Statistics for Machine Leaning

NASSMA 2019

Saad Benjelloun

June 27, 2019

Objective

Give the outlines of statistical testing techniques, in the context of ML algorithms.

- What is a statistical test ? What is it used for ?
- How a statistical test is conducted ?
- What are the outputs of a statistical test ? How to interpret them ?

We will give examples in the context of ML algorithms.

Outline : statistics for machine learning

- Reminder about random variables, probability distributions...
- Introduction to Statistical Tests
  - The general philosophy.
  - The methodology.
- T-tests.
- Statistical Tests for linear regression.

Random variables and probability densities

A Random Variable (RV) is a mathematical object/framework to model the outcome of a function with uncertainties.

- Example : when rolling a dice, the top face is a random variable $\in \{1, 2, 3, 4, 5, 6\}$.

- Other examples : the number of connection to your website, arrival time of your train/flight...
- RVs are a way to model what you cannot control, but still want to quantify the uncertainties on it.
- Generally, a RV is either discrete or continuous (categorical or numerical variable/feature).

- A RV is either discrete or continuous (categorical or numerical variable/feature).
- The uncertainties on random variables are quantified through a Probability Distribution.
- A probability is a value between 0 and 1. 0 for impossible events, and 1 for certain events.
  - Example : when rolling a dice, the probability of each face $\in \{1, 2, 3, 4, 5, 6\}$ is $1/6$.

Random variables and probability densities

Example of probability densities

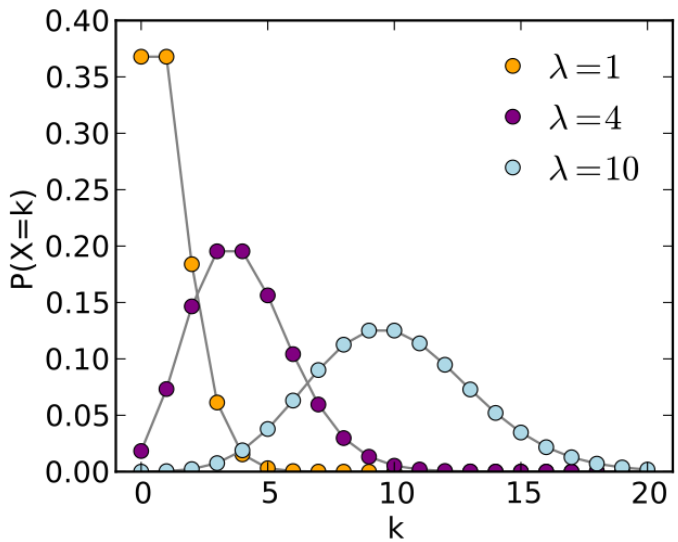
- Uncertainty on **discrete** random variables is modeled by **discrete** probability densities/probability distribution (P.D.).
- Uncertainty on **continuous** random variables is modeled by **continuous** probability densities/probability distribution.

Random variables and probability densities

Examples of discrete probability densities

A discrete P.D. gives the probability P of each possible value.

- Uniform distribution : rolling a dice, $P(i) = 1/6, i \in 1, 2, ., 6$.
- Bernoulli distribution : $\mathbb{P}(0) = p$, $P(1) = 1 - p$.
- Poisson : $P(k) = e^{-\lambda} \frac{\lambda^k}{k!}$ $k \in 0, 1, 2, 3, ...$

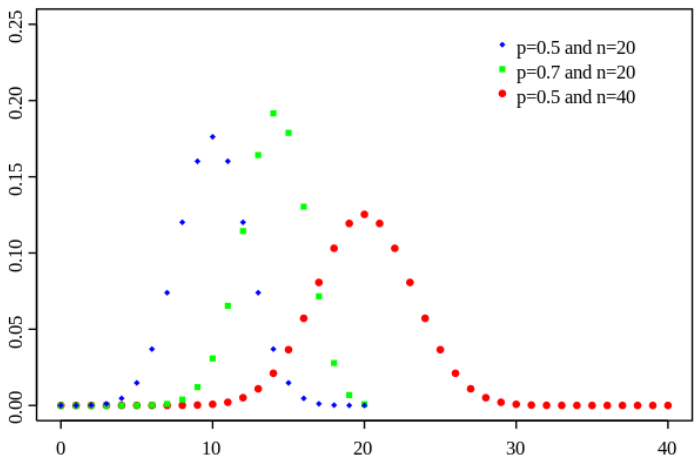


Random variables and probability densities

Examples of discrete probability densities

A discrete P.D. gives the probability P of each possible value.

