The Chi Square

Computational Social Intelligence - Lecture 04

Prof. Alessandro Vinciarelli
School of Computing Science &
Institute of Neuroscience and Psychology

http://www.dcs.gla.ac.uk/vincia
Alessandro.Vinciarelli@glasgow.ac.uk

This lecture is based on the following text (available on Moodle):

- D.C.Howell, "Statistical Methods for Psychology", Chapter 6, Sections 6.1, 6.3 and 6.4 (excluding subsection "Correcting for Continuity)", Cengage Learning, 2009.

The extra-material available in the pdf of the text does not need to be studied

Outline

- Very Quick Recap
- Goodness of Fit Test (One Way Classification)
- Two Way Classification
- Conclusions

Outline

- Very Quick Recap
- Goodness of Fit Test (One Way Classification)
- Two Way Classification
- Conclusions

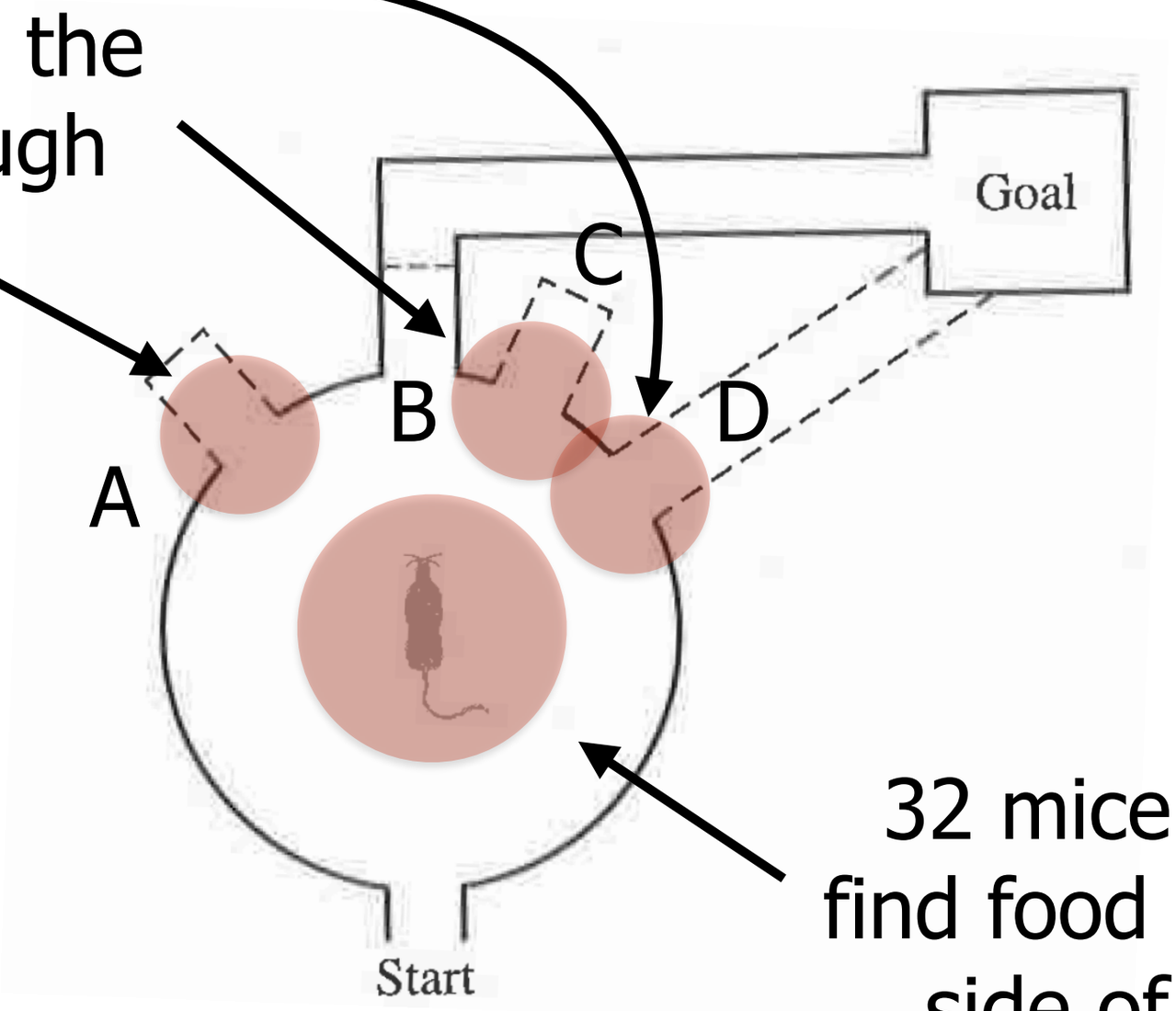
Hypothesis Testing (Main) Ingredients

- Statistic: Any measurement that can be extracted from a data sample;
- Sampling Distribution: probability density function of the statistic when the Null Hypothesis is true;
- Confidence Level: an acceptable probability of doing a Type I Error (rejecting the Null Hypothesis when it should not), typically 5%.

