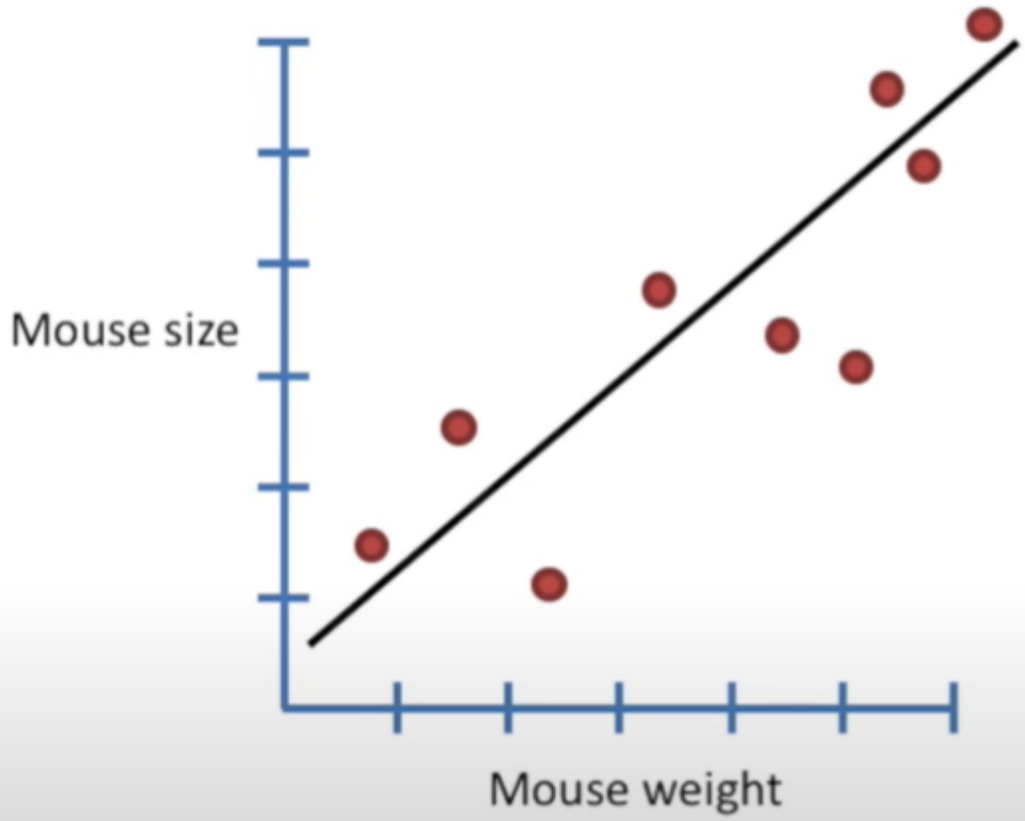
In this case, the relationship is pretty obvious. The heavier the mouse, the bigger the mouse...



...and the lighter the mouse, the smaller the mouse

In this context, “model” refers to a relationship.

The model is a way to explore the relationship between weight and size.

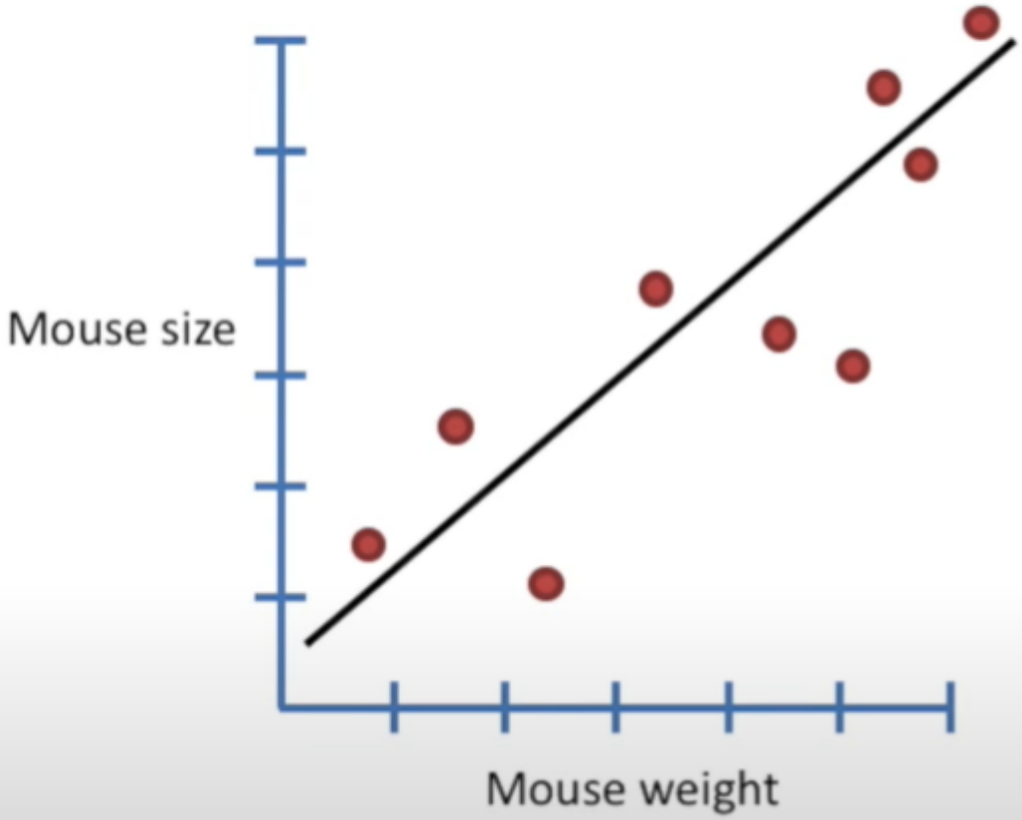


A “model” can also be an equation.

