An intuitive approach for the teaching of Statistics

Susana MONTES

http://unimode.uniovi.es
University of Oviedo, Spain

Novi Sad - December 13, 2011

Overview

Overview

Introduction

- The teaching of Statistics has changed a lot in Spain for the last twenty years.

Introduction

- The teaching of Statistics has changed a lot in Spain for the last twenty years.
- We had a systems based on:

Introduction

- The teaching of Statistics has changed a lot in Spain for the last twenty years.
- We had a systems based on:
  - a lot of hours a week,

Introduction

- The teaching of Statistics has changed a lot in Spain for the last twenty years.
- We had a systems based on:
  - a lot of hours a week,
  - very theoretical subjects.

Introduction

- The teaching of Statistics has changed a lot in Spain for the last twenty years.
- We had a systems based on:
  - a lot of hours a week,
  - very theoretical subjects.
- Nowadays:

Introduction

- The teaching of Statistics has changed a lot in Spain for the last twenty years.
- We had a systems based on:
  - a lot of hours a week,
  - very theoretical subjects.
- Nowadays:
  - the number of hours a week has decreased a lot;

Introduction

- The teaching of Statistics has changed a lot in Spain for the last twenty years.
- We had a systems based on:
  - a lot of hours a week,
  - very theoretical subjects.
- Nowadays:
  - the number of hours a week has decreased a lot;
  - moreover, an important part of them are taught in computer rooms.

Introduction

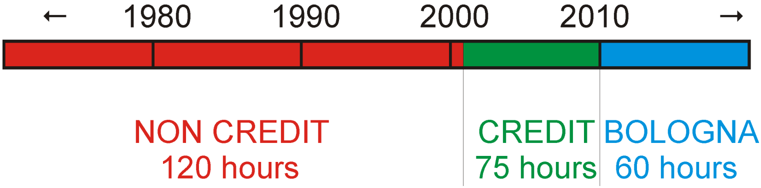
- The teaching of Statistics has changed a lot in Spain for the last twenty years.
- We had a systems based on:
  - a lot of hours a week,
  - very theoretical subjects.
- Nowadays:
  - the number of hours a week has decreased a lot;
  - moreover, an important part of them are taught in computer rooms.
- Let us see in detail these changes and our final proposal, based on our experience.

Evolution

In Spain ...

- Until 90s, we had a system which was not based on credits.
- From the middle of the nineties all the studies started with a new system based on credits. This implied a total change of the Syllabus of all the studies in Spain.
- From 2010 all the studies are according to the Bologna Process. Again this implies a change of the Syllabus.
- We will study the influence of these changes in the teaching of Statistics for non-mathematicians.
- Along this presentation, we will consider, as an example, the case of Statistics for Engineering.

Statistics for Engineering



Overview

Non-credit system

- Until 2001, the Syllabus of Statistics for Engineering established 120 hours of statistical lessons.
- That means 4 hours a week during all the academic year (3rd year) or, equivalently, 8 hours a week during one semester.
- All of them were in the classical classroom: 2 hours with theory and 2 hours solving problems in the blackboard.
- No lessons with computers.
- The content was mainly based on the theoretical aspect of the Statistics: Cramer-Rao bound, convergence in probability, etc.

Content

Introduction

Non-credits system

Credits system

Bologna Process

1. PROBABILITY MODELS: Sample Spaces, Events and Set Operations; Probability: Definition, Axioms and Properties; Probabilistic Space; Discrete Sample Space; Conditional Probability; Bayes Theorem.
2. UNIDIMENSIONAL RANDOM VARIABLES: Concept, Operations and Properties; Distribution Function; Discrete and Continuous Random Variables; Characteristics of a Random Variable.
3. REMARKABLE DISTRIBUTION MODELS: Bernouilli; Binomial; Geometric; Negative Binomial; Hypergeometric; Poisson; Uniform; Gamma; Beta; Normal; Weibull.
4. N-DIMENSINAL RANDOM VARIABLES: Concept; Classification (Discrete and Continuous); Distribution Function; Marginal and Conditional Distributions; Independence of Random Variables; Functions of Random Variables: distribution of the sum, product and quotient; Moments of the function of random vectors.

