Multiple regression and conclusions

Overview

Review and continuation of inference for linear regression

Quick discussion of multiple regression

How to determine which CI/hypothesis test to use

Conclusions

Announcement

Final exam review session

- Tuesday December 9th from 1-2:15pm
- In this classroom (Marsh)

Final exam:

- Dec 15th (Monday) at 2pm
- Location…

Final project due Sunday Dec 7th

Review and continuation of inference in regression using simulation methods

Review of regression (class 6 and 7)

Regression is method of using one variable ***x*** to predict the value of a second variable ***y***

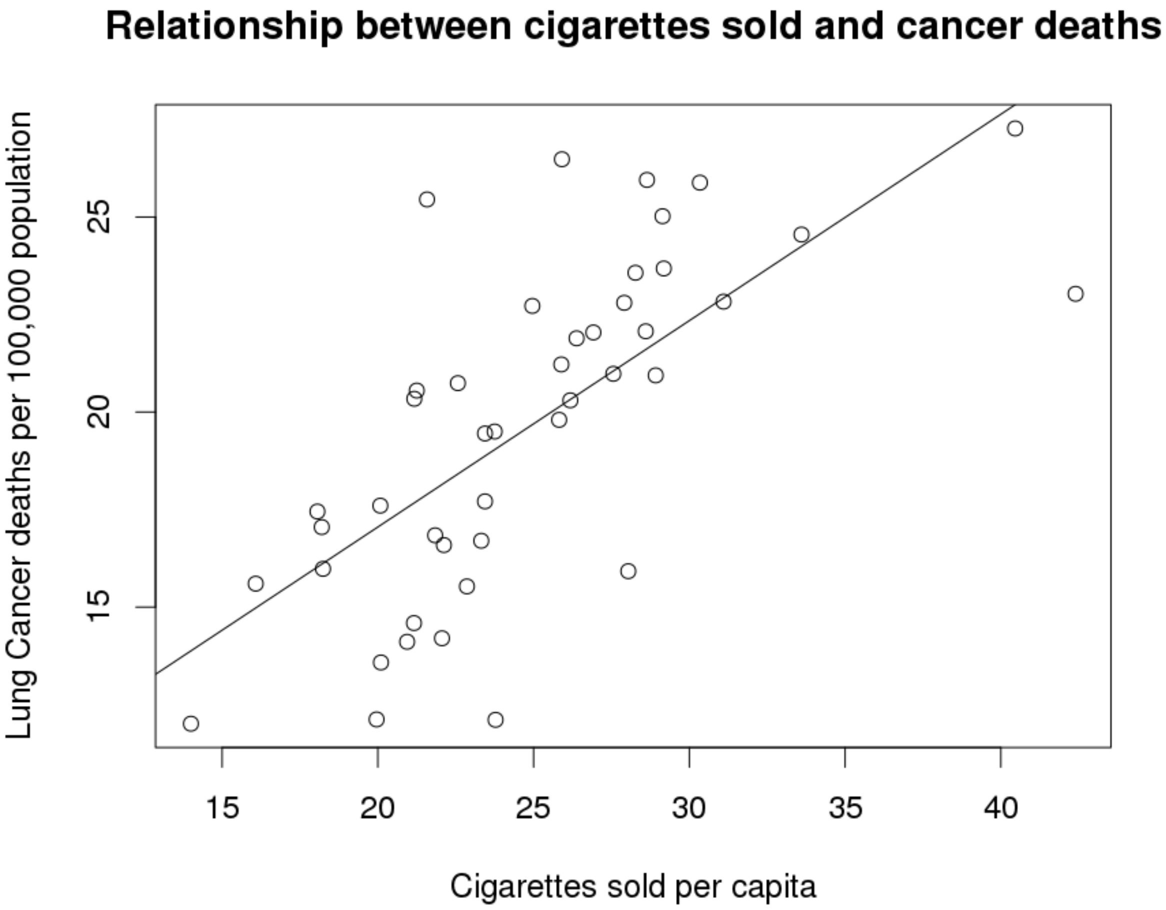
- i.e., $\hat{y} = f(x)$

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

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

$$Response = a + b \cdot Explanatory$$

Review cancer smoking regression line



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

```
R: my_fit <- lm(y ~ x)
   coef(my_fit)
```

