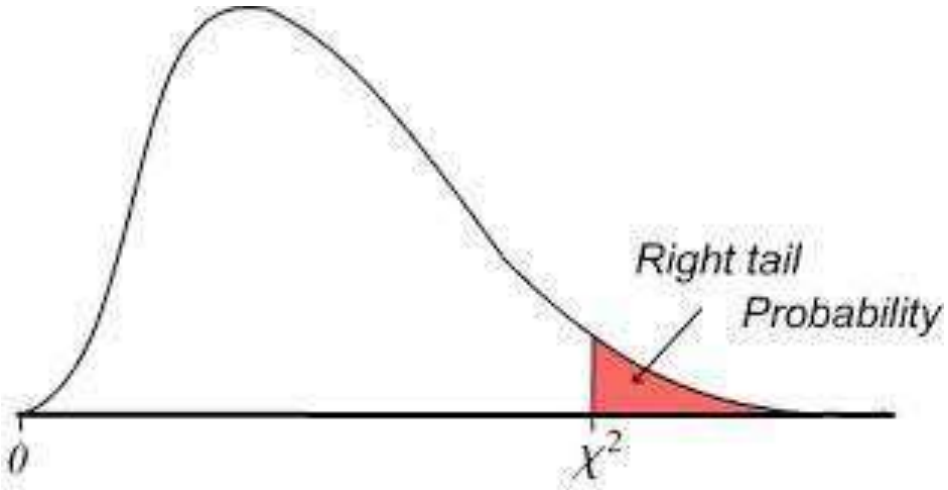
$$X^2 = \sum \frac{(o-e)^2}{e}$$

Chi-Square Test

By:
Dr. Girima Nagda

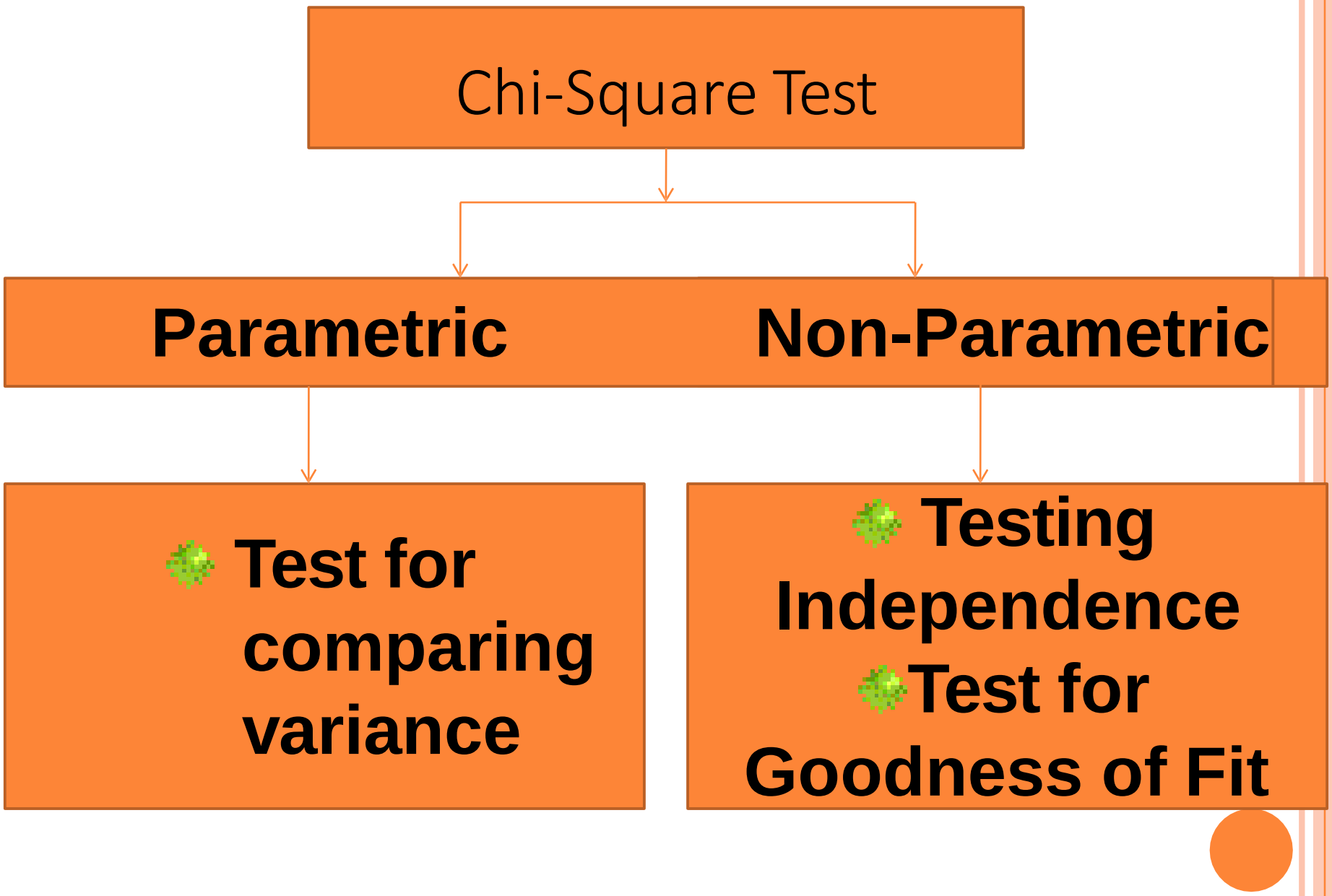


CONTENTS

- PARAMETRIC VS NON PARAMETRIC TEST
- INTRODUCTION
- CHARACTERISTICS OF THE TEST
- CHI SQUARE DISTRIBUTION
- APPLICATIONS OF CHI SQUARE TEST
- CALCULATION OF THE CHI SQUARE
- CONDITION FOR THE APPLICATION OF THE TEST
- LIMITATIONS OF THE TEST
- YATE’S CORRECTION FOR CONTINUITY

Parametric vs non parametric

- PARAMETRIC TEST: The test in which, the population constants like mean, std deviation, std error, correlation coefficient, proportion etc. and data tend to follow one assumed or established distribution such as normal, binomial, poisson etc.
- 2) NON PARAMETRIC TEST: the test in which no constant of a population is used. Data do not follow any specific distribution and no assumption are made in these tests. E.g. to classify good, better and best we just allocate arbitrary numbers or marks to each category.


Chi-Square Test
Parametric
Non-Parametric
Test for comparing variance
Testing Independence
Test for Goodness of Fit

Introduction

- Student’s T-test is applied for quantitative characters
- In biological experiments and field surveys, apart from quantitative data one has to deal with qualitative data for eg., in genetic studies, the variables like seed color, eye color etc., which do not have numerical values
- In such case chi square test is used

Chi-Square Test

- **Karl Pearson introduced a test to distinguish whether an observed set of frequencies differs from a specified frequency distribution**
- **The chi-square test uses frequency data to generate a statistic**

Karl Pearson

A chi-square test is a statistical test commonly used for testing independence and goodness of fit. Testing independence determines whether two or more observations across two populations are dependent on each other (that is, whether one variable helps to estimate the other). Testing for goodness of fit determines if an observed frequency distribution matches a theoretical frequency distribution.

Introduction

- The Chi-square test is one of the most commonly used non-parametric test, in which the sampling distribution of the test statistic is a ***chi-square distribution,*** when the null hypothesis is true.

- It was introduced by ***Karl Pearson*** as a test of association. The Greek Letter $\chi 2$ is used to denote this test.