Content

5. CONVERGENCE OF RANDOM VARIABLES: Moment-generating function, Properties; Types of Convergences; Weak Law of Large Numbers; Central Limit Theorem.
6. SAMPLING AND DESCRIPTIVE STATISTICS: Sample and Population; Simple Random Sampling, Montecarlo Method; Graphic Representation of Data; Numerical Description of Data, Measures of Central Tendency and Dispersion; Sampling Distributions for Normal Populations; Fisher Theorem.
7. POINT ESTIMATION: Estimation; Properties of the Estimators (Unbiased, Efficient, Sufficient and Consistent); Methods of point estimation: the Method of Moments and the Method of Maximum Likelihood.

Content

8. INTERVAL ESTIMATION: Random Intervals and Confidence Intervals; Procedures for obtaining confidence intervals: pivot method; Confidence Intervals for Normal Populations.
9. HYPOTHESIS TESTING: Definition; Neyman-Pearson Test; Errors, Power and Critical Value; Parametric test and confidence interval; Likelihood-ratio test; Chi-Square Test for fit of a distribution; Kolmogoroff-Smirnoff Test for fit of a distribution; Independence and Homogeneity Test.
10. ANALYSIS OF VARIANCE: One-way ANOVA; Two-way ANOVA: without and with interaction.

Overview

Non-credit system

- From 2001 to 2009, the Syllabus of Statistics for Engineering established 75 hours of statistical lessons.
- That means 5 hours a week during one semester. Almost the middle!
- It was taught at the second academic year of the studies.
- Four of them were in the classical classroom: 3 hours with theory and 1 hour solving problems.
- One hour a week was in a computer room.

Kind of lesson	← 2001		2001→
Theoretical	4	⟹	3
Practical	4	⟹	1
With computers	0	⟹	1

Content

Content

- However the number of proofs in the classroom decreased a lot.
- Also some parts were explained with less detail. For example,
  - CHAPTER 4. DISCRETE AND CONTINUOUS DISTRIBUTION MODELS: Bernouilli; Binomial; Geometric; Negative Binomial; Hypergeometric; Poisson; Uniform; Gamma; Beta; Normal; Weibull.
- The practical lessons usually were not coordinated with the theorical lessons.
- However, the content was still mainly based in the theoretical aspect of the Statistics. As an example:
  - CHAPTER 6. CONVERGENCE OF RANDOM VARIABLES: Types of Convergences; Chebychev Theorem; Weak Law of Large Numbers; Central Limit Theorem.

Assessment system

- The final mark of this course will be based on the written official exam and the mark obtained at the practical sessions in the computer rooms.
- The written exam will be divided in two parts: a first theoretical part, formed by a test valued in four points and a second with exercises, valued in five points. In total, the written exam is valued with 9/10 points.
- The practical sessions in the computer rooms, which will be compulsory in order to be allowed to do the written exam, will be valued with 1/10 points.

Overview

And them...Bologna arrives!!!

- From 2010 any Universitary Syllabus in Spain has to be adapted to the Bologna System.
- This supposed a new change of the Syllabus. Usually this implies a decresing number of hours for Mathematics.
- In particular, the Syllabus of Statistics for Engineering established 60 hours of statistical lessons.

	← 2001		2001–2010		2010→
N° hours	120	⟹	75	⟹	60

- That means 4 hours a week during one semester.

Just the middle that 9 years before!

Bologna System

- Three hours a week are in the classical classroom: 2 hours with theory and 1 hour solving problems.
- One hour a week was in a computer room.

Kind of lesson	← 2001		2001–2010		2010→
Theoretical	4	⟹	3	⟹	2
Practical	4	⟹	1	⟹	1
With computers	0	⟹	1	⟹	1

- Moreover, it is taught at the second semester of the first academic year of the studies. Thus, the level of Mathematics is still very low for the students.

