Practical Business Python

Lecture 9: Statistical Methods in Python: Inferences and Regressions

Iegor Vyshnevskyi

Woosong University

November 16, 2023

Agenda

1. Intro to Statistical Inference
2. Intro to Confidence Interval
3. Calculation of Confidence Interval
4. Basic Concepts of Hypothesis Testing
5. Hypothesis Testing Practical Examples
6. Intro to Regression
7. Assumptions of Linear Regression
8. Regression: practical use
9. In-class Assignment

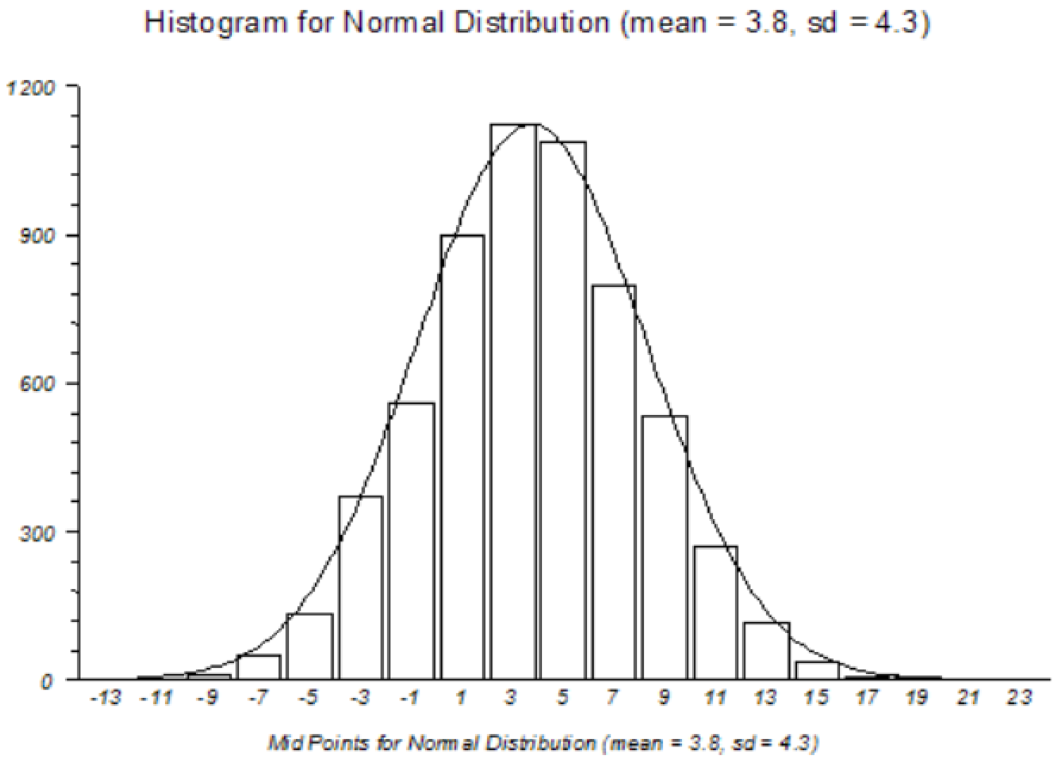
1. Intro to Statistical Inference

Terminology

- A ***variable*** is what we measure and want to study in our project. It could be employee salaries or the transaction value of customers, for example
- A ***population*** is a set of all units we want to draw conclusions about. For example: All the employees in an organization.
- And a ***sample*** is a subset of employees (in other words a specific group of employees) in the organization.

Terminology (cont.)

- A statistical distribution gives us an idea about how these values are distributed in a population. The most common distribution is a ***normal distribution***.



- A ***factor*** defines sub groups in a study such as the gender or location of employees.
- ***Descriptive Statistics*** typically include the mean, median, and standard deviation of a variable under study.

Statistical Inference

the process of drawing conclusions about unknown population properties, using a sample drawn from the population.

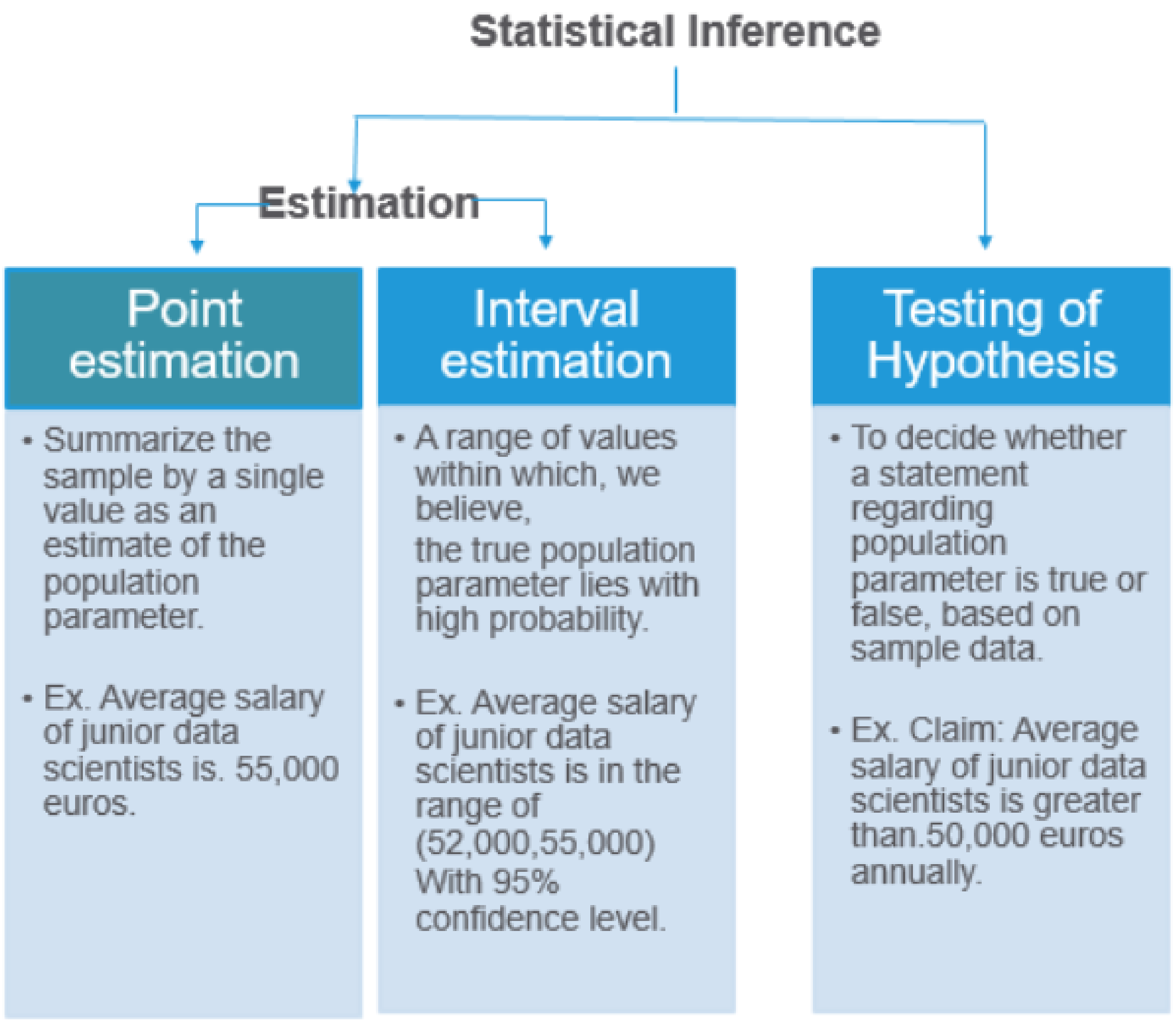
Unknown population properties can be, for example, mean, proportion or variance. These are also called ***parameters***.

Type of Statistical Inference

- ***statistical estimation*** is concerned with best estimating a value or range of values for a particular population parameter, and
- ***hypothesis testing*** is concerned with deciding whether the study data are consistent at some level of agreement with a particular population parameter.

Resource: https://www.engati.com/glossary/statistical-inference

Type of Statistical Inference (cont.)



Statistical Inference

Estimation

Point estimation

- Summarize the sample by a single value as an estimate of the true population parameter.
- Ex. Average salary of junior data scientists is. 55,000 euros.

Interval estimation

- A range of values within which, we believe, the true population parameter lies with high probability.
- Ex. Average salary of junior data scientists is in the range of (52,000,55,000) With 95% confidence level.

Testing of Hypothesis

- To decide whether a statement regarding population parameter is true or false, based on sample data.
- Ex. Claim: Average salary of junior data scientists is greater than.50,000 euros annually.

