Parametric inference for regression and multiple regression

Overview

Parametric inference for regression

Multiple regression

Strategies for identifying the appropriate methods to use

Practice problems

Parametric inference for regression

Review of regression (class 6 and 7)

In **linear regression** we fit a line to the data, called the **regression line**

$$\hat{y} = a + b \cdot x$$

$$\textit{Predicted response} = a + b \cdot \textit{Explanatory}$$

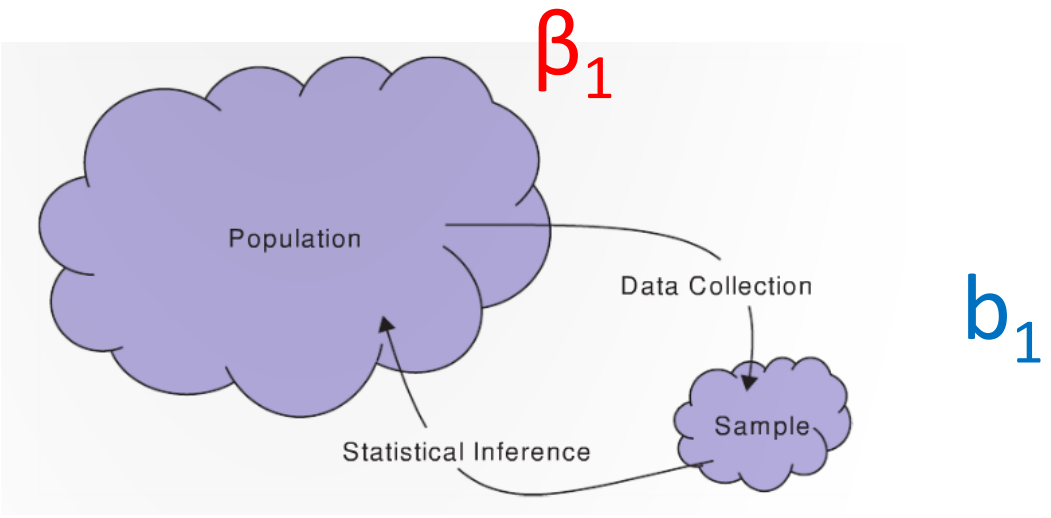
Change in notation to be consistent with the Lock5 and what most statisticians use

$$\textit{Predicted response} = b_0 + b_1 \cdot \textit{Explanatory}$$

Inference on simple linear regression

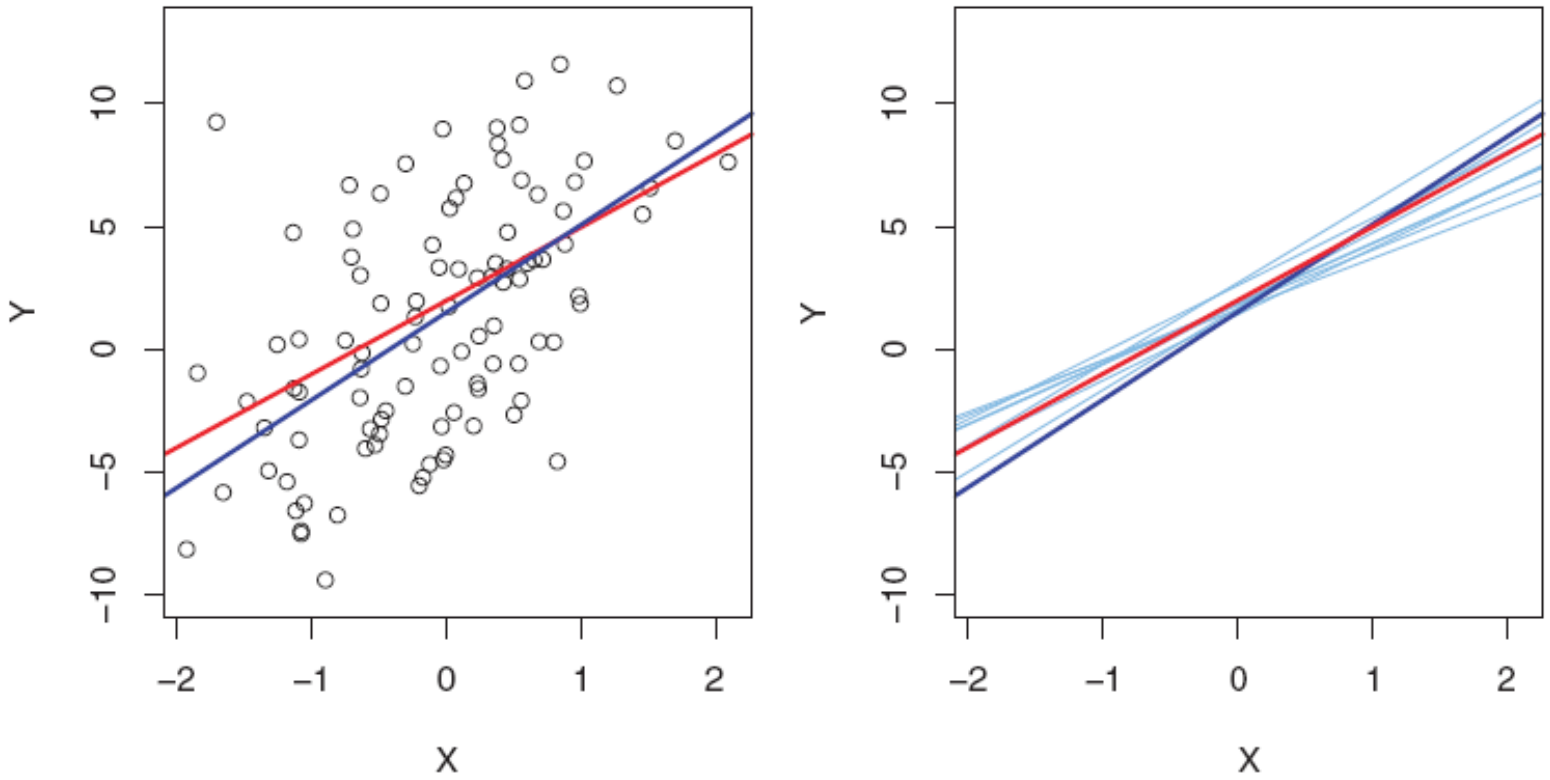
The Greek letter β_1 is used to denote the slope of the population

