Inference for regression continued, class review, and conclusions

Overview

Parametric inference for regression

Quick review of class and next steps

Conclusions

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**

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

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

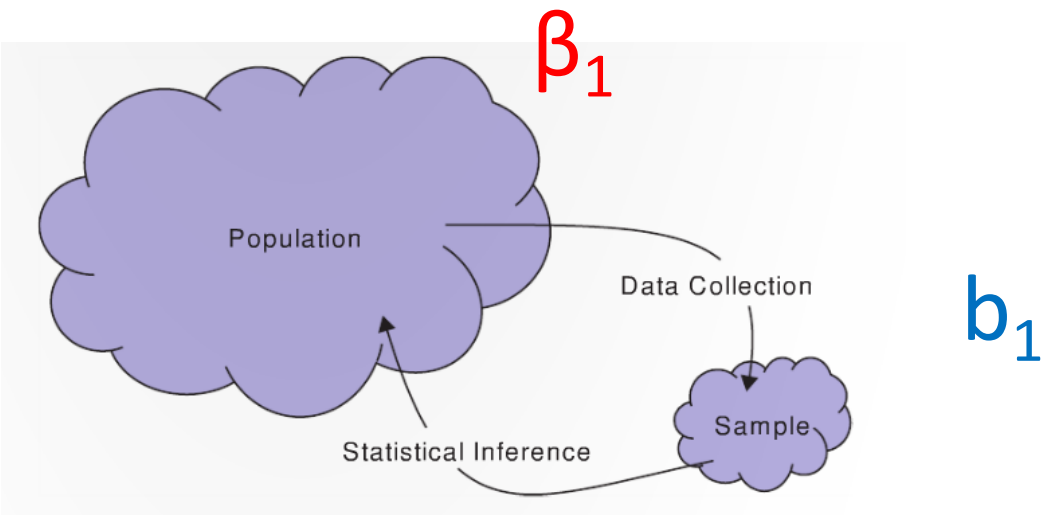
Change in notation to be consistent with the Lock5 and what most statisticians use

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

Inference on simple linear regression

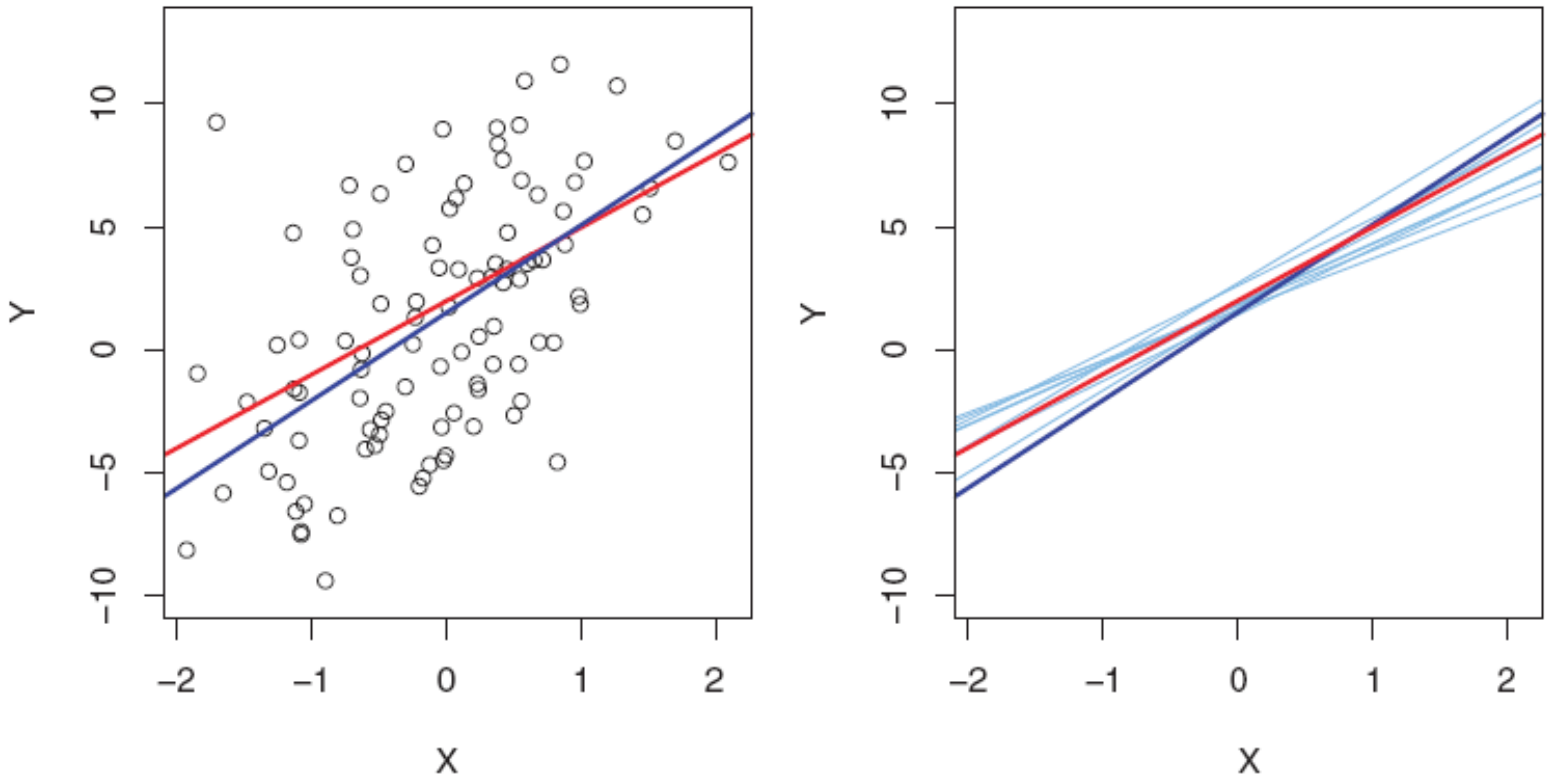
The Greek letter β_1 is used to denote the slope of the population