- Uniform distribution : rolling a dice, $P(i) = 1/6, i \in 1, 2, ., 6$.
- Bernoulli distribution : P(0) = p, P(1) = 1-p.
- Poisson : $P(k) = e^{-\lambda} \frac{\lambda^k}{k!}$ $k \in 0, 1, 2, 3, ...$
- Binomial : $B(k) = \binom{N}{k} p^k (1-p)^{N-k}, k \in 0, 1, 2, .., N$



Random variables and probability densities

Examples of discrete probability densities

A discrete P.D. gives the probability P of each possible value.

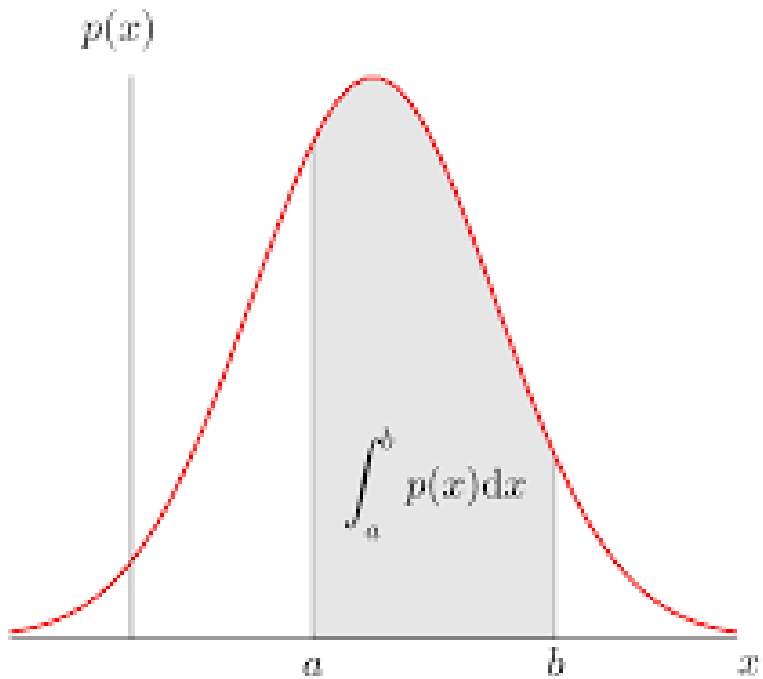
- Uniform distribution : rolling a dice, $P(i) = 1/6, i \in 1, 2, ., 6$.
- Bernoulli distribution : P(0) = p, P(1) = 1-p.
- Poisson : $P(k) = e^{-\lambda} \frac{\lambda^k}{k!}$ $k \in 0, 1, 2, 3, ...$
- Binomial. $B(k) = \binom{N}{k} p^k (1-p)^{N-k}, k \in 0, 1, 2, .., N$
- Other examples : Negative binomial, ... etc

Random variables and probability densities

Examples of continuous probability densities

A continues P.D. p gives the probability of each possible range of values as the integral over that range :

$$P([a,b]) = \int_a^b p(x)dx$$



Random variables and probability densities

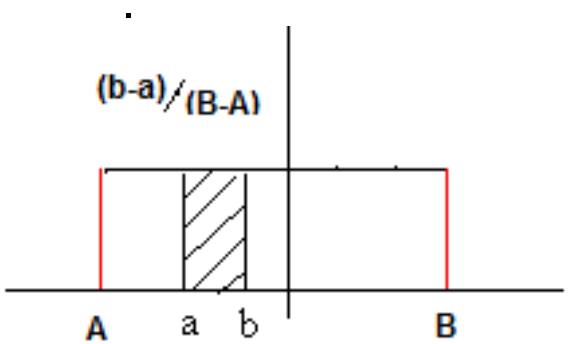
Examples of continuous probability densities