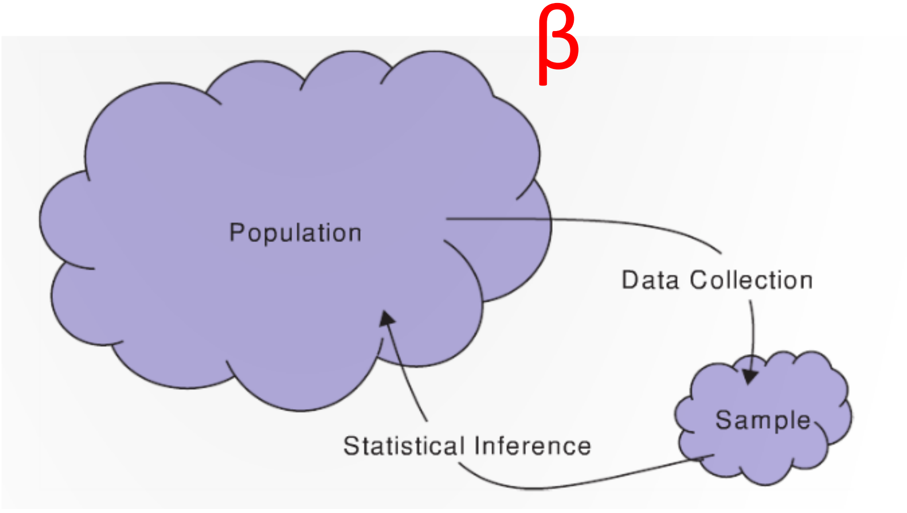
a = 6.47   b = 0.53

$\hat{y} = 6.47 + .53 \cdot x$

Review regression notation

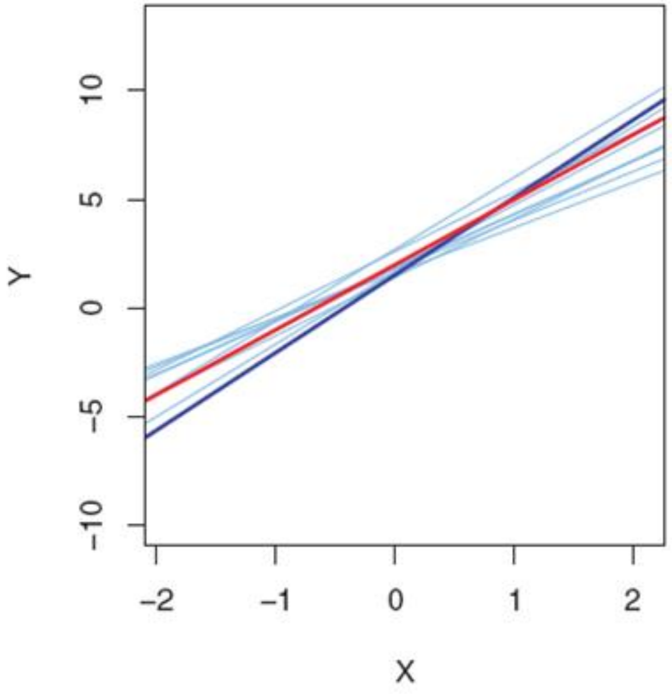
The Greek letter β is used to denote the slope of the **population**

The letter **b** is typically used to denote the slope of the **sample**



β

b



Using the bootstrap to create confidence intervals

We could use the bootstrap to create confidence intervals by:

1. Creating a bootstrap sample by sampling with replacement from our *paired data*
   - SDS1000: resample_pairs(v1, v2)
2. Fitting a regression line to our bootstrap sample and extracting the slope b
3. Repeat 10,000 times to get a bootstrap distribution of b's
4. Taking the standard deviation of the bootstrap distribution to get SE*
5. Using our confidence interval formula:

$$\textit{Statistic} \pm 1.96 \cdot SE^*$$

State	Cig per capita	Lung
AL	18.2	17.05
AZ	25.82	19.8
AR	18.24	15.98
CA	28.6	22.07
CT	31.1	22.83
DE	33.6	24.55
DC	40.46	27.27

Using permutation hypothesis tests

If we wanted to run a hypothesis tests for the regression slope, how would we write the null and alternative hypotheses using symbols?

H_0: $\beta = 0$

H_A: $\beta \neq 0$

Any ideas how to run a permutation test to assess whether H_0: $\beta = 0$?

State	Cig per capita	Lung
AL	18.2	17.05
AZ	25.82	19.8
AR	18.24	15.98
CA	28.6	22.07
CT	31.1	22.83
DE	33.6	24.55
DC	40.46	27.27

Using permutation hypothesis tests

We could use run a permutation test for H_0: $\beta = 0$ by creating a null distribution using:

1. Shuffle one of the columns of data
2. Fitting a regression line to our bootstrap sample and extracting the slope b
3. Repeat 10,000 times to get a null distribution of b's

We can obtain a p-value by seeing how many points in the null distribution are greater than the observed statistic value of b

State	Cig per capita	Lung
AL	18.2	17.05
AZ	25.82	19.8
AR	18.24	15.98
CA	28.6	22.07
CT	31.1	22.83
DE	33.6	24.55
DC	40.46	27.27

Let's try it in R!

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

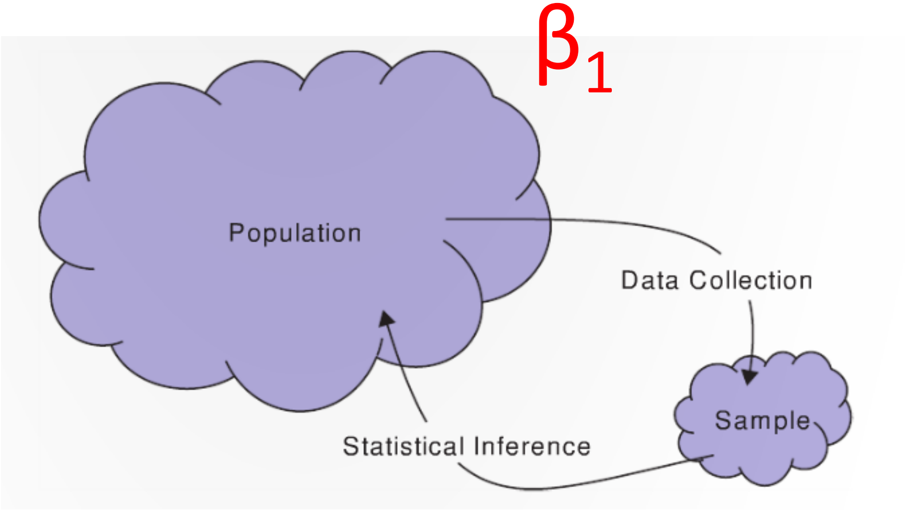
$$\textit{Predicted response} = b_0 + b_1 \cdot \textit{Explanatory}$$