New purposes!!!

- At that moment some of us consider we have the opportunity of a total change in our methodology.

New purposes!!!

- Until 2001 (Very high theoretical level!!!):
  - The students had a good theoretical level.
  - Their practical abilities were really poor.
  - Most of them were not able to connect the subject with their profession.
- From 2001 until 2010 (Neither high theoretical nor practical level!!!):
  - The students had not a good theoretical level.
  - Their practical habilities were not realistic.
  - Besides that, the link between the theoretical and practical lessons was really unestable.
  - Some of them were able to connect the subject with their profession.
  - In fact, an elective course about Statistical Quality Control was the most elected, with more than 100 students (more than the double of the next one!!!).

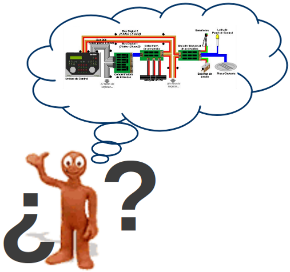
New purposes from 2010!!!

- We hope the subject is useful to them for solving real problems in their future jobs.

- Everyone knows how to call by phone, but almost no one knows the theory of this process.
- We decided to use a similar system for teaching Statistics for No-Mathematicians.

New purposes from 2010!!!

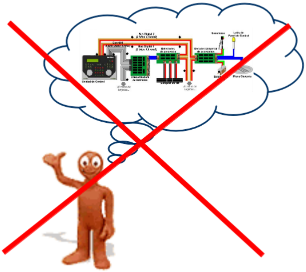
- We hope the subject is useful to them for solving real problems in their future jobs.



- Everyone knows how to call by phone, but almost no one knows the theory of this process.
- We decided to use a similar system for teaching Statistics for No-Mathematicians.

New purposes from 2010!!!

- We hope the subject is useful to them for solving real problems in their future jobs.



- Everyone knows how to call by phone, but almost no one knows the theory of this process.
- We decided to use a similar system for teaching Statistics for No-Mathematicians.

New purposes from 2010!!!

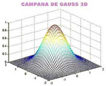
- We will sacrifice the theory in order to give the students high practical abilities.
- Any concept is introduced from an intuitive point of view and we devote our time to the ideas behind this concept and its practical use with the computer.
- Students participe all the time in the introduction of the new concepts and they deduce with the help of the teacher the utility and behaviour of any new concept.
- Computer lessons are the main part of the subject. Any other lessons can be considered as a preparation for this lessons.
- In computer lessons we consider a real data base and we use the concepts introduced in the "theorical" lessons to analyse these data and obtain real conclusions using Statistics.

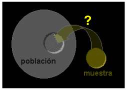
Assessment system

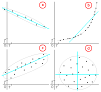
- The final mark of this course will be based on the written official exam and the mark obtained at the practical sessions in the computer rooms.
- The written exam is based on ideas and analysis of the results obtained by the computer. Students have not to develept the theory any time. Only to understand the concept and to know its applicability. The written exam is valued with 7/10 points.
- The practical sessions in the computer rooms, which will not be compulsory in order to be allowed to do the written exam, will be valued with 3/10 points.

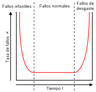
COMPUTER LESSONS	2001–2010		2010→
Valuation	1	⟹	3
Attendance	Compulsory	⟹	Elective

Content

1. DESCRIPTIVE STATISTICS 
2. PROBABILITY MODELS 

3. STATISTICAL INFERENCE 

4. MULTIVARIATE ANALYSIS 

5. RELIABILITY ANALYSIS 

Descriptive Statistics

1. Introduction
2. Basic concepts
3. Frequency Tables
4. Graphical Representation
   - BarChart
   - PieChart
   - Histogram
   - Boxplot
5. Univariate Descriptive Measures
   - Measures of Central Tendency
   - Other Location Measures
   - Measures of Dispersion

A lot of Sketches!!!