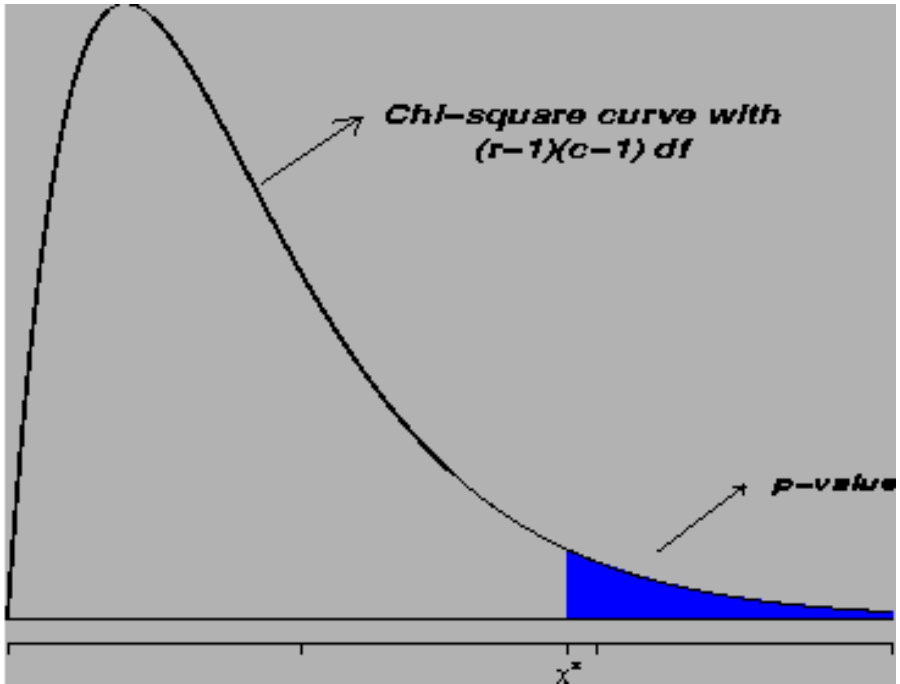
- It can be applied when there are few or no assumptions about the population parameter.

- It can be applied on categorical data or qualitative data using a contingency table.

- Used to evaluate ***unpaired/unrelated samples and proportions.***

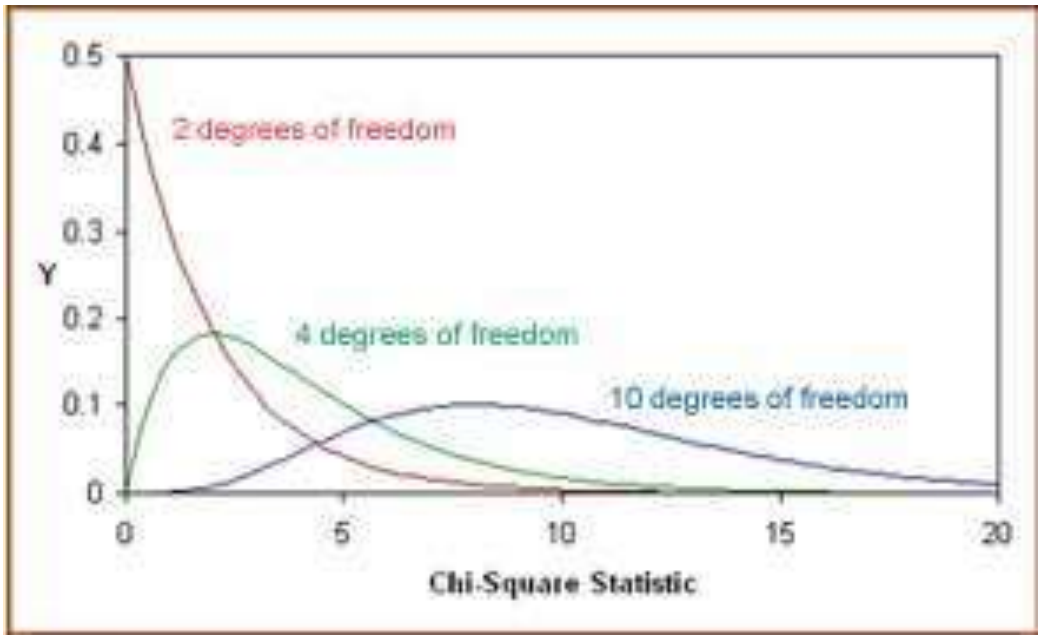
Chi-squared distribution

- The distribution of the chi-square statistic is called the chi-square distribution.
- The **chi-squared distribution** with *k* degrees of freedom is the distribution of a sum of the squares of *k* independent standard normal random variables. It is determined by the *degrees of freedom.*
- The simplest chi-squared distribution is the square of a standard normal distribution.
- The chi-squared distribution is used primarily in hypothesis testing.

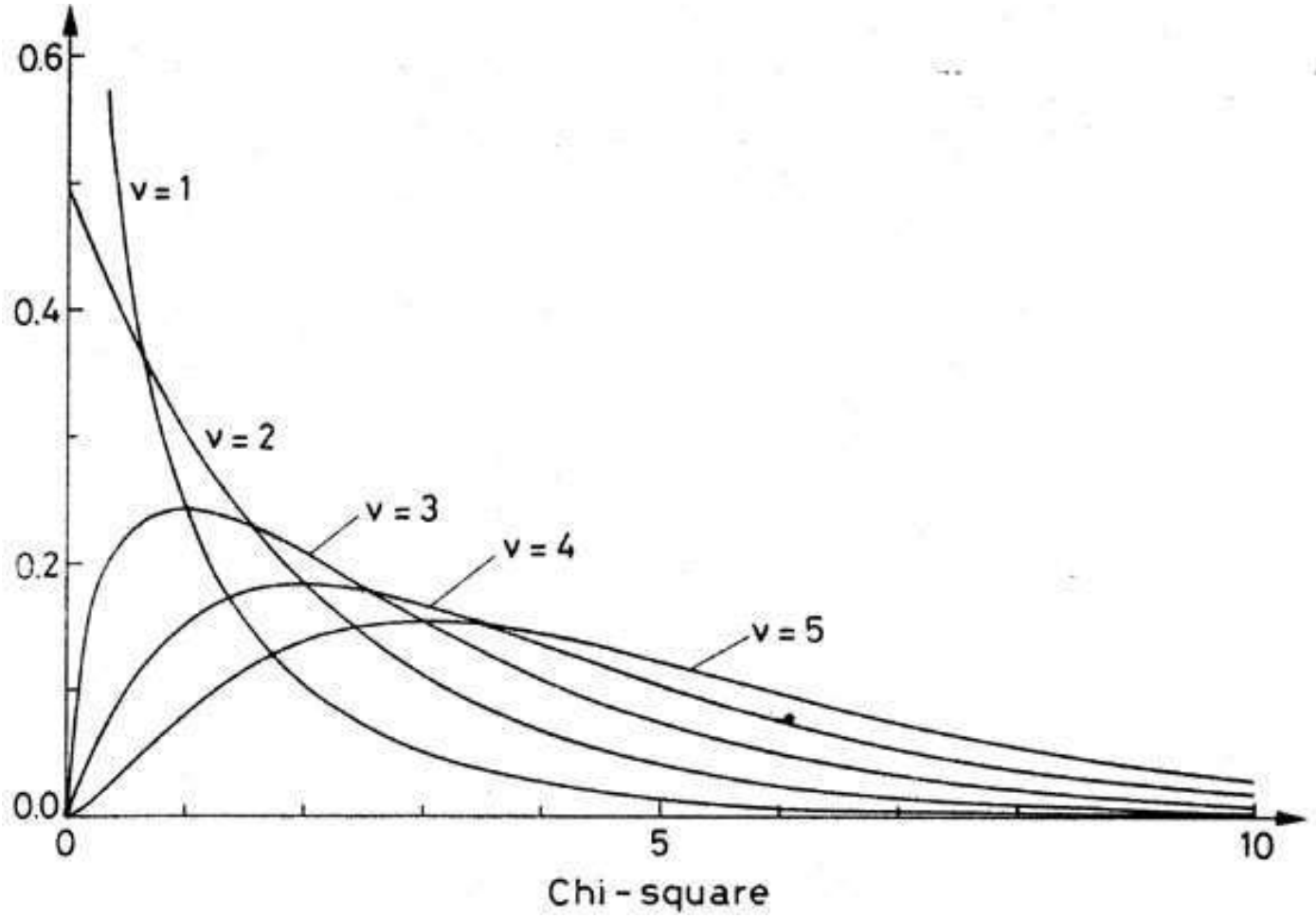


- The chi-square distribution has the following properties:

1. The mean of the distribution is equal to the number of degrees of freedom: $\mu = v$.
2. The variance is equal to two times the number of degrees of freedom: $\sigma^2 = 2 * v$



3. The χ^2 distribution is not symmetrical and all the values are positive. The distribution is described by degrees of freedom. For each degrees of freedom we have asymmetric curves.



