Statistical Approaches for Pattern Recognition

Unit objectives

After completing this unit, you should be able to:

- Understand the concept of probability distributions
- Gain knowledge on example of statistical approaches
- Understand linear models for regression
- Learn about linear models for classification

Understanding statistics

- Statistics is a method of statistical research utilizing computational models, descriptions, and excerpts for theoretical or real-life data studies.

- Statistics is the analysis of how to draw inference from the evidence, feedback, and conclusions. Such statistical indicators involve.

Moment number	Name	Measure of	Formula
1	Mean	Central tendency	$\bar{X} = \frac{\sum_{i=1}^{N} X_i}{N}$
2	Variance (Volatility)	Dispersion	$\sigma^2 = \frac{\sum_{i=1}^{N} (X_i - \bar{X})^2}{N}$
3	Skewness	Symmetry (Positive or Negative)	$Skew = \frac{1}{N} \sum_{i=1}^{N} \left[\frac{(X_i - \bar{X})}{\sigma} \right]^3$
4	Kurtosis	Shape (Tall or flat)	$Kurt = \frac{1}{N} \sum_{i=1}^{N} \left[\frac{(X_i - \bar{X})}{\sigma} \right]^4$

Where X is a random variable having N observations (i = 1,2,…,N).

T-test

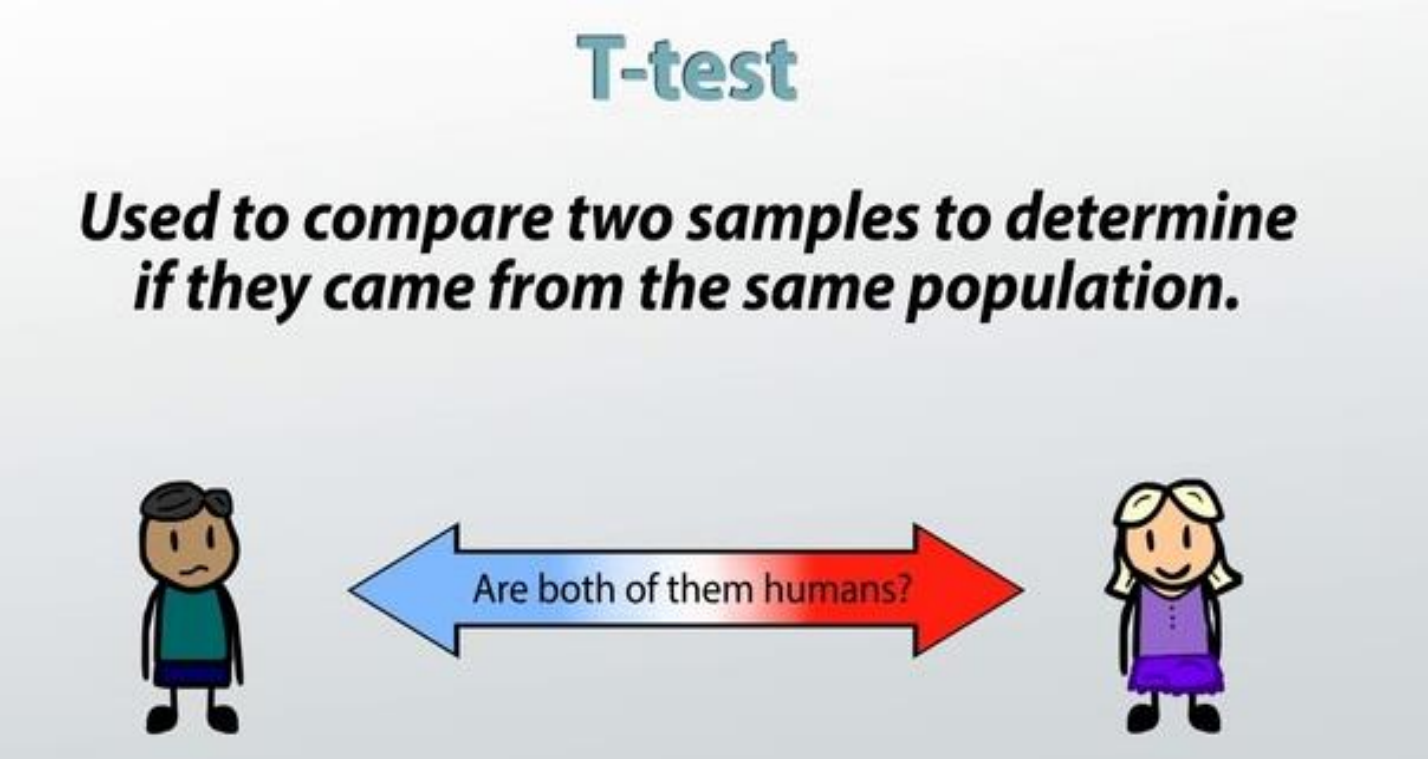


Figure: T-test example

Source: https://images.app.goo.gl/PAJFuF1hSWmx9y5o7

Z-test

- It is a method that tests how two population variables differ in the defined variances and the sample size.
- It is believed that the test statistics have a normal distribution.
- To do an accurate z-test, irritating parameters such as standard deviation need to be understood.
- Z-statistical or z-value is a sum that represents the amount, over or below the general population of standard variations generated by the Z-test.

Self evaluation: Exercise 6

- To continue with the training, after learning the various steps involved in pattern recognition and anomaly detection, it is instructed to utilize the concepts to perform the following activity.