Outline

- Very Quick Recap
- Goodness of Fit Test (One Way Classification)
- Two Way Classification
- Conclusions

When the other alleys are open, the experimenters count how many times the mice pass through them



32 mice "learn" to find food on the right side of the alley

Tolman, Ritchie and Kalish, "Studies in spatial learning. II. Place Learning vs Response Learning", Journal of Experimental Psychology, 36(3):221, 1946.

Research Hypothesis

- Research Hpothesis: The mice learn that the food is on the right and tend to select alleys that go in such a direction;
- Null Hypothesis: The mice select randomly one of the alleys.

Results

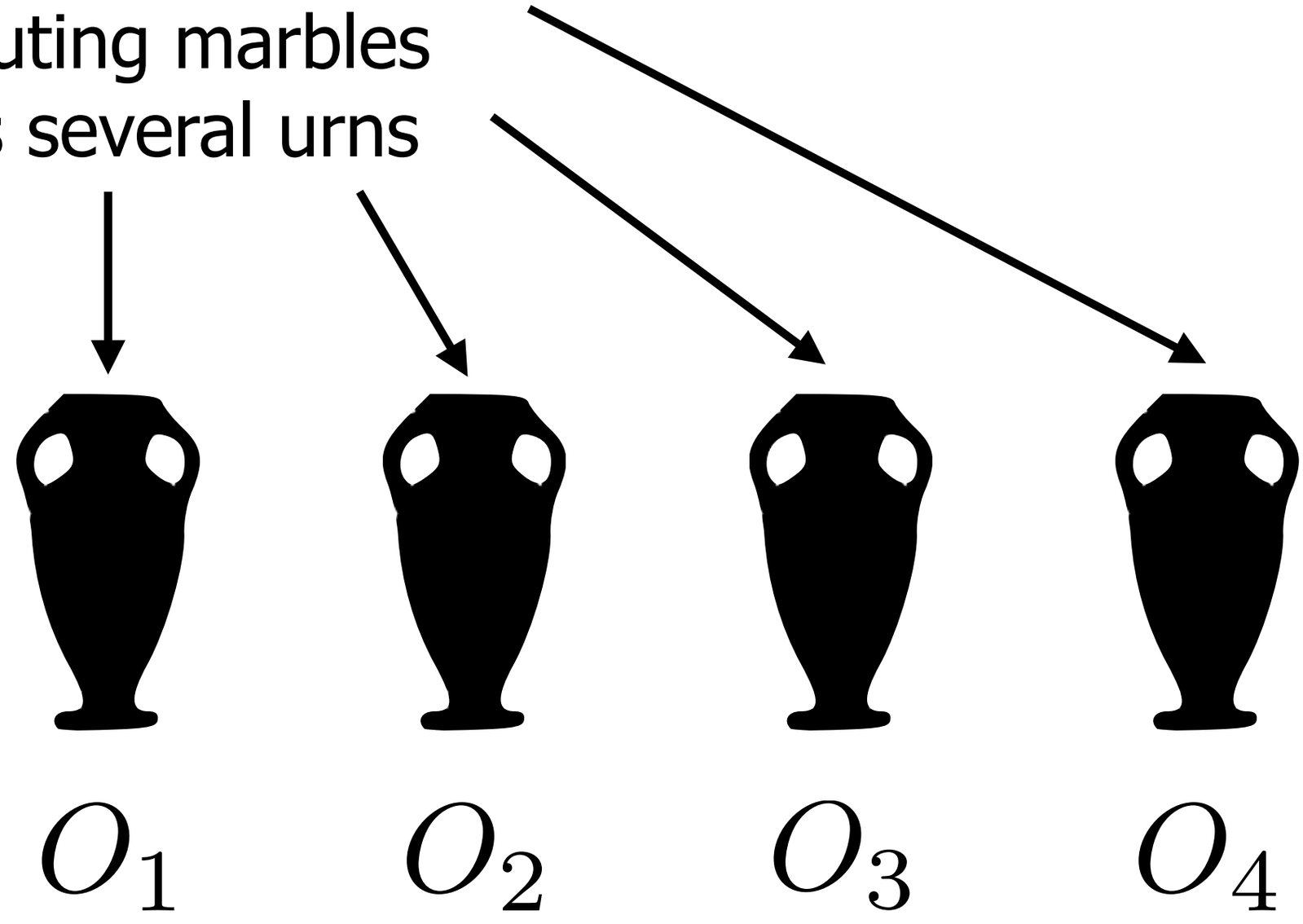
Alley Chosen	A	B	C	D
Observed (O)	4	5	8	15
Expected (E)	8	8	8	8

Tolman, Ritchie and Kalish, “Studies in spatial learning. II. Place Learning vs Response Learning”, Journal of Experimental Psychology, 36(3):221, 1946.