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



Population: β_1

Sample estimates: b_1

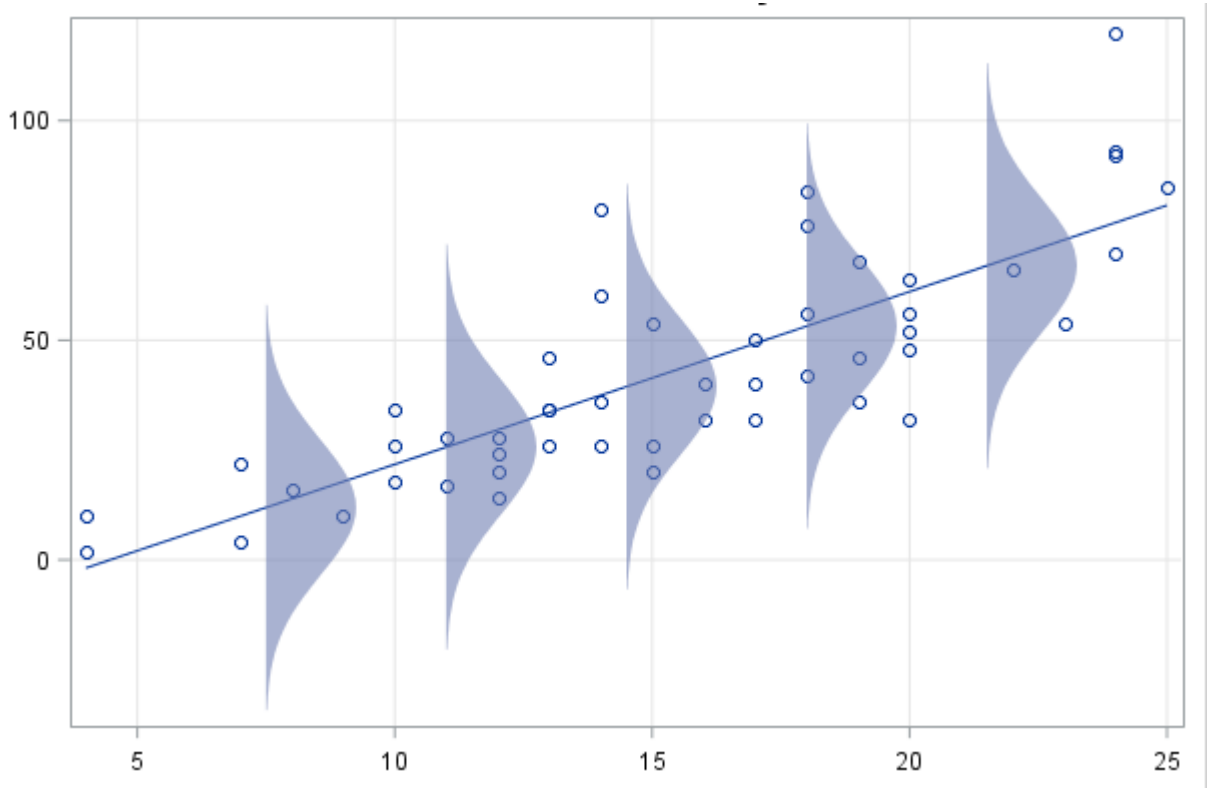


Simple linear regression underlying model

Intercept Slope } *Parameters*

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

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



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

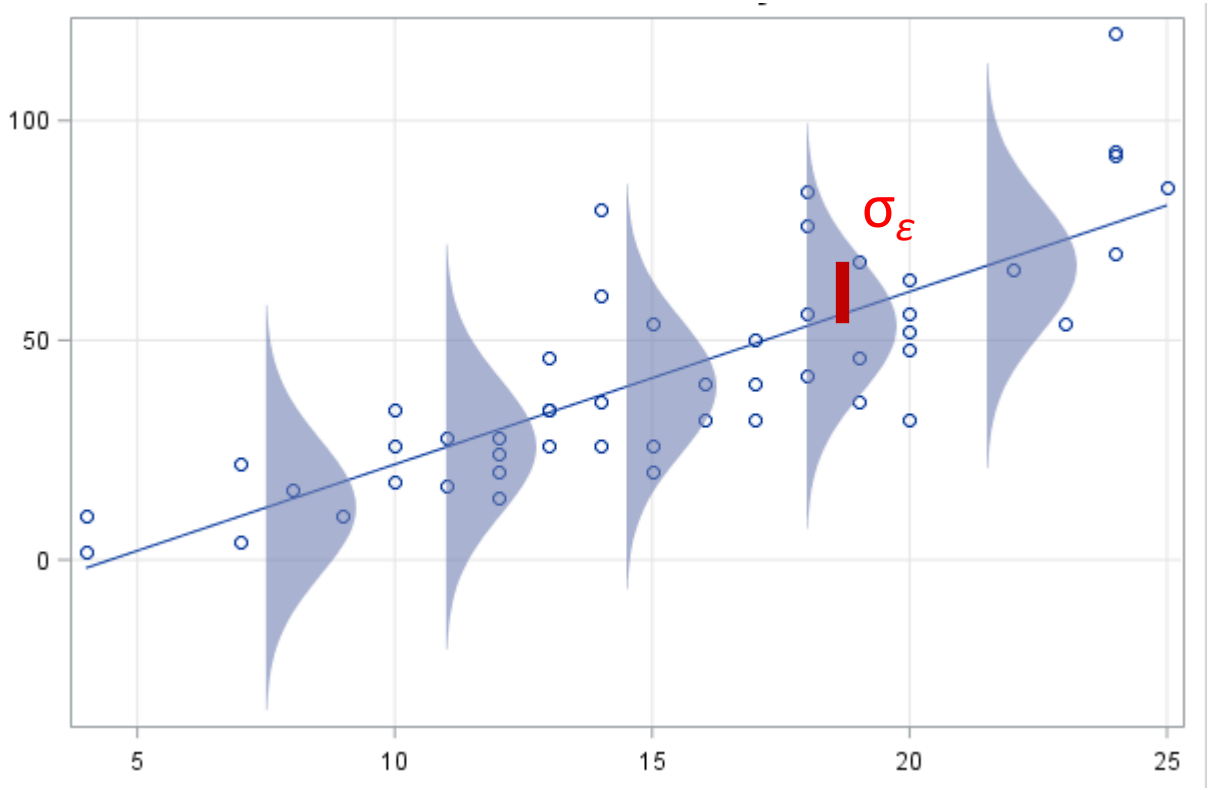
$$\hat{y} = b_0 + b_1 x$$

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

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

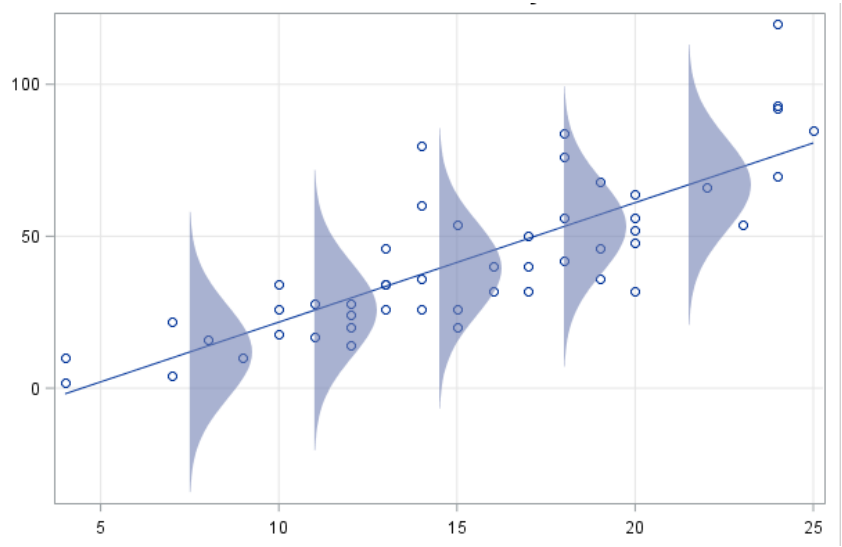
$$SE_{b_1} = \frac{\sigma_\epsilon}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

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

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

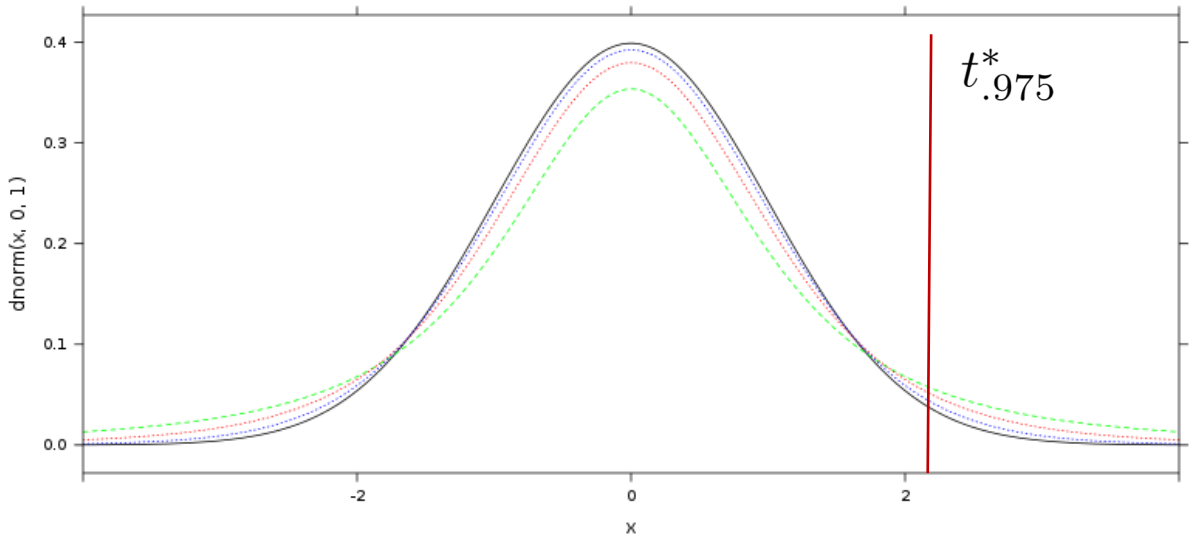
t* is the critical value for the t_{n-2} density curve needed to obtain a desired confidence level

N(0, 1)