The letter b_1 is typically used to denote the slope of the sample



Population: β_1

Sample estimates: b_1

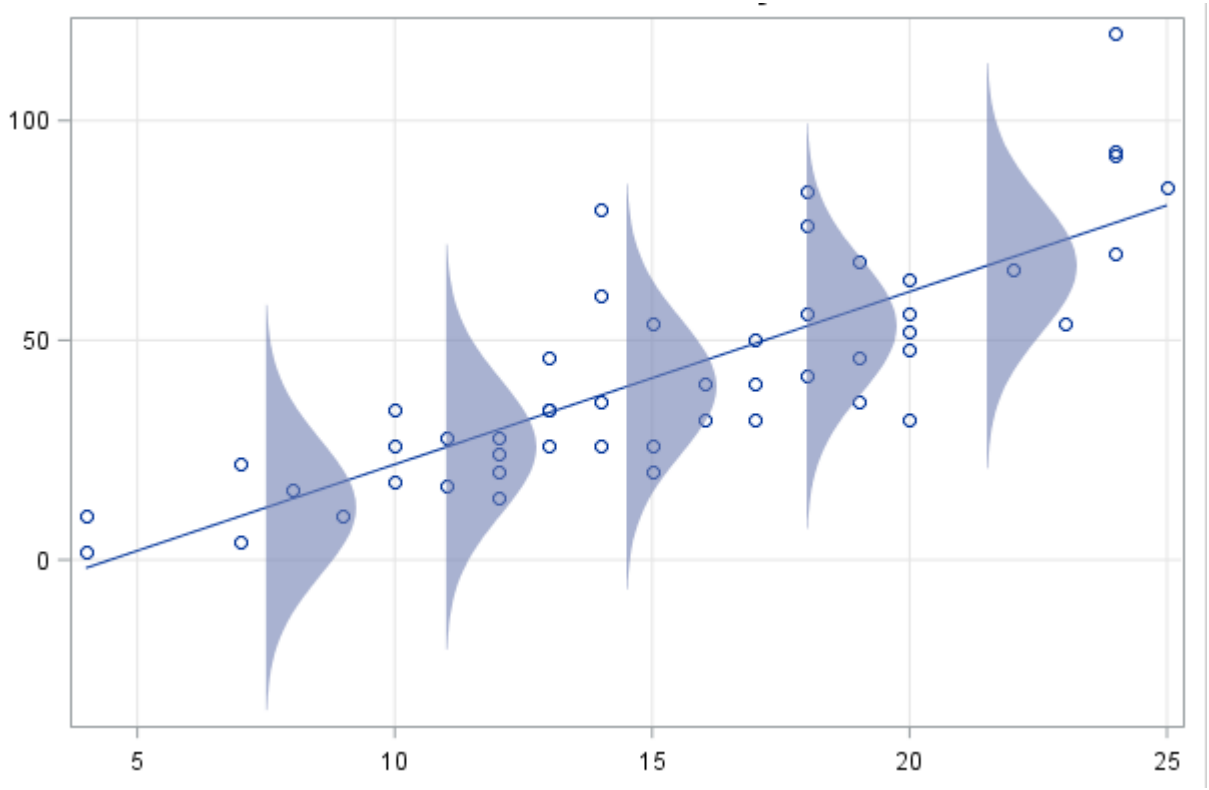


Simple linear regression underlying model

Intercept Slope } *Parameters*

$$Y \approx \beta_0 + \beta_1 x$$

$$Y = \beta_0 + \beta_1 x + \epsilon$$



$$\epsilon \sim N(0, \sigma_\epsilon)$$

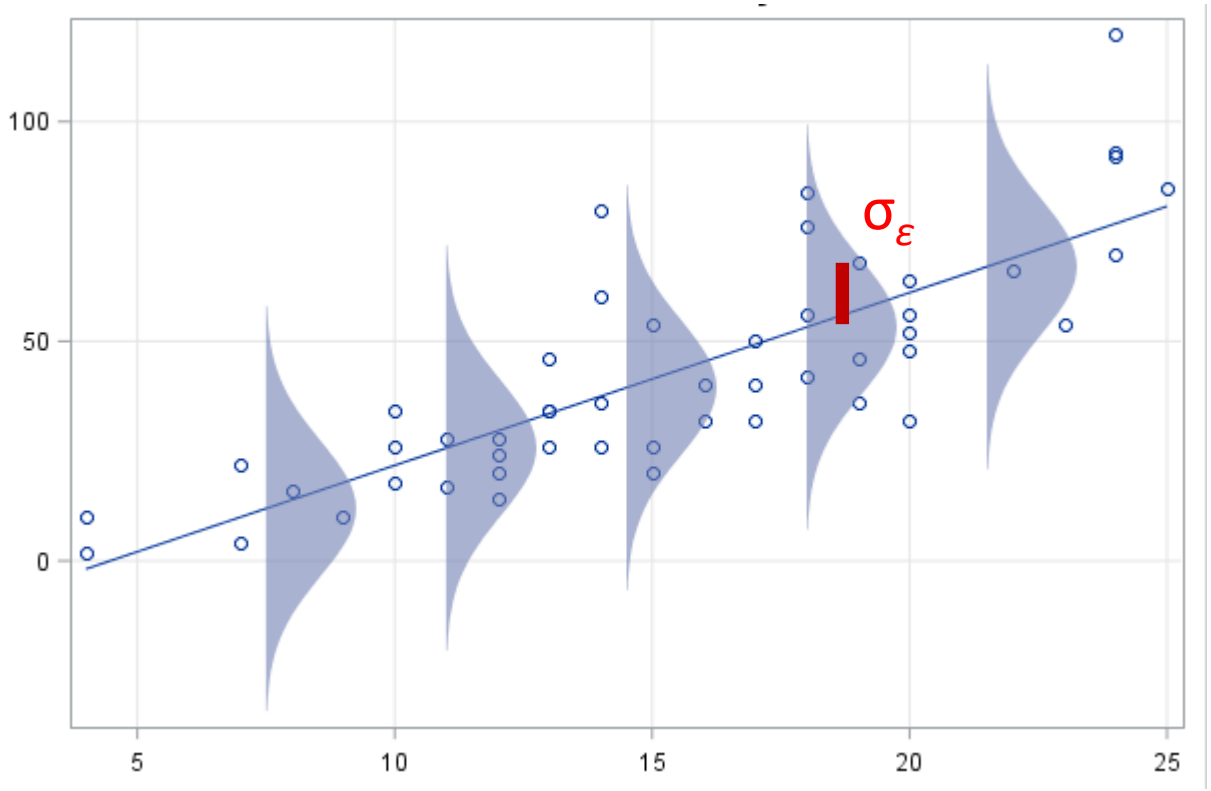
$$\hat{y} = b_0 + b_1 x$$

$$SSE = \sum_{i=1}^{n} (y_i - (b_0 + b_1 x))^2$$

Estimating σ_ε

We can also use the **standard deviation of errors** as an estimate standard deviation of irreducible noise $\boldsymbol{\sigma_\varepsilon}$

