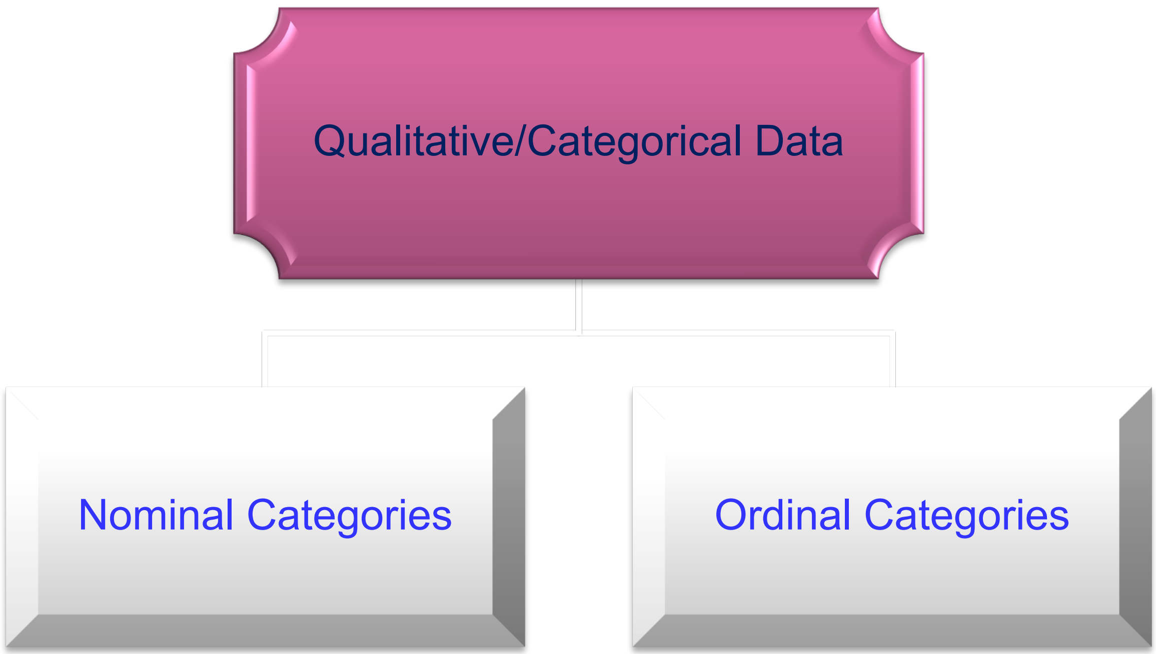
Statistical tests to observe the statistical significance of Categorical variables