mouse size = 0.1 + 0.8 × mouse weight

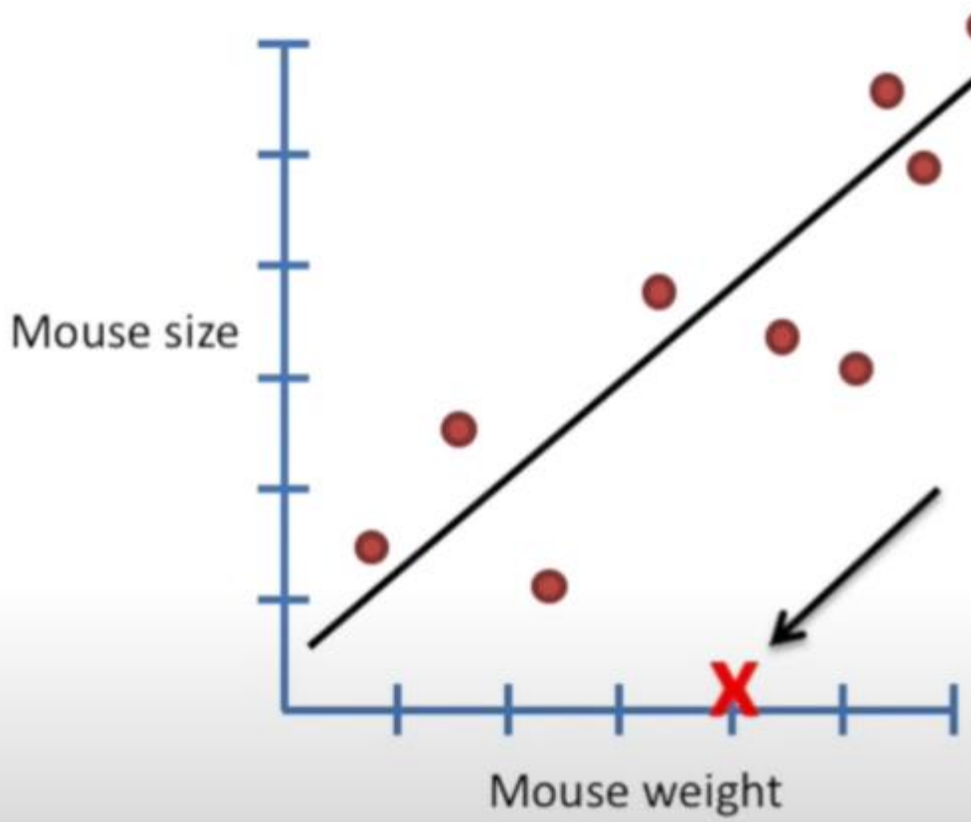
A mathematical model

The model (or equation) can tell us about mice we haven't measured yet.



mouse size = 0.1 + 0.8 × mouse weight

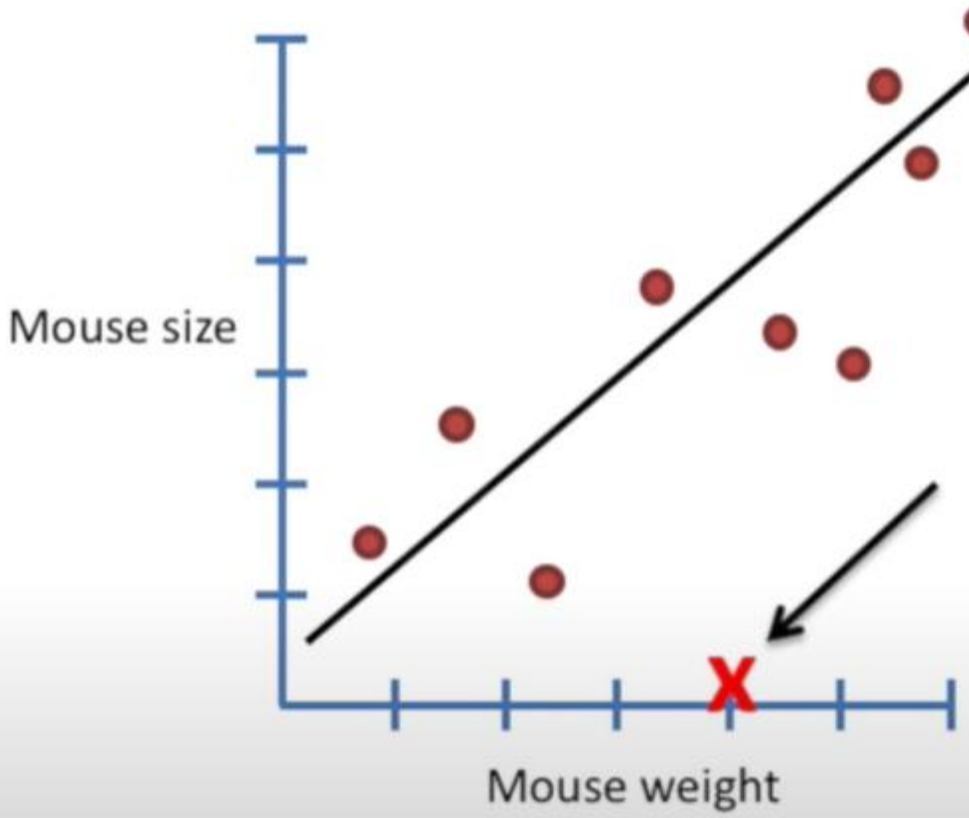
The model (or equation) can tell us about mice we haven't measured yet.



mouse size = 0.1 + 0.8 x mouse weight

Someone might want to know how large a mouse might be if it weighs 4 units.

The model (or equation) can tell us about mice we haven't measured yet.

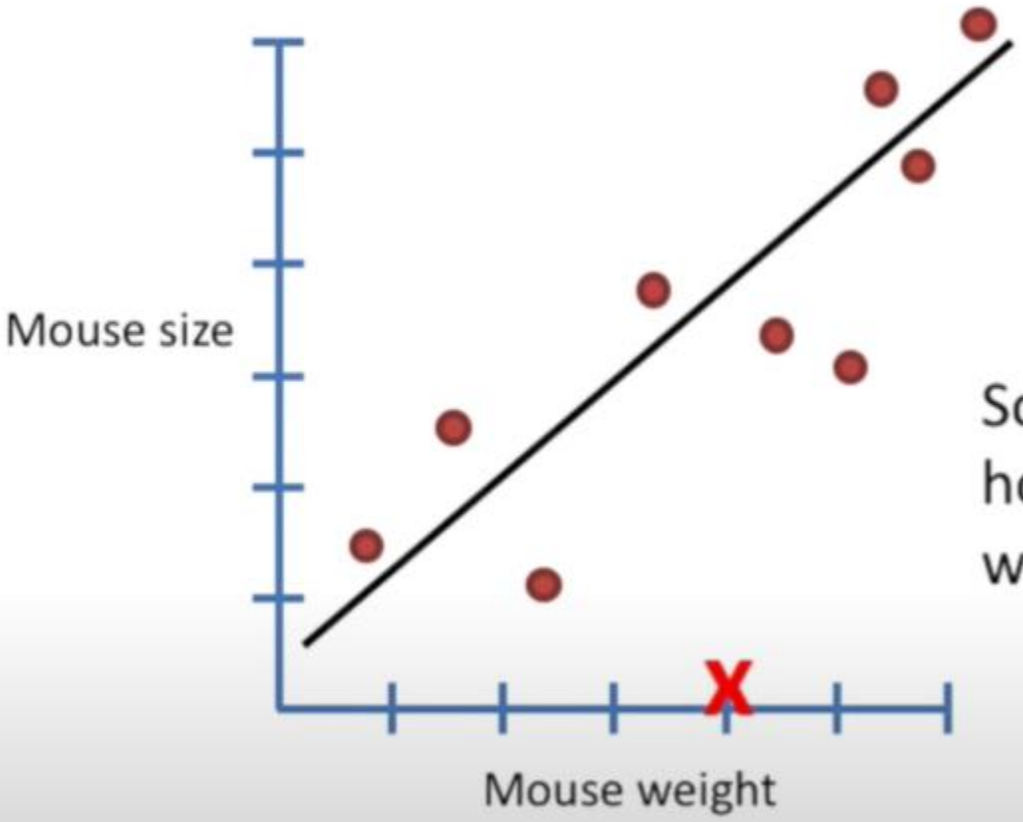


mouse size = 0.1 + 0.8 x mouse weight

Someone might want to know how large a mouse might be if it weighs 4 units.

mouse size = 0.1 + 0.8 x 4

The model (or equation) can tell us about mice we haven't measured yet.