KIND OF VARIABLE	FREQ. TABLE	USUAL GRAPHS	LOCATION MEASURES	DISP. MEAS.
Qualitative-Nominal (sex, machine, ...)	Yes	BarChart PieChart	Mode	
Qualitative-Ordinal (Studies level, ...)	Yes	BarChart PieChart	Mode Median Percentile	
Quantitative discrete (N° days, ...)	Yes	BarChart PieChart	Mode Mean Median Percentile	All
Quantitative continuous (Weight, ...)	No	Histogram Box-Plot	Mean Median Percentile	All

Methodology

- Students obtain by themselves the definition of mean.
- They also arrive to its drawbacks:
  - Example 1: A grandfather has 4 grandchildren who are 4, 5, 6 and 30 years old. The mean is 11.25 years all. Is this a right central tendency measure?
  - Example 2: The mean salary in Spain in 2002 was 19808 euro by year. However a half of the Spaniars earned less than 15832 euro. Is the mean a right central tendency measure in this case?

Methodology

- We work with students with the idea of Median.
- We also present some examples were the median is not the best choice:
  - Academic results:

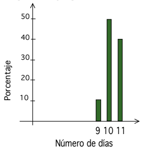
Anne	
Mark	nº of credits
1	150
2	45
3	18
4	12

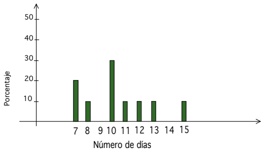
Susan	
Mark	nº of credits
1	180
2	36
3	9
4	0

  - Anne is better than Susan, but the median is 1 in both cases.
- Critical Thinking in Learning!!!

Necessity of Measures of Dispersion

- Example 1: Two providers have an average of 10.3 days to provide the order. The number of working days used in the last orders was:





  Who is our preference?

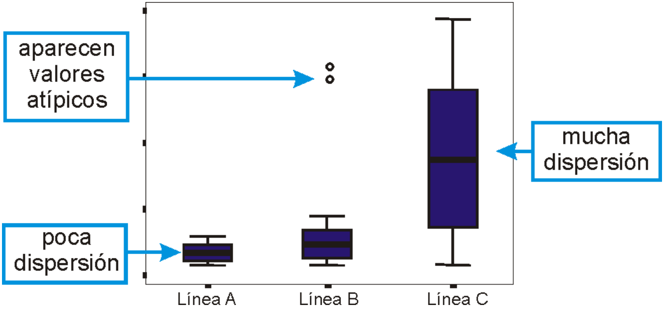
- Example 2: Hours devoted to study by four students.

	John	Anne	Emma	Paul
October	3	0	0	0
November	3	2	0	0
December	3	3	6	3
January	4	8	7	10

  Who is more regular? And less?

- Active participation of Students!!!

Interpretations



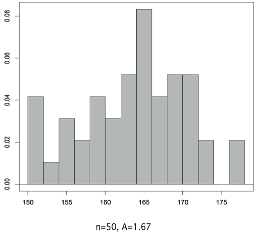
- They devote their time to undertand the results obtained with the computer!!!

Probability Models

1. Introduction
2. The concept of Density Function
3. Distribution Function of a Continuous Random Variable
4. Expected value, variance and standard deviation for a continouos random variable
5. Remarkable continous distributions:
   - Normal or Gaussian
   - Exponential
   - Weibull
6. Introduction to discrete models
7. The concept of probability mass function
8. Expected value, variance and standard deviation for a discrete random variable
9. Remarkable continous distributions:
   - Binomial
   - Poisson

Histogram and Density Function

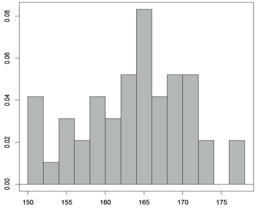
X =height; n =n$^{\text{o}}$ of women;
A =interval width



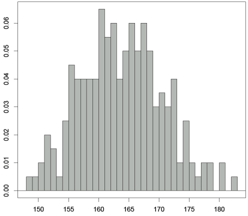
Histogram and Density Function

X =height; n =n^{o} of women;

