Statistical Computation
Math 475

Jimin Ding

Department of Mathematics
Washington University in St. Louis
www.math.wustl.edu/ jmding/math475/index.html

September 17, 2013

Part II

Contingency Tables for Categorical Data

Outline I

Outline II

Statistical Modeling

	Explanatory Variable		
Response	Continuous	Categorical	Mixed
Continuous	Regression	ANOVA	ANCOVA
Categorical	Categorical Data Analysis		

Categorical Data Analysis

- Contingency tables: summarize categorical data in a tabular form.
- Transformation model (example: log linear models)
- Generalized linear models (GLM) (example: logistic regression)

Categorical Data

Introduction
Models
Data
Distributions
Two-way Table
Other Concepts for 2×2 Tables
Ordered data
Stratified Contingency Tables
Matched Pairs
SAS Programs

Type of data:

- Dichotomous: normal/abnormal blood pressure
- Ordinal: age range (Young, Middle, Old)
- Nominal: color (Red, Green, Black)
- Discrete counts: the number of heart attacks
- Grouped data: the number of deaths in 10-years intervals

Sampling framework: $\rightarrow$ model and inference

- Historical data
- Sample survey
- Experimental data

Distributions for Categorical Data

Introduction
Models
Data
Distributions
Two-way Table
Other Concepts for 2×2 Tables
Ordered data
Stratified Contingency Tables
Matched Pairs
SAS Programs

- Binomial distribution:
  The total number of trials, N, is fixed and known. Failure happens independently in each identical trial with a certain probability π. Then the number of failure(success), Y, is a random variable with Binomial distribution, $Bin(N, \pi)$.
  $$P(Y = k) = \frac{N!}{k!(N-k)!}\pi^k(1-\pi)^{N-k}, \quad k = 0, 1, 2, \cdots, N.$$
  $$E(Y) = N\pi, \quad Var(Y) = N\pi(1-\pi).$$
- Multinomial distribution:
  Eextension of binomial distribution to multiple possible outcomes. For example, in an auto-fatality study, "uninjured", "injury not requiring hospitalization", "injury requiring hospitalization".

Distribution for Categorical Data

Introduction
Models
Data
Distributions
Two-way Table
Other Concepts for 2×2 Tables
Ordered data
Stratified Contingency Tables
Matched Pairs
SAS Programs

- Poisson distribution:
  The total number of trials is unknown. The number of some possible outcome, Y, is a random variable with Poisson distribution $Pois(\pi)$. For example, the number of fatal accidents in a given week in St.Louis. (Note, the total number of accidents is unknown.)
  $$P(Y = k) = \frac{e^{\pi}\pi^k}{k!}, \quad k = 0, 1, 2, \cdots.$$
  $$E(Y) = Var(Y) = \pi.$$

Distribution for Categorical Data

- Product multinomial distribution:
  The number of trials for each given covariate (the row total) is fixed as well as the total number of trials.
- Hypergeometric distribution:
  The number of trials for each given covariate (the row total) is fixed and the number of each possible outcome from all covariates (the column total) is fixed.
  Then the number of failure in a sequence of n draws from a population with size N which contains m defective items (without replacement), X, is a random variable with Hypergeometric distribution, $HyperGeom(N, n, m)$.

$$P(X = k) = \frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}}, \quad , k = 0, 1, \cdots, \min(n, m)$$

Two-way $I \times J$ Contingency Table

Definition of Contingency Tables

Categorical data consist of frequency counts of observation occurring in each of response category. Let X and Y denote two categorical variables, with I and J levels respectively. We display the frequency counts of IJ possible combinations of outcomes in a table with I rows and J columns. This type of table is called two-way contingency table. A contingency table that cross classifies three variables is called a three-way table, and so fourth.

An Example of Contingency Table

- Example: American's belief in afterlife.
  (Source: edited from table 2.1 in Agresti (2002))

	Yes	No	Total
Females	435	147	582
Males	375	134	509
Total	810	281	1091

- Key interest: Do females more believe in afterlife? Is there any association between gender and opinion about afterlife?
- If the actual probability of believing in an afterlife is same for both male and female, then belief in an afterlife can be regarded as independent of gender.

2 × 2 Table

Denote n as frequency counts and π as proportion of each combination of outcomes. Classify the response variable in columns and explanatory variable in rows.

