Departamento de Ciencia de la Computación (DC-UCSP)

Maestría en Ciencia de Datos

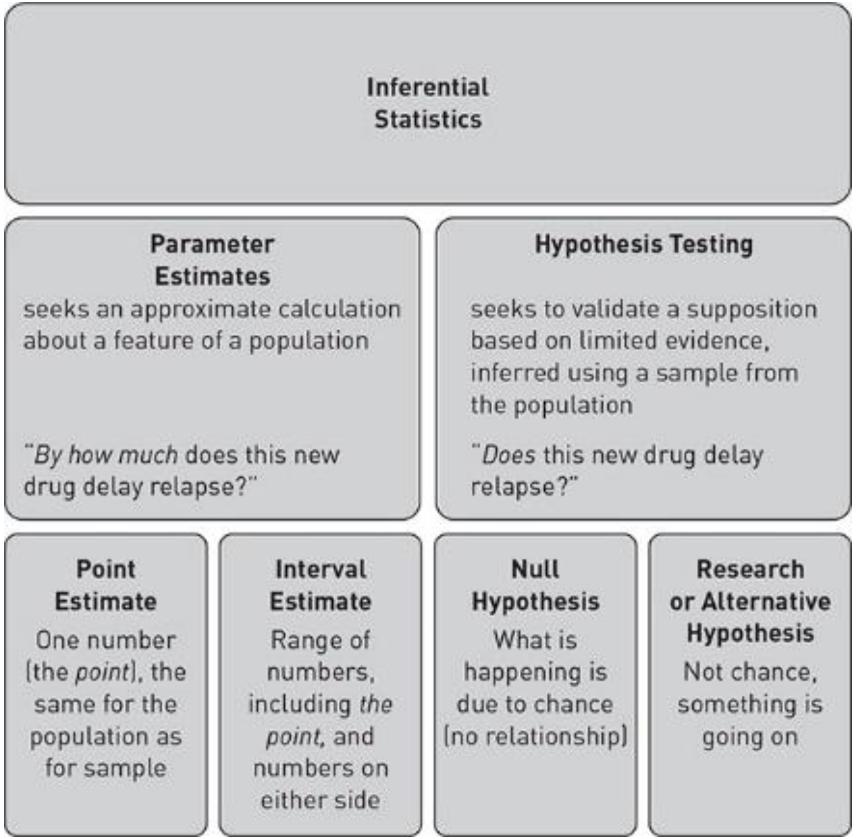
Inferential Statistics

Daniel Alexis Gutierrez Pachas, PhD
dgutierrezp@ucsp.edu.pe

Universidad Católica San Pablo | Departamento de Ciencia de la Computación

Inferential Statistics

Inferential statistics is used to analyse the results and draw conclusions. Experts described inferential statistics as the mathematics and logic of how this **generalization from sample to population**.



Inferential Statistics

Parameter Estimates
seeks an approximate calculation about a feature of a population

"*By how much* does this new drug delay relapse?"

Hypothesis Testing

seeks to validate a supposition based on limited evidence, inferred using a sample from the population

"*Does* this new drug delay relapse?"

Point Estimate
One number (the *point*), the same for the population as for sample

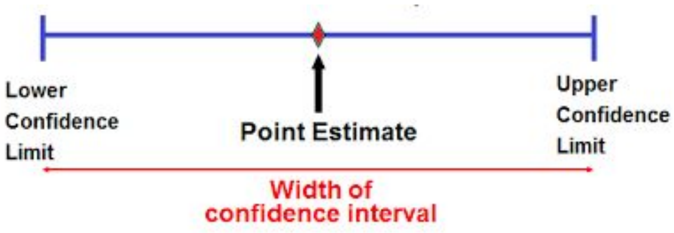
Interval Estimate
Range of numbers, including *the point*, and numbers on either side