df = 2

df = 5

df = 15

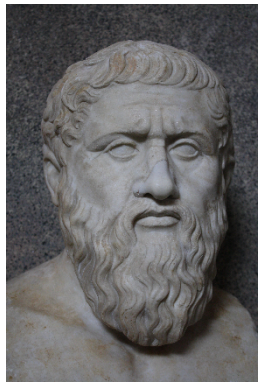


Let’s try it in R…

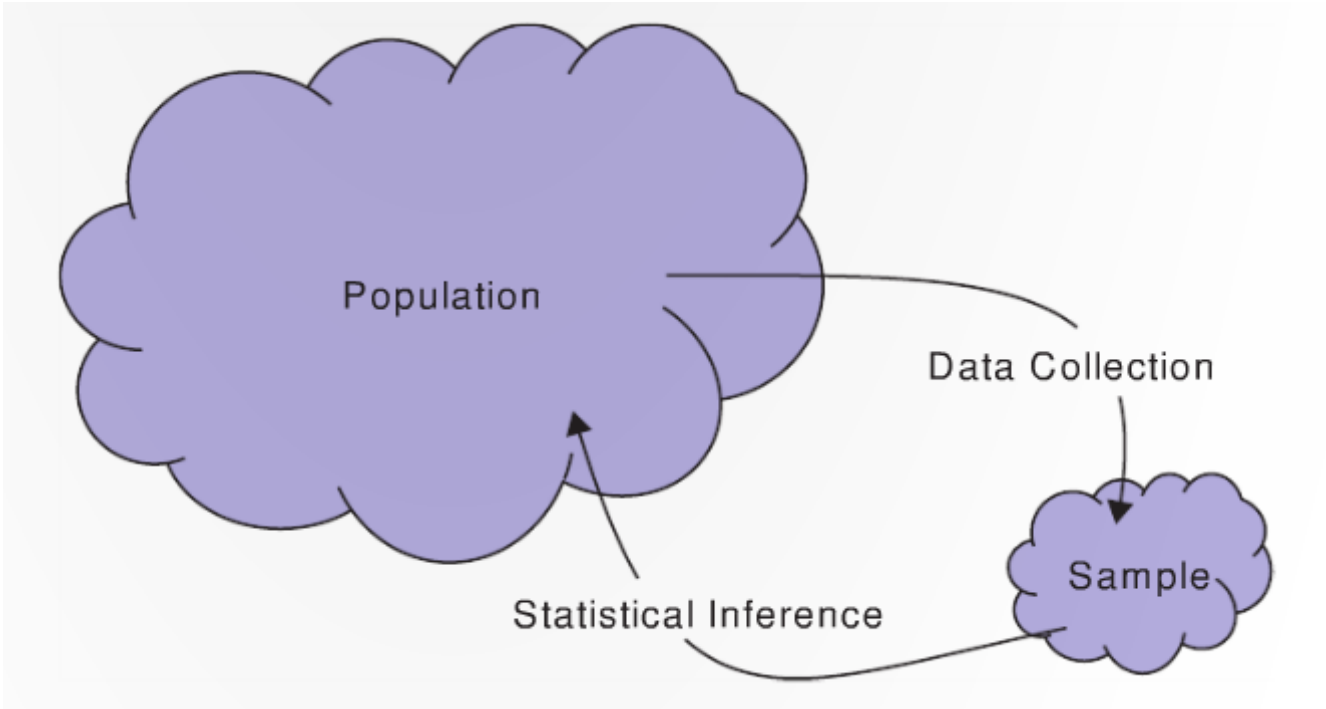
Quick review of the class

Central concepts in Statistics

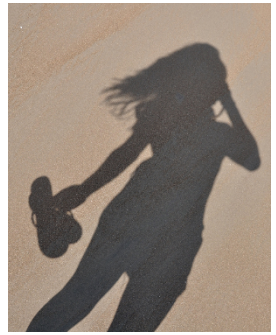
Population parameters vs. sample statistics



$\pi, \mu, \sigma, \rho, \beta$



$\hat{p}, \bar{x}, s, r, b$

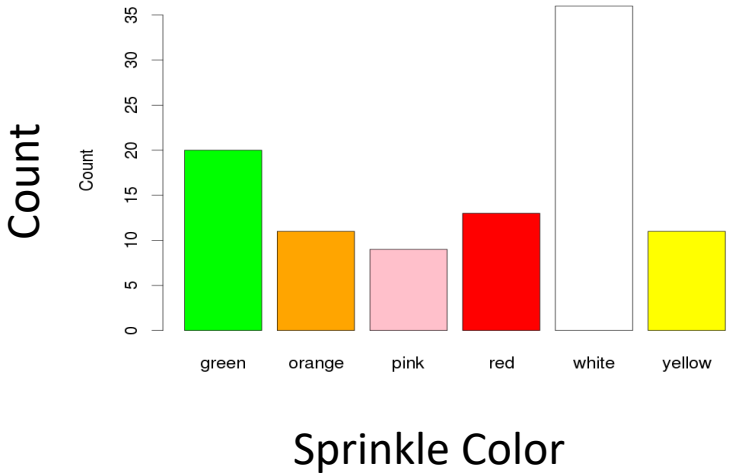


Descriptive statistics: exploring the shadows

Describing structured data

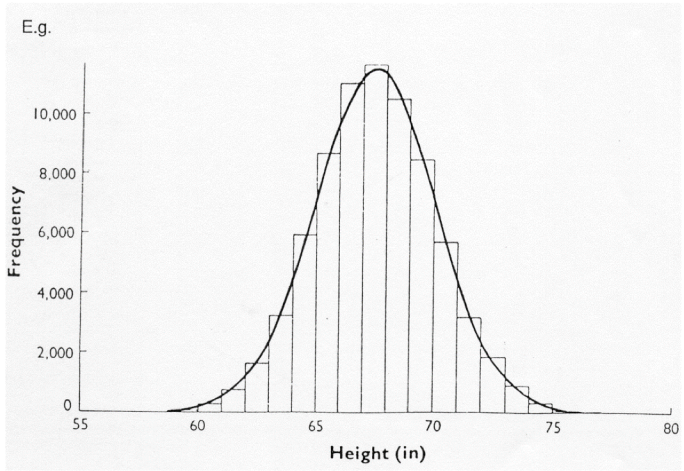
	transactionid	date_sold	make_bought	price_bought	zip_bought	mileage_bought
1	16966151	2014-09-27	Acura	30892.00	21043	40
2	16914863	2014-09-27	Toyota	25566.00	15108	297
3	15977620	2014-07-31	Nissan	34300.00	8753	0
4	18666685	2015-01-27	Subaru	30059.00	7446	10
5	14383133	2014-04-27	Honda	32508.00	97027	21
6	18196788	2014-12-18	Toyota	10819.66	95117	55246
7	15722278	2014-07-24	Audi	59630.00	90401	143

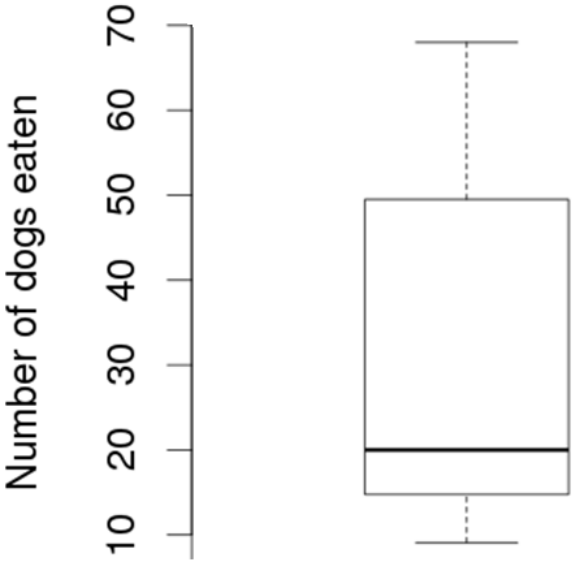
Categorical data



Proportion $\hat{p}$

Quantitative data

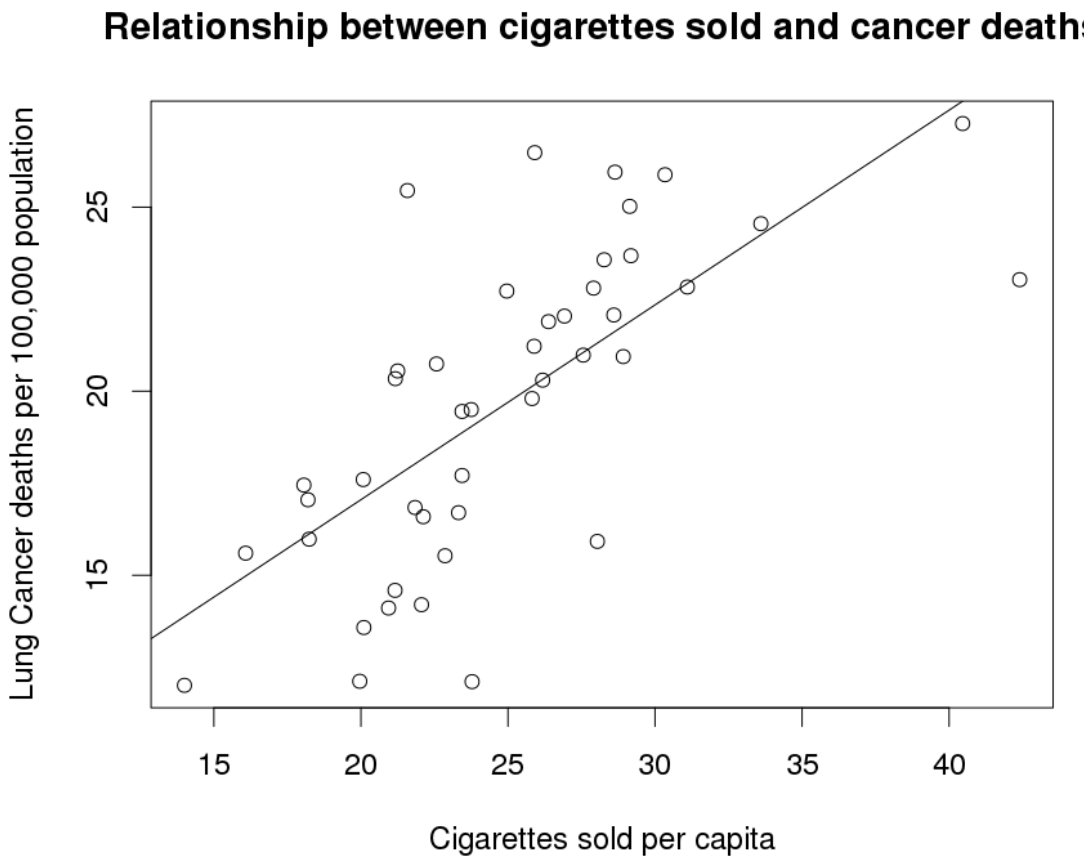




Center: Mean $\bar{x}$, median

Spread: Standard deviation (s), IQR

Relationships between 2 quantitative variables



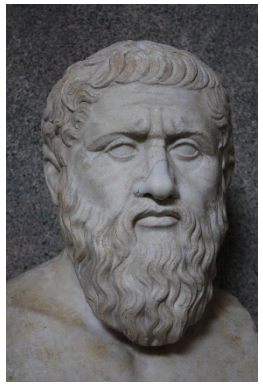
Correlation:

$$r = \frac{1}{(n-1)} \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{s_x} \right) \left(\frac{y_i - \overline{y}}{s_y} \right)$$