	Col1	Col2	Row Total
Row1	$n_{11}(\pi_{11})$	$n_{12}(\pi_{12})$	$n_{1+}(\pi_{1+})$
Row2	$n_{21}(\pi_{21})$	$n_{22}(\pi_{22})$	$n_{2+}(\pi_{2+})$
Col Total	$n_{+1}(\pi_{+1})$	$n_{+2}(\pi_{+2})$	$n_{++}(\pi_{++}=1)$

Let $\pi_{j|i} = \pi_{ij}/\pi_{i+}$, for $i = 1, 2; j = 1, 2$. For example, in the previous example,

- π_{21} is the probability of a random selected participant is a male and believes in afterlife.
- π_{2+} is the probability of a male participant is selected.
- $\pi_{j|i}, i = 2, j = 1$ is the probability that a male believes in aferlife.

Association in Two-way Contingency Tables

- The test of whether males and females believe in afterlife equivalently can be written as
$$H_0 : \pi_{1|1} = \pi_{1|2} \quad vs. H_a : \pi_{1|1} \neq \pi_{1|2},$$
which, in a 2x2 table, is same as
$$H_0 : \pi_{1|1} = \pi_{+1} \quad vs. H_a : \pi_{1|1} \neq \pi_{+1},$$
$$H_0 : \pi_{11} = \pi_{1+}\pi_{+1} \quad vs. H_a : \pi_{11} \neq \pi_{1+}\pi_{+1}.$$
- In general, such a "homogeneity test of rows" in a $I \times J$ table can be written as
$$H_0 : \pi_{j|1} = \pi_{j|2} = \cdots = \pi_{j|I} = \pi_{+j}, \quad j = 1, \cdots, J.$$
That is the probability of falling into any particular column response is the same in each row. This is equivalent to the "independence test",
$$H_0 : \pi_{ij} = \pi_{i+}\pi_{+j}, \text{ for } i = 1, \cdots, I, \quad j = 1, \cdots, J.$$

Association in Two-way Contingency Tables

- The "homogeneity test" is often used in the situation that the column variable, Y, can be viewed as a response variable and the row variable, X, can be viewed as an explanatory variable (covariate, predictor).
- The "independence test", which is that all joint probabilities equal to the product of their marginal probabilities, is more generally used when the column and row variables are at the same level.

Pearson Chi-Square Test

- For a sample of size n with cell counts $\{n_{ij}\}$, the expected frequencies are defined as $\mu_{ij} = \frac{n_{i+}n_{+j}}{n}$. Note under the null hypothesis that two variables are independent, μ_{ij} represents $E(n_{ij})$.
- The test statistic: Pearson
$$Q^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(n_{ij} - \mu_{ij})^2}{\mu_{ij}}$$
has approximately a $\chi^2_{(I-1)(J-1)}$ distribution for a large sample. Larger value is stronger against H_0.
- Large sample: $n_{ij} \geq 5$ and $\mu_{ij} \geq 1$. If the sample size of some cells are too small, one need to combine some cells to make Pearson χ^2 test appropriate.

Exercise

For 2×2 table, Pearson χ^2 can be simplied as:

$$
\begin{aligned}
Q^2 &= \sum_{i=1}^{2}\sum_{j=1}^{2}\frac{(n_{ij}-\mu_{ij})^2}{\mu_{ij}} \\
&= \frac{n(n_{11}n_{22}-n_{12}n_{21})^2}{n_{1+}n_{2+}n_{+1}n_{+2}}.
\end{aligned}
$$

Likelihood Ratio χ^2 Test

Introduction
Two-way Table
Definition
2×2 Table
Association Tests
Pearson Test
LRT
Association Measures
Fisher Test
Other Concepts for 2×2 Tables
Ordered data
Stratified Contingency Tables
Matched Pairs
SAS

- It is a variant of Pearson χ^2 test.
- Asymptotically it is equivalent to Pearson χ^2 test.
- The test statistic:

$$G^2 = 2\sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}\log(\frac{n_{ij}}{\mu_{ij}}),$$

  has approximately a χ^2 distribution with degree of freedom $(I-1)(J-1)$.
- Larger value provides stronger evidence against independence.
- It performs even worse than Pearson χ^2 test for small sample and sparse table.

