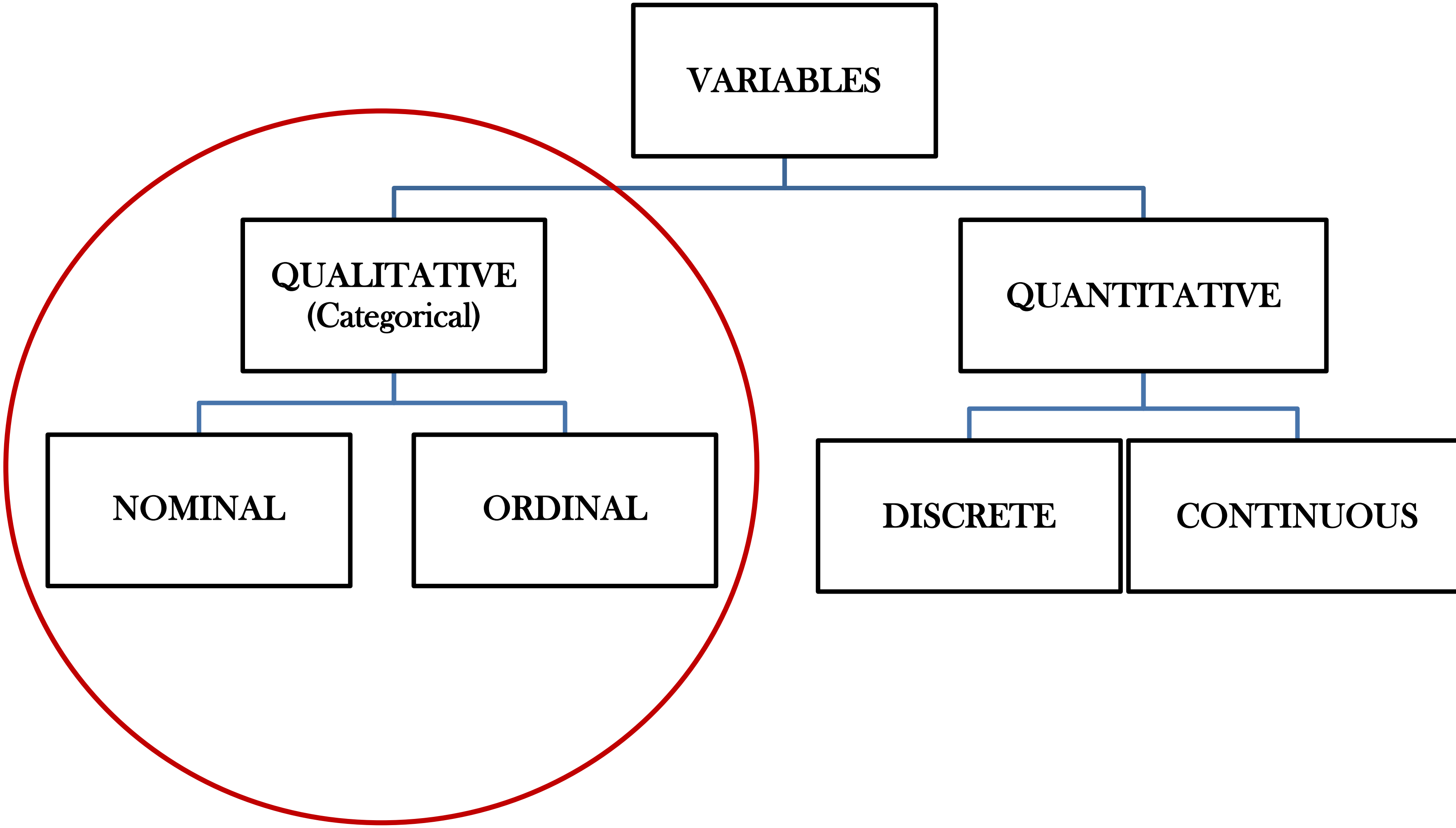
48th ICRO-SUN PG TEACHING PROGRAMME
26 & 27 OCTOBER, 2024
MAX SUPERSPECIALITY HOSPITAL, BATHINDA

CLINICAL TRIAL & CANCER STATISTICS

Categorical data analysis

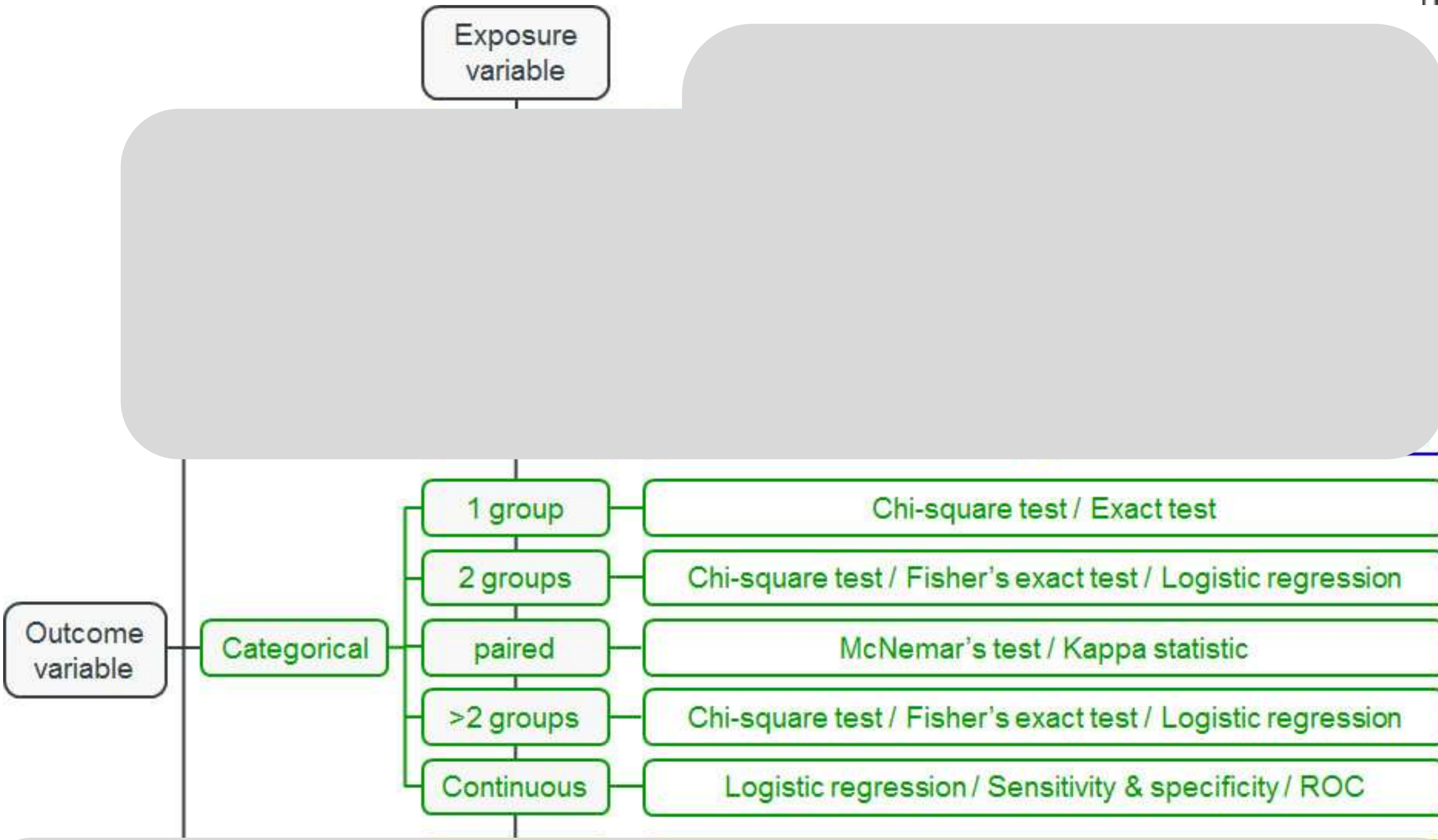
Dr. Madhur Verma
Associate Professor of Community and Family Medicine
AIIMS Bathinda
drmadhurverma@gmail.com
9466445513