- You are instructed to write the following activities using python code.

- Exercise 6: Polynomial regression for classification.

Z-test and t-test difference

- A typical and simplified method of statistical research is a z-test that measures the statistical validity of a sample mean to the predicted mean population but needs awareness of the standard model variance, which is not always feasible.

- The t-test is a more practical form of study since it needs just the standard deviation of the sample in comparison to the norm of the population.

P-value

- In mathematics, the p-value is the chance of producing outcomes as extreme as the findings of a mathematical experiment test obtained, given the null hypothesis is accurate.

- A lower p-value indicates better support for the alternate hypothesis.

Descriptive statistics

- Descriptive statistics are short descriptive equations summarizing a certain collection of results, which may either reflect the whole population or a subset of a community.

- Descriptive statistics are divided into core pattern measurements and volatility measurements (spread).

- Measures of central tendency include the mean, median, and mode, while measures of variability include the standard deviation, variance, the minimum and maximum variables, and the kurtosis and skewness.

Self evaluation: Exercise 7

- To continue with the training, after learning the various steps involved in pattern recognition and anomaly detection , it is instructed to utilize the concepts to perform the following activity.

- You are instructed to write the following activities using python code.

- Exercise 7: Neural networks.

Type I error

- Type I error always use If the null hypothesis is dismissed, even though it is true and should not be dismissed during the hypothesis testing phase.

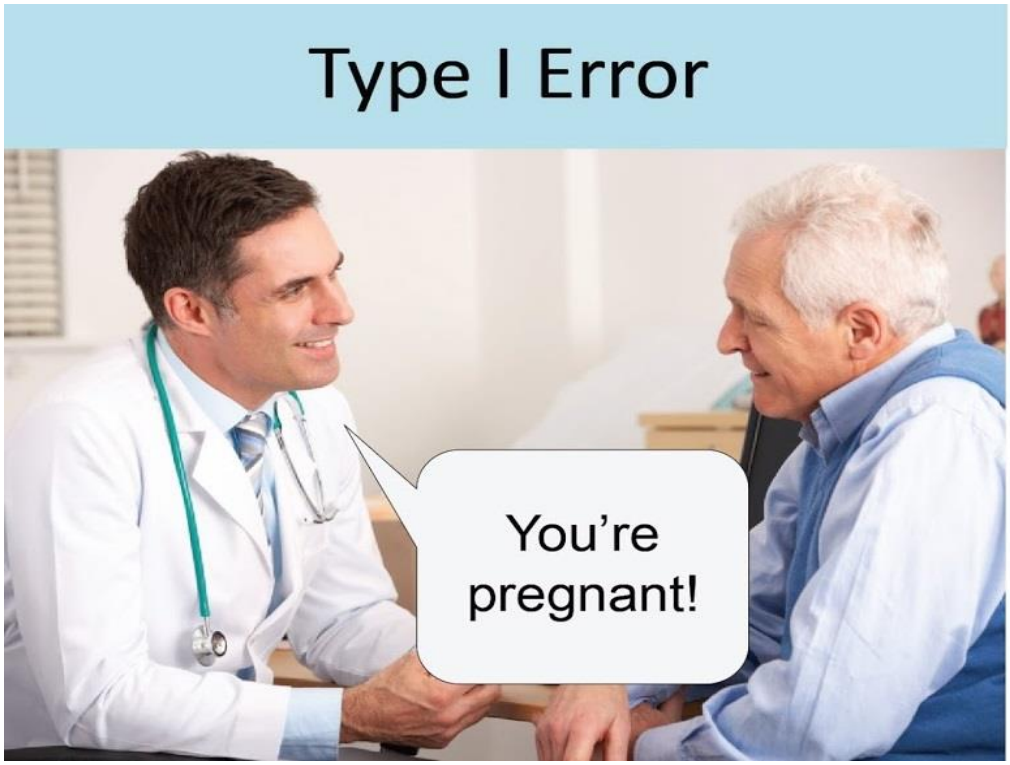


Figure: Type I error

Source: https://images.app.goo.gl/h3GAMnAgikgkrQ3GA

Type II error

- Throughout mathematical research, the dismissal of a real null hypothesis is a Type I, while the error of type II defines the mistake that happens when a null hypothesis is not discarded and is simply incorrect.

- Or put things another way, things generate a false statement. The fallacy denies the alternate explanation, although it does not arise out of chance.