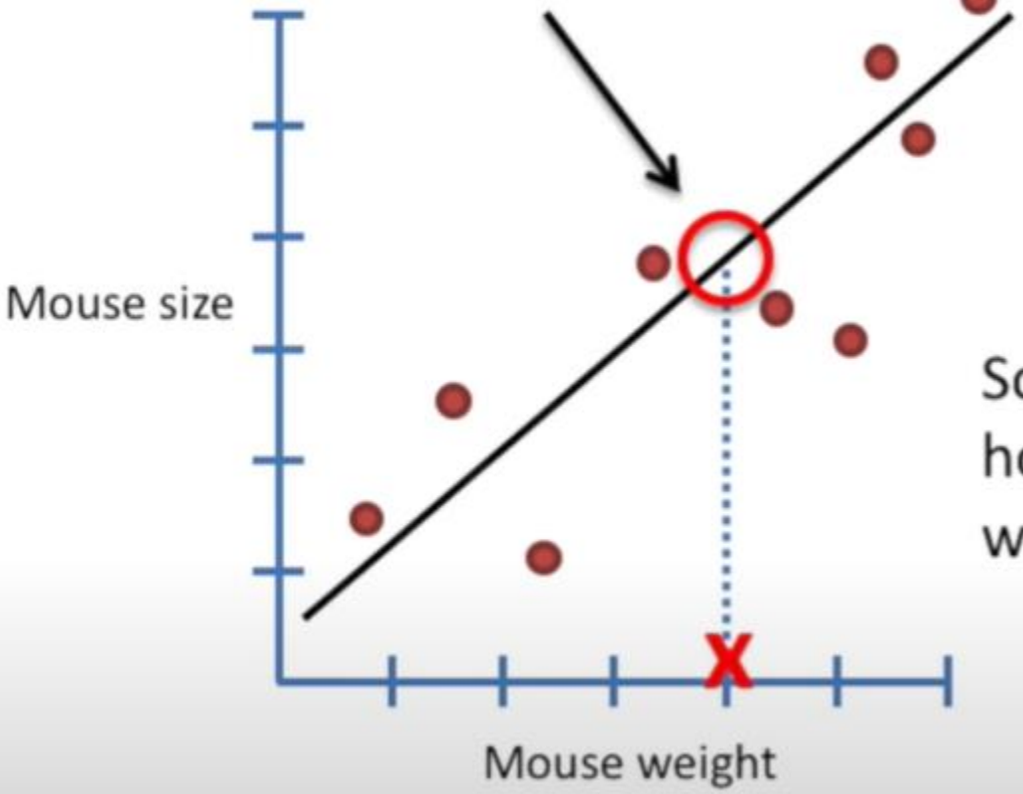
mouse size = 0.1 + 0.8 x mouse weight

Someone might want to know how large a mouse might be if it weighs 4 units.

mouse size = 0.1 + 0.8 x 4

mouse size = 3.3

The model predicts that a mouse that weighs 4 units will be 3.3 units big.



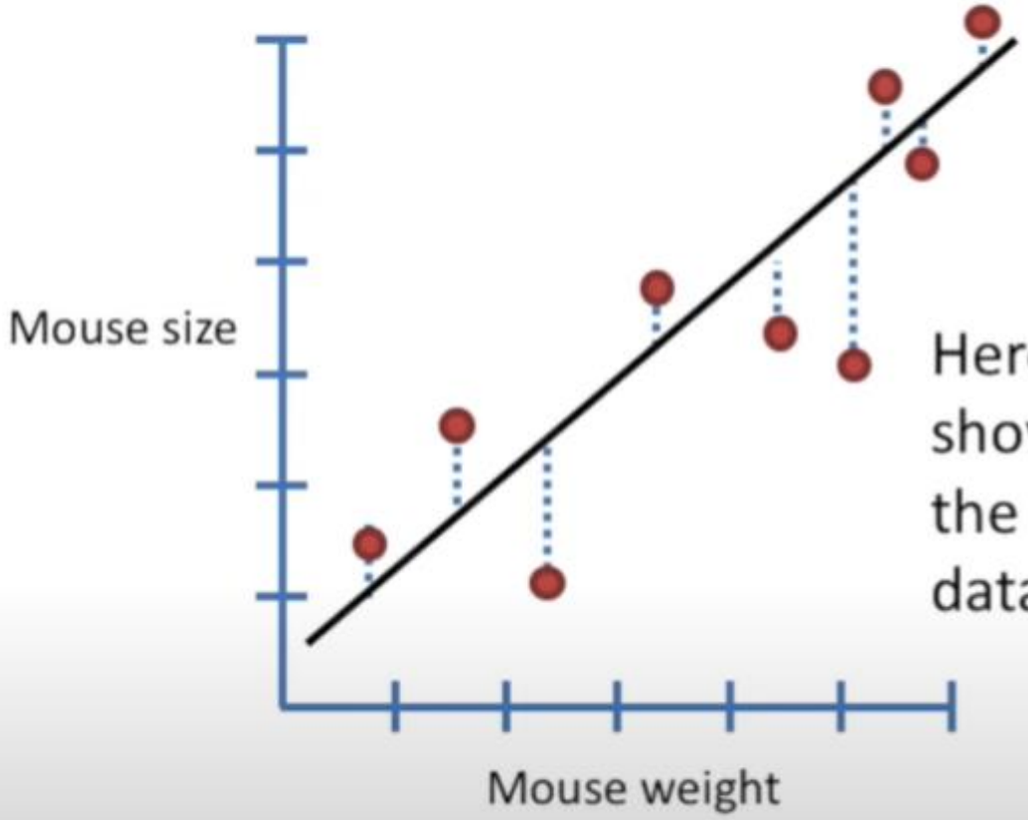
mouse size = 0.1 + 0.8 × mouse weight

Someone might want to know how large a mouse might be if it weighs 4 units.