Observations and Expectations

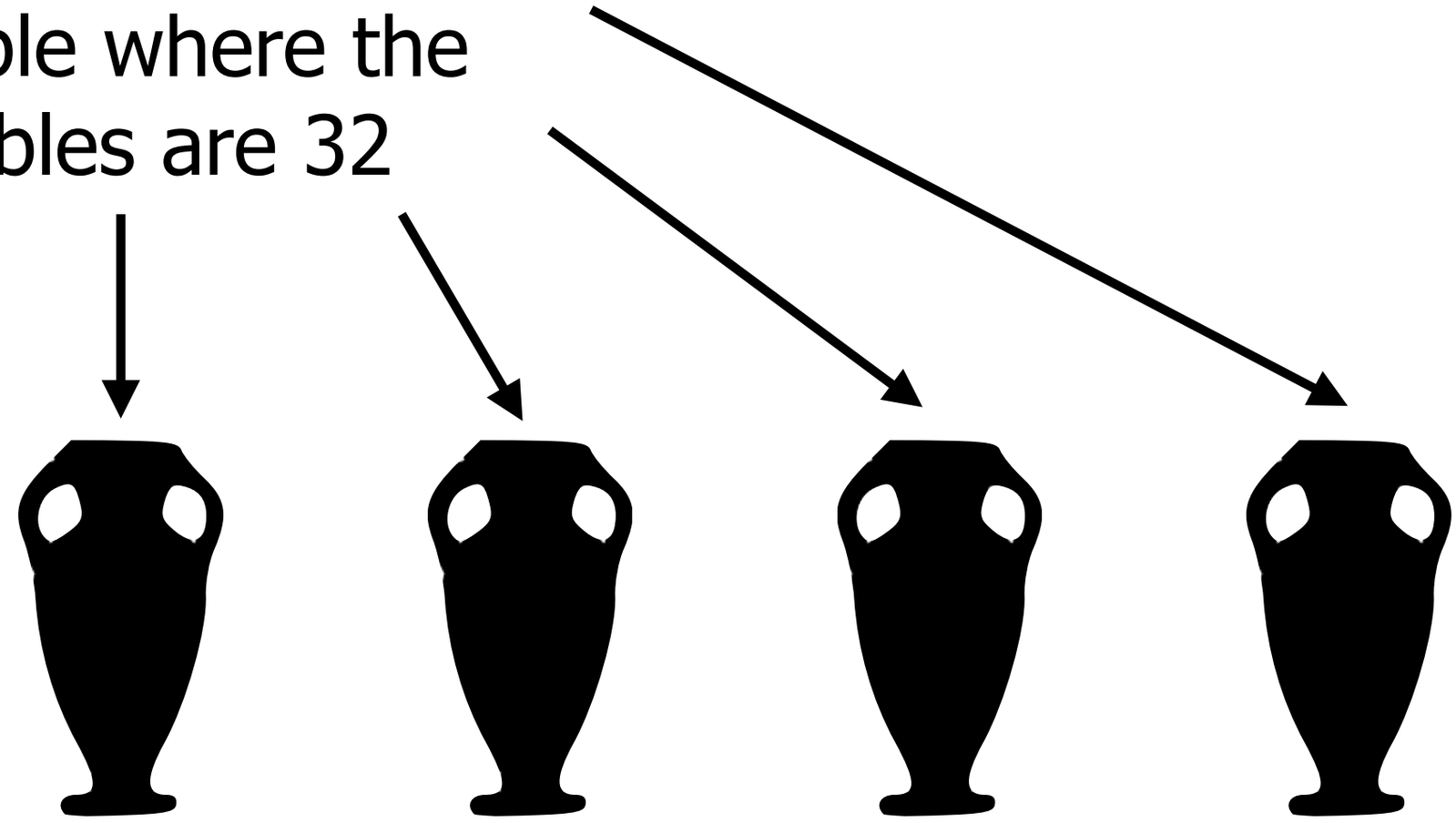
- Observed frequencies are those actually observed in the data (they do not stem from a decision of the experimenter);
- Expected frequencies are those that would be observed if the null hypothesis was true (they stem from the experimental design);
- It is necessary to find a suitable statistic and its sampling distribution.

