CHI-SQUARED TEST

Pau Fonseca i Casas; pau@fib.upc.edu

Introduction

- “Study the past if you would divine the future ."
  - **Confucius** Chinese philosopher & reformer (551 BC - 479 BC).

Test of homogeneity

- We work with k classes on P populations, and we test if is a common population.
- The data (observed values) are represented in a contingency table:

	C_1	…	C_i	…	C_k	
X_1	$O_{1,1}$		$O_{i,1}$			
…						
X_j	$O_{1,j}$		$O_{i,j}$			$n_{.,j}$
…				…		
X_P					$O_{k,P}$	
			$n_{i,.}$			n

Test of homogeneity

- If exists homogeneity on the populations, exists a common probabilities for each class. Taking $E_{i,j} = n_{\cdot,j}\, n_{i,\cdot}/n$, we can see that:

$$X^2 = \sum_{j=1}^{P} \sum_{i=1}^{k} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}}$$

- Follows a chi-square distribution with $(k\text{-}1)(P\text{-}1)$ degrees of freedom.

Chi-squared test

- Compare the observed frequencies in empirical experience with theoretical frequencies, derived from some hypothetical distribution.
- Objective:
  - Determine the distribution in the population.
  - Check if several groups share the same distribution.
  - Studying independence of two (or more) factors.

Chi-squared test

- Divide interval into k segments with the same density.
- Calculate F^{-1} function we are evaluating.
- From this function to calculate the intervals corresponding to each segment.
- Calculate X^2.

Measure the difference

- The discrepancy between the observed frequencies (O_i) and expected las (E_i) is evaluated by:

$$X^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

- k is the number of classes you take.
- Can be brought on by a discrete variable.
- Or to discretize a continuous variable.

Distribution of the Pearson statistic

- If the data come from a population described by the model given above:

$$X^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

- It follows a Chi-Square law with k-1 degrees of freedom:

$$X^2 \sim \chi^2_{k-1}$$

Example

- We have in a reparation service 4 priority levels for the works to be done:
  - Urgent
  - High
  - Mean
  - Low
- The probabilities for each reparation is urgent 10%, high 20%, mean 30% and low 40%.

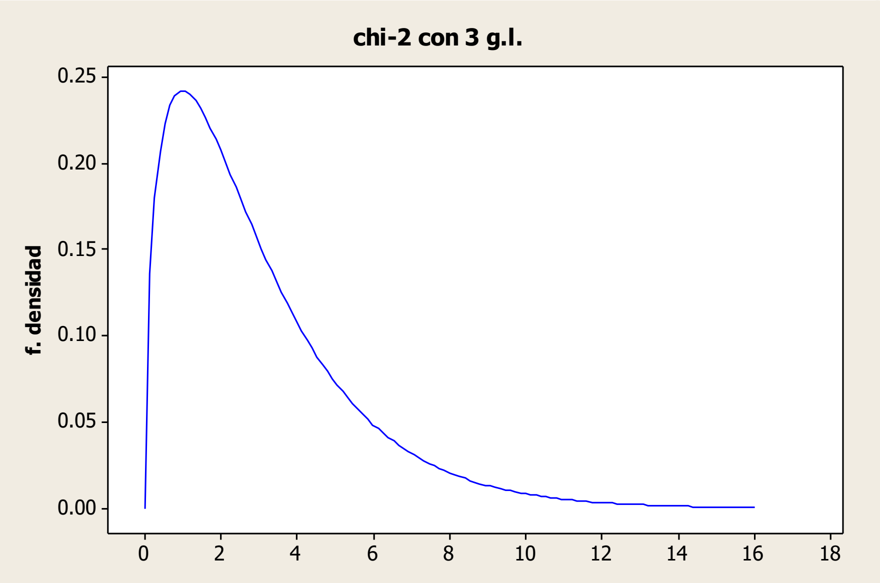
Example

- From historical data we obtain the next table.
- The question is, are the differences between the observed values and our proposed probabilities for each category due to the random nature of the phenomenon?.

	urgent	high	mean	low
Observed	78	107	145	223
Expected	55.3	110.6	165.9	221.2
Diffrerence	22.7	-3.6	-20.9	1.8

Example

- The model we are using is this.
  - 43% between 0 and 2
  - 31% between 2 and 4
  - 15% between 4 and 6
  - 7% between 6 and 8



Example

- Calculating the statistic

Clase	Prob. p_i	Número esperado (sobre 553) E_i	Número observado O_i	Diferencia estandarizada	Dif. estand. al cuadrado
urgente	0.10	0.10·553=55.3	78	3.0526	9.3181
alta	0.20	0.20·553=110.6	107	-0.3423	0.1172
media	0.30	0.30·553=165.9	145	-1.6226	2.6330
baja	0.40	0.40·553=221.2	223	0.1210	0.0146
				Suma:	12.0829