Other Measures Based on Pearson χ^2 Test

Introduction
Two-way Table
Definition
2×2 Table
Association Tests
Pearson Test
LRT
Association Measures
Fisher Test
Other Concepts for 2×2 Tables
Ordered data
Stratified Contingency Tables
Matched Pairs
SAS

- Continuity adjusted χ^2: intend to correct Q^2 to approximate Fisher's Exact Test. It was proposed by Yates and not commonly used now.
- Phi coefficient: a measure of Pearson correlation based on the rank of observations for categorical data. It is only meaningful for 2×2 table or for ordinal data. For a 2×2 table,

$$\phi = \frac{n_{11}n_{22}-n_{12}n_{21}}{\sqrt{n_{1+}n_{2+}n_{+1}n_{+2}}}.$$

  Note $-1 \le \phi \le 1$. Furthermore, $\phi > 0$ implies the positive association between the 1st row and the 1st column. One can prove in 2×2 table $n\phi^2 = Q^2$.
- Contingency coefficient: $c = \sqrt{\frac{Q^2}{Q^2+n}}$.
- Cramer's V: $V = \sqrt{\frac{Q^2}{n\min(I-1,J-1)}}$.

Example: Fisher's Tea Drinker

Introduction
Two-way Table
Definition
2×2 Table
Association Tests
Pearson Test
LRT
Association Measures
Fisher Test
Other Concepts for 2×2 Tables
Ordered data
Stratified Contingency Tables
Matched Pairs
SAS

Ms. Bristol, a colleague of R.A. Fisher, claimed that when drinking tea she should distinguish whether milk or tea was added to the cup first and she preferred milk first. To test her claim, Fisher asked her to taste 8 cups of tea, 4 of which had milk added first and 4 of which had tea added first. She knew there were 4 cups of each type and had to predict which four had milk added first. The order of presenting the cups to her was randomized.

	Guess added first		
Added first	Mike	Tea	Total
Milk	3	1	4
Tea	1	3	4
Total	4	4	8

How would you test Ms. Bristol's claim? What do you find out?

Fisher Exact Test

When the sample size is small, Fisher exact test is preferred than χ^2 test. We illustrate this test in 2×2 table.

- A small-sample probability distribution for the cell counts is defined for the set of tables having the same row and column totals as the observed data.
- Under Poisson, binomial, or multinomial sampling assumptions for the cell counts, the distribution of such restricted set of tables with fixing the row and column totals is hypergeometric.

Fisher Exact Test

- Under H_0,

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

- Depends on H_α, the p-value is the sum of hypergeometric prob. for outcomes at least as favorable to the alternative hypothesis as the observed outcome.

Difference of Proportions

Consider the column variable, Y, is the response variable (such as success/failure) and the row variable, X, is the explanatory variable indicating two different groups. The interest may be comparing the proportion of success in two groups

$$\pi_{1|2} - \pi_{1|1} = \frac{\pi_{21}}{\pi_{2+}} - \frac{\pi_{11}}{\pi_{1+}},$$

which is between -1 and 1. For example, it can interpreted as the difference between the success rate of the 2nd group and that of the 1st group.

Relative Risk

Sometimes, the absolute difference of two success rates is not important but their relative difference or their ratio is interested. The relative risk is defined as the ratio of the success proportion of the first group to that of the second group:

$$\frac{\pi_{1|1}}{\pi_{1|2}} \geq 0.$$

Odds Ratio

- Odds: Prob. of success/Prob. of failure,

$$\frac{\pi_{1|1}}{\pi_{2|1}} = \frac{\pi_{11}/\pi_{1+}}{\pi_{12}/\pi_{1+}} = \pi_{11}/\pi_{12}$$

- Odds ratio: Odds of the 1st group/Odds of the 2nd group

$$\theta = \frac{\pi_{11}/\pi_{12}}{\pi_{21}/\pi_{22}} = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}} \geq 0.$$

- Remarks:
  - $\theta = 1$ is equivalent to "independence of X and Y".
  - $\frac{1}{\theta}$ is also an odds ratio.
  - For inference, it is more convenient to use symmetric measure $\log\theta$. ($\log\theta = 0 \Leftrightarrow$ Independence.)
  - When X and Y change, θ ($\log\theta$) doesn't change.
  - It is a natural parameter in logistic regression.
  - It is a good approximation of relative risk in the cross-sectional study when success is rare.

Sensitivity & Specificity

These measures are of interest when one is determining the efficacy of screening tests for various disease outcomes.

- Sensitivity: the true proportion of positive result when a subject has disease;
- Specificity: the true proportion of negative result when a subject doesnot have disease.

Example: Breast Cancer Diagnoses

	Diagnosis		
	+	-	Total
Cancer	82	18	100
No cancer	1	99	100
Total	83	107	200

Sensitivity:$\pi_{1|1} = 0.82$
Specificity:$\pi_{2|2} = 0.99$

Two-way Table for Ordered Data

Mantel-Haenszel χ^2 Test

When both response and explanatory variables are both ordinal, a positive or negative trend in the association is common. To test the monotone trend, Mantel (1963) proposed this test. If the test is significant, we say that increases in one variable are associated with increases (or decreases for negative relationships) in the other variable greater than would be expected by chance.

