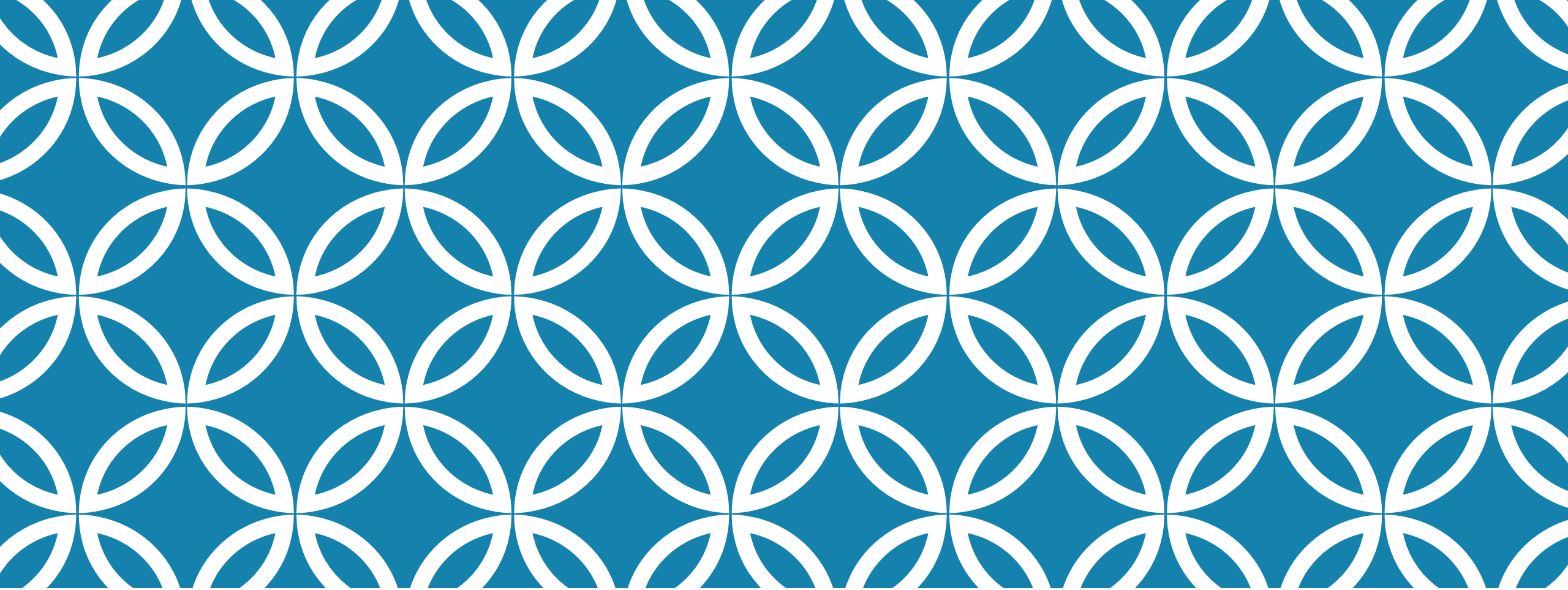

CHI-SQUARE TEST OF INDEPENDENCE

INSTRUCTOR
SANTOSH CHHATKULI

FUNCTION OF THE TEST

This test is used to determine whether two nominal scale categorical variables are independent or not.

For this test, the data must be presented in R x C contingency table, where R is the row and C is the column.

The R X C contingency table is given by,

Row Variable	Column Variable				Total
	1	2	…	C	
1	O_{11}	O_{12}	…	O_{1C}	R_1
2	O_{21}	O_{22}	…	O_{2C}	R_2
…	…	…	…	…	…
R	O_{R1}	O_{R2}	…	O_{RC}	R_R
Total	C_1	C_2	…	C_C	n

Where,

O_{ij} = observed cell frequency of ith row and jth column. (i = 1, 2, …, R ; j = 1, 2, …, C)

n = The total size of the sample which is fixed in advance in independence test. The row and column totals are determined by chance.