Consider a process distributing marbles across several urns



The O values are the numbers of marbles observed in the urns

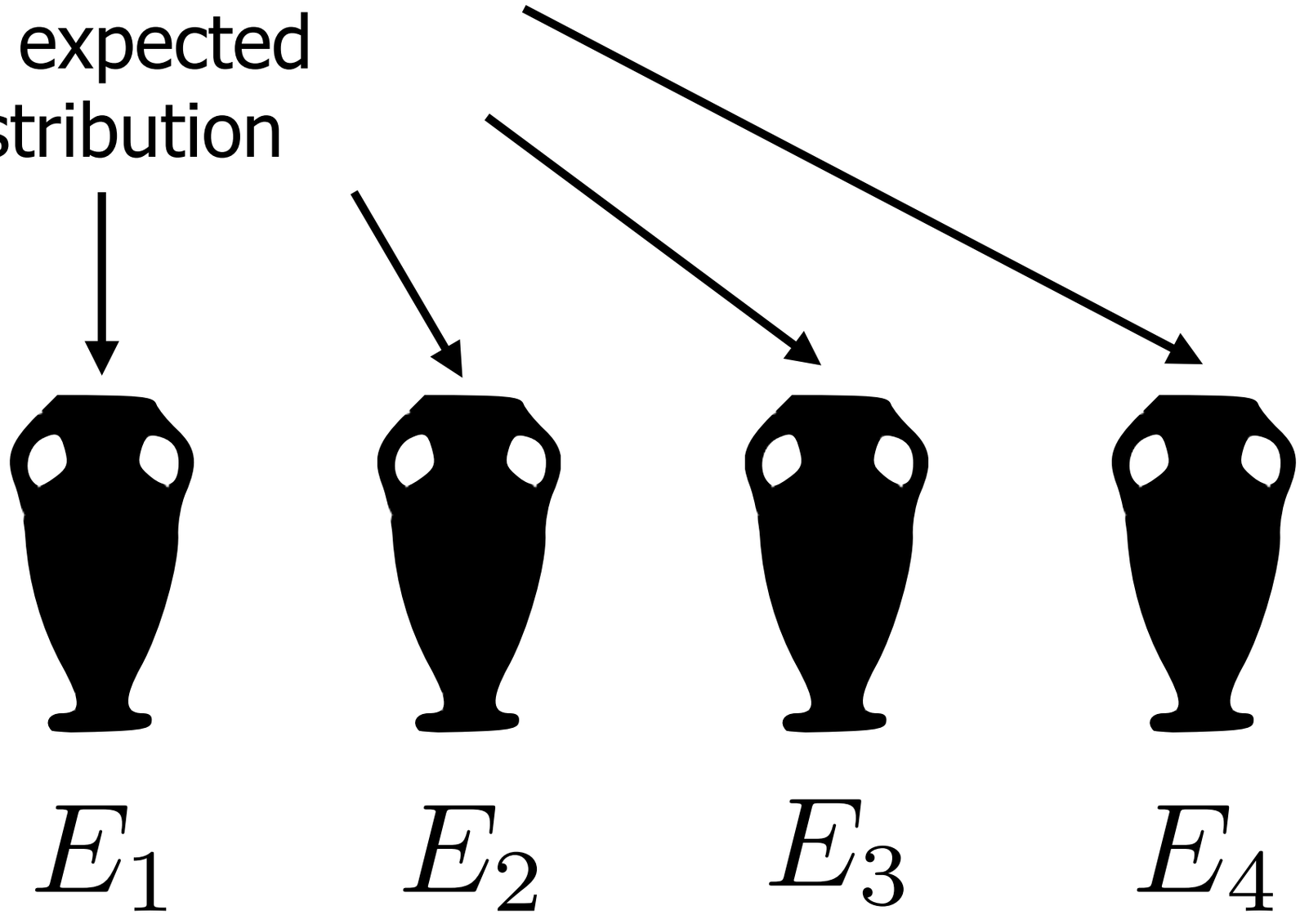
Consider a particular example where the marbles are 32



