χ^2 The Chi-Squared Test of Independence χ^2

This test determines if two types of *categorical* variables are associated with each other. (if not associated then they are independent of one another). The test does <u>not</u> prove causation. *There are two ways to run the test:*

A. *using the χ^2 statistic*   **B.** Using a probability value, *p*

Note: The term *correlation* is not appropriate to use with this test or with categorical data.

A. The Chi Squared Test of Independ. (using the χ^2 statistic)

Follow the steps below if performing the Chi-Square Test of Independence *for a project* or *for an assignment*.

> 1. **Make a Contingency Table and State the Null Hypothesis**. Place the data into a contingency table showing the respective observed frequencies. (must be integers)

Here is an example question:

A student was going to investigate whether favorite subject *is associated* with gender. So, the student gave his 150 classmates a questionnaire to fill out. The results for the question on the gender of the student and their favorite subject are given in the table below, which is a 2×3 contingency table of **observed** frequencies.

State a Null Hypothesis (H$_0$) and an Alternative Hypothesis (H$_1$).

    H$_0$: Favorite subject is **independent** of gender.

    H$_1$: Favorite subject and gender are **associated** (or you can say Favorite subject and gender are <u>not</u> independent). Note: Being associated does <u>not</u> mean one variable **causes** the other variable to happen.

Then, show a Contingency Table

	History	Biology	French	
Female	22	20	18	
Male	20	11	9	

> 2. Find row and column totals. Then <u>Calculate the **expected** values</u> in a separate table or combined table. You will need column and row frequency totals to get you started.

	History	Biology	French	
Female	22	20	18	(60)
Male	20	11	9	(40)
	(42)	(31)	(27)	TOTAL 100

Expected values:

	History	Biology	French
Female	**n**	18.6	16.2
Male	q	r	10.8

The expected values for each cell are calculated from row and column totals from the original contingency table as follows:

To find n for example : $\dfrac{60}{100} \cdot \dfrac{42}{100} \cdot 100 = 25.2$ which is equal to $\dfrac{Row\ total \cdot Column\ total}{n}$

On IB exams, you will want to know how to produce these values using the above method _and_ from your GDC

3. __Check if the test will be valid__ (results trustworthy?):
   All **EXPECTED** values must be greater than 5. If not, the test will be _invalid_.   Do not continue the test.

   If your contingency table is larger than a 2 by 2, then you might be able to condense the number of columns or rows in a sensible way and re-start the test. If so, be sure to show your new contingency table and then re-start step 2. (On projects -- be sure to discuss the reasoning behind your new category divisions). To avoid this problem, always collect an abundance of data !

4. State **significance level**. This is usually 5%. _Values such as 1% or 10% are also common. On most problems this value will be given to you. If not, choose 5%._

   EXAMPLE STATEMENT: **_"This test will be run using a 5% significance level"._**

5. Find the number of __degrees of freedom__ and the __critical value__. This is simple:

   d.f. = **(number of rows – 1) x (number of columns – 1)**, in our example, (2-1) x(3-1) = 1 x 2 = __2__

   The critical value: Will be given to you on IB exams. On projects and problems from class you will have to cite and use a table.

   a.   Go down the left column until you reach the number of degrees of freedom (2 in this example)
   b.   Go across until you reach the column 0.95 and read off your value. It is 5.99
        (note: we use 0.95 because there is usually a 5% significance level and 0.95 is 100% – 5%)
        Note: If the test was being run at a 1% contingency level, the go across to the 0.01 column.