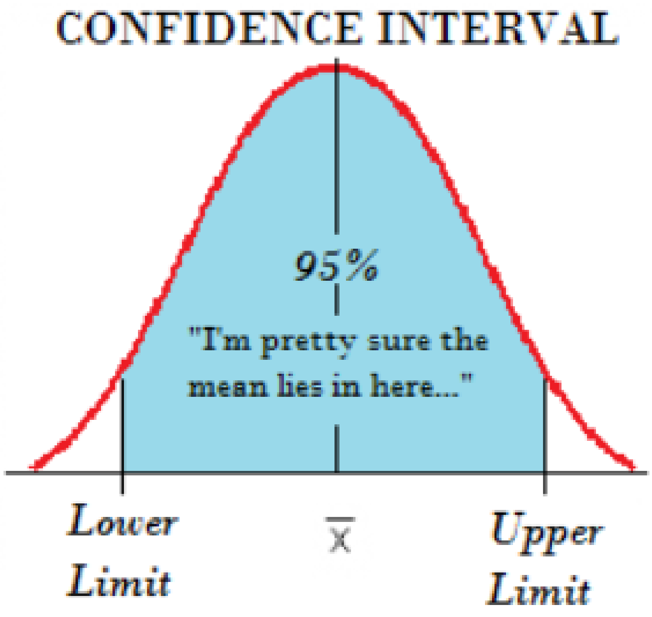
Resource: https://digitaschools.com/what-is-statistical-inference/

2. Intro to Confidence Interval

Confidence Interval

- the mean of your estimate plus and minus the variation in that estimate. This is the range of values you expect your estimate to fall between if you redo your test, within a certain level of confidence.

- ***Confidence***, in statistics, is another way to describe probability. For example, if you construct a confidence interval with a 95% confidence level, you are confident that 95 out of 100 times the estimate will fall between the upper and lower values specified by the confidence interval.



Resource: https://www.statisticshowto.com/confidence-limits-definition/

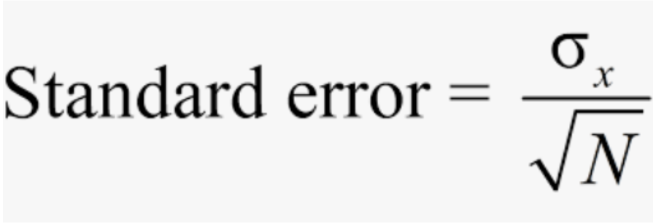
Confidence Interval Formula

Confidence interval =

Mean of sample ± Test Statistic * Standard Error

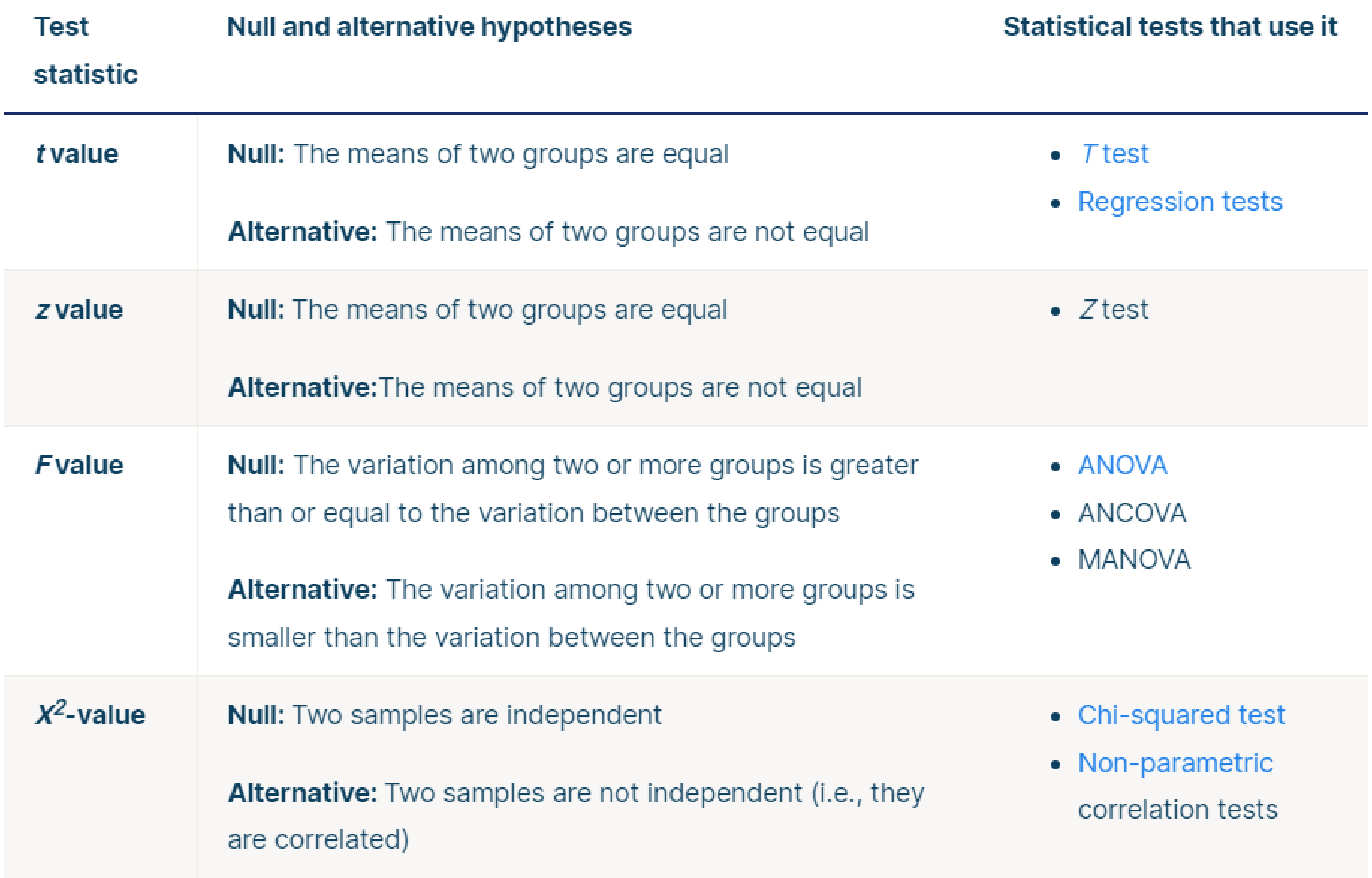
Test Statistic is a number calculated from a statistical test of a hypothesis. It shows how closely your observed data match the distribution expected under the null hypothesis of that statistical test.

Standard error = standard deviation / squared number of observations



$$\text{Standard error} = \frac{\sigma_x}{\sqrt{N}}$$

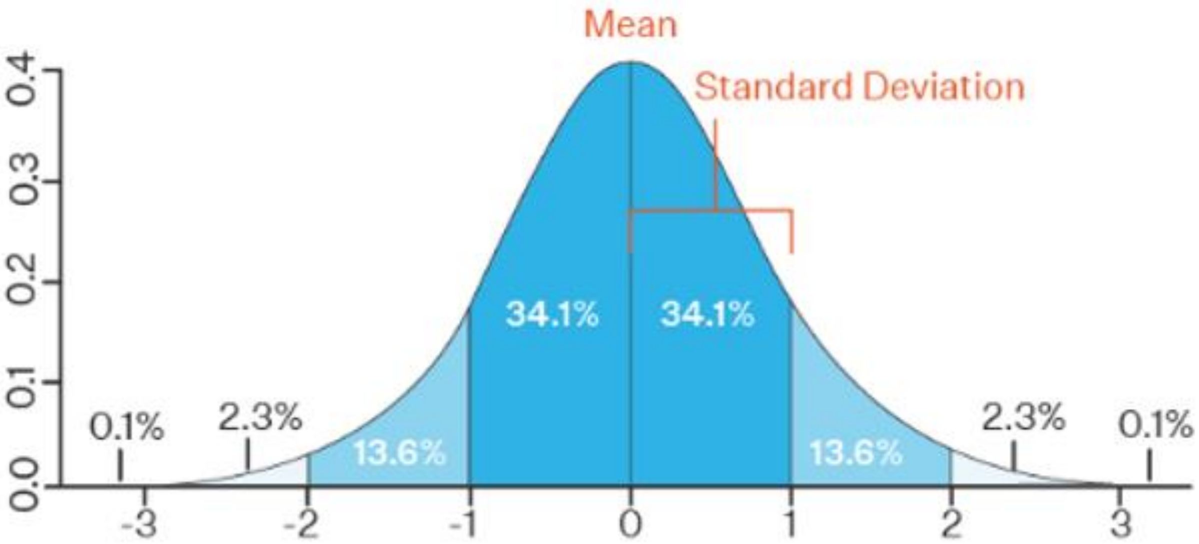
Type of Most Common Test Statistics



Test statistic	Null and alternative hypotheses	Statistical tests that use it
***t* value**	**Null:** The means of two groups are equal **Alternative:** The means of two groups are not equal	- *T* test - Regression tests
***z* value**	**Null:** The means of two groups are equal **Alternative:**The means of two groups are not equal	- *Z* test
***F* value**	**Null:** The variation among two or more groups is greater than or equal to the variation between the groups **Alternative:** The variation among two or more groups is smaller than the variation between the groups	- ANOVA - ANCOVA - MANOVA
X^2-value	**Null:** Two samples are independent **Alternative:** Two samples are not independent (i.e., they are correlated)	- Chi-squared test - Non-parametric correlation tests