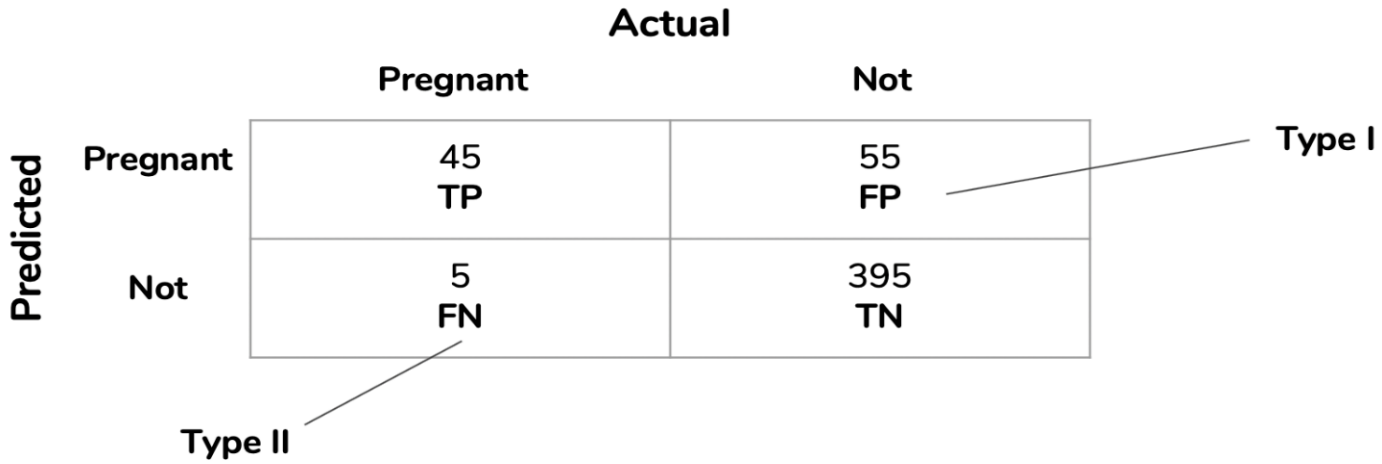
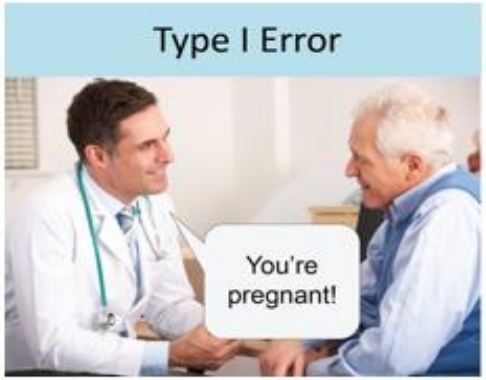


Figure: Confusion Matrix

Source: https://images.app.goo.gl/vjRF7VdSvThqaGbY9

Differences between type I and type II errors



	Null hypothesis is TRUE	Null hypothesis is FALSE
Reject null hypothesis	Type I Error (False positive)	Correct outcome! (True positive)
Fail to reject null hypothesis	Correct outcome! (True negative)	Type II Error (False negative)

