National Institute of Allergy and Infectious Diseases

Introduction to **Categorical Data Analysis**

RML Workshop, Sept 12, 2019

Contact our team: bioinformatics@niaid.nih.gov

NIAID

National Institute of Allergy and Infectious Diseases

Jingwen Gu (jingwen.gu@nih.gov)

Clinical Statistician

Bioinformatics and Computational Biosciences Branch (BCBB)/OCICB/NIAID

What is Categorical data?

- Categorical data arises when individuals are categorized into one of two or more mutually exclusive groups.
- Continuous data could be transformed to categorical data with respect to some predefined criteria.
- Categorical data on either a nominal or ordinal scale.
  - ✓ **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
- Understanding the types of data is important as they determine which method of data analysis to use and how to report the results.

Anthony & Raghavendra, (2011)

Outline

1. Contingency table
   - Positive/Negative Predictive Value, Sensitivity, Specificity, Type I/II error
   - Joint, Marginal, Conditional Probability

2. Strength of association
   - Odds ratio
   - Relative Risk

3. Test of independence
   - Cases with nominal large sample size and small sample size
   - Cases with stratified and paired data
   - Cases with ordinal data

Contingency table

Contingency tables summarize results where you compared two or more groups and the outcome is a categorical variable (such as disease vs. no disease, pass vs. fail, artery open vs. artery obstructed).

Cases: A data frame where each row represents one case. (e.g. patient-level data)

Count

Smoking	Lung Cancer	Count
Yes	Case	688
Yes	Control	650
No	Case	21
No	Control	59

Contingency table

	Lung Cancer	
Smoking	Case	Control
Yes	688	650
No	21	59

How to display data?

Most contingency tables have two rows (two groups) and two columns (two possible outcomes).

The top row usually represents exposure to a risk factor or treatment, and bottom row is mainly for control. The outcome is entered as column on the right side with the positive outcome as the first column and the negative outcome as the second column.

A particular subject or patient can be only in one column but not in both.

Dakhale, G. N., Hiware, S. K., Shinde, A. T., & Mahatme, M. S. (2012)

Contingency table application

Contingency table displays data from these five kinds of studies:

- Cross-sectional study
- Prospective study
- Retrospective study
- Experiment
- Assess the accuracy of a diagnostic test

GraphPad Statistics Guide.

PPV and NPV

Positive predictive value (PPV) of a test is the probability of an individual with a positive test has the disease. $PPV = Pr(Disease + | Test+)$.

Negative predictive value (NPV) of a test is the probability of an individual with negative test does not have the disease. $NPV = Pr(Disease - | Test -)$.

	Diseased	Not Diseased	Total
Test Positive	100	900	1000
Test Negative	50	5000	5050
Total	150	5900	6050

Here, $PPV = \frac{100}{1000} = 0.1$ and $NPV = \frac{5000}{5050} = 0.99$

Sensitivity and Specificity

Sensitivity The probability of the test finding disease among those who have the disease or the proportion of people with disease that are positive with the test.

Specificity The probability of the test finding no disease among those who do not have the disease or the proportion of people free of a disease who have a negative test.

Type I error is the rejection of a true null hypothesis of no disease, also known as false positive.

Type II error is fail to reject a false null hypothesis of no disease, also known as false negative.

- $True\ positive\ (TP) = \Pr(\ Test + \mid Disease +) = Sensitivity$
- $True\ negative\ (TN) = \Pr(\ Test - \mid Disease -) = Specificity$
- $False\ positive\ (FP) = \Pr(\ Test + \mid Disease -) = Type\ I\ error$
- $False\ negative\ (FN) = \Pr(\ Test - \mid Disease +) = Type\ II\ error$

Example – Screening test

Screening Test Result	Diseased	Not Diseased	Total
Positive	10	400	410
Negative	1	4500	4501
Total	11	4900	4911

The sensitivity of a screening test $= \mathrm{TP} = Pr(Test+ \mid Disease+) = 10/11 = 0.910$).