A continues P.D. p gives the probability of each possible range of values as the integral over that range :

$$P([a,b]) = \int_a^b p(x)dx$$

- Uniform distribution over [A,B]: $p(x) = \frac{1}{B-A}$.

$$P([a,b]) = \frac{b-a}{B-A}, \text{ if } a,b \in [A,B]$$



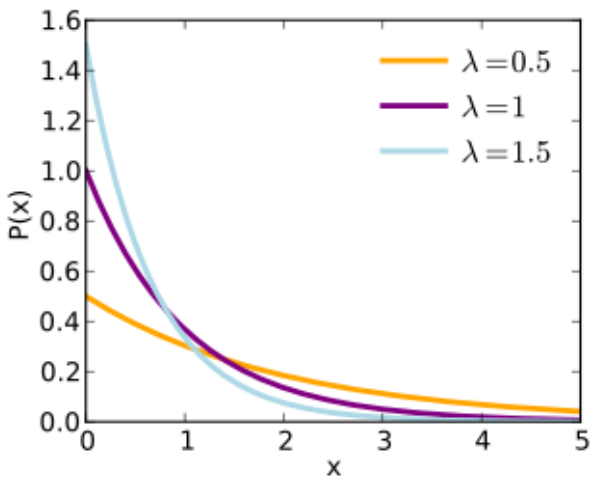
Random variables and probability densities

Examples of continuous probability densities

A continues P.D. p gives the probability of each possible range of values as the integral over that range :

$$P([a,b]) = \int_a^b p(x)dx$$

- Uniform distribution over [A,B]: $p(x) = \frac{1}{B-A}$.
- Exponential distribution : $p(x) = \lambda e^{-\lambda x}$.



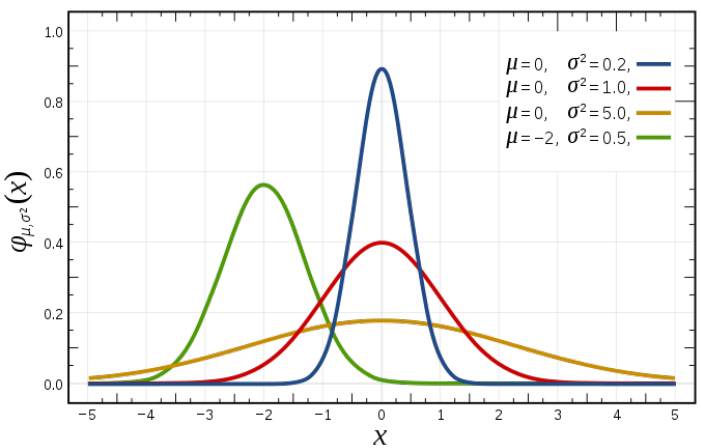
Random variables and probability densities

Examples of continuous probability densities

A continues P.D. p gives the probability of each possible range of values as the integral over that range :

$$P([a,b]) = \int_a^b p(x)dx$$

- Uniform distribution over [A,B]: $p(x) = \frac{1}{B-A}$.
- Exponential distribution : $p(x) = \lambda e^{-\lambda x}$.
- Normal distribution : $p(x) = \phi_{\mu,\sigma}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$



Random variables and probability densities

Examples of continuous probability densities

A continues P.D. p gives the probability of each possible range of values as the integral over that range :

$$P([a,b]) = \int_a^b p(x)dx$$

- Uniform distribution over $[A,B]$: $p(x) = \frac{1}{B-A}$.
- Exponential distribution : $p(x) = \lambda e^{-\lambda x}$.
- Normal distribution : $p(x) = \phi_{\mu,\sigma}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$.
- Other examples : Chi2, Student, Fisher ... etc

Introduction to statistical tests

Introduction to statistical testing

Introduction to statistical tests

Hypothesis

A hypothesis is any affirmation regarding (the probability distribution of) a Random Variable.

