INTRODUCTION TO STATISTICAL MODELING

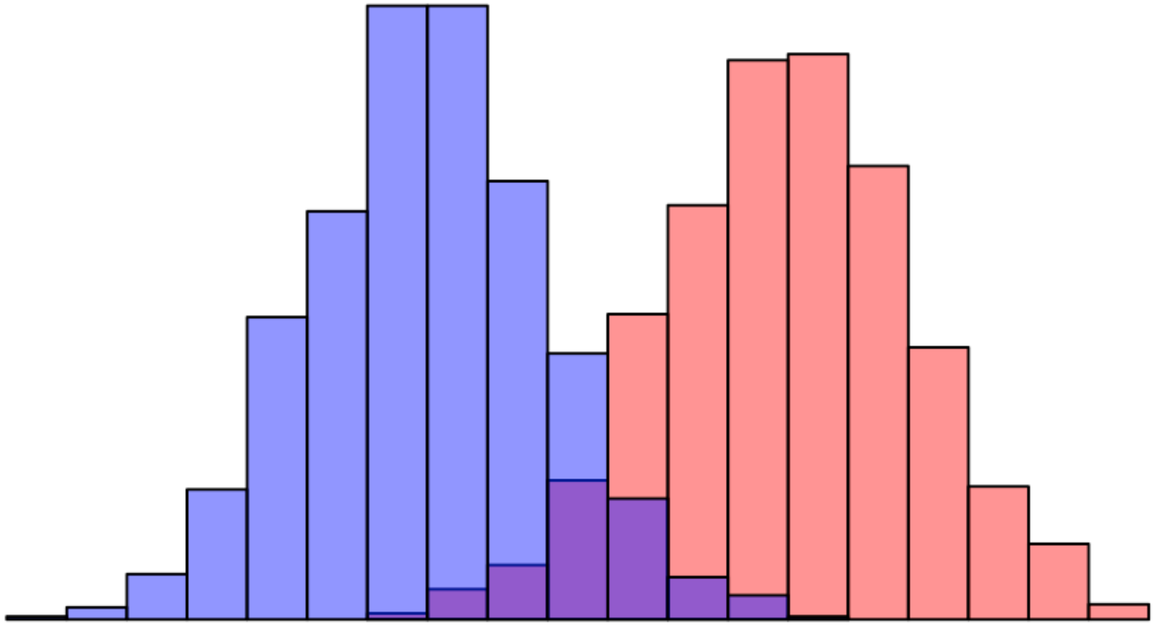
Welcome to statistical modeling!

Statistical models are used for...

- Identifying patterns in data
- Classifying events
- Untangling multiple influences
- Assessing strength of evidence

The t-test: a statistical skateboard

Are two groups different?



The t-test: a statistical skateboard

A statistical model?