Types of Variables



Categorical data

Nominal	
Locality	Rural/Urban
Gender	M, F
Diagnosis	Normal, Abnormal
Ordinal	
Age (years)	< 15, 15 -30, 30-45, 45 +
SES	Low, Medium, High
Improvement	Mild, Moderate, Fair
Tumor grade	Grade 1, 2, 3


Exposure variable
Outcome variable
Categorical
1 group
Chi-square test / Exact test
2 groups
Chi-square test / Fisher's exact test / Logistic regression
paired
McNemar's test / Kappa statistic
>2 groups
Chi-square test / Fisher's exact test / Logistic regression
Continuous
Logistic regression / Sensitivity & specificity / ROC

Initiating the categorical data analysis

To initiate a categorical data analysis, it is recommended to follow a systematic approach:

1. Understand Your Data
2. Summarise the Data
3. Examine Relationships Between Variables.
4. Advanced Analysis.
5. Test for Homogeneity or Independence
6. Interpretation

Categorical data analysis

1. **Understand Your Data**

- **Identify variables**

  Review your dataset and identify the categorical variables you want to analyse.

- **Define levels:** Ensure that each categorical variable has clearly defined levels or categories.

- **Check for missing data:** Handle missing data appropriately (e.g., using imputation techniques or omitting cases)

Categorical data analysis

2. Summarize the Data

- **Frequency tables:** Start by creating frequency tables for each categorical variable to understand the distribution of categories.

- **Bar plots or pie charts:** Visualize the frequency distribution using bar plots or pie charts to give a clear picture of how the categories are distributed.

- **Collapse tables if necessary:** reduce categories if necessary.

Categorical data analysis

3. **Examine Relationships Between Variables**

- **Contingency tables:** For two or more categorical variables, create contingency tables (cross-tabulation) to explore relationships.
- **Chi-square test:** Use the chi-square test to assess whether there's a significant association between two categorical variables.
- **Cramér's V:** If the chi-square test is significant, use Cramér's V to measure the strength of the association.

  Its value ranges from 0 to 1, where **0** indicates no association, while **1** indicates a perfect association (complete dependence).

Contingency Tables

Contingency Tables

- Cross-classifications of categorical variables in which
  - Rows (typically): categories of EXPLANATORY variables
  - Columns: categories of OUTCOME variables.

- Counts in the "cells" of the table give the numbers of individuals at the corresponding combination of levels of the two variables.

- Contingency tables enable us to compare one characteristic of the sample, e.g. Oral cancer, defined by another categorical variable, e.g. Smoking.

Contingency table (Bivariate)
Example 1: Gender and smoking status

S. No	Gender	Smoking
1	M	Y
2	F	Y
3	M	N
4	F	Y
5	M	N
6	F	N
7	M	Y
8	F	N
9	F	N
10	M	Y

Gender	Smoking Status		
	Yes n(row%)	No n(row%)	Total n(col%)
Male	3 (60)	2 (40)	5 (50)
Female	2 (40)	3 (60)	5 (50)
Total	5 (50)	5 (50)	10 (100)

Contingency table (Bivariate)
Example 2: Education status and Cervical cancer screening readiness

Education status	Cervical cancer screening readiness			
	Very Eager	Pretty	Not too	Row Total
Above Average	164	**233**	26	423
Average	293	473	117	883
Below Average	132	383	172	687
Col Total	589	1089	315	1993

Row and column totals are called **Marginal counts**

What can a contingency table do ?

Can summarize by percentages on response variable (happiness)

Education status	Cervical cancer screening readiness			
	Very Eager	Pretty	Not too	Row Total
Above Average	164	233	26	423
Average	293	473	117	883
Below Average	132	383	172	687
Col Total	589	1089	315	1993

Example: Percentage "**readiness**" is

39% for above average. education (164/423 = 0.39) 33% for average education (293/883 = 0.33) 19% for below average education (??)

What can a contingency table do ?

2. **Association between two categorical variables.**

For example, *you want to know*

- if there is any association between gender and smoking.
- Is there any association between hepatitis C infection and the population's HCC risk?
- To test whether lung cancer is associated with smoking or not.
- Obesity is associated with colon cancer.

Chi-Square Tests

Chi-Square Tests

Simplest & most widely used non-parametric test in statistical work.
Chi-Square Tests: These tests check whether the differences or patterns between two groups are real or just **random.**

- Chi-square is basically a measure of ***significance.***
- It is not a good measure of **the strength of the** association.
- It can help you decide if an association exists but not tell how strong it is.

Chi-Square Tests

Assumptions

1. The sample must be randomly drawn from the population.
2. Data must be reported in **raw frequencies (not percentages)**.
3. Categories of the variables must be mutually exclusive & exhaustive.
4. Expected frequencies cannot be too small; **expected frequency should be more than *5* in at least 80% of the** cells, and all individual expected counts should be $\geq$1.

Logic of the chi-square

The total number of observations in each column and the total number of observations in each row are considered to be **given or fixed.**

If we assume that columns and rows are independent, we can calculate - **expected frequencies.**

Disease			
Exposure	Yes	No	Total
Yes	37	13	50
No	17	53	70
Total	54	66	120

Logic of Chi square

If a relationship (or dependency) does occur

↓

The observed frequencies will vary from the expected frequencies

↓

The value of the chi-square statistic will be large.

Steps for Chi-square test

- Define Null and alternative hypothesis
- State alpha
- Calculate degree of freedom
- State decision rule
- Calculate test statistics
- State and Interpret results

Hypothesis Testing

- Tests a claim about a parameter using evidence (data in a sample) ⇒ gives causal relationships

Steps

1. Formulate a Hypothesis about the population
2. Random sample
3. Summarizing the information (descriptive statistic)
4. Does the information given by the sample support the hypothesis? Are we making any errors? (inferential stat.)

Decision rule: Convert the research question to null and alternative hypothesis

Null Hypothesis

H0 = No difference between observed and expected observations

H1 = difference is present between observed and expected observations

What is statistical significance?