- When rolling a dice, the probability of each face $\in \{1, 2, 3, 4, 5, 6\}$ is $1/6$.

Introduction to statistical tests

Statistical hypothesis

A hypothesis is any affirmation regarding the probability distribution of a Random Variable.

- When rolling a fair dice, the probability of each face $\in \{1, 2, 3, 4, 5, 6\}$ is $1/6$.
- Other examples :
- The probability of defect in the production is 10%.
- The human height is normally distributed ?
- Two algorithms A and B has the same performance over the test-cases, or over last months of production. Algorithms C is better.

Introduction to statistical tests

Statistical tests

We have a Hypothesis that we call H_0.

- Can we adopt H_0 or do we have a good reason to reject hypothesis H_0 ?
- Is the accused person innocent or do we have a good reason to condemn him ?
- Dice example : Your competitor launched a dice 10 times, and got the sequence $[6, 5, 6, 6, 6, 6, 6, 5, 6, 6]$. Do you think he has a fair dice ?
- What if he got this : $[6, 5, 6, 6, 4, 6, 1, 5, 6, 6]$?
- The idea is to evaluate the likelihood of the observed data under the hypothesis H_0.

Introduction to statistical tests

Statistical tests

- "The idea is to evaluate the likelihood of the data under the hypothesis H_0"
- "If this likelihood is (very) small, we better reject H_0".
- We summarize the whole data into a single variable called the Test Statistic.
- For the dice example (N trials), a possible test statistic is :

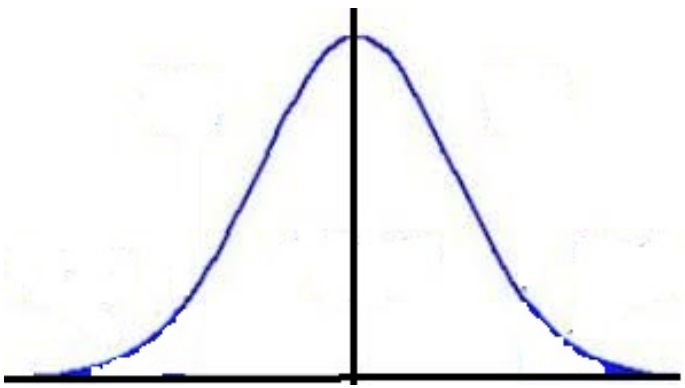
$$D = \sum_{i=1}^{6} \frac{(n_i - N/6)^2}{N/6}$$

  n_i the number of time you got the value i.
- D is called the Pearson statistics.

Introduction to statistical tests

Statistical tests

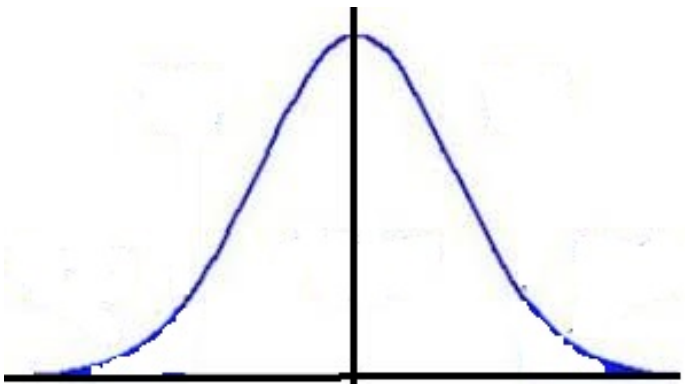
- The Test Statistics is a function of the data.
- We have to know the probability density f of the Test Statistics under the hypothesis H0.



Introduction to statistical tests

Statistical tests

- The Test Statistics is a function of the data.
- We have to know at least approximately = asymptotically the probability density f of the Test Statistics under the hypothesis H0.

