

# CHI-SQUARED TEST

Nihan Acar-Denizli, Pau Fonseca i Casas

# Introduction

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

# Chi-squared test

## □ Objectives:

- To check if several groups share the same distribution.
  - To determine the distribution in the population.
  - To study the independence of two (or more) factors.
- 
- Compare the observed frequencies in an empirical experience with theoretical frequencies, derived from some hypothetical distribution.

# Chi-squared test

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

# Measure the difference

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

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

- $k$  is the number of classes.
- Can be brought on by a discrete variable.
- Or by discretizing a continuous variable.

# Distribution of the Pearson statistic

- If the data come from a population described by the model given below,

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

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

$$X^2 \sim \chi_{k-1}^2$$

# Test of homogeneity

- We work with  $k$  classes on  $P$  populations, and we test if there 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 homogeneity exists on the populations, then a common probability exists for each class.

Taking  $E_{i,j} = n_{i.} n_{.j} / n$ , 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  $(k-1)(P-1)$  degrees of freedom.



# Example: Test of homogeneity

- 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: Test of homogeneity

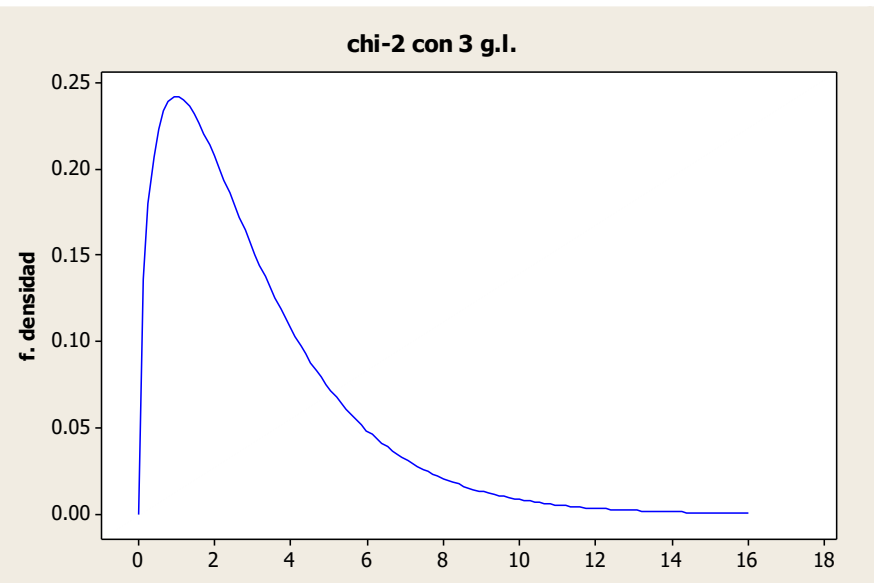
- From historical data we obtain the following table.

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

- «Are the differences between the observed values and our proposed probabilities for each category due to the random nature of the phenomenon?»

# Example: Test of homogeneity

- 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



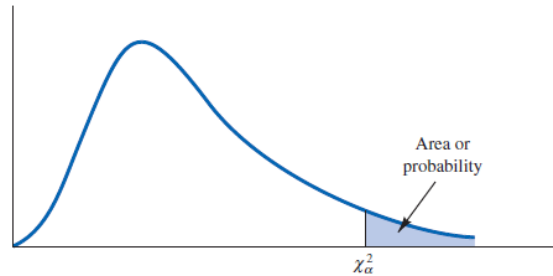
# Example: Test of homogeneity

## □ Calculating the statistic:

Class	Prob (pi)	Expected Number (Ei)	Observed Number (Oi)	Standardized Difference	Squared Std. Diff.
Urgente	0.10	$(0.10) \cdot 553 = 55.3$	78	515.29	9.31808318
High	0.20	$(0.20) \cdot 553 = 110.6$	107	12.96	0.11717902
Medium	0.30	$(0.30) \cdot 553 = 165.9$	145	436.81	2.63297167
Low	0.40	$(0.40) \cdot 553 = 221.2$	223	3.24	0.01464738
				<b>SUM= 12.0829</b>	

