Statistical Tests to Analyze the Categorical Data

- Types of Categorical Data:
  - Nominal categories
  - Ordinal categories
- Types of analysis for categorical data:
  - Descriptive: rate and ratio
  - Analytic: confidence interval and test of significance

THE CHI-SQUARE TEST (χ^2)

BACKGROUND AND NEED OF THE TEST

- Data collected in the field of medicine is often qualitative.--- For example:
  - the presence or absence of a symptom
  - classification of pregnancy as 'high risk' or 'non-high risk'
  - the degree of severity of a disease (mild, moderate, severe)
- The measure computed in each instance is a proportion, corresponding to the mean in the case of quantitative data such as height, weight, BMI, serum cholesterol.
- Comparison between two or more proportions, and the test of significance employed for such purposes is called the "Chi-square test"
- KARL PEARSON IN 1889, DEVISED AN INDEX OF DISPERSION OR TEST CRITERIOR DENOTED AS "CHI-SQUARE ". (χ^2).

Introduction

- What is the χ2 test?
  - χ2 is a non-parametric test of statistical significance for bi variate tabular analysis.
  - Any appropriately performed test of statistical significance lets you know that degree of confidence you can have in accepting or rejecting a hypothesis.
  - The hypothesis tested with chi square is whether or not two different samples are different enough in some characteristics or aspects of their behavior that we can generalize from our samples that the populations from which our samples are drawn are also different in the behavior or characteristic.
  - The χ2 test is used to test a distribution observed in the field against another distribution determined by a null hypothesis.

- (If you find the part below difficult, go to the example to make things easy)
- Being a statistical test, $\chi 2$ can be expressed as a formula. When written in mathematical notation the formula looks like this:

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

  - The summation is over all cells of the contingency table consisting of r rows and c columns
  - *o* is the observerd frequency,
  - *e* is the expected frequency, $e = \frac{\text{total of row in which the cell lines} \times \text{total of row in which the cell lines}}{\text{total of cells}}$
  - The null hypothesis is rejected if the calculated $\chi^2 > \chi^2_{,a,df}$
  - *df*= degrees of freedom. *Df*= (# of rows – 1) × (# of columns -1)

    = (r-1) (c-1)

- When using the chi square test, the researcher needs a clear idea of what is being investigate.
- It is customary to define the object of the research by writing a hypothesis.
- Chi square is then used to either prove or disprove the hypothesis.
- The hypothesis is the most important part of a research project. It states exactly what the researcher is trying to establish. It must be written in a clear and concise way so that other people can easily understand the aims of the research project.
- 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)

Requirements

- 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 be exceed 20.
- The expected frequency under the H0 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

Example

- Objective: To determine if smoking is a risk factor for MI

  Null Hypothesis: smoking does not cause MI

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

- To calculate the expected value for **smokers** who **got MI**:

$$\text{Expected value (e)} = \frac{\text{total of row in which the cell lines}}{\text{total of cells}} \times \text{total of row in which the cell lines}$$

$$= \frac{50 * 45}{100}$$

$$= 22.5$$

If we complete the entire table in this manner, we get:

	Disease (MI)		No Disease(No MI)		Total
Smokers	o=29	$e=\frac{50*45}{100}$ =22.5	o=21	$e=\frac{50*55}{100}$ =27.5	50
Non-smokers	o=16	$e=\frac{50*45}{100}$ =22.5	o=34	$e=\frac{50*55}{100}$ =27.5	50
Total	45		55		100

(o is the observed frequency, e is the expected frequency)

- $\chi 2 = \sum \frac{(o-e)^2}{e} = \frac{(29-22.5)^2}{22.5} + \frac{(16-22.5)^2}{22.5} + \frac{(21-27.5)^2}{27.5} + \frac{(34-27.5)^2}{27.5}$

  = **6.84**

- After we calculated the chi square ($\chi 2$=**6.84**), we calculate whether this value accepts the null hypothesis (Ho) or reject it, so we do the followings:
  - Calculate the degrees of Freedom:

    df = (r-1) ×(c-1) r=rows, c= columns

    = (2-1) × (2-1) =**1**
  - Decide on the alpha (α) value (in this case we choose **0.05**).
  - Look in the chi square table for the degrees of freedom of the alpha value chosen:

    In this case, df=1, α=0.05, so the Critical (table) Value = 3.84

df	ά = 0.05	ά = 0.01	ά = 0.001
1	**3.84**	6.64	10.83
2	5.99	9.21	13.82
3	7.82	11.35	16.27
4	9.49	13.28	18.47
5	11.07	15.09	20.52
6	12.59	16.81	22.46
7	14.07	18.48	24.32
8	15.51	20.09	26.13
9	16.92	21.67	27.88
10	18.31	23.21	29.59

  - Calculated value(6.84) is greater than critical value (3.84) at 0.05 level with 1 df
  - Hence we reject our Ho (null hypothesis) and conclude that there is highly statistically significant association between smoking and MI.

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 o(e)	Non-responders o(e)	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

- X^2 = 0.439+ 2.571+ 0.003+ 0.020+ 0.013+ 0.075+ 0.022+ 0.125+ 0.219+ 1.231 = 4.718
- Degrees of Freedom

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

  = (5-1) (2-1) = 4
- Critical Value of X2 with 4 d.f.f at 0.05= 9.49 (we got this value from the table in the previous page)
- Calculated value (4.718) is less than critical (table) value (9.49) at 0.05 level with 4 d.f.f
- Hence we cannot reject our Ho and conclude that the distributions are similar, that is non responders do not differ from responders.

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.

McNemar's test

- What to do when we have paired samples and both the exposure and outcome variables are qualitative variables (Binary). They form 2 × 2 Table as in matched case control or cross-over trial

- 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. (every case has a correspondent control)

	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

	Control: exposed	**Control**: not exposed	Total
CASES: exposed	12	43	55
CASES: not exposed	7	121	128
Total	19	164	183

    - **formulas:**

	Control: exposed	**Control**: not exposed	Total
CASES: exposed	E	F	E+f
CASES: not exposed	G	H	G+h
Total	E+G	F+H	N

      - Odds raio = f/g = 43/7 = 6.1
      - $X^2 = \frac{(|f-g|-1)^2}{f+g} = \frac{(|43-7|-1)^2}{43+7} = 24.5$
    - Degrees of Freedom

      df = (r-1) (c-1) = (2-1) (2-1) =1
    - Critical Value (Table A.6) = 3.84
    - 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.

In Conclusion !

- When both the study variables and outcome variables are categorical (Qualitative): Apply
  - Chi square test
  - Fisher's exact test (Small samples)
  - Mac nemar's test (for paired samples)