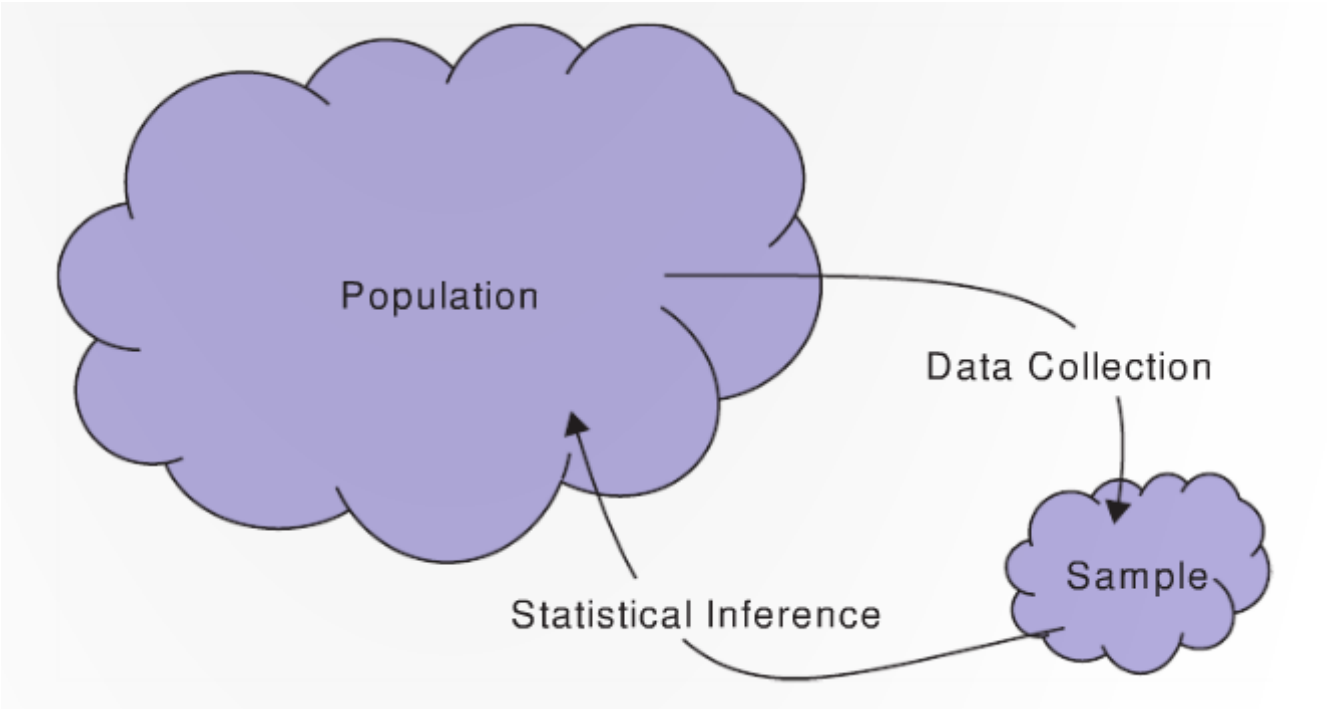
Regression:

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

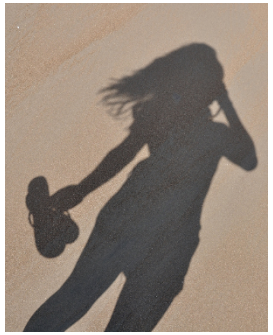
Statistical inference:
Confidence Intervals and Hypothesis Tests



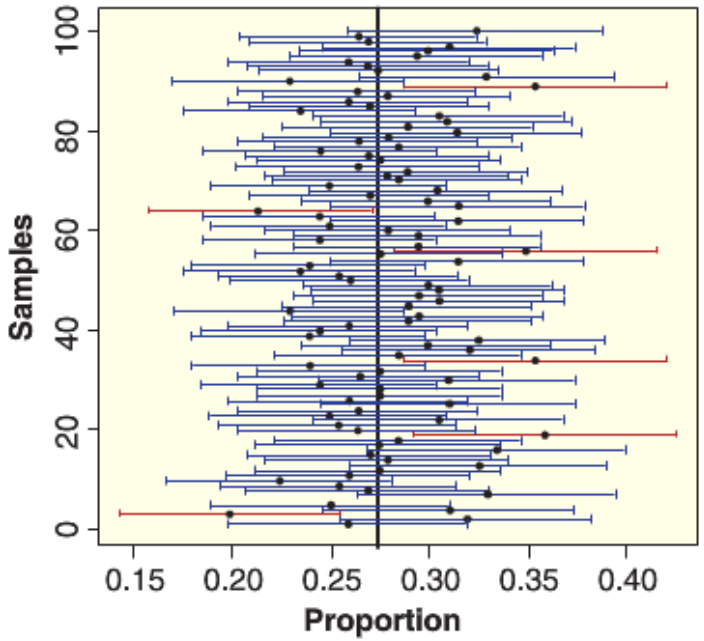
$\pi, \mu, \sigma, \rho, \beta$



$\hat{p}, \bar{x}, s, r, b$



Confidence Intervals

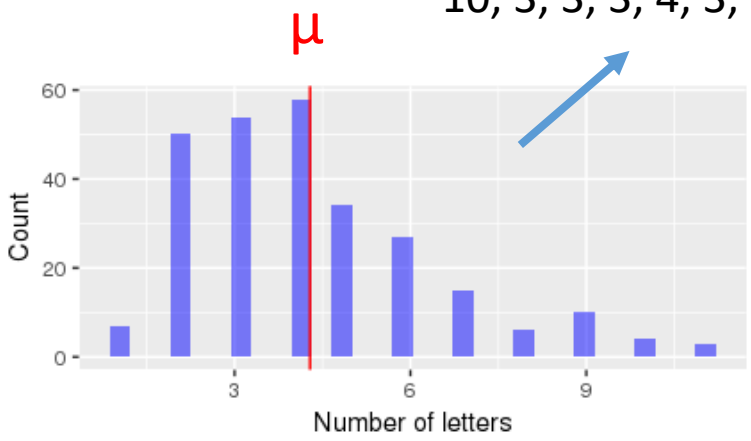


A **confidence interval** is an interval computed by a method that will contain the *parameter* a specified percent of times

The **confidence level** is the percent of all intervals that contain the parameter

Computational methods for CIs: The Bootstrap

The sample (n = 10)
10, 3, 3, 3, 4, 3, 2, 6, 4, 5



3, 3, 3, 5, 3,
4, 5, 2, 2, 10

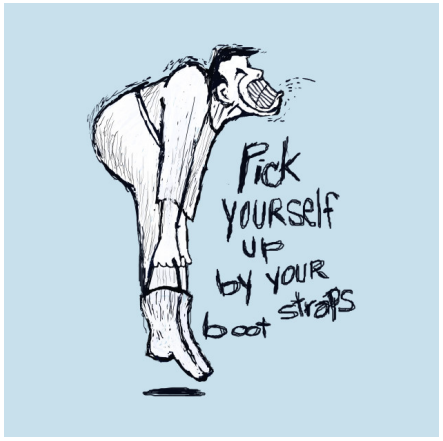
$\bar{x}^* = 4$

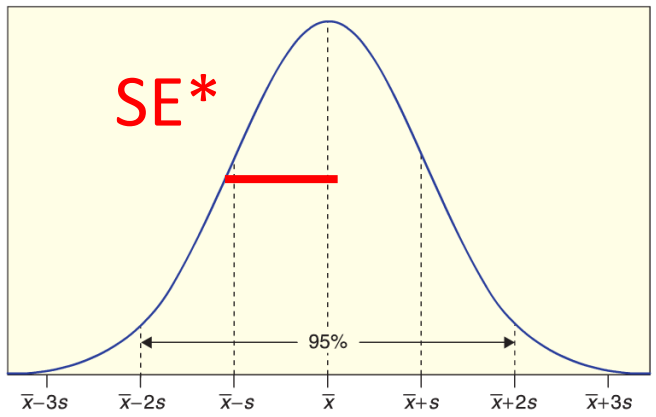
3, 3, 2, 3, 6,
4, 6, 5, 3, 6

$\bar{x}^* = 4.1$

5, 3, 2, 3, 3,
3, 10, 3, 4, 3

$\bar{x}^* = 3.9$





Bootstrap distribution!

