CATEGORICAL VARIABLES
CHI SQUARED TEST FOR ASSOCIATION
OR TEST FOR INDEPENDENCE

PLOUGHING THROUGH BIOMETRY

OBJECTIVES

Types of data- Two variables both categorical and categorical

Appreciate the type of data, how to summarise and undertake Chi square tests

Data Summary for count data

1.Chi squared test for independence

Principle, and example

OVERVIEW

	Type of Predictors		
Type of Response	**Categorical**	**Continuous**	**Categorical and Continuous**
Continuous	Analysis of Variance	Linear Regression	
Categorical	?		

OVERVIEW

	Type of Predictors		
Type of Response	**Categorical**	**Continuous**	**Categorical and Continuous**
Continuous	Analysis of Variance	Linear Regression	
Categorical	Test of association		

COUNT DATA
EXAMPLE OF SOME FLOWER COLOURS FROM A GENETICS EXPERIMENT.

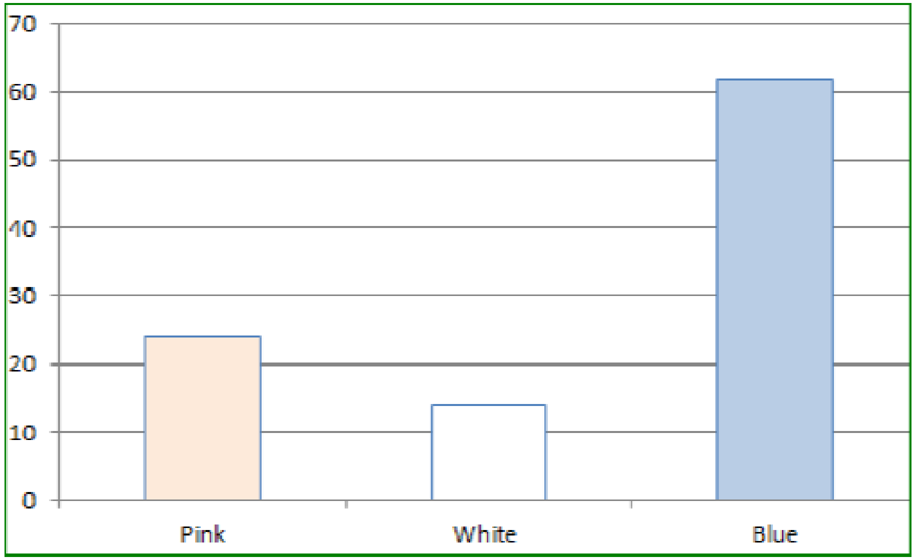


Figure 59: Flower colour for testing a ratio of 3 : 2 : 5

Colour	Pink	White	Blue	Total
Number of plants	24	14	62	100

COUNTS AND PROPORTIONS

Proportion = count / total

These are given as values between 0 and 1

Calculate the proportion for this data:

Colour	Pink	White	Blue	Total
Number of plants	24	14	62	100

WHAT HAVE WE LEARNED SO FAR

- Understand that for type of analysis we need a count. This is made up of the frequency of items within a category.
- Understand the terminology of count, and understand how to look at is as a proportion , A proportion is a value between 0 and 1.

TYPES OF DATA
…. APPROPRIATE WITH THIS METHOD

Binary (such as presence /absence)

Nominal (categories with no order)

Ordinal (categorise with an order)

Many types of research questions:

Are the wildlife species associated with particular vegetation types ?

Is the infection of cattle associated with whether they have been vaccinated or not.

Is my presence and absence of a weed species associated with whether I sprayed a herbicide or not?

BINARY PROPORTIONS AS RESPONSE DATA

Scientist may work with the binary data of the success/failure type.

For example, binary data are used in the following

-- detection test of certain bacteria which gives the

results in the false/true form;

-- in a food sensory panel which asks consumers to decide

between liking/disliking a certain test product and so on.

WHAT IS A FREQUENCY TABLE

- A frequency table shows the number of observations that fall in certain categories or intervals. A one-way frequency table examines one variable.
- Here we are looking at a categorical variable of income

Income	Frequency	Percent	Cumulative Frequency	Cumulative Percent
High	155	36	155	36
Low	132	31	287	67
Medium	144	33	431	100

HERE WE HAVE ANOTHER VARIABLE - HEALTH STATUS

- In this table we have added a column showing the cumulative frequency. Since Health Status has some sort of order from poor health to Excellent health this is called an *ordinal categorical variable*.

