Is education level associated with colour preference?

Using the chi-squared test in biology and psychology

Despite biology and psychology landing in separate IB groups, they share a number of mathematical methods that are designed to show that data are statistically signficant. **Germán Tenorio** investigates the use of one of these methods, the chi-squared test, showing how knowledge in the sciences can be said to be genuinely multidisciplinary

Exam context

IBDP biology and psychology students are expected to apply an inferential statistical test in their internal assessment to determine the significance of their results. In addition, the IBDP biology guide requires the use of the chi-squared test as a skill in subtopics 4.1 and 10.2. After reading this article you will be able to apply the chi-squared test to meet these requirements.

Although psychology and biology are not the same discipline, they have several things in common. One of them is the use of inference statistical tests when analysing results obtained by experimental research.

There are different inferential statistical tests that can be applied depending on what you are investigating and the type of data obtained, such as the t-test, Pearson test, chi-squared test and many others. These tests help you to determine if the results you have obtained are significant or not.

Inferential statistical tests

Biology and psychology students have to conduct experimental research for their internal assessments. In this investigation, you have to come up with a question that the experiment is trying to answer (the research question, RQ), where you state both the independent and dependent variables. An RQ in psychology could be: 'Do less empathetic people have lower academic achievements?' An RQ in biology could be: 'Do rabbits prefer to make their dens near pine trees?'

As you can see, research questions can be about populations, but there is a problem, as we can't analyse the entire population due to the lack of time. For this reason, the data we use to answer research questions come from relatively small samples from populations. We use descriptive statistics to summarise the data from a sample, such as the mean or standard deviation, among others. However, when we try to estimate values of the population that is represented by the sample, we need to use inferential statistics, as we are inferring information about the population from the sample.

There are two types of inferential statistical tests:

- **Parametric tests** are applied to data we assume have a normal distribution (or bell curve), such as the t-test and analysis of variance (ANOVA) test. A normal distribution is characterised as having most data around the mean, and two small groups of data on both ends, resembling a bell, which is symmetrical. In other words, these tests are applied to quantitative or numerical variables.
- **Nonparametric tests** such as the chi-squared test are applied to data without a normal distribution or when both variables are categorical (non-numerical).

Chi-squared test

The chi-squared test is a statistical hypothesis test that makes a comparison between the data collected in an experiment vs the data expected. For instance, it is frequently used to test how well the results of genetic crosses fit predictions based on Mendel's laws of inheritance. The aim is to determine whether the variation in the results from the expected values is due to chance or not.

The chi-squared test is also used to determine whether there is a significant association between two categorical variables, which is useful in psychology when you have to investigate an association between two factors, such as gender and higher studies preference, or education level and memory. In addition, it is also used in ecology to decide whether there is an association between two species that live closely, or whether this is due to a random distribution.

As in all hypothesis tests, a null as well as an alternative hypothesis must be formulated. The **null hypothesis (H_0)** states that there is no association between the two variables you are investigating. In other words, you assume that both variables are independent and the possible relation observed occurs purely by chance. The **statistical alternative hypothesis (H_a)** states that both variables are significantly dependent or related, so that one influences the result of the other.

Collecting data

Data must be collected by random sampling in order to apply the chi-squared test. Chi-square (χ^2) is calculated based on the following formula:

$$\chi^2 = \frac{\Sigma (\text{Observed} - \text{Expected})^2}{\text{Expected}}$$

The calculated chi-square value is looked up in a chi-square distribution table (Table 1). The first column in the table refers to degrees of freedom, which is related to the number of different levels for each variable. For example, it could be the number of possible phenotypes in a genetic cross or the number of possible education levels.

The first row in the table refers to the probability (*p*-value). A *p*-value of 0.05 means that there is a 5% chance that the relation observed between both variables is random and a 95% chance that this relation is real and that both variables are dependent.

If the calculated chi-squared value is greater than the value for *p* 0.05, the null hypothesis is rejected, so the relation observed between both variables is not due to random chance and therefore it is significant.

