Practical Training on GraphPad Prism for Statistical Testing (part 2)

Jingwen Gu (jingwen.gu@nih.gov)
Biostatistician
Bioinformatics and Computational Biosciences Branch (BCBB)
Office of Cyber Infrastructure and Computational Biology (OCICB)

Type of Categorical Data

Nominal and ordinal are categorical data.

- Nominal: unorder categories
  - Examples: Gender, race, hair color
  - Measures: counts, frequency, mode
- Ordinal: order categories
  - Examples: Highest education degree, levels of satisfaction
  - Measures: counts, frequency, mode, median

Outline

Dependent variable is Categorical

- Chi-square test
- Fisher exact test
- Chi-square test for trend
- McNemar's test

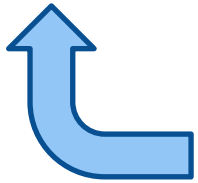
Test for data with categorical data

Patient ID	Smoking status	Lung cancer
1	Yes	Case
2	No	Control
3	Yes	Case
4	No	Control
5	Yes	Control
6	No	Control
7	No	Control
8	No	Case
9	Yes	Control
...	...	...

What is the appropriate test for seeing whether smoking is associated with lung cancer in this study?

Contingency table

	Lung Cancer		
Smoker	Case	Control	Total
Yes	688	650	1338
No	21	59	80
Total	709	709	1418



Is there a significant association between smoking and lung cancer?

Patient ID	Smoking status	Lung cancer
1	Yes	Case
2	No	Control
3	Yes	Case
4	No	Control
5	Yes	Control
6	No	Control
7	No	Control
8	No	Case
9	Yes	Control
...	...	...

Note

1. In Prism, the input data should be ”**contingency**” form. Prism could not cross tabulate data.

2. The categories defining the rows and columns must be mutually exclusive, with each subject (or experimental unit) contributing to one cell only.

3. Input Data format:

- Prospective and experimental studies: (row) top to bottom– risk factor/treatment and placebo; (column) left to right – disease and non disease.
- Case-control retrospective studies: (row) top to bottom- exposed to risk factor and not exposed; (column) left to right – cases and control.

Chi-square test

Chi-square test are used for testing independence by evaluating the closeness between observed and expected frequencies.

Assumption: **large samples and independence** of individual observation.

H_0: Two variables are independent. H_a: They are not independent.

$X^2 = \sum_i \sum_j \frac{(n_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}}$ where $\hat{\mu}_{ij} = n\pi_{i+}\pi_{+j}$, degree of freedom: $(I-1)(J-1)$

X^2 follow $\chi^2_{(I-1)(J-1)}$. The larger the values of X^2 is, the more evidence exists against independence. If p-value less than significance level, then reject null hypothesis.

Fisher Exact test

Fisher's exact test, as its name implies, always gives an exact p-value and works fine with **small sample sizes**.

Assumption: **Independence** of individual observation and **fixed totals**.

H_0: Two variables are independent. H_a: They are not independent.

The exact possibility assigned to each of the possible outcomes:

$$p = \frac{n_{1+}!\, n_{2+}!\, n_{+1}!\, n_{+2}!}{n!\, n_{11}!\, n_{12}!\, n_{21}!\, n_{22}!}$$

Calculate p-value as the total probability of observing data as extreme and more extreme cases. Reject null hypothesis if p-value less than significant level.

Alternative: Barnard test

Chi-square test OR Fisher's exact test?

- Why not always Fisher's exact test?

The problem is that the Fisher's test is based on assuming that the row and column totals are fixed by the experiment.

- In Prism, if you enter huge numbers (the sum is greater than 1,000,000), it will perform the chi-square test even if you chose Fisher's test.

Let’s practice in Prism

- Use sample data provided by Prism
- Perform Chi-square test and Fisher’s exact test to evaluate association.

Interpreting result:
Chi-square test and Fisher’s exact test

Chi-square test for Trend

The **chi-square test for trend** tests whether there is a linear trend between row number and the fraction of subjects in the left column. It is also called the **Cochran-Armitage** *method*.

It only makes sense when the rows are arranged in a **natural order** (such as by age, dose, or time), and are **equally spaced**.

Enter the risk factor/exposure as rows and outcome as column.

McNemar test

McNemar's test is used for comparing categorical responses for two samples that are **statistically dependent**.

Commonly occur in studies with **repeated measurement of subjects**.

$$X^2 = \frac{(|n_{21} - n_{12}| - 1)^2}{n_{21} + n_{12}}$$

For large samples, X^2 has a chi-squared distribution with $df = 1$, p-value less than significance level, reject the null hypothesis of independence.

National Institute of Allergy and Infectious Diseases

Read more

NIAID

Example – McNemar's test

For individuals observed as a pair, they are matched in age, gender and location, interested in whether poverty impact visiting the doctor.

H_0: no association between poverty and visit the doctor in past 12 month

H_a : there is association between poverty and visit the doctor in past 12 month

	Below poverty		
Above Poverty	Exposed	Unexposed	Total
Exposed	42	15	57
Unexposed	2	4	8
Total	44	19	63

$$X^2 = \frac{(|2 - 15| - 1)^2}{2 + 15} = 8.47$$

With 1 degree of freedom, p-value is 0.0036, reject the null hypothesis of no association at 0.05 significance level, indicating that there is association between poverty and visit the doctor in past 12 months.

Source

Let's practice in Prism

- Use sample data provided
- Perform Chi-square test with trend in Prism and McNemar's test in Prism QuickCals to evaluate association.

- Prism QuickCals: https://www.graphpad.com/quickcalcs/mcNemar1/