mouse size = 0.1 + 0.8 × 4

mouse size = 3.3

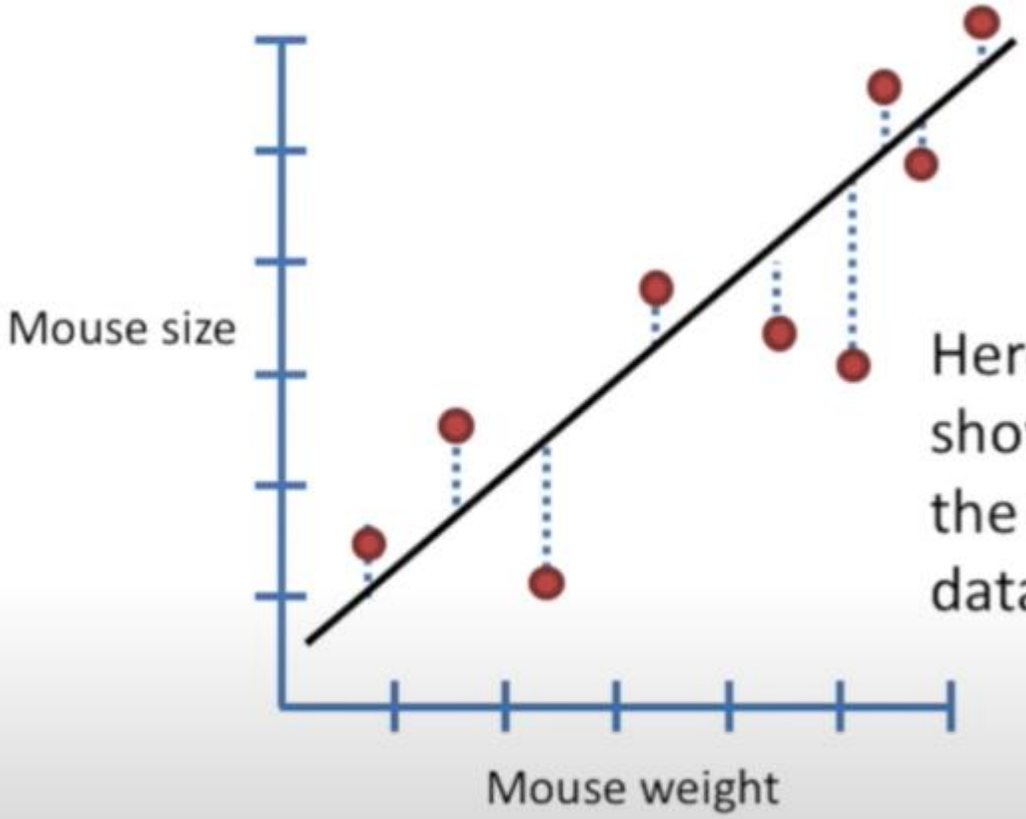
The model (equation) is an approximation of the real data.



mouse size = 0.1 + 0.8 × mouse weight

Here, the dotted lines show the distance from the model to the actual data points.

The model (equation) is an approximation of the real data.

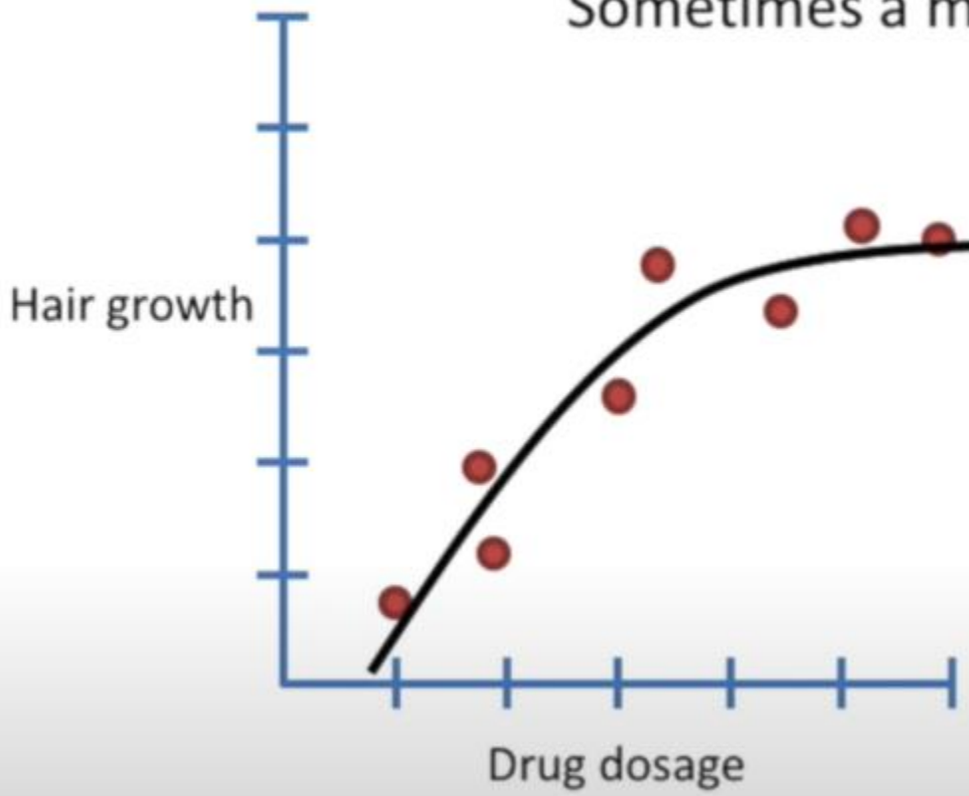


mouse size = 0.1 + 0.8 x mouse weight

Here, the dotted lines show the distance from the model to the actual data points.