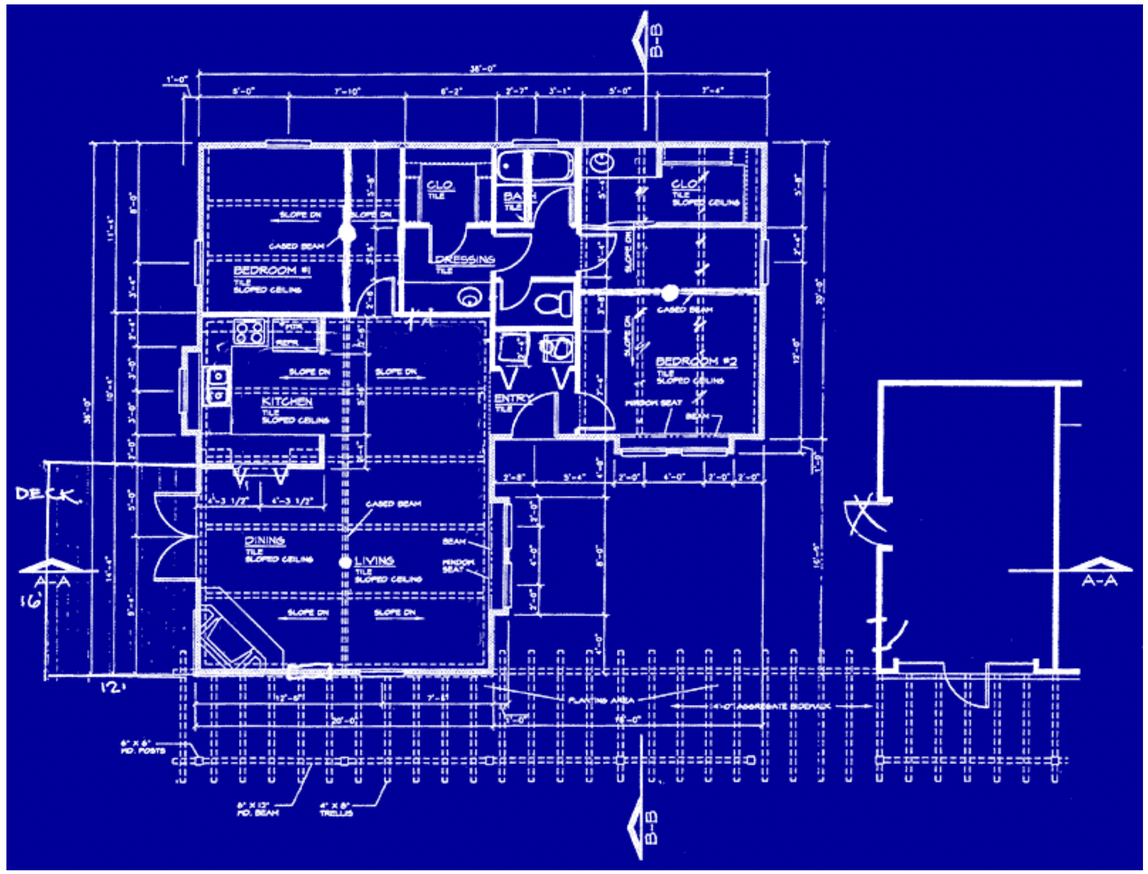
The t-test in practice

Defining "model"

A model is a **representation** for a **purpose**

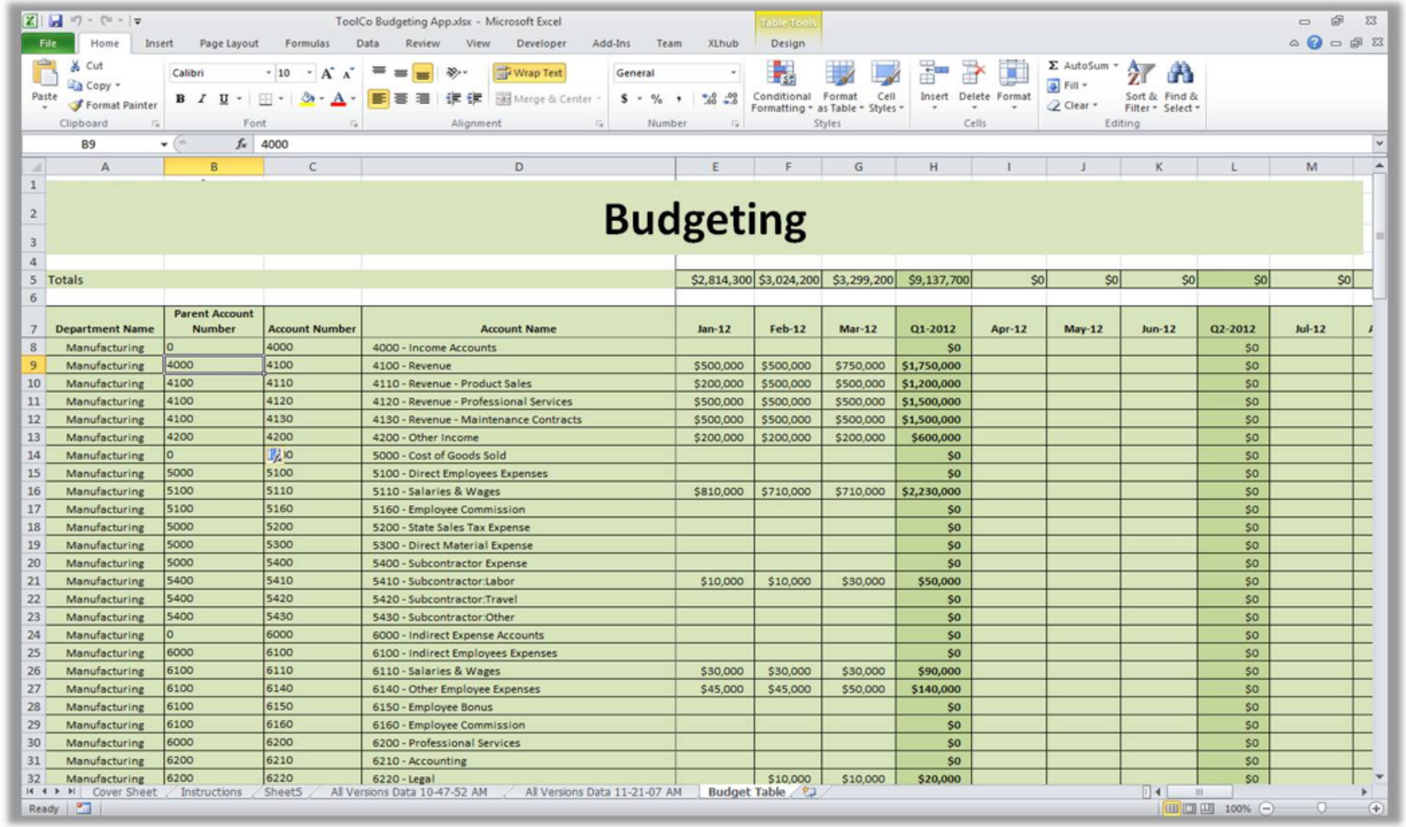
- **Representation**: it stands for something in the real world
- **Purpose**: *YOUR* particular use for the model