The specificity of a screening test: $= \mathrm{TN} = Pr(Test- \mid Disease-) = 4500/4900 = 0.918$).

$$\mathrm{FP} = Pr(Test+ \mid Disease-) = 400/4900 = 0.082$$

$$\mathrm{FN} = Pr(Test- \mid Disease+) = 1/11 = 0.091$$

A good screening exam has both high sensitivity and specificity.

Partial table

In a three-way contingency table cross-classifies X, Y, and Z, we control for Z by studying the XY relationship at fixed levels of Z.

- Partial table splits the original three-way table according to levels of Z.
- The associations in partial tables are called conditional associations. It refers to the association between X and Y conditional on fixing Z at some level.
- The two-way contingency table obtained by combining the partial tables is called the XY marginal table.

Gender	Smoke	Case	Control
Male	Yes		
	No		
Female	Yes		
	No		
Total	Yes		
	No		

Probability distribution

Joint distribution

- Let π_{ij} denote the probability that (X, Y) occurs in the cell in row i and column j. π_{ij} is the joint distribution of X and Y.

Marginal distribution

- The marginal distribution that $X = i$ or $Y = j$, $\pi_{i+} = \sum_j \pi_{ij}$ and $\pi_{+j} = \sum_i \pi_{ij}$.
- Sum of the marginal distribution is 1.

Conditional distribution

- Given that a subject is classified in row i of X, we use $\pi_{j|i}$ to denote the probability of classification in column j of Y, j = 1 , . . . , J. Then, $\sum_j \pi_{j|i} = 1$.

Example - Probability distribution

A new drug is being tested on a group of 800 people (400 men and 400 women) with a particular disease. We wish to establish whether there is a link between taking the drug and recovery from the disease.

Drug trial results:

	Drug taken	
Recovered	Yes	No
Yes	200	160
No	200	240
Recovery rate	50%	40%

We can conclude that the drug has positive effect. But if the result break down by gender…

Cont

Gender	Male		Female	
Drug taken	Yes	No	Yes	No
Recovered				
Yes	180	70	20	90
No	120	30	80	210
Recovery rate	60%	70%	20%	30%

Both for male and female, the recovery rates are better without drug. Gender influences drug taken because men are much more likely in this study to have been given the drug than women.

The result that a marginal association can have a different direction from conditional association is called Simpson's paradox.

Avoid? Yes if we are certain that we know every possible variable that can impact the outcome variable. If we are not certain – and in general we simply cannot be – then Simpson's paradox is theoretically unavoidable.

Confounding

Confounding variable is a variable that influences both the dependent and independent variable causing a spurious association.

To reduce effects of confounding variable:

- In experimental studies: randomly assigning subjects to different levels of confounding variable.
- In observational studies: control confounding variable that can influence relationship.

Strength of association

To measure the strength of association, use other methods like:

- Odds ratio
- Relative risk

Above methods also measures of risk, they can be useful in safety and efficacy studies.

Measures are effective when confounding variables are controlled.

Odds ratio (OR)

The odds ratio (OR) is the ratio of the odds of an event occurring in one group to the odds of it occurring in another group.

	Outcome	
Exposure	Yes	No
Yes	n_{11}	n_{12}
No	n_{21}	n_{22}

The odds of outcome when exposure presents is:

$$\frac{n_{11}/(n_{11}+n_{12})}{n_{12}/(n_{11}+n_{12})} = \frac{n_{11}}{n_{12}}$$

Similarly, the odds of outcome when exposure absent is: $\frac{n_{21}}{n_{22}}$

Odds ratio: $\hat{\theta} = \frac{n_{11}n_{22}}{n_{12}n_{21}}$

The asymptotic standard error of the $\log\hat{\theta}$: $\sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}$

The odds ratio is a relative measure of risk, telling us how much more likely it is that someone who is exposed to the factor under study will develop the outcome as compared to someone who is not exposed.

Odds ratio

OR=1 means there is an equal chance or likelihood of getting the disease among exposed group compared to unexposed group.

OR>1 means there is more chance of likelihood of getting the disease among exposed group compared to unexposed group.

OR<1 means there is a less chance or likelihood of getting the disease among exposed group compared to unexposed group.

Anthony & Raghavendra, (2011)

Confidence intervals

The general form of confidence interval is:

$$estimate \pm critical\ value \times SE(estimate)$$

The confidence level of a confidence interval is the probability that the true parameter is between this interval.