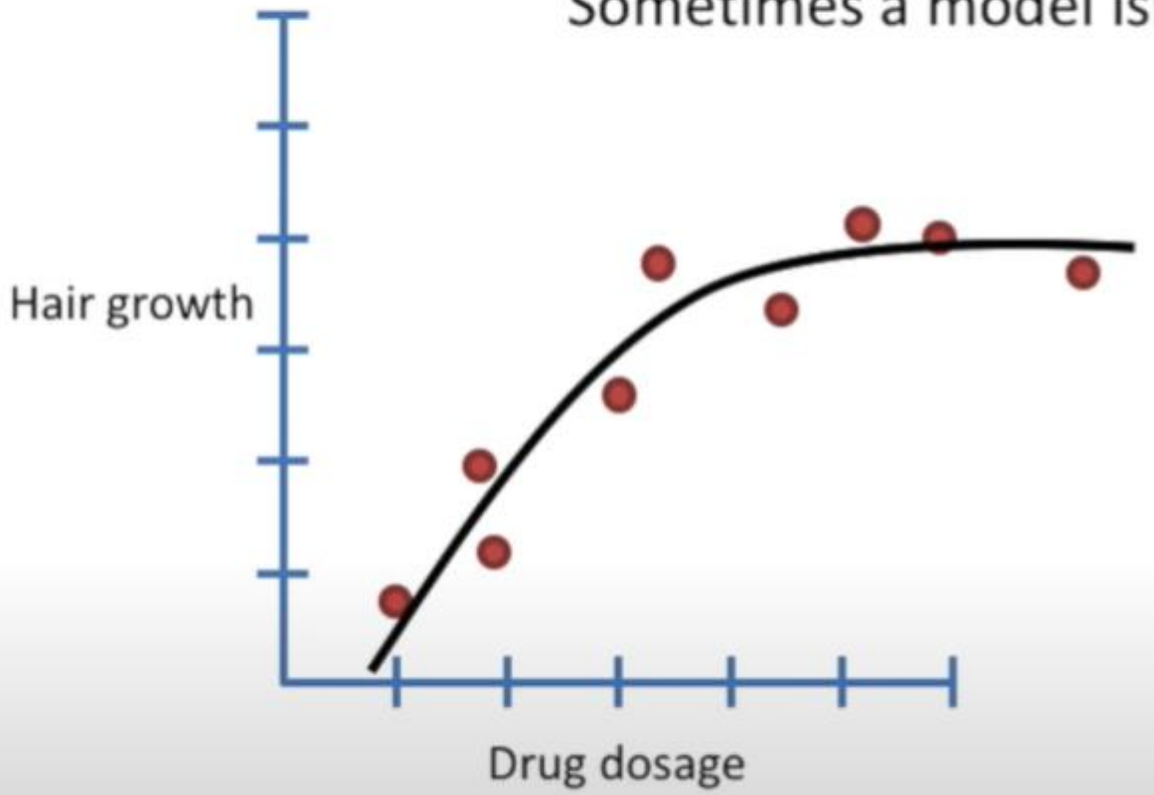
A lot of statistics is dedicated to determining if a model makes a good or bad approximation of the data.

Sometimes a model isn't a straight line.



In this case, the model helps us understand the relationship between a drug and hair growth for aging men.

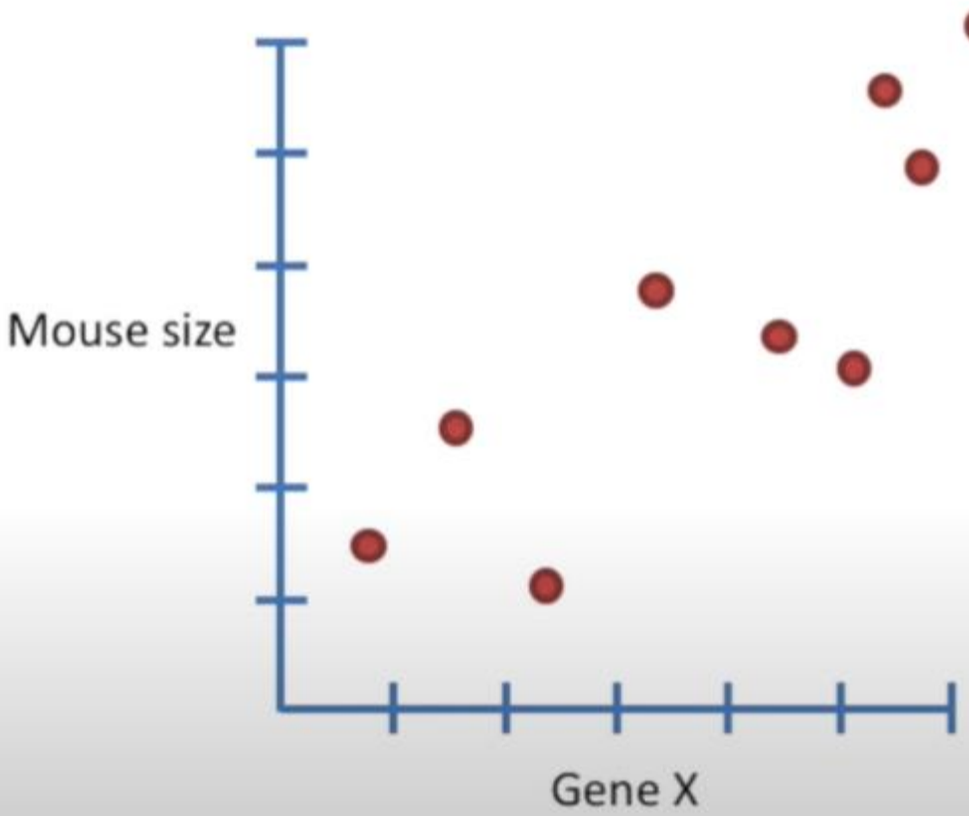
Sometimes a model isn't a straight line.



In this case, the model helps us understand the relationship between a drug and hair growth for aging men.

We see that after a point, increasing the dosage doesn't help grow any more hair.

In summary…



1) We use models to explore relationships.

For example, I might be interested in the relationship between Gene X and mouse size.

2) We use statistics to determine how useful and how reliable our model is.

Statistics: 4 categories

1

Description

Summarize data using simple statistics or charts (mean, sd, boxplots..

Statistics: 4 categories

1	2
Description	Exploration
Summarize data using simple statistics or charts (mean, sd, boxplots..	Extract information from a large dataset without having a precise question to answer (PCA..)

Statistics: 4 categories

1	2	3
Description	Exploration	Test
Summarize data using simple statistics or charts (mean, sd, boxplots..	Extract information from a large dataset without having a precise question to answer (PCA..)	Accept/reject a very precise hypothesis assuming error risk (t-test, ANOVA, correlation tests..)

Statistics: 4 categories

1	2	3	4
Description	Exploration	Test	Modeling
Summarize data using simple statistics or charts (mean, sd, boxplots..	Extract information from a large dataset without having a precise question to answer (PCA..)	Accept/reject a very precise hypothesis assuming error risk (t-test, ANOVA, correlation tests..)	Understand the way a variable evolves according to a set of other variables (regression, ANOVA, ANCOVA..)