Dr.Shaikh Shaffi Ahamed Ph.D.,

Professor

Dept. of Family & Community Medicine

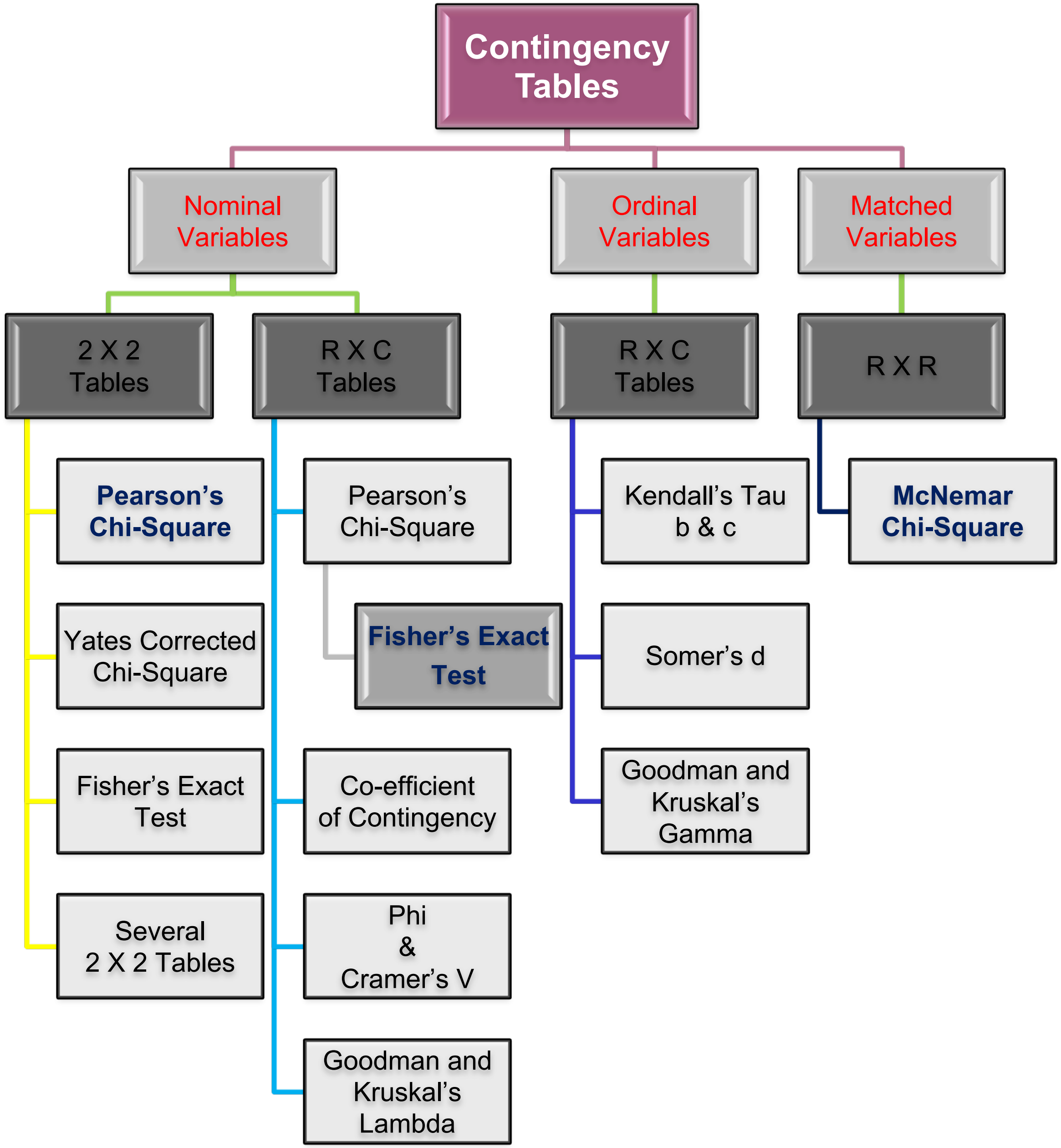
Types of Categorical Data



Types of Analysis for Categorical Data

Type of Analysis

- Descriptive
  Rate and Ratio
- Analytic
  Confidence Interval
  and
  Test of Significance



Contingency Tables
Nominal Variables
Ordinal Variables
Matched Variables
2 X 2 Tables
R X C Tables
R X C Tables
R X R
Pearson's Chi-Square
Yates Corrected Chi-Square
Fisher's Exact Test
Several 2 X 2 Tables
Pearson's Chi-Square
Fisher's Exact Test
Co-efficient of Contingency
Phi & Cramer's V
Goodman and Kruskal's Lambda
Kendall's Tau b & c
Somer's d
Goodman and Kruskal's Gamma
McNemar Chi-Square

Choosing the appropriate Statistical test

- Based on the three aspects of the data
  - Types of variables
  - Number of groups being compared &
  - Sample size

Statistical tests

Chi-square test:

Study variable: Qualitative
Outcome variable: Qualitative
Comparison: two or more proportions
Sample size: > 20
Expected frequency: > 5

Fisher's exact test:

Study variable: Qualitative
Outcome variable: Qualitative
Comparison: two proportions
Sample size:< 20

Macnemar's test: (for paired samples)

Study variable: Qualitative
Outcome variable: Qualitative
Comparison: two proportions
Sample size: Any

Chi-square test

Purpose

To find out whether the association between two categorical variables are statistically significant

Null Hypothesis

There is no association between two variables

Chi-Square test

$$X^2 = \sum \left[\frac{(o - e)^2}{e} \right]$$

Figure for Each Cell

1. *The summation is over all cells of the contingency table consisting of r rows and c columns*

2. *O is the observed frequency*

3. $\hat{E}$ *is the expected frequency*

$$\hat{E} = \frac{\begin{pmatrix}\text{total of row in}\\ \text{which the cell lies}\end{pmatrix} \cdot \begin{pmatrix}\text{total of column in}\\ \text{which the cell lies}\end{pmatrix}}{(\text{total of all cells})}$$

reject H_o if $\chi^2 > \chi^2_{.\alpha, df}$

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

where df = $(r-1)(c-1)$

4. *The degrees of freedom are df = (r-1)(c-1)*

Requirements

- Prior to using the chi square test, there are certain requirements that must be met.
  - The data must be in the form of frequencies counted in each of a set of categories. Percentages cannot be used.
  - The total number observed must exceed 20.

