

## UNIVARIATE ANALYSIS OF CATEGORICAL DATA

### 1. Categorical data analysis

- One proportion
  - Chi-square goodness of fit
- Two proportion (independent sample)
  - Pearson chi-square/ fisher Exact test
- Dependent sample (matched or paired)
  - Mc Nemar's test
- Stratified sampling to control cofounder effect
  - Mantel-Haenszel test

### 2. Two proportion (independent sample) – Pearson Chi-square & Fisher Exact test.

- To test the association between two categorical variable
- IHD vs. Gender
  - Does gender associated with IHD status?
- Result of test
  - Not significant → no association
  - Significant → an association
- Step 1: State the hypothesis
  - $H_0$ : There is no association between gender and IHD
  - $H_A$ : There is an association between gender and IHD
- Step 2: set the significance level
  - How much? – accept the error in estimating the proportion in the population
  - Usually:  $\alpha = 0.05$
- Step 3: check the assumption
  - Two variables are independent

- Two variables are categorical
- Expected count of less than 5 is > 20% (take fisher exact test) and if < 20% (take pearson chi-square test).
  - Expected count = [row total x column total]/grand total
- Step 4: statistical test
  - Chi-square test or
  - Fisher exact test
  - $X^2 = \sum (O - E)^2 / E$
  - Chi-square value:
    - When the difference between observed and expected increase  
 → Value of chi-square increase → p-value decrease → significant increase
- Step 5: Interpretation
  - p value = 0.123
    - do not reject  $H_0$
  - There is no significant association between gender and IHD status.
- Step 6: conclusion
  - There is no significance association between gender and IHD status using Pearson Chi-square tests (p-value = 0.123)
- Data presentation

Table 1: Association between IHD and gender

Variable	IHD		z stat	p-value
	Yes n (%)	No n (%)		
Gender				
Male	15 (60)	10 (40)	2.381	*0.123
Female	20 (80)	5 (20)		

\* Pearson Chi-square test

3. Two proportion (dependent sample) – Mc Nemar’s test

- Dependent sample (matched or pair sample)
- $X^2 = (|b+c|) / (b+c)$
- Discordant pair
  - Is pair of different outcome
  - Use to test the difference in the outcome
- Sample of 25 pair patient with breast cancer
  - Matched for age
  - Undergone
    - Simple Mastectomy (SM)
    - Radical Mastectomy (RM)
  - Difference of 5-year survival proportion between two group
- Step 1: state the null and alternative hypothesis
  - $H_0$ : there is no association between type between type of mastectomy and 5-year survival proportion in patients with breast cancer
  - $H_A$ : there is an association between type of mastectomy and 5-year survival proportion in patients with breast cancer.
- Step 2: set the significance level
  - $\alpha = 0.05$
- Step 3: check the assumption
  - Categorical data
  - Dependent or matched sample
- Step 4: statistical test
  - Mc Nemar’s test
- Step 5: interpretation
  - p-value = 0.021
    - reject  $H_0$
  - there is significant association between type of mastectomy and 5-year survival proportion in patients with breast cancer.

		SM	
		Live	Die
RM	Live	a**	*b
	Die	*c	d**

\* Discordant  
 \*\* Concordant

- Step 6: conclusion
  - There is significant association between type of mastectomy and 5-year survival proportion in patients with breast cancer using Mc Nemar's test (p-value = 0.021)
- Data presentation

Table 2: Association between type of mastectomy and 5-year survival proportion in patients with breast cancer

Variable	Simple mastectomy		p-value
	Live n (%)	Die n (%)	
Radical			
Live	13 (%)	1 (%)	*0.021
Die	9 (%)	2 (%)	

\* Mc Nemar's test