4. **As the degrees of freedom increase, the chi-square curve approaches a normal distribution.**

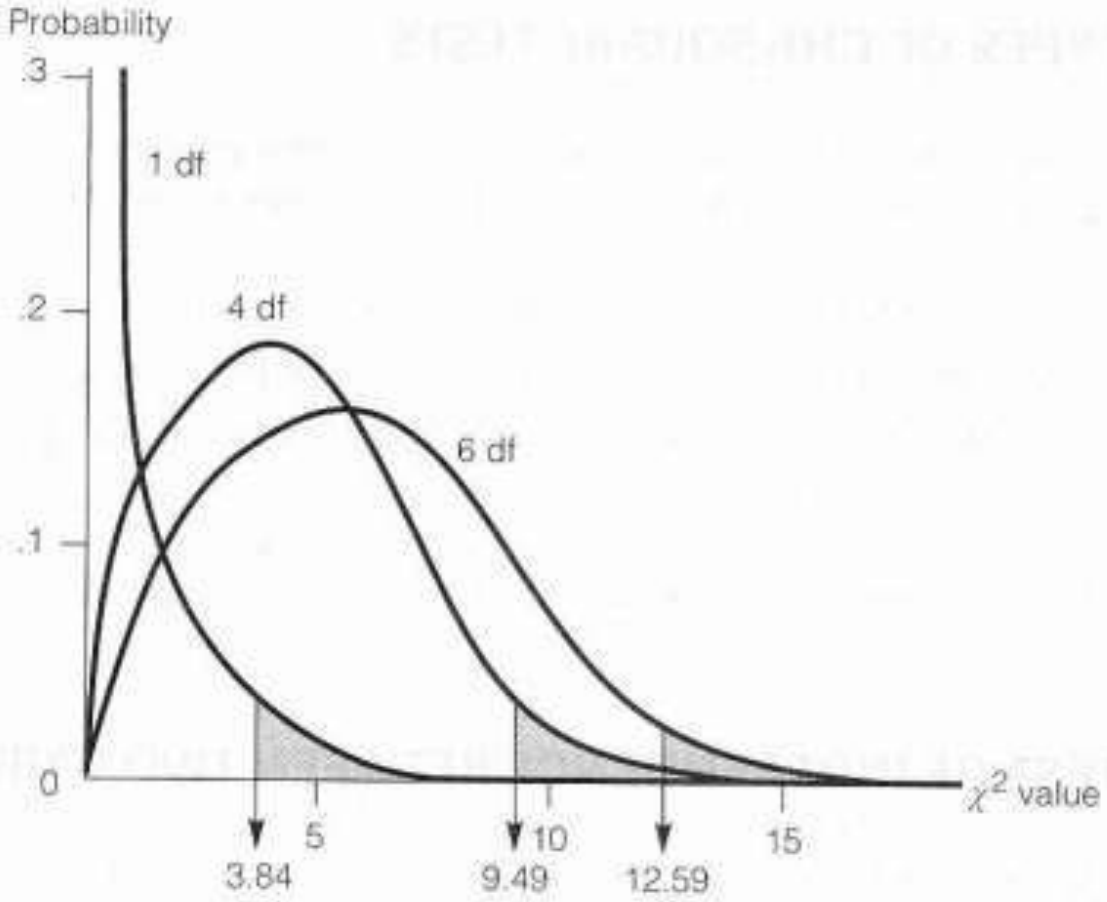
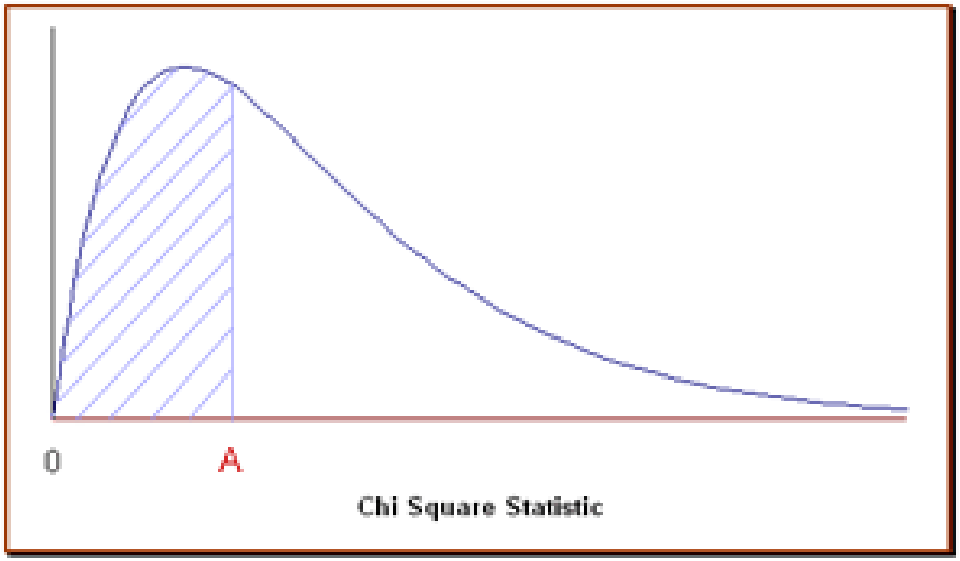


Figure 12.1 The Chi-Square Distribution for Varying Degrees of Freedom.

Cumulative Probability and the Chi- Square Distribution

- The chi-square distribution is constructed so that the total area under the curve is equal to 1. The area under the curve between 0 and a particular chi-square value is a ***cumulative probability associated with that chi-square value.***
- Ex: The shaded area represents a cumulative probability associated with a chi-square statistic equal to *A*; that is, it is the probability that the value of a chi-square statistic will fall between 0 and *A*.



Contingency table

- A **contingency table** is a type of table in a matrix format that displays the frequency distribution of the variables.
- They provide a basic picture of the interrelation between two variables and can help find interactions between them.

	Column 1	Column 2	Totals
Row 1	A	B	R1
Row 2	C	D	R2
Totals	C1	C2	N

- The chi-square statistic compares the observed count in each table cell to the count which would be expected ***under the assumption of no association between the row and column classifications.***

Degrees of freedom

- The number of independent pieces of information which are free to vary, that go into the estimate of a parameter is called the degrees of freedom.
- In general, the degrees of freedom of an estimate of a parameter is equal to ***the number of independent scores that go into the estimate minus the number of parameters used as intermediate steps in the estimation of the parameter itself*** (i.e. the sample variance has *N*-1 degrees of freedom, since it is computed from *N* random scores minus the only 1 parameter estimated as intermediate step, which is the sample mean).
- The number of degrees of freedom for 'n' observations is 'n-k' and is usually denoted by 'ν ', where 'k' is the number of independent linear constraints imposed upon them. It is the only parameter of the chi-square distribution.
- The degrees of freedom for a chi squared contingency table can be calculated as:

$$\nu = (\text{Number of rows} - 1) * (\text{Number of columns} - 1)$$

Chi Square formula

- The chi-squared test is used to determine whether there is a significant difference between the expected frequencies and the observed frequencies in one or more categories.
- The value of χ 2 is calculated as:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} + \frac{(O_3 - E_3)^2}{E_3} + \ldots\ldots + \frac{(O_n - E_n)^2}{E_n}$$

Where, O_1, O_2, O_3….On are the observed frequencies and E_1, E_2, E_3…E_n are the corresponding expected or theoretical frequencies.

The observed frequencies are the frequencies obtained from the observation, which are sample frequencies.
The expected frequencies are the calculated frequencies.