Requirements

- The expected frequency under the H_0 hypothesis in any one fraction must not normally be less than 5.
- All the observations must be independent of each other. In other words, one observation must not have an influence upon another observation.

APPLICATION OF CHI-SQUARE TEST

- TESTING INDEPENDCNE (or ASSOCATION)
- TESTING FOR HOMOGENEITY
- TESTING OF GOODNESS-OF-FIT

Chi-square test

- **Objective** : **Smoking is a risk factor for MI**
- **Null Hypothesis: Smoking does not cause MI**

	D (MI)	No D(No MI)	Total
Smokers	29	21	50
Non-smokers	16	34	50
Total	45	55	100

Chi-Square test

	MI	**Non-MI**
Smoker	**29** O E	**21** O E
Non-Smoker	**16** O E	**34** O E

Chi-square test

	MI	Non-MI	
Smoker	29 O E	21 O E	50
Non-smoker	16 O E	34 O E	50
	45	55	100

Chi-square test

	MI	Non-MI	
Smoker	29 O, 22.5 E	21 O, E	50
Non-smoker	16 O, E	34 O, E	50
	45	55	100

$$22.5 = \frac{50 \text{ X } 45}{100}$$

Chi-square test

	MI	No MI	
smoker	29 O E 22.5	21 O E 27.5	50
Non smoker	16 O E 22.5	34 O E 27.5	50
	45	55	100

Chi-Square

Degrees of Freedom

$df = (r-1)(c-1)$

$= (2-1)(2-1) = 1$

Critical Value (Table A.6) = 3.84

$X^2 = 6.84$

Calculated value(6.84) is greater than critical (table) value (3.84) at 0.05 level with 1 d.f.f

Hence we reject our Ho and conclude that there is highly statistically significant association between smoking and MI.

Chi- square test

Find out whether the gender is equally distributed among each age group

	Age			
Gender	<30	30-45	>45	Total
Male	60 (60)	20 (30)	40 (30)	120
Female	40 (40)	30 (20)	10 (20)	80
Total	100	50	50	200

Test for Homogeneity (Similarity)

To test similarity between frequency distribution or group. It is used in assessing the similarity between non-responders and responders in any survey

Age (yrs)	Responders	Non-responders	Total
<20	76 (82)	20 (14)	96
20 – 29	288 (289)	50 (49)	338
30-39	312 (310)	51 (53)	363
40-49	187 (185)	30 (32)	217
>50	77 (73)	9 (13)	86
Total	940	160	1100

Association between Diabetes and Heart Disease?

- **Background:**

Contradictory opinions:

- 1. A diabetic's risk of dying after a first heart attack is the same as that of someone without diabetes. There is no link between diabetes and heart disease.

vs.

- 2. Diabetes takes a heavy toll on the body and diabetes patients often suffer heart attacks and strokes or die from cardiovascular complications at a much younger age.

- So we use hypothesis test based on the latest data to see what's the right conclusion.
- There are a total of 5167 managed-care patients, among which <u>1131 patients are non-diabetics</u> and <u>4036 are diabetics</u>. Among the non-diabetic patients, <u>42%</u> of them had their blood pressure properly controlled (therefore it's <u>475 of 1131</u>). While among the diabetic patients only <u>20%</u> of them had the blood pressure controlled (therefore it's <u>807 of 4036</u>).

Association between Diabetes and Heart Disease?

- Data

	Controlled	Uncontrolled	Total
Non-diabetes	475	656	1131
Diabetes	807	3229	4036
Total	1282	3885	5167

Association between Diabetes and Heart Disease?

Data:
Diabetes: 1=Not have diabetes, 2=Have Diabetes
Control: 1=Controlled, 2=Uncontrolled

DIABETES * CONTROL Crosstabulation

Count

		CONTROL		
		1.00	2.00	Total
DIABETES	1.00	475	656	1131
	2.00	807	3229	4036
Total		1282	3885	5167

Association between Diabetes and Heart Disease?