Figure: type I and II error

Source: https://images.App.Goo.Gl/sr5bvbv93kw9hm378

Null hypothesis

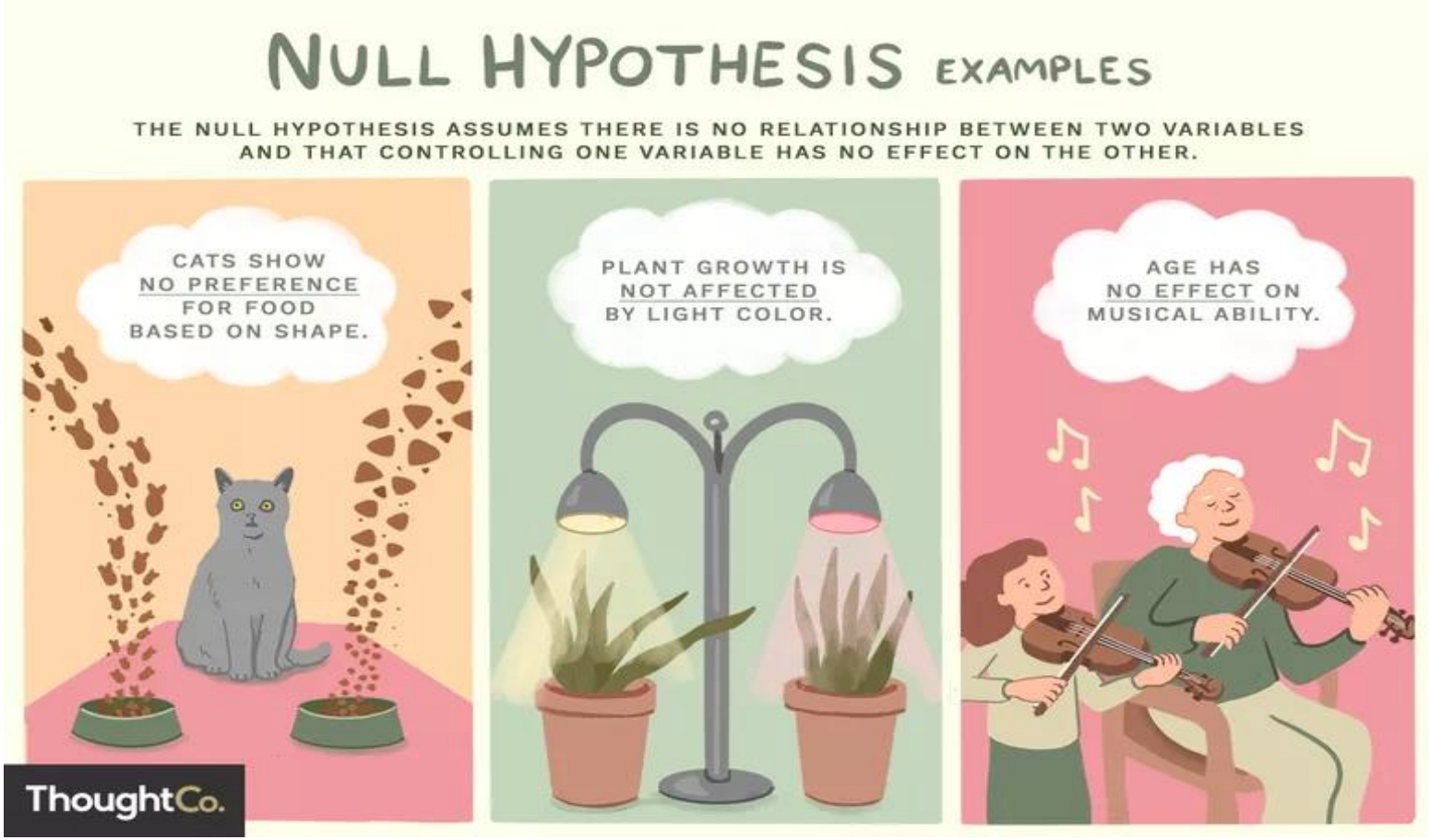


Figure: Null Hypothesis

Source: https://www.thoughtco.com/null-hypothesis-examples-609097

Statistical significance

- Statistical importance is an analyst's conviction that the findings in the data cannot be interpreted by chance alone.

- The tool by which the analyst makes the decision is mathematical hypothesis testing.

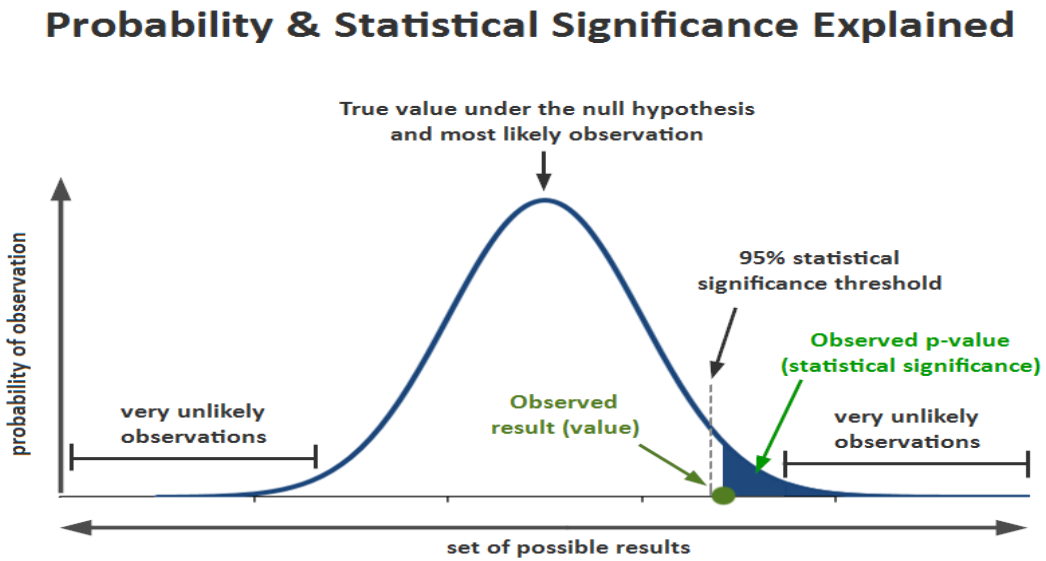


Figure: Probability and Significance overview

Source: https://images.app.goo.gl/jAxPTMP9VVxHhdHs6

Self evaluation: Exercise 8

- To continue with the training, after learning the various steps involved in pattern recognition and anomaly detection, it is instructed to utilize the concepts to perform the following activity.

- You are instructed to write the following activities using python code.