$O_1 = 4$ $O_2 = 5$ $O_3 = 8$ $O_4 = 15$

The O values are the numbers of marbles observed in the urns

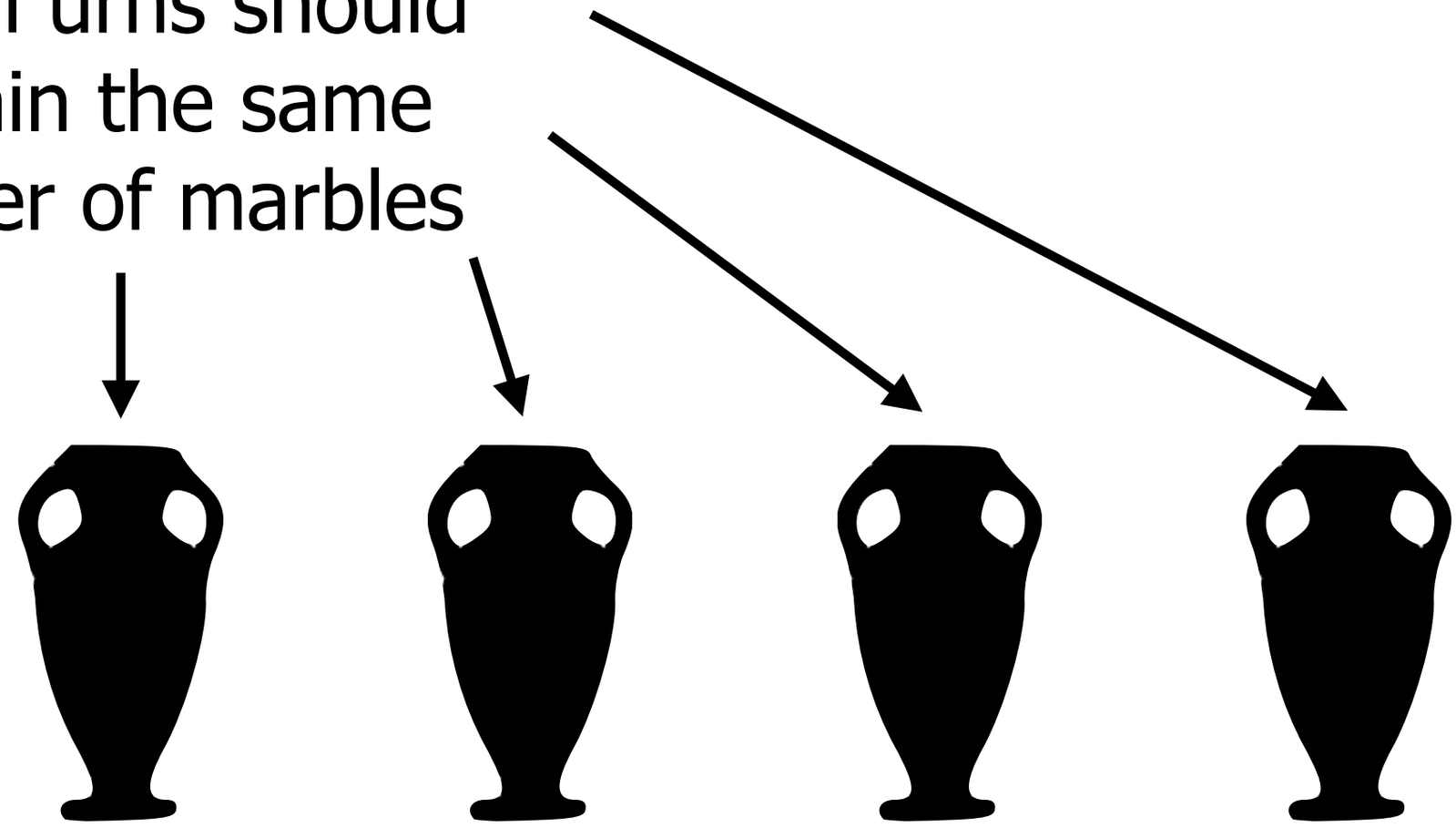
The null hypothesis is the expected distribution



The E values are the numbers of marbles expected in the urns

The question is whether O and E values are different

The null hypothesis is that all urns should contain the same number of marbles



$$E_1 = E_2 = E_3 = E_4 = 8$$

The E values are all the same to reflect the null hypothesis

The E values are set according to the null hypothesis

The Chi Square

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

This variable tests whether there is a matching between the observations (O) and the expectations (E)