DIABETES * CONTROL Crosstabulation

			CONTROL		
			1.00	2.00	Total
DIABETES	1.00	Count	475	656	1131
		% within DIABETES	42.0%	58.0%	100.0%
		% within CONTROL	37.1%	16.9%	21.9%
		% of Total	9.2%	12.7%	21.9%
	2.00	Count	807	3229	4036
		% within DIABETES	20.0%	80.0%	100.0%
		% within CONTROL	62.9%	83.1%	78.1%
		% of Total	15.6%	62.5%	78.1%
Total		Count	1282	3885	5167
		% within DIABETES	24.8%	75.2%	100.0%
		% within CONTROL	100.0%	100.0%	100.0%
		% of Total	24.8%	75.2%	100.0%

Association between Diabetes and Heart Disease?

Hypothesis test:

1) H_0: There is no association between diabetes and heart disease. (There is no association between diabetes and heart disease. (or) Diabetes and heart disease are independent.)
2) H_A: There is a associaton between diabetes and heart disease. (There is an association between diabetes and heart disease. (or) Diabetes and heart disease are dependent.)
3) Assume a significance level of .05

Association between Diabetes and Heart Disease?

SPSS Output

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	229.268[b]	1	.000		
Continuity Correction[a]	228.091	1	.000		
Likelihood Ratio	212.149	1	.000		
Fisher's Exact Test				.000	.000
Linear-by-Linear Association	229.224	1	.000		
N of Valid Cases	5167				

a. Computed only for a 2x2 table

b. 0 cells (.0%) have expected count less than 5. The minimum expected count is 280.62.

Association between Diabetes and Heart Disease?

4) The computer gives us a Chi-Square Statistic of 229.268
5) The computer gives us a p-value of .000 i.e., (<0.0001).
6) Because our p-value is less than alpha (0.05), we would reject the null hypothesis.
7) There is sufficient evidence to conclude that there is an association between diabetes and heart disease.

Example

- The following data relate to suicidal feelings in samples of psychotic and neurotic patients:

	Psychotics	Neurotics	Total
Suicidal feelings	2	6	8
No suicidal feelings	18	14	32
Total	20	20	40

Example

- The following data compare malocclusion of teeth with method of feeding infants.

	Normal teeth	Malocclusion
Breast fed	4	16
Bottle fed	1	21

Fisher’s Exact Test:

- The method of Yates's correction was useful when manual calculations were done. Now different types of statistical packages are available. Therefore, it is better to use Fisher's exact test rather than Yates's correction as it gives exact result.

$$Fisher's\ Exact\ Test = \frac{R_1!R_2!C_1!C_2!}{n!a!b!c!d!}$$

What to do when we have a paired samples and both the exposure and outcome variables are qualitative variables (Binary).

Problem

- A researcher has done a matched case-control study of endometrial cancer (cases) and exposure to conjugated estrogens (exposed).
- In the study cases were individually matched 1:1 to a non-cancer hospital-based control, based on age, race, date of admission, and hospital.

McNemar's test

Situation:

- **Two paired binary variables that form a particular type of 2 x 2 table**
- **e.g. matched case-control study or cross-over trial**

Data

	Cases	Controls	Total
Exposed	55	19	74
Not exposed	128	164	292
Total	183	183	366

- can’t use a chi-squared test - observations are not independent - they’re paired.
- we must present the 2 x 2 table differently
- each cell should contain a count of the number of pairs with certain criteria, with the columns and rows respectively referring to each of the subjects in the matched pair
- the information in the standard 2 x 2 table used for unmatched studies is insufficient because it doesn’t say who is in which pair - ignoring the matching

Data

	Controls		
Cases	Exposed	Not exposed	Total
Exposed	12	43	55
Not exposed	7	121	128
Total	19	164	183

We construct a matched 2 x 2 table:

	Controls		
Cases	Exposed	Not exposed	Total
Exposed	e	f	e+f
Not exposed	g	h	g+h
Total	e+g	f+h	n

Formula

The odds ratio is: f/g

The test is:

$$X^2 = \frac{(|f - g| - 1)^2}{f + g}$$

Compare this to the χ^2 distribution on 1 df

$$X^2 = \frac{(|43 - 7| - 1)^2}{43 + 7} = \frac{1225}{50} = 24.5$$

P <0.001, Odds Ratio = 43/7 = 6.1

$p_1 - p_2$ = (55/183) – (19/183) = 0.197 (20%)

s.e.($p_1 - p_2$) = 0.036

95% CI: 0.12 to 0.27 (or 12% to 27%)

- Degrees of Freedom

  df = (r-1) (c-1)

  = (2-1) (2-1) =1
- Critical Value (Table A.6) = 3.84
- X^2 = 25.92
- Calculated value(25.92) is greater than critical (table) value (3.84) at 0.05 level with 1 d.f.f
- Hence we reject our Ho and conclude that there is highly statistically significant association between Endometrial cancer and Estrogens.