- This is also called the **residual standard error (RSE)**



$$\hat{\sigma}_\epsilon = \sqrt{\frac{1}{n-2}SSE}$$

$$= \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Hypothesis test for regression coefficients

We can run hypothesis tests to assess whether there is a relationship between y and x, and calculate p-values

- H_0: $\beta_1 = 0$ (slope is 0, so no relationship between x and y
- H_A: $\beta_1 \neq 0$

One type of hypothesis test we can run is based on a t-statistic: $t = \frac{b_1 - 0}{SE_{b_1}}$

- The t-statistic comes from a t-distribution with n - 2 degrees of freedom

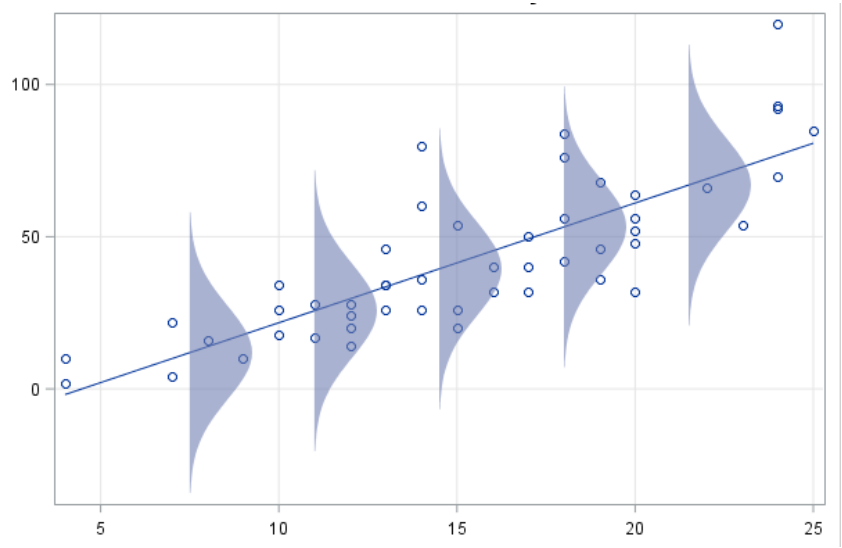
$$SE_{b_1} = \frac{\sigma_\epsilon}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

$$SE_{b_0} = \sigma_\epsilon \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

Inference using parametric methods

When using parametric methods, we make the following assumptions:

- **Normality**: residuals are normally distributed around the regression line
- **Homoscedasticity**: constant variance over the whole range of x values
- **Linearity**: A line can describe the relationship between x and y
- **Independence**: each data point is independent from the other points



These assumptions are usually checked after the models are fit using ‘regression diagnostic’ plots

Confidence intervals for regression coefficients

For the slope coefficient , the confidence interval is: $b_1 \pm t^* \cdot SE_{b_1}$

Where: $SE_{b_1} = \frac{\sigma_\epsilon}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$

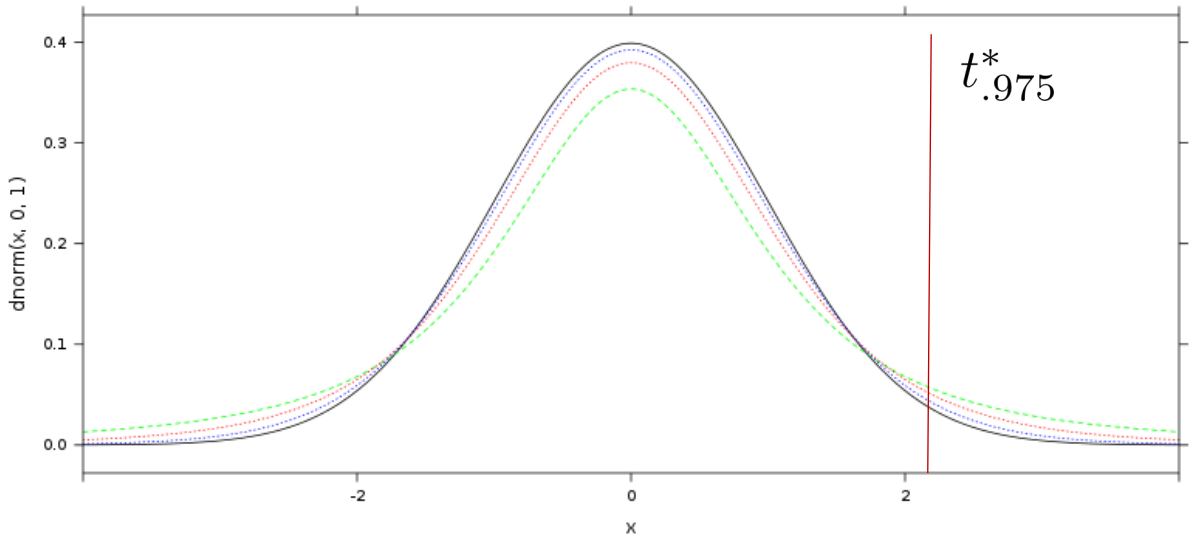
t* is the critical value for the t_{n-2} density curve needed to obtain a desired confidence level

N(0, 1)

df = 2

df = 5

df = 15



Let’s try it in R…

Multiple regression

Multiple regression

In multiple regression we try to predict a quantitative response variable *y* using several predictor variables $x_1, x_2, \ldots, x_k$

For multiple linear regression, the underlying model is:

$$y = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \ldots \beta_k \cdot x_k + \epsilon$$

We estimate coefficients b_i using a data set to make predictions $\hat{y}$

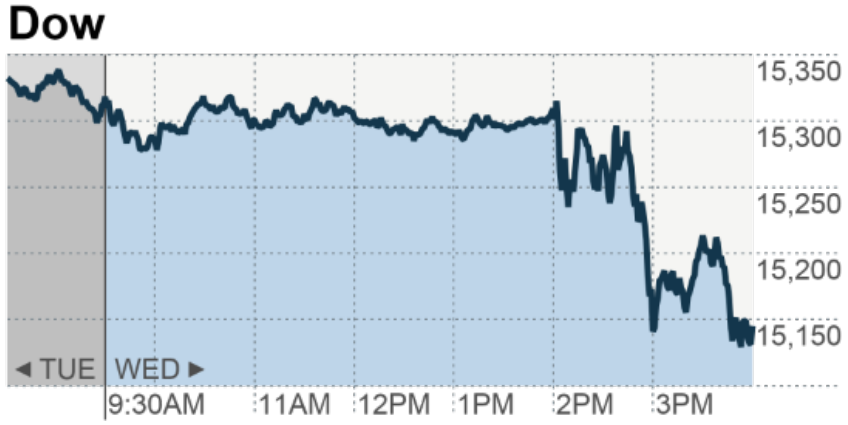
$$\hat{y} = b_0 + b_1 \cdot x_1 + b_2 \cdot x_2 + \ldots + b_k \cdot x_k$$