Usually use 95% confidence interval, $\alpha = 0.05$. In this lecture, critical value is usually $z_{\alpha/2}$. For a 95% confidence interval, $z_{\alpha/2}$ = 1.96.

Sometimes a confidence interval for θ can be obtained indirectly, we first calculate (L, U), which is a confidence interval for $\log(\theta)$, and then a confidence interval for θ is obtained as (e^L, e^U).

Example – Odds ratio

In a study, patients admitted with lung cancer in the preceding year were queried about their smoking behavior.

For each of the 709 patients admitted, they recorded the smoking behavior of a noncancer patient at the same hospital of the same gender and within the same 5-year grouping on age.

The 709 cases in the first column of the table below are those having lung cancer and the 709 controls in the second column are those not having it. A smoker was defined as a person who had smoked at least one cigarette a day for at least a year.

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

Source: data reported in Table IV, R. Doll and A.
Source: data B. Hill, *Br. Med. J.,* 739-748, Sept. 30, 1950.

Evaluate the strength of association by calculating odds ratio and 95% confidence interval. Is the odds ratio significantly different from 1?

Cont

From example, $Odds\ ratio = \frac{688 \times 59}{650 \times 21} = 2.97 \approx 3$

The odds of patient to have lung cancer are three times that if smoke compare to if did not smoke.

The asymptotic standard error of the log odds is:

$$ASE(log\hat{\theta}) = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}} = 0.26$$

95% CI for $\log\theta$ is $\log\hat{\theta} \pm 1.96\ ASE(log\hat{\theta})$=(0.58,1.60)

95% CI for θ is $(e^{\log\hat{\theta} - 1.96\ ASE(log\hat{\theta})}, e^{\log\hat{\theta} + 1.96\ ASE(log\hat{\theta})}) = (1.79, 4.95)$

The 95% confidence interval is $(1.79, 4.95)$. Since it does not include one, odds ratio is significantly different from 1. The odds of the smoking group to have lung cancer is between 1.79 and 4.95 times compared to the non smoking group.

Relative risk

Relative risk is a measure of the risk of a certain event happening in one group compared to the risk of the same event happening in another group.

Relative risk equals to:

$$r = \frac{\hat{\pi}_1}{\hat{\pi}_2} = \frac{n_{11}\,(n_{21} + n_{22})}{n_{21}(n_{11} + n_{12})}$$

	Outcome	
Exposure	Yes	No
Yes	n_{11}	n_{12}
No	n_{21}	n_{22}

An estimated standard error for log r:

$$\hat{\sigma}(\log r) = \sqrt{\frac{1}{n_{11}} - \frac{1}{n_{11} + n_{12}} + \frac{1}{n_{21}} - \frac{1}{n_{21} + n_{22}}}$$

In clinical trails, it is used to compare the risk of developing a disease in people not receiving the treatment (or receiving a placebo) versus people who are receiving the treatment.

Relative risk

RR=1 means there is no difference in the risk of getting disease between two groups.

RR>1 means there is more risk of getting disease among exposed group compared to unexposed group.

RR<1 means there is a less risk of getting disease among exposed group compared to unexposed group.

Example– Aspirin and Heart Attacks

Below table is from a report on the relationship between aspirin use and myocardial infraction (heart attacks) by the Physicians' Health Study Research Group at Harvard Medical School. The Physicians' Health Study was a five-year randomized study testing whether regular intake of aspirin reduces mortality from cardiovascular disease. Every other day, the male physicians participating in the study took either one aspirin tablet or a placebo. The study was "blind" – the physicians in the study did not know which type of pill they were taking.

	Myocardial Infarction		
Group	Yes	No	Total
Aspirin	104	10933	11037
Placebo	189	10845	11034

Source: Preliminary Report: Findings from the Aspirin Component of the Ongoing Physicians' Health Study. *New Engl. J. Med.*, **318**: 262–264, 1988.

Example – Aspirin and Heart Attacks

In this example, the proportion having MI were $\hat{\pi}_1 = \frac{104}{11037} = 0.0094$ for Aspirin and $\hat{\pi}_2 = \frac{189}{11034} = 0.0171$ for placebo.

