Session 3: Statistical inference using R

Overview

Statistical inference using parametric methods: Confidence intervals and hypothesis tests

Computational methods

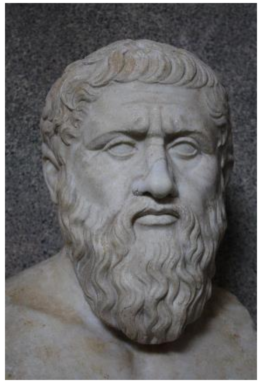
- Permutation tests (example correlation)

Simple linear regression

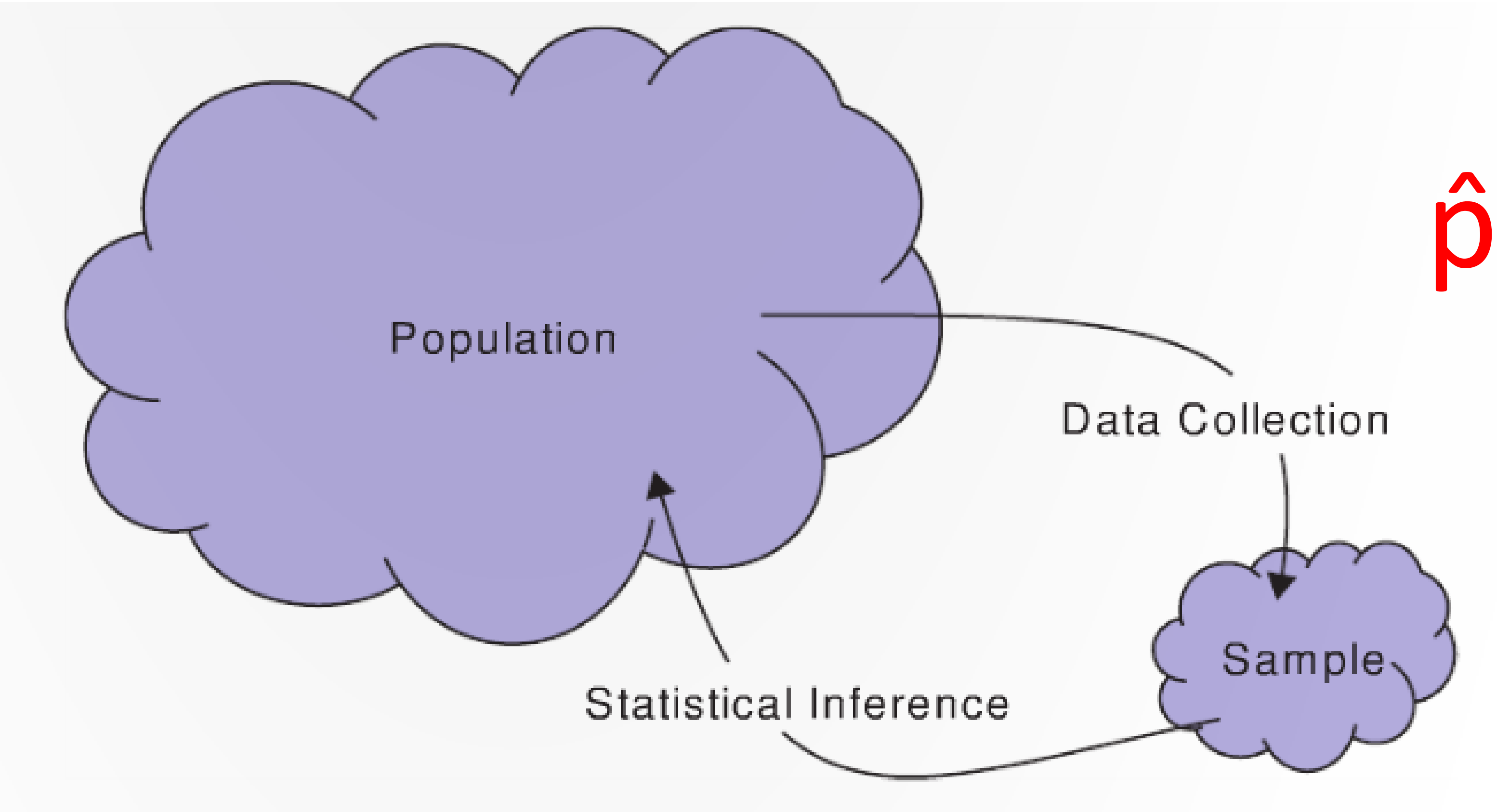
Multiple linear regression

What is statistical inference?

