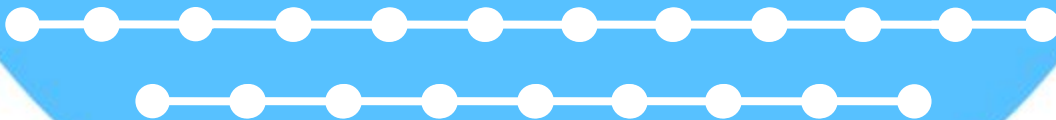




21

STATISTICAL TESTS FOR CATEGORICAL VARIABLES



KSU COLLEGE OF MEDICINE
2019 - 2020

ACKNOWLEDGMENTS

DONE BY

Shahad Alzahrani

Hadeel Awartani

REVIEWER

Afnan Almustafa

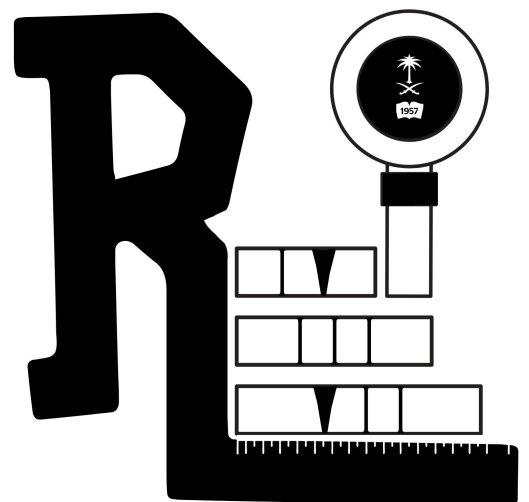
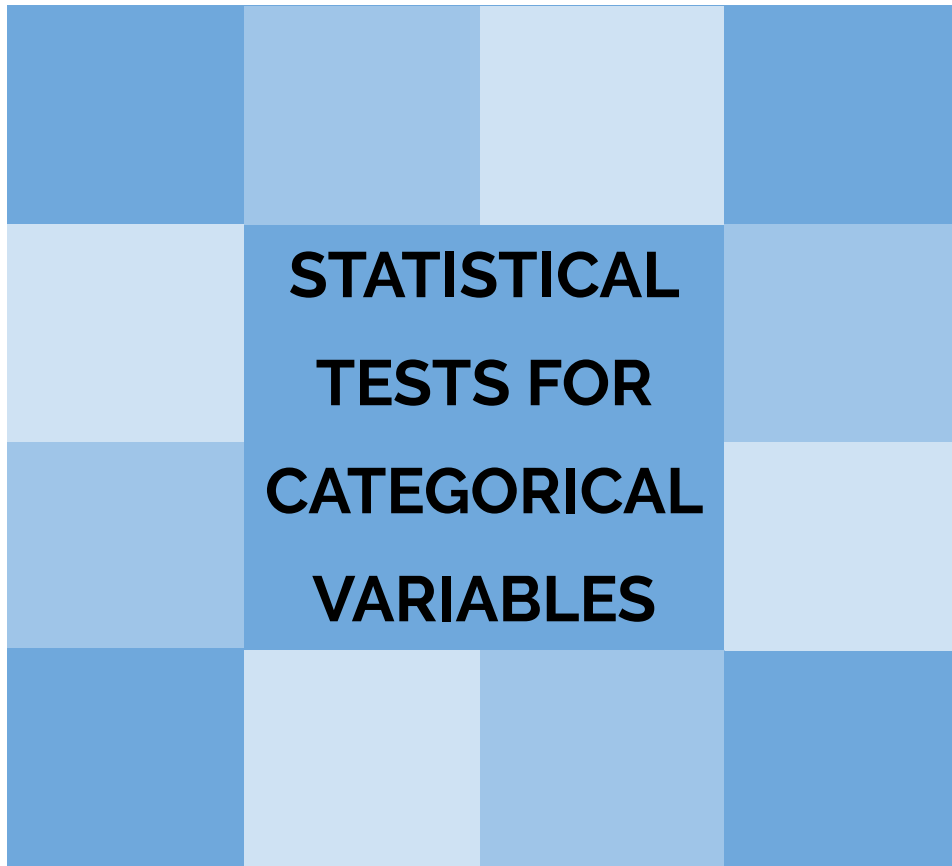