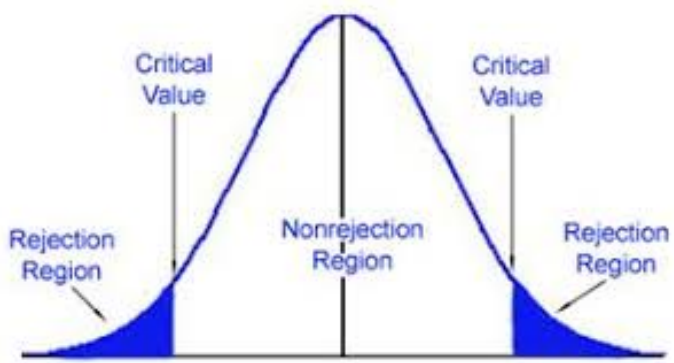


Introduction to statistical tests

Statistical tests

"If the likelihood under H_0, of the observed value of the test statistic is (very) small, we better reject H_0".

- A test defines a critical region W of the test statistic values.
- To this region corresponds low values of the P.D. f under H_0.
- If the value of the test statistic falls inside the critical region, the null hypothesis is rejected.
- If the value of the test statistic falls outside the critical region, then there is not enough evidence to reject the null hypothesis.

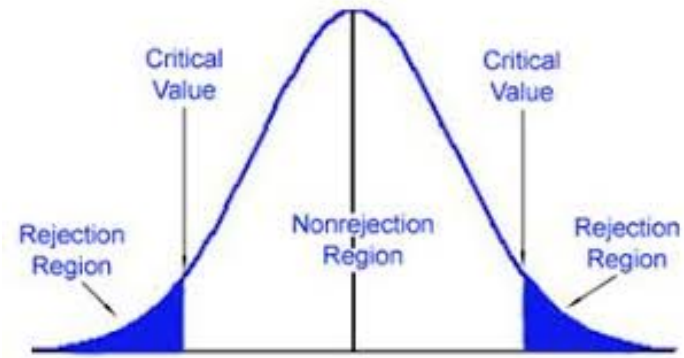


Introduction to statistical tests

Conducting a statistical test

- The probability associated to the critical region is called significance level and is noted α.

$$\alpha = P(T \in W/H_0) = \int_W f(x)dx$$



- α is the risk you want to take of rejecting H_0 while it is true in the reality.
- We always start by setting a value of α and compute the corresponding critical region.

Introduction to statistical tests

Conducting a statistical test

- You have your data and you want to conduct a test T for hypothesis H_0.
- You set your significance level α (5%?).
- You compute the corresponding critical region.
- You calculate the value of your statistical test.
- If the value of the test statistic falls inside the critical region, the null hypothesis is rejected.
- If the value of the test statistic falls outside the critical region, then there is not enough evidence to reject the null hypothesis.

Introduction to statistical tests

Example statistical test : χ^2 test

- You roll a dice 30 times and you obtain the following results :

Face	Count
1	3
2	7
3	5
4	10
5	2
6	3

- The Pearson test statistic is :

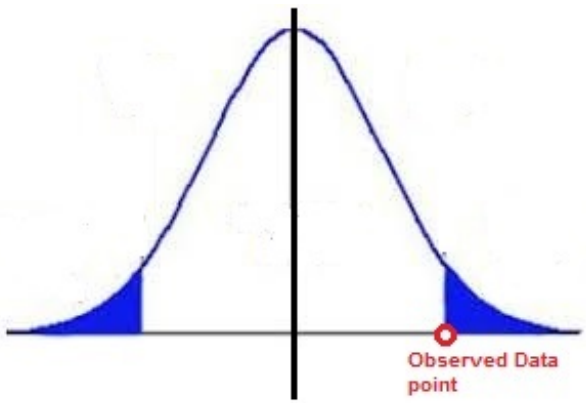
$$d = \sum_{1}^{6} \frac{(n_i - N/6)^2}{N/6} = 9.2$$

- and it asymptotically has a χ^2_5 probability density.
- for $\alpha = 0.5$, $W = [11,07, +\infty]$

Introduction to statistical tests

Another way of conducting a statistical test

- Another way of conducting a statistical tests is through the p-value of the observed test statistic.
- The p-value is the probability of all the events that are more extreme than the observed one.
- The p-value is the probability the critical region "starting" at observed value.