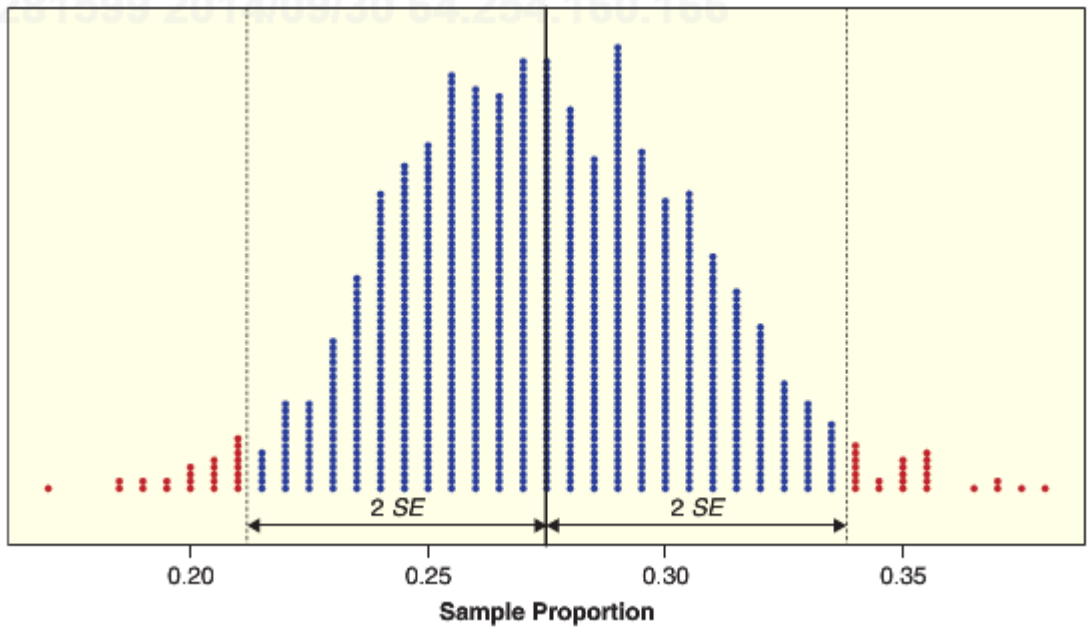
do_it(10000) * {...}

95% Confidence Intervals

When a bootstrap distribution for a sample statistic is approximately normal, we can estimate a 95% confidence interval using:

$$Statistic \pm 2 \cdot SE^*$$

Where SE* is the standard error estimated using the bootstrap

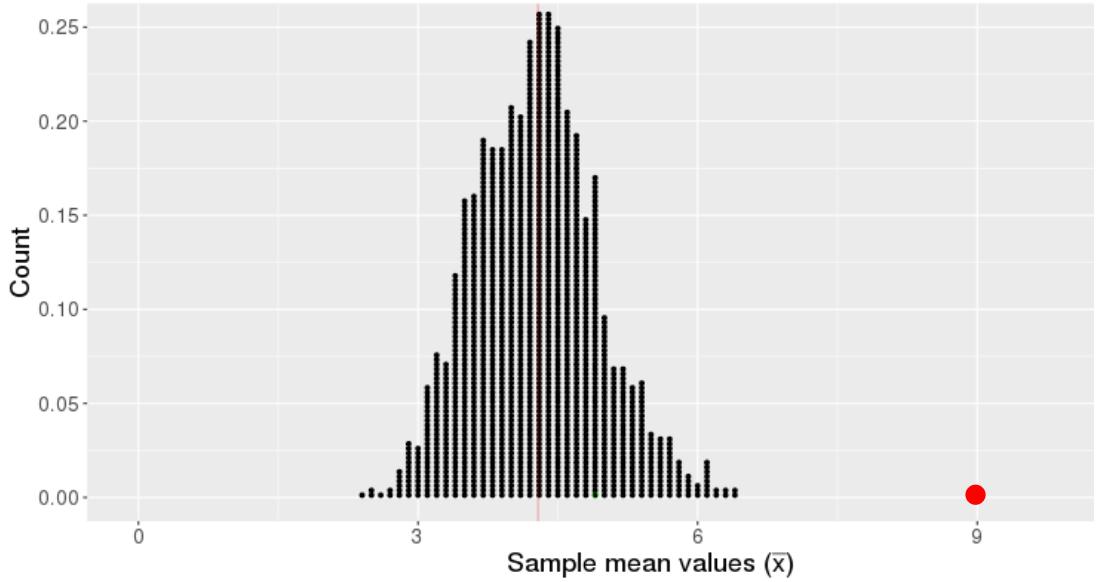


Hypothesis test logic

We start with a claim about a population parameter

- E.g., μ = 4

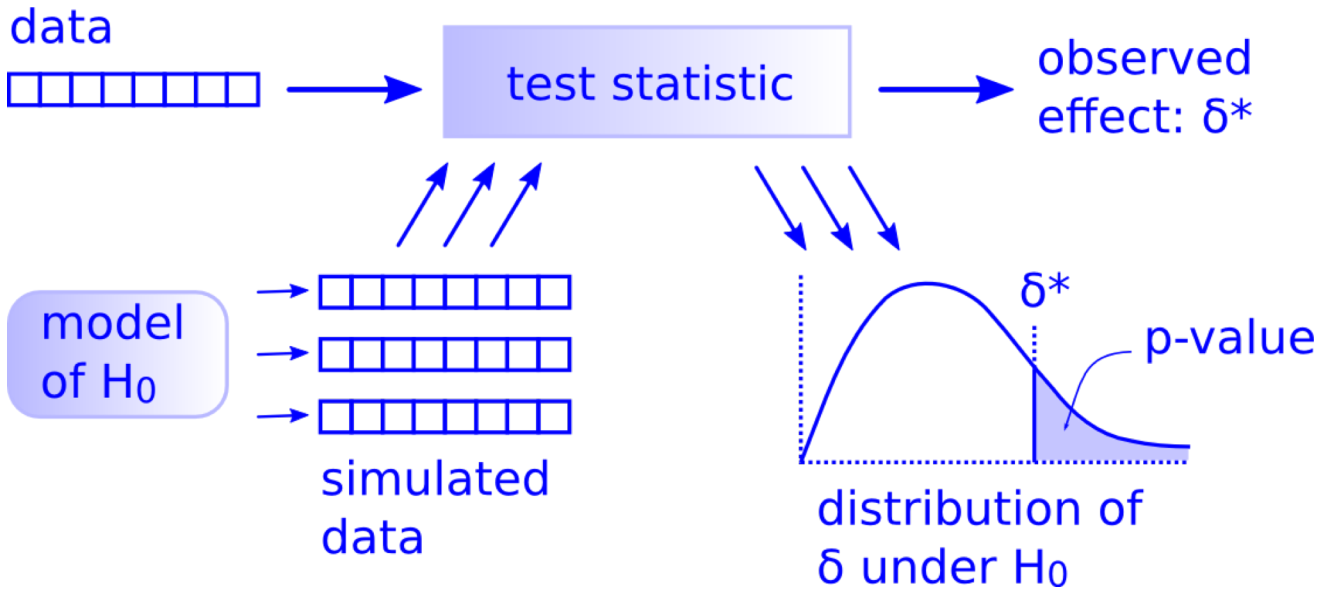
This claim implies we should get a certain distribution of statistics

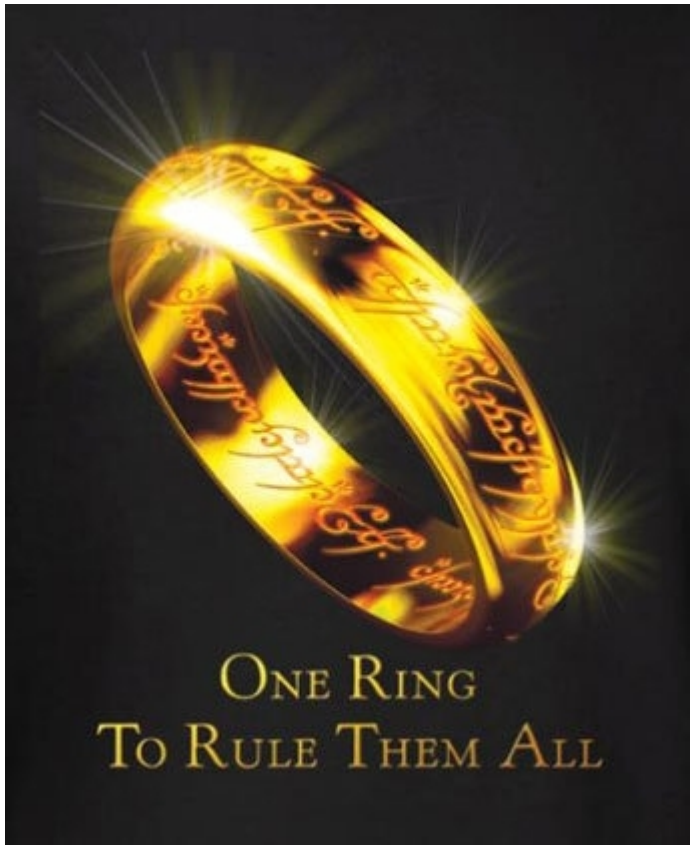


If our observed statistic is highly unlikely, we reject the claim

One test to rule them all

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





Just follow the 5 hypothesis tests steps!