Alternate χ 2 Formula

	Disease		
Exposure	Yes	No	Total
Yes	a	b	a+b
No	c	d	c+d
Total	a+c	b+d	n

$$\chi_1^2 = \frac{n(ad-bc)^2}{(a+c)(b+d)(a+b)(c+d)}$$

The alternate χ 2 formula applies only to 2x2 tables

Characteristics of Chi-Square test

1. It is often regarded as a ***non-parametric test*** where no parameters
   - regarding the rigidity of populations are required, such as mean and SD.
2. It is based on ***frequencies***.
3. It encompasses the ***additive property*** of differences between observed and expected frequencies.
4. It tests the hypothesis about the ***independence of attributes***.
5. It is preferred in analyzing complex contingency tables.

Steps in solving problems related to Chi-Square test

STEP 1

- **Calculate the expected frequencies**

$$E = \frac{\text{Row Total} \times \text{Column Total}}{\text{Grand Total}}$$

STEP 2

- **Take the difference between the observed and expected frequencies and obtain the squares of these differences $(O-E)^2$**

STEP 3

- **Divide the values obtained in Step 2 by the respective expected frequency, E and add all the values to get the value according to the formula given by:**

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

Conditions for applying Chi-Square test

1. The data used in Chi-Square test must be ***quantitative*** and in the form of ***frequencies***, which must be ***absolute*** and not in relative terms.
2. The total number of observations collected for this test must be ***large*** (at least 10) and should be done on a ***random*** basis.
3. Each of the observations which make up the sample of this test must be ***independent*** of each other.
4. The expected frequency of any item or cell must not be ***less than 5***; the frequencies of adjacent items or cells should be polled together in order to make it more than 5.
5. This test is used only for ***drawing inferences*** through test of the hypothesis, so it ***cannot be used for estimation*** of parameter value.

Practical applications of Chi-Square test

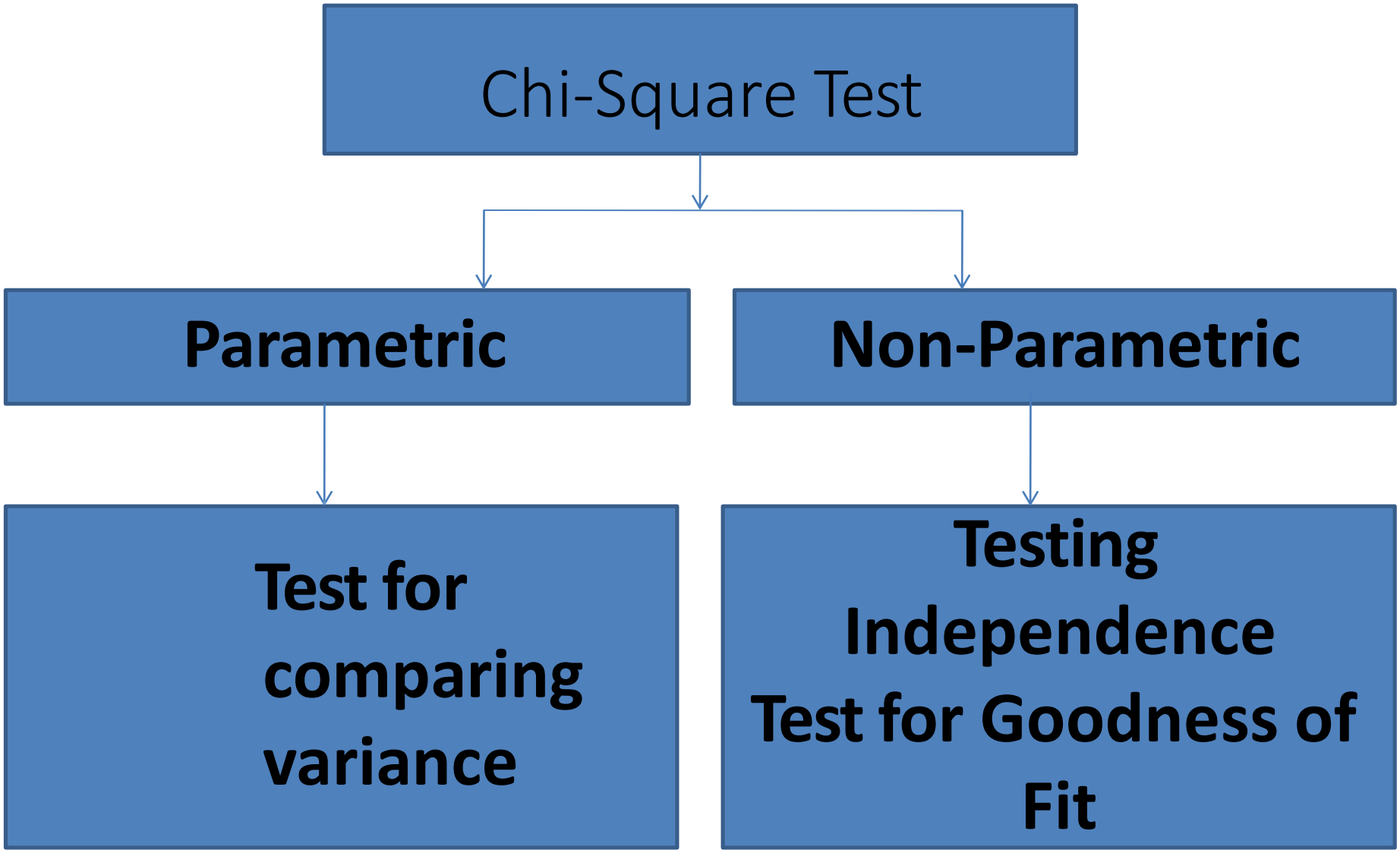
- The applications of Chi-Square test include testing:

1. The significance of ***sample & population variances*** **[σ^2 s & σ^2 p]**
2. The ***goodness of fit*** of a theoretical distribution: Testing for goodness of fit determines if an observed frequency distribution fits/matches a theoretical frequency distribution (**Binomial distribution, Poisson distribution or Normal distribution)**. These test results are helpful to know whether the samples are drawn from identical distributions or not. **When the calculated value of χ^2 is less than the table value at certain level of significance, the fit is considered to be good one and if the calculated value is greater than the table value, the fit is not considered to be good.**

Table/Critical values of χ^2

Degrees of Freedom	Probability										
	0.95	0.90	0.80	0.70	0.50	0.30	0.20	0.10	0.05	0.01	0.001
1	0.004	0.02	0.06	0.15	0.46	1.07	1.64	2.71	3.84	6.64	10.83
2	0.10	0.21	0.45	0.71	1.39	2.41	3.22	4.60	5.99	9.21	13.82
3	0.35	0.58	1.01	1.42	2.37	3.66	4.64	6.25	7.82	11.34	16.27
4	0.71	1.06	1.65	2.20	3.36	4.88	5.99	7.78	9.49	13.28	18.47
5	1.14	1.61	2.34	3.00	4.35	6.06	7.29	9.24	11.07	15.09	20.52
6	1.63	2.20	3.07	3.83	5.35	7.23	8.56	10.64	12.59	16.81	22.46
7	2.17	2.83	3.82	4.67	6.35	8.38	9.80	12.02	14.07	18.48	24.32
8	2.73	3.49	4.59	5.53	7.34	9.52	11.03	13.36	15.51	20.09	26.12
9	3.32	4.17	5.38	6.39	8.34	10.66	12.24	14.68	16.92	21.67	27.88
10	3.94	4.86	6.18	7.27	9.34	11.78	13.44	15.99	18.31	23.21	29.59
	Nonsignificant								Significant		

3. The ***independence*** in a contingency table:
   - Testing independence determines whether two or more observations across two populations are dependent on each other.
   - If the **calculated value is less than the table value** at certain level of significance for a given degree of freedom, then it is concluded that null hypothesis is true, which means that two attributes are independent and hence not associated.
   - If **calculated value is greater than the table value**, then the null hypothesis is rejected, which means that two attributes are dependent.
4. The chi-square test can be used to test the strength of the association between exposure and disease in a ***cohort study, an unmatched case-control study, or a cross-sectional study***.