Statistics: know your data

Quantitative data is measured

Ratio

Interval

Qualitative data is categorized

Ordinal

Nominal

Statistics: know your data

Ratio Data: Properties

Examples: Age, Height, Weight

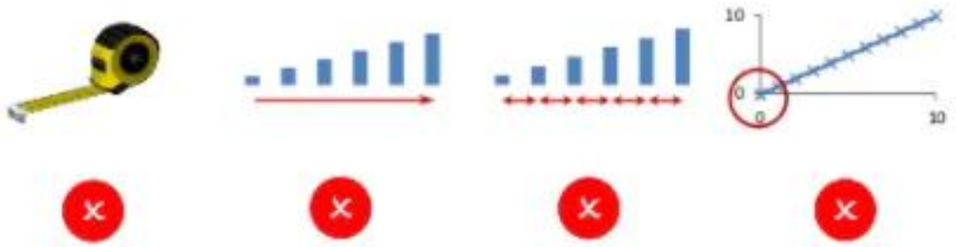
Ordinal Data: Properties

Examples: Health Status, Level of Education, Customer Satisfaction

Interval Data: Properties

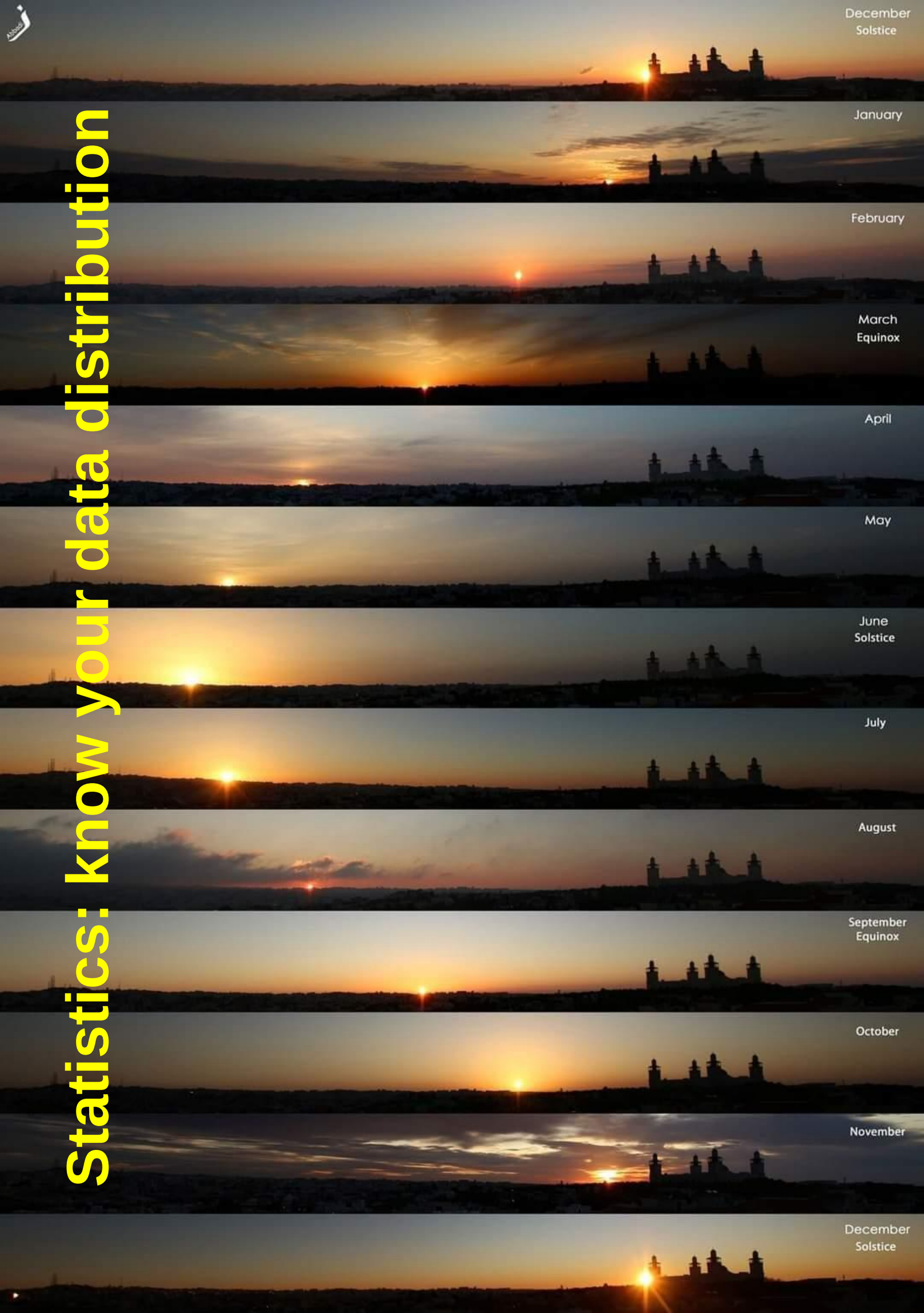
Examples: Temperature, Dates

Nominal Data: Properties



Examples: Gender, Nationality, Religion

Statistics: know your data distribution



Statistics: know your data distribution

Continuous data

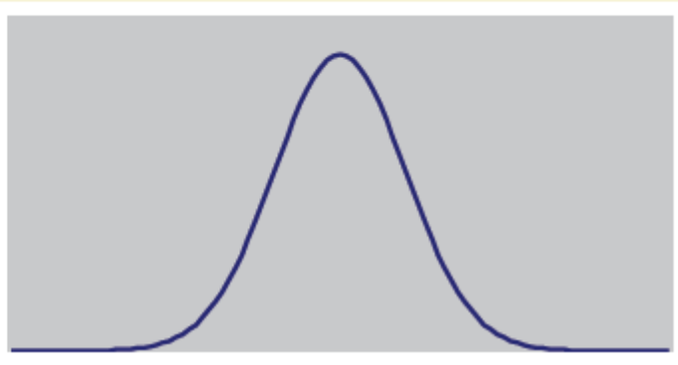
- Normal distribution
- Skewed distribution
- Log normal

