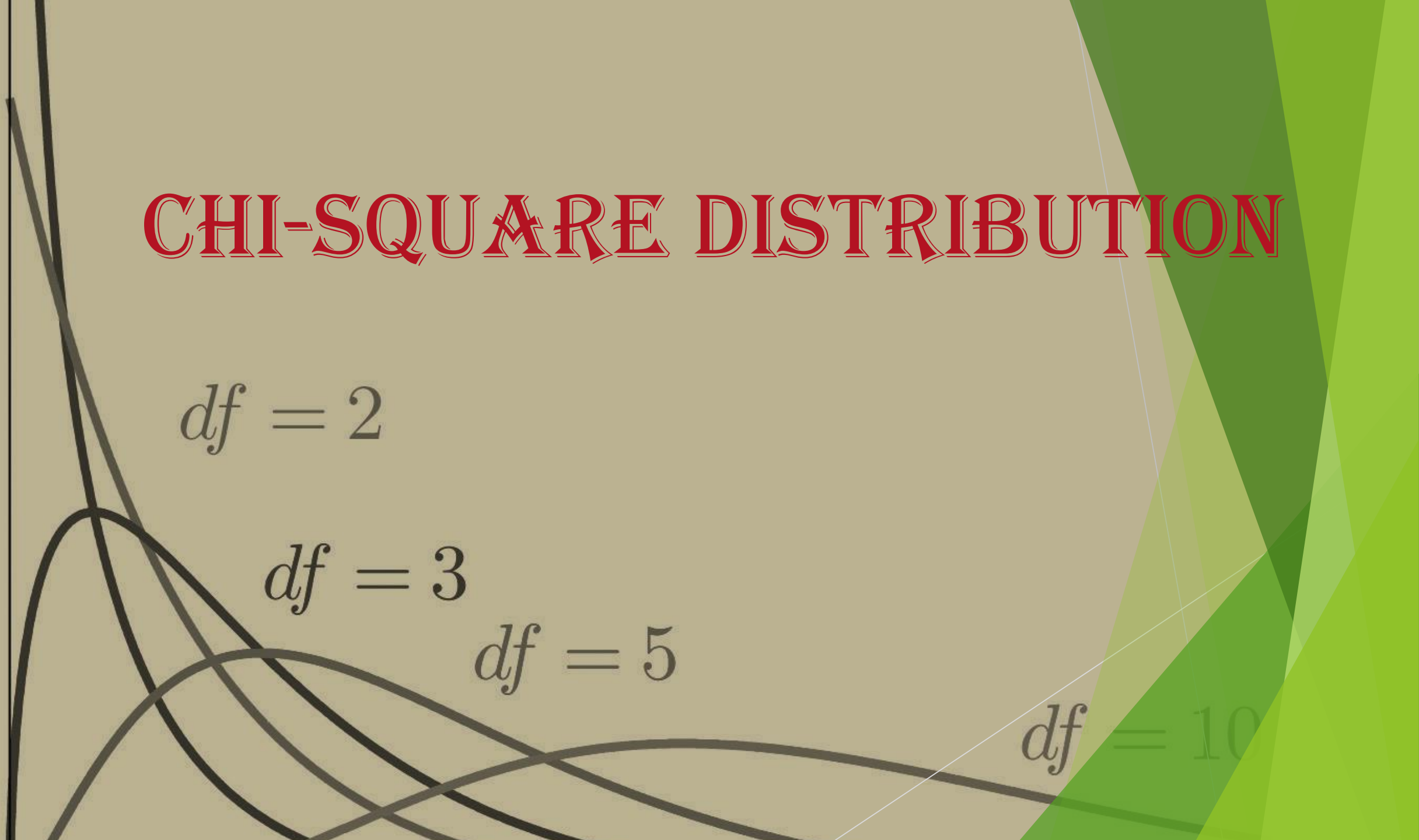
CHI-SQUARE DISTRIBUTION



Properties of the Chi-square Distribution

- Chi-square (χ^2) distribution is not symmetric
- Ranges from 0 to infinity – no negative values
- Skewed to the right
- Total area under the curve = 1
- Family of distributions – different chi-square distribution for each value of the degrees of freedom
- Degrees of freedom = (#rows – 1) * (# columns – 1)

In general if $X_i, (i = 1, 2, 3 \ldots\ldots, n)$ are n independent normal variables with mean μ_i and variances ${\sigma_i}^2$, $(i = 1, 2, 3 \ldots\ldots, n)$ then

$$\chi^2 = Z^2 = \left(\frac{X_i - \mu_i}{\sigma_i}\right)^2$$

If o_i and E_i be the set of observed and expected frequencies then

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

$$= \sum \left(\frac{O_i}{E_i}\right)^2 - N, \quad \text{where} \quad N = \sum O_i$$

NOTE: $\sum O_i = \sum E_i$

Conditions for validity of χ^2-test

- A chi-square test is an approximate test for large values of n.
- For validity of chi-square test of ‘ goodness of fit’ between theoretical and experiment, the following are assumption
- The sample observation should be independent
- Constraints on the cell frequencies, if any, should be linear.eg. $\sum O_i = \sum E_i$
- The total frequency should be reasonably large, say greater than 50.