Table 1 Chi-square distribution table

Degrees of freedom	Levels of significance for one-tailed test			
	$\chi^2_{0.1}$	$\chi^2_{0.05}$	$\chi^2_{0.01}$	$\chi^2_{0.005}$
1	2.706	3.841	6.635	7.879
2	4.606	5.991	9.210	10.597
3	6.251	7.815	11.345	12.838
4	7.779	9.488	13.277	14.860
5	9.236	11.070	15.086	16.750
6	10.645	12.592	16.812	18.548
7	12.017	14.067	18.475	20.278
8	13.362	15.507	20.090	21.955

Questions and activities

1 One of your classmates is investigating whether there is an association between hand dominance and success of students in their extended essay (EE). He randomly obtained records of 100 EEs and grouped them in the following categories:
- Right-handed students: 14 A, 15 B, 19 C, five D and two E.
- Left-handed students: 14 A, 8 B, 12 C, eight D and three E.

Apply the chi-squared test to determine if both variables are independent.

2 Your biology teacher crossed two pure breeding plants of genotypes AABB and aabb. All the F1 were dihybrid (AaBb), showing the same phenotype. An F1 plant was then crossed with a homozygous recessive plant (aabb). The resulting F2 were 140 plants AaBb, 115 Aabb, 110 aaBb and 135 aabb. Apply the chi-squared test to determine if observed data differs from expected data.

Psychology example

Imagine you want to carry out an experiment for your internal assessment, investigating whether people's education level is associated to their colour preference.

You have surveyed a random sample of 96 people, and respondents were classified by education level (primary, secondary, university or doctoral) and by primary-colour preference (red, yellow or blue). Your collected data were as follows:
- Primary: 13 red, five yellow and five blue.
- Secondary: 12 red, six yellow and seven blue.
- University: six red, six yellow and nine blue.
- Doctoral: eight red, seven yellow and 12 blue.

It seems there is some kind of trend in the election of colour, as people with higher-level studies tend to choose blue more frequently than those with a lower education level. Are both variables dependent? Is it just due to the chance? The chi-squared test is applied to answer this question.

Step 1: state the null and alternative hypotheses

- Null hypothesis: the education level and colour preference are independent. The distribution we see occurred purely by chance.
- Alternative hypothesis: the education level and colour preference are not independent. Colour election depends on education level.

Step 2: calculate the expected data

The data collected are the observed data and have to be put in a contingency table (Table 2).

To calculate the expected data assuming independent distribution, multiply the row total by the column total for each of the four education levels and divide by the grand total. So for calculating

Table 2 Contingency table

Education level/ colour preference	Red	Yellow	Blue	Row total
Primary	13	5	5	23
Secondary	12	6	7	25
University	6	6	9	21
Doctoral	8	7	12	27
Column total	39	24	33	**TOTAL: 96**

the expected data for people with primary education choosing red, you just have to multiply the column total (39) by the row total (23) and divide by 96. We always assume independent distribution for the expected data because it is our null hypothesis: both variables are independent and the distribution we see occurred purely by chance.

Draw a new table with the observed and the calculated expected data. The sum of the observed data must coincide with the sum of the expected data for each colour (Table 3).

Table 3 Observed and expected data

Education level/ colour preference	Red		Yellow		Blue	
	Observed	Expected	Observed	Expected	Observed	Expected
Primary	13	9.34	5	5.75	5	7.9
Secondary	12	10.16	6	6.25	7	8.6
University	6	8.5	6	5.25	9	7.2
Doctoral	8	11	7	6.75	12	9.3
Total	**39**		**24**		**33**	

Step 3: calculate the chi-squared value.

To calculate the chi-squared value, just apply the formula

$$\chi^2 = \frac{\sum (\text{Observed} - \text{Expected})^2}{\text{Expected}}$$

$$\chi^2 = (13 - 9.34)^2/9.34 + (12 - 10.16)^2/10.16 + (6 - 8.5)^2/8.5 + (8 - 11)^2/11 + (5 - 5.75)^2/5.75 + (6 - 6.25)^2/6.25 + (6 - 5.25)^2/5.25 + (7 - 6.75)^2/6.75 + (5 - 7.9)^2/7.9 + (7 - 8.6)^2/8.6 + (9 - 7.2)^2/7.2 + (12 - 9.3)^2/9.3 = 6.141$$

The calculated χ^2 value is 6.14

Step 4: calculate degrees of freedom.