- Exercise 8: Sparse kernel machines

Hypothesis testing

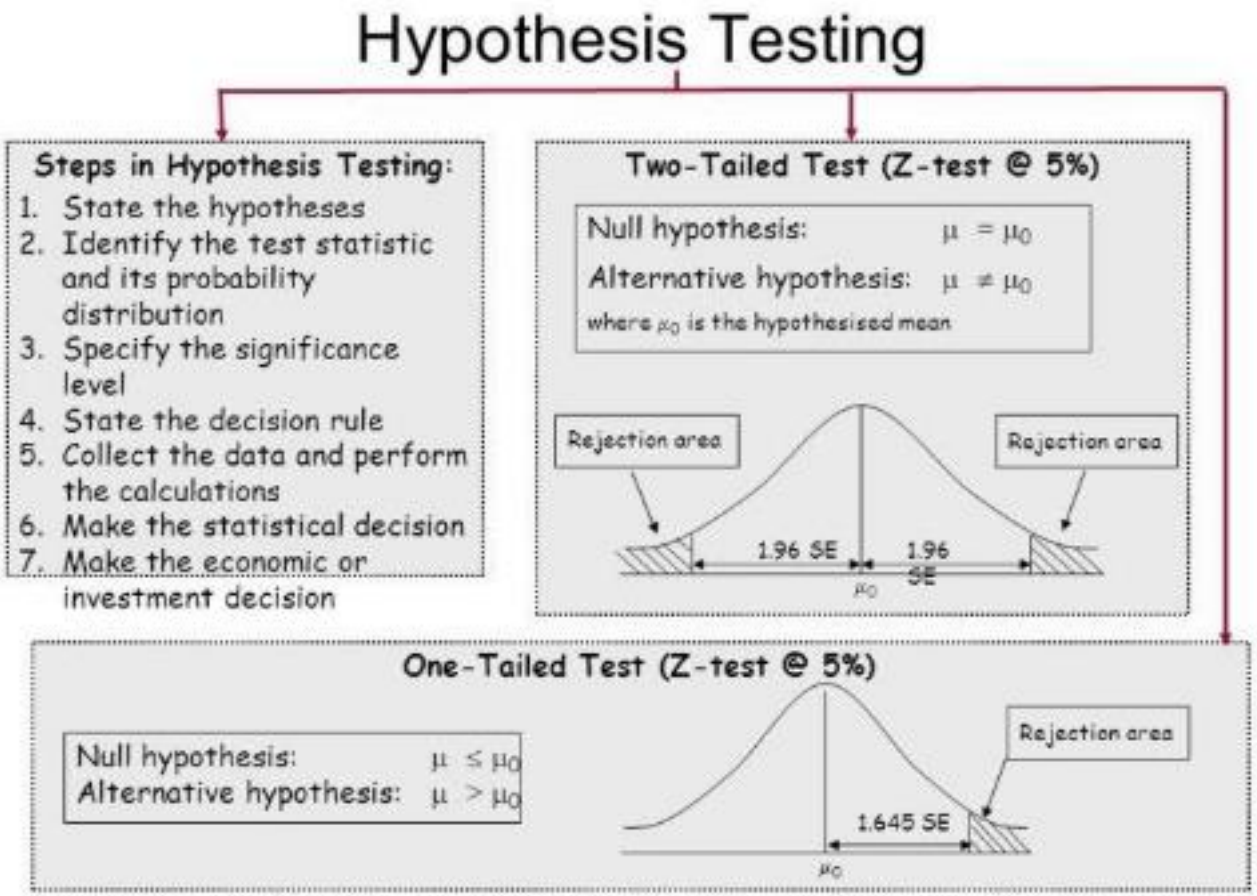


Figure: Hypothesis testing Process

Source: https://images.app.goo.gl/DcJE1FPHEm9xLeT89

Four steps of hypothesis testing

- The first step is to state the two assumptions so that only one hypothesis is right.
- The next step is to establish a method of analysis to analyze the data.
- The third stage is the implementation of the software and the realistic evaluation of the sample data.
- The fourth and final stage is to determine and dismiss the null hypothesis or to suggest that the null hypothesis is plausible if the information is given.

Real-world example of hypothesis testing

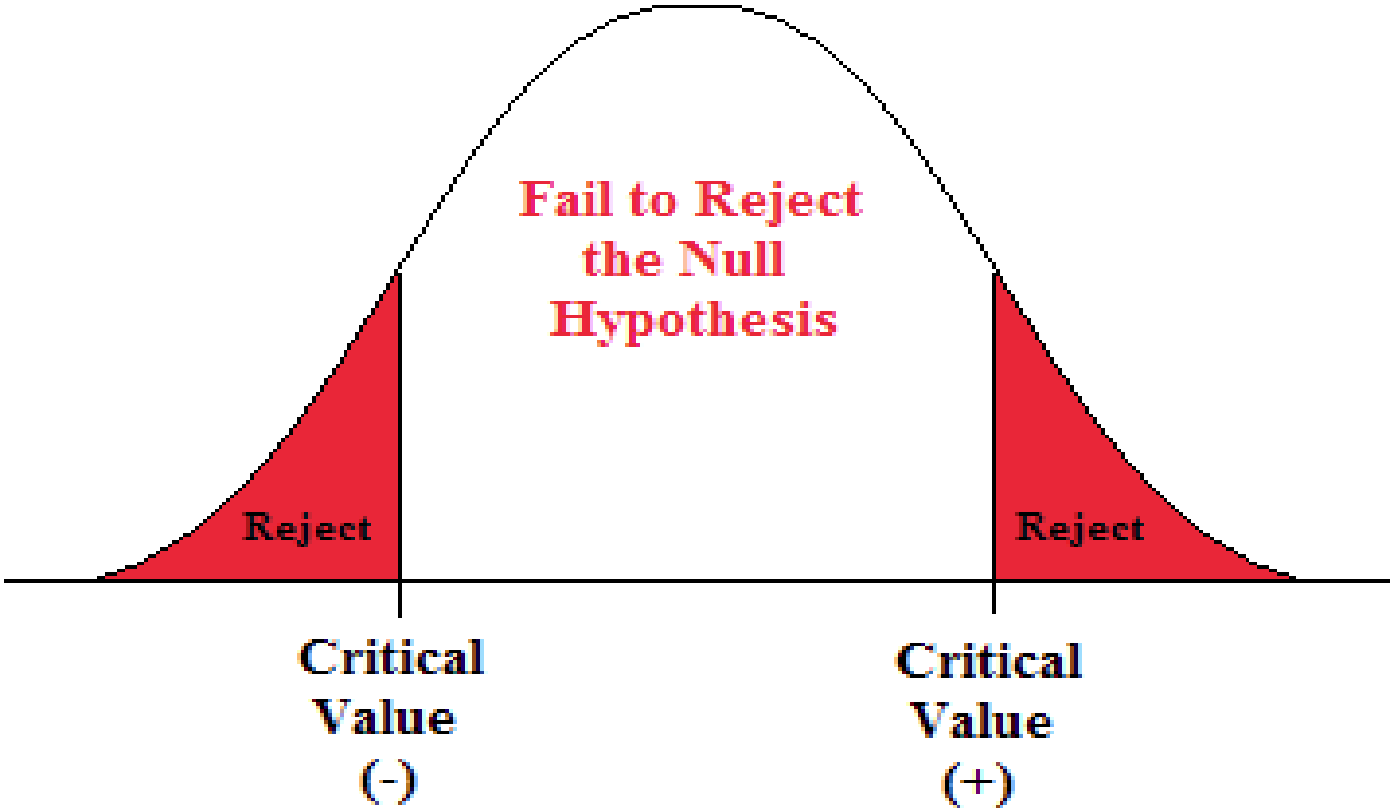


Figure: Beta Risk

Source: https://images.app.goo.gl/d8JcqwQ4Lws3oXN19

Bonferroni test

- A Bonferroni test is multiple forms of comparison used for statistical evaluation. Finally, a result can emerge from many hypothesis experiments with different variables, suggesting the dependent variable's statistical significance, although none exists.