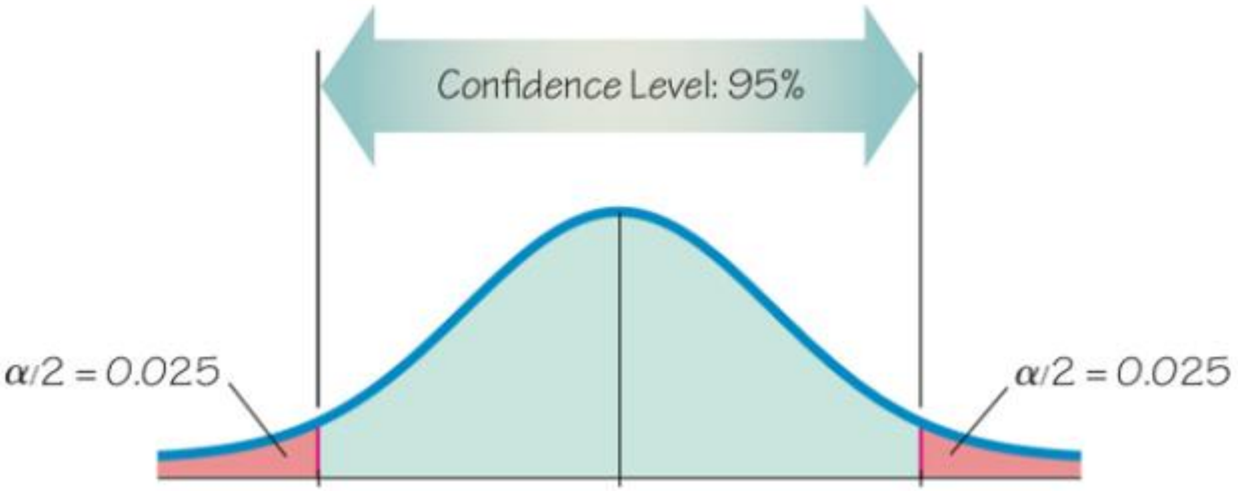
Resource: https://www.scribbr.com/statistics/test-statistic/

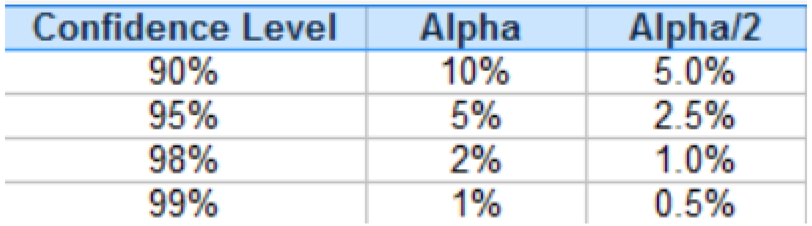
Standard deviation



- Around 68% of scores are within 1 standard deviation of the mean,
- Around 95% of scores are within 2 standard deviations of the mean,
- Around 99.7% of scores are within 3 standard deviations of the mean.

Standard deviation



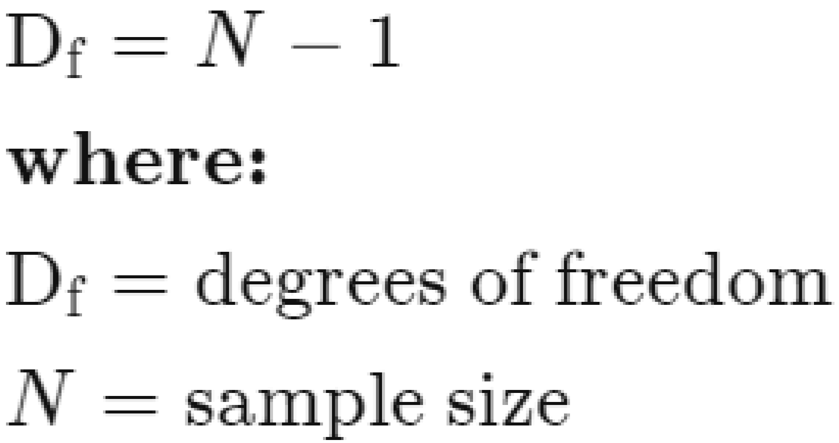


Confidence Level	Alpha	Alpha/2
90%	10%	5.0%
95%	5%	2.5%
98%	2%	1.0%
99%	1%	0.5%

Resource: https://math.stackexchange.com/questions/2835809/one-tailed-confidence-interval-1-2-alpha-rationale
https://www.statisticshowto.com/probability-and-statistics/confidence-interval/

Degree of Freedom

- the maximum number of logically independent values, which are values that have the freedom to vary, in the data sample.



$$\mathrm{D_f} = N - 1$$

where:

$\mathrm{D_f} =$ degrees of freedom

$N =$ sample size

3. Basic Concepts of Hypothesis Testing

*What is **Hypothesis Testing***

a type of statistical analysis in which you put your assumptions about a population parameter to the test.

First, we need to state the ***null*** and ***alternative*** hypothesis.

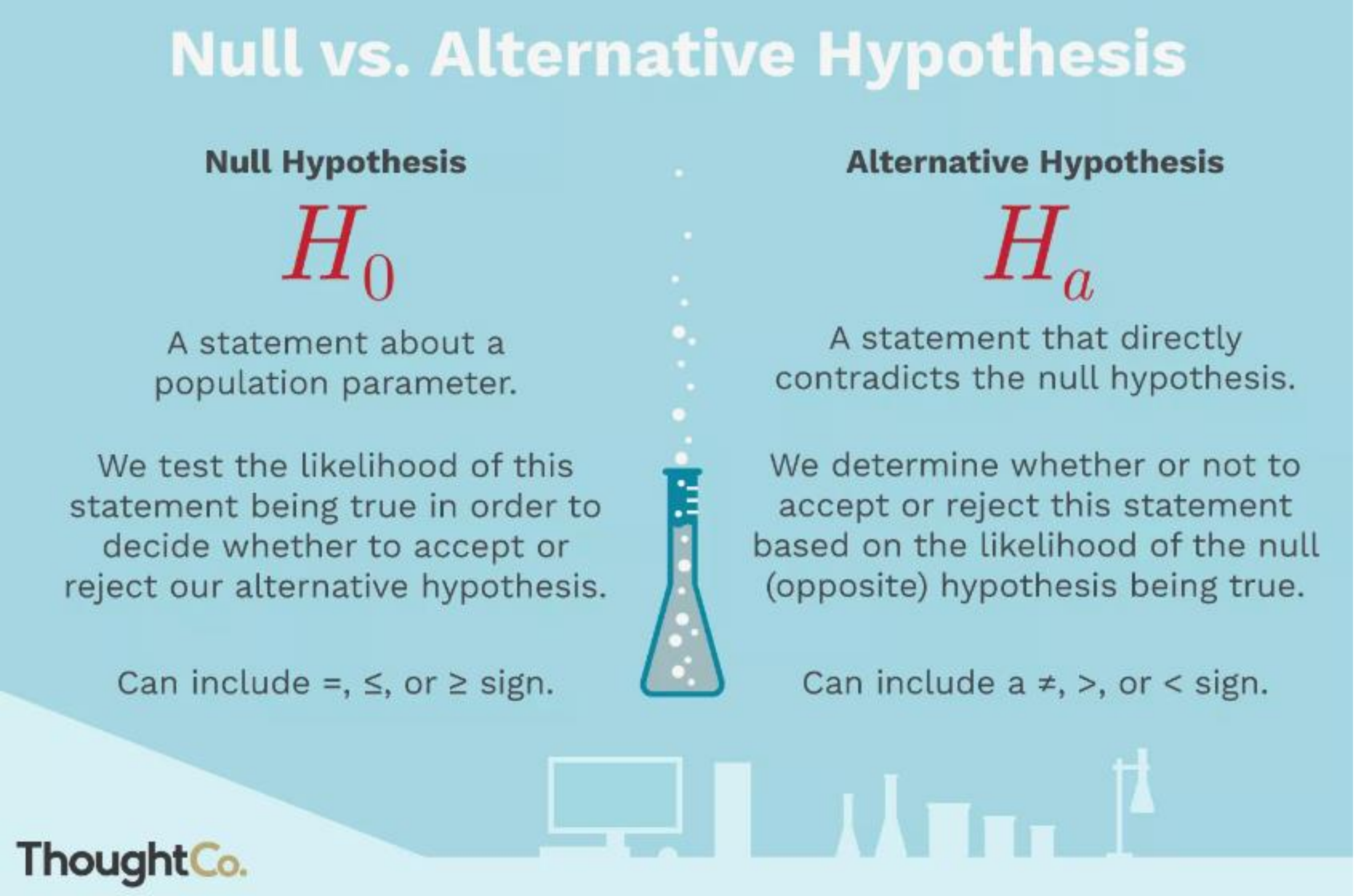
The ***Alternative Hypothesis*** is the Hypothesis which we are maintaining and would like to prove.