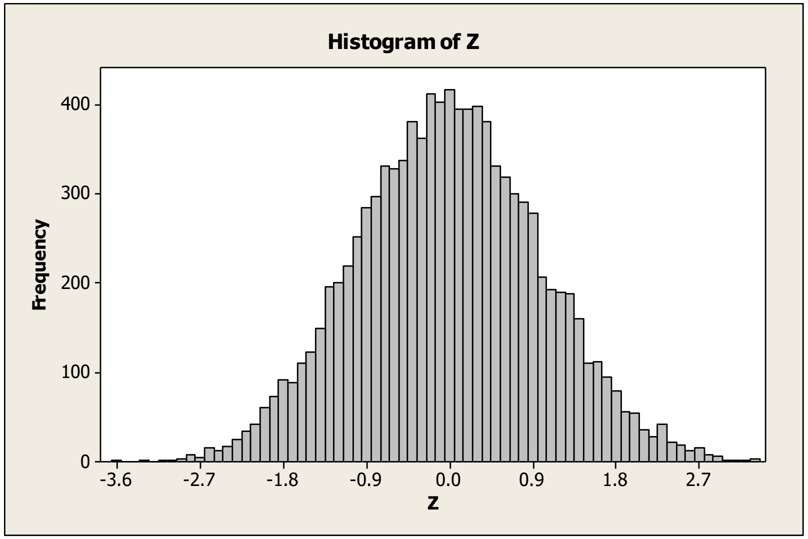
- In our case X^2 is 12.08, looking the table we detect that this value is so high. The proposed classification is not correct.

The Chi-square table

df	0,1	0,05	0,025	0,01
1	3	4	5	7
2	5	6	7	9
3	6	8	9	11
4	8	9	11	13
5	9	11	13	15

For continuous variables

- We want to test the fitness of an algorithm that generates a normal distribution N(0,1).
- We generate 1000 values (empirically generated), it seems that the generation process is correct.
- To assure this we perform a Chi-suqare test.



Arbitrarily define a discretization of the variable in 10 classes to obtain a significant frequency.

Clas	Prob.	Expected number (over 10000)	Observed number	Standardized difference
< –3	0.001350	13.50	6	-2.04
[–3, –2]	0.021400	214.00	195	-1.30
[–2, –1.5]	0.044057	440.57	456	0.74
[–1.5, –1]	0.091848	918.48	934	0.51
[–1, 0]	0.341345	3413.45	3477	1.09
[0, 1]	0.341345	3413.45	3402	-0.20
[1, 1.5]	0.091848	918.48	886	-1.07
[1.5, 2]	0.044057	440.57	417	-1.12
[2, 3]	0.021400	214.00	220	0.41
> 3	0.001350	13.50	7	-1.77

X2 val13.59. P-value (with 9 g.l.) = 0.14. There is no evidences against this.

Table value=14.6837

Other considerations

- Beware rare classes: so that the test is reliable, it must have a minimum number of expected observations (usually, more than 5).
- If the sample is used to determine the value of a parameter of the reference distribution, reducing the number of degrees of freedom.
- It is important to review the differences standardized , considered (under H0) as N (0,1). Notable values are significant signs of abnormalities (eg, the normal las tails).

Proposed exercices

Exercise 1: Epilachna varivestis

- Determine whether the following amount of insects (*Epilachna varivestis* by bean plants) can be represented by a Poisson distribution:

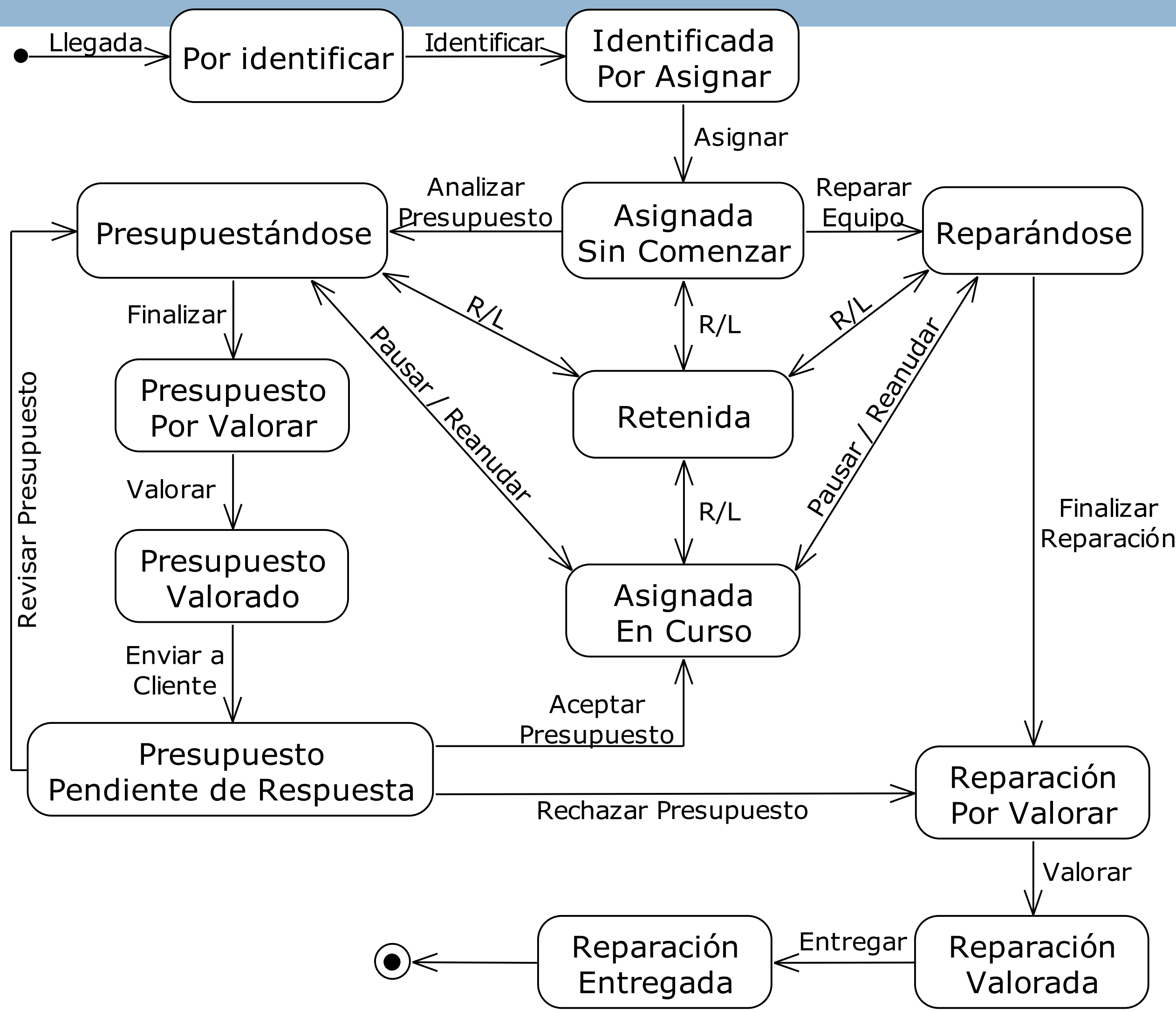
Y	fy (=Oi)	Pi	Ei	(Oi-Ei)^2/Ei
0	12		23.03	
1	56		36.01	
2	23		28.15	
3	10		14.67	
4	5		5.73	
5	4		2.39	
total	110		110	

$$\chi^2_{q-(p+1)} = \sum_{i=0}^{q} \frac{(Oi - Ei)^2}{Ei}$$

Exercise 2: Reparations SA

- **Reparacions SA** is an small business that works in the reparation of electronic components.
- Very technician, young and dynamic team.

The process



Llegada
Por identificar
Identificar
Identificada Por Asignar
Asignar
Analizar Presupuesto
Reparar Equipo
Presupuestándose
Asignada Sin Comenzar
Reparándose
R/L
R/L
R/L
Finalizar
Pausar / Reanudar
Pausar / Reanudar
Presupuesto Por Valorar
Retenida
Revisar Presupuesto
Valorar
R/L
Finalizar Reparación
Presupuesto Valorado
Asignada En Curso
Enviar a Cliente
Aceptar Presupuesto
Presupuesto Pendiente de Respuesta
Rechazar Presupuesto
Reparación Por Valorar
Valorar
Reparación Entregada
Entregar
Reparación Valorada