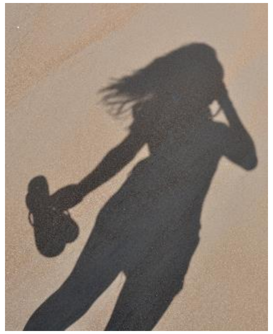
$\pi, \mu, \sigma, \rho, \beta$



Parameter: a number characterizing a property of a population



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



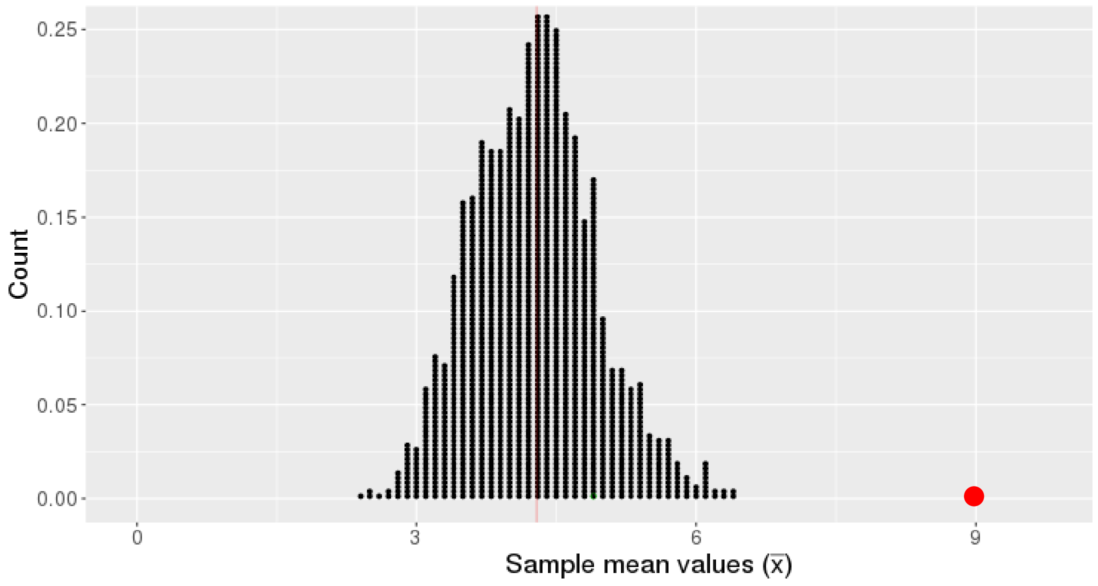
Statistic: A number computed from a sample