TABLE OF CONTENTS



LECTURE OBJECTIVES



By the end of this lecture, I am able to:

- No Objectives
-
-

Type of Qualitative/Categorical Data :

1- Nominal Categories

2- Ordinal Categories

These categories are ordered

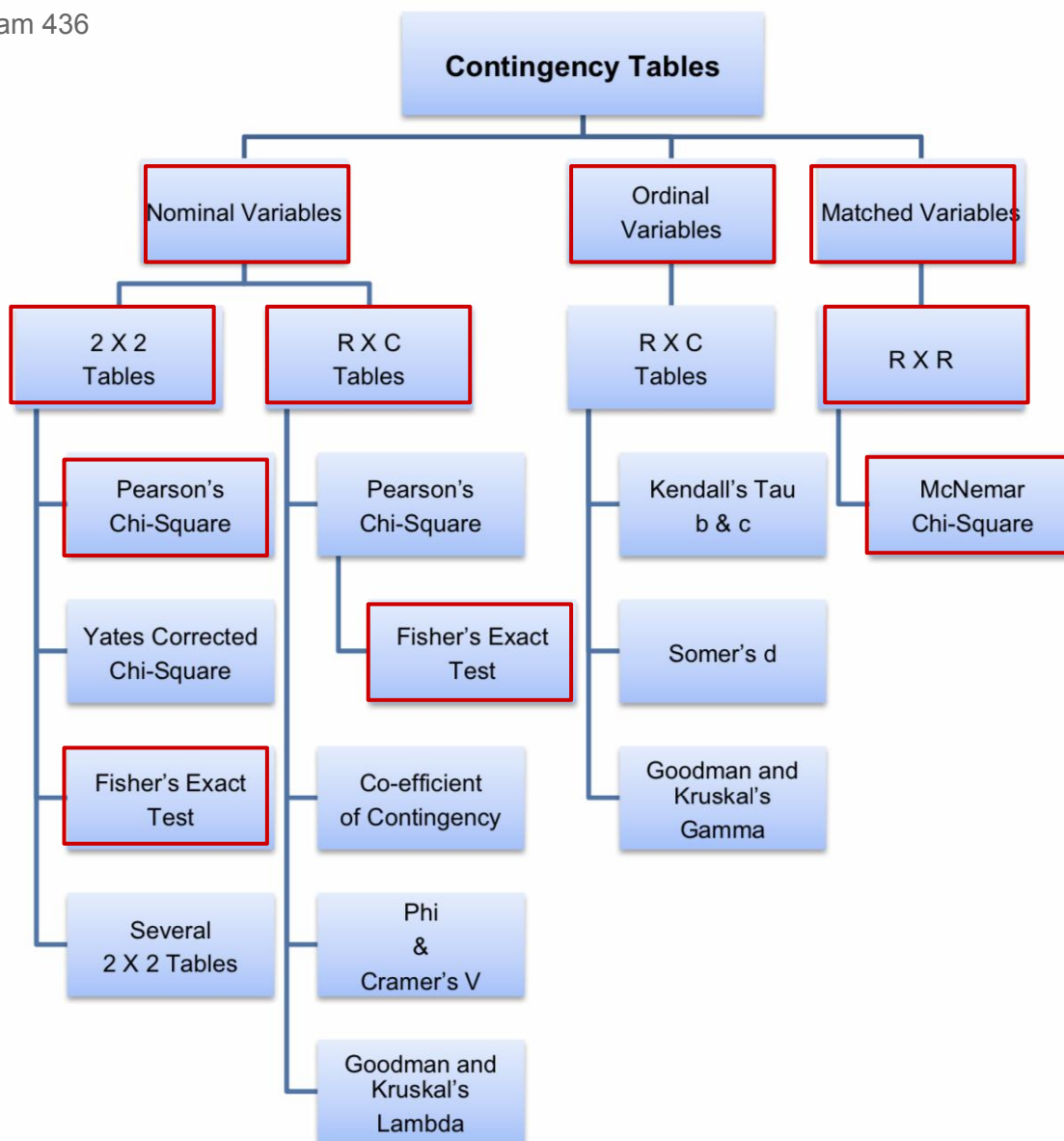
Type of Analysis for Categorical Data :