Health status	Number (f)	Cumulative freq (f/n)	Cum. number	Cum. frequency
Excellent	19	0.38	19	0.38
Very Good	12	0.24	31	0.62
Good	9	0.18	40	0.80
Fair	6	0.12	46	0.92
Poor	4	0.08	50	1.0
Total (n)	50	1.00		

DATA CAN BE ARRANGED AND PRESENTED AS A TABLE

	Columns		
Rows	A cell, contains the count for row 1 and column 1		
			Total

TERMINOLOGY

Classes of data - for the groups

Frequencies of data in the cells

Row, columns, cells, margins

Expected values

Observed values

Contingency table

EXAMINING AND SUMMARISING CATEGORICAL VARIABLES

By examining the distribution of categorical variables, you can

- determine the frequency of data values
- recognize possible associations among variables.
- Test whether the counts in your sample fit an expected distribution (*This is Goodness of Fit, and not covered in this presentation)*

LETS LOOK AT A TWO WAY TABLE

CHI SQUARE TEST FOR INDEPENDENCE

2 BY 2 CONTINGENCY TABLE

		Characteristic A		
		Yes	No	Total
	Yes	n_1	n_2	$n_1 + n_2$
Characteristic B	No	n_3	n_4	$n_3 + n_4$
	Total	$n_1 + n_3$	$n_2 + n_4$	$N = n_1 + n_2 + n_3 + n_4$

- They are also referred to as r x c tables, row by columns.

EXPECTED FREQUENCIES AND GENERAL CONTINGENCY TABLES

The table 20 shows the form of a general contingency table with cells (n_1, n_2, n_3 and n_4), and row and column totals on the margins and a grand total (N) in the bottom corner. The row totals sum across columns and column totals sum over rows.

The table contains rows and columns, and the principle of analysis is assume under the null hypothesis that the proportions in the different columns are the same as in the different rows. The expected frequencies for a particular row and column are:

$$\text{expected frequency} = \frac{\text{column total}}{\text{overall total}} \times \text{total in row} \tag{113}$$

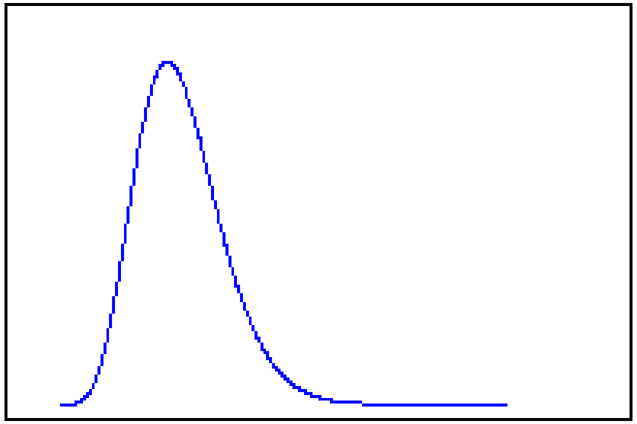
Chi-square (χ^2) distribution

A common distribution used in tests of statistical significance to:

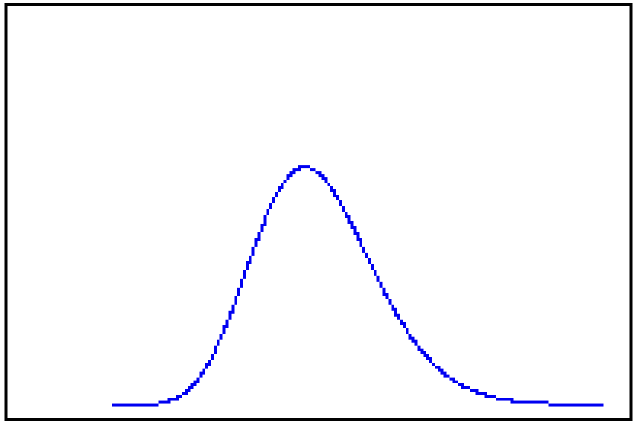
- Test how well a sample fits a theoretical distribution. For example, you can use a goodness-of-fit test to determine whether your sample data fit a Poisson distribution.
- Test the independence between categorical variables. For example, a manufacturer wants to know if the occurrence of four types of defects (missing pin, broken clamp, loose fastener, and leaky seal) is related to shift (day, evening, overnight).