Basic hypothesis test logic

We start with a claim about a population parameter

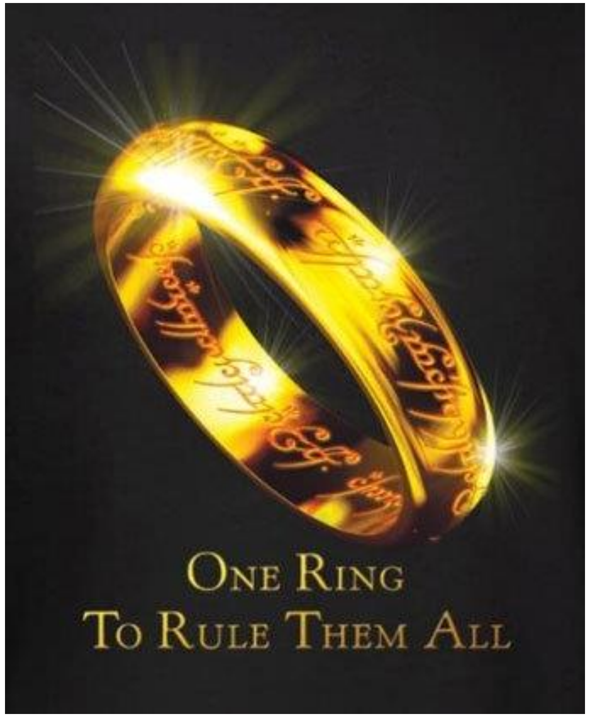
- E.g., μ = 4.2

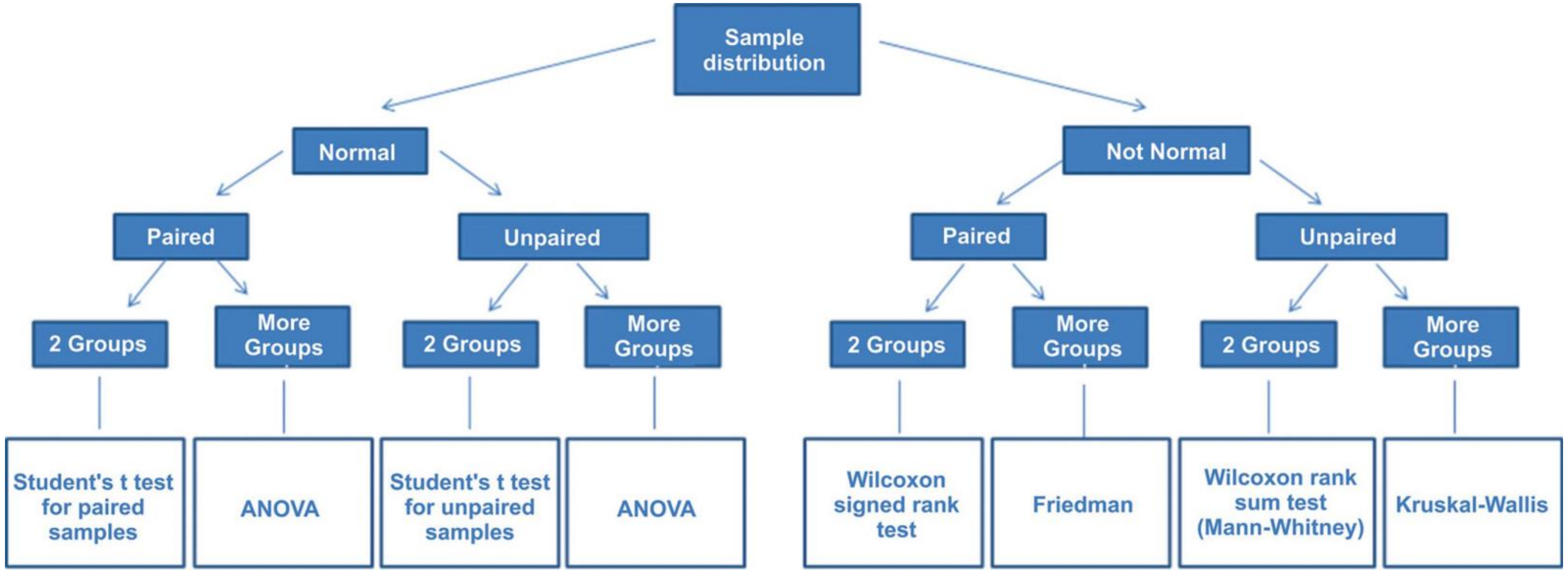
This claim implies we should get a certain distribution of statistics



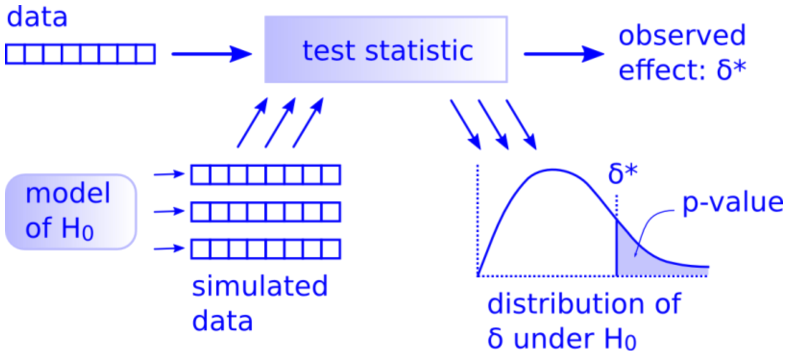
If our observed statistic is highly unlikely, we reject the claim

The big picture: There is only one hypothesis test!





Just need to follow 5 steps!



What are confidence intervals?

What are Confidence intervals?

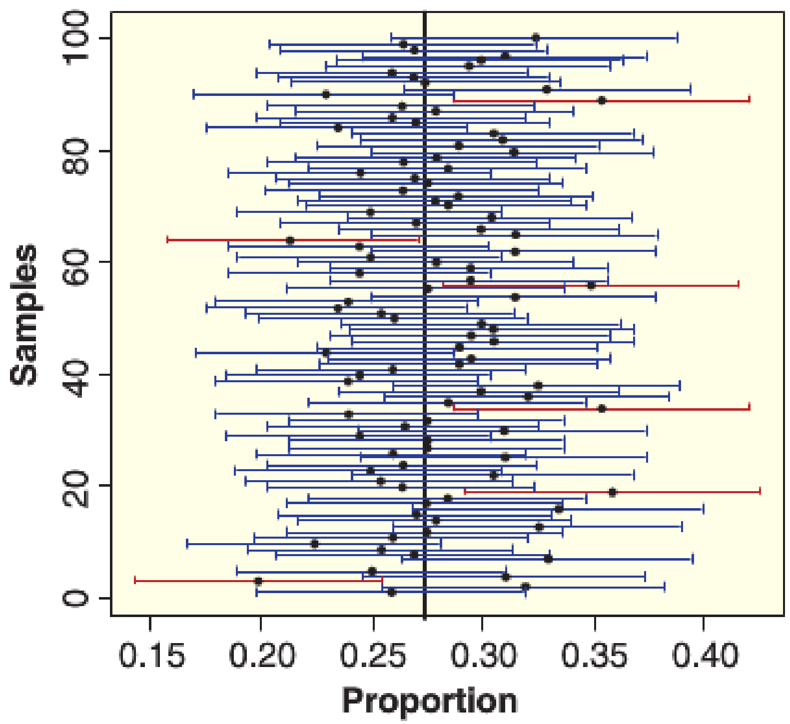
- Range of plausible values that capture the parameter a fixed % of the time