Computational methods for hypothesis tests

Type of parameter	Simulation method
A single proportion π	
Comparing 2 or more means μ_1, μ_2, μ_k	
Correlation and regression ρ, β	

do_it(10000) * {…}

Computational methods for hypothesis tests

Type of parameter

- A single proportion π
- Comparing 2 or more means μ_1, μ_2, μ_k
- Correlation and regression ρ, β

Simulation method

- Coin flipping
- Combine and reassign
- Shuffle one of the columns

do_it(10000) * {…}

Computational methods for hypothesis tests

Type of parameter

A single proportion π

Comparing 2 or more means μ_1, μ_2, μ_k

Correlation and regression ρ, β

Simulation method

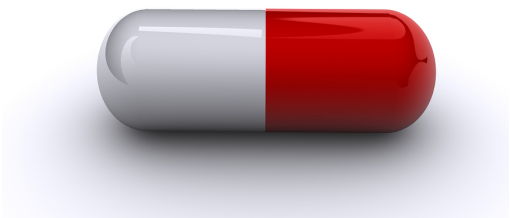
Coin flipping

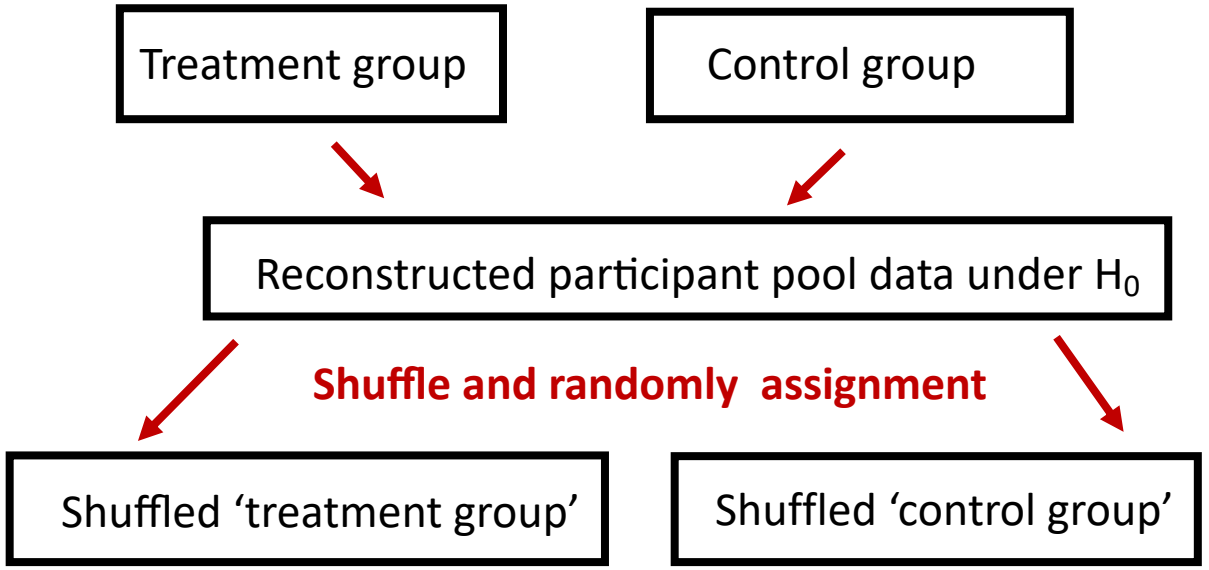
Combine and reassign

Shuffle one of the columns

do_it(10000) * {…}

	5	3	2		7			8
6		1	5					2
2			9	1	3		5	
7	1	4	6	9	2			
	2						6	
			4	5	1	2	9	7
	6		3	2	5			9
1					6	3		4
8			1		9	6	7	





Computational methods for hypothesis tests

Type of parameter

A single proportion π

Comparing 2 or more means μ_1, μ_2, μ_k

Correlation and regression ρ, β

Simulation method

Coin flipping

Combine and reassign

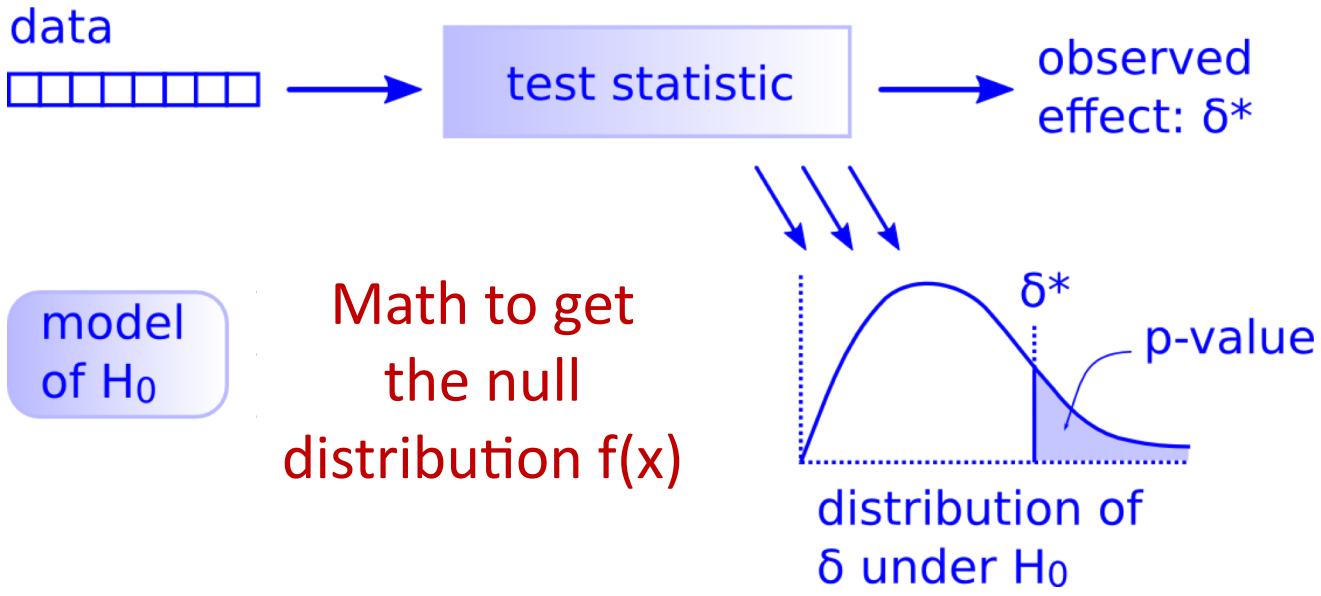
Shuffle one of the columns

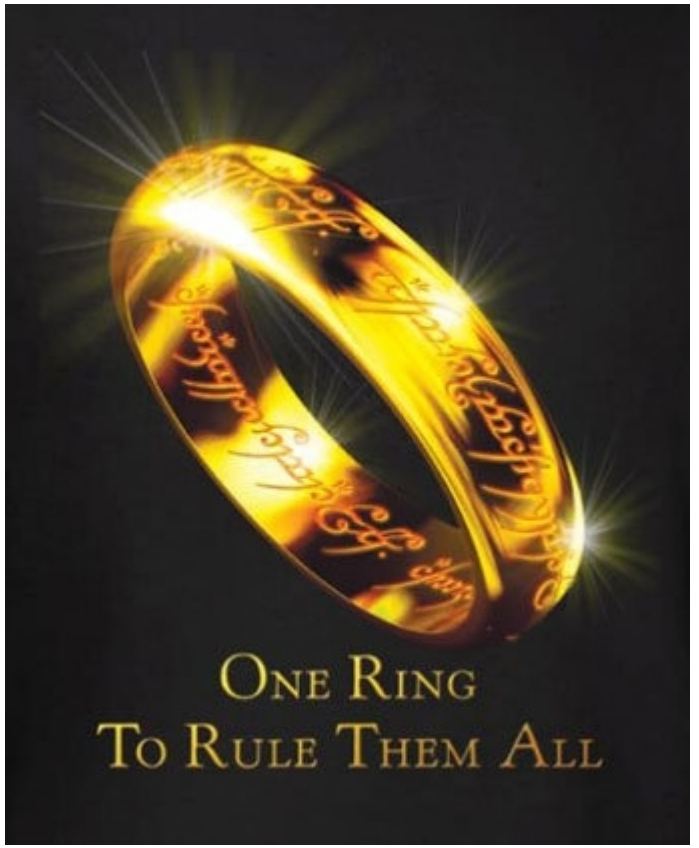
do_it(10000) * {…}

	Calories	Carbohydrates
AppleJacks	117	27
Boo Berry	118	27
Cap'n Crunch	144	31
Cinnamon Toast Crunch	169	32

Parametric methods for hypothesis tests

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





Just follow the ~5 hypothesis tests steps!

