

χ^2 TEST FOR INDEPENDENCE

Data is usually given for two attributes (variables) and displayed in a contingency (two-way) table. Examples of **attributes** are: gender, hair colour, social background, eyesight etc. These are qualitative variables but the test can be used for quantitative variables such as length of seashells if the lengths, measured in millimeters, can be subdivided into “small”, “medium” and large.

Let’s take the dreaded drink and with it, a theory that rich kids tend to drink more. To investigate this theory we need a null hypothesis **Ho**, written in words. For our null hypothesis we state that there is **no** link between the two attributes.

Ho: Child alcohol consumption is independent of parent’s social background.

We split the parent’s social standing into three groups; the alcohol consumption into three groups and put the results of a survey in a 3x3 table. 3 rows horizontally and 3 columns vertically.

OBSERVED VALUES		CHILDREN			TOTALS
		FREQUENTLY	OCCAISONALLY	NONE	
PARENT	LOWER CLASS	10	21	9	40
	MIDDLE CLASS	24	90	24	138
	UPPER CLASS	13	19	7	39
TOTALS		47	130	40	217

The expected values under the null hypothesis now have to be calculated: If the null hypothesis were true and there was no association between the two attributes, we would expect equal proportions of frequent users across the social groups. The overall frequent user proportion can be found to be $\frac{47}{217}$

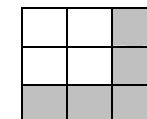
from the marginal totals. The values we would **expect** in column one would be:

Looking at the calculations in the cells more carefully, we note that we are using

$\frac{\text{row} \cdot \text{total} \times \text{column} \cdot \text{total}}{\text{grand} \cdot \text{total}}$ **and we will use this formula to calculate the expected values in the other cells.**

In fact we only need to calculate the values in four cells because the others can be worked out by subtraction. The marginal totals have to be the same as before.

FREQUENTLY
$\frac{47}{217} \times 40 = 8.7$
$\frac{47}{217} \times 138 = 29.9$
$\frac{47}{217} \times 39 = 8.4$



PARENT	EXPECTED VALUES	CHILDREN			TOTALS
		FREQUENTLY	OCCAISONALLY	NONE	
	LOWER CLASS	8.7	24.0	7.3	40
	MIDDLE CLASS	29.9	82.7	25.4	138
	UPPER CLASS	8.4	23.3	7.3	39
	TOTALS	47	130	40	217

We need a **statistic** to measure the degree of discrepancy between the observed and expected values.

If we look at the differences (observed – expected) and sum these, the answer would be zero.

If we **square** these before adding, we would not get zero and to finally, express these as a proportion of the expected values before summing gives the

measure of discrepancy. This is the calculated χ^2 statistic. χ^2 (calc) = $\sum \frac{(O-E)^2}{E} = \frac{1.3^2}{8.7} + \frac{3^2}{24} + \frac{1.7^2}{7.3} + \dots + \frac{0.3^2}{7.3} = 6.10$

We need to decide if this value is **significant** if we are to make a statement about alcohol consumption and its link with social background based on statistical evidence.

A chi-squared table of values will help us find the critical value.

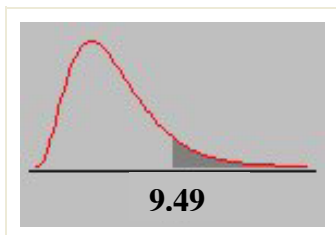
If the **expected** values are largely different from the **observed** values, this will be reflected in a large calculated value and if it is larger than the critical value found from tables, we will conclude that the null hypothesis is false and there **is** a link between the two attributes.

Now to find the critical value look to the tables. The values are entered and we look at the 5% value but must establish the row to look along.

The rows give the degrees of freedom: the degree of play we have when calculating the expected values.

We have seen that only four cells need to be calculated and the rest worked out by subtraction so we use 4 degrees of freedom for this test.

$$\chi^2_{5\% (4)} = 9.49$$



If our calculated value lies in the shaded region which it has a 5% chance of doing, then we reject the null hypothesis of independence.

Since χ^2 (calc) < 9.49, we have no grounds to reject Ho and conclude that **Child alcohol consumption is independent of parent's social background.**

Right tail areas for the *Chi-square* Distribution

df	P = 0.05	P = 0.01	P = 0.001
1	3.84	6.64	10.83
2	5.99	9.21	13.82
3	7.82	11.35	16.27
4	9.49	13.28	18.47
5	11.07	15.09	20.52