A =interval width



n=50, A=1.67

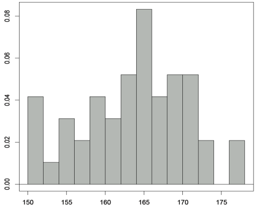


n=200, A=1

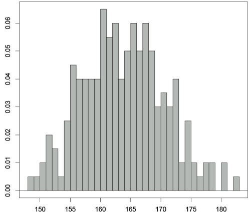
Histogram and Density Function

X =height; n =n$^{\text{o}}$ of women;

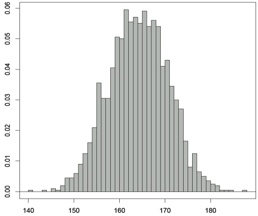
A =interval width



n=50, A=1.67



n=200, A=1



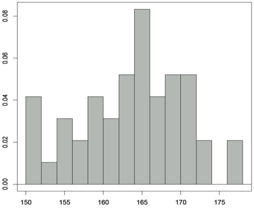
N=2000, A=1

$n \to \infty$

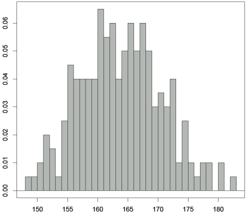
$A \to 0$

Histogram and Density Function

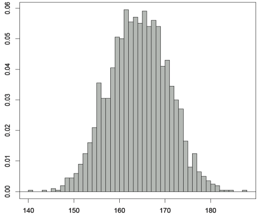
X =height; n =n^o of women;
A =interval width



n=50, A=1.67

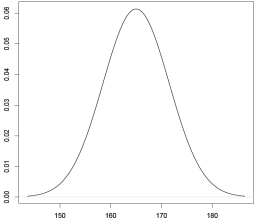


n=200, A=1

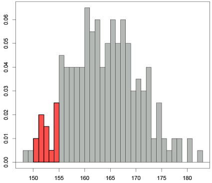


N=2000, A=1

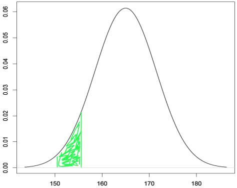
$n \to \infty$
$A \to 0$



Idea of probability: continuous r.v.



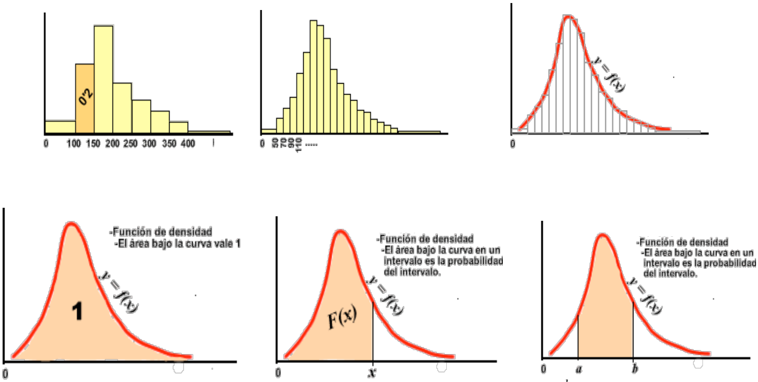
área( )=proporción de mujeres en la muestra cuya estatura está comrpendida entre 150 y 155 cm.



área( )= probabilidad de que, al elegir al azar una mujer, su estatura esté comprendida entre 150 y 155 cm.

- Density function: measures probabilities or "theorical proportions".
- When $n \to \infty$ and $A \to 0$, red area $\approx$ green area.
  (If the theoretical model describes properly the population.)