Inference on simple linear regression

The Greek letter $\boldsymbol{\beta_1}$ is used to denote the slope of the population

The letter $\mathbf{b_1}$ is typically used to denote the slope of the sample

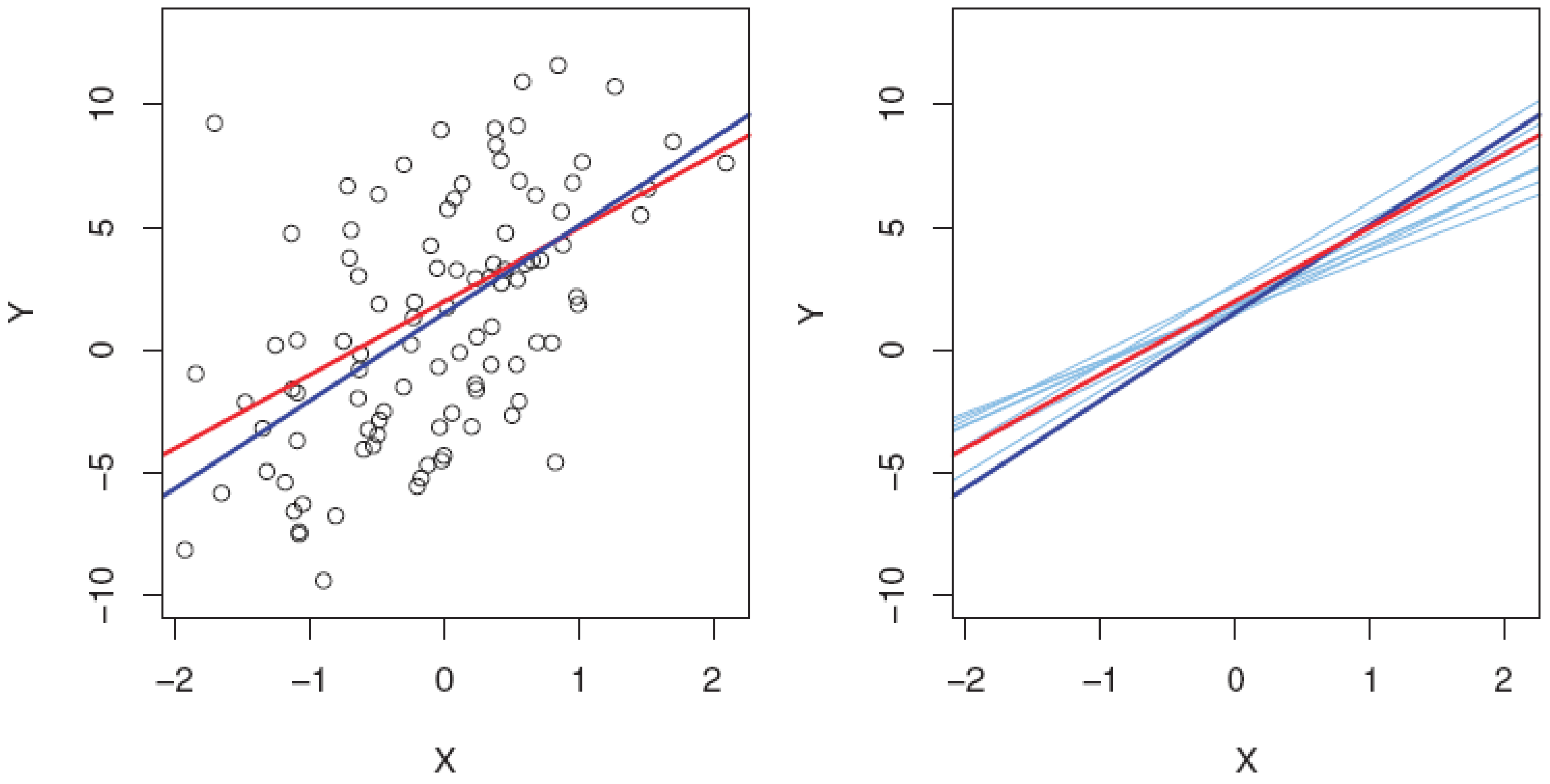
β_1



b_1

Population: β_1

Sample estimates: b_1

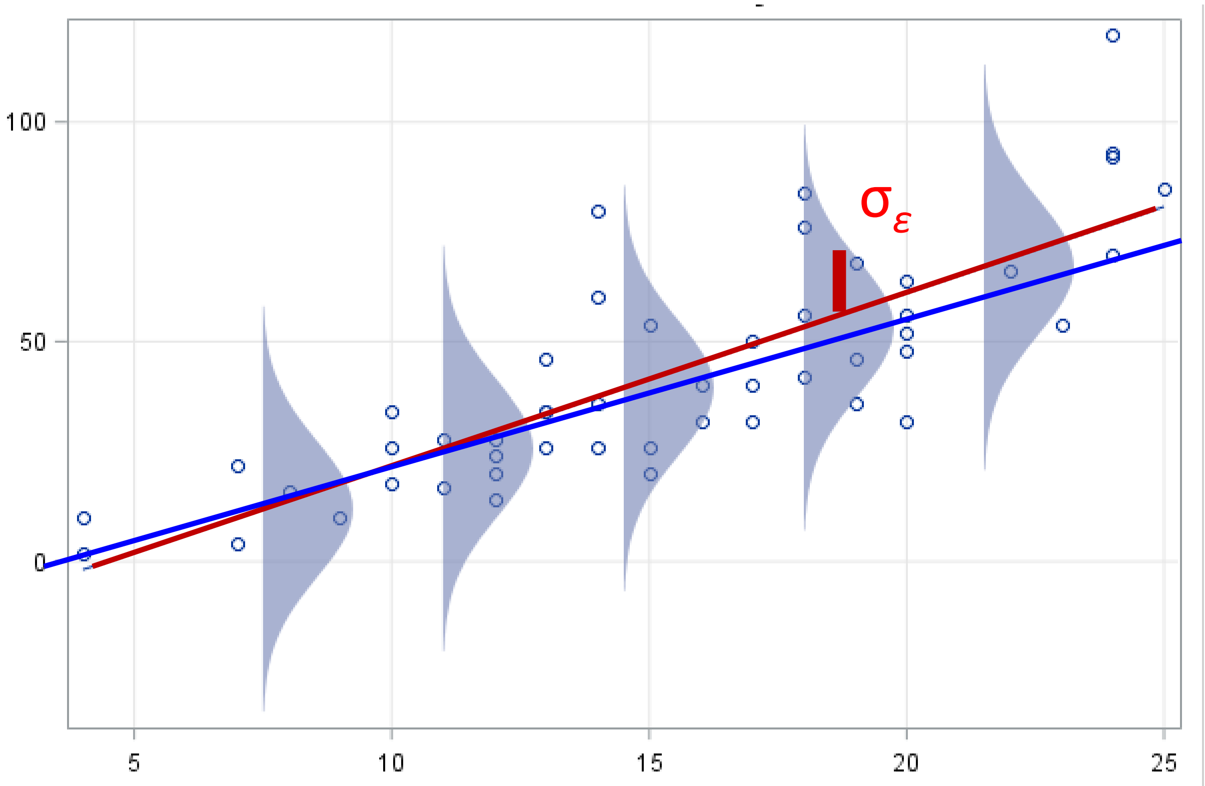


Simple linear regression underlying model

Intercept Slope } *Parameters*

