Introduction

- Concept of comparison of quantities:
  - + Do AVE trips from Madrid to Seville cost > 60 € ?
  - + Do people who sleep more have better health ?
  - + Is effective a certain treatment against some illness?

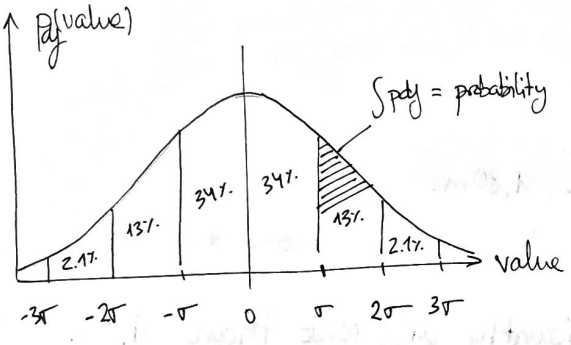
Key concepts to understand previously

Population → All existing elements in a dataset. All spaniards. It is not possible to know the truth of a population without examining all its elements.

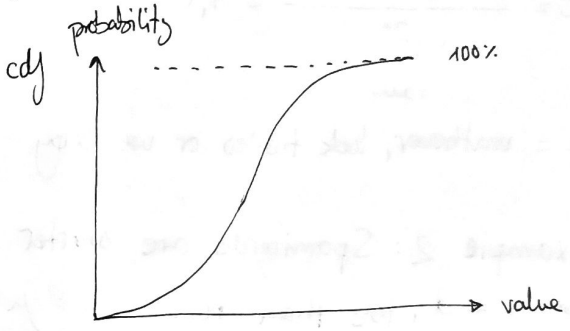
Sample → Subset of a population usually selected randomly by a statistician to infer some information about.

Distribution → mathematical function that provides the probabilities of occurence of different outcomes of an experiment: pdf (probability density function), and cdf (cumulative distribution function), for continuous distributions.

pdf



cdf



Normal Distribution, T Student Distribution →

- Standarization (normal distribution): $z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$
  - $\bar{x}$ → mean of sample
  - μ → mean of distribution
  - σ → standard deviation of distribution
  - n → number of elements in sample
- Standarization (T student): $t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$
  - s → estimated standard deviation from sample

‖ z & t have no dimensions ‖

Degrees of freedom

Number of decisions that can be made while computing a statistic.

- computing a sample mean: n observations - 1
- choosing hat every weekday: all weekdays - 1

Types of Hypothesis tests

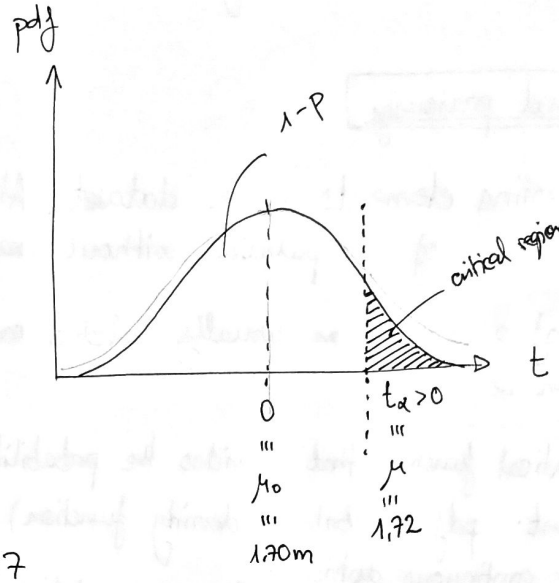
+ One sample vs constant : now
+ Two or more samples to each other ← Related samples, Independent samples, ANOVA (>2 samples) } after

Example 1: Spaniards are taller than 1,70 m.
One sided, greater than test

a) Get a sample of Spaniards, imagine $\bar{x} = 1,72$ with a sample size of $n = 5000$ and $s = 0,2$

$H_0 : \mu \leq \mu_0$ → "Spaniards are not taller than 1,70m"

$H_1 : \mu > \mu_0$



b) $$t = \frac{1,72 - 1,70}{\frac{0,20}{\sqrt{5000}}} = 7,07$$

p = wathever, look tables or use scipy

α test significance, is a threshold we set, p obtained is the probability of a result obtained by chance.

If that p is less than 5%, test is successful and null hypothesis H_0 is rejected: $t > 0$ is mandatory as well as $p < \alpha$

Example 2: Spaniards are shorter than 1,80m.
One sided, less than test. $H_0 : \mu \geq \mu_0$, $H_1 : \mu < \mu_0$, idem but with $t < 0$.

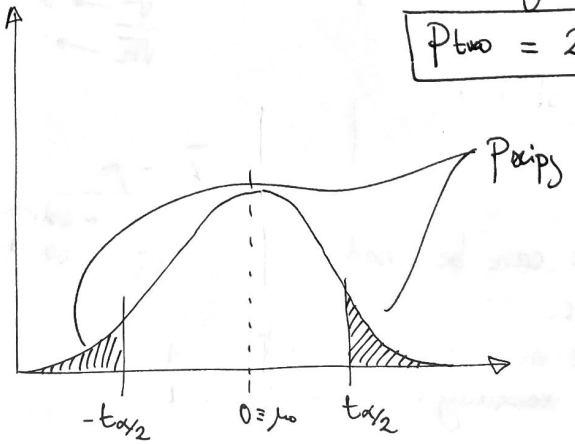
Example 3: Spaniards' Height is significantly different than 1,80.
Two sided test

$H_0 : \mu = \mu_0$

$H_1 : \mu \neq \mu_0$

+ In scipy, all test are two sided by default:

$$P_{two} = 2 \cdot P_{one}$$



+ Confidence interval:

$$\bar{x} \pm t_{\alpha/2} * \frac{s}{\sqrt{n}}$$

contains population mean with α% confidence (95% for example).

Test assumptions, must be true to be valid

- Observations must be independent of each other in the sample.
- Data distribution is normal.
- Sample size is at least 30.
- For z test, σ is known (never happens!), otherwise use t test.