Probability of an interval from the density function



$$\int_{-\infty}^{\infty} f(t)dt = 1$$

$$\int_{-\infty}^{x} f(t)dt = P(X \leq x) = F_X(x)$$

$$\int_{a}^{b} f(t)dt = P(a < X < b) = P(a \leq X \leq b)$$

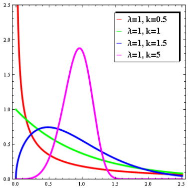
Methodology

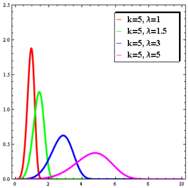
- No integrals!
- We obtain any probability from the expression of the Distribution Function, since it is similar to the procedure used with a computer.

Example

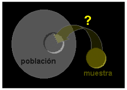
The r.v. X has a Weibull distribution with parameters $k > 0$ *(shape)* and $\lambda > 0$ *(scale)*, if its density function is:

$$f(x) = \begin{cases} \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-\left(\frac{x}{\lambda}\right)^{k}} & \text{si } x \geq 0 \\ 0 & \text{si } x \leq 0. \end{cases}$$





Statistical Inference



1. Introduction
2. Point Estimation
3. Interval Estimation
4. Statistical Hypothesis Testing

Necessity of the Statistical Inference

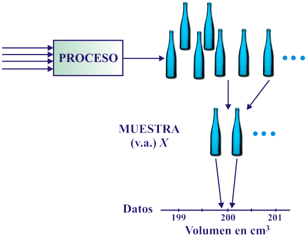
Introduction

Non-credits system

Credits system

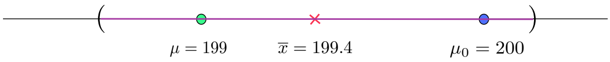
Bologna Process

Is the volume of these bottle according to the specifications?



- The sample **ONLY** provides us information about the sample.
- For general conclusions (for all the population) the techniques of Statistical Inference are **NECESSARY**.

Again ideas!!!

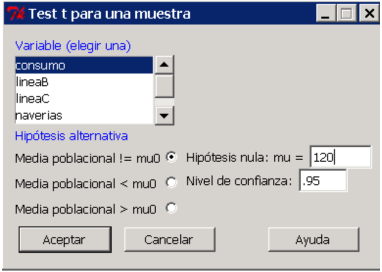


- Point estimation of μ: $\overline{x} = 199'4$.
  Error of estimation (unknown by the controlers): $0'4\text{cm}^3$.
- Interval estimation of μ: violet interval obtained from a sample of 100 bottles. Controlers have a high security (or confidence level) that μ is in this interval. But, they don't know the value of μ.
- Hypothesis Testing about μ: we conclude if μ can be igual to $\mu_0 = 200$. From this sample, we decide don't reject this hypothesis.

Intrepretation vs. manual development

Results of the one-sample test t

Estadísticos ⟶Medias ⟶Test t para una muestra



One Sample t-test

data: Acero$consumo

t = 3.8136, df = 116, **p-value = 0.0002210**

alternative hypothesis: **true mean is not equal to 120**

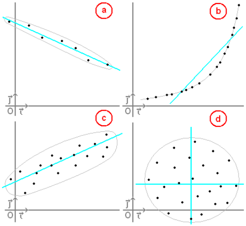
95 percent confidence interval: 129.3516 149.5614

sample estimates: mean of x 139.4565

Studied tests

- Goodness-of-fit test.
  - Shapiro-Wilk normality test
- Central tendency measures.
  - Test t for one sample (normal population)
  - Wilcoxon Test for one sample (NO normal pop.)
- Proportion.
- Comparison of two proportions.
- Comparison of two dispersions.
  - Test F (normal)
  - Levene Test (no normal)
- Difference of central tendency measures.
  - Test t for independent samples (normal, independence)
    - Equal variability
    - Non-equal variability
  - Test t for related data (normal, paired samples)
  - Wilcoxon Test for two samples (NO normal, independent samples)
  - Wilcoxon Test for paired samples (NO normal, paired samples)

Multivariate Analysis



1. Relation between populations
   - Chi-Square Independence Test
   - Tests of Pearson's Correlation
2. Lineal Regression
   - Lineal Regression Models
   - Estimation of the parameters
   - Model Adequacy
   - Forecasting

Multivariate Analysis

- They obtain the chi-square coefficient, the contigency coefficient and interprete the chi-square test of independence.
- They obtain covariance, the Pearson's product-moment correlation coefficient and interprete the p-value of the Pearson's correlation test
- They are able to do a complete simple linear regression analysis with real data (with residual analysis).