- A statistical concept indicates that the result is very unlikely due to chance and, therefore, likely represents a true relationship between the variables.
- Statistical significance is usually indicated by the alpha value (or probability value), which should be smaller than a chosen significance level.

State alpha value

- Alpha error (type I) is Rejecting a true null hypothesis (which says that there is no difference between observed and expected).

For the majority of the studies, alpha is 0.05

Meaning: that the investigator has set 5% as the maximum chance of incorrectly rejecting the null hypothesis.

Degree of freedom

Calculation

- For Goodness of Fit = Number of levels (outcome)-1
- For independent variables / Homogeneity of proportion : (No. of columns – 1) (No. of rows – 1)

The Chi-Square Distribution

- No negative values
- Mean is equal to the degrees of freedom
- The standard deviation increases as degrees of freedom increase, so the chi-square curve spreads out more as the degrees of freedom increase.
- As the degrees of freedom become very large, the shape becomes more like the normal distribution.

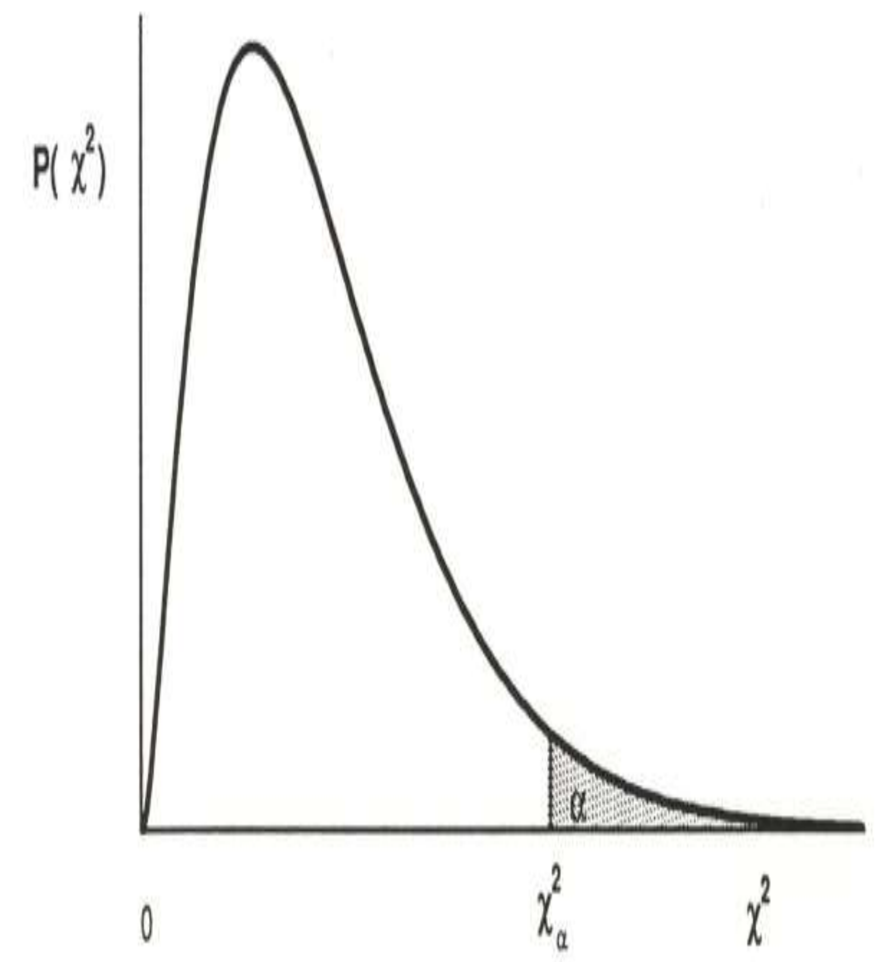


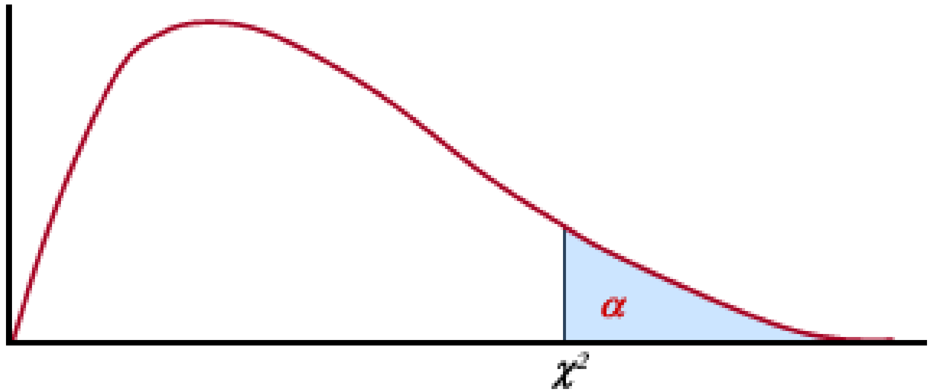
Table of the chi square distribution

df	0.200	0.100	0.075	0.050	0.025	0.010	0.005	0.001	0.0005
				Level of Significance α					
1	1.642	2.706	3.170	3.841	5.024	6.635	7.879	10.828	12.116
2	3.219	4.605	5.181	5.991	7.378	9.210	10.597	13.816	15.202
3	4.642	6.251	6.905	7.815	9.348	11.345	12.838	16.266	17.731
4	5.989	7.779	8.496	9.488	11.143	13.277	14.860	18.467	19.998
5	7.289	9.236	10.008	11.070	12.833	15.086	16.750	20.516	22.106
6	8.558	10.645	11.466	12.592	14.449	16.812	18.548	22.458	24.104
7	9.803	12.017	12.883	14.067	16.013	18.475	20.278	24.322	26.019
8	11.030	13.362	14.270	15.507	17.535	20.090	21.955	26.125	27.869
9	12.242	14.684	15.631	16.919	19.023	21.666	23.589	27.878	29.667
10	13.442	15.987	16.971	18.307	20.483	23.209	25.188	29.589	31.421
11	14.631	17.275	18.294	19.675	21.920	24.725	26.757	31.265	33.138
12	15.812	18.549	19.602	21.026	23.337	26.217	28.300	32.910	34.822
13	16.985	19.812	20.897	22.362	24.736	27.688	29.820	34.529	36.479
14	18.151	21.064	22.180	23.685	26.119	29.141	31.319	36.124	38.111
15	19.311	22.307	23.452	24.996	27.488	30.578	32.801	37.698	39.720
16	20.465	23.542	24.716	26.296	28.845	32.000	34.267	39.253	41.309
17	21.615	24.769	25.970	27.587	30.191	33.409	35.719	40.791	42.881
18	22.760	25.989	27.218	28.869	31.526	34.805	37.157	42.314	44.435
19	23.900	27.204	28.458	30.144	32.852	36.191	38.582	43.821	45.974
20	25.038	28.412	29.692	31.410	34.170	37.566	39.997	45.315	47.501

The Chi-Square Distribution

The chi-square distribution is different for each value of the degrees of freedom, different critical values correspond to degrees of freedom.

We find the critical value that separates the area defined by α from that defined by 1 – α.