Check of one-tailed

- A one-tail test is a statistical test in which a distribution's critical area is unilateral so that the value is either greater or lower than a certain value but not both.

- If the test sample falls into the critical unilateral zone, an alternative hypothesis rather than the null hypothesis is accepted.

Probability distributions

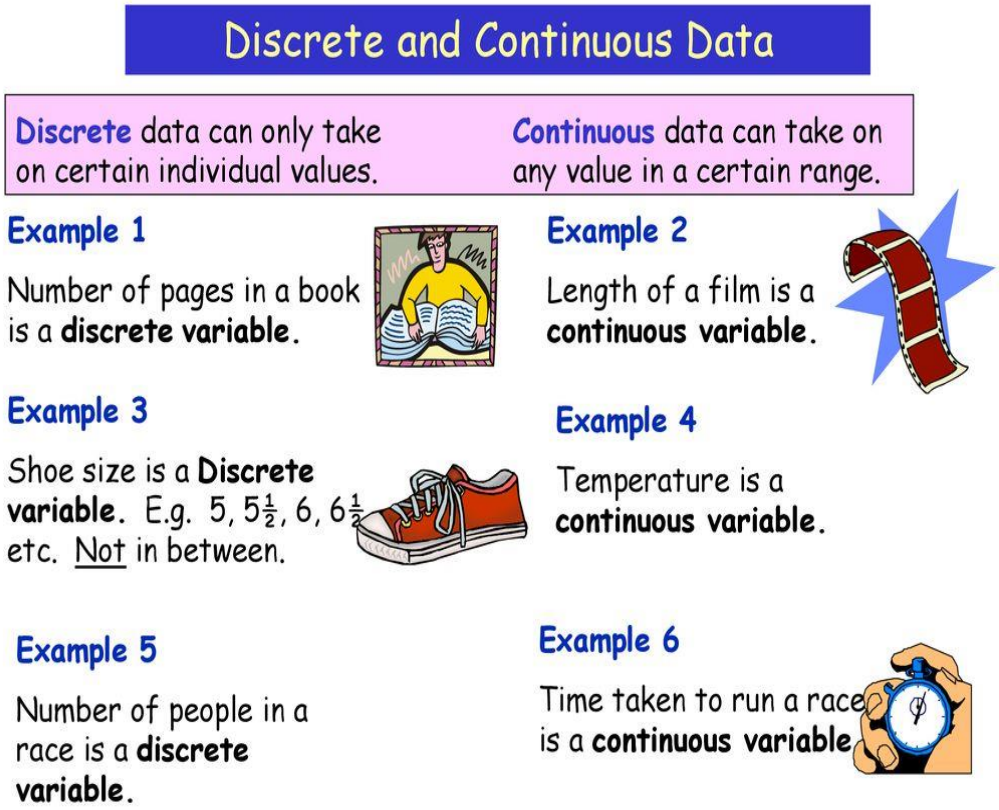


Figure: Discrete data and continuous data

Source: https://images.app.goo.gl/JdPuZvCNUxvNRV9S7

Self evaluation: Exercise 9

- To continue with the training, after learning the various steps involved in pattern recognition and anomaly detection, it is instructed to utilize the concepts to perform the following activity.

- You are instructed to write the following activities using python code.

- Exercise 9: Sampling methods for pattern recognition.

Types of distributions

- **Bernoulli distribution:**
  - Bernoulli equation defines events with precisely two real-life outcomes. Several illustrations of such activities are as follows: a team wins a tournament or not, a student passes or fails an assessment and a roll-out dice shows either a 6 or another number.
  - Only two possible tests, namely 1 (success) and 0 (failure) have a Bernoulli distribution and only one analysis.

- Therefore, a random X variable with a Bernoulli distribution will have value 1 at the probability of success (p), and value 0 at the probability of failure(q or 1-p).

- An incident of the head here is a success, and a tail occurrence is a deception. Having a head = 0.5 = probability to get a neck because only two of these scenarios are likely.