The sample relative risk is $\frac{\hat{\pi}_1}{\hat{\pi}_2} = \frac{0.094}{0.171} = 0.55$

Therefore, the participant who taking aspirin are 0.55 times the risk of myocardial infarction compared to participant taking the placebo. OR participant taking aspirin had 45% less risk of having a myocardial infarction compared to participant taking the placebo.

$$\hat{\sigma}(\log r) = \sqrt{\frac{1}{n_{11}} - \frac{1}{n_{11}+n_{12}} + \frac{1}{n_{21}} - \frac{1}{n_{21}+n_{22}}} = \sqrt{\frac{1}{104} - \frac{1}{11037} + \frac{1}{189} - \frac{1}{11034}} = 0.121$$

The 95% confidence interval of $\log r$ is

$$\log\left(\frac{\hat{\pi}_1}{\hat{\pi}_2}\right) \pm 1.96\,\hat{\sigma}(\log r) = log0.55 \pm 1.96 \times 0.121 = (-0.84, -0.36)$$

The 95% confidence interval of r is $e^{\log\left(\frac{\hat{\pi}_1}{\hat{\pi}_2}\right) \pm 1.96\,\hat{\sigma}(\log r)} = (0.43, 0.70)$.

Similarity and difference between OR and RR

- The relationship between odds ratio and relative risk:

$$Odds\ ratio = Relative\ risk \cdot \frac{1-\pi_2}{1-\pi_1}$$

when π_1 and π_2 small, odds ratio is approximately equals to relative risk.

- OR can be used to describe the results of case-control study as well as prospective cohort studies.

- RR is based upon the incidence of an event given that we already know the study participants' exposure status. In cancer research, relative risk is used in prospective (forward looking) studies, such as cohort studies and clinical trials.

- Dealing with small probability, RR is better in interpretation.

Test of Independence

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

Large sample size

	Age < 50		
	CVD	Non-CVD	Total
Obese	10	90	100
Not Obese	35	465	500
Total	45	555	600

Stratified data

	Age ≥ 50		
	CVD	Non-CVD	Total
Obese	36	164	200
Not Obese	25	175	200
Total	61	339	400

```
> AOH
    y alc     obe hyp
1   5   0     low yes
2   9 1-2     low yes
3   8 3-5     low yes
4  10  6+     low yes
5  40   0     low  no
6  36 1-2     low  no
7  33 3-5     low  no
8  24  6+     low  no
9   6   0 average yes
10  9 1-2 average yes
11 11 3-5 average yes
12 14  6+ average yes
13 33   0 average  no
14 23 1-2 average  no
15 35 3-5 average  no
16 30  6+ average  no
17  9   0    high yes
18 12 1-2    high yes
19 19 3-5    high yes
20 19  6+    high yes
21 24   0    high  no
22 25 1-2    high  no
23 28 3-5    high  no
24 29  6+    high  no
```

Ordinal data

	Guess Poured First		
Poured First	Milk	Tea	Total
Milk	3	1	4
Tea	1	3	4
Total	4	4	

Source: Based on experiment described by Fisher (1935a).

Small sample size

	2008 Election		
2004 Election	Democrat	Republican	Total
Democrat	175	16	191
Republican	54	188	242
Total	229	204	433

Paired data

Response	Dosage				
	0	1	2	3	Total
1	10	9	10	7	36
0	0	1	0	3	4
Total	10	10	10	10	40

Two categories and ordinal variables

Test of Independence

Pearson's chi-square test and likelihood ratio test are used for testing independence by evaluating the closeness between observed and expected frequencies.

- Assumption: Independent random sampling, no more than 20% of the cells has an expected frequency less than five, and no empty cells.
- H_0: Two variables are independent. H_a: They are not independent.
- $\hat{\mu}_{ij}$ is the expected frequencies, n_{ij} is the observed.