1- Descriptive Rate and Ratio

2- Analytic Confidence Interval and Test of Significance

Team 436

Column= outcome
Row=exposure



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 :

Study variables can be multiple, outcome only 1	Chi-square test	Fisher's exact test	Macnemar's test: (for paired samples)
Study variable Age, gender, income	Qualitative	Qualitative	Qualitative
Outcome variable	Qualitative	Qualitative	Qualitative
Comparison	two or more proportions	two proportions	two proportions
Sample size	> 20	< 20	Any
Expected frequency	> 5		

Qualitative data- categorical data

Chi-square test :


Purpose :

To find out whether the associated between two categorical variable are statistically significant

Null Hypothesis:

There is no association between two variables

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

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. **E** is the expected frequency.

3. \hat{E} is the expected frequency

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

reject H_0 if $\chi^2 > \chi^2_{\alpha,df}$

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

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

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

How many cells are independent

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. Use the frequency (actual #, not %)
- The total number observed must exceed 20. (Sample size)
- The expected frequency under the H 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 independence (or association)
- 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

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

Is there an association between smoking & MI?

- MI \ smoker = 22.5
- NonMI \ smoker = 27.5
- MI \ Non smoker = 22.5
- NonMI \ Non smoker = 22.5

Remember; $E > S$
When will $E < S$? If sample size is small that's why S must be > 20

O= observed, what frequency I saw from the data
E= expected frequency, what I expect

-Degrees of Freedom
 $df = (r-1)(c-1)$
 $= (2-1)(2-1) = 1$

-Critical Value (Table A.6) = 3.84
 $\chi^2 = 6.84$

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

-Hence we reject our H_0 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

Gender	Age			Total
	<30	30-45	>45	
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 1131 patients are non-diabetics and 4036 are diabetics. Among the non-diabetic patients, 42% of them had their blood pressure properly controlled (therefore it's 475 of 1131). While among the diabetic patients only 20% of them had the blood pressure controlled (therefore it's 807 of 4036).

-Data :

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

-Date: [Convert into codes](#)

Diabetes:

1=Not have diabetes
2=Have Diabetes

Control:

1=Controlled,
2=Uncontrolled Count

DIABETES * CONTROL Crosstabulation

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

DIABETES * CONTROL Crosstabulation

			CONTROL		Total
			1.00	2.00	
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) H₀: 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 an association 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

SPSS output :

The test we are using

Chi-Square Tests			P value		
	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.

- 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.

The higher the difference, the larger calculated value from Chi test the smaller p value the more the significance

Example:

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

Small sample; when you calculate E < S can't do Chi, use Fisher's Exact
 If only one < S you can still use Chi

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

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:

V. Complicated, know when to apply;

- in matching (case control)
- In cross over trials (small # of pts, give tx then watch then another tx- pt acts as intervention & placebo)

- 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 :**
 - 1- A researcher has done a matched case- control study of endometrial cancer (cases) and exposure to conjugated estrogen (exposed).
 - 2- 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. *The more the matching requirements the harder*

McNemar's test :

Situation:

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

Data:

Wrong way to put 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.

McNemar's test :

Data:

Correct way to put data

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

Then apply test

We construct a matched 2x2 table :

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

Test only uses f&g (mismatched), unlike Chi which takes all 4 cells

Formula :

-The odds ratio is : f/g

-The test is :

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

-Compare this to the 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

-p1-p2 = (55/183) - (19/183) = 0.197 (20%)

- s.e.(p1-p2)= 0.036

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

In case control, you will interpret the ODDS RATIO: the odds of EXPOSURE to _____ is 6.1 higher in case than control.

In COHORT you interpret the OUTCOME

-Degrees of Freedom: $df = (r-1)(c-1)$

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

-Critical Value (Table A.6) = 3.84

-X2 = 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

Z-test: normal distribution

Study variable	Qualitative
Outcome variable	Qualitative
Comparison	-Sample proportion with population proportion; -two sample proportions
Sample size	larger in each group(>30)

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

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 sample

D.comparison with theoretical value :

- $Z \sim N(0,1)$; $Z_{0.05} = 1.96$

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)

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

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

Middle= larger area = > 5% = not significant
 Extreme= smaller area = < 5% = 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)
-