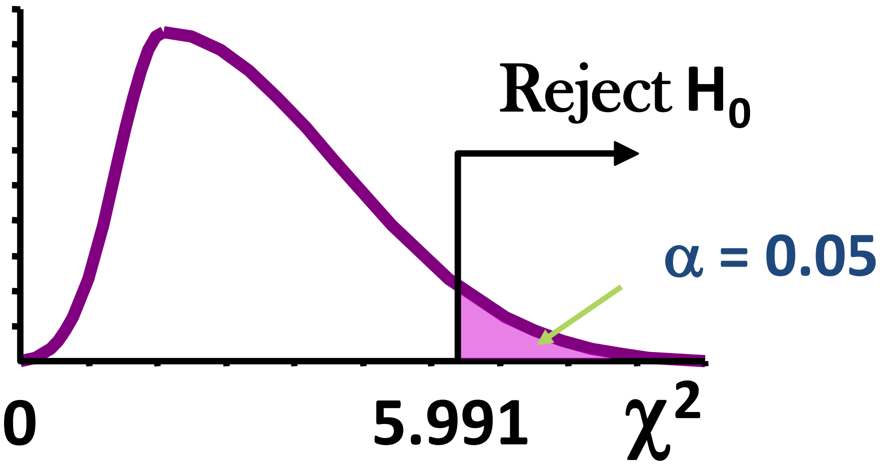
Finding Critical Value

Q. What is the critical χ^2 value if *df = 2*, and α =0.05?

If $n_i = E(n_i)$, $\chi^2 = 0$

Do not reject H_0

df =2

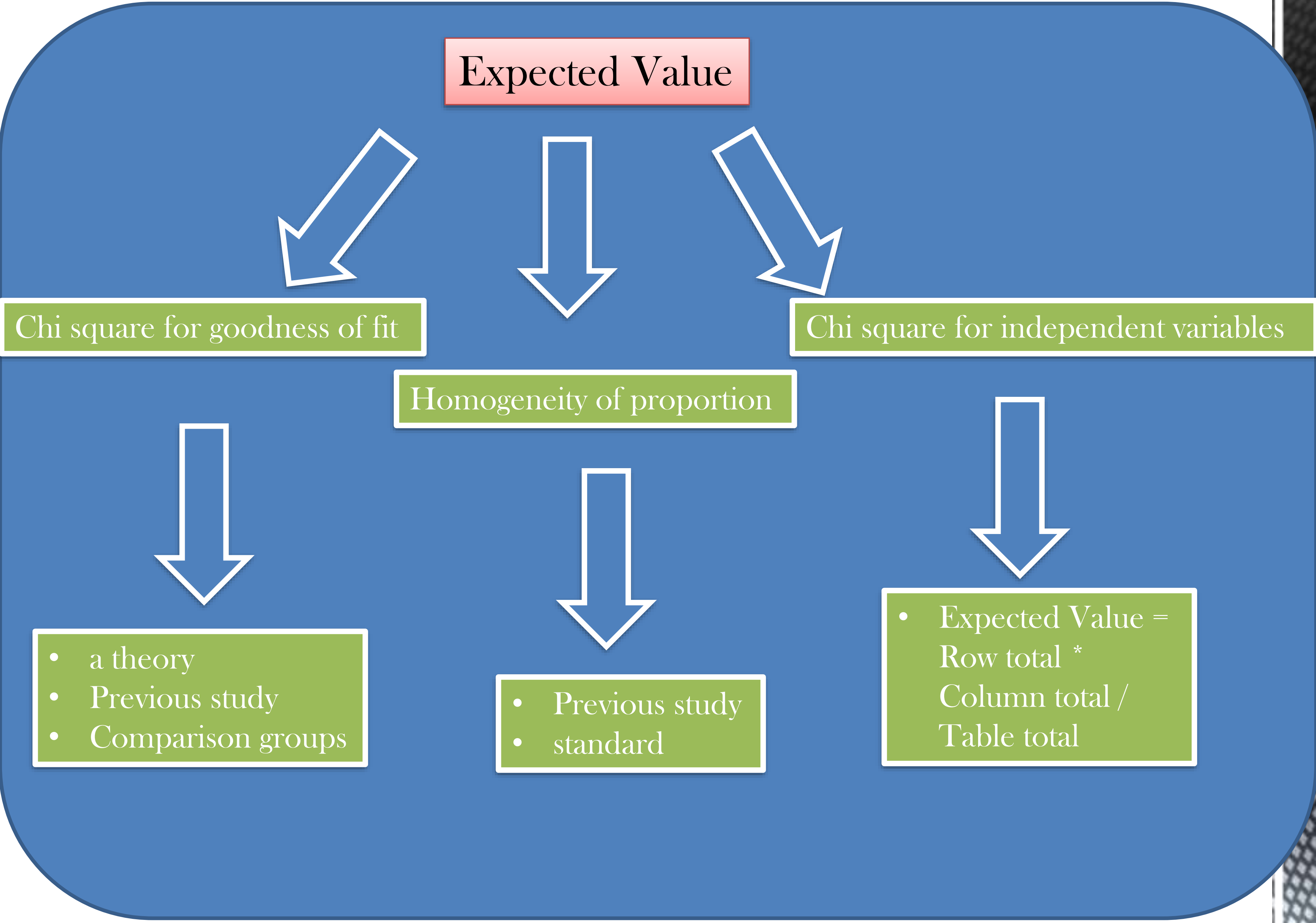


χ^2 Table (Portion)

	Significance level				
DF	0.995	…	0.95	…	0.05
1	…	…	0.004	…	3.841
2	0.010	…	0.103	…	5.991

State decision rule

If the value obtained is greater than the critical value of chi square , the null hypothesis will be rejected


Expected Value
Chi square for goodness of fit
Homogeneity of proportion
Chi square for independent variables
• a theory
• Previous study
• Comparison groups
• Previous study
• standard
• Expected Value = Row total * Column total / Table total

State and interpret results

See whether the value of chi square is more than or less than the critical value

If the value of chi square is less than the critical value we accept the null hypothesis

If the value of chi square is more than the critical value the null hypothesis can be rejected

Chi-Square Tests

Chi-Square Tests: These tests check whether the differences or patterns between two groups are real or just **random.**

Types:

- **Chi-Square Goodness of Fit Test:** Tests whether the observed frequencies in a categorical dataset match the expected frequencies based on a specific hypothesis.

- **Chi-Square Test of Independence**: Assesses whether two categorical variables (in row and columns) are independent of each other.

- **Chi-Square Test for Homogeneity of Proportions:** Compares the distributions of a categorical variable across different populations.

Take Home Message

1. The chi-square test applied to Qualitative data may be nominal or ordinal.

2. Before applying the Chi-square test, see all assumptions are met.

3. If the value of chi-square is large >>>, there is a high probability of rejecting the null hypothesis.

4. If the value of chi-square is small >>>, there is less probability of rejecting the null hypothesis

Fisher's Exact Test

- Used when sample sizes are small, and the Chi-square test may not be appropriate.
- It tests for independence between two categorical variables in a 2x2 contingency table.