Multiple regression

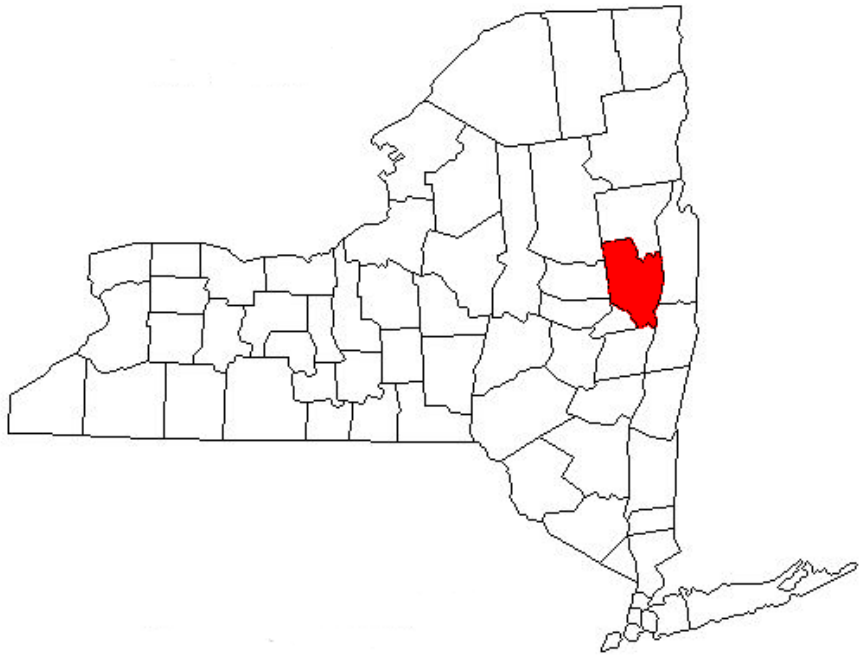
$$\hat{y} = b_0 + b_1 \cdot x_1 + b_2 \cdot x_2 + \ldots + b_k \cdot x_k$$

There are many uses for multiple regression models including:

- To make predictions as accurately as possible
- To understand which predictors (x) are related to the response variable (y)



Predicting house prices in Saratoga NY



Predicting house prices in Saratoga NY

We build a linear model to predict the **price** of houses based on:

- y-intercept
- Living area (sq feet?)
- Number of bathrooms
- Number of fireplaces
- Lot size (sq acres?)
- Age of the house (years)

$$\hat{y} = b_0 + b_1 \cdot x_1 + b_2 \cdot x_2 + b_3 \cdot x_3 + b_4 \cdot x_4 + b_5 \cdot x_5$$

$$\hat{price} = b_0 + b_1 \cdot area + b_2 \cdot bathrooms + b_3 \cdot fireplaces + b_4 \cdot lot + b_5 \cdot age$$

Predicting house prices in Saratoga NY

We build a linear model to predict the **price** of houses based on:

- y-intercept
- Living area (sq feet?)
- Number of bathrooms
- Number of fireplaces
- Lot size (sq acres?)
- Age of the house (years)

b_0	5670
b_1	69
b_2	17851
b_3	9799
b_4	890
b_5	-170

We can find the coefficients (b_i's) by minimizing the sum of the squared residuals

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$\hat{y} = b_0 + b_1 \cdot x_1 + b_2 \cdot x_2 + b_3 \cdot x_3 + b_4 \cdot x_4 + b_5 \cdot x_5$$

$$\hat{price} = b_0 + b_1 \cdot area + b_2 \cdot bathrooms + b_3 \cdot fireplaces + b_4 \cdot lot + b_5 \cdot age$$

Predicting house prices in Saratoga NY

We build a linear model to predict the **price** of houses based on:

- y-intercept
- Living area (sq feet?)
- Number of bathrooms
- Number of fireplaces
- Lot size (sq acres?)
- Age of the house (years)

b_0	5670
b_1	69
b_2	17851
b_3	9799
b_4	890
b_5	-170

We can find the coefficients (b_i's) by minimizing the sum of the squared residuals

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$\hat{y} = b_0 + b_1 \cdot x_1 + b_2 \cdot x_2 + b_3 \cdot x_3 + b_4 \cdot x_4 + b_5 \cdot x_5$$

$$\hat{price} = 5670 + 69 \cdot area + 17851 \cdot bathrooms + 9799 \cdot fireplaces + 890 \cdot lot - 170 \cdot age$$

Predicting house prices in Saratoga NY

Suppose we wanted to predict the price of a house that had:

- Living area = 2,000
- Number of bathrooms = 2
- Number of fireplaces = 0
- Lot size = .1
- Age of the house = 30

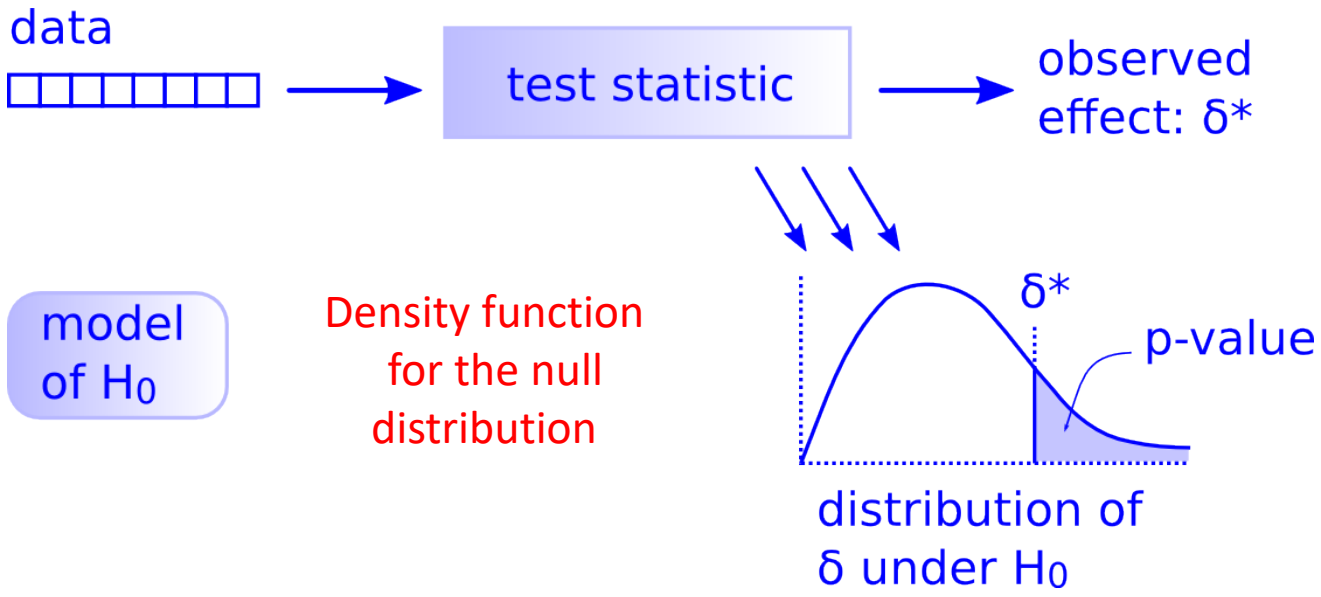
$$\hat{price} = 5670 + 69 \cdot area + 17851 \cdot bathrooms + 9799 \cdot fireplaces + 890 \cdot lot - 170 \cdot age$$