Example question: A company claims that their new product is more effective than the current market leader.

> ***Null hypothesis:*** The new product is **not** more effective than the current market leader.
>
> ***Alternative hypothesis***: The new product is more effective than the current market leader.

Example question: A researcher wants to investigate if there is a difference in the mean weight of two different breeds of dogs.

> ***Null hypothesis:*** There is **no** significant difference in the mean weight of the two breeds of dogs.
>
> ***Alternative hypothesis:*** There is a significant difference in the mean weight of the two breeds of dogs.



Null vs. Alternative Hypothesis

Null Hypothesis

H_0

A statement about a population parameter.

We test the likelihood of this statement being true in order to decide whether to accept or reject our alternative hypothesis.

Can include =, ≤, or ≥ sign.

Alternative Hypothesis

H_a

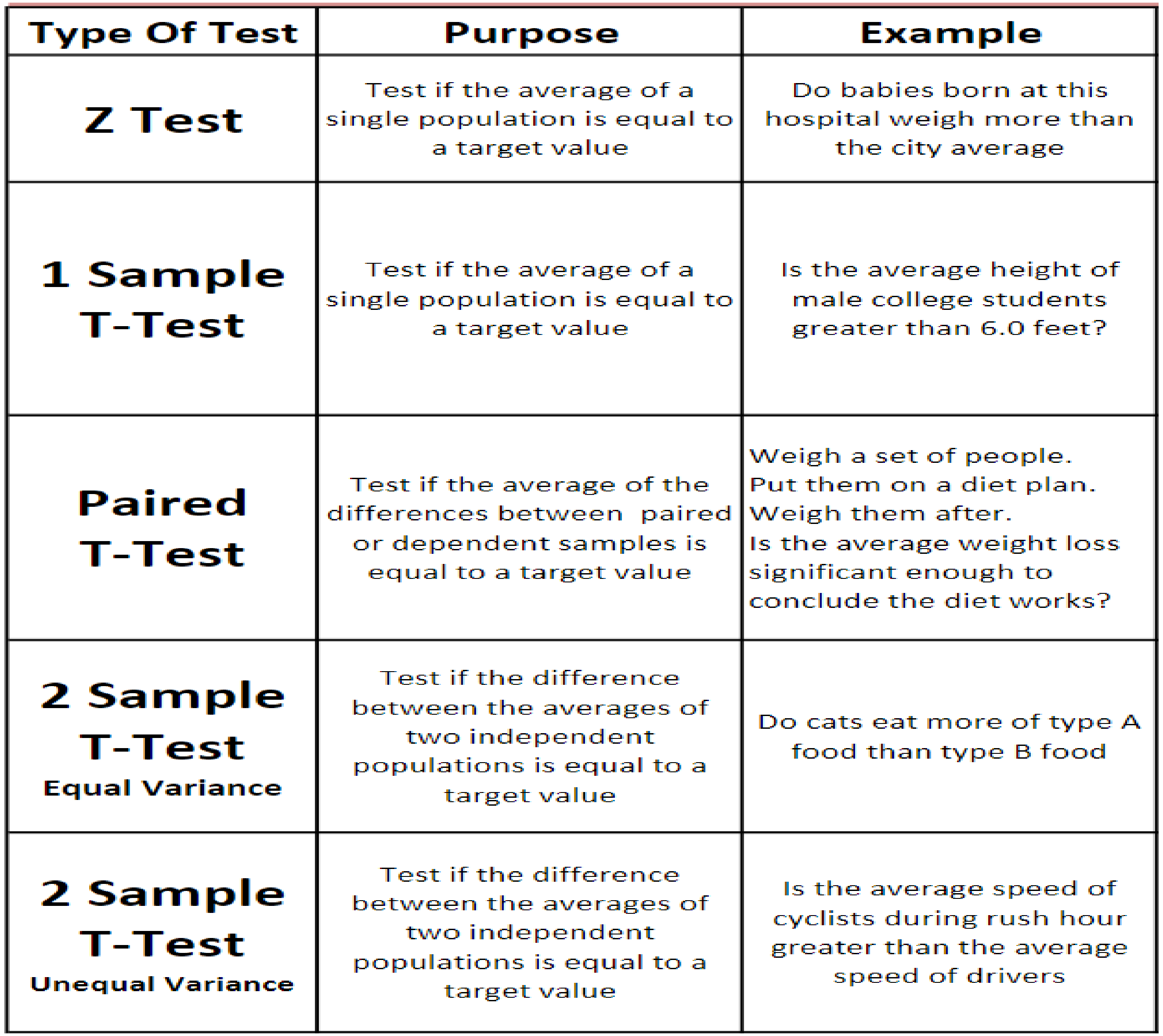
A statement that directly contradicts the null hypothesis.

We determine whether or not to accept or reject this statement based on the likelihood of the null (opposite) hypothesis being true.

Can include a ≠, >, or < sign.

ThoughtCo.

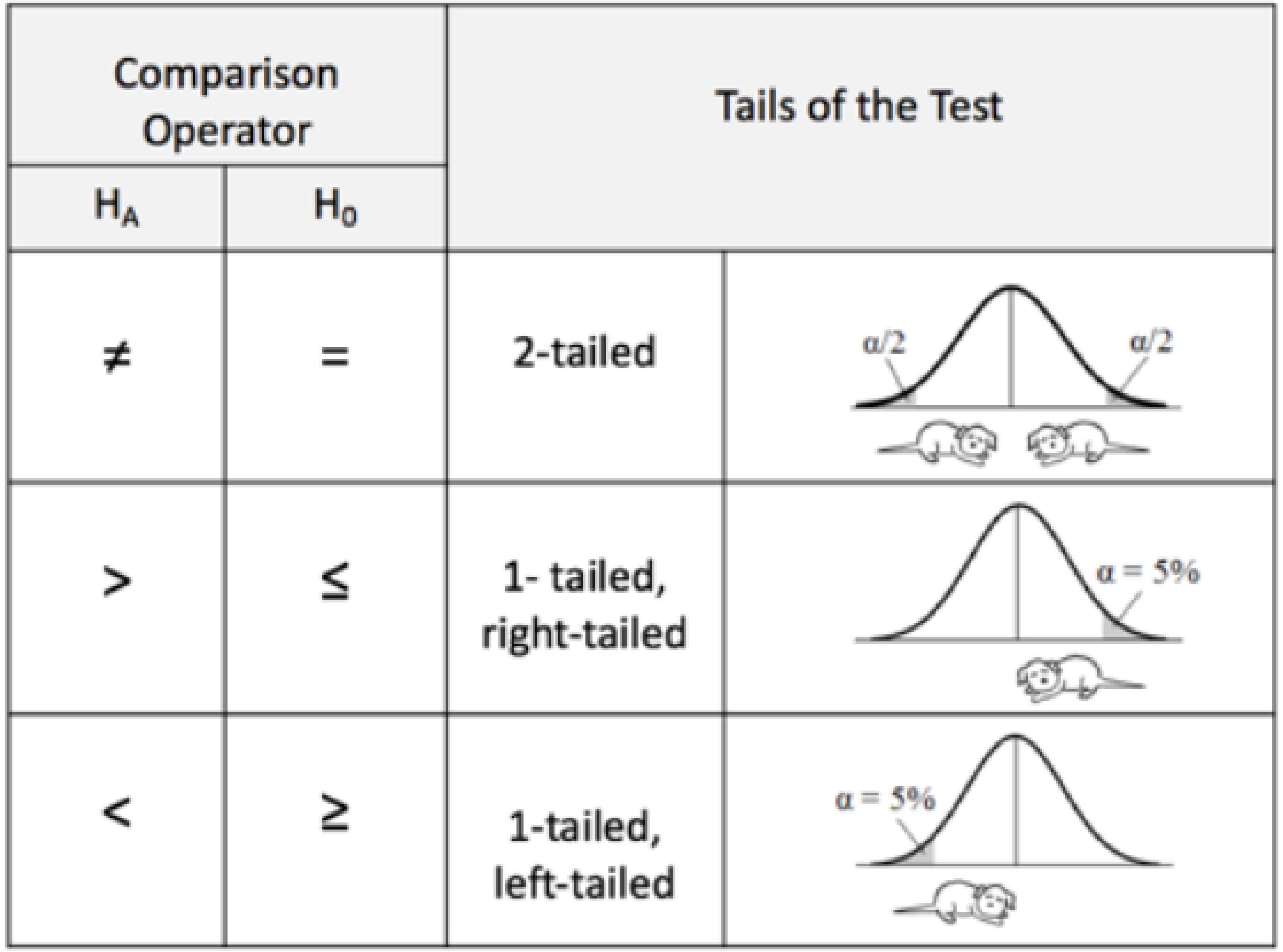
Type of hypothesis tests (cheat sheet)