6. State the _rejection inequality_

   "If x^2 > 5.99 , then I will reject H_0 "

   Translation: _You will reject the null hypothesis if the chi-square statistic (that you calculate on the next step) is greater than the critical value from the table. (This means the differences between the expected and observed values are large enough to reject independence. If not, you will accept the null hypothesis if the chi-square statistic is not greater than the critical value._

7. Calculate the **Chi Squared Statistic**
   You need to know how to calculate this statistic using the formula and by your GDC, depending on the question.

a) By "formula" (you may at times need to show a process with the formula)

$$X^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

where f_o = observed frequency (*a natural number*)

f_e = expected frequency (*a natural number*)

1. For each cell in the table, subtract the *expected value* <u>from</u> the *observed value* and square the difference. Divide this by the expected value.
2. Find the sum, which is the *Chi-Square* statistic.
3. On a project this process needs to be shown clearly.

b) With your GDC:

a. Choose **MATRIX** and go to **EDIT**
b. Make sure your matrix is the right size
c. Enter your **Observed** values in **Matrix A**
d. Choose **STAT** and go to **TESTS**
e. Scroll down to χ^2-**Test** and press **ENTER**
f. Choose **Calculate**.
g. Your **expected** values can now also be found in **Matrix B**

.............................Ok, in our favorite subject/gender example: χ^2 = *1.78*

Note when working on projects: If the degree of freedom is 1 (in other words for *all 2 by 2 tables*) the ***Yates Continuity Correction*** <u>must</u> be used. In this case, do <u>not</u> use the statistic from your calculator. It will be invalid. here is the revised formula.

If $df = 1$, we use

$$\chi^2_{calc} = \sum \frac{(|f_o - f_e| - 0.5)^2}{f_e}$$

where $|f_o - f_e|$ is the **absolute value** or **modulus** of $f_o - f_e$

---

8. **Make two summary statements** about the test.

A. First make a statement, in a complete sentence, using the following guidelines:

> If the χ² value is **less than or equal to** the critical value, we **fail to reject the null hypothesis**.
>
> or if If the χ² value is **greater than** the critical value, we **reject the null hypothesis**.

B. **Then**, in a complete sentence, make a general statement of what the results tell you about the particular situation you are studying using the vocabulary of the situation.

<u>In our Example:</u>
"*Since 1.78 is less than the critical value, 5.99, we fail to reject H_0 which indicates independence between the gender and favorite subject .* "

<u>and the general statement could be something like:</u>
"*Based on these results, there is no apparent association between favorite subject at my school and one's gender.*"

(do not use the word "***correlation***" or "***affect***" when using this test)

Note:

Be cautious of the wording of your conclusions during those times when you reject the Null Hypothesis , **H₀**.   The Chi Square Test of Independence does not prove anything. It supplies evidence to support if two categorical variables are associated with each other or not. If our example had had the opposite result, the general statement might read as:

> "*Since 9.72 is greater than the critical value, 5.99, I reject H_0 which indicates that favorite subject and gender are <u>not</u> independent.*"
>
> *These results supply evidence to support the possibility that favorite subject and gender of the students I studied are associated.*

B. The Chi Squared Test of Indep.
(using Probability)

You may be asked to perform the test using **probability** instead of the Chi-Square statistic. In that case, you will follow steps 1 through 4. But continue with the following steps instead:

---

5. State the *rejection inequality*

"If the p-value is **less than** the significance level " (5% or 0.05 for many cases), then I will reject H_0.

---

6. **Calculate the *p-value*.**
This step must be done on a calculator. Follow the same calculator steps as if you are calculating the Chi-square statistic.

"*From my GDC, p = 0.41* "

---

7. **Make two summary statements**

A. First make a statement, in a complete sentence, using the following guidelines:

If the *p value* is **greater than or equal to** the signifance level (5% in most problems), we **fail to reject the null hypothesis**.

or if If the *p value* is **LESS THAN** than the signifance level, we **reject the null hypothesis**.

B. **Then**, in a complete sentence, make a general statement of what the results tell you about the particular situation you are studying using the vocabulary of the situation.

Example:
"*Since my p-value of 0.41 is greater than the 0.05 significance level, I fail to reject H_0 which indicates independence between gender and favorite subject.* "

and the general statement could be something like:
 "*Based on these results, there is no apparent association between favorite subject at my school and one's gender and favorite subject.*"

(do not use the word *correlation* or *affect* when using this test)