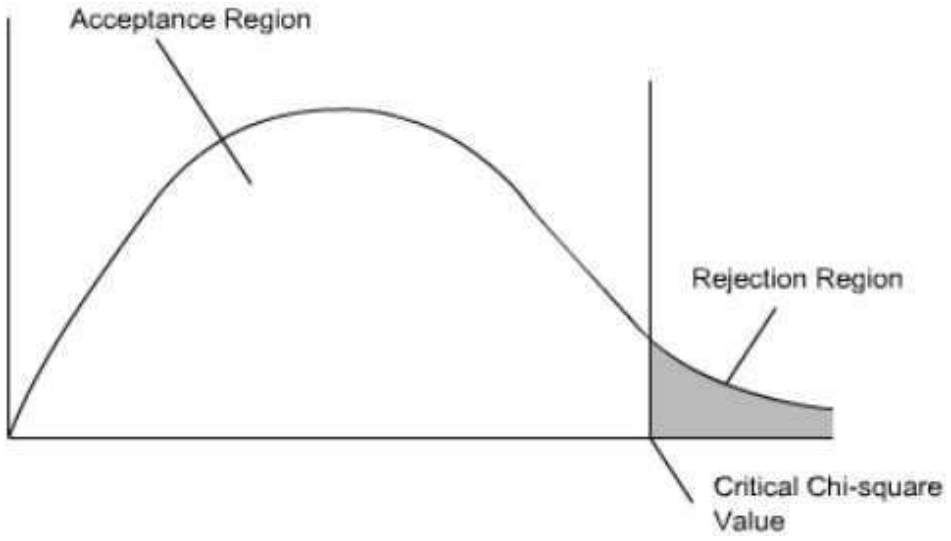

Chi-Square Test
Parametric
Non-Parametric
Test for comparing variance
Testing Independence
Test for Goodness of Fit

Interpretation of Chi-Square values

- The χ 2 statistic is calculated under the ***assumption of no association***. „
- **Large value of χ 2 statistic** ⇒ ***Small probability*** of occurring by chance alone ($p < 0.05$) ⇒ Conclude that ***association exists*** between disease and exposure. „**(Null hypothesis rejected)**
- **Small value of χ 2 statistic** ⇒ ***Large probability*** of occurring by chance alone ($p > 0.05$) ⇒ Conclude that ***no association exists*** between disease and exposure. **(Null hypothesis accepted)**

Interpretation of Chi-Square values

- The left hand side indicates the degrees of freedom. If the calculated value of χ2 falls in the acceptance region, the null hypothesis ‘Ho’ is accepted and vice-versa.



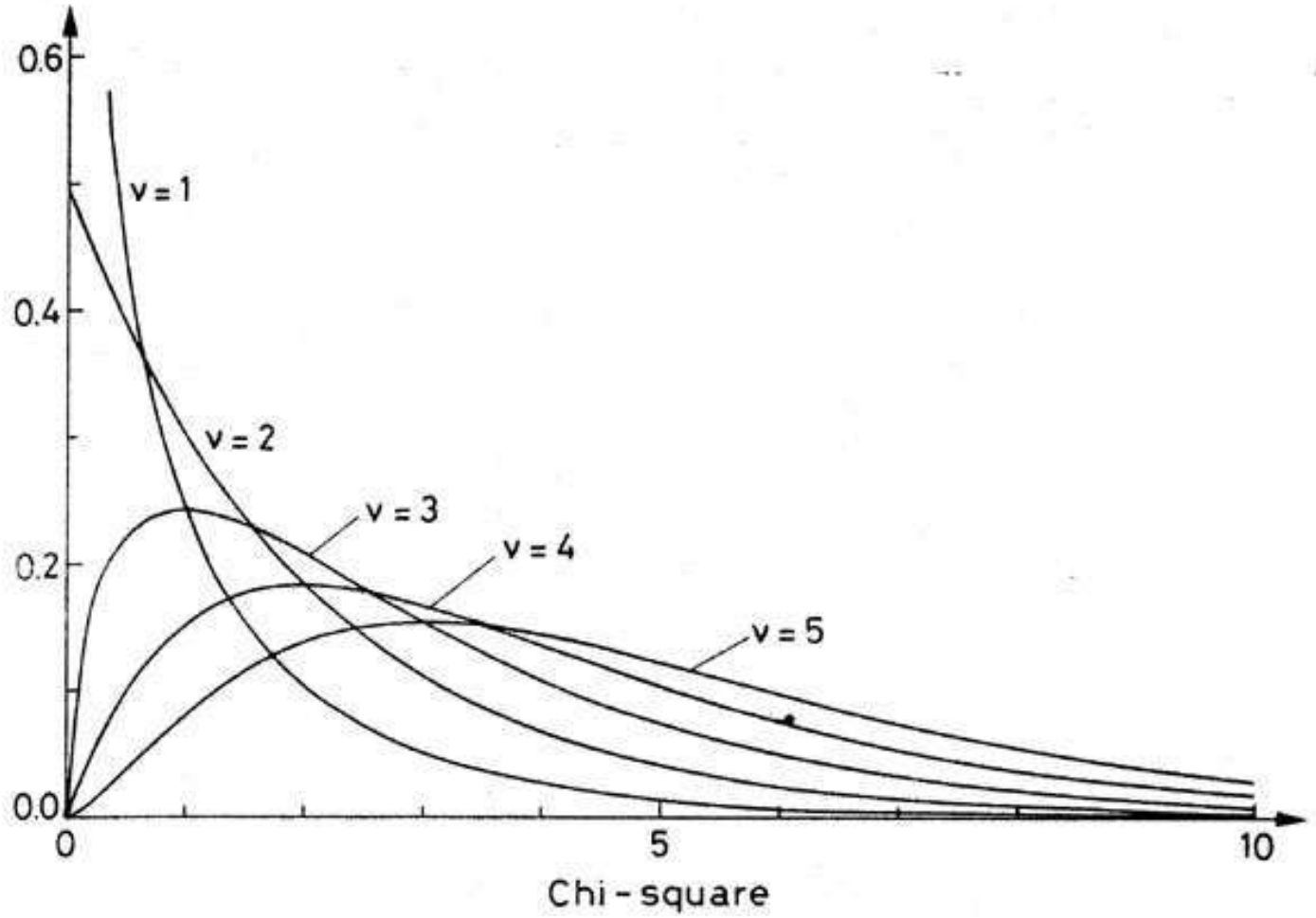
Limitations of the Chi-Square Test

1. The chi-square test does ***not give us much information about the strength of the relationship***. It only conveys the existence or nonexistence of the relationships between the variables investigated.
2. The chi-square test is ***sensitive to sample size***. This may make a weak relationship statistically significant if the sample is large enough. Therefore, chi-square should be used together with measures of association like ***lambda, Cramer's V or gamma*** to guide in deciding whether a relationship is important and worth pursuing.
3. The chi-square test is also ***sensitive to small expected frequencies***. It can be used only when not more than **20%** of the cells have an ***expected frequency of less than 5.***
4. Cannot be used when samples are ***related or matched***.

Conditions for the application of χ^2 test

- **Observations recorded and collected are collected on random basis.**
- **All items in the sample must be independent.**
- **No group should contain very few items, say less than 10. Some statisticians take this number as 5. But 10 is regarded as better by most statisticians.**
- **Total number of items should be large, say at least 50.**

The χ^2 distribution is not symmetrical and all the values are positive. For each degrees of freedom we have asymmetric curves.