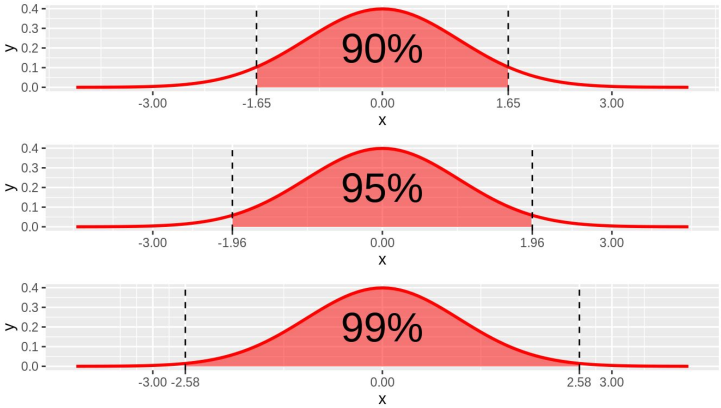
Null Hypothesis
What is happening is due to chance (no relationship)

Research or Alternative Hypothesis
Not chance, something is going on

Parameter Estimation is a branch of statistics that involves using sample data to estimate the parameters of a distribution. Point estimation involves the use of sample data to calculate a single value which is to serve as a “best guess” of an unknown population (for example, the population mean).

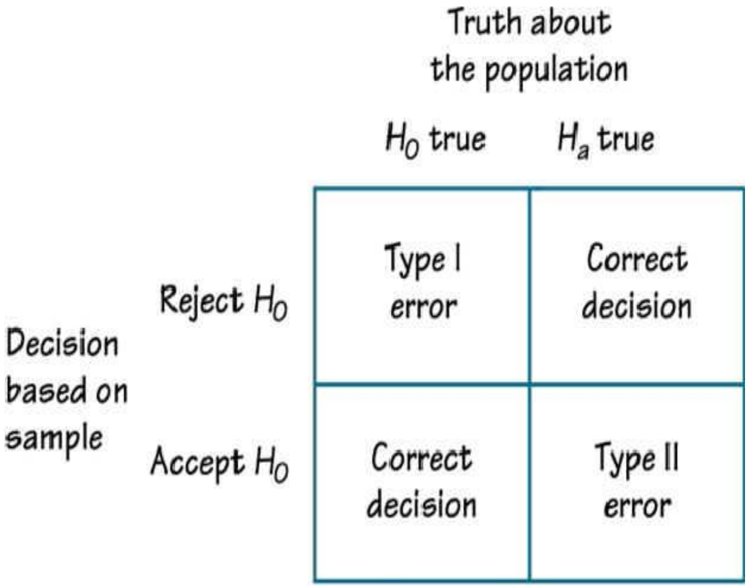
Interval Estimation is the use of sample data to estimate an interval of plausible values of a parameter of interest. Also, the interval is kown as **Confidence Interval**.





Hypothesis Testing

The process of distinguishing between the null hypothesis and the alternative hypothesis is aided by considering two conceptual types of errors. The first type of error occurs when the null hypothesis is wrongly rejected. The second type of error occurs when the null hypothesis is wrongly not rejected. The two types are known as Type I and Type II errors.



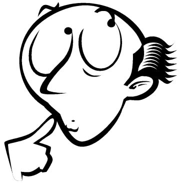
		Truth about the population	
		H_0 true	H_a true
Decision based on sample	Reject H_0	Type I error	Correct decision
	Accept H_0	Correct decision	Type II error

If p < α we reject the null Hypothesis. Otherwise, we accept this. Usually, α=0.05.

Significance test

Up to now, we have only considered two variables. However, in many situations, there might be more than one covariate which affects Y and consequently all of them are relevant to the analysis. We define a multiple linear model as follows:

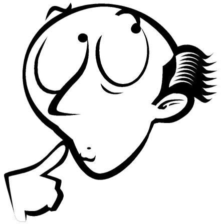
$$Y = \beta_0 + \beta_1 X_1 + \ \dots + \beta_n X_n + e, \quad e \sim N(0, \sigma^2 I)$$



$$\begin{cases} \mathrm{H}_0 : \ \beta_i = 0. \\ \mathrm{H}_a : \ \beta_i \neq 0. \end{cases}$$

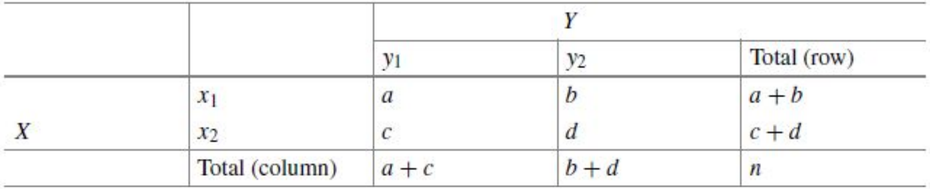
Normality Test

$$Y = \beta_0 + \beta_1 X_1 + \ \ldots + \beta_n X_n + e, \quad e \sim N(0, \sigma^2 I)$$



