**AN EXAMPLE BASED ON CHI SQUARE AND ENTROPY**

**What is chi square test?**

The c2 test is used to determine whether an association (or relationship) between 2 categorical variables in a sample is likely to reflect a real association between these 2 variables in the population.

EXAMPLE:

A group of students were classified in terms of personality (introvert or extrovert) and in terms of colour preference (red, yellow, green or blue) with the purpose of seeing whether there is an association (relationship) between personality and colour preference. Data was collected from 400 students and presented in the 2 (rows) x 4 (cols) contingency table below:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **(Observed counts)** | **Colours** | | | | |
|  | **Red** | **Yellow** | **Green** | **Blue** | **Totals** |
| **Introvert personality** | 20 | 6 | 30 | 44 | 100 |
| **Extrovert personality** | 180 | 34 | 50 | 36 | 300 |
| **Totals** | 200 | 40 | 80 | 80 | 400 |

Suitable null and alternative hypotheses might be:

* H0: Colour preference is not associated with personality, and
* H1: Colour preference is associated with personality

To perform a chi-squared test, the number of students expected in each cell of the table if the null hypothesis is true, is calculated.

The expected numbers (under the null hypothesis) in each cell are equal to

Thus for the introvert/red cell the expected number is

= 50

To calculate the chi-squared (chi 2) statistic the value of

needs to be calculated for each cell in the table. For the introvert/red cell this is

= 18.00

The chi-square statistic is calculated to be total of these values

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **(Expected counts)** | **Colours** | | | | |
|  | **Red** | **Yellow** | **Green** | **Blue** | **Totals** |
| **Introvert personality** | 50 | 10 | 20 | 20 | 100 |
| **Extrovert personality** | 150 | 30 | 60 | 60 | 300 |
| **Totals** | 200 | 40 | 80 | 80 | 400 |

From these expected and the observed values the chi-squared test-statistic is computed, and the resulting p-value is examined.