The problem

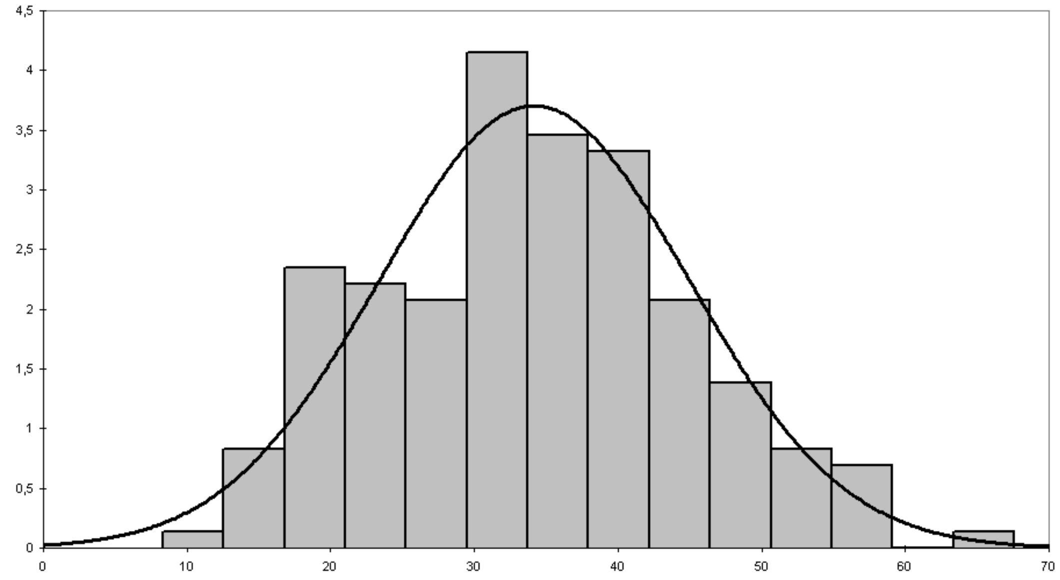
- We want to create a DSS (Decision Support System) to simplify the initial allocation of tasks of Reparacions SA, (defining an algorithm for optimal assignment of tasks).

The data

- The data represent the full operation of the DB model.

Example of the model distribution

Using N (34.17, 10.77) to model the probability that a task requires a budget or not.

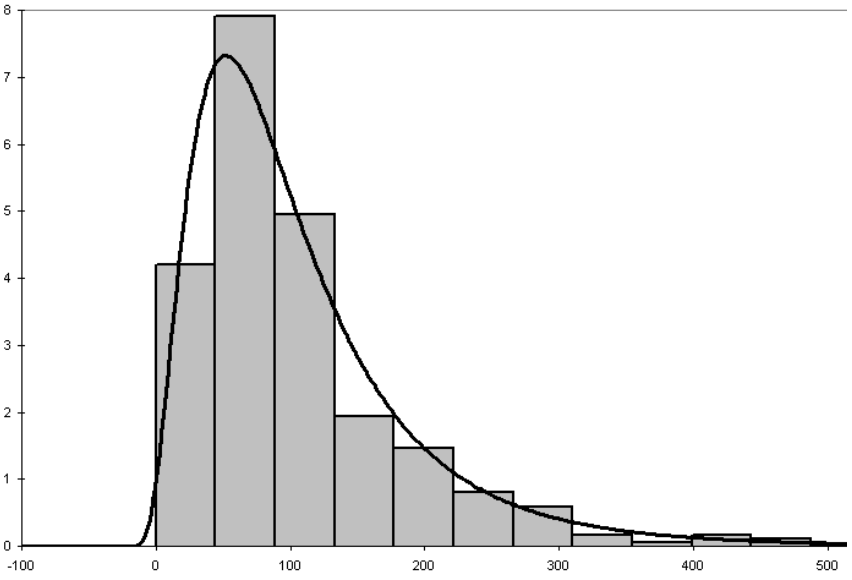


The delays

Type of delay	**Distribution function**
Not assigned	Exponential (209,75)
Retained	Exponential (5837)
Budget pending a response	Exponential (2966,6)

Example of delay

- Time to repair a computer by operator number 3 (generic reparation) without any budget.
- Lognormal distribution (4,68 ; 0,61)



Our example

- Working with a Chi-square test to determine which is most appropriate fdp.
- The data that ill represent new repair arrivals company (time in minutes).
  - Mean: 152.92
  - Deviation: 160.4
  - N = 225

1	1	2	2	3	4	4	5	5	5	5	6	7	7	8
10	10	11	11	11	12	14	14	14	15	15	15	16	16	16
17	18	18	19	21	21	21	22	22	24	24	25	25	25	26
28	29	30	31	31	31	32	32	33	34	36	36	37	37	40
40	40	41	41	43	43	43	44	45	47	47	47	48	48	51
53	54	54	54	55	56	57	57	58	59	61	61	64	67	68
69	69	75	77	77	77	77	78	78	78	79	81	82	82	84
84	86	90	94	95	96	97	97	98	103	106	106	106	108	108
112	112	113	118	119	121	122	125	126	127	128	130	132	133	133
134	134	135	139	142	146	149	149	154	155	156	160	164	169	169
170	171	180	186	186	187	193	196	201	203	208	209	210	213	214
218	220	221	222	226	227	229	233	239	239	239	241	249	249	253
256	257	261	270	272	273	286	288	310	323	324	326	329	335	338
340	343	348	350	350	353	354	358	371	375	382	382	387	411	418
423	427	437	445	458	464	496	538	545	567	587	723	753	769	991