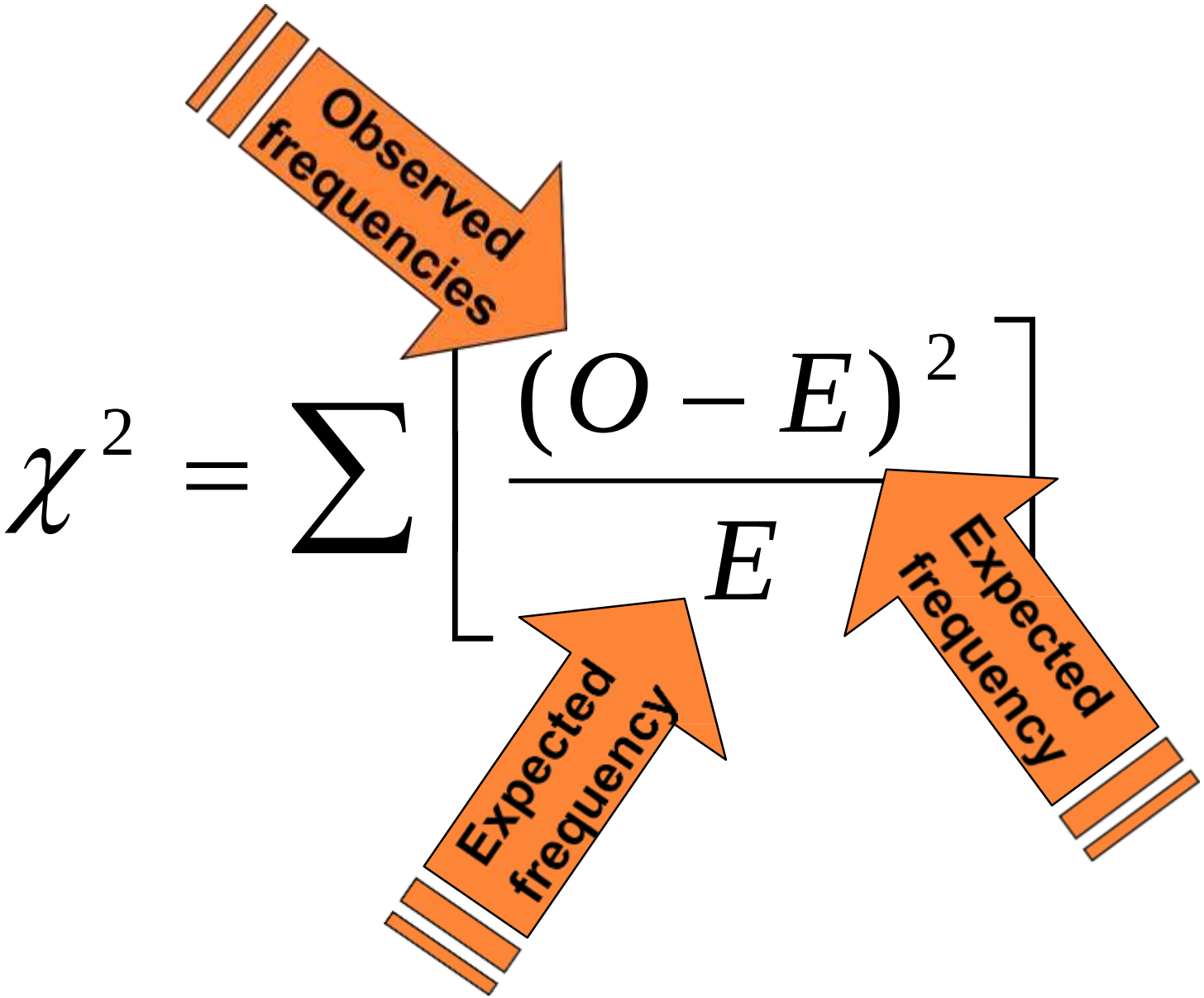
1. Test for comparing variance

$$\chi^2 = \frac{\sigma_s^2}{\sigma_p^2}(n-1)$$

Chi- Square Test as a Non-Parametric Test

- **Test of Goodness of Fit.**
- **Test of Independence.**

$$\chi^2 = \sum \left[\frac{(O - E)^2}{E} \right]$$

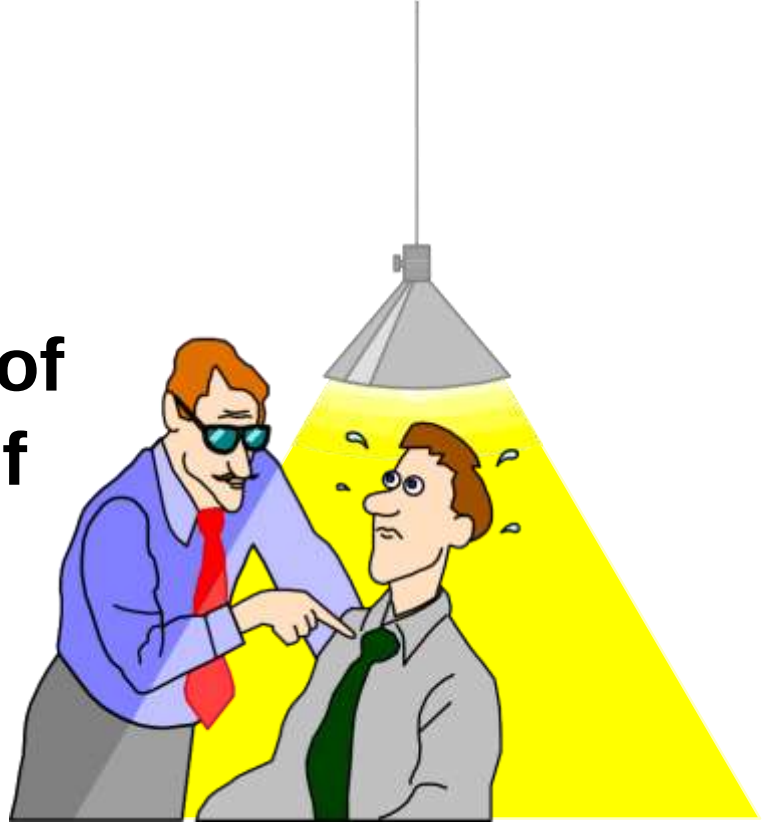

Observed frequencies
$\chi^2 = \sum \left[\frac{(O-E)^2}{E} \right]$
Expected frequency
Expected frequency

2. As a Test of Goodness of Fit

It enables us to see how well does the assumed theoretical distribution(such as Binomial distribution, Poisson distribution or Normal distribution) fit to the observed data. When the calculated value of χ^2 is less than the table value at certain level of significance, the fit is considered to be good one and if the calculated value is greater than the table value, the fit is not considered to be good.

Example

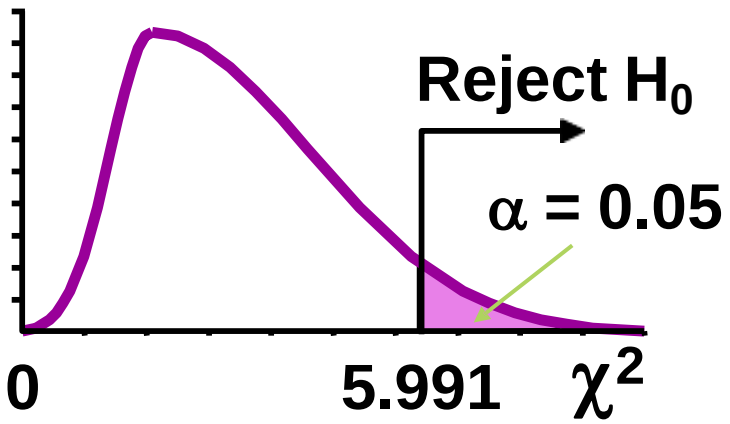
As personnel director, you want to test the perception of fairness of three methods of performance evaluation. Of 180 employees, 63 rated Method 1 as fair, 45 rated Method 2 as fair, 72 rated Method 3 as fair. At the 0.05 level of significance, is there a difference in perceptions?



Solution

Observed frequency	Expected frequency	(O-E)	$(O-E)^2$	$\frac{(O-E)^2}{E}$
63	60	3	9	0.15
45	60	-15	225	3.75
72	60	12	144	2.4
				6.3

- H_0: $p_1 = p_2 = p_3 = 1/3$
- H_1: At least 1 is different
- $\alpha = 0.05$
  - $n_1 = 63$ $n_2 = 45$ $n_3 = 72$
  - Critical Value(s):



Test Statistic:

$\chi^2 = 6.3$

Decision:

Reject H_0 at sign. level 0.05

Conclusion:

At least 1 proportion is different

3.As a Test of Independence

χ^2 test enables us to explain whether or not two attributes are associated. Testing independence determines whether two or more observations across two populations are dependent on each other (that is, whether one variable helps to estimate the other. If the calculated value is less than the table value at certain level of significance for a given degree of freedom, we conclude that null hypotheses stands which means that two attributes are independent or not associated. If calculated value is greater than the table value, we reject the null hypotheses.