How can we create confidence intervals?

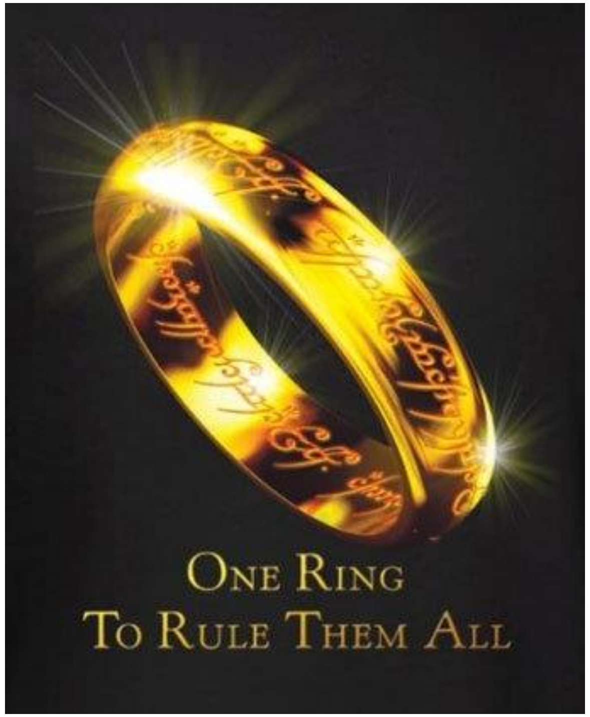
- Use the bootstrap (to estimate the SE)
- Use formulas (to estimate the SE)