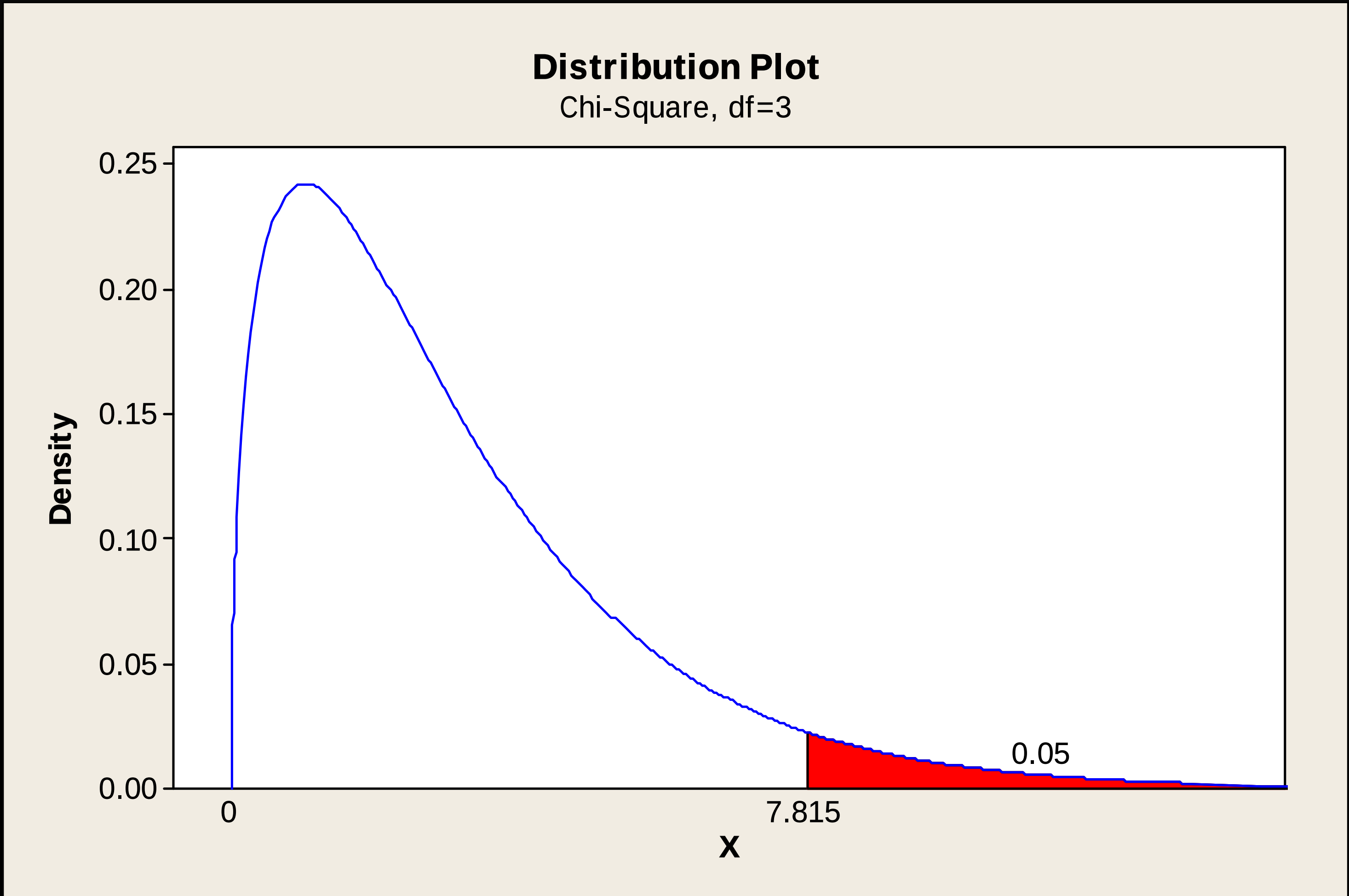
The shape of the chi-square distribution depends on the number of degrees of freedom. The distribution is positively skewed, but skewness decreases with more degrees of freedom. When the degrees of freedom are 30 or more, the distribution can be approximated by a normal distribution.