Pearson chi-square statistic: $X^2 = \sum_i \sum_j \frac{(n_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}}$ where $\hat{\mu}_{ij} = n\pi_{i+}\pi_{+j}$

Likelihood ratio statistic: $G^2 = -2log\Lambda = 2\sum_i \sum_j n_{ij} \log(n_{ij}/\hat{\mu}_{ij})$

Degree of freedom: $(I-1)(J-1)$

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

Example – Chi-square test

In the example of case-control study of lung cancer and smoking:

Is there a significant association between smoking and lung cancer?

H_0: Smoking and lung cancer are independent. H_a: not independent
Assume significance level=0.05. $\hat{\mu}_{ij} = n\pi_{i+}\pi_{+j}$

Observed table

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

Expected table

	Lung Cancer	
Smoker	Case	Control
Yes	669	669
No	40	40
Total	709	709

Cont – Pearson Chi-square test

The Pearson statistic equals to

$$X^2 = \frac{(688-669)^2}{669} + \frac{(650-669)^2}{669} + \frac{(21-40)^2}{40} + \frac{(59-40)^2}{40}$$
$$= 19.129$$

Degree of freedom is $(2-1)(2-1) = 1$

P-value is 0.0000126, less than 0.05, reject null hypothesis. *(calculate using* $pchisq(19.129,1,lower.tail = FALSE)$ *in R)*

We have 95% confidence to reject the null hypothesis that smoking and lung cancer are independent.

Cont – Likelihood ratio test

The likelihood ratio statistic equals to,

$$G^2 = 2\left[688\log\left(\frac{688}{669}\right) + 650\log\left(\frac{650}{669}\right) + 21\log\left(\frac{21}{40}\right) + 59\log\left(\frac{59}{40}\right)\right]$$
$$= 19.878$$

Degree of freedom is $(2-1)(2-1) = 1$

P-value is 0.00000825, less than 0.05, reject null hypothesis.

We have 95% confidence to reject the null hypothesis that smoking and lung cancer are independent.

Properties

X^2 and G^2 have the same limiting chi-squared distribution.

$X^2 - G^2$ converges in probability to zero (asymptotically equivalent).

The order of the row or column vector does not change for the result of a chi-square or likelihood ratio test of independence.

Small sample size

Fisher's exact test can be used for test of independence when n is small.

- Assumption: Independence of individual observation and fixed totals. (the row and column totals are fixed, or "conditioned.") When row or column totals are unconditioned, makes this test less powerful.

The probability mass function is:

$$p(t) = P(n_{11} = t) = \frac{\binom{n_{1+}}{t}\binom{n_{2+}}{n_{+1}-t}}{\binom{n}{n_{+1}}}$$

	Outcome	
Exposure	Yes	No
Yes	n_{11}	n_{12}
No	n_{21}	n_{22}

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.

Example – Fisher's test for small size data

Example: Lady Tasting Tea

R. A. Fisher described the following experiment from his days working at Rothamsted Experimental Station. His colleague, Dr. Muriel Bristol declares that by tasting a cup of tea made with milk she can discriminate whether the milk or the tea infusion was first added to the cup.

Experiment design: consist of eight cups of tea, four pouring milk first and four pouring tea first; serve in a random order. She knew there were four cups of each type and had to predict which four had the milk added first.

	Guess Poured First		
Poured First	Milk	Tea	Total
Milk	3	1	4
Tea	1	3	4
Total	4	4	

Source: Based on experiment described by Fisher (1935a).

Milk poured first (4 cups) Tea poured first (4 cups)

NIAID

Cont – Fisher's test for small size data

Distinguishing the order of pouring better than with pure guessing corresponds to $\theta > 1$, reflecting a positive association between order of pouring and the prediction.

H_0: $\theta = 1$ against H_a: $\theta > 1$

All Incorrect

0	4
4	0

1	3
3	1

Even

2	2
2	2

3	1
1	3

All Correct

4	0
0	4

The probability is $\mathrm{p}(n_{11} = 3) = \frac{\binom{4}{3}\binom{4}{1}}{\binom{8}{4}} = \frac{4!4!4!4!}{8!3!1!3!1!} = 0.229$. The more extreme in the direction of H_a has $n_{11} = 4$ correct. $\mathrm{p}(n_{11} = 4) = \frac{\binom{4}{4}\binom{4}{0}}{\binom{8}{4}} = \frac{4!4!4!4!}{8!4!0!4!0!} = 0.014$

The P-value is $\mathrm{p}(n_{11} \geq 3) = 0.243$.