Parametric hypothesis tests and CIs

Type of parameter	Null distribution
One or two proportions π_1, π_2	
One or two means, regression μ_1, μ_2, β	
More than 2 proportions π_1, π_2, π_3	
More than 2 means μ_1, μ_2, μ_k	

Parametric hypothesis tests and CIs

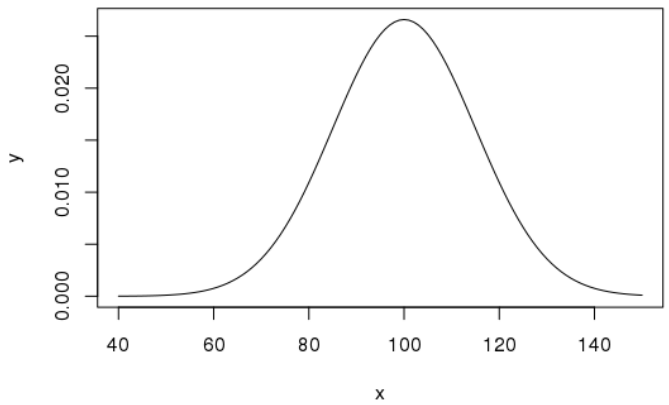
Type of parameter

- One or two proportions π_1, π_2
- One or two means, regression μ_1, μ_2, β
- More than 2 proportions π_1, π_2, π_3
- More than 2 means μ_1, μ_2, μ_k

Null distribution

- Normal distribution (z-test)
- t-distribution (t-test)
- χ^2-distribution (χ^2-test)
- F-distribution (ANOVA)

N(μ, σ)



$$\hat{p} \pm z^* \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$z = \frac{\hat{p} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$$

Parametric hypothesis tests and CIs

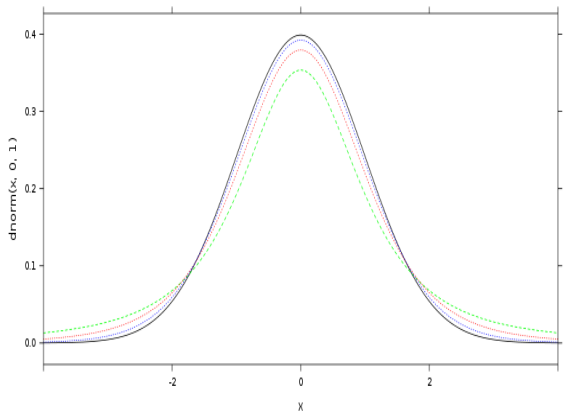
Type of parameter

- One or two proportions π_1, π_2
- One or two means, regression μ_1, μ_2, β
- More than 2 proportions π_1, π_2, π_3
- More than 2 means μ_1, μ_2, μ_k

Null distribution

- Normal distribution (z-test)
- t-distribution (t-test)
- χ^2-distribution (χ^2-test)
- F-distribution (ANOVA)

t(df)



$$(\overline{x_1} - \overline{x_2}) \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$t = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

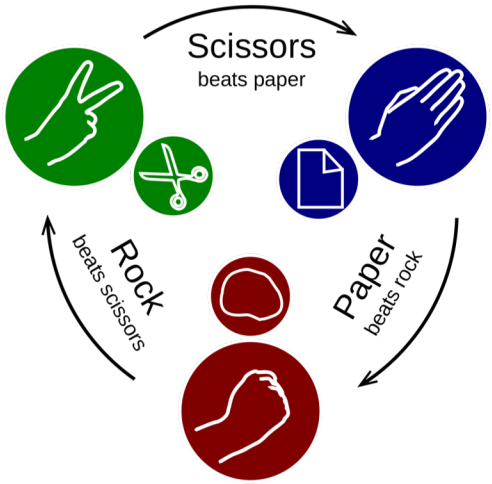
Parametric hypothesis tests and CIs

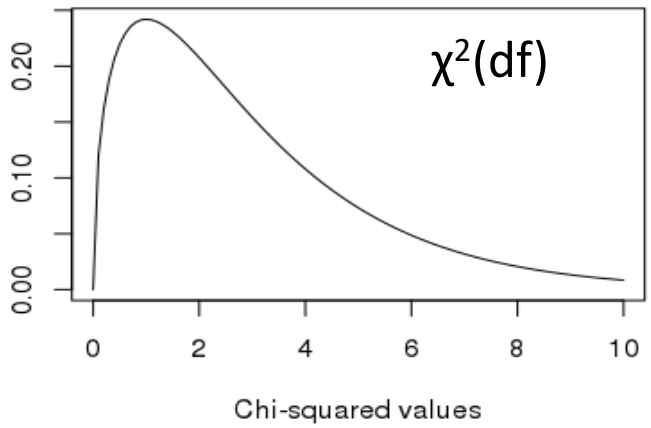
Type of parameter

- One or two proportions π_1, π_2
- One or two means, regression μ_1, μ_2, β
- More than 2 proportions π_1, π_2, π_3
- More than 2 means μ_1, μ_2, μ_k

Null distribution

- Normal distribution (z-test)
- t-distribution (t-test)
- χ^2-distribution (χ^2-test)
- F-distribution (ANOVA)





$$\chi^2 = \sum_{i=1}^{n} \frac{(Observed_i - Expected_i)^2}{Expected_i}$$

Parametric hypothesis tests and CIs

Type of parameter

One or two proportions π_1, π_2

One or two means, regression μ_1, μ_2, β

More than 2 proportions π_1, π_2, π_3

More than 2 means μ_1, μ_2, μ_k

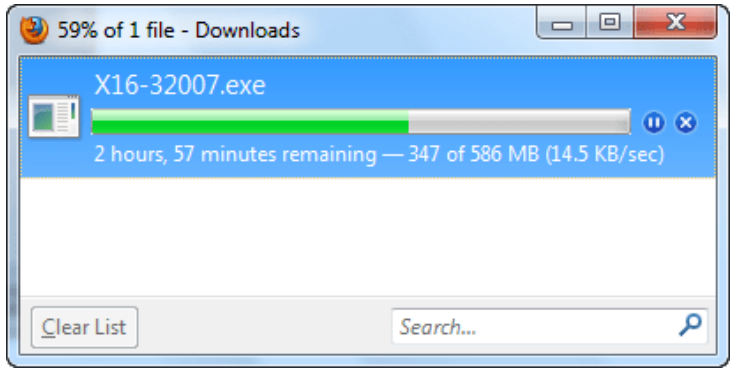
Null distribution

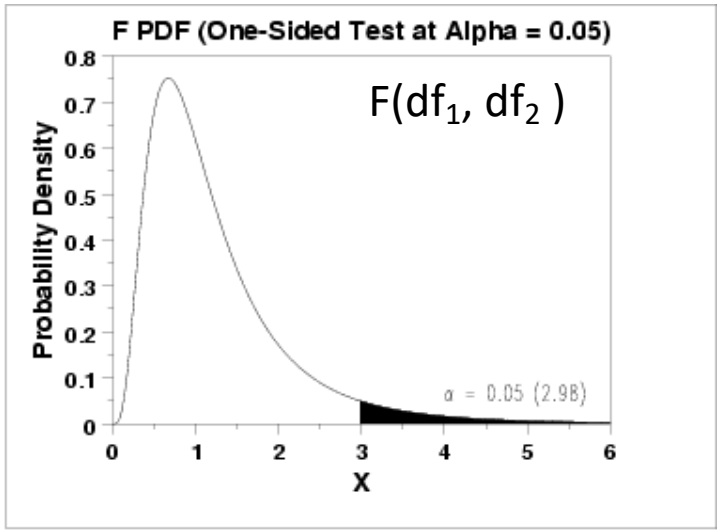
Normal distribution (z-test)

t-distribution (t-test)

χ^2-distribution (χ^2-test)

F-distribution (ANOVA)