Type Of Test	Purpose	Example
Z Test	Test if the average of a single population is equal to a target value	Do babies born at this hospital weigh more than the city average
1 Sample T-Test	Test if the average of a single population is equal to a target value	Is the average height of male college students greater than 6.0 feet?
Paired T-Test	Test if the average of the differences between paired or dependent samples is equal to a target value	Weigh a set of people. Put them on a diet plan. Weigh them after. Is the average weight loss significant enough to conclude the diet works?
2 Sample T-Test Equal Variance	Test if the difference between the averages of two independent populations is equal to a target value	Do cats eat more of type A food than type B food
2 Sample T-Test Unequal Variance	Test if the difference between the averages of two independent populations is equal to a target value	Is the average speed of cyclists during rush hour greater than the average speed of drivers

Resource: https://studylib.net/doc/25486595/hypothesis-testing-cheat-sheet-1

One and Two-tailed tests



Comparison Operator		Tails of the Test	
H_A	H_0		
≠	=	2-tailed	α/2 α/2
>	≤	1- tailed, right-tailed	α = 5%
<	≥	1-tailed, left-tailed	α = 5%

Example:

To test whether the Mean lifetime of the lightbulbs we manufacture is more than 1,300 hours.

For two-tailed test:

H0: Mean = 1300

Ha: Mean ≠ 1300

For right tailed test:

H0: Mean ≤ 1300

Ha: Mean > 1300

For left tailed test:

H0: Mean ≥ 1300

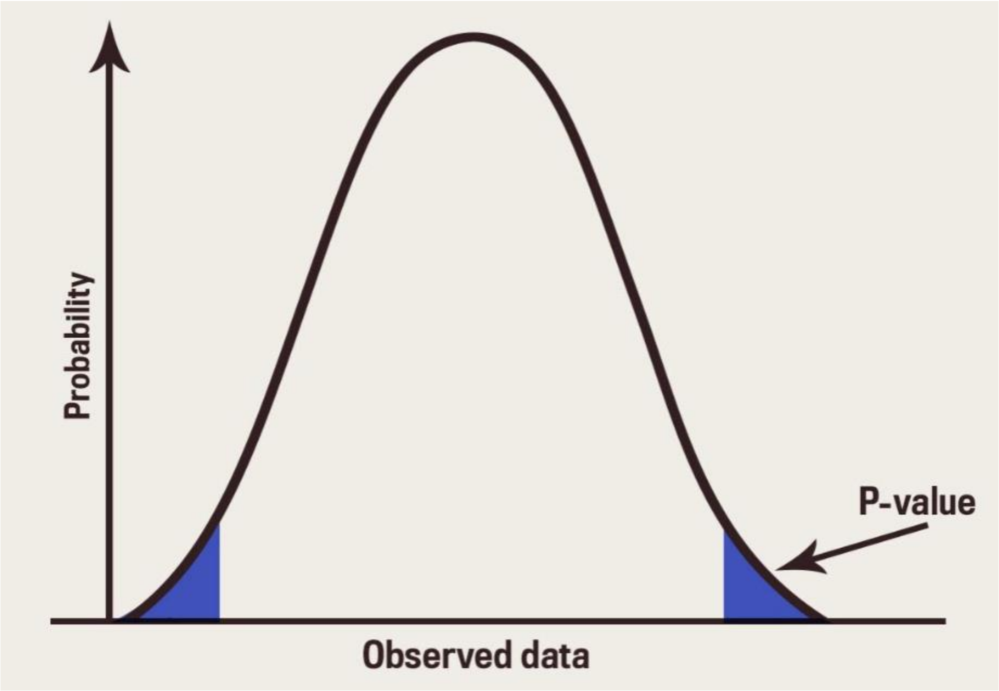
Ha: Mean < 1300

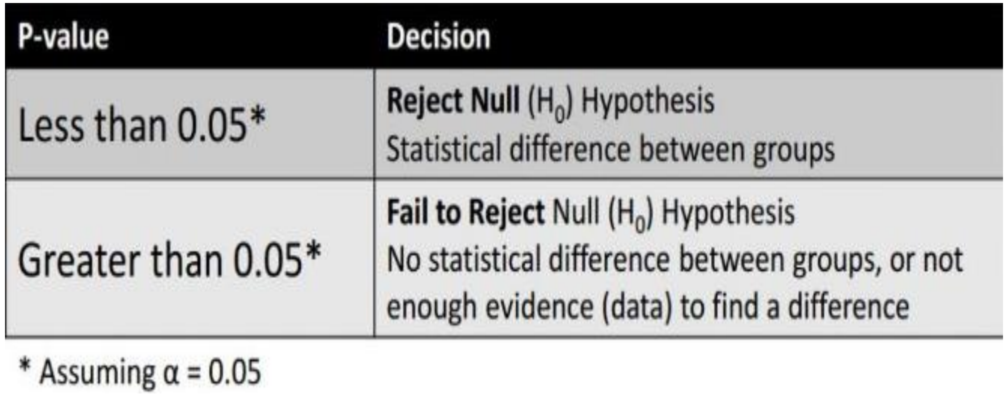
P-value

The p-value is a probability.

When the p-value is very small, it means it is very unlikely (small probability) that the observed spatial pattern is the result of random processes, so you can reject the null hypothesis.

An easier way to remember the decision of a hypothesis test is by using the phrase ***"when p is low, the null must go."***



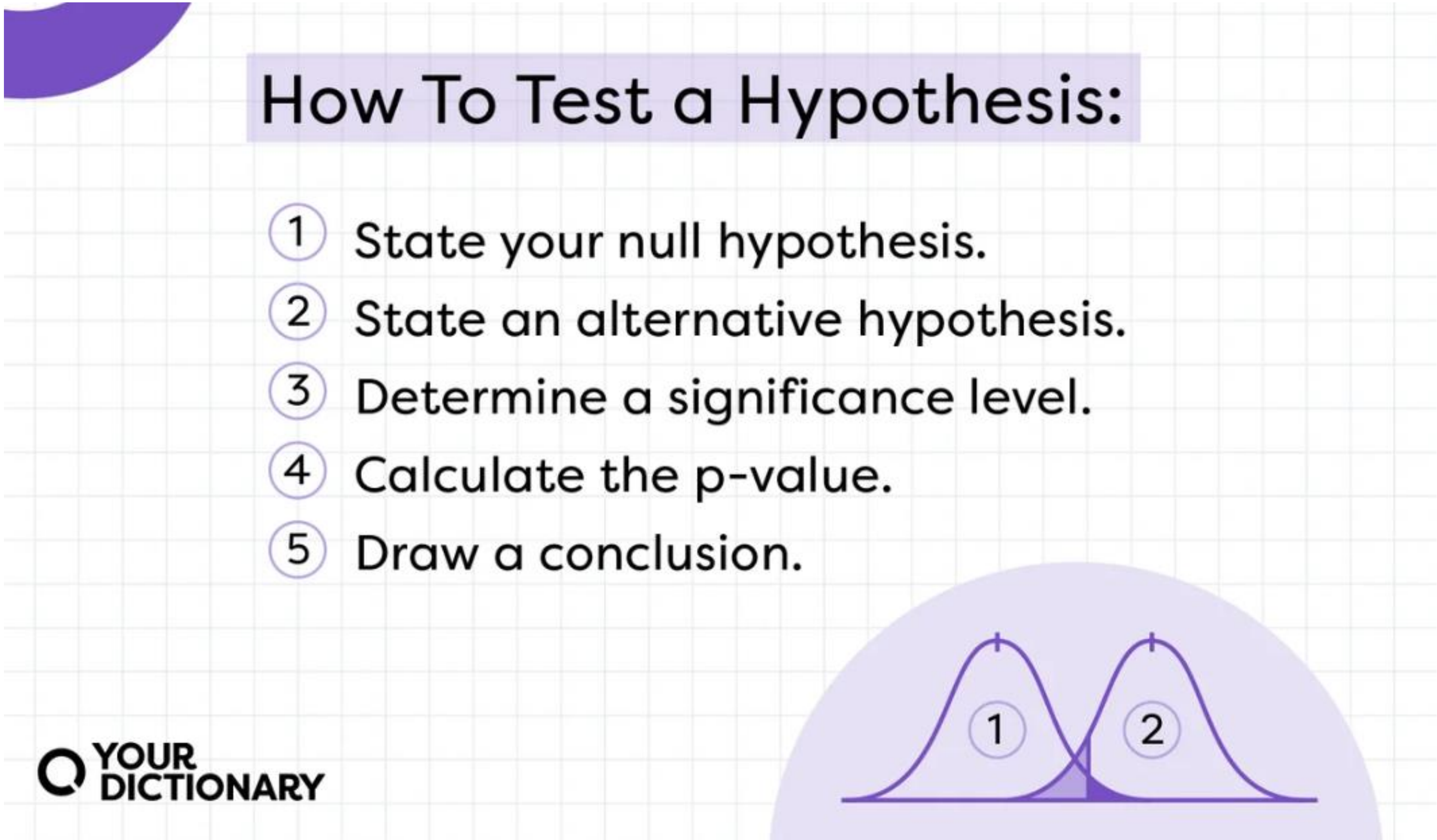


P-value	Decision
Less than 0.05*	**Reject Null** (H_0) Hypothesis Statistical difference between groups
Greater than 0.05*	**Fail to Reject** Null (H_0) Hypothesis No statistical difference between groups, or not enough evidence (data) to find a difference