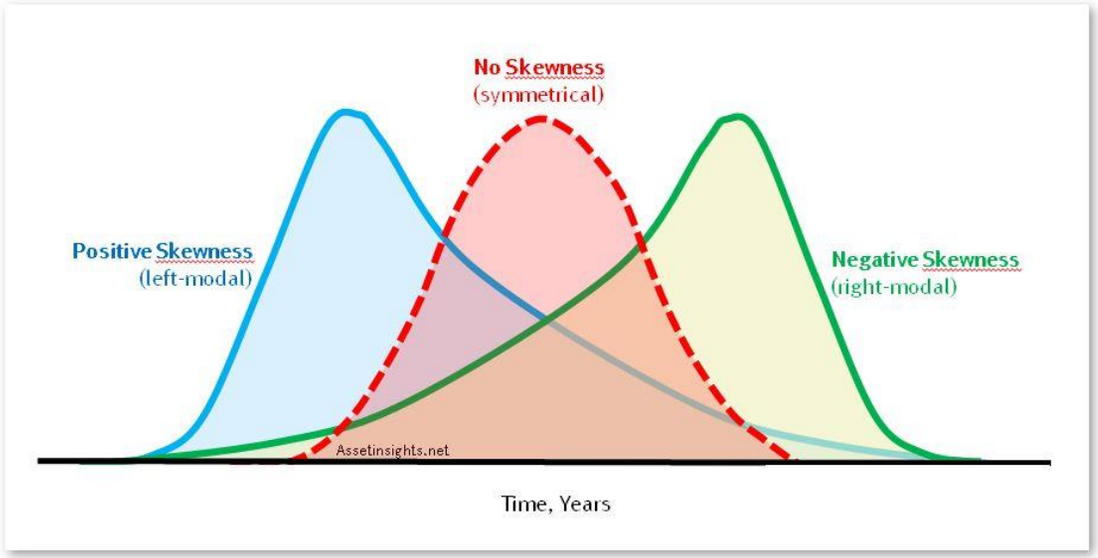
Categorical data

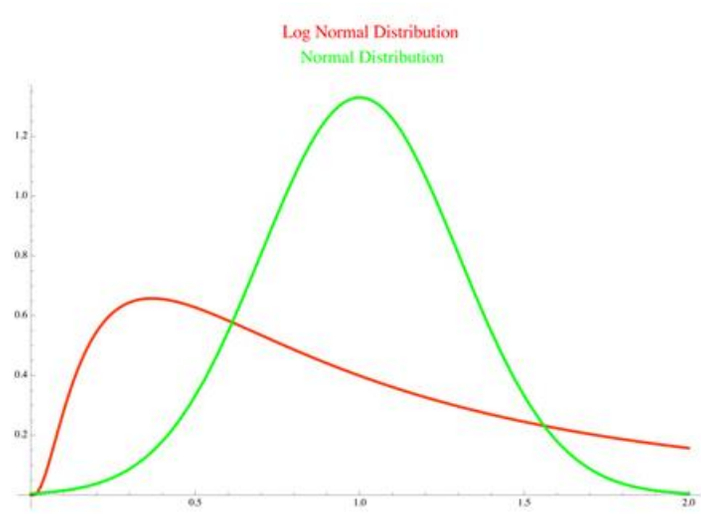
- Binary
- Ordered
- Unordered

Statistics: know your data distribution

- Symmetric: Mean = Median = Mode
- "BELL-SHAPED"
- Standard Deviation Rule:
  - 68% of data within 1 SD of mean
  - 95% of data within 2 SD of mean
  - 99.7% of data within 3 SD of mean







Statistics: know your data

Categorical

Binary: Classify a person with a characteristic or not (male/female, sick/healthy, alive/death)

Ordered:

Likert scales: strongly disagree / disagree / neither / agree / strongly agree

Rankings: top 5 cricketing nations: NZ, UK, Australia…

Unordered: (nominal, multinomial, generalized logit)

Learning style: Self / Class / Team

TB Treatment outcomes: Completed, Cured, Failed, Died

Comparison between variables

Continuous vs. Continuous

Correlation: Pearson, Spearman

Regression (y=mx + b)

Continuous vs. Categorical

Parametric: t-test

Non-parametric: Mann–Whitney *U* test

Regression:

Continuous outcome: ANOVA, ANCOVA

Categorical outcome: Binary: Logistic

Ordered: ordered logit; Unordered: multinomial

Categorical vs. Categorical

Chi square: Mantel-Haenszel, Fisher's Exact, McNemar,

Risk difference - Odds Ratios (case-control studies) vs.

Relative Risks (cohort studies)

Why do we need statistical modeling

Example research question:

Impact of iron supplementation on hemoglobin level among rural pregnant women in 2022

Why do we need statistical modeling

Example research question:

Impact of iron supplementation on hemoglobin level among rural pregnant women in 2022

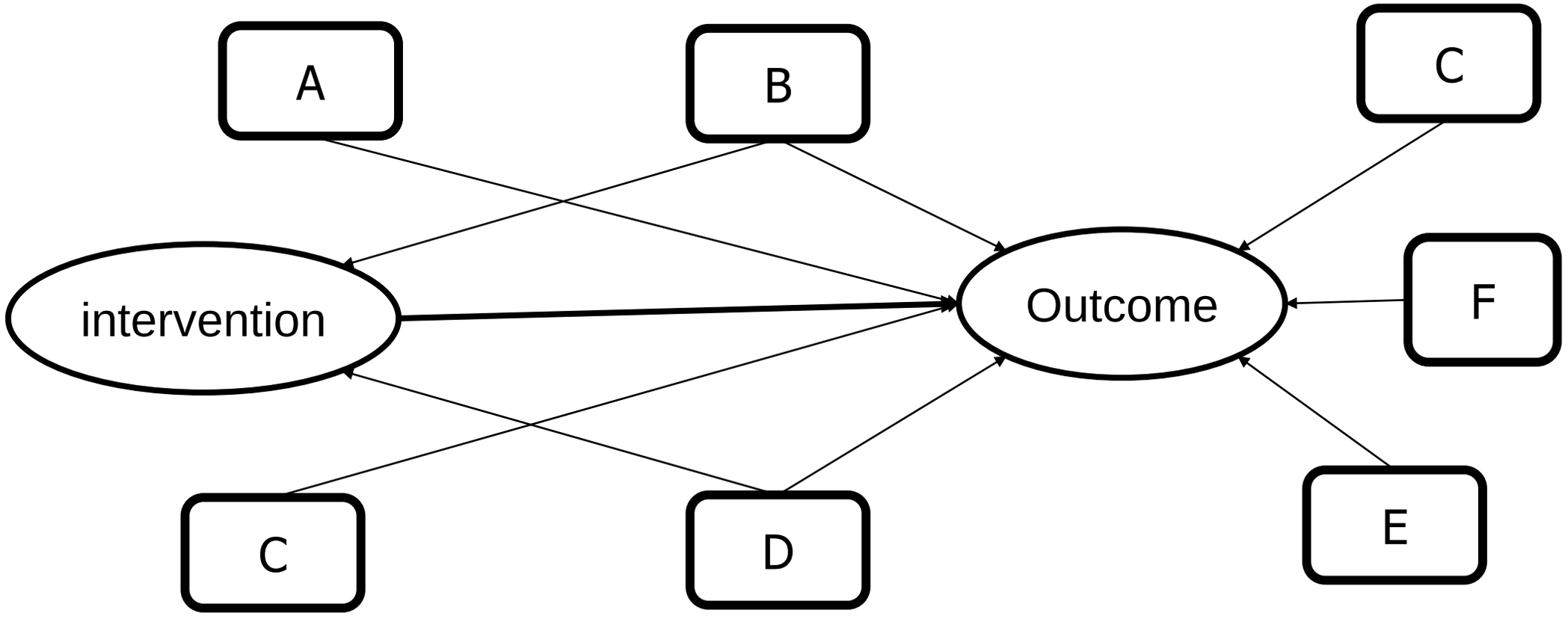
Simply, we can collect data from intervention and control group and do the ***t-test*** between two groups, or chi-square test for binary outcome

Why do we need statistical modeling

Example research question:

Impact of iron supplementation on hemoglobin level among rural pregnant women in 2022

Simply, we can collect data from intervention and control group and do the ***t-test*** between two groups, or chi-square test for binary outcome