Chi-square distribution with 20 degrees of freedom



Chi-square distribution with 40 degrees of freedom



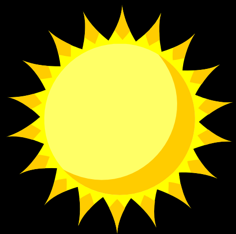

Distribution Plot
Chi-Square, df=3
Density
0.25
0.20
0.15
0.10
0.05
0.00
0
7.815
0.05
X

OBJECTIVE

To perform a chi-square test for independence

Lets have a look at what we are testing

NO ASSOCIATION - AN EXAMPLE. THE NUMBER OF COUNTS OF DAYS WHEN YOUR FRIEND WAS GRUMPY RELATED TO THE TWO TYPES OF WEATHER.

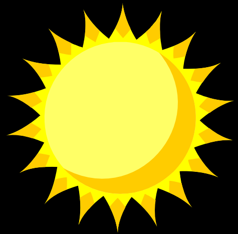


72%   28%

72%   28%

- Is your friend’s mood associated with the weather?

ASSOCIATION

		
	82%	18%
	60%	40%

Is your friend’s mood associated with the weather here ?

ASSUMPTIONS FOR THE DATA

- Present the percentages
- Analyse the counts
- The observations are independent
- Each experimental unit occurs only once in the table
- The smallest *expected value* for a cell should be greater than 5.

- There are *exact tests* for small samples.

(For small samples can use Fishers Exact test – not used very often and not covered in this website)

ASSOCIATION- WHAT DOES IT MEAN

- An association exists between two variables if the distribution of one variable changes when the level (or value) of the other variable changes.

- If there is no association, the distribution of the first variable is the same regardless of the level of the other variable.

DATA FROM A WILDLIFE STUDY

- Two regions of different vegetation, and two levels of abundance of a mammal species.
- From our data we want to test if the species is associated with a particular type of vegetation?

CONTINGENCY TABLE FOR REGION BY ABUNDANCE OF SPECIES

Two way table in the worksheet. Also called a *2 by 2* contingency table showing abundance of species

	Abundant Areas	Sparse Areas	Total number of areas
Region 1	14	6	20
Region 2	8	12	20
Combined Regions	22	18	40

RELATING TO OUR QUESTION, WE CAN ALSO LOOK AT THE PERCENT WITHIN A REGION

	Abundant Areas	Sparse Areas	Total number of areas
Region 1	14 (70%)	6 (30%)	20
Region 2	8 (40%)	12 (60%)	20
Combined Regions	22	18	40

HYPOTHESIS

A 2×2 contingency table (see 21)for areas and abundance of a species, show the data to be analysed. The research question is to check whether species abundance is associated with or independent of the regions The hypothesis states:

H_0 there is no association between abundance and region.

H_a there is an association between abundance and region.

If there is no association then there is independence.

In this example we are interested in whether the 70% in Region 1 is the same as 40% in region 2 by pure chance or is it a real difference between regions.

EXPECTED VALUES FOR A CELL

$$\text{expected frequency} = \frac{\text{column total}}{\text{overall total}} \times \text{total in row}$$

CALCULATING O-E

- Notation – observed is your data values
- The four cells have been rearranged in this table to show the figures to calculate
- Step 1 is to get the difference of observed and expected

	Observed	Expected	Observed-Expected
Region 1, abundant	14	11	+3
Region 1, sparse	6	9	-3
Region 2, abundant	8	11	-3
Region 2, sparse	12	9	+3

NEXT STEP CALCULATE THE CHI SQUARE STATISTIC

Follow through with the last columns calculations and add all four as shown below to get Chi square.

	Observed	Expected	Observed-Expected	$(O-E)^2/E$
Region 1, abundant	14	11	+3	
Region 1, sparse	6	9	-3	
Region 2, abundant	8	11	-3	
Region 2, sparse	12	9	+3	

$$\chi^2 = \frac{+3^2}{11} + \frac{-3^2}{9} + \frac{-3^2}{11} + \frac{+3^2}{9}$$

CALCULATING CHI SQUARE

The test statistic is calculated by summing these values to get the χ^2 statistic.

$$\chi^2 = \sum \frac{(O - E)^2}{E} \tag{115}$$

This is calculated in equation 116:

$$\chi^2 = \frac{+3^2}{11} + \frac{-3^2}{9} + \frac{-3^2}{11} + \frac{+3^2}{9} \tag{116}$$