Cannot reject H_0 at significant level 0.05, which means that the result does not establish an association between actual order of pouring and her predictions.

Not independent (stratified) data?

For example, suppose that the rows of the table represent two different kinds of preoperative antibiotics and the columns denote whether or not there was a postoperative infection.

There are 100 subjects. These subjects are not independent if the table combines results from 50 subjects in one hospital with 50 subjects from another hospital. Any difference between hospitals, or the patient groups they serve, would affect half the subjects but not the other half. You do not have 100 independent observations.

To analyze this kind of data, use the Cochran-Mantel-Haenszel test or logistic regression.

Anthony & Raghavendra, (2011)

Stratified data

Cochran-Mantel-Haenszel test for the analysis of stratified or matched categorical data. It is used to test the conditional independence in $2\times2\times K$ tables and can be generalized to $\mathrm{I}\times\mathrm{J}\times K$ table.

- Often used in observational studies where random assignment of subjects to different treatments cannot be controlled, but confounding covariates can be measured.
- H_0: There is no association between the two inner variables.

According to stratification, create 2×2 contingency tables. Assume there are k tables.

The test statistic can be calculated by:

$$X^2_{CMH} = \frac{[\left|\sum_{i=1}^{k}(A_i - \frac{N_{1i}M_{1i}}{T_i})\right| - 0.5]^2}{\sum_{i=1}^{k}\frac{N_{1i}N_{2i}M_{1i}M_{2i}}{T_i^2(T_i-1)}}$$

	Y1	Y2	Total
X1	A_i	B_i	N_{1i}
X2	C_i	D_i	N_{2i}
Total	M_{1i}	M_{2i}	T_i

Follow χ^2 distribution asymptotically with one degree of freedom under H_0.

Example – CHM test for stratified data

To examine the association between obesity and cardiovascular diseases (CVD), data is stratified into two categories with age< 50 and age≥50:

	Age < 50		
	CVD	Non-CVD	Total
Obese	10	90	100
Not Obese	35	465	500
Total	45	555	600

	Age ≥ 50		
	CVD	Non-CVD	Total
Obese	36	164	200
Not Obese	25	175	200
Total	61	339	400

H_0:There is no association between obesity and CVD.

$$\chi^2_{CMH}=\frac{[|10-\frac{100*45}{600}+36-\frac{200*61}{400}|-0.5]^2}{\frac{45*555*100*500}{600^2\,(600-1)}+\frac{61*339*200*200}{400^2\,(400-1)}}=3.0$$

p-value=0.08, fail to reject null hypothesis that there is no association between obesity and CVD.

Paired data?

In some experiments, subjects are matched for age and other variables. One subject in each pair receives one treatment while the other subject gets the other treatment. These data should be analyzed by special methods such as McNemar's test. Paired data should not be analyzed by chi-square or Fisher's test.

Anthony & Raghavendra, (2011)

Paired data

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 (e.g. test the improvement in response rate after a particular treatment).

Assumption: sample randomly drawn from the population and no expected frequencies should be less than five.

McNemar's statistic with continuity correction:

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

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.

Example – McNemar's test for matched pair data

In the 2010 General Social Survey, subjects were asked who they voted for democrat or republican in the 2004 and 2008 Presidential elections. Was there a shift in this direction?

	2008 Election		
2004 Election	Democrat	Republican	Total
Democrat	175	16	191
Republican	54	188	242
Total	229	204	433

The McNemar statistic is: $\frac{(|54-16|-1)^2}{54+16} = 19.557$ with $df = 1$, p-value $9.764e-06$, extremely strong evidence of a shift in the Democrat direction.

Trend data?

Cochran-Armitage trend test is a frequently used test in association between a variable with two categories and an ordinal variable with k categories (e.g. dose-response studies).

Response	Dosage			
	1	2	k	K
1	n_{11}	n_{12}	n_{1k}	n_{1K}
0	n_{01}	n_{02}	n_{0k}	n_{0K}

H_0: There is no linear trend in binomial proportions of response across increasing levels of dosage.