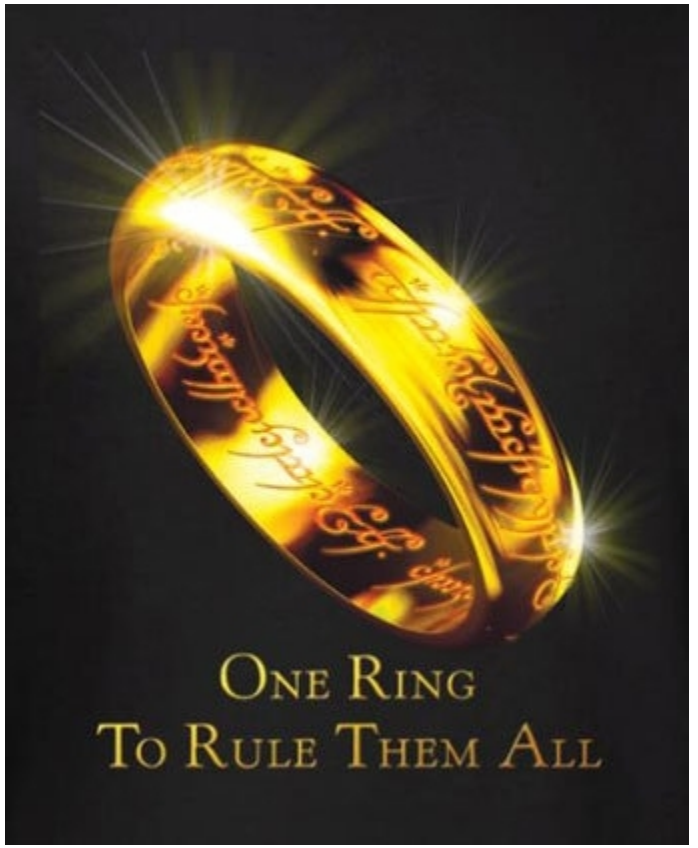
Let's try fitting multiple regression models in R!

Choosing the appropriate hypothesis test and confidence interval

One test to rule them all

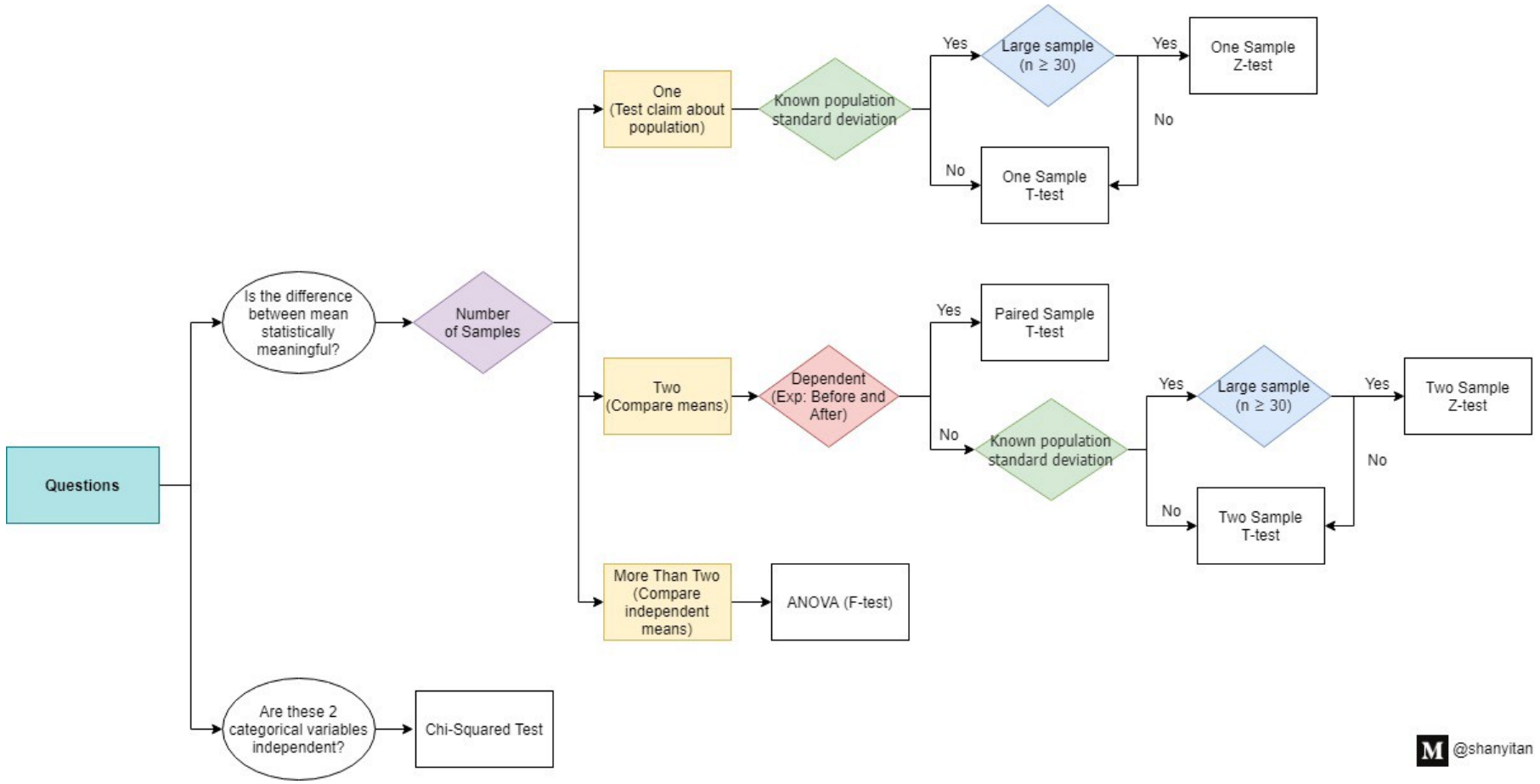
There is only one [hypothesis test](#)!