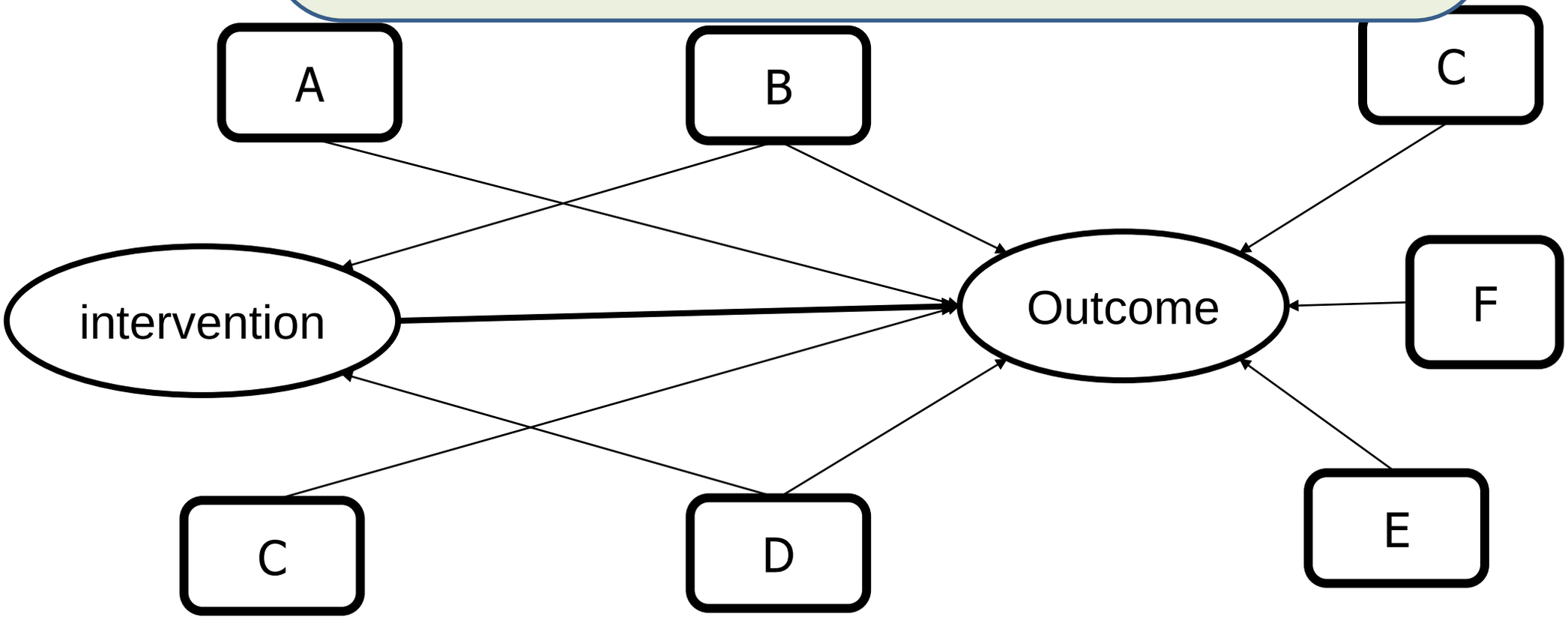
Why do we need statistical modeling

Example research question:

Impact of … rural pregnant w…

Simply, we c… the ***t-test*** b…

How adjust covariates
How to adjust confounder
How to adjust cluster if needed
How adjust time for longitudinal data



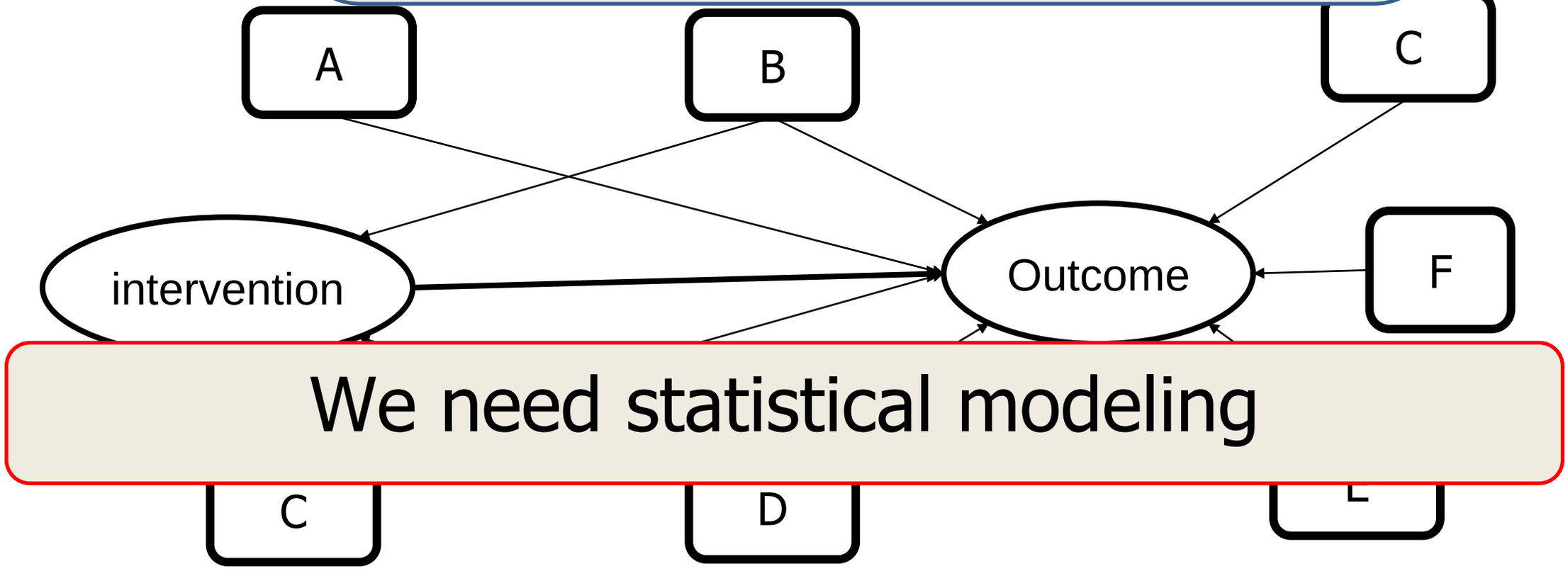
Why do we need statistical modeling

Example research question:

Impact of … rural pregnant w…

Simply, we c… the ***t-test*** b…

How adjust covariates
How to adjust confounder
How to adjust cluster if needed
How adjust time for longitudinal data



We need statistical modeling

Statistical modeling depends on design

- Cross-sectional study
- Cohort study
- Case-control study
- RCT
- Follow-up study

Which model is needed ???

Statistical modeling depends on hypothesis

p1=p0

m1=m0

OR=1

RR=1

HR=1

Which model is needed ???

Statistical modeling depends on variable

- Binary
- Quantitative
- Count
- Categorical
- Time series

Which model is needed ???

Major shortcomings in icddr,b analysis

- Data driven analysis/modeling rarely used
- Selection of variables in regression model: conceptual framework not followed
- Post-estimation not always done
- Sensitivity analysis, rare

“All models are wrong…but some are useful”

George E.P. Box

"The more accurate the map, the more it resembles the territory. The most accurate map possible would be the territory, and thus would be perfectly accurate and perfectly useless."

Neil Gaiman