Regression models

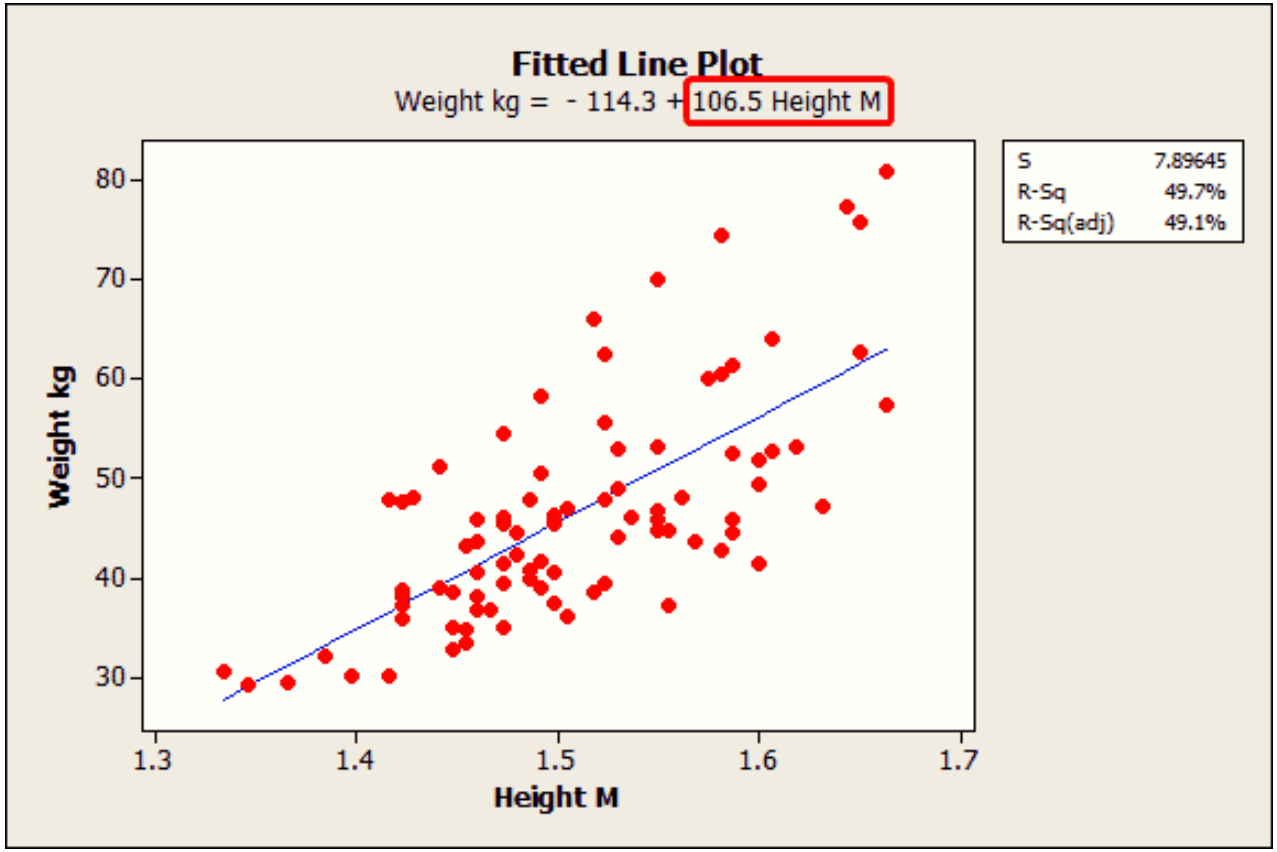


Figure: Regression Analysis

Source: https://images.app.goo.gl/yAcBF7tq1zwh4gWaA

Self evaluation: Exercise 10

- To continue with the training, after learning the various steps involved in pattern recognition and anomaly detection, it is instructed to utilize the concepts to perform the following activity.

- You are instructed to write the following activities using python code.

- Exercise 10: Decision tree.

Types of regression

Type of regression	Dependent variable and its nature	Independent variable and its nature	Relationship between variables
Simple linear	One, continuous, normally distributed	One, continuous, normally distributed	Linear
Multiple linear	One, continuous	Two or more, may be continuous or categorical	Linear
Logistic	One, binary	Two or more, may be continuous or categorical	Need not be linear
Polynomial (logistic) [multinomial]	Non-binary	Two or more, may be continuous or categorical	Need not be linear
Cox or proportional hazards regression	Time to an event	Two or more, may be continuous or categorical	Is rarely linear

Figure: Regression types

Source: https://images.app.goo.gl/59CM8MmjMP1sGPQJA

How to select the best model for regression?

- A research institute asks its students to perform linear regression-whether the outcome is constant.

- When you have a conditional regression logistics requirement! However, the more options you have, the easier it is to pick one.

Common questions

- How many regression types do we have?
- How much mathematical knowledge is required to understand regression?
- Ridge vs. Lasso Regression - what is the difference?
- Which types of problems can be solved using regression?
- What are the major challenges faced by regression techniques?
- Is Regression Analysis relevant in the industry?
- Which programming language works best for regression?