* Assuming $\alpha = 0.05$

Resource: https://www.leansixsigmadefinition.com/glossary/p-value/

To sum up



How To Test a Hypothesis:

1. State your null hypothesis.
2. State an alternative hypothesis.
3. Determine a significance level.
4. Calculate the p-value.
5. Draw a conclusion.

YOUR DICTIONARY

4. Intro to Regression

Regression

is a statistical method used to study the relationship between a dependent variable (usually denoted as Y) and one or more independent variables (usually denoted as X).

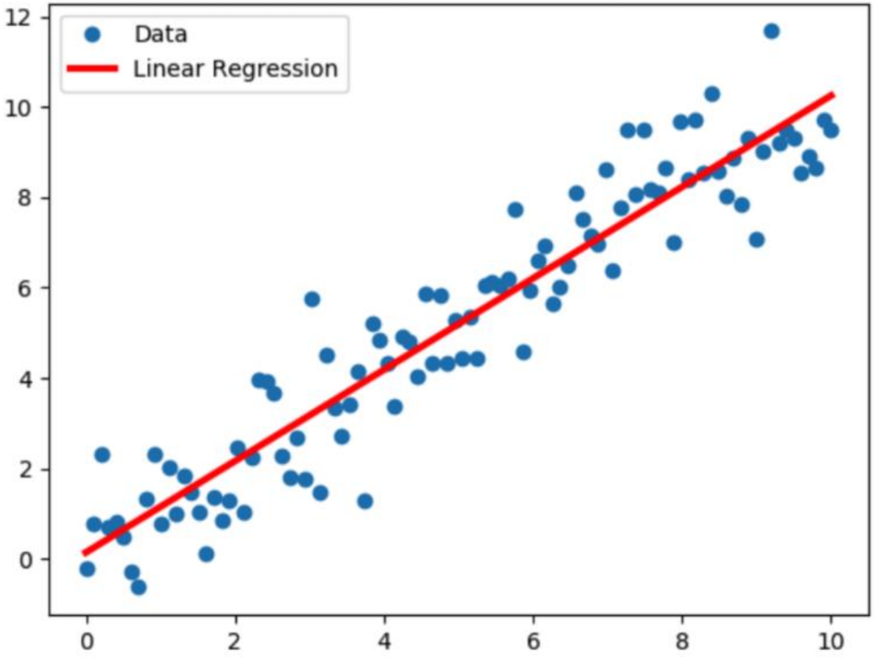
- It is commonly used for predictive modeling, to estimate the value of the dependent variable based on the values of one or more independent variables.
- The goal of regression is to find the best-fitting line or curve that can describe the relationship between the variables, which can then be used to make predictions or to understand how changes in the independent variable(s) affect the dependent variable.

Type of regressions

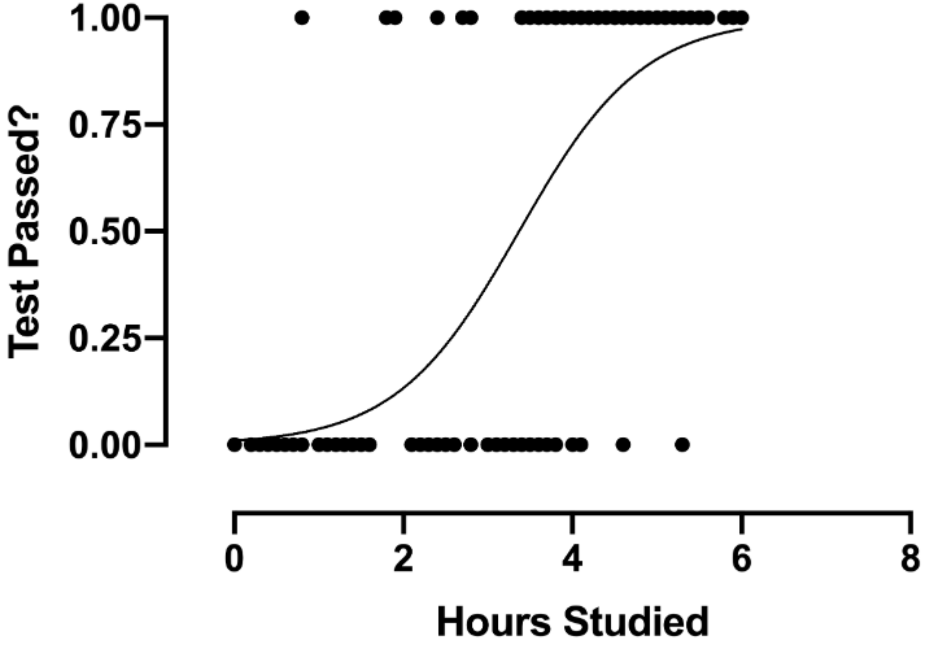
- Linear Regression
- Logistic Regression
- Polynomial Regression
- Ridge Regression
- Lasso Regression
- Quantile Regression
- Bayesian Linear Regression
- Principal Components Regression
- Partial Least Squares Regression
- Elastic Net Regression

Type of regressions (cont.)

Linear Regression: A method that models the relationship between a dependent variable and one or more independent variables as a straight line.



Logistic Regression: A method used to model the probability of a binary outcome (i.e., yes or no, true or false) based on one or more predictor variables.