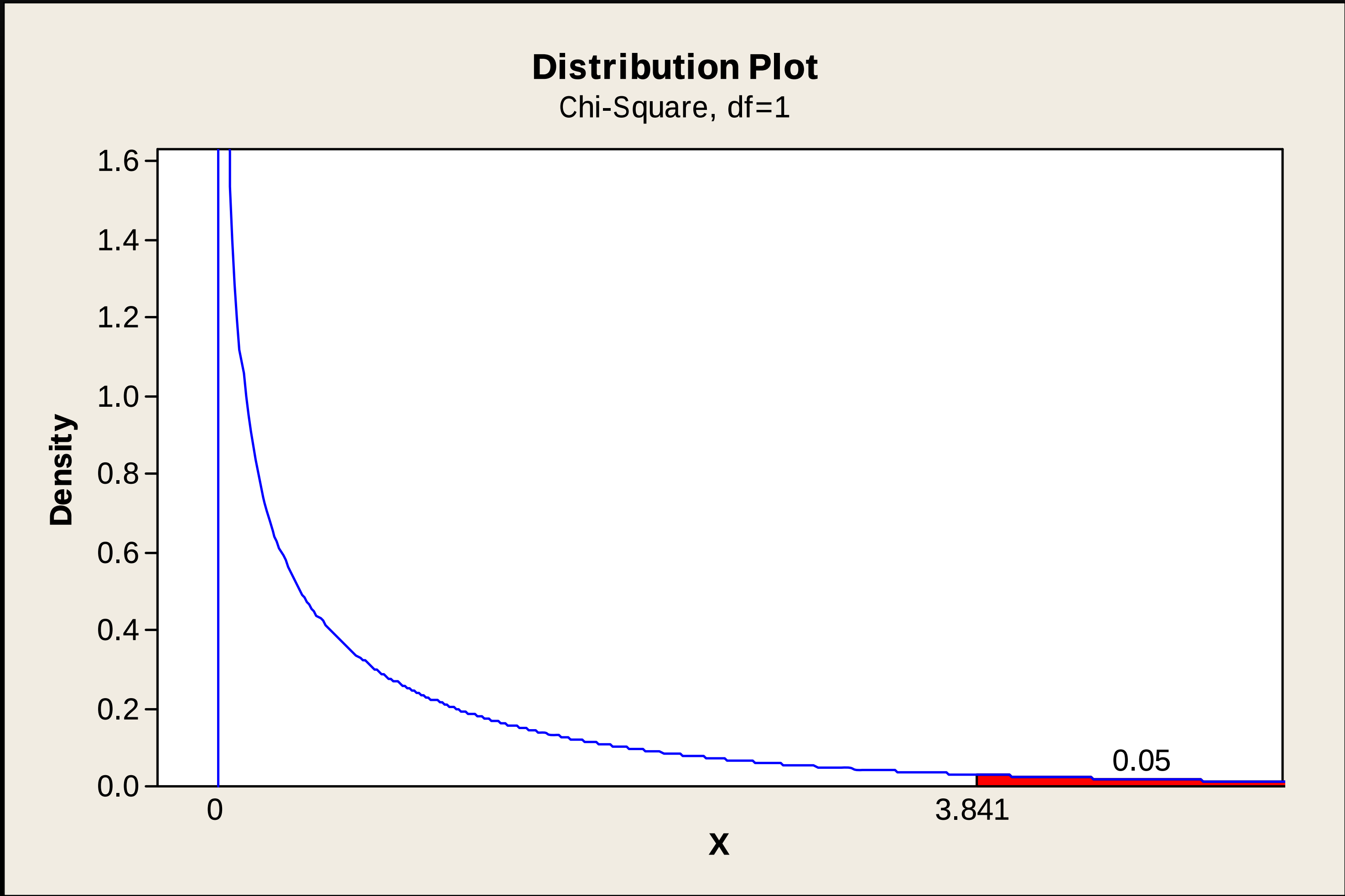
$$\chi^2 = 0.818 + 1.00 + 0.818 + 1.00 = 3.636 \tag{117}$$

CHI SQUARE DEGREES OF FREEDOM

The df = (number of rows-1) $\times$ (number of columns-1) $= (2-1) \times (2-1) = 1$
So we consider a χ^2 distribution on 1 df for a 2×2 table.