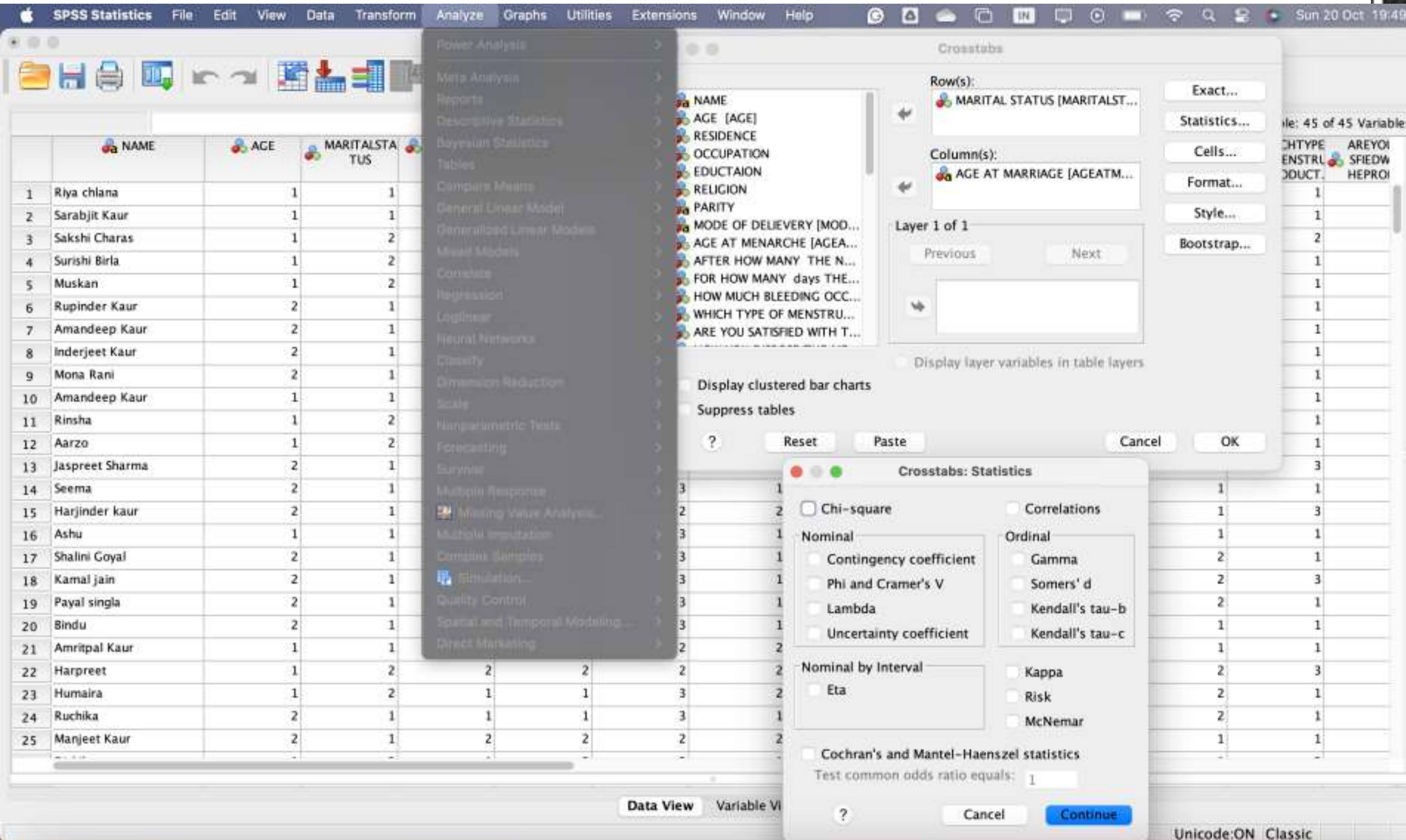
Cochran (1954) suggests

The decision regarding the use of Chi-square should be guided by the following considerations:

1. When **N > 40**, use Chi-square corrected for continuity.
2. When **N is between 20 and 40**, the Chi-square test may be used if all the expected frequencies are $\geq$five.

   If any expected frequency is less than 5, use the Fisher's Exact probability test.

3. When **N < 20**, use Fisher's test in all cases.


SPSS Statistics File Edit View Data Transform Analyze Graphs Utilities Extensions Window Help
Sun 20 Oct 19:49
Crosstabs
NAME
AGE [AGE]
RESIDENCE
OCCUPATION
EDUCTAION
RELIGION
PARITY
MODE OF DELIEVERY [MOD...
AGE AT MENARCHE [AGEA...
AFTER HOW MANY THE N...
FOR HOW MANY days THE...
HOW MUCH BLEEDING OCC...
WHICH TYPE OF MENSTRU...
ARE YOU SATISFIED WITH T...
Row(s):
MARITAL STATUS [MARITALST...
Column(s):
AGE AT MARRIAGE [AGEATM...
Layer 1 of 1
Previous
Next
Display layer variables in table layers
Exact...
Statistics...
Cells...
Format...
Style...
Bootstrap...
Display clustered bar charts
Suppress tables
Reset
Paste
Cancel
OK
Crosstabs: Statistics
Chi-square
Correlations
Nominal
Contingency coefficient
Phi and Cramer's V
Lambda
Uncertainty coefficient
Ordinal
Gamma
Somers' d
Kendall's tau-b
Kendall's tau-c
Nominal by Interval
Eta
Kappa
Risk
McNemar
Cochran's and Mantel-Haenszel statistics
Test common odds ratio equals: 1
Cancel
Continue
NAME
AGE
MARITALSTATUS
Riya chlana
Sarabjit Kaur
Sakshi Charas
Surishi Birla
Muskan
Rupinder Kaur
Amandeep Kaur
Inderjeet Kaur
Mona Rani
Amandeep Kaur
Rinsha
Aarzo
Jaspreet Sharma
Seema
Harjinder kaur
Ashu
Shalini Goyal
Kamal jain
Payal singla
Bindu
Amritpal Kaur
Harpreet
Humaira
Ruchika
Manjeet Kaur
Data View
Unicode:ON Classic

Yate's Correction

Chi-square distribution is a continuous distribution, and it fails to maintain its **continuity** even if any one of the expected frequencies is less than 5.

In such cases, Yates Correction for continuity is applied to maintain the character of continuity of the distribution.

The formula for the Chi-square test with Yates correction is:

$$\text{Chi-square} = \frac{N\,(\,|\text{observed} - \text{expected}| - 0.5\,)^2}{\text{expected}}$$

Phi-Coefficient

- It is only used on 2X2 contingency tables.