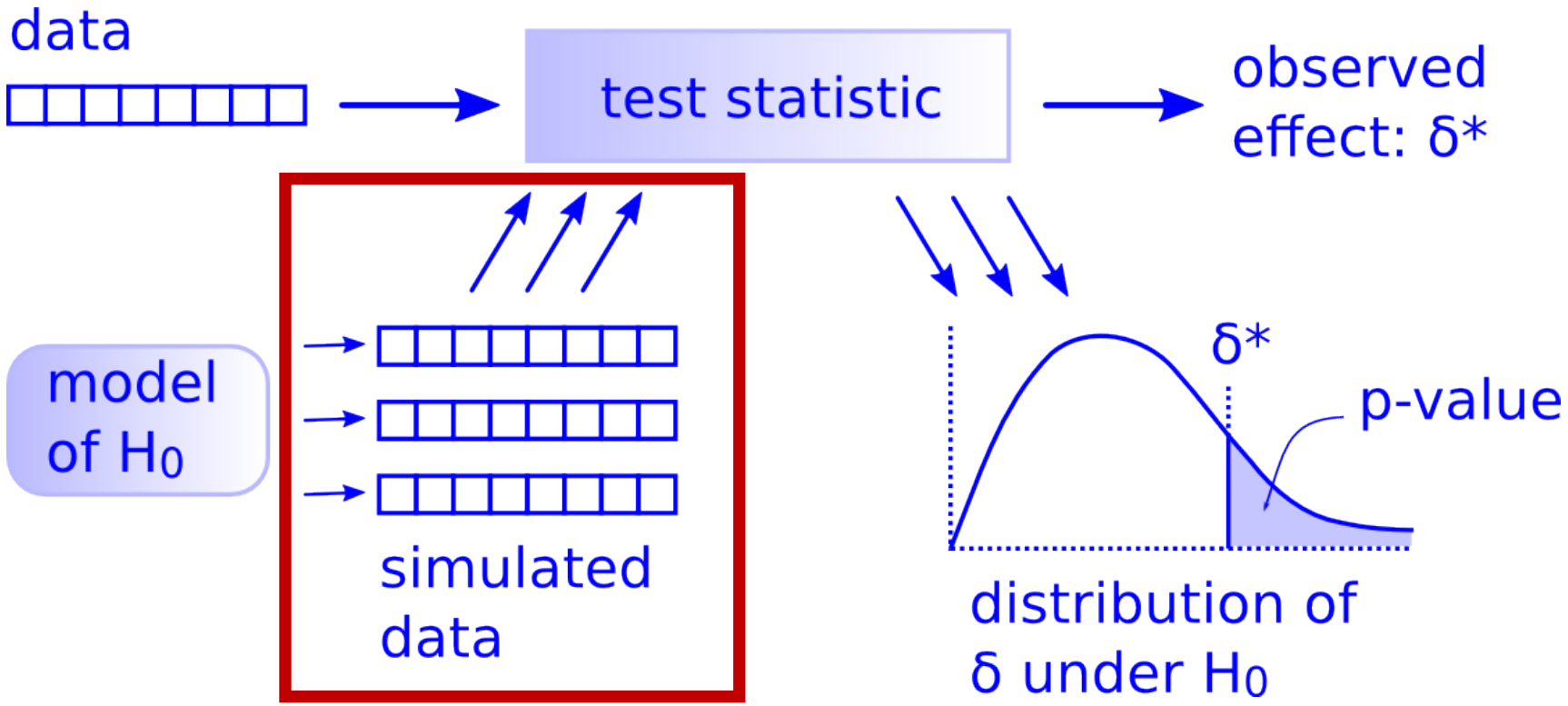
Just follow the 5 hypothesis tests steps!

Choosing the appropriate parametric test



Data	1 Sample	2 Samples	> 2 Samples
Categorical data	H_0: $\pi = p_0$ H_A: $\pi \neq p_0$ z-test $z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$	H_0: $\pi_1 = \pi_2$ H_A: $\pi_1 \neq \pi_2$ z-test or a chi-square $z = \frac{\hat{p_1} - \hat{p_2}}{\sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{n_1} + \frac{\hat{p_2}(1-\hat{p_2})}{n_2}}}$	H_0: $\pi_1 = p_1,\ \pi_2 = p_2,\ \ldots\ , \pi_k = p_k$ H_A: At least one p_i is different than specified chi-square test $\chi^2 = \sum_{i=1}^{k} \frac{(Observed_i - Expected_i)^2}{Expected_i}$ df = k - 1
Quantitative data	H_0: $\mu = v_0$ H_A: $\mu \neq v_0$ One sample t-test $t = \frac{\bar{x} - v_0}{s/\sqrt{n}}$ df = n - 1	H_0: $\mu_1 = \mu_2$ H_A: $\mu_1 \neq \mu_2$ Two sample t-test $t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$ df = min $n_1 - 1,\ n_2 - 1$	H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$ H_A: At least one μ_i is different Analysis of Variance $F = \frac{\frac{1}{K-1}\sum_{i=1}^{K} n_i(\bar{x}_i - \bar{x}_{tot})^2}{\frac{1}{N-K}\sum_{i=1}^{K}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2}$ $df_1 = k,\ \ df_2 = n - k$

Choosing the appropriate resampling method



Data	1 Sample	2 Samples	> 2 Samples
Categorical data	H_0: $\pi = p_0$ H_A: $\pi \neq p_0$ Flip "coins" rflip_count()	H_0: $\pi_1 = \pi_2$ H_A: $\pi_1 \neq \pi_2$ Flip "coins" rflip_count()	H_0: $\pi_1 = p_1$, $\pi_2 = p_2$, … , $\pi_k = p_k$ H_A: At least one p_i is different than specified Flip coins rflip_count()
Quantitative data	H_0: $\mu = v_0$ H_A: $\mu \neq v_0$ resample sample(… , replace = TRUE)	H_0: $\mu_1 = \mu_2$ H_A: $\mu_1 \neq \mu_2$ Shuffle data shuffle()	H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$ H_A: At least one μ_i is different Shuffle data shuffle()