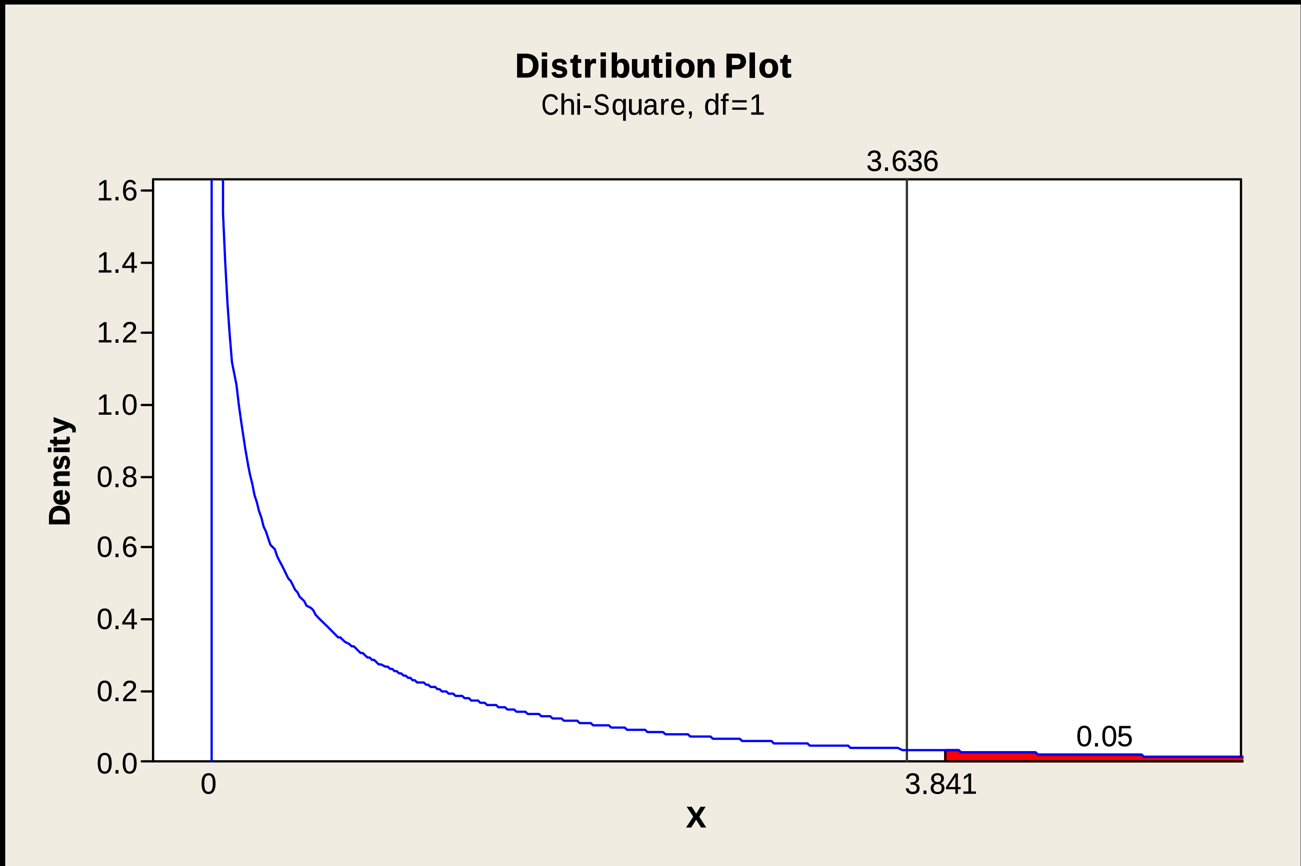
$$\chi^2 = 0.818 + 1.00 + 0.818 + 1.00 = 3.636$$

CHECK WITH THE CHI SQUARE DISTRIBUTION ON 1 DF. WHAT IS THE CRITICAL VALUE? 3.841



CHECK OUR CONCLUSIONS

- Our Calculated value is 3.636 and since it lies in the acceptance region, we accept out null hypothesis of no association.
- If we wanted to do another study, we may decide to count more species as a larger sample size would give more power to the analysis.



SUMMARY OF CHI-SQUARE TESTS

Chi-square tests and the corresponding *p*-values

- determine whether an association exists

Chi-square tests and the corresponding *p*-values

- do not measure the strength of an association
- depend on and reflect the sample size.