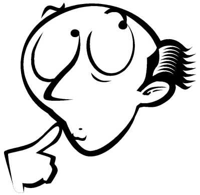
$$\begin{cases} H_0 : & \text{The residual (e) has normal distribution.} \\ H_a : & \text{The residual (e) has no normal distribution.} \end{cases}$$

Independence Test



		Y		
		y_1	y_2	Total (row)
X	x_1	a	b	$a+b$
	x_2	c	d	$c+d$
	Total (column)	$a+c$	$b+d$	n

$$\chi^2 = \frac{n(ad - bc)^2}{(a+b)(c+d)(a+c)(b+d)}$$

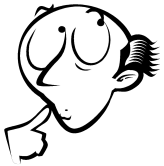


$$\begin{cases} H_0 : \text{X and Y are independents.} \\ H_a : \text{X and Y are dependents.} \end{cases}$$

The idea behind the Pearson coefficient is that when the relationship between two variables is stronger, the deviations between observed and expected frequencies are expected to be higher (because the expected frequencies assume independence), indicating a stronger relationship between the two variables. If observed and expected frequencies are identical or similar, then this suggests that the association between the two variables is weak, and the variables may even be independent.

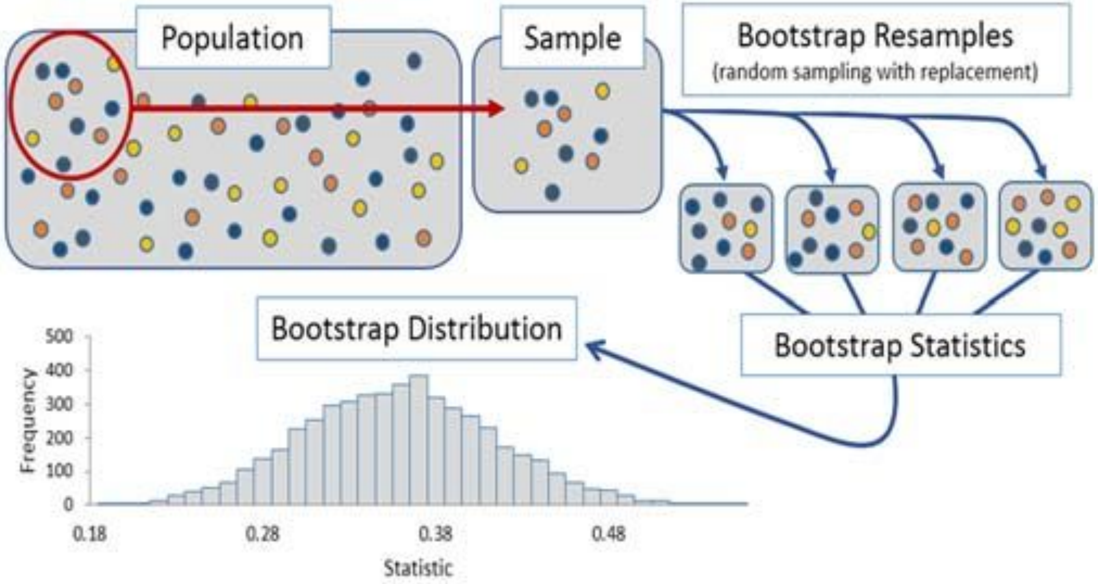
Anova Test

When performing regression analysis is the analysis of variance (ANOVA) table. This table can have several meanings and interpretations and may look slightly different depending on the context. We apply Anova test to compare k means.



$$\begin{cases} H_0 : \text{There are no differences between the means of the different groups } (\mu_1=\mu_2=\ldots=\mu_k=\mu). \\ H_a : \text{At least one pair of means are significantly different from each other} \end{cases}$$

The **bootstrapping method** is based on generating new pseudo-samples of the same size as the original sample, through repeated sampling of the available data. If the original is representative of the population, the distribution of the statistic calculated from the pseudo-samples is close to the sampling distribution that would be obtained if the population could be accessed to generate new samples.



Population

Sample

Bootstrap Resamples
(random sampling with replacement)

Bootstrap Statistics

Bootstrap Distribution