Some everyday models



Mathematical model

- Constructed out of **mathematical entities**
  - Numbers
  - Model formulas
  - Equations
  - ...



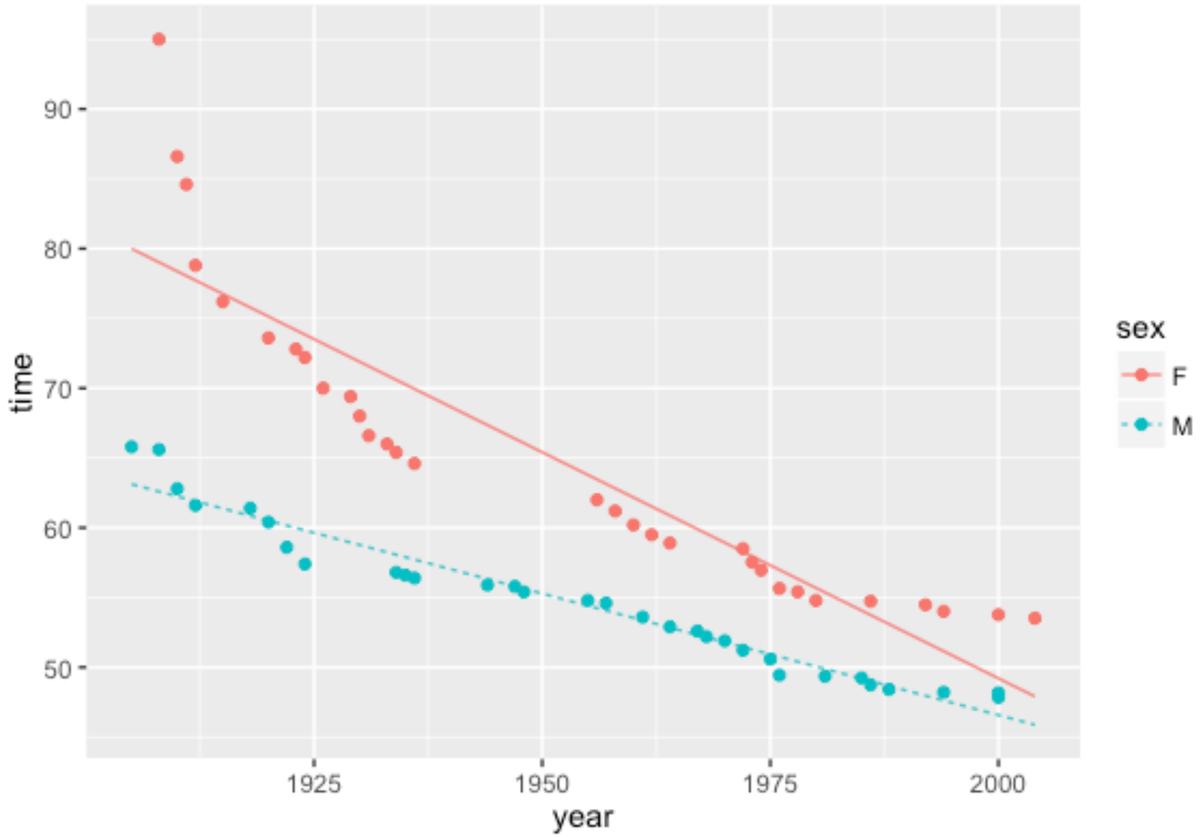
ToolCo Budgeting App.xlsx - Microsoft Excel