Steps involved

- **Determine The Hypothesis:**

  H_o : The two variables are independent
  H_a : The two variables are associated

- **Calculate Expected frequency**

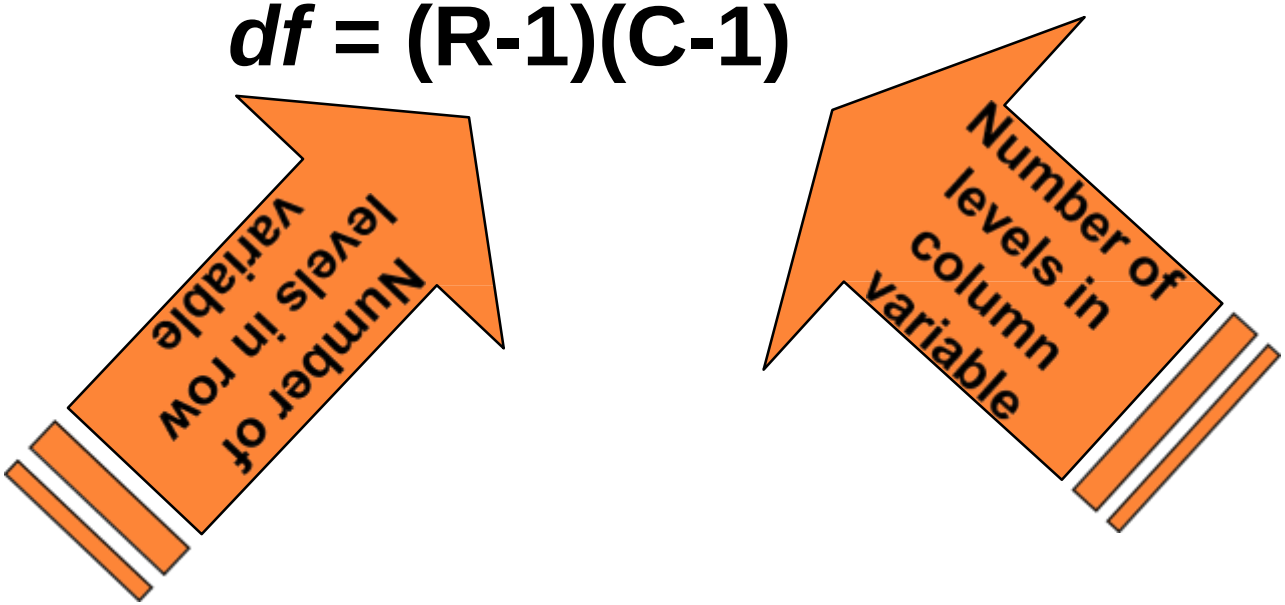
$$E = \frac{(\text{Row total})\,(\text{Column total})}{\text{Grant total}}$$

- **Calculate test statistic**

$$\chi^2 = \sum \left[\frac{(O - E)^2}{E} \right]$$

- **Determine Degrees of Freedom**

$$df = (R-1)(C-1)$$



- Compare computed test statistic against a tabled/critical value

The computed value of the Pearson chi-square statistic is compared with the critical value to determine if the computed value is *improbable*

The critical tabled values are based on sampling distributions of the Pearson chi-square statistic.

If calculated χ^2 is greater than χ^2 table value, reject H_o

Critical values of χ^2

Degrees of Freedom	Probability										
	0.95	0.90	0.80	0.70	0.50	0.30	0.20	0.10	0.05	0.01	0.001
1	0.004	0.02	0.06	0.15	0.46	1.07	1.64	2.71	3.84	6.64	10.83
2	0.10	0.21	0.45	0.71	1.39	2.41	3.22	4.60	5.99	9.21	13.82
3	0.35	0.58	1.01	1.42	2.37	3.66	4.64	6.25	7.82	11.34	16.27
4	0.71	1.06	1.65	2.20	3.36	4.88	5.99	7.78	9.49	13.28	18.47
5	1.14	1.61	2.34	3.00	4.35	6.06	7.29	9.24	11.07	15.09	20.52
6	1.63	2.20	3.07	3.83	5.35	7.23	8.56	10.64	12.59	16.81	22.46
7	2.17	2.83	3.82	4.67	6.35	8.38	9.80	12.02	14.07	18.48	24.32
8	2.73	3.49	4.59	5.53	7.34	9.52	11.03	13.36	15.51	20.09	26.12
9	3.32	4.17	5.38	6.39	8.34	10.66	12.24	14.68	16.92	21.67	27.88
10	3.94	4.86	6.18	7.27	9.34	11.78	13.44	15.99	18.31	23.21	29.59
	Nonsignificant								Significant		

EXAMPLE

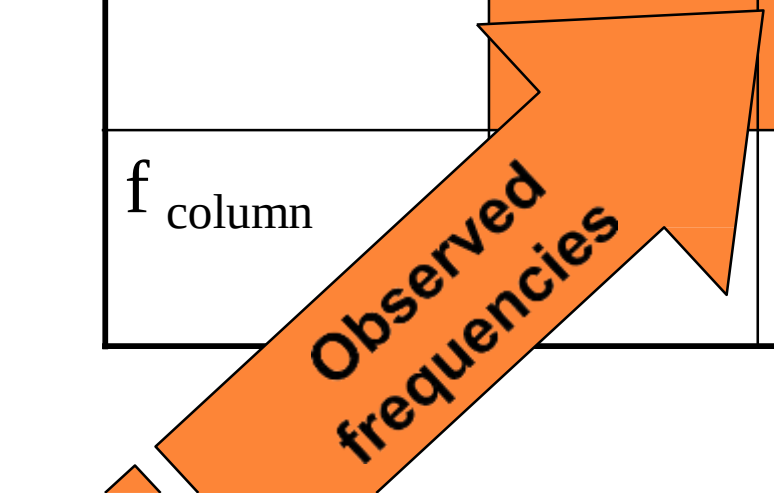
- **Suppose a researcher is interested in voting preferences on gun control issues.**
- **A questionnaire was developed and sent to a random sample of 90 voters.**
- **The researcher also collects information about the political party membership of the sample of 90 respondents.**

Bivariate Frequency Table or Contingency Table

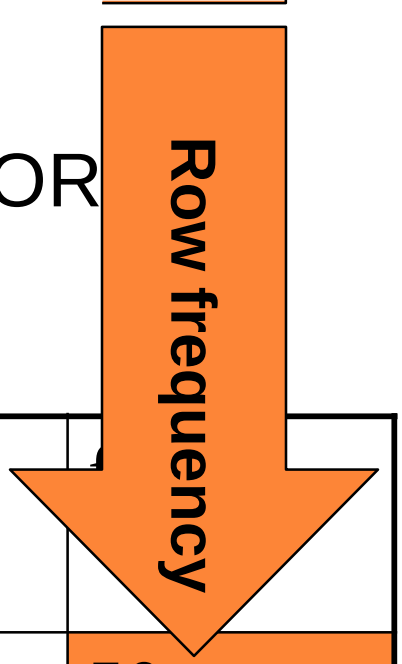
	Favor	Neutral	Oppose	f_{row}
Democrat	10	10	30	50
Republican	15	15	10	40
f_{column}	25	25	40	n = 90

Bivariate Frequency Table or Contingency Table

	Favor	Neutral	Oppose	f_{row}
Democrat	10	10	30	50
Republican	15	15	10	40
f_{column}		25	40	n = 90



Bivariate Frequency Table or Contingency Table



	Favor	Neutral	Oppose	
Democrat	10	10	30	50
Republican	15	15	10	40
f_{column}	25	25	40	n = 90

Bivariate Frequency Table or Contingency Table

	Favor	Neutral	Oppose	f_{row}
Democrat	10	10	30	50
Republican	15	15	10	40
f_{column}	25	25	40	n = 90

Column frequency

Determine The Hypothesis

- Ho : There is no difference between D & R in their opinion on gun control issue.

- Ha : There is an association between responses to the gun control survey and the party membership in the population.