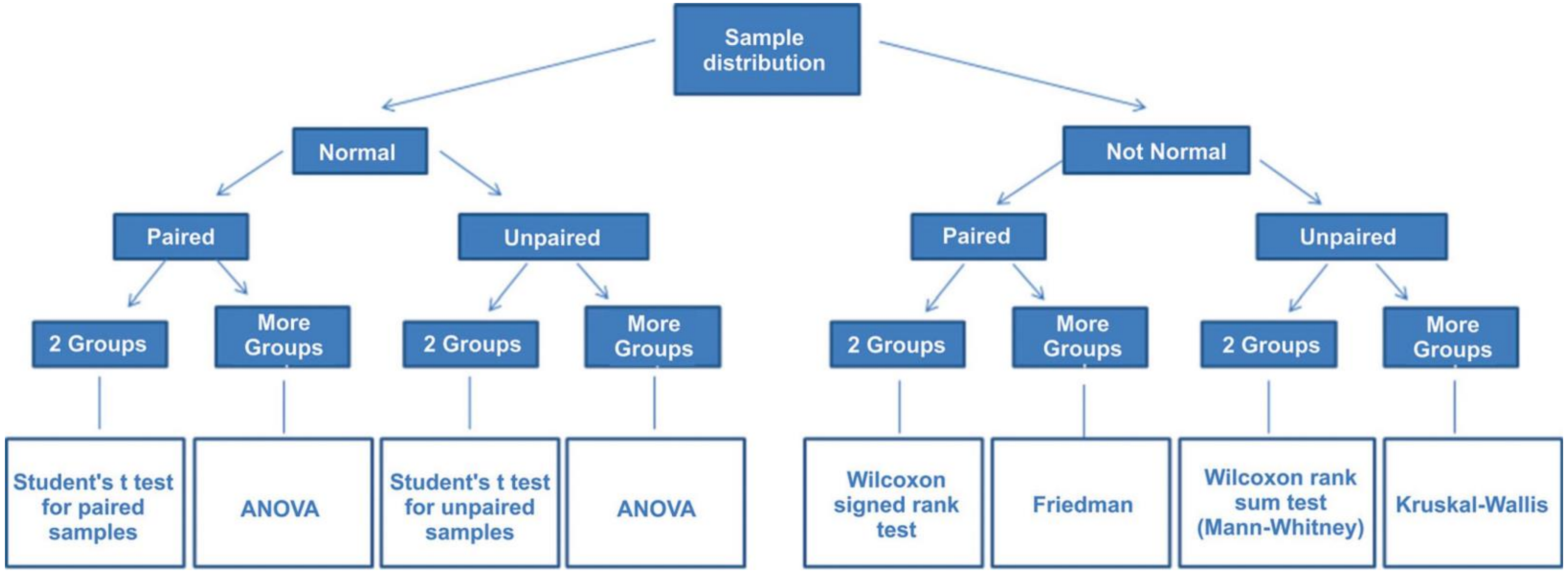
μ

μ

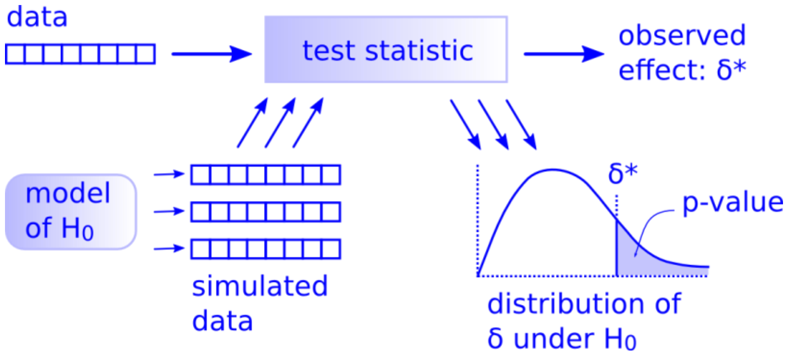


The big picture: There is only one hypothesis test!





Zoo of hypothesis tests…



Parametric methods

t-tests for a single mean

- H_0: $\mu = v$
- H_A: $\mu \quad = v$

t-tests for two means

- H_0: $\mu_{Treatment} = \mu_{Control}$ or $\mu_{Treatment} - \mu_{Control} = 0$
- H_A: $\mu_{Treatment} > \mu_{Control}$ or $\mu_{Treatment} - \mu_{Control} > 0$

We can run these tests using the function t.test()

Parametric methods

Hypothesis test for proportions

- H_0: $\pi = v$
- H_A: $\mu \;\; = v$

In R: prop.test()

Hypothesis test for correlation

- H_0: $\rho = 0$
- H_A: $\rho > 0$

In R: cor.test()

There are other hypothesis test functions

- chisq.test(), wilcox.test(), etc.

Simulation methods

Permutation/randomization tests are hypothesis tests that work by randomly shuffling the data

The bootstrap is a computational way to create confidence intervals

Advantages:

- They rely on fewer "assumptions" than parametric tests
- They can be applied to many more situations where parametric tests are unknown