Working by examples!!!

EMPRESA 1 - INDEPENDENCIA

	SEXO		
Sueldo	Hombre	Mujer	*TOTAL*
20 – 35	240	60	300
36 – 50	160	40	200
51 – 65	400	100	500
TOTAL	800	200	1000

⇓

$\chi^2 = 0$

⇓

Pearson's Chi-squared test

data: .Table

X-squared = 0, df = 2, p-value = 1

⇓

No rechazo H_0: No hay evidencias estadísticas de relación entre el sexo de la persona y el sueldo que cobran por hora

EMPRESA 2 - RELACIÓN

	SEXO		
Sueldo	Hombre	Mujer	*TOTA*
20 – 35	120	180	300
36 – 50	185	15	200
51 – 65	495	5	500
TOTAL	800	200	1000

⇓

$\chi^2 = 432'34$

⇓

Pearson's Chi-squared test

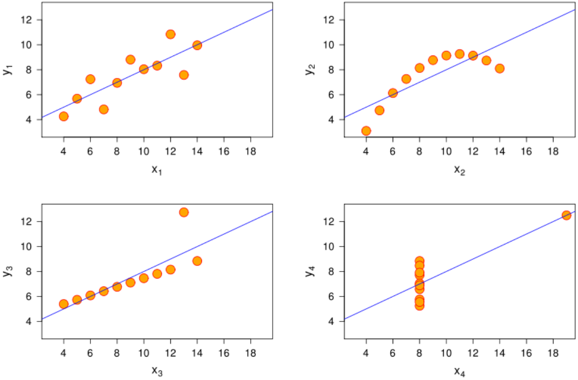
data: .Table

X-squared = 432.3438, df = 2, p-value ¡

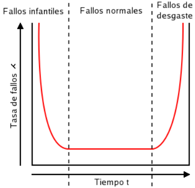
⇓

Rechazo H_0: Hay evidencias estadísticas de relació
sexo de la persona y el sueldo que cobran por hora

Again critical learning: Anscombe

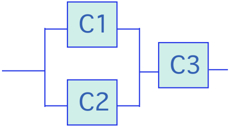


Reliability Analysis



1. Introduction
2. Reliability Function
3. Reliability of a System
4. Set Theory
5. Stochastic Independence
6. Formal aspect of the Probability Theory: Conditional Probability, Total Probability and Bayes Theorems.

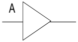
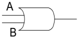
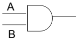
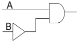
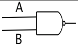
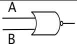
An excuse for introducing set theory



C1	C2	C3	S
0	0	0	0
0	0	1	0
0	1	0	0
0	1	1	1
1	0	0	0
1	0	1	1
1	1	0	0
1	1	1	1

- $F_1 = \{100, 101, 110, 111\}$
- $F = \{011, 101, 111\}$

An excuse for introducing set theory

OPERACIÓN	REPR. OPERACIÓN	PUERTA	REPR. PUERTA
Complementario	$\overline{A}$	NOT	
Unión	$A \cup B$	OR	
Intersección	$A \cap B$	AND	
Diferencia	$A \setminus B$	– – –	
Dif. simétrica	$A \triangle B$	XOR	
– – –	$\overline{A \cap B}$	NAND	
– – –	$\overline{A \cup B}$	NOR	

Conclusions

1. More participation
2. More interest
3. An audience more critical
4. They relate the subject with their studies
5. They undestand the utility of the statistics for them
6. They are able to do real data analysis at the end of the course

Thank you for your attention!

http://unimode.uniovi.es
University of Oviedo, Spain