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

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

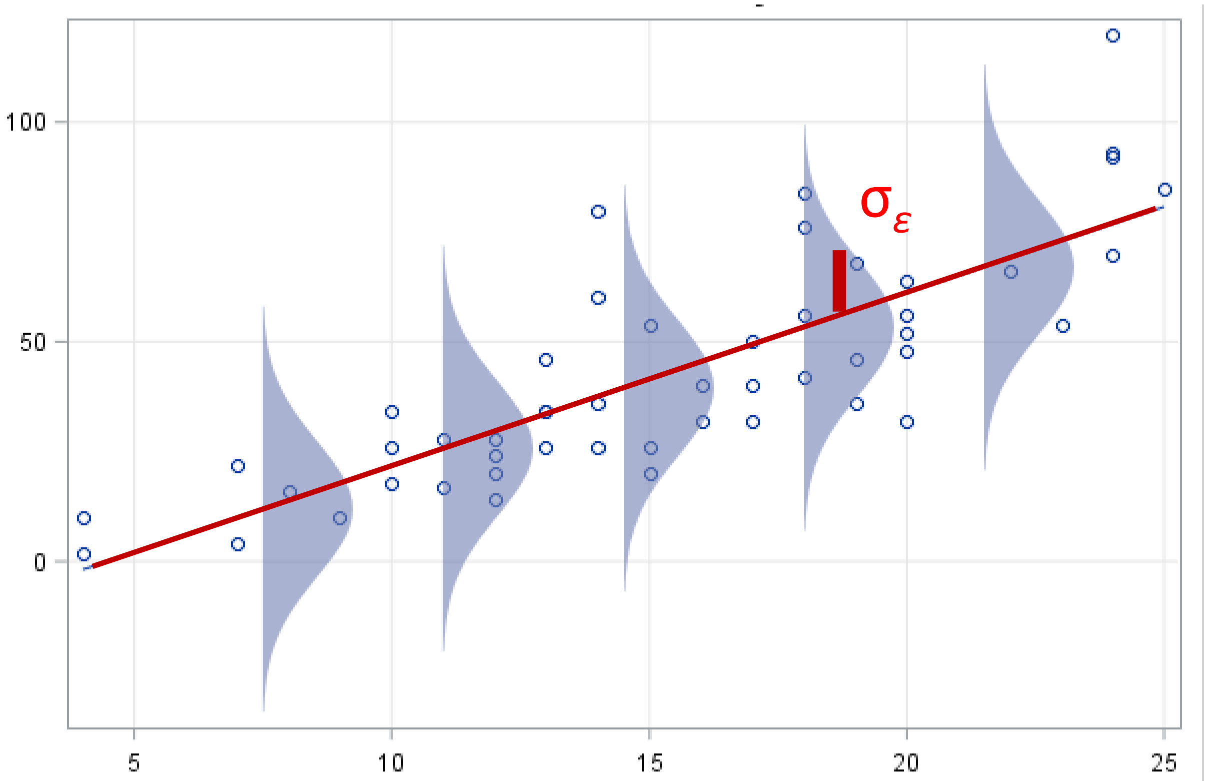


$$\epsilon \sim N(0, \sigma_\epsilon)$$

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

Estimating σ_ε

We can also use the **standard deviation of residuals σ_e** as an estimate standard deviation of irreducible noise **σ_ε**