H_1: There is a linear (increasing/decreasing) trend in binomial proportions of response across increasing levels of dosage.

Anthony & Raghavendra, (2011)

Cochran-Armitage trend test

It modifies the Pearson Chi-square test to incorporate a suspected ordering in the effects of the *k* categories of the second variable.

$p_{1+} = \frac{n_{1+}}{n_{++}}$, and $\bar{s} = \frac{\sum n_{+i} s_i}{n}$, $b = \frac{\sum n_{+i}(p_{1|i} - p_{1+})(s_i - \bar{s})}{\sum (s_i - \bar{s})^2}$

$$z^2 = \frac{b^2}{p_{1+} p_{0+}} \sum n_{+i} (s_i - \bar{s})^2$$

where $p_{0+} = 1 - p_{1+}$, z has an asymptotic chi-squared distribution with df=1.

OR

$$z = \frac{\sum n_{+i}(p_{1|i} - p_{1+})(s_i - \bar{s})}{\sqrt{p_{1+} p_{0+} \sum n_{+i}(s_i - \bar{s})^2}}$$

Has an asymptotic normal distribution. Results of the Cochran-Armitage trend test are similar to those obtained by testing that the slop is zero in a linear logistic model.

Liu, H.

Example – Cochran-Armitage trend test

Suppose a study is conducted to test the effect of a drug on 40 subjects. The subjects are randomized into four balanced groups receiving 0 mg, 1 mg, 2 mg and 3 mg of the drug, respectively. The results for each of the two responses are recorded for each subject, and the raw data can be categorized into:

Response	**Dosage**				
	0	1	2	3	Total
1	10	9	10	7	36
0	0	1	0	3	4
Total	10	10	10	10	40

$z = -1.8856$ and $p - value = 0.02967$. Reject the null hypothesis of no linear trend in binomial proportions of response across increasing levels of dosage.

Ordinal data

The chi-square tests ignore some information when used to test independence between ordinal classifications. Taking the ordering into account are usually more powerful.

In ordinal data analysis, we can assign scores to the levels for ordinal variables by using:

- Average of category interval
- Midrank

Then Linear trend test statistic can be used:

$$M^2 = (n-1)r^2$$

where r is Pearson correlation between two variables.

For large samples, it is approximately chi-squared with df = 1.

Example – Linear trend test for ordinal data

```
> AOH
    y alc      obe hyp
1   5   0      low yes
2   9 1-2      low yes
3   8 3-5      low yes
4  10  6+      low yes
5  40   0      low  no
6  36 1-2      low  no
7  33 3-5      low  no
8  24  6+      low  no
9   6   0  average yes
10  9 1-2  average yes
11 11 3-5  average yes
12 14  6+  average yes
13 33   0  average  no
14 23 1-2  average  no
15 35 3-5  average  no
16 30  6+  average  no
17  9   0     high yes
18 12 1-2     high yes
19 19 3-5     high yes
20 19  6+     high yes
21 24   0     high  no
22 25 1-2     high  no
23 28 3-5     high  no
24 29  6+     high  no
```

491 subjects are cross-classified according to the three factors: hypertension (hyp; 2 levels), obesity (obe; 3 levels) and alcohol (alc; 4 levels).

- Alc: the classification of alcohol intake of drinks per day (0, 1-2, 3-5, 6+)
- Obe: the classification of obesity (low, average, high)
- Hyp: the classification of hypertension (yes, no)

Objective: whether correlation between two ordinal variables.

H_0: $\rho = 0$ versus H_A: $\rho \neq 0$

- Use linear trend test $M^2 = (n-1)r^2$ to test for independence between two variables.
- Check p-value of statistic with degree of freedom 1.