Parametric confidence intervals

Confidence intervals have the form: $statistic \pm q^* \cdot SE$

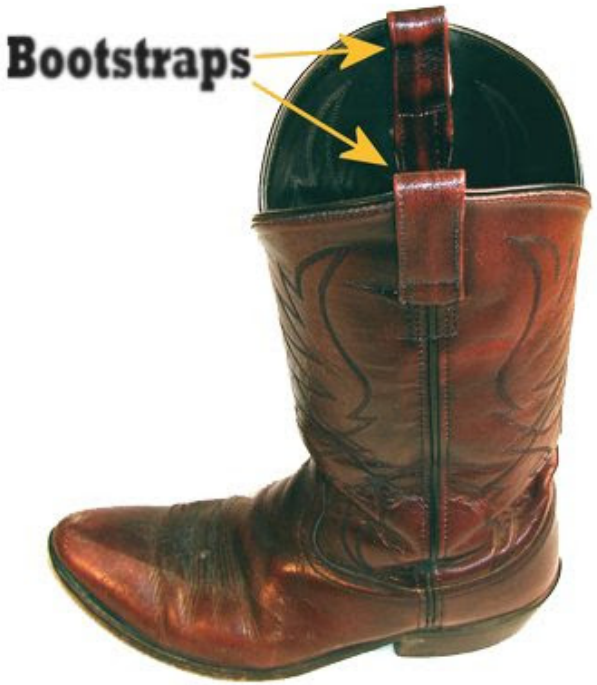
We just need the appropriate standard error (SE) formula

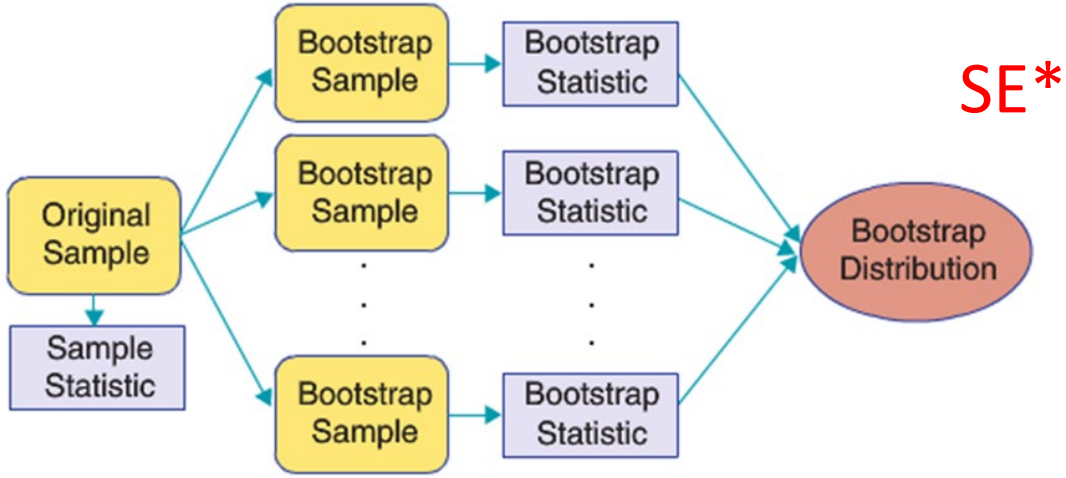
- (and to determine if we should use t* or z*)

Data	1 Sample	2 Samples
Categorical Data	$SE = \sqrt{\frac{\pi(1-\pi)}{n}}$ $\hat{p} \pm z^* \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$	$SE = \sqrt{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}}$ $\hat{p_1} - \hat{p_2} \pm z^* \sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{n_1} + \frac{\hat{p_2}(1-\hat{p_2})}{n_2}}$
Quantitative Data	$SE = \frac{s}{\sqrt{n}}$ $\overline{x} \pm t^* \frac{s}{\sqrt{n}}$	$SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ $(\overline{x_1} - \overline{x_2}) \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

Computational confidence intervals

The bootstrap!





$$Statistic \pm z^* \cdot SE^*$$

Additional hypothesis tests

Suppose in the future you want to test a hypothesis we have not covered in this class. What should you do?

- For example, H_0: $\sigma^2_1 = \sigma^2_2$

Write null and alternative hypotheses in symbols and then look up an appropriate test

- For example, H_0: $\sigma^2_1 = \sigma^2_2$ Levene's test, Bartlett's test, or the Brown–Forsythe test
- Make sure the conditions/assumptions for the test are met
  - See how robust the test is to violations to these assumption

Side note: **non-parametric** tests are another type of hypothesis test that does makes fewer assumptions than

- i.e., they do not assume that data comes from a normal distribution
  - Based on ranks, similar to the relationship between the mean and the median

How to use statistical methods to analyze real data

Know the scientific questions that you want to address and have data analysis plan ***before*** you collect data!

- i.e., state H_0's and H_A's for the questions you want to address, find the appropriate test, etc.

Run a pilot study to get a sense of the data you will collect in your real study

- Will give a sense of the distribution of the data (is the data normal, etc.)
- You can do power calculations as well to estimate the samples size n that you will need

Ideally can pre-register your data analysis plan before collecting the data

- Can help with the replication crisis


THE TRUTH IS OUT THERE

Practice finding appropriate data analysis methods

1. Identify the type of data you are dealing with:
   - A. What are the cases/observational units
   - B. What are the variable(s) of interest
   - C. How many samples/parameters are you comparing
   - D. Are the statistics/parameters categorical or quantitative