Computer Output

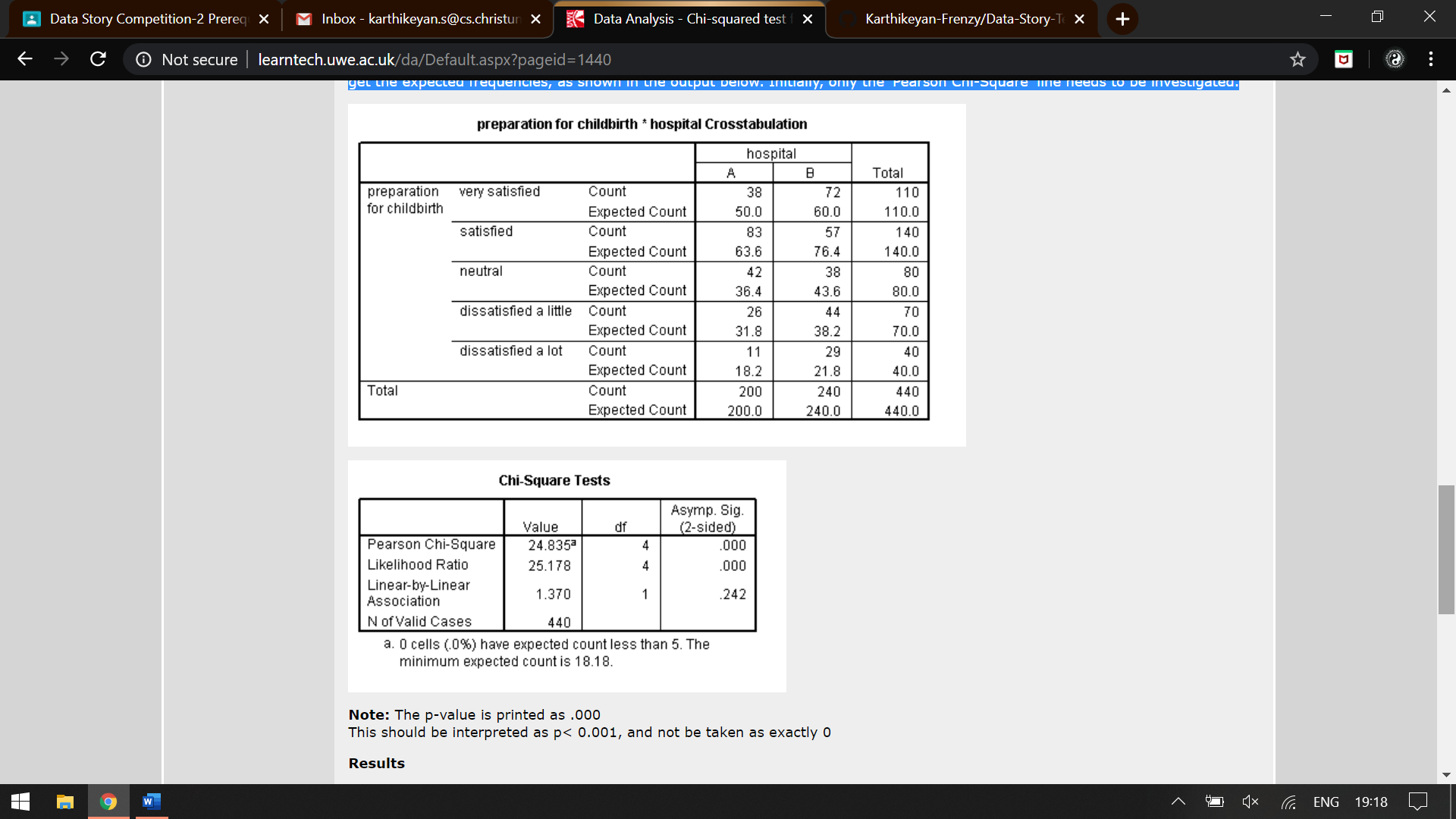
Chi-squared test in SPSS

Data should be entered in 2 columns, then select

Analyse >Descriptive Statistics>Crosstabs

SPSS can only be used for raw data

Some choices need to be made from the Statistics and Cells buttons in the dialogue box, to get the chi-squared test results, and to get the expected frequencies, as shown in the output below. Initially, only the 'Pearson Chi-Square' line needs to be investigated.



Note: The p-value is printed as .000

This should be interpreted as p< 0.001, and not be taken as exactly 0

Results

The chi-squared test statistic is 71.20 with an associated p < 0.001

Note: .000 should not be interpreted as exactly zero, as in the computer print-out.

The null hypothesis is rejected, since p < 0.001, and a conclusion is made that colour preference is associated with personality. Examining the pattern of numbers, it is noted that more introverts prefer blue than expected and less preferred red. The extroverts tend to favour red more than blue.

**What is entropy?**

**It is a measure of the randomness in the information being processed. The higher the entropy, the harder it is to draw any conclusions from that information.**

Probability Distribution Considering a categorical variable as a discrete random variable X, the probability mass function p(x) defines the probability that X is exactly equal to a categorical value, x. The functions for all possible categorical values thus define the frequency distribution of all the data records within this dimension:

= p(x)

where count(x) is the number of records has the value x and count(X) is the number of all records of X. Entropy The entropy is computed as

which provides a measure of uncertainty (in the context of information theory) of X. It provides information about the variation, or diversity, of the information contained in one data dimension. Joint Entropy Joint entropy is defined over two variables X and Y as