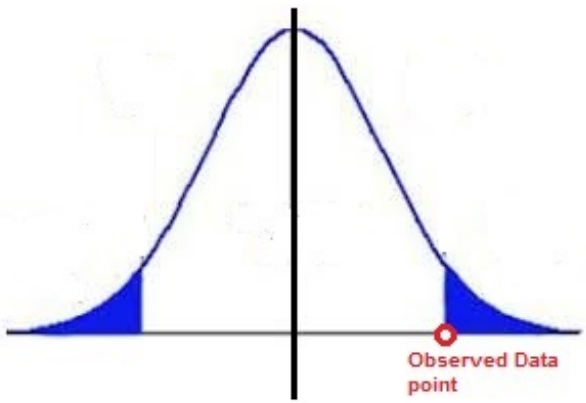


Introduction to statistical tests

Another way of conducting a statistical test

- The p-value is the probability of all the events that are more extreme than the observed one.
- The p-value is the probability the critical region W_t "starting" at observed value t.

$$p\_value = \int_{W_t} f \ dx$$



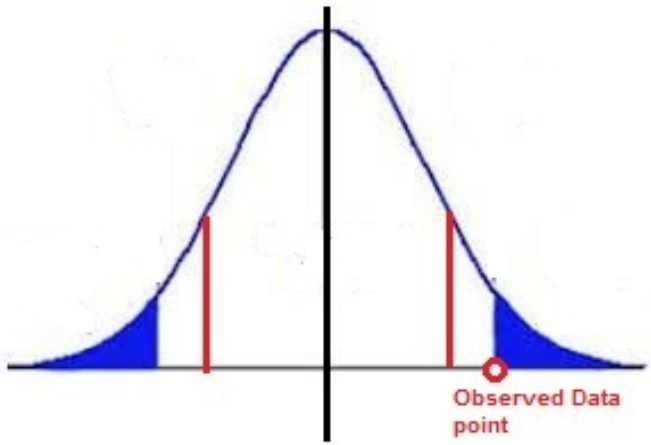
Introduction to statistical tests

Another way of conducting a statistical test

- The p-value is either smaller or bigger than the wished significance level α.

$$p\_value = \int_{W_t} f \; dx, \quad \alpha = \int_{W_c} f \; dx$$

Here we reject H_0 at significance level α.

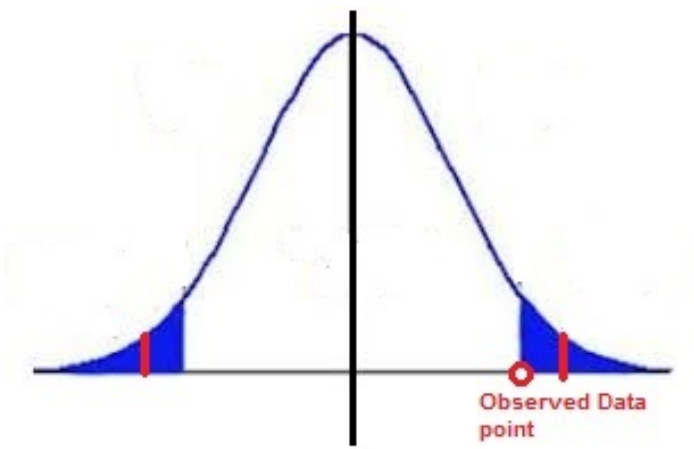


Introduction to statistical tests

Another way of conducting a statistical test

- The p-value is either smaller or bigger than the wished significance level α.

$$p\_value = \int_{W_t} f \; dx, \quad \alpha = \int_{W_c} f \; dx$$



Here we cannot reject H_0 at the significance level α.

Introduction to statistical tests

Conducting a statistical test

- You have your data and you want to conduct a test T for hypothesis H_0.
- You decide what is your significance level α.
- You calculate the value of your statistical test.
- You calculate the p-value of your observed test statistic.
- If the p-value is less than α the null hypothesis is rejected.
- Else If the p-value is larger than α you cannot reject the null hypothesis.