Calculating Test Statistics

	Favor	Neutral	Oppose	f_{row}
Democrat	f_o =10 f_e =13.9	f_o =10 f_e =13.9	f_o =30 f_e=22.2	50
Republican	f_o =15 f_e =11.1	f_o =15 f_e =11.1	f_o =10 f_e =17.8	40
f_{column}	25	25	40	n = 90

Calculating Test Statistics

	Favor	Neutral	Oppose	f_{row}
Democrat	f_o =10 f_e =13.9	f_o =10 f_e =13.9	f_o =30 f_e=22.2	50
Republican	f_o =15 f_e =11.1	f_o =15 f_e = 40* 25/90	f_o =10 3	40
f_{column}	25	25	40	n = 90

Calculating Test Statistics

$$\chi^2 = \frac{(10-13.89)^2}{13.89} + \frac{(10-13.89)^2}{13.89} + \frac{(30-22.2)^2}{22.2} +$$

$$\frac{(15-11.11)^2}{11.11} + \frac{(15-11.11)^2}{11.11} + \frac{(10-17.8)^2}{17.8}$$

$$= 11.03$$

Determine Degrees of Freedom

- $df = (R-1)(C-1) =$
  $(2-1)(3-1) = 2$

Compare computed test statistic against table value

- **α = 0.05**
- **df = 2**
- **Critical tabled value = 5.991**
- **Test statistic, 11.03, exceeds critical value**
- **Null hypothesis is rejected**
- **Democrats & Republicans differ significantly in their opinions on gun control issues**

χ^2 Test of Independence Thinking Challenge

- You're a marketing research analyst. You ask a random sample of **286** consumers if they purchase Diet Pepsi or Diet Coke. At the **0.05** level of significance, is there evidence of a **relationship**?

	Diet Pepsi		
Diet Coke	**No**	**Yes**	**Total**
No	**84**	**32**	**116**
Yes	**48**	**122**	**170**
Total	**132**	**154**	**286**

χ^2 Test of Independence Solution*

✓ $E_{ij} \geq 5$ in all cells

	Diet Pepsi				
	No		Yes		
Diet Coke	Obs.	Exp.	Obs.	Exp.	Total
No	84	53.5	32	62.5	116
Yes	48	78.5	122	91.5	170
Total	132	132	154	154	286

116·132 / 286

154·132 / 286

170·132 / 286

170·154 / 286

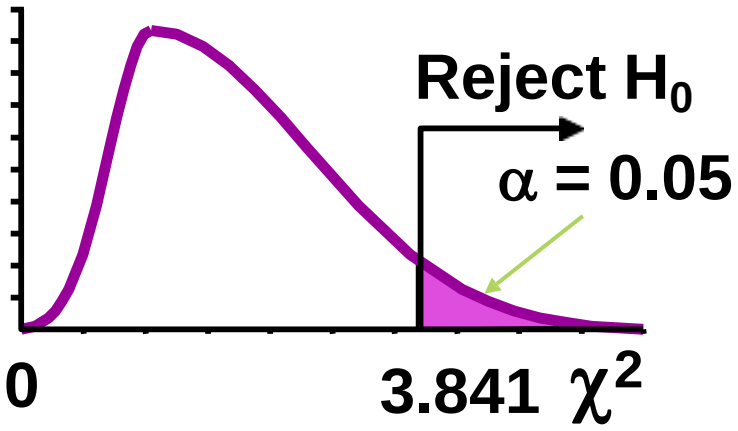
χ^2 Test of Independence Solution*

$$\chi^2 = \sum_{\text{all cells}} \frac{\left[\left[n_{ij} - E_{ij}\right]\right]^2}{E_{ij}}$$

$$= \frac{\left[\left[n_{11} - E_{11}\right]\right]^2}{E_{11}} + \frac{\left[\left[n_{12} - E_{12}\right]\right]^2}{E_{12}} + \cdots + \frac{\left[\left[n_{22} - E_{22}\right]\right]^2}{E_{22}}$$

$$= \frac{\left[\left[84 - 53.5\right]\right]^2}{53.5} + \frac{\left[\left[32 - 62.5\right]\right]^2}{62.5} + \cdots + \frac{\left[\left[122 - 91.5\right]\right]^2}{91.5} = 54.29$$

- H_0: No Relationship
- H_1: Relationship
- $\alpha =$ 0.05
- df = (2 - 1)(2 - 1) = 1
- Critical Value(s):



Test Statistic: $\chi^2 = 54.29$

Decision:

Reject at sign. level 0 .05

Conclusion:

There is evidence of a relationship

χ^2 Test of Independence Thinking Challenge 2

There is a statistically significant relationship between purchasing Diet Coke and Diet Pepsi. So what do you think the relationship is? Aren't they competitors?

	Diet Pepsi		
Diet Coke	**No**	**Yes**	**Total**
No	**84**	**32**	**116**
Yes	**48**	**122**	**170**
Total	**132**	**154**	**286**

YOU RE-ANALYZE THE DATA

High Income

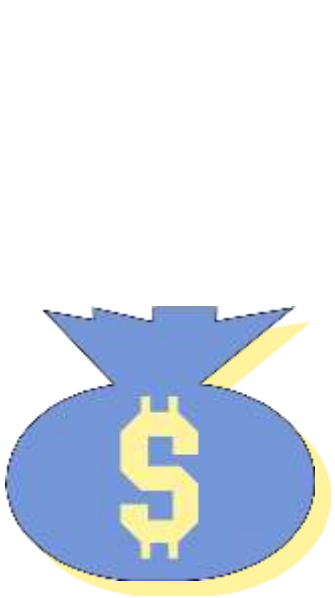
	Diet Pepsi		
Diet Coke	No	Yes	Total
No	4	30	34
Yes	40	2	42
Total	44	32	76

Low Income

	Diet Pepsi		
Diet Coke	No	Yes	Total
No	80	2	82
Yes	8	120	128
Total	88	122	210

Data mining example: no need for statistics here!

TRUE RELATIONSHIPS*



Control or intervening variable (true cause)

Underlying causal relation

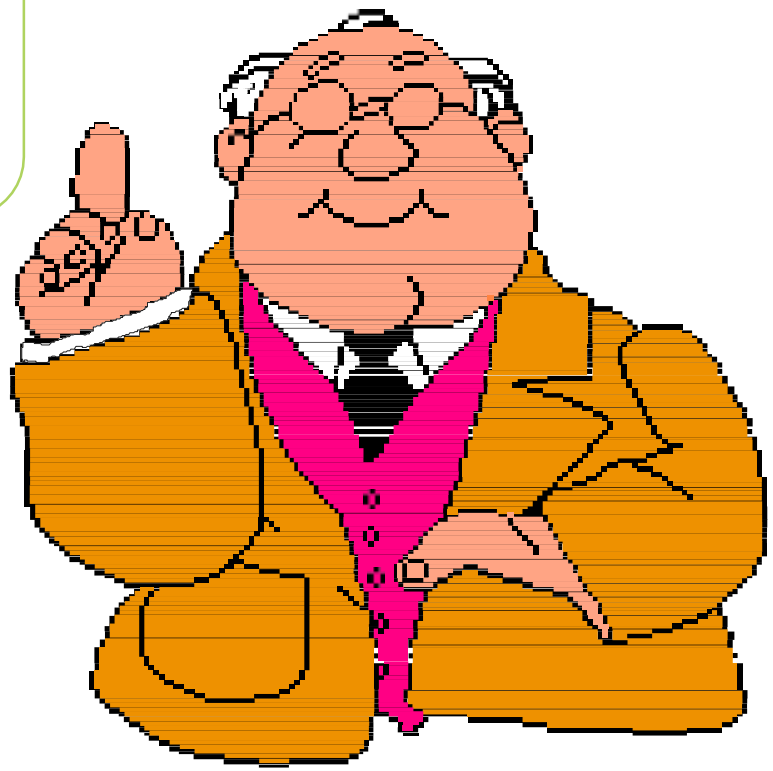
Diet Coke

Apparent relation

Diet Pepsi

Moral of the Story

Numbers don't think - People do!



Conclusion

1. **Explained χ^2 Test for Proportions**
2. **Explained χ^2 Test of Independence**


Thank you!