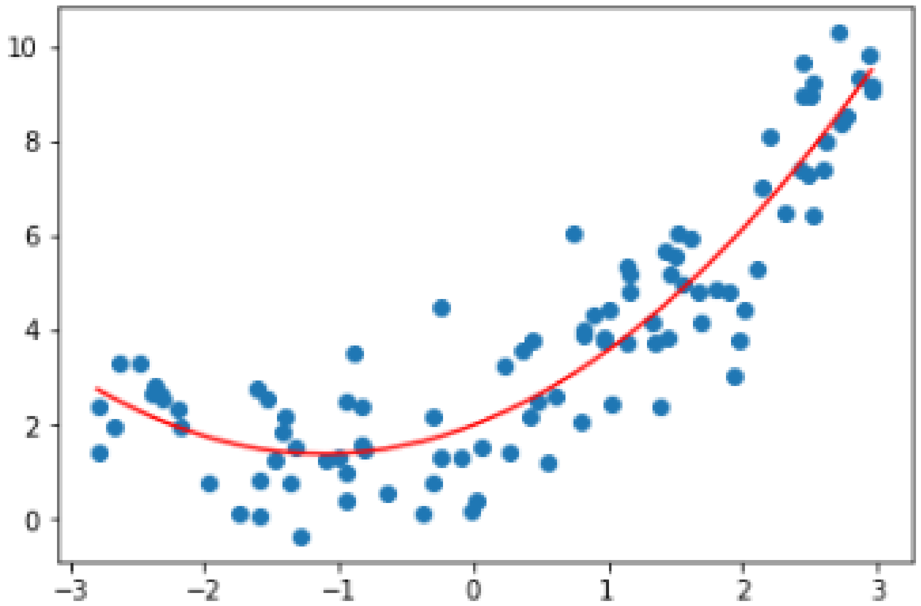


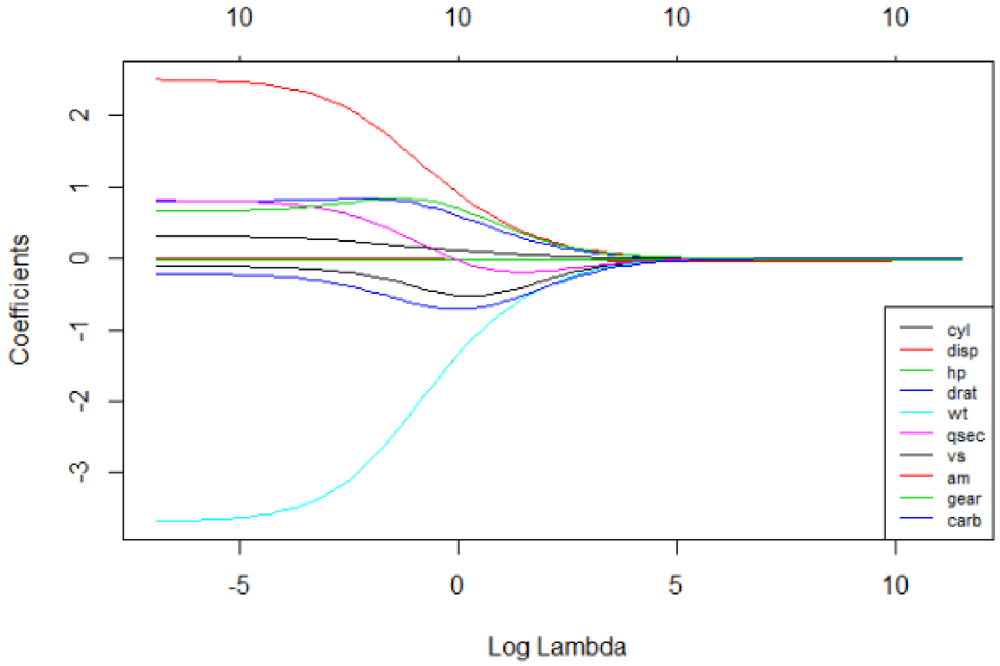
Resource: https://www.analyticsvidhya.com/blog/2022/01/different-types-of-regression-models/

Type of regressions (cont.)

Polynomial Regression: A method that models the relationship between a dependent variable and one or more independent variables using a polynomial function.

Ridge Regression: A method used to avoid overfitting in linear regression by adding a penalty term to the regression equation.



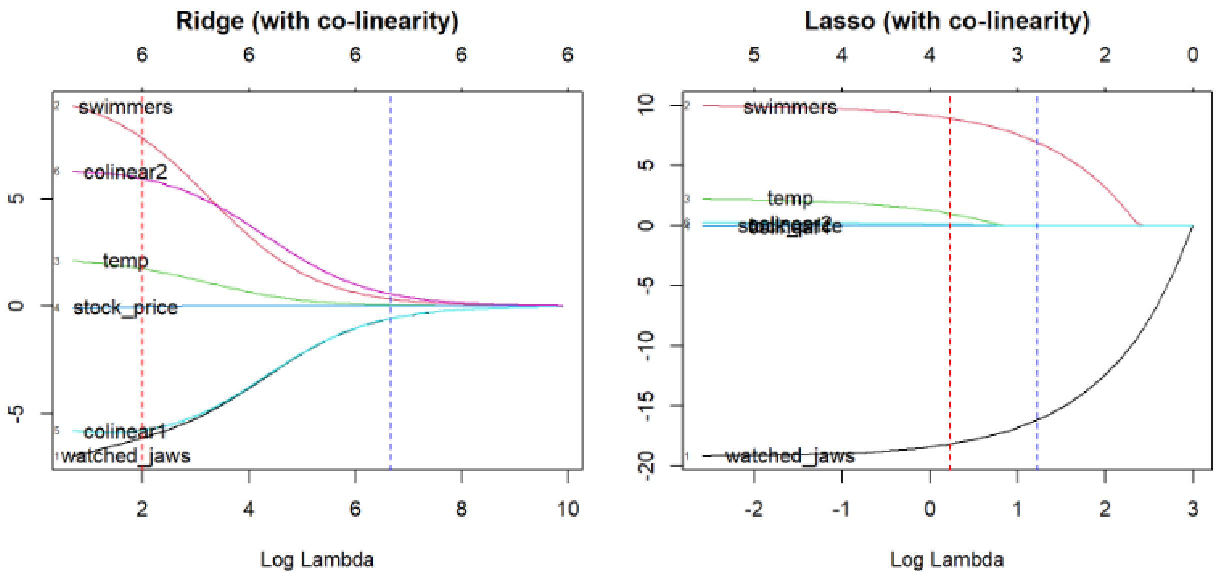


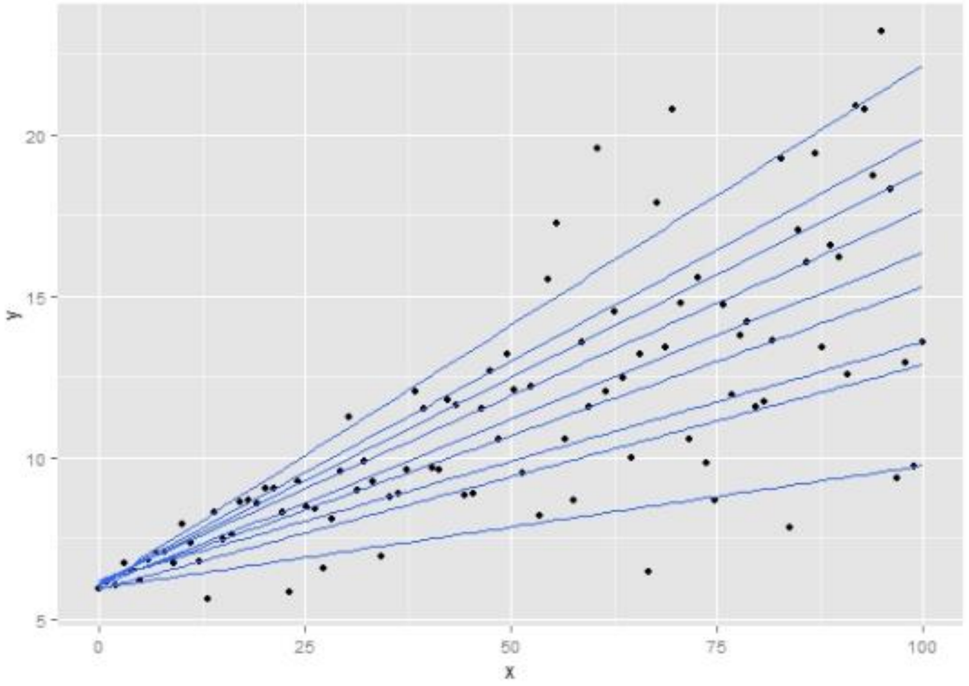
Resource: https://www.analyticsvidhya.com/blog/2022/01/different-types-of-regression-models/

Type of regressions (cont.)

Lasso Regression: A method used to select important predictor variables and avoid overfitting in linear regression by shrinking the coefficients of less important variables to zero.

Quantile Regression: A method that estimates the relationship between a dependent variable and one or more independent variables at different quantiles of the dependent variable.

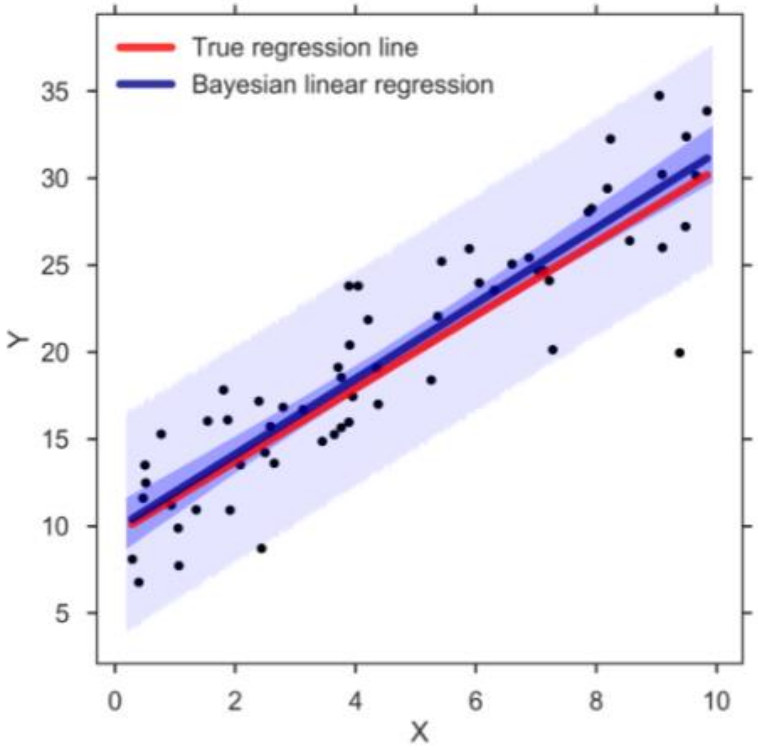




Resource: https://www.analyticsvidhya.com/blog/2022/01/different-types-of-regression-models/

Type of regressions (cont.)