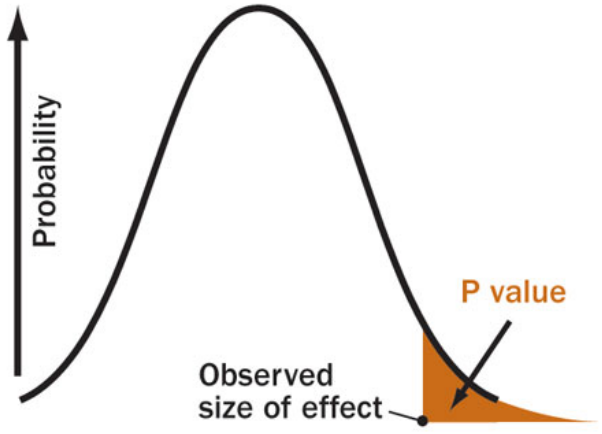




$$F = \frac{\text{between-group variability}}{\text{within-group variability}}$$

$$= \frac{\frac{1}{K-1}\sum_{i=1}^{K} n_i(\bar{x}_i - \bar{x}_{tot})^2}{\frac{1}{N-K}\sum_{i=1}^{K}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2}$$

Two theories of hypothesis testing



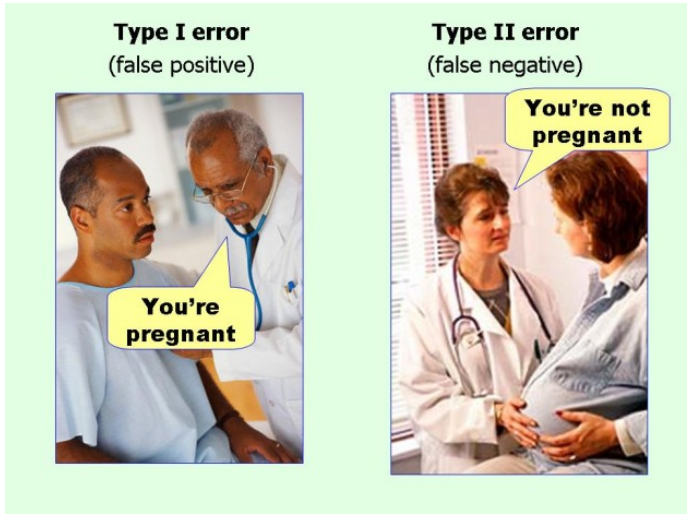
Significance testing

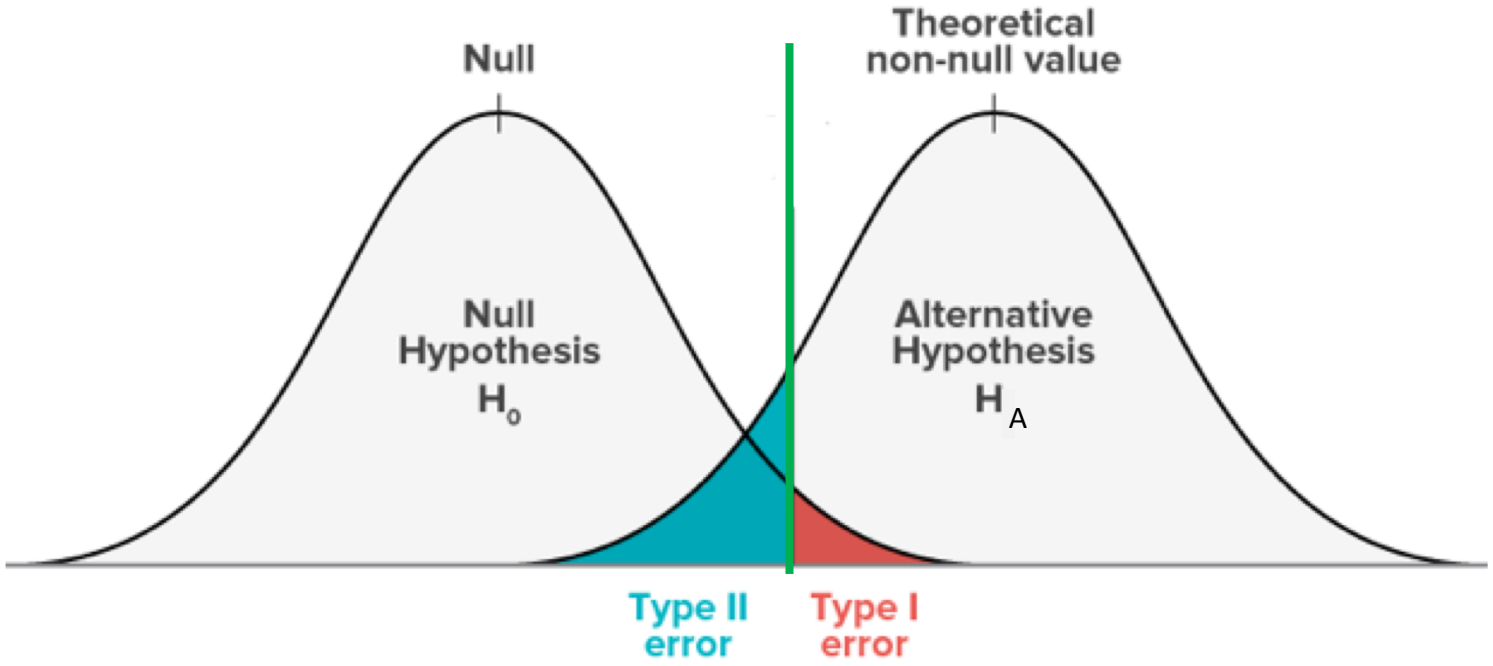
- Fisher
- P-value as strength of evidence

Hypothesis testing

- Neyman and Pearson
- Make a formal decision to reject or not reject (p-value < α)

Neyman-Pearson Type I and Type II Errors



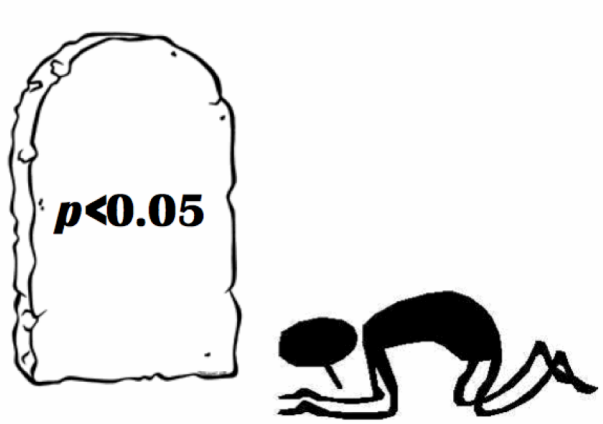


	Reject H_0	**Do not reject H_0**
H_0 is true	Type I error (α) (false positive)	No error
H_A is true (H_o is false)	No error	Type II error (β) (false negative)

Problems with the NP hypothesis tests

Problem 1: we are interested in the results of a specific experiment, not whether we are right most of the time

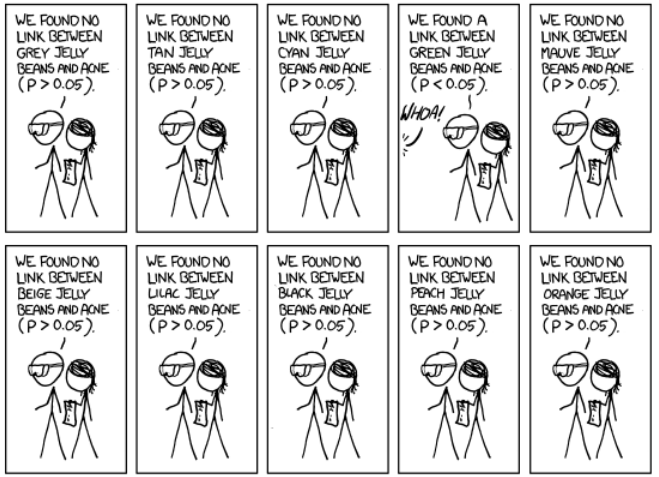
- E.g., 95% of these statements are true:
  - Calcium is good for your heart, Paul is psychic, Buzz and Doris can communicate, …



Problem 2: Arbitrary thresholds for alpha levels

- P-value = 0.051, we don't reject H_0?

Problem 3: running many tests can give rise to a high number of type 1 errors



Next steps in Statistics

Probability and Statistical theory (S&DS 238, 240, 241, 242)

- Parametric probability models and theory

Data Science (S&DS 123, 230, 262)

- Learn more advanced ways to visualize and manipulate data in R and Python

Linear models class (S&DS 312; Stats2 at other schools)

- Multiple regression
- Learn more advanced forms of ANOVA (multi-way/repeated measures)

Machine Learning (S&DS 355, 363, 365)

- Algorithms for making predictions

Many more advanced classes!

One last question…

What was the worst joke of the semester?

That's all folks!