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

# Chi Square Table



Degrees of Freedom	Area in Upper Tail									
	.995	.99	.975	.95	.90	.10	.05	.025	.01	.005
1	.000	.000	.001	.004	.016	2.706	3.841	5.024	6.635	7.879
2	.010	.020	.051	.103	.211	4.605	5.991	7.378	9.210	10.597
3	.072	.115	.216	.352	.584	6.251	7.815	9.348	11.345	12.838
4	.207	.297	.484	.711	1.064	7.779	9.488	11.143	13.277	14.860
5	.412	.554	.831	1.145	1.610	9.236	11.070	12.832	15.086	16.750
6	.676	.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.647	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.041	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801

# Is the die biased?

X	1	2	3	4	5	6
Number of times	20	15	12	17	9	17

$H_0$  : The die is fair (The distribution of counts follows a Uniform distribution ( $F = F_0$ )).

$H_1$  : The die is not fair (The distribution of counts does not follow a Uniform distribution ( $F \neq F_0$ )).

The expected value is s

$$E[X] = np = 90 * \left(\frac{1}{6}\right) = 15.$$

# Is the die biased?

$$\begin{aligned}\chi^2 &= \sum_i \frac{(O_i - E_i)^2}{E_i^2} \\ &= \frac{(20 - 15)^2}{15} + \frac{(15 - 15)^2}{15} + \dots + \frac{(17 - 15)^2}{15} = 5.2\end{aligned}$$

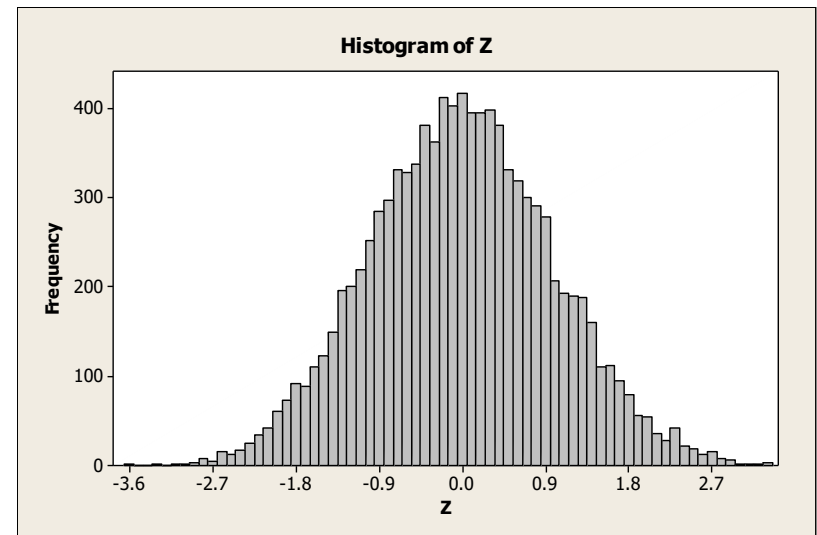
$$P(\chi^2 > 5.2) = \text{between } 0.1 \text{ and } 0.9$$

$$\text{or } \chi_{3,0.05}^2 = 7.815 > \chi_{obs}^2 = 5.2$$

There is no evidence to reject the null hypothesis.

# Testing a family of distributions for continuous variables

- We want to test the fitness of an algorithm that generates a normal distribution  $N(0,1)$ .
- We generate 10000 values (empirically generated), it seems that the generation process is correct.
- To assure this we perform a Chi-square test.



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



# Testing a family of distributions for continuous variables

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

## Testing a family of distributions for continuous variables

$$\chi^2_{9,0.05} = 14.6837 > \chi^2_{obs} = 13.59$$

Besides, the p value of the value 14.68 would be around 0.10. That is greater than  $\alpha = 0.05$ .

(You might check p-value calculator for chi-square table: <https://www.socscistatistics.com/pvalues/chidistribution.aspx>)

So, there is no evidences against that the distribution does not follow a Normal distribution.