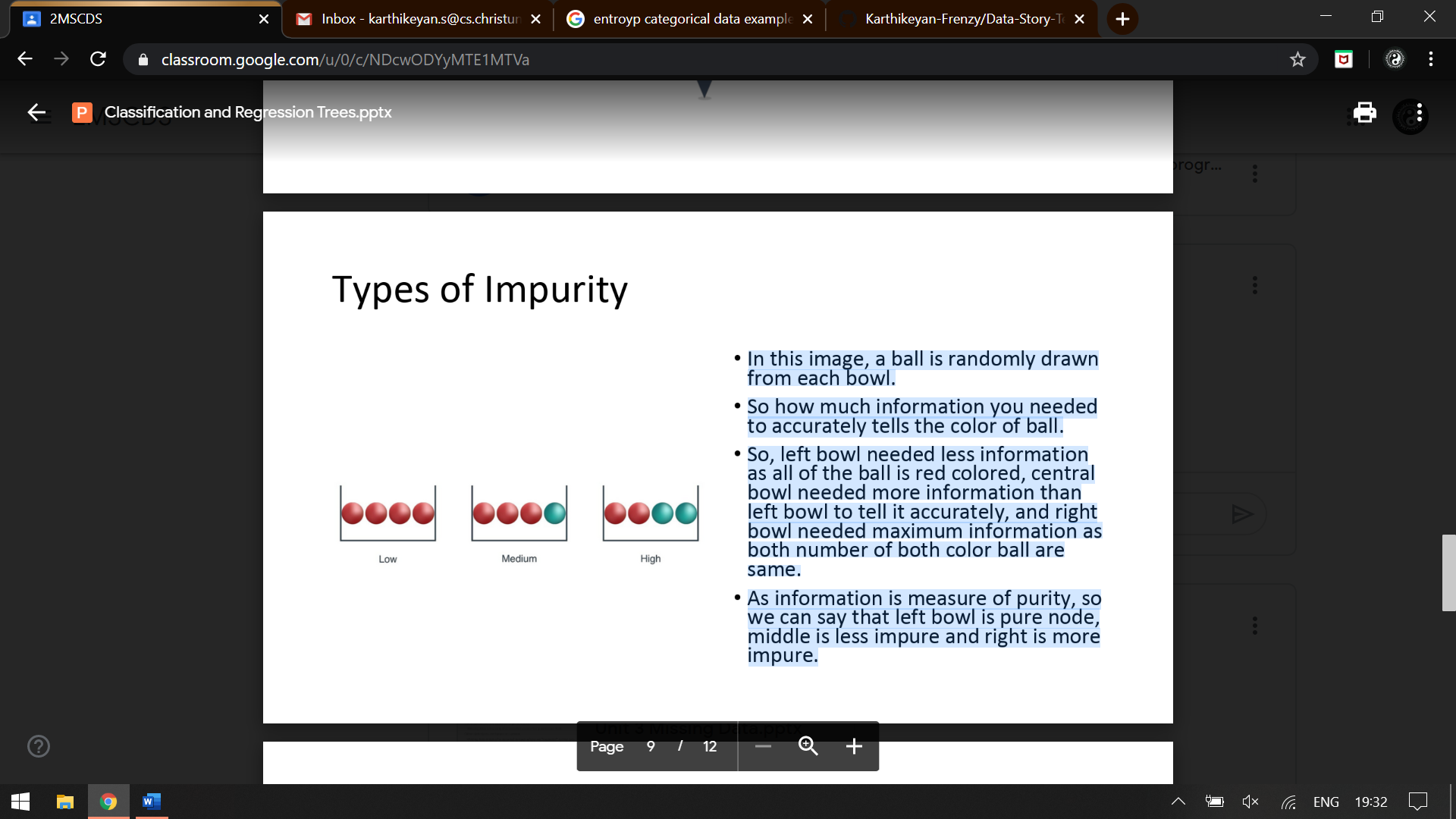
where x and y are particular values of X and Y, respectively. p(x, y) is the probability of these values occurring together. Joint entropy measures data diversity associated with two variables.

Entropy is amount of information is needed to accurately describe some sample.

So, if sample is homogeneous, means all the element are similar than Entropy is 0, else if sample is

equally divided than entropy is maximum 1.

So, left bowl has lowest entropy, middle bowl has more entropy and right bowl has highest entropy.



In this image, a ball is randomly drawn from each bowl.

So how much information you needed to accurately tells the colour of ball. So, left bowl needed less information as all of the ball is red coloured, central bowl needed more information than left bowl to tell it accurately, and right bowl needed maximum information as both number of both colour ball are

same. As information is measure of purity, so we can say that left bowl is pure node, middle is less impure and right is more impure.