- Interpreted as a measure of the relative (strength) of an association between two variables ranging from 0 to 1

$$Phi\ (\phi) = \sqrt{\frac{\chi^2}{n}}$$

n = total number of observation

$$= \frac{ad-b}{\sqrt{(a+b)(a+c)(c+d)(b+d)}}$$

Sex	Smoking		Total
	Yes	No	
Male	a	b	a+b
Female	c	d	C+d
Total	a+c	b+d	n

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma y^2 - (\Sigma y)^2]}}$$

Pearson's Contingency Coefficient (C)

- It is interpreted as a measure of relative (strength) of an association between two variables
- The coefficient will always be less than 1 and varies according to the number of rows and columns.
- This can be used for general rXc tables.
- It ranges between 0 to 1

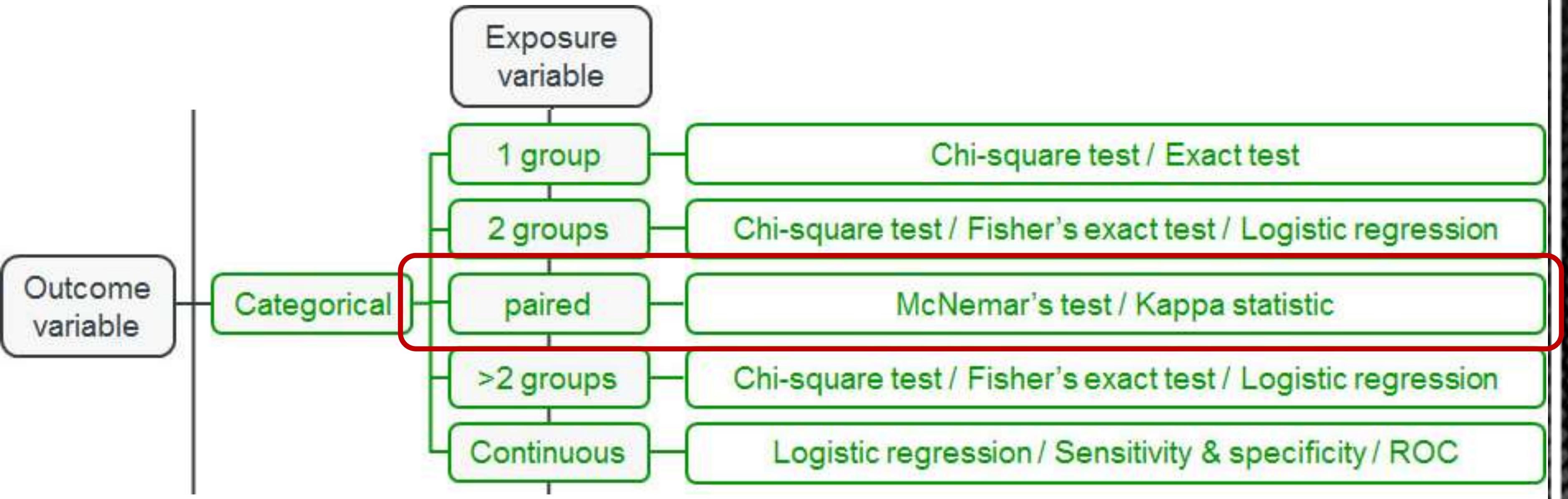
$$C = \sqrt{\frac{\chi^2}{n + \chi^2}} = \sqrt{\frac{}{1 + \phi}}$$

Cramer's V Coefficient (V)

- It is useful for comparing multiple $\chi 2$ test statistics and is generalizable across tables of varying sizes.
- It is not affected by sample size and, therefore is very useful in situations where you expect statistically significant chi-square was the result of large sample size instead of any substantive relationship between the variables.
- It is interpreted as a measure of the relative (strength) of an association between variables.
- The coefficient ranges from 0 to 1 (perfect association).
- In practice, you may find that a Cramer's V of 0.10 provides a good minimum threshold for suggesting there is a substantive relationship two variables

$$C = \sqrt{\frac{\chi^2}{n(q-1)}}$$

Where, q= smaller number of rows or columns


Exposure variable
Outcome variable
Categorical
1 group
Chi-square test / Exact test
2 groups
Chi-square test / Fisher's exact test / Logistic regression
paired
McNemar's test / Kappa statistic
>2 groups
Chi-square test / Fisher's exact test / Logistic regression
Continuous
Logistic regression / Sensitivity & specificity / ROC

McNemar's Test

- Used in case of two paired/related samples or there are repeated measurements.
- It can be used to test for the significance of changes in "before-after" designs in which each person is used as his own control.

- Thus, the test can be used
  - ❖ to test the effectiveness of a treatment /training/ program/ therapy/intervention

    or

  - ❖ to compare the ratings of two judges on the same set of individuals.

Kappa Statistic

- The Kappa Statistic measures the agreement between the evaluations of two examiners when both are rating the same objects.

- It describes agreement achieved beyond chance as a proportion of that agreement that is possible beyond chance.

- The value of the Kappa Statistic ranges from -1 to 1, with larger values indicating better reliability.

  A value of 1 indicates perfect agreement.

  A value of 0 indicates that agreement is no better than chance.

- Generally, a Kappa > 0.60 is considered satisfactory.