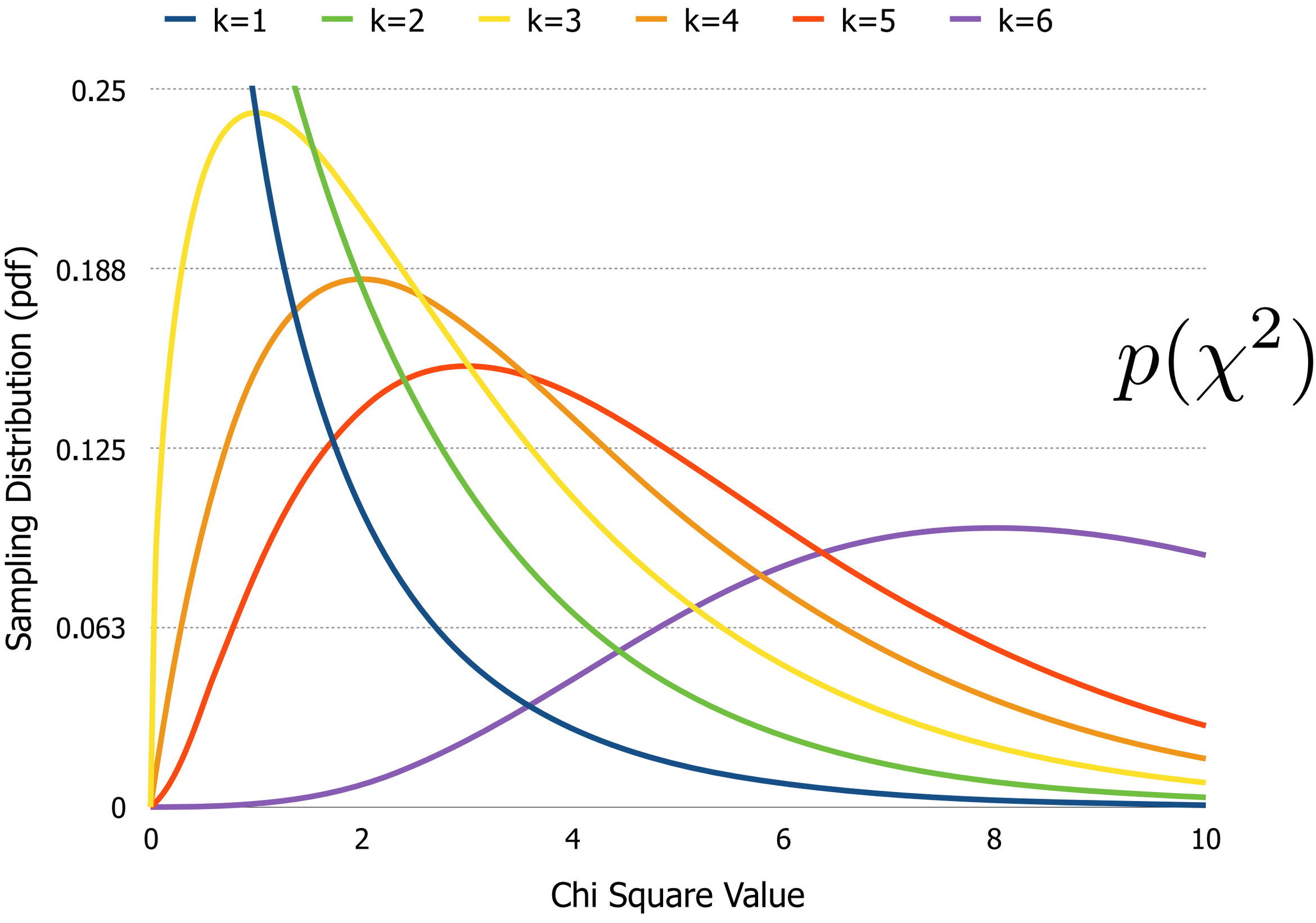
Sum over all values being compared

The probability density function is known when the null hypothesis is true

The parameter “k” corresponds to the degrees of freedom

$$p(\chi^2) = \frac{1}{2^{\frac{k}{2}}\Gamma\left(\frac{k}{2}\right)} \left(\chi^2\right)^{\frac{k}{2}-1} e^{-\frac{1}{2}\chi^2}$$

The value of “k” corresponds to the number of observations decreased by one


k=1
k=2
k=3
k=4
k=5
k=6
0.25
0.188
0.125
0.063
0
Sampling Distribution (pdf)
0
2
4
6
8
10
Chi Square Value
$p(\chi^2)$

Degrees of Freedom

- The values of the Observations have to respect constraints;
- In the case of the One Way Classification, the constraint is the sum, no more than 32 mice (or marbles) can be observed;
- If there are N observations, the number of degrees of freedom is then N-1.

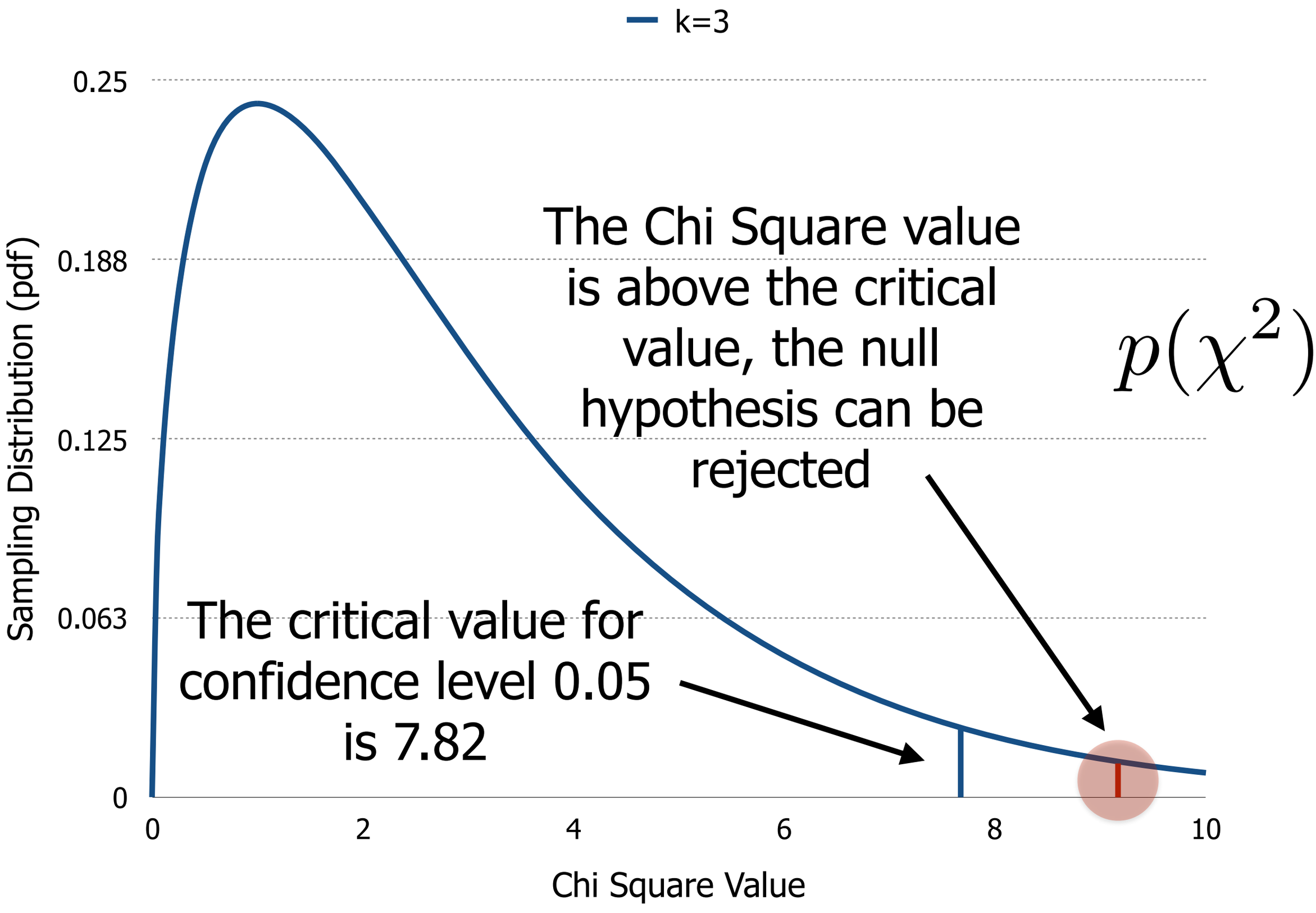
The O values are inserted in the expression of the Chi Square variable