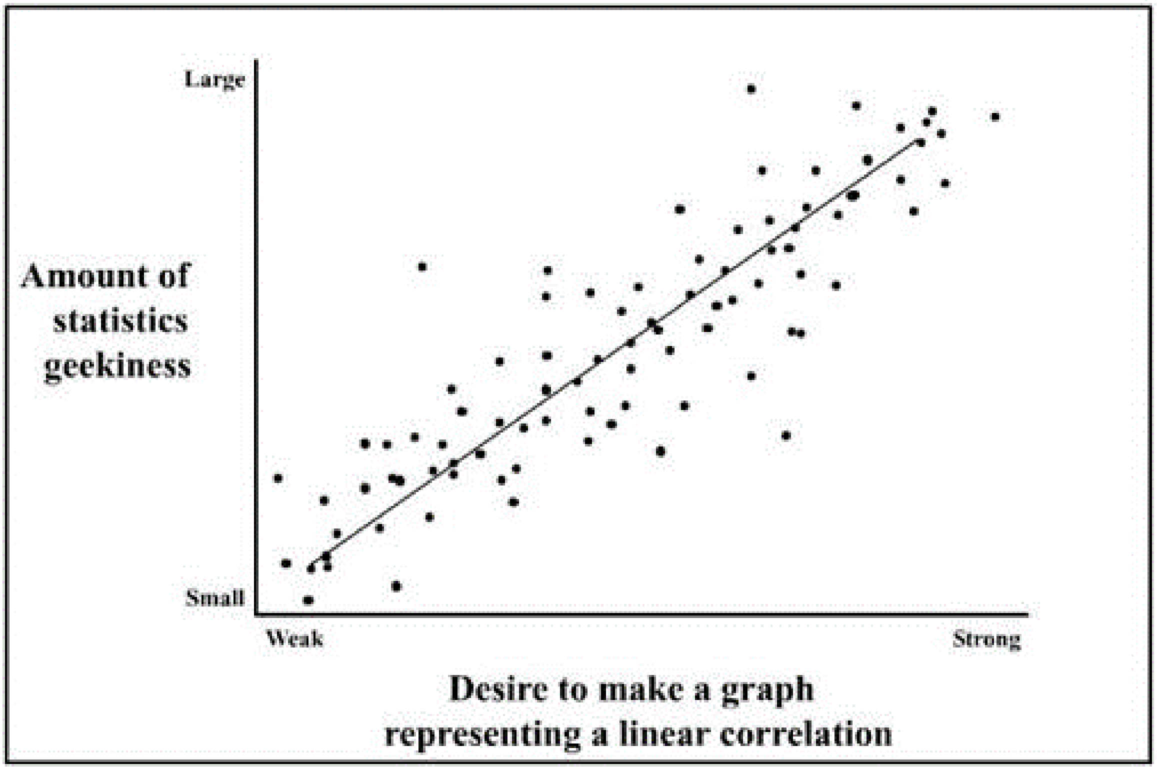
Linear regression

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

$$\hat{y} = f(x)$$

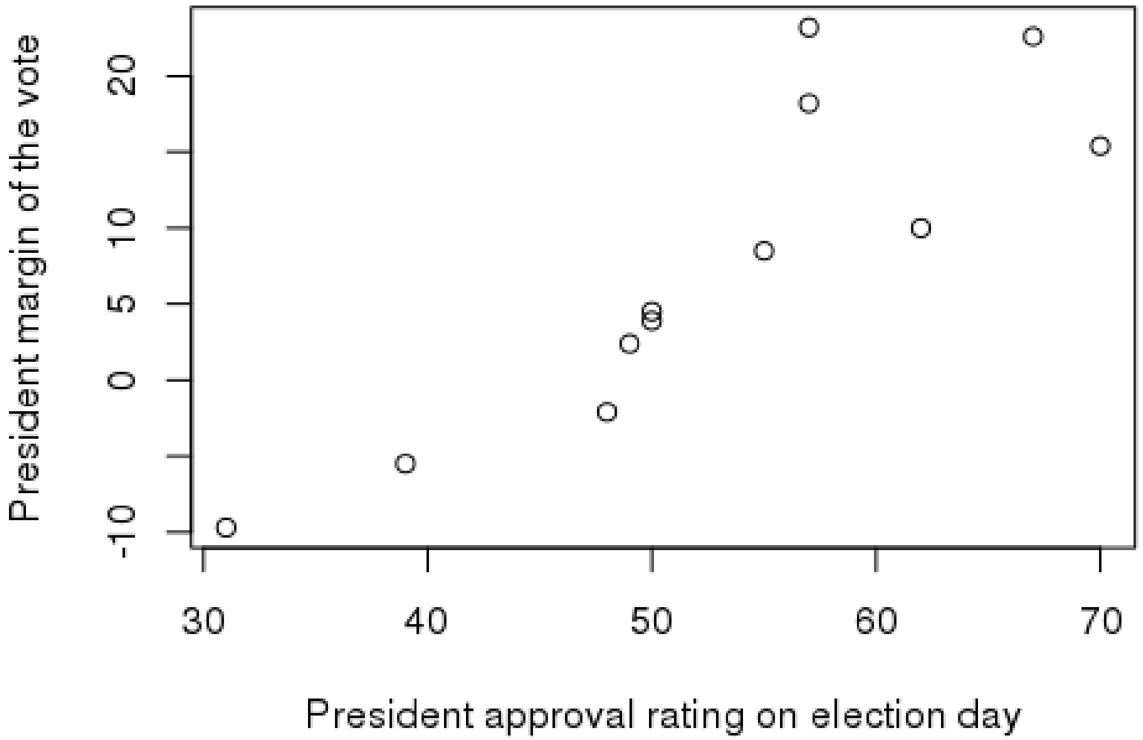
In **linear regression** we fit a line to the data, called the **regression line**