- No theoretical cell frequency less than 5 (the chi-square distribution is essentially a continuous distribution but it can not maintain it character if cell frequency is less than 5) If any theoretical frequency is then for the application of chi-square test, it is pooled with the preceding or succeeding frequency so that polled frequency is more than5 and finally adjust for the d.f.lost in pooling.
- It may be noted that chi-square test is depend on the set of observed and expected frequency and on d.f.
- It does not make any assumption regarding the parent population from which the observation are taken.

- It does not involve any population parameter, it is termed as a statistic and the test is known as NON-PARAMETRIC TEST or DISTRIBUTION FREE TEST.
- The chi-square test for goodness of fit and the chi-square test for independence.
- The most obvious difference between the chi-square tests and the other hypothesis tests we have considered (t and ANOVA) is the nature of the data.
- For chi-square, the data are frequencies rather than numerical scores.

The Chi-Square Test for Goodness-of-Fit

- The **chi-square test for goodness-of-fit** uses frequency data from a sample to test hypotheses about the shape or proportions of a population.
- Each individual in the sample is classified into one category on the scale of measurement.
- The data, called **observed frequencies**, simply count how many individuals from the sample are in each category.

The Chi-Square Test for Independence

- The second chi-square test, the **chi-square test for independence**, can be used and interpreted in two different ways:

1. Testing hypotheses about the relationship between two variables in a population, or
2. Testing hypotheses about differences between proportions for two or more populations.

- Although the two versions of the test for independence appear to be different, they are equivalent and they are interchangeable.

The first version of the test emphasizes the relationship between chi-square and a correlation, because both procedures examine the relationship between two variables.

- The second version of the test emphasizes the relationship between chi-square and an independent-measures t test (or ANOVA) because both tests use data from two (or more) samples to test hypotheses about the difference between two (or more) populations.

Both chi-square tests use the same statistic. The calculation of the chi-square statistic requires two steps:

1. The null hypothesis is used to construct an idealized sample distribution of **expected frequencies** that describes how the sample would look if the data were in perfect agreement with the null hypothesis.

2. A chi-square statistic is computed to measure the amount of discrepancy between the ideal sample (expected frequencies from H_0) and the actual sample data (the observed frequencies = f_o).

A large discrepancy results in a large value for chi-square and indicates that the data do not fit the null hypothesis and the hypothesis should be rejected.

- χ^2-criterion is based on observed frequencies O and expected frequencies E.
- Assuming that there is no association between the given attributes,
- We calculate frequencies in each cell
- These frequencies are called expected frequency of the cell
- We denote the observed frequencies of the (I, j) the cell by O_{ij} and corresponding expected frequencies by E_{ij}
- If table has r-rows and c-column there will be r*c cell in the table
- Such a table is called Contingency Table
- If $A_1, A_2, A_3, \ldots \ldots A_r$ $are\ total\ r-rows\ and\ B_1, B_2, B_3 \ldots\ldots B_c$
- Then $E_{ij} = \frac{A_i B_j}{N}$
- $\chi^2 = \frac{\sum_i^r \sum_j^c O^2{}_{ij}}{Eij} - N$
- d.f.= (r-1)(c-1)

DEGREE OF FREEDOM

- The degree of freedom means the number of values which can be chosen arbitrary under the given restrictions
- Eg: If we have to choose 5 number whose sum is 50, we cannot choose all the 5 numbers arbitrarily because of the restriction
- We can choose four number arbitrarily
- In general if there are n numbers to be chosen and k independent constraints then the d.f. is given n-k

YATE'S CORRELATION

- In 2*2 table the d.f. is 1
- If any one of the frequency is less than 5,
- We have to use pooling method
- But this will result in χ^2 with zero degree of freedom
- This is meaningless
- In this case in 1934 Yate's suggested to use

a	b
c	d

$$\chi^2 = \frac{N(|ad-bc| - \frac{N}{2})^2}{(a+b)(a+c)(b+d)(c+d)}$$

Else

$$\chi^2 = \frac{N(ad-bc)^2}{(a+b)(a+c)(b+d)(c+d)}$$