The E values are inserted in the expression of the Chi Square variable

$$\chi^2 = \frac{(4-8)^2}{8} + \frac{(5-8)^2}{8} + \frac{(8-8)^2}{8} + \frac{(15-8)^2}{8}$$

$$\chi^2 = 9.25$$

The Chi Square is a random variable and its value depends on the O and E values, with the O being observed and the E reflecting the null hypothesis


k=3
0.25
0.188
0.125
0.063
0
Sampling Distribution (pdf)
0
2
4
6
8
10
Chi Square Value
The Chi Square value is above the critical value, the null hypothesis can be rejected
The critical value for confidence level 0.05 is 7.82
$p(\chi^2)$

Fake Example

- The outcome of the test depends on both Observations and Expectations;
- Imagine a (fake) experiment in which the expectations are different because the apparatus is different (e.g., there is food at the entrance of the alleys);
- The observations might remain the same, but the expectations change.

The O values are inserted in the expression of the Chi Square variable

The E values are inserted in the expression of the Chi Square variable

$$\chi^2 = \frac{(4-5)^2}{5} + \frac{(5-5)^2}{5} + \frac{(8-9)^2}{9} + \frac{(15-13)^2}{13}$$

$$\chi^2 = 0.61$$

The Chi Square is a random variable and its value depends on the O and E values, with the O being observed and the E reflecting the null hypothesis

k=3

The Chi Square value is below the critical value, the null hypothesis fails to be rejected

$p(\chi^2)$

The critical value for confidence level 0.05 is 7.82

Sampling Distribution (pdf)

0.25
0.188
0.125
0.063
0

0
2
4
6
8
10

Chi Square Value

Recap

- The value of the Chi Square depends on the data (through the O values) and on the null hypothesis (through the E values);
- The O values cannot be changed because they correspond to the data observed in an experiments;
- The E values must be set according to the null hypothesis to be tested.

Outline

- Very Quick Recap
- Goodness of Fit Test (One Way Classification)
- Two Way Classification
- Conclusions

Research Hypothesis

- Research Hpothesis: An aggressor tends to be considered guilty more frequently when the victim appears to be less faulty;
- Null Hypothesis: An aggressor tends to be considered guilty irrespectively of how faulty the victim appears to be.

Contingency Tables (I)

Fault	Guilty	Not Guilty	Total
Low	153	24	177
High	105	76	181
Total	258	100	358

R_1

R_2

C_1 C_2

Marginals

Pugh, "Contributory fault and rape convictions: Loglinear models for blaming the victim", Social Psychology Quarterly, 46(3):233-242, 1983

Contingency Tables (II)

- The elements of the table are Observations (how many times two characteristics coexist in a sample);
- The table shows whether one characteristic (a variable) is contingent or associated to the other;
- The sums over the values of a row or a column are called marginals.