Kappa Statistic

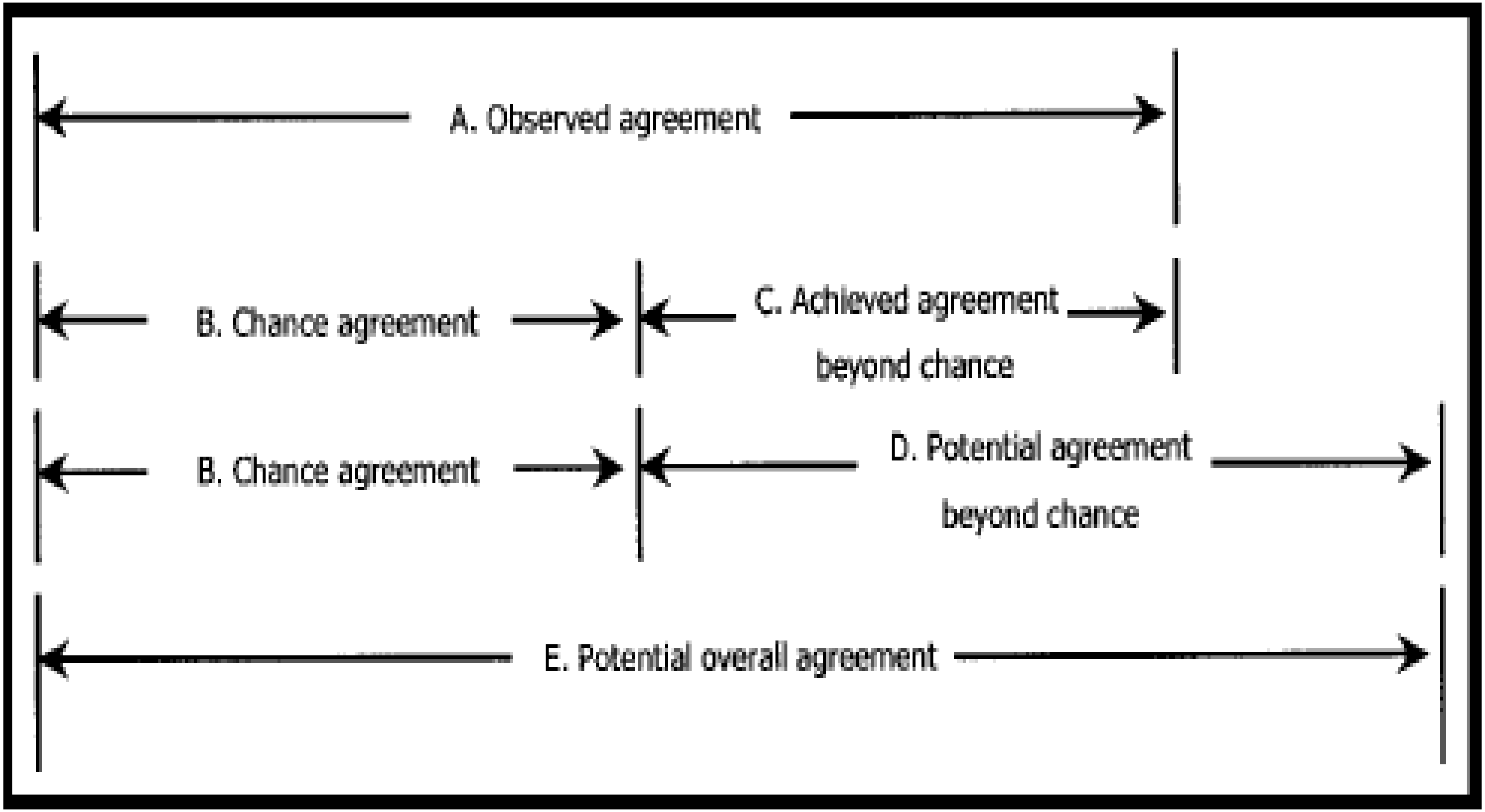


Figure.
Schematic representation of the relationship of kappa to overall and chance agreement. Kappa=C/D. Adapted from Rigby.[24]

Kappa Statistic

$$Kappa = \frac{P_0 - P_E}{1 - P_E}$$

Where:
P_0 = proportion of observed agreement
P_E = proportion of expected agreement by chance

0.00	Agreement is no better than chance
0.01-0.20	Slight agreement
0.21-0.40	Fair agreement
0.41-0.60	Moderate agreement
0.61-0.80	Substantial agreement
0.81-0.99	Almost perfect agreement
1.00	Perfect agreement

Initiating the categorical data analysis

4. Advanced Analysis

- **Logistic regression:** If you're analyzing the relationship between categorical and continuous variables or a binary outcome, logistic regression is appropriate.

- **Chi-square automatic interaction detector (CHAID):** This method builds decision trees using categorical data, often used in market research and healthcare to identify significant predictors.

3. Logistic Regression

- **Binary Logistic Regression**: Used when the dependent variable has two categories (e.g., success/failure). It estimates the probability of an outcome based on one or more predictor variables.

- **Multinomial Logistic Regression:** Used when the dependent variable has more than two categories, but the categories do not have an inherent order.

- **Ordinal Logistic Regression**: Used when the dependent variable has ordered categories (e.g., low, medium, high).

Initiating the categorical data analysis

5. Test for Homogeneity of proportions or Independence

- Chi-square test for association tests (2x2): whether two categorical variables are associated
- Chi-square test of independence (R x C): used to test a variety of sizes of contingency tables
- Chi-square goodness-of-fit test: whether the distribution of cases in a single categorical variable follows a known/hypothesised distribution
- Chi-square test of homogeneity: whether the proportions in each group are equal in the population
- **Fisher's exact test:** If sample sizes are small, this test can be more accurate than the chi-square test for testing independence.

Chi-square goodness-of-fit test

It is also called Pearson's chi-square goodness-of-fit test.

The chi-square goodness-of-fit test is a single-sample nonparametric test.

Q: How "close" are the observed values to those which would be expected in a study

OR

Q: An administrator at a hospital may want to determine whether an equal number of people are hospitalised each day of the week to better plan staffing levels.?

Expected frequency can be based on

- theory
- previous experience
- comparison groups

Example: Are cancer-related deaths affected by seasonal variations??

Null Hypothesis: The proportion of deaths due to cancer in winter, summer, autumn, spring is equal = ¼ = 25%
Alternative: Not all probabilities stated a in null hypothesis is correct

Cancer deaths	Observed	Expected = 322*1/4
Summer	78	80.5
Spring	71	80.5
Autumn	87	80.5
Winter	86	80.5
Total	322	