# Example: Test of Independency

- Recently, a study has determined that the zodiac sign does not play, against that deep, no influence when looking spouse.
- The study classified a large number of couples and found no evidence of heterogeneity between signs.
- For simplicity we work with 4 signs instead of 12.

# Example: Test of Independency

		Spouse 2			
Spouse 1		Pisces	Aquarius	Taurus	Leo
	Pisces	888	1025	1032	1055
	Aquarius	861	1005	1065	1069
	Taurus	914	1031	1008	1047
	Leo	850	991	1039	1120

# Example: Test of Independency

- $H_0$ : The distribution of the zodiacal sign of the 2nd spouse is independent from the sign of the first spouse.

	Piscis	Acuario	Tauro	Leo	
Piscis	888	1025	1032	1055	4000
Acuario	861	1005	1065	1069	4000
Tauro	914	1031	1008	1047	4000
Leo	850	991	1039	1120	4000
	3513	4052	4144	4291	16000
	21.96%	25.32%	25.9%	26.82%	
Piscis	878.25	1013	1036	1072.75	
Acuario	878.25	1013	1036	1072.75	
Tauro	878.25	1013	1036	1072.75	
Leo	878.25	1013	1036	1072.75	

# Example: Test of Independency

- $\chi^2$  statistic related to the table is 8.41 (p-value = 0,493).
- According to the start hypotheses we must follow a  $\chi^2$  of 9 d.f. where the table value is 16.919.
- Since  $p > 0.05$ , we can not reject the null hypothesis.
- The zodiacal sign of the pair is distributed evenly, regardless of the sign. The sign of the second spouse is independent from the sign of the first spouse.

# Other considerations

- Beware rare classes: so that the test is reliable, it must have a minimum number of expected observations (usually, more than 5).
- It is important to review the standardized differences, considered (under  $H_0$ ) as  $N(0,1)$ . Notable values are significant signs of abnormalities.



# Proposed exercises



# Exercise 1: *Epilachna varivestis*

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

Y	$f_y (=O_i)$	$P_i$	$E_i$	$(O_i - E_i)^2 / E_i$
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	

df	0,05	0,025	0,01	0,005
4	9	11	13	14,8602
5	11	13	15	16,7496
6	13	14	17	18,5475

# Exercise 1: *Epilachna varivestis*

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

Y	$f_y (=O_i)$	$P_i$	$E_i$	$(O_i - E_i)^2 / E_i$
0	12	0.21	23.03	5.28
1	56	0.33	36.01	11.09
2	23	0.26	28.15	0.942
3	10	0.13	14.67	1.487
4	5	0.05	5.73	0.093
5	4	0.02	2.39	1.084
total	110	1	110	19.976

df	0,05	0,025	0,01	0,005
4	9	11	13	14,8602
5	11	13	15	16,7496
6	13	14	17	18,5475

# Exercise 2: Internet Shopping

- A web designer suspects that the chance for an internet shopper to make a purchase through her web site varies depending on the day of the week. Test this claim.

Observed	Mon	Tues	Wed	Thurs	Fri	Sat	Sun	TOTAL
No purchase	399	261	284	263	393	531	502	
Single purchase	119	72	97	51	143	145	150	
Multiple purchases	39	50	20	15	41	97	86	
TOTAL								3758

# Exercise 2: Internet Shopping

- A web designer suspects that the chance for an internet shopper to make a purchase through her web site varies depending on the day of the week. Test this claim.

Observed	Mon	Tues	Wed	Thurs	Fri	Sat	Sun	TOTAL
No purchase	390.2 6	268.3 4	280.9 6	230.5 1	404.2 7	541.5 9	517.0 7	2633
Single purchase	115.1 6	79.19	82.91	68.02	119.3 0	159.8 2	152.5 9	777
Multiple purchases	51.58	35.47	37.13	30.47	53.43	71.58	68.34	348
TOTAL	557	383	401	329	577	773	738	3758