Expected value in cell "ij" when the null hypothesis is true

Total number of "marbles" in row "i"

$$E_{ij} = R_i \frac{C_j}{N}$$

"N" is the total number of marbles in the table

Fraction of "marbles" in column "j"

Expected value in cell "ij" when the null hypothesis is true

Total number of "marbles" in column "j"

$$E_{ij} = \frac{R_i}{N} C_j$$

"N" is the total number of marbles in the table

Fraction of "marbles" in row "i"

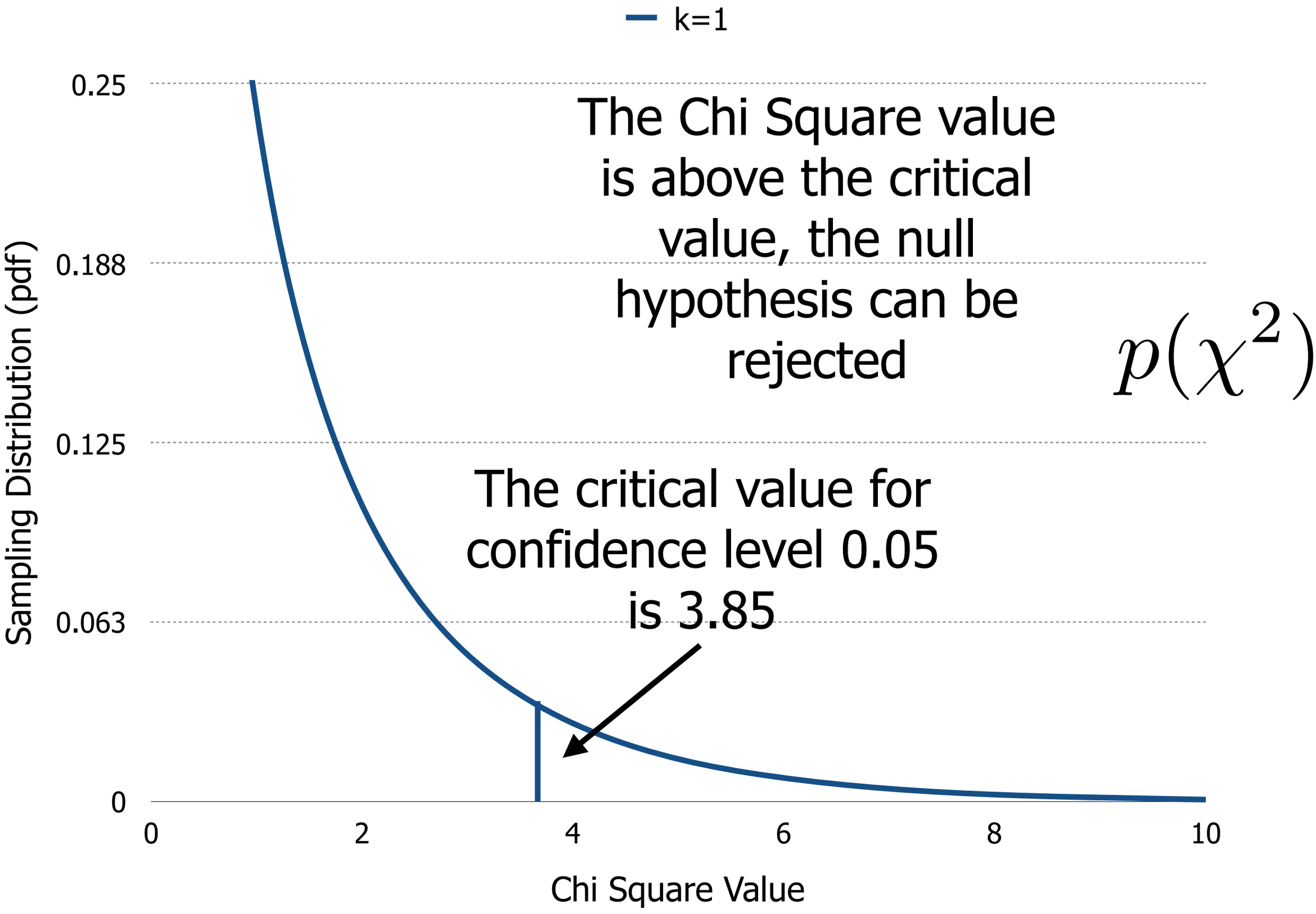
Number of Rows

Number of Columns

$$k = (R - 1)(C - 1) = 1$$

$$\chi^2 = \sum_{i=1}^{R} \sum_{j=1}^{C} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = 35.9$$

Sum over all elements of the contingency table


k=1
0.25
0.188
0.125
0.063
0
Sampling Distribution (pdf)
0
2
4
6
8
10
Chi Square Value
The Chi Square value is above the critical value, the null hypothesis can be rejected
$p(\chi^2)$
The critical value for confidence level 0.05 is 3.85

Outline

- Very Quick Recap
- Goodness of Fit Test (One Way Classification)
- Two Way Classification
- Conclusions

Conclusions

- The Chi Square test is useful when the observations take the form of counts (how many times an event of interest occurs);

- One way classification can show how well the observations fit the expectations;

- Two way classification can show how much two variables of interest are associated.

Thank You!