Bayesian Linear Regression: A method that uses Bayesian inference to estimate the parameters of a linear regression model.



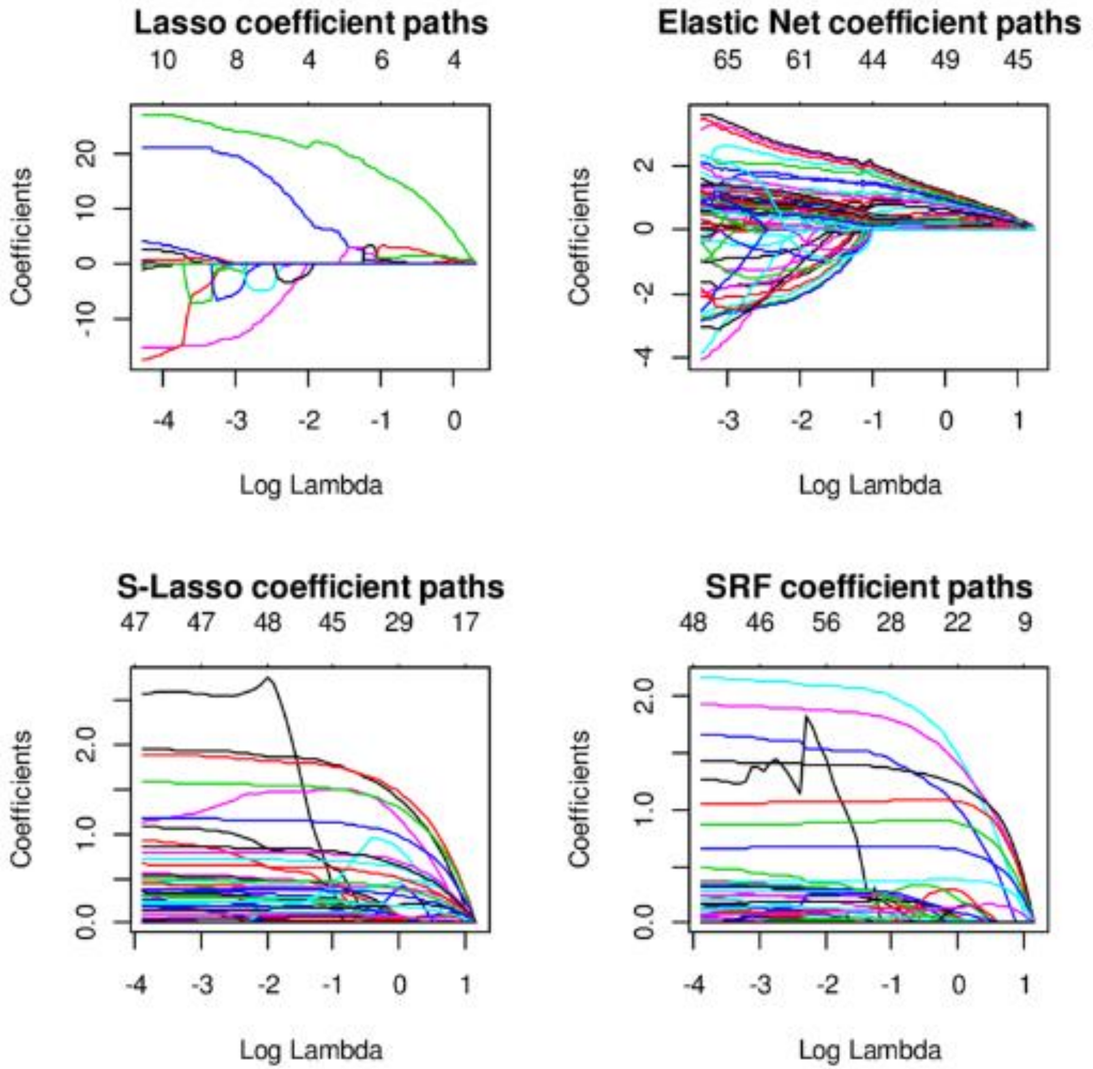
Principal Components Regression: A method that uses principal component analysis to reduce the dimensionality of the predictor variables before performing linear regression.

Partial Least Squares Regression: A method that uses partial least squares regression to reduce the dimensionality of the predictor variables before performing linear regression.

Resource: https://www.analyticsvidhya.com/blog/2022/01/different-types-of-regression-models/

Type of regressions (cont.)

Elastic Net Regression: A method that combines the penalty terms of ridge regression and lasso regression to overcome their limitations and select important predictor variables while avoiding overfitting.



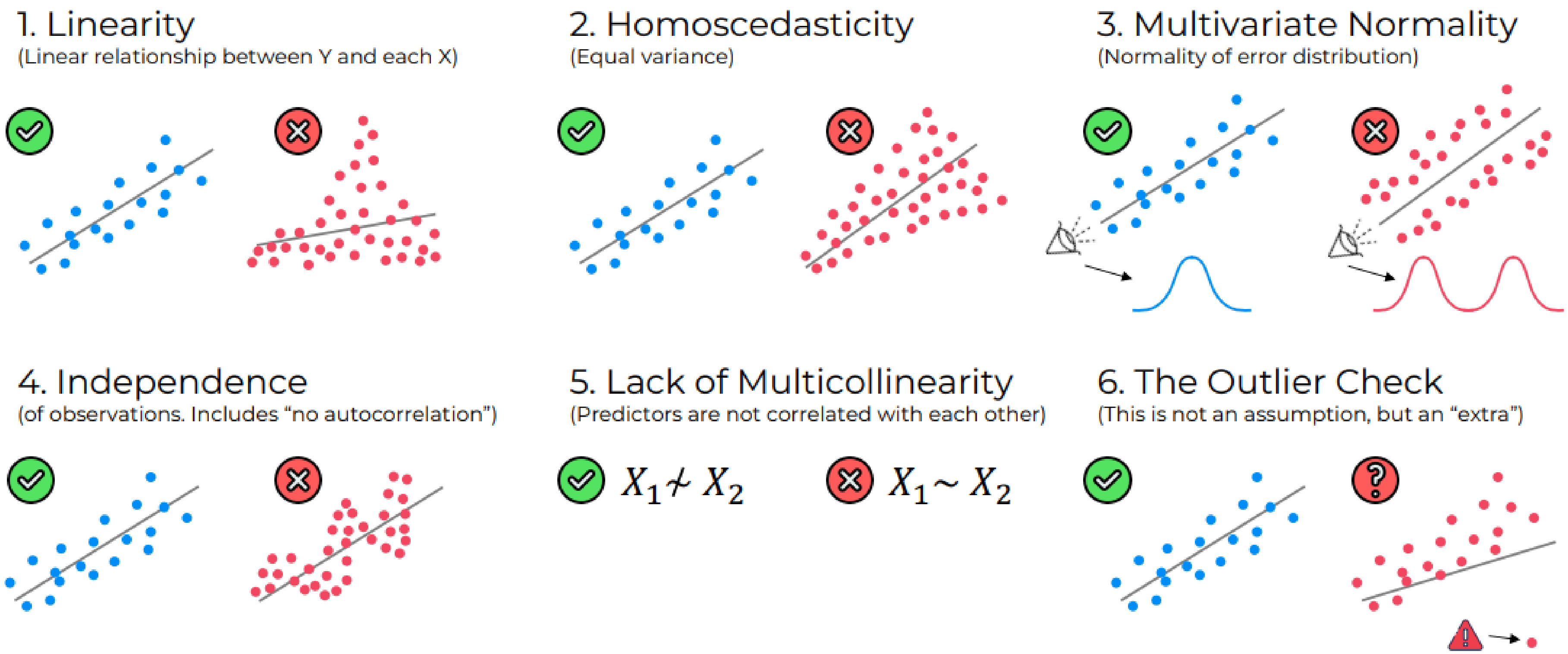
Resource: https://www.analyticsvidhya.com/blog/2022/01/different-types-of-regression-models/

5. Assumptions of Linear Regression

Linear regression is a widely used statistical technique for modeling the relationship between a dependent variable and one or more independent variables.

However, the accuracy of the regression model depends on several assumptions that need to be satisfied for the model to be valid.

Key Assumptions of Linear Regression


1. Linearity
(Linear relationship between Y and each X)
2. Homoscedasticity
(Equal variance)
3. Multivariate Normality
(Normality of error distribution)
4. Independence
(of observations. Includes "no autocorrelation")
5. Lack of Multicollinearity
(Predictors are not correlated with each other)
$X_1 \nsim X_2$
$X_1 \sim X_2$
6. The Outlier Check
(This is not an assumption, but an "extra")

If these assumptions are not met, the regression model may produce biased or inconsistent estimates, and the results may be invalid.

Therefore, it is important to check for these assumptions before using linear regression for modeling the data.

Resource: https://www.superdatascience.com/blogs/assumptions-of-linear-regression