Linear models for classification

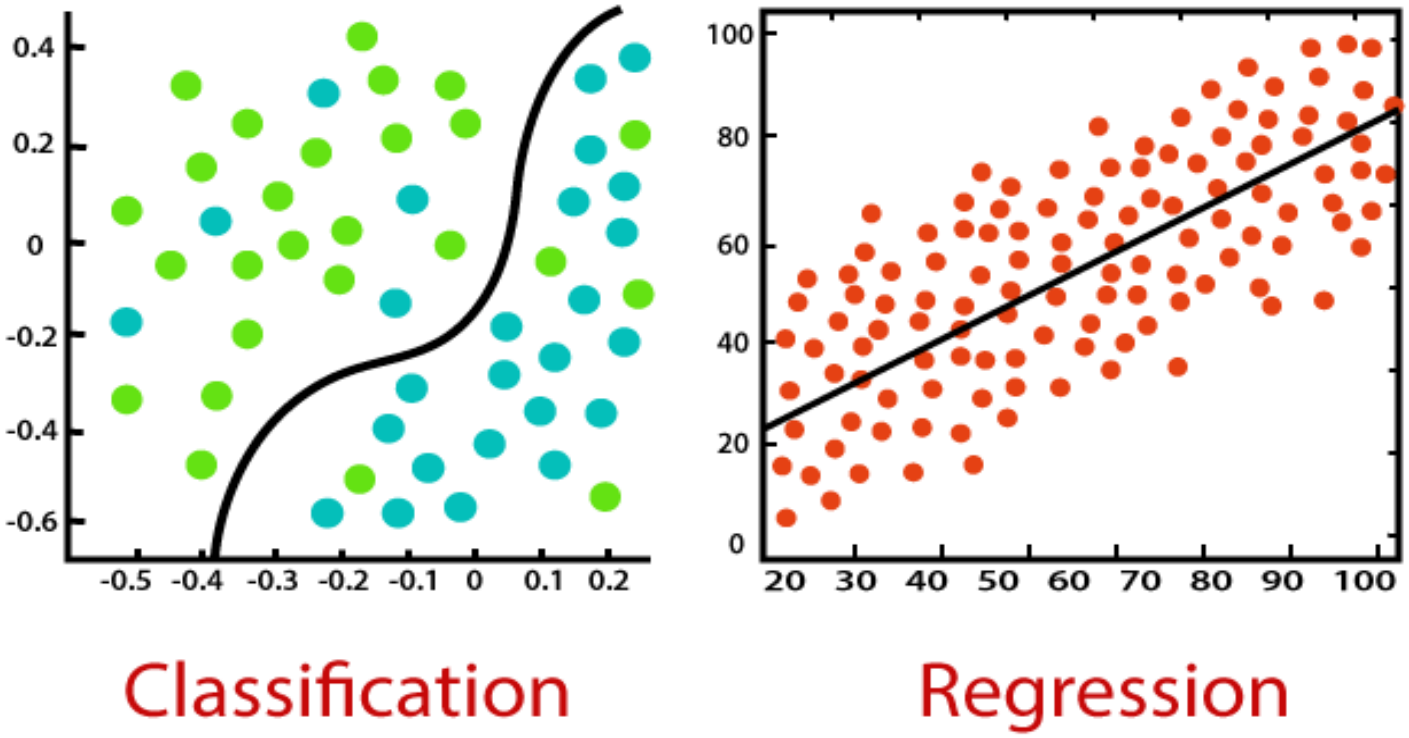


Figure: Classification and regression Model

Source: https://images.app.goo.gl/fzwieTpyn2DHXAwB9

Example of positive linear regression

- Price elasticity research.
- Risk evaluation in an insurance company.
- Sports analysis.

Checkpoint (1 of 2)

Multiple choice questions:

1. Memory decay affects what kind of memory?
   a) Short tem memory in general
   b) Older memory in general
   c) Can be short term or older
   d) None of the mentioned

2. How is pattern information distributed?
   a) It is distributed across the weights
   b) It is distributed in localized weights
   c) It is distributed in certain proactive weights only
   d) None of the above

3. What are the requirements of learning laws?
   a) Learning should be able to capture more & more patterns
   b) Convergence of weights
   c) All the mentioned
   d) None of the above

Checkpoint solutions (1 of 2)

Multiple choice questions:

1. Memory decay affects what kind of memory?
   - **a) Short tem memory in general**
   - b) Older memory in general
   - c) Can be short term or older
   - d) None of the mentioned

2. How is pattern information distributed?
   - **a) It is distributed across the weights**
   - b) It is distributed in localized weights
   - c) It is distributed in certain proactive weights only
   - d) None of the above

3. What are the requirements of learning laws?
   - a) Learning should be able to capture more & more patterns
   - b) Convergence of weights
   - **c) All the mentioned**
   - d) None of the above

Checkpoint (2 of 2)

Fill in the blanks:

1. _____factors affect the performance of learner system does not include?
2. In language understanding, the levels of knowledge that does not include ___.
3. _____consists of the categories which does not include structural units.
4. A search algorithm takes _____as an input and returns ____as an output.

True or False:

1. In pattern mapping problem in neural nets, is there any kind of generalization involved between input & output? True/False
2. Linear neurons can be useful for application such as interpolation, is it true? True/False
3. Does pattern classification & grouping involve same kind of learning? True/False

Checkpoint solutions (2 of 2)

Fill in the blanks:

1. **Good data structures** factors affect the performance of learner system does not include?
2. In language understanding, the levels of knowledge that does not include **Empirical**.
3. **A model of language** consists of the categories which does not include structural units.
4. A search algorithm takes **problem** as an input and returns **solution** as an output.

True or False:

1. In pattern mapping problem in neural nets, is there any kind of generalization involved between input & output? **True**
2. Linear neurons can be useful for application such as interpolation, is it true? **True**
3. Does pattern classification & grouping involve same kind of learning? **False**

Question bank

Two mark questions:

1. What is probability distributions ?
2. What are the components of probability distributions ?
3. List any 3 types linear models for regression.
4. What is y=mx+c formula for linear regression?

Four mark questions:

1. What is multiple regression model?
2. Describe r-squared method.
3. Describe classification techniques.
4. Describe any 3 types of classification methods.

Eight mark questions:

1. Explain linear models for classification.
2. Explain probability distributions with chart.

Unit summary

Having completed this unit, you should be able to:

- Understand the concept of probability distributions
- Gain knowledge on example of statistical approaches
- Understand linear models for regression
- Learn about linear models for classification