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

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

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

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

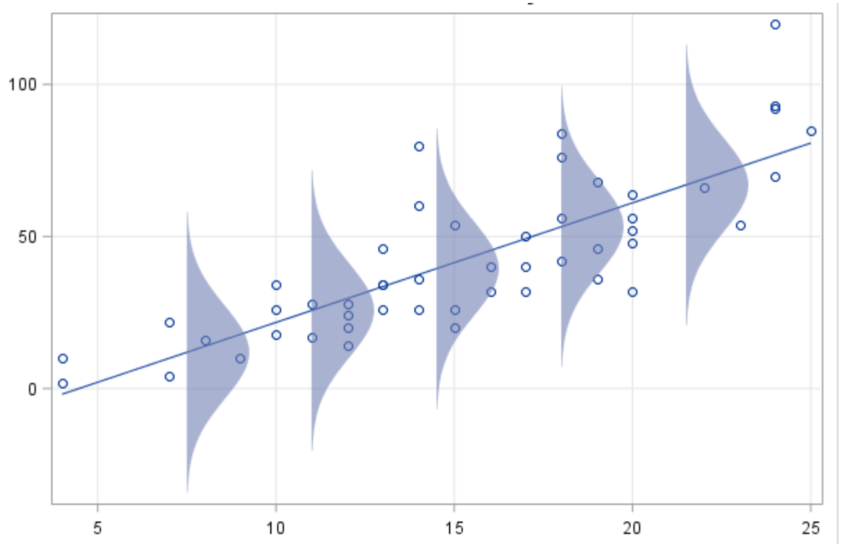
Inference using parametric methods

When using parametric methods, we make the following (LINE) assumptions:

- **Linearity**: A line can describe the relationship between x and y
- **Independence**: each data point is independent from the other points
- **Normality**: errors are normally distributed
- **Equal variance (homoscedasticity)**: constant variance of errors over the whole range of x values

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

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



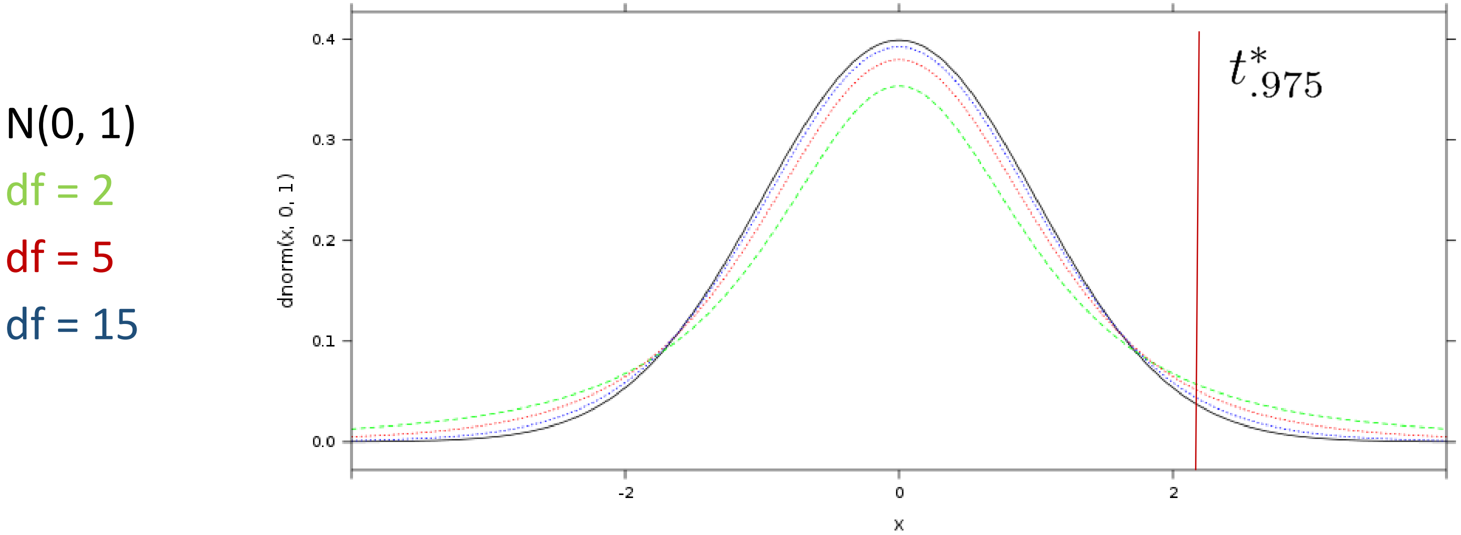
These assumptions are usually checked after the models are fit using ‘regression diagnostic’ plots.

Confidence intervals for regression coefficients

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

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

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



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 ŷ

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

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)

Dow
15,350
15,300
15,250
15,200
15,150
◀ TUE
WED ▶
9:30AM
11AM
12PM
1PM
2PM
3PM

Multiple regression

Let’s predict first-year college GPA based data from 219 students using the following variables:

- High school GPA   (HSGPA)
- Verbal SAT scores (SATV)
- Number of humanities credits (HU)