Budgeting

Totals				$2,814,300	$3,024,200	$3,299,200	$9,137,700	$0	$0	$0	$0	$0	
Department Name	Parent Account Number	Account Number	Account Name	Jan-12	Feb-12	Mar-12	Q1-2012	Apr-12	May-12	Jun-12	Q2-2012	Jul-12	
Manufacturing	0	4000	4000 - Income Accounts				$0				$0		
Manufacturing	4000	4100	4100 - Revenue	$500,000	$500,000	$750,000	$1,750,000				$0		
Manufacturing	4100	4110	4110 - Revenue - Product Sales	$200,000	$500,000	$500,000	$1,200,000				$0		
Manufacturing	4100	4120	4120 - Revenue - Professional Services	$500,000	$500,000	$500,000	$1,500,000				$0		
Manufacturing	4100	4130	4130 - Revenue - Maintenance Contracts	$500,000	$500,000	$500,000	$1,500,000				$0		
Manufacturing	4200	4200	4200 - Other Income	$200,000	$200,000	$200,000	$600,000				$0		
Manufacturing	0	0	5000 - Cost of Goods Sold				$0				$0		
Manufacturing	5000	5100	5100 - Direct Employees Expenses				$0				$0		
Manufacturing	5100	5110	5110 - Salaries & Wages	$810,000	$710,000	$710,000	$2,230,000				$0		
Manufacturing	5100	5160	5160 - Employee Commission				$0				$0		
Manufacturing	5000	5200	5200 - State Sales Tax Expense				$0				$0		
Manufacturing	5000	5300	5300 - Direct Material Expense				$0				$0		
Manufacturing	5000	5400	5400 - Subcontractor Expense				$0				$0		
Manufacturing	5400	5410	5410 - Subcontractor:Labor	$10,000	$10,000	$30,000	$50,000				$0		
Manufacturing	5400	5420	5420 - Subcontractor:Travel				$0				$0		
Manufacturing	5400	5430	5430 - Subcontractor:Other				$0				$0		
Manufacturing	0	6000	6000 - Indirect Expense Accounts				$0				$0		
Manufacturing	6000	6100	6100 - Indirect Employees Expenses				$0				$0		
Manufacturing	6100	6110	6110 - Salaries & Wages	$30,000	$30,000	$30,000	$90,000				$0		
Manufacturing	6100	6140	6140 - Other Employee Expenses	$45,000	$45,000	$50,000	$140,000				$0		
Manufacturing	6100	6150	6150 - Employee Bonus				$0				$0		
Manufacturing	6100	6160	6160 - Employee Commission				$0				$0		
Manufacturing	6000	6200	6200 - Professional Services				$0				$0		
Manufacturing	6200	6210	6210 - Accounting				$0				$0		
Manufacturing	6200	6220	6220 - Legal		$10,000	$10,000	$20,000				$0		

Statistical model

- A special type of mathematical model
- Informed by data
- Incorporates uncertainty and randomness



In this course, you will...

- Design a model
- Use data to train your model
- Interpret and use your model

INTRODUCTION TO STATISTICAL MODELING

Let’s practice!

INTRODUCTION TO STATISTICAL MODELING

R objects for statistical modeling

Let's review model definitions

- A **model** is a representation for a purpose
- A **mathematical model** is built from mathematical stuff
- A **statistical model** is trained on data, built on objects

Data frames

- Columns are *variables*, which have *names*
- Contents of variables are *values*
- Rows are *cases* (e.g. people)
- The case is the object from which values for variables are measured

name	prof_goal	age	fav_color
Daniel	Astronaut	5	Blue
Nicky	Rock star	7	Green
Tom	Dancer	4	Pink

Functions

- Model training functions

```
> my_model <- rpart(formula, data = DF, ...)
```

- Functions to evaluate models

```
> predict(my_model, newdata = DF, ...)
```

Formulas

```
wage ~ exper + sector
```

Basic statistics with R formulas

```
# Use the mosaic package
> library(mosaic)

> mean(wage ~ sector, data = CPS85)
 clerical     const     manag     manuf     other      prof
 7.422577  9.502000 12.704000  8.036029  8.500588 11.947429
 sales     service
 7.592632  6.537470
```

Model formulas

```
# Describes how to relate variables together
> my_model <- lm(wage ~ exper + sector, data = CPS85)
```

Formulas in English

`wage ~ sector` can be read as any of these:

- wage as a function of sector
- wage accounted for by sector
- wage modeled by sector
- wage explained by sector
- wage given sector
- wage broken down by sector

INTRODUCTION TO STATISTICAL MODELING

Let’s practice!