Two-tailed critical ratios of χ^2

Degrees of freedom df	.10	.05	.02	.01
1	2.706	3.841	5.412	6.635
2	4.605	5.991	7.824	9.210
3	6.251	7.815	9.837	11.341
4	7.779	9.488	11.668	13.277
5	9.236	11.070	13.388	15.086
6	10.645	12.592	15.033	16.812
7	12.017	14.067	16.622	18.475
8	13.362	15.507	18.168	20.090
9	14.684	16.919	19.679	21.666
10	15.987	18.307	21.161	23.209
11	17.275	19.675	22.618	24.725
12	18.549	21.026	24.054	26.217
13	19.812	22.362	25.472	27.688
14	21.064	23.685	26.873	29.141
15	22.307	24.996	28.259	30.578

Stata Output

```
                 | Controls               |
Cases            |   Exposed    Unexposed |       Total
-----------------+------------------------+------------
         Exposed |        12           43 |          55
       Unexposed |         7          121 |         128
-----------------+------------------------+------------
           Total |        19          164 |         183

McNemar's chi2(1) =      25.92    Prob > chi2 = 0.0000
Exact McNemar significance probability       = 0.0000

Proportion with factor
        Cases       .3005464
        Controls    .1038251     [95% Conf. Interval]
                   ---------     --------------------
        difference  .1967213     .1210924    .2723502
        ratio       2.894737     1.885462    4.444269
        rel. diff.  .2195122     .1448549    .2941695

        odds ratio  6.142857     2.739772    16.18458  (exact)
```

Statistical Tests

Z-test:

Study variable: Qualitative

Outcome variable: Qualitative

Comparison: Sample proportion with population proportion; two sample proportions

Sample size: larger in each group(>30)

Test for sample proportion with population proportion

Problem

In an otological examination of school children, out of 146 children examined 21 were found to have some type of otological abnormalities. Does it confirm with the statement that 20% of the school children have otological abnormalities?

a . Question to be answered:

Is the sample taken from a population of children with 20% otological abnormality

b. Null hypothesis : The sample has come from a population with 20% otological abnormal children

Test for sample prop. with population prop.

c. Test statistics

$$z = \frac{p - P}{\sqrt{\frac{pq}{n}}} = \frac{14.4 - 20.0}{\sqrt{\frac{14.4 * 85.6}{146}}} = 1.69$$

P – Population. Prop.
p- sample prop.
n- number of samples

d.Comparison with theoritical value

Z ~ N (0,1); $Z_{0.05}$ = 1.96

The prob. of observing a value equal to or greater than 1.69 by chance is more than 5%. We therefore do not reject the Null Hypothesis

e. Inference

There is a evidence to show that the sample is taken from a population of children with 20% abnormalities

Example

Researchers wished to know if urban and rural adult residents of a developing country differ with respect to prevalence of a certain eye disease. A survey revealed the following information

Residence	Eye disease		Total
	Yes	No	
Rural	24	276	300
Urban	15	485	500

Test at 5% level of significance, the difference in the prevalence of eye disease in the 2 groups

Z-test for
(two independent sample proportions)

$$Z = \frac{P_1 - P_2}{\sqrt{\dfrac{P_1\,(1-P_1)}{n_1} + \dfrac{P_2\,(1-P_2)}{n_2}}}$$

P1= proportion in the first group

P2= proportion in the second group

n1= first sample size

n2= second sample size

Critical z =

- **1.96 at 5% level of significance**
- **2.58 at 1% level of significance**

Answer

P1 = 24/300 = 0.08   p2 = 15/500 = 0.03

$$Z = \frac{0.08 - 0.03}{\sqrt{\frac{0.08(1-0.08)}{300} + \frac{0.03(1-0.03)}{500}}} = 2.87$$

2.87 > 1.96 (from Z-table at α=0.05)

Hence we can conclude that,

the difference of prevalence of eye disease between the two groups is statistically significant

In Conclusion !

When both the study variables and outcome variables are categorical (Qualitative):

Apply

(i) Chi square test (for two and more than two groups)

(ii) Fisher's exact test (Small samples)

(iii) Mac Nemar's test (for paired samples)

(iv) Z-test for single sample(comparing sample proportion with population proportion) and two samples(two sample proportions)