Source: Knuiman, M.W. & Speed, T.P. (1988)

NIAID

Cont

```
> AOH_1
    y alc      obe hyp
1   5 0.0      low yes
2   9 1.5      low yes
3   8 4.0      low yes
4  10 7.0      low yes
5  40 0.0      low  no
6  36 1.5      low  no
7  33 4.0      low  no
8  24 7.0      low  no
9   6 0.0  average yes
10  9 1.5  average yes
11 11 4.0  average yes
12 14 7.0  average yes
13 33 0.0  average  no
14 23 1.5  average  no
15 35 4.0  average  no
16 30 7.0  average  no
17  9 0.0     high yes
18 12 1.5     high yes
19 19 4.0     high yes
20 19 7.0     high yes
21 24 0.0     high  no
22 25 1.5     high  no
23 28 4.0     high  no
24 29 7.0     high  no
```

Assign scores to the level of ordinal variables

- Average of the category interval
  - Obesity: low, median, high assign to 1, 2, 3
  - High BP: no, yes assign to 0, 1
  - Alcohol: 0, 1-2, 3-5, 6+ assign to 0, 1.5, 4, 7

  (Left graphic shown the recoding data)

- Midrank

  Rank the observations and applies midrank as scores.
  - Obesity: 83 for low (average among 1-165), 246 for median (from 166-326), 409 for high (from 327-491)

Create contingency table

	0	1.5	4	7	Sum
1	45	45	41	34	165
2	39	32	46	44	161
3	33	37	47	48	165
Sum	117	114	134	126	491

(rows for obesity and columns for alcohol)

Pearson correlation between obesity and alcohol is $r^2 = 0.103$, $M^2 = 5.235$, $p - value = 0.0221$

Compare with significant level and make conclusion.

We have 95% confidence to conclude that there is association between obesity and alcohol.

Compare result with other statistic

	X^2	G^2	M^2
Obesity vs Alcohol	0.325	0.317	0.022
High BP vs Alcohol	0.026	0.022	0.003
Obesity vs High BP	0.003	0.003	0.001

Sensitivity to choice of scores

- Scores that are linear transforms of each other, such as (1, 2, 3, 4) and (0, 2, 4, 6), have the same absolute correlation and hence the same M^2.
- Results may depend on the scores when the data are highly unbalanced.

Summary of Statistical Testing Applied to Specific Cases

Large and independent data → Pearson Chi-square test/
Likelihood ratio test

Small and conditioned data→ Fisher's exact test

Stratified data → Cochran-Mantel-Haenszel

One variable has two levels and the other variable ordinal → Cochran Armitage trend test

Paired data → McNemar's test / CMH test

Ordinal data → Linear trend test

** There are much more than listed statistical testing can be applied to categorical data, compare and select one most fit to your cases! Check assumption before applying it.

Survey

Thank you for attending the class!

Your feedback is important to us. Please complete survey so we can improve next time.

References

Agresti, A. (2018). *An introduction to categorical data analysis*. Wiley.

Anthony, G.M & Raghavendra, D. (2011) *Applied Clinical Trials.* Retrieved from http://www.appliedclinicaltrialsonline.com/categorical-data-analysis

Dakhale, G. N., Hiware, S. K., Shinde, A. T., & Mahatme, M. S. (2012). Basic biostatistics for post-graduate students. *Indian journal of pharmacology*, *44*(4), 435–442. doi:10.4103/0253-7613.99297

Fenton, N., Neil, Martin., Constantinou, A. (2015) *Simpson's Paradox and the implications for medical trials.* Retrieve from https://www.eecs.qmul.ac.uk/~norman/papers/simpson.pdf

GraphPad Statistics Guide. *Key concepts: Contingency table.* Retrieved from https://www.graphpad.com/guides/prism/7/statistics/index.htm?stat_interpreting_results_contingen_2.htm

Knuiman, M.W. & Speed, T.P. (1988) *Incorporating Prior Information into the Analysis of Contingency Tables. Biometrics*, **44 (4)**, 1061–1071.

Liu, H. *Cochran-Armitage Trend Test Using SAS* Retrieve from https://www.lexjansen.com/pharmasug/2007/sp/SP05.pdf

NCI Dictionary of Cancer Term. *Relative risk.* Retrieve from https://www.cancer.gov/publications/dictionaries/cancer-terms/def/relative-risk

Sullivan, L. M. (2011). *Essentials of biostatistics in public health*. Jones & Bartlett Publishers.

Wikipedia *Cochran-Mantel-Haenszel test* https://en.wikipedia.org/wiki/Cochran–Mantel–Haenszel_statistics