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$



Approval rating vote margin regression line

From previous 12 US president's running for reelection



$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

R: `lm(y ~ x)`

$\hat{\beta}_0$ = -36.76

$\hat{\beta}_1$ = 0.84

$\hat{y}$ = -36.76 + 0.84 · x

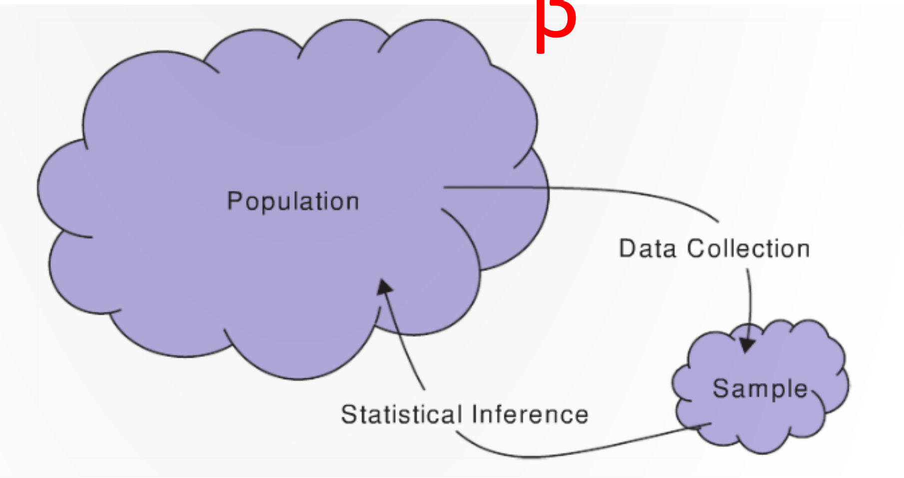
The **least squares line**, minimizes the sum of squared residuals

Inference for linear regression

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

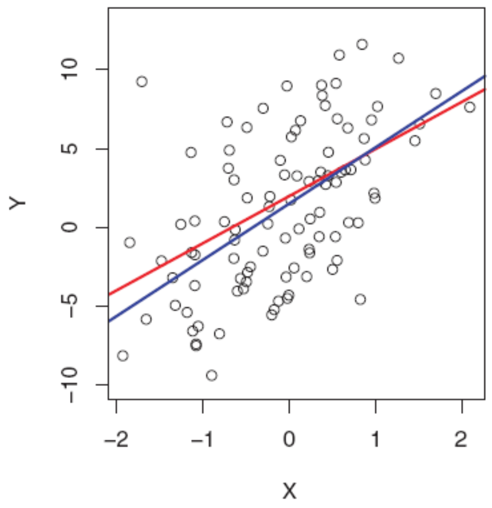
The letter $\hat{\beta}$ is typically used to denote the slope ***of the sample***

β



$\hat{\beta}$

True line Estimated line



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: $t = \frac{\hat{\beta}_1 - 0}{\hat{SE}_{\hat{\beta}_1}}$

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

$$\hat{SE}_{\hat{\beta}_1} = \frac{\hat{\sigma}_\epsilon}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}} \qquad \hat{SE}_{\hat{\beta}_0} = \hat{\sigma}_\epsilon \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

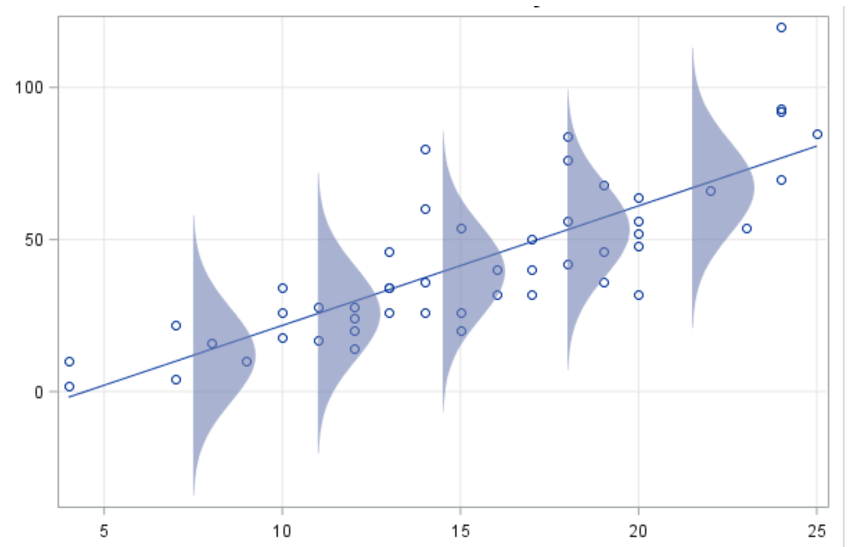
Inference using parametric methods

When using parametric methods, we make the following (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