Degrees of freedom for a contingency table is always calculated by multiplying the number of rows in a table less one by the number of columns less one. In this case, there are four rows corresponding to education level and three columns corresponding to the three primary colours, so that the calculated degrees of freedom is $(4 - 1) \times (3 - 1) = 6$

Step 5: look up the calculated χ^2 value

With six degrees of freedom, the significant p 0.05 value in Table 1 has a χ^2 value of 12.592. If the calculated χ^2 value is greater than 12.592, the null hypotheses can be rejected, and the difference between expected and observed data are not just by chance and both variables are not independent. In this example, as the calculated χ^2 value (6.141)

Theory of knowledge

1 Does using statistical analysis techniques make the knowledge gained in the human sciences more reliable?
2 Is it possible for the human sciences ever to reach the level of reliability of the natural sciences?
3 This article suggests that to be a genuine member of the community of biologists, you need to have a fairly strong mathematical background. What other non-biological knowledge do you think you would need to be an expert biologist? Make a list, and interview biology teachers to see what they say. Do you agree?

References and resources

Use of the chi-squared test in Mendelian genetics: **www.tinyurl.com/ybrc3fap**

Use of the chi-squared test in psychology: **www.tinyurl.com/y8q3fofc**

The chi-squared test can be used to find out if there is an association between two plant species that usually live close to each other

Table 4 Contingency table

	Species Y present	Species Y absent	Row total
Species X present	55	3	58
Species X absent	17	25	42
Column total	72	28	**TOTAL: 100**

Table 5 Observed and expected data

	Species Y present		Species Y absent	
	Observed	**Expected**	**Observed**	**Expected**
Species X present	55	41.76	3	16.24
Species X absent	17	30.24	25	11.76
Total	**72**		**28**	

is not greater than 12.592, we can affirm that the differences between the observed and expected results are not statistically significant, so the education level and colour preference are independent.

Biology example

Imagine you are investigating, as part of subtopic 4.1, if there is an association between two plant species that usually live close to each other. You recorded species presence or absence in 100 quadrats (a square frame with a surface of 1 m^2). Data were as follows:

- species X growing alone: 3
- species Y growing alone: 17
- both species X and Y growing together: 55
- neither species X nor Y growing: 25

Step 1: state the null and alternative hypotheses

- Null hypothesis: there is no association between both species. The distribution observed occurred by chance.
- Alternative hypothesis: there is a real association between both species, so species X usually lives close to species Y.

Step 2: calculate the expected data

The data collected are the observed data and have to be put in the contingency table where the numbers of quadrats containing or not containing the two species are shown (Table 4).

To calculate the expected data assuming independent distribution, multiply the row total by the column total for each of the four species combinations and divided by the grand total. So for calculating the expected data for both absent, you just have to multiply the column total (28) by the row total (42) and divide by 100.

Draw a new table with the observed and the calculated expected data (Table 5).

Step 3: calculate the chi-squared value

To calculate the chi-squared value, apply the formula:

$$\chi^2 = (55 - 41.76)^2/41.76 + (17 - 30.24)^2/30.24 + (3 - 16.24)^2/16.24 + (25 - 11.76)^2/11.76 = 35.695$$

The calculated χ^2 value is 35.695.

Step 4: calculate degrees of freedom

In this example, there are two rows (species Y present and absent) and two columns (species X present and absent), so that the calculated degree of freedom is $(2 - 1) \times (2 - 1) = 1$

Step 5: look up the calculated χ^2 value

With one degree of freedom, the significant p 0.05 value has a χ^2 value of 3.841. If the calculated χ^2 value is greater than 3.841, the null hypothesis can be rejected, and the difference between expected and observed data are not just by chance and species X is associated with sites where species Y lives more than expected by chance alone. In this example, as the calculated χ^2 value (35.695) is greater than 3.841, we can affirm that the association between species X and Y is statistically significant.

Key points

- Inferential statistical tests are used to determine if the values estimated for a population represented by a sample are statistically significant.
- The chi-squared test is a nonparametric test that is applied to categorical or non-numerical data.
- In the chi-squared test, the null hypothesis (H_0) states that both variables you are investigating are independent and the possible relation observed occurs purely by chance. The alternative hypothesis (H_a) states that both variables are significantly dependent.
- When the calculated χ^2 value is greater than the significant probability 0.05 value, the null hypothesis is rejected, concluding that both variables are not independent.

Germán Tenorio has a PhD in molecular biology and teaches IB biology at Colegio Internacional San Francisco de Paula, Spain. He is also an IB examiner and workshop leader.