Assign scores to the row and column variable. The test statistics is defined as $M^2 = (n-1)r^2$, where r is the Pearson correlation between row variable and column variables. With large samples, M^2 has approximately χ^2_1 distribution.

When the association truly has a positive or negative trend, the test is more powerful than Pearson χ^2 and Likelihood ratio tests.

Scores of Ordinal Data

In SAS, one can use "SCORES=" option to specify the score type among: TABLE, RANK, RIDIT, and MODRIDIT scores (Spearman's ρ).
The default score is TABLE. For numerical variables, table scores are the values for the row and column levels. For character variables, table scores are the row and column numbers.
For a 2×2 table with character variables, using the table scores, we have

$$M^2 = \frac{(n_{11} - \mu_{11})^2}{Var(n_{11})}, \quad Var(n_{11}) = \frac{n_{1+}n_{+1}n_{2+}n_{+2}}{n^2(n-1)}.$$

Exercise: Derive the relationship between M^2 and Q^2 in a 2×2 table.

More Examples

- RANK scores are continuous version of the ranks. For $2 \times J$ table with RANK scores for Y, this Mantel-Haenszel test is equivalent to Wilcoxon test of nonparametric two sample test.
- RIDIT scores are standardized (by the sample size) version of the rank scores.
- MODRIDIT are modified ridit scores, which are the rank scores $/(n+1)$ respresenting the expected values of order statistics of $U(0,1)$.

Choice of Scores

The choice of scores is a subjective call. Different score choices will lead to different test statistics and so may produce different conclusion.

> *"Any set of scores gives a valid test, provided that they are constructed without consulting the results of the experiment. If the set of scores is poor, in that it badly distorts a numerical scale that really does underlie the ordered classification, the test will not be sensitive. The scores should therefore embody the best insight available about the way in which the classification was constructed and used." – Cochran (1954)*

Stratified Contingency Tables

Example: Mortality of Sunny City & Happy City

City	Death	Survival	Total
Sunny	1,475	98,525	100,000
Happy	1,125	98,875	100,000
Total	2,600	197,400	200,000

Which city have a higher mortality rate?

	Sunny City		Happy City	
Age	Death	Total	Death	Total
<25	25(1.00)	25,000	110(2.0)	55,000
25-44	50(1.25)	40,000	50(2.5)	20,000
45-64	200(10.00)	20,000	315(15.0)	21,000
≥ 65	1,200(80.00)	15,000	650(162.5)	4,000
Total	1,475(14.7)	100,000	1,125(11.25)	100,000

Which city have a higher mortality rate?

Simpson's Paradox

- This is an example of so called Simpson's Paradox.
- Even though the mortality rates of Happy city is higher than those of Sunny city at all age ranges, the overall mortality rate of Happy city is lower than that of Sunny city.
- In fact, this is because there are more young people in Happy city, who are at lower risk of death.
- Hence the 2×2 contingency table which combines all age groups is misleading.
- The independence test on the association of row and column variables for the combined table is also misleading.

Introduction
Two-way Table
Other Concepts for 2 × 2 Tables
Ordered data
Stratified Contingency Tables
Example: Mortality of Sunny City & Happy City
Simpson's Paradox
Cochran-Mantel-Haenszel (CMH) Test
Brewslow-Day Test
Matched Pairs
SAS Programs

Cochran-Mantel-Haenszel (CMH) Test

- We should control "age" in above example and construct a set of 2×2 tables stratifying on covariate "age" to analyze the mortality rates.
- The Mantel-Haenszel strata test for a set of 2×2 tables is based on summing the upper-left entries for all strata.
- For the kth table, the estimated expected counts under no association assumption is:

$$\mu_{11k} = \frac{n_{1+k}n_{+1k}}{n_{++k}},$$

and the variance of the counts in cell $(1,1)$ is:

$$Var(n_{11k}) = \frac{n_{1+k}n_{+1k}n_{2+k}n_{+2k}}{n_{++k}^2(n_{++k}-1)}.$$

Introduction
Two-way Table
Other Concepts for 2 × 2 Tables
Ordered data
Stratified Contingency Tables
Example: Mortality of Sunny City & Happy City
Simpson's Paradox
Cochran-Mantel-Haenszel (CMH) Test
Brewslow-Day Test
Matched Pairs
SAS

Cocharn-Mantel-Haenszel (CMH) Test

- The test statistic summarizes the information from K 2×2 stratified tables using:

$$CMH = \frac{[\sum_{k=1}^K (n_{11k} - \mu_{11k})]^2}{\sum_{k=1}^K Var(n_{11k})},$$

which has approximately χ^2 distribution with df=1.
- The CMH test removes the confounding variable by stratifying the other covariates and provides bigger power for detecting association in a random study. It does not assume homogeneity of odds ratio across strata.
- This test also needs large sample as in Mantel-Haenszel trend test. ($\sum_k n_{11k} > 30$)
- This CMH test can be generalized for a set of $I \times J$ stratified tables for nominal and ordinal data.