2. Write down the null and alternative hypothesis

3. Identify which test is appropriate for you null and alternative hypothesis

Practice problems

Let's go through some practice problems

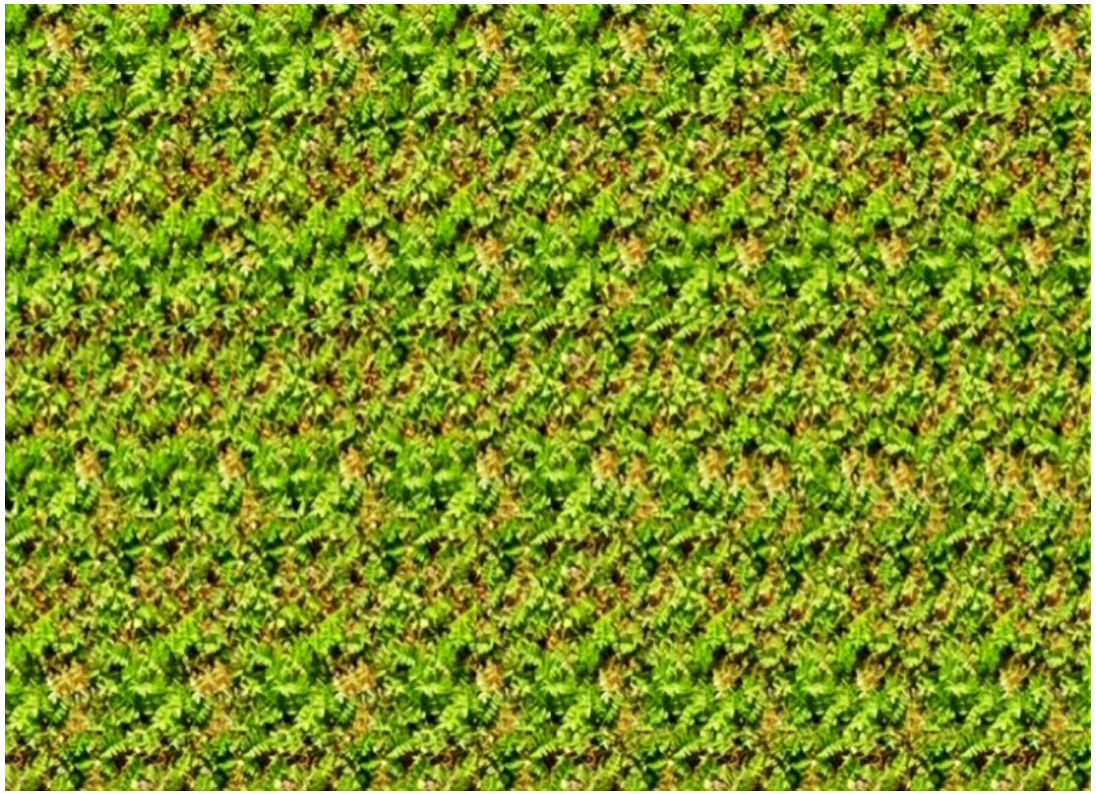
1. Identify what are the cases and variables of interest
2. Write down the null and alternative hypotheses
3. Identify the appropriate hypothesis test to use
4. Optionally: run the appropriate hypothesis test

These examples are from the review classes, so see those videos for those classes for more information

Stereograms

Stereograms appear to be composed entirely of random dots. However, they contain separate images that a viewer can “fuse” into a three-dimensional (3D) image by staring at the dots while defocusing the eyes.

An experiment was performed to determine whether knowledge of the embedded image affected the time required for subjects to fuse the images.



Stereograms

One group of subjects (group NV) received no information about the shape of the embedded object.

A second group (group VV) received a drawing of the object

The experimenters measured how many seconds it took for the subject to report that he or she saw the 3D image.

1. Identify what are the cases and variables of interest
2. Write down the null and alternative hypotheses
3. Identify the appropriate hypothesis test to use
4. Optionally: run the appropriate hypothesis test

```
SDS100::download_data("stereograms.txt")

stereograms <- read.table("stereograms.txt",
                          header = TRUE)
```

The freshman 15

Is it true that students tend to gain weight during their first year in college?

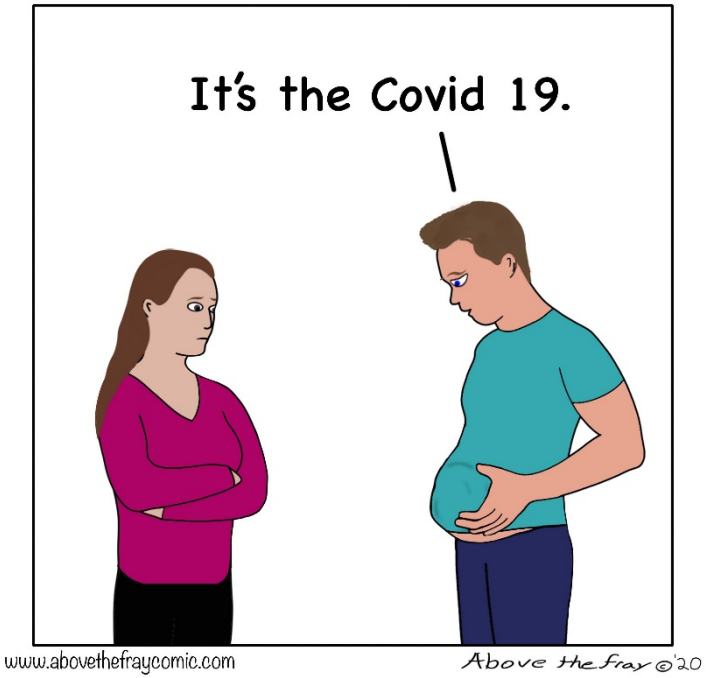
Cornell Professor of Nutrition David Levitsky recruited students from two large sections of an introductory health course. Although they were volunteers, they appeared to match the rest of the freshman class in terms of demographic variables such as sex and ethnicity.

The students were weighed during the first week of the semester, then again 12 weeks later.

Based on Professor Levitsky’s data, let's run a hypothesis test to see if students did indeed gain weight and create a confidence interval as well.

The freshman 15

1. Identify what are the cases and variables of interest
2. Write down the null and alternative hypotheses
3. Identify the appropriate hypothesis test to use
4. Optionally: run the appropriate hypothesis test




```
SDS100::download_data("freshman-15.txt")
freshman <- read.table("freshman-15.txt", header = TRUE)
```

Zodiac CEOs

Fortune magazine collected the zodiac signs of 256 heads of the largest 400 companies.

Question: Are more CEOs born under particular zodiac signs?

1. Identify what are the cases and variables of interest
2. Write down the null and alternative hypotheses
3. Identify the appropriate hypothesis test to use
4. Optionally: run the appropriate hypothesis test

```
SDS100::download_file("zodiac.csv")
zodiac <- read.csv("zodiac.csv", header = TRUE)
```

Do children's preferences change as they age?

Students in grades 4-6 in selected schools in Michigan, were asked the following question:

What would you *most* like to do at school?

A. Make good grades.

B. Be good at sports.

C. Be popular.

Are students' preferences different in different grades?

- Do steps 1-4

```
SDS100::download_data("popularkids.txt")
kids <- read.table("popularkids.txt", header = TRUE)
```

Migraine medication

A pharmaceutical company tested three formulations of a pain relief medicine for migraine headache sufferers.

For the experiment, 27 volunteers were selected and 9 were randomly assigned to one of three drug formulations.

The subjects were instructed to take the drug during their next migraine headache episode and to report their pain on a scale of 1 = no pain to 10 =extreme pain 30 minutes after taking the drug.

Is there a difference on average between the drugs effectiveness?

- Do steps 1-4

```
SDS100::download_data("analgesics.txt")
drugs  <- read.table("analgesics.txt", header = TRUE)
```