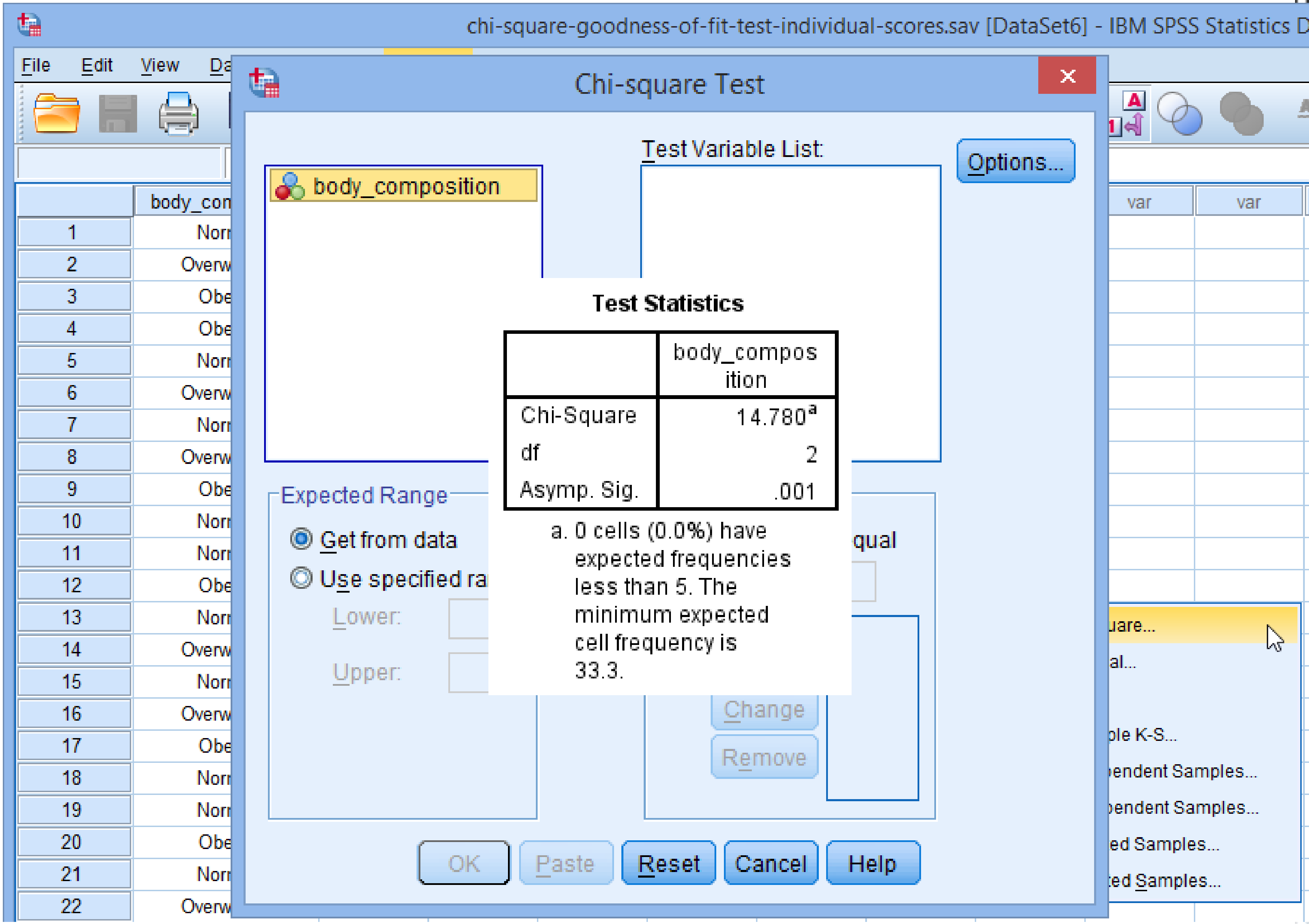
Degree of freedom = k-1 = 4-1 =3

For α =0.05 for df =3 critical value X^2 = 7.81

$$\chi^2 = \sum \frac{(observed\ frequency - expected\ frequency)^2}{expected\ frequency}$$

$X^2 = (78-80.5)^2/80.5 + (71-80.5)^2/80.5 + (87.5 - 80.5)^2/80.5 + (86 - 80.5)^2/80.5 = 2.09$

Conclusion: As calculated X^2 value is less than Critical value we can **accept the null hypothesis** and state that deaths due to cancer across seasons are not statistically different from what's expected by chance (i.e. all seasons being equal)



Test Statistics

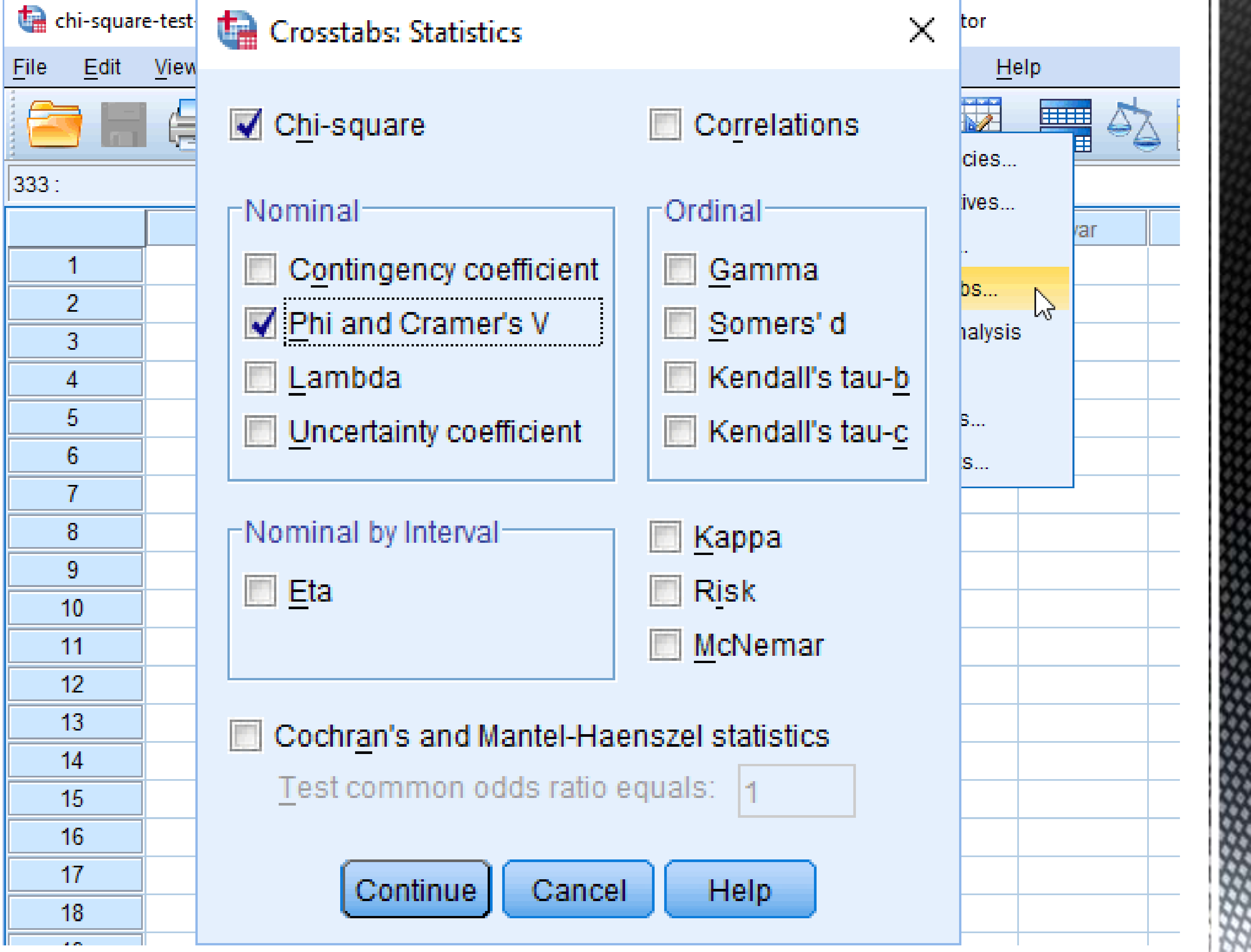
	body_composition
Chi-Square	14.780[a]
df	2
Asymp. Sig.	.001

a. 0 cells (0.0%) have expected frequencies less than 5. The minimum expected cell frequency is 33.3.

Chi-square for independence

It focuses on contingency tables that are greater than 2 x 2, which are often referred to as r x c contingency tables.

It tests whether two variables measured at the nominal level are independent (i.e., whether there is an association between the two variables).


chi-square-test
File
Edit
View
Help
333 :
Crosstabs: Statistics
Chi-square
Correlations
Nominal
Contingency coefficient
Phi and Cramer's V
Lambda
Uncertainty coefficient
Ordinal
Gamma
Somers' d
Kendall's tau-b
Kendall's tau-c
Nominal by Interval
Eta
Kappa
Risk
McNemar
Cochran's and Mantel-Haenszel statistics
Test common odds ratio equals:
1
Continue
Cancel
Help

Other analytical approaches

- **Probit Regression:** Similar to logistic regression, but it assumes a normal cumulative distribution function instead of a logistic one, often used in binary outcome models.

- **Cochran-Mantel-Haenszel Test:** Tests for an association between two categorical variables while controlling for a third variable (stratification).

- **Log-Linear Models:** Used to model the relationships between three or more categorical variables by modeling the logarithm of the expected cell frequencies in a contingency table.

- **Cluster Analysis (for Categorical Data):** Methods like k-modes or latent class analysis (LCA) group observations into clusters based on categorical attributes.

Initiating the categorical data analysis

6. Interpretation

- Analyse p-values (typically < 0.05 for significance).
- Assess effect sizes (e.g., Cramér's V for associations).
- Interpret visualisations (e.g., bar plots, mosaic plots).

Thank You!

http://dilbert.com/strips/comic/2007-05-10/