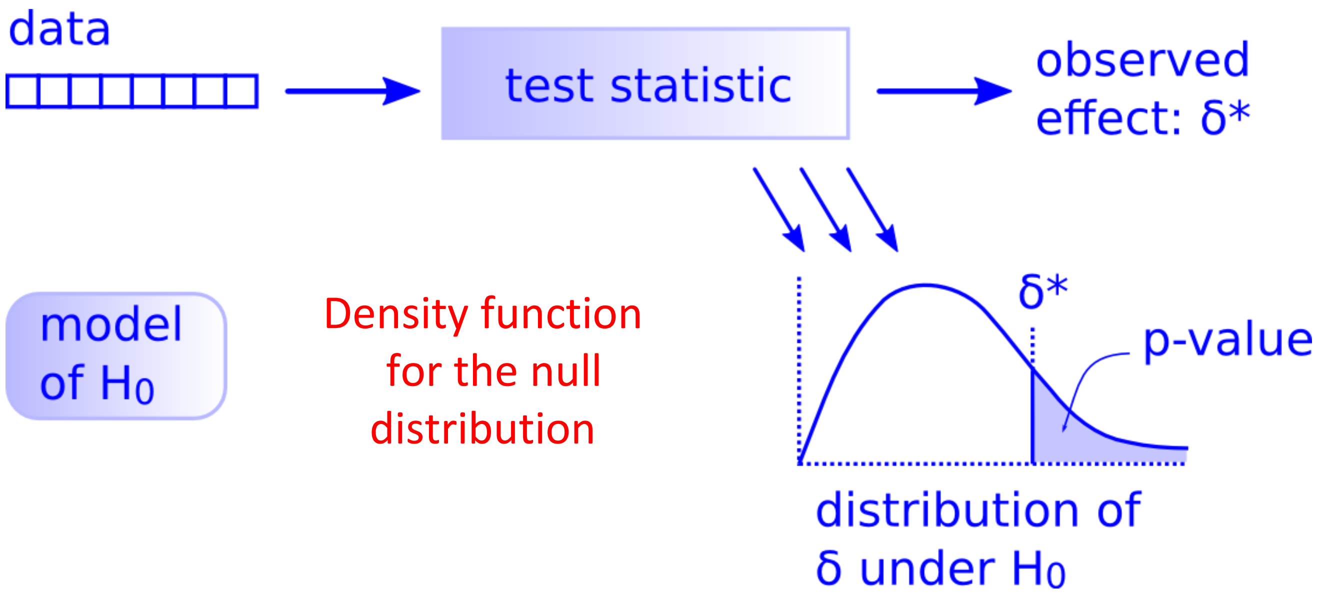
$$\hat{y}_{GPA} = b_0 + b_1 \cdot x_{HSGPA} + b_2 \cdot x_{SATV} + b_3 \cdot x_{HU}$$

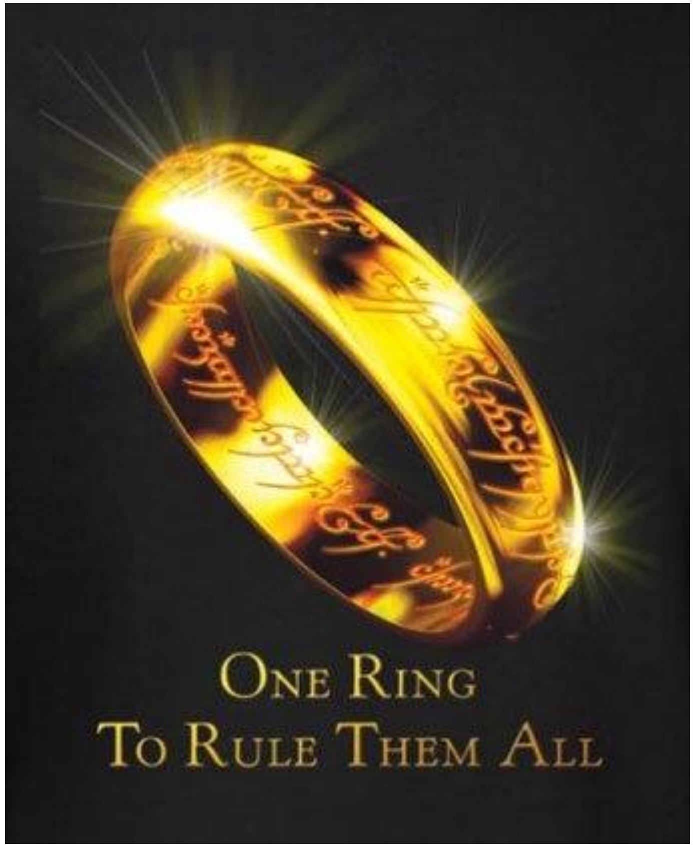
Let’s quickly try it in R!