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

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

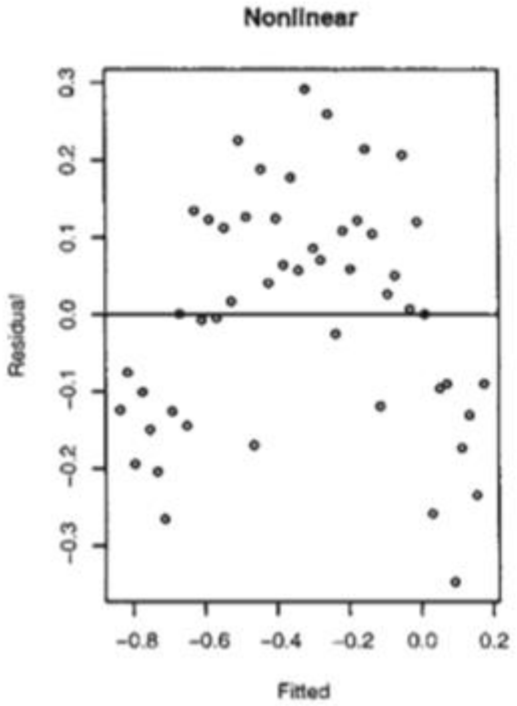


These assumptions are usually checked after the models are fit using 'regression diagnostic' plots.

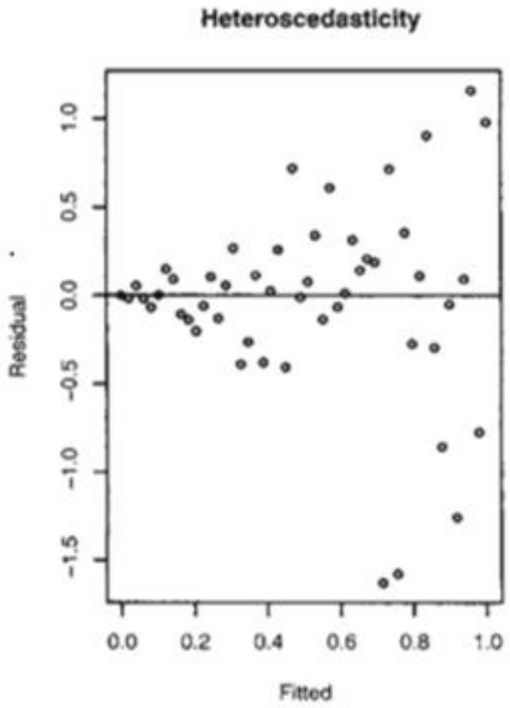
Regression diagnostics

Linearity, Independence, Normality, Equal variance of errors

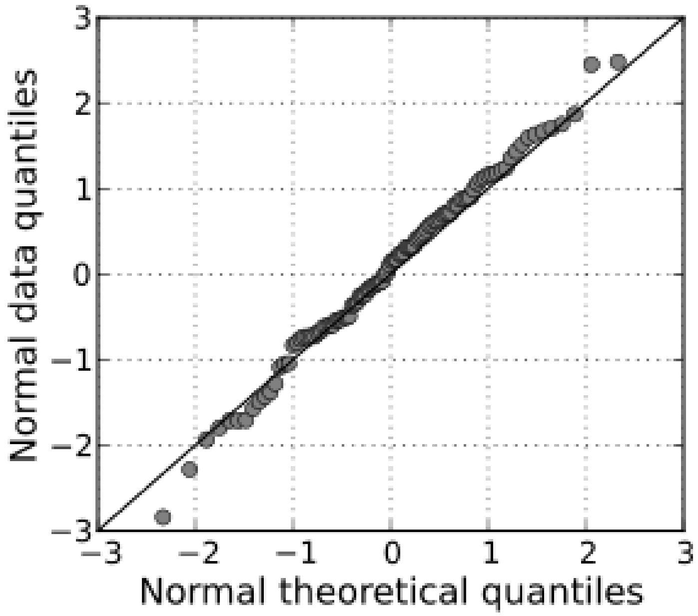
$e = y - \hat{y}$



$\hat{y}$



$\hat{y}$



Let's try it in R...

Let's try to predict faculty salary's based on a school's endowment

Multiple regression

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

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \cdot x_1 + \hat{\beta}_2 \cdot x_2 + \ldots + \hat{\beta}_k \cdot x_k$$

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)

Multiple regression

$$\text{salary} \;=\; \hat{\beta}_0 \;+\; \hat{\beta}_1 \cdot \text{f(endowment)} \;+\; \hat{\beta}_2 \cdot \text{g(enrollment)}$$

Let's explore this in R…