Introduction
Two-way Table
Other Concepts for 2 × 2 Tables
Ordered data
Stratified Contingency Tables
Example: Mortality of Sunny City & Happy City
Simpson's Paradox
Cochran-Mantel-Haenszel (CMH) Test
Brewslow-Day Test
Matched Pairs
SAS

Brewslow-Day Test

- The CMH test is inappropriate when the association varies dramatically among stratified tables.
- Test the homogeneity of odds ratio in a set of 2×2 tables.
- The test statistic has a similar form as Pearson χ^2 test but summing over all cells from all tables. It has approximately χ^2 distribution with df=$K-1$.
- When there is only one strata (one table), CMH=Mantel Haenszel Trend Test.

Matched Pairs

Example: Votes

For a poll of a random sample of 1600 Canadian citizens, 944 indicated approval of the Prime Minister's performance in office. Six month later, of these same 1600 voters, 880 indicated approval.

	Second Survey		
First Survey	Approval	Disapproval	Total
Approval	794	150	944
Disapproval	86	570	656
Total	880	720	1600

Note there are two samples in this data set and each sample has the same subject (or there is a natural match between subjects in two samples). The responses in the two samples are hence dependent.

McNemar's Test

The researchers are interested in comparing the proportions for approval in the first and second survey. (NOTE: They are dependent!)
The hypothesis on the marginal homogeneity :$H_0 : \pi_{1+} = \pi_{+1}$ is equivalent to the hypothesis on the equivalence of off-main-diagonal: $H_0 : \pi_{12} = \pi_{21}$.
Let $n^* = n_{12} + n_{21}$. Under the null hypothesis, each of n^* observation has equal probability to contribute to n_{12} or n_{21}. In another words, $n_{12} \sim Bin(n^*, 0.5)$. When n^* is large, the binomial distribution can be approximated by normal distribution with mean $0.5n^*$ and variance $0.25n^*$. Hence the standardized normal test statistic is:

$$z = \frac{n_{12} - 0.5n^*}{\sqrt{0.25n^*}} = \frac{n_{12} - n_{21}}{\sqrt{n_{12} + n_{21}}}.$$

McNemar's Test

- Under H_0, the square of this statistic, z^2, has approximately χ^2 distribution with df=1.
- This test for a comparison of two dependent proportions is called "McNemar's Test" (1947).
- When there are more than two categories, one can use Bowker's Symmetry Test (1948). To test $H_0 : \pi_{ij} = \pi_{ji}$, we calculate the test statistic

$$Q_B = \sum_{i<j} \frac{(n_{ij} - n_{ji})^2}{n_{ij} + n_{ji}},$$

  which approximately follows $\chi^2_{I(I-1)}$ under H_0.
- Cohen's Kappa Coefficient is a measure of such agreement.

$$\kappa = \frac{sum_i \pi_{ii} - \sum_i \pi_{i+}\pi_{+i}}{1 - \sum_i \pi_{i+}\pi_{+i}}.$$

References

- A. Agresti (2002), Categorical Data Analysis. Chap 1-3.
- Stokes, Davis, Koch (2000), Categorical Data Analysis Using the SAS Program, Chap 2-6.

SAS Programs

PROC FREQ;

- WEIGHT;
  Use it when data are already counts.
- TABLES X*Y*Z; (TABLE X*Y*Z;)
  The two rightmost variables will be displayed in the table as rows and columns, while the others are strata;
- RISKDIFF;
  Report the difference of proportion;
- MEASURES;
  Report the odds ratio and relative risk;
- NOCOL; NOPCT; Suppress the column percentage or no percentage at all.
- AGREE;
  It is for matched pairs and includes McNemar's test.
- CMH;
  Cochran-Mantel-Haenszel tests.

Reading Data

- Input statement in DATA step
  Reading: Chap 12, Page 353-370.
- External files:
  Reading: Chap 13, Page 376-400.