Choosing the appropriate hypothesis test and confidence interval

One test to rule them all

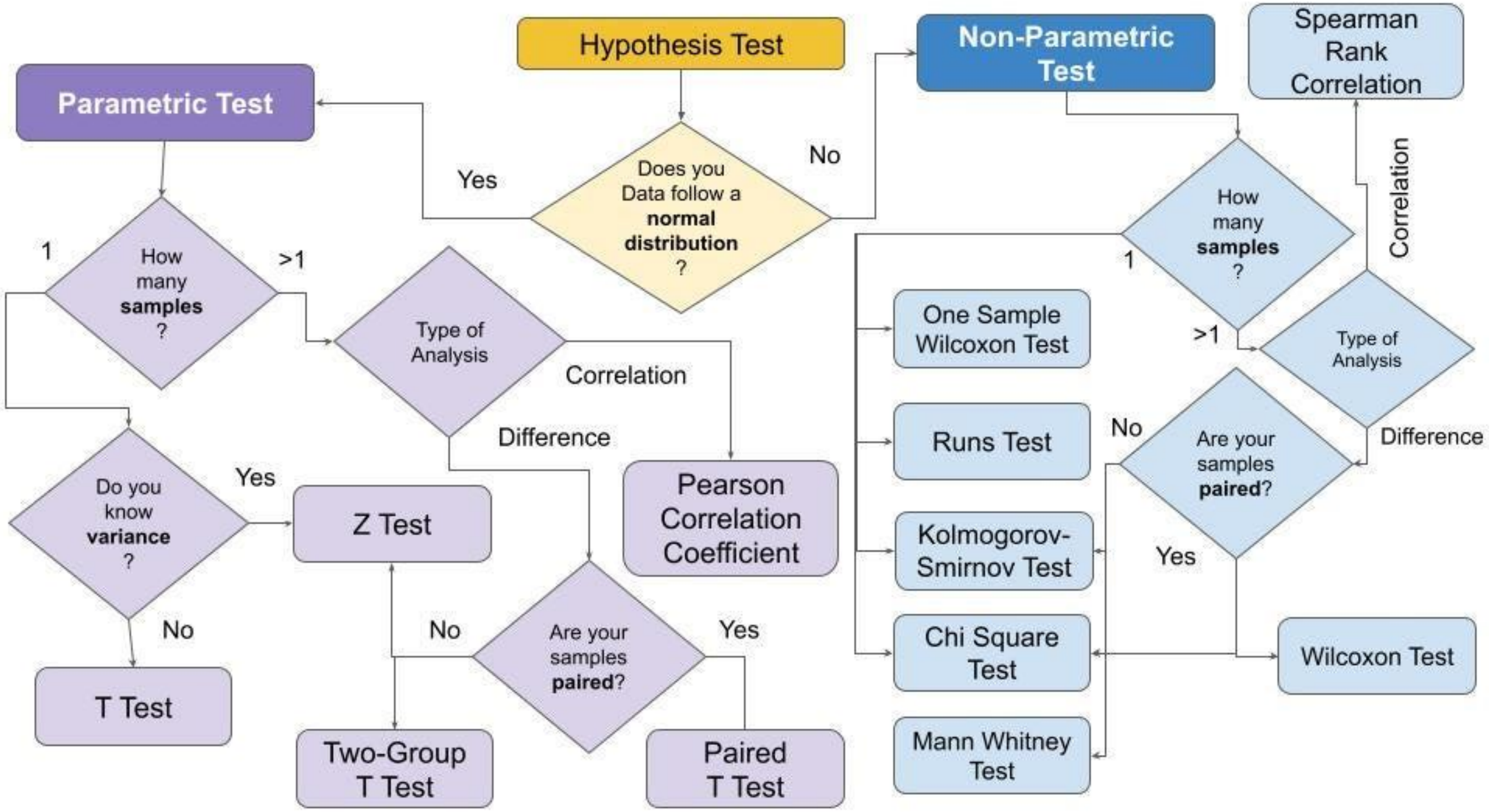
There is only one hypothesis test!


data
test statistic
observed effect: δ*
model of H0
Density function for the null distribution
δ*
p-value
distribution of δ under H0



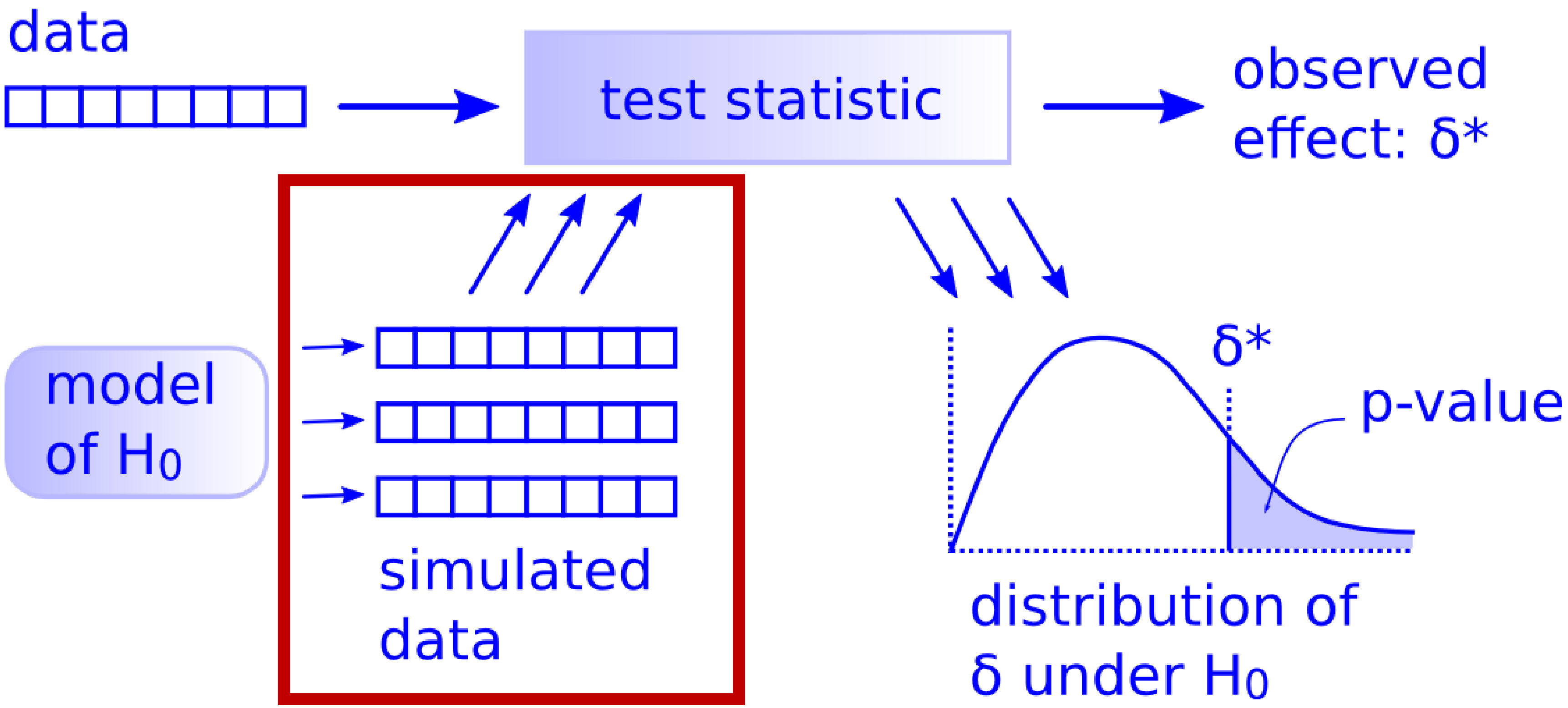
Just follow the 5 hypothesis tests steps!

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 $z = \frac{\hat{p_1} - \hat{p_2}}{\sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{n_1} + \frac{\hat{p_2}(1-\hat{p_2})}{n_2}}}$	$H_0: \pi_1 = p_1,\ \pi_2 = p_2,\ \dots\ ,\pi_k = p_k$ H_A: At least one p_i is different than specified chi-square test $\chi^2 = \sum_{i=1}^{k} \frac{(Observed_i - Expected_i)^2}{Expected_i}$
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	$H_0: \mu_1 = \mu_2 = \dots = \mu_k$ 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 -1, 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: $\pi_1 = p_1,\ \pi_2 = p_2,\ \ldots\ , \pi_k = p_k$ H_A: At least one p_i is different than specified Roll a k-sided die n times rmultinom(1, n, prob =)
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: $statistic \pm q^* \cdot SE$

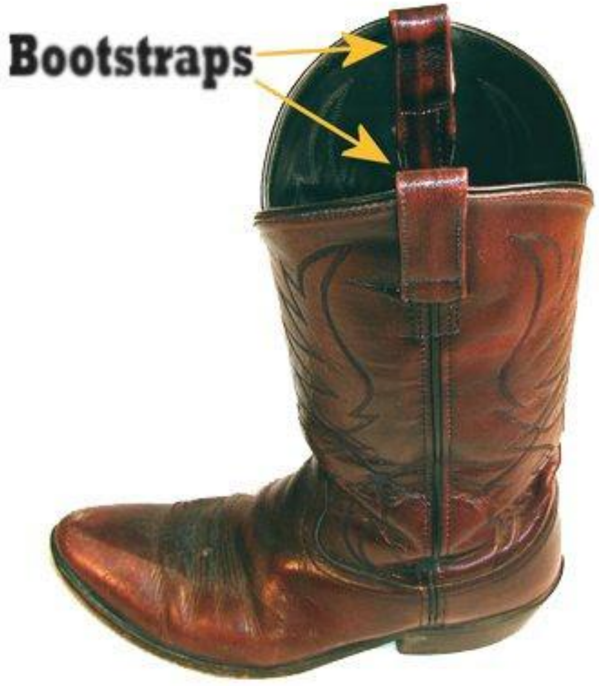
We just need the appropriate standard error (SE) formula

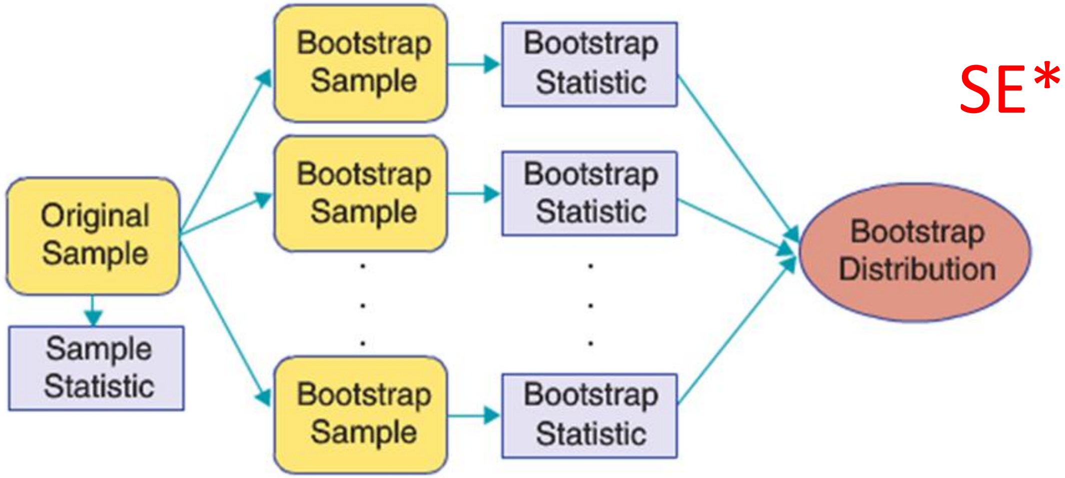
(and to determine if we should use t* or z*)

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!





SE*

$$\textit{Statistic} \pm z^* \cdot SE^*$$

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 makes fewer assumptions

- 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

Conclusions

THE TRUTH IS OUT THERE

Teaching Staff

Preceptors

- Lynda Aouar
- Addison McGhee

Course Manager

- Brian Xiang

Teaching Fellow

- Vladimir Averin

Undergraduate Learning Assistants

- Cindy Cai
- Alyssa Chang
- Jessica Huang
- Eric Lin
- Jasmine Garcia
- Asher Mehr
- Sarah Lepkowitz
- Lucas Papamitsakis
- Aryav Bothra

Good luck studying for the finals!