Introduction to statistical tests

Example statistical test : χ^2 test

- You roll a dice 30 times and you obtain the following results :

Face	Count
1	3
2	7
3	5
4	10
5	2
6	3

- The Pearson test statistic is :

$$D = \sum_{1}^{6} \frac{(n_i - N/6)^2}{N/6} = 9.2$$

- and it asymptotically has a χ^2_5 probability density.
- The p-value of 9.2 is 0.1.

Statistical tests for ML

T-tests : One-sample t-test

- One sample $(x_1, x_2..x_n)$ from a population.
- One-sample t-test:

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

H0 : The mean of X over the population is μ.

$$T \propto T(n-1).$$

e.g. H0 : The mean delay/advance of production (or shuttle arrival ?!) is zero.

Statistical tests for ML

T-tests : Unpaired samples t-test

- Two separate sets of independent samples :
  $(x_1, x_2..x_{nA})$, $(y_1, y_2..y_{nB})$.
  H_0 : The two sub-groups of the population have the same mean.
- Unpaired samples t-test :

$$t = \frac{m_A - m_B}{\sqrt{\frac{S^2}{n_A} + \frac{S^2}{n_B}}}$$

$$S^2 = \frac{\sum (x - m_A)^2 + \sum (x - m_B)^2}{n_A + n_B - 2}$$

$$T \propto T(n_A + n_B - 2)$$

- E.g. : H0 : Life expectancy of women is not significantly different life expectancy of men.

Statistical tests for ML

T-tests: paired samples t-test

- Two separate sets of n-paired samples :
  $(x_1, x_2..x_n)$, $(y_1, y_2..y_n)$.
  E.g. : weight of each person before and after a medical treatment. H_0 : The two samples have the same mean.
- Paired-samples t-test

$$t = \frac{m}{s/\sqrt{n}}$$

- Where we have computed the differences serie $d_i = y_i - x_i$. m and s are the empirical mean and s.d. of teh serie d.

$$T \propto T(n).$$

E.g. H0 : The weight before and after the medical treatment is significally the same.

Statistical tests for ML

Linear regression

Tests for linear regression

Statistical tests for ML

Linear regression

- When conducting a linear regression between X and Y.
- Data $= \{(x_1, y_1), (x_2, y_2), (x_3, y_3)..,(x_N, y_N)\}$.
- Your model is

$$Y = \alpha + \beta X + \epsilon$$

.

- Testing the significance of the linear regression, means testing if we have enough evidence against the null hypothesis :

$$\beta = 0$$

.

- The statistics $(N-2)\frac{R^2}{1-R^2}$ has a $F(1, N-2)$ as PD. (Fisher).
- In python, using Scikit-learn, you get the p-value with :

lm.pvalues

Statistical tests for ML

Multi-variate Linear regression

- When conducting a linear regression between Y and X^1, X^2... X^n.
- Data $= \{(x_1^1, x_1^2.., y_1), (x_2^1; .., y_2), (x_3^1, ..y_3).., (x_N^1, y_N)\}$.
- Your model is

$$Y = \beta_0 + \sum \beta_i X_i + \epsilon$$

.

- Testing the significance of a feature X_i, means testing if we have enough evidence against the null hypothesis :

$$\beta_i = 0$$

.

- A partial Fisher test generalizes the previous test.

Statistical tests for ML

To go further

- Comparing machine learning methods and selecting a final model is a common operation in ML.
- Resampling methods like k-fold cross-validation and comparing a mean score can be misleading as result of a statistical fluke.
- Statistical significance tests are designed to address this problem.

Statistical tests for ML

To go further

- We test the null hypothesis : samples of skill scores being observed are drawn from the same distribution.
- The commonly used tests are
  - McNemars test.
  - 5 X 2 cross-validation paired t-test.

  Reference : T. G. Dietterich, "Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms," in Neural Computation, vol. 10, no. 7, pp. 1895-1923, 1 Oct. 1998.

Statistical tests for ML

Thank you.