TEST ASSUMPTIONS

The Chi-Square Test of Independence is a non-parametric test used to determine if there is a significant association between two categorical variables.

1. Both variables should be measured at a nominal or ordinal level (e.g., gender, educational level, etc.).
2. The data should be collected through random sampling to ensure that the sample represents the population.
3. The sample size should be large enough to ensure that the expected frequencies in each cell of the contingency table are reasonably large. A common guideline is that at least 80% of the cells should have expected frequencies of 5 or more, and no cell should have an expected frequency of less than 1.
4. The sample size must be large enough to support the Chi-Square approximation. Small sample sizes can lead to unreliable results.

HYPOTHESIS TO TEST

H_0: row categorical variable is independent of column categorical variable

H_1: row categorical variable is not independent of column categorical variable

If null hypothesis is true, then the relative frequencies are same for each row categories or relative frequencies are same for each column categories.

TEST STATISTICS

The test statistic for the chi-square test of independence is,

$$\chi^2 = \sum_{i=1}^{R} \sum_{j=1}^{C} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Where,

Oij = observed cell frequency for the cell in ith row and jth column

Eij = expected cell frequency for the cell in ith row and jth column

$$= \frac{R_i \, x \, C_j}{n} = \frac{ith\ row\ total\ x\ jth\ column\ total}{Sample\ size}$$

The test statistics has Chi-square distribution with (R-1)(C-1) degrees of freedom.

Yate's corrected formula is,

$$\chi_C^2 = \sum_{i=1}^{R} \sum_{j=}^{C} \frac{\{|O_{ij} - E_{ij}| - 0.5\}^2}{E_{ij}}$$

2 x 2 contingency table

It's a special case of R x C table when R = 2 and C = 2. Its smallest possible dimension of R x C table.

Row Variable	**Column Variable**		**Total**
	Category 1	Category 2	
Category 1	O11	O12	R1
Category 2	O21	O22	R2
Total	C1	C2	n

In 2 x 2 we can conveniently calculate chi-square statistic using following formula:

$$\chi^2 = \frac{n\{O_{11} \cdot O_{22} - O_{12} \cdot O_{21}\}^2}{R_1 \cdot R_2 \cdot C_1 \cdot C_2}$$

The Yate's corrected formula is given by,

$$\chi_c^2 = \frac{n\left\{|O_{11} \cdot O_{22} - O_{12} \cdot O_{21}| - \frac{n}{2}\right\}^2}{R_1 \cdot R_2 \cdot C_1 \cdot C_2}$$

YATE'S CORRECTION

Yate's formula adjusts for overestimation of the Chi-square statistic when the sample size is small, or when some expected cell frequencies are close to or below 5. In small samples, the uncorrected Chi-square test may lead to **false positives** (Type I errors), rejecting the null hypothesis when it should not be rejected.

When to use yate's corrected formula

1. Yates' correction is used specifically for 2×2 contingency tables (tables with two rows and two columns).
2. Especially used when one or more expected frequencies are **close to 5**
3. Some statisticians recommend using the correction when the total sample size $n < 40$

When not to use Yate's corrected formula

1. IT is not used in Contingency tables larger than 2x2: The Yates correction was originally designed for 2x2 tables and its application to larger tables can be problematic. It may lead to overly conservative results, especially when the sample size is large
2. Large sample sizes ($n > 40$): With large sample sizes, the chi-square distribution itself provides a good approximation, and the Yates correction may not be necessary. In fact, it can sometimes lead to a loss of power, making it harder to detect true differences.
3. When all expected cell counts are 5 or greater: The Yates correction is primarily used to address the issue of small expected cell counts. If all expected cell counts are 5 or greater, the chi-square test is generally considered valid without the correction.

DECISION RULE

The chi-square test of independence is a right sided test, hence calculated chi-square must exceed critical chi-square in order to reject the null hypothesis H0. Hence, the decision rule is that:

If calculated $\chi^2 \geq \chi^2_\alpha\{(R-1)(C-1